EUCLID

EdUcational Curriculum for the usage of LInked Data

Deliverable 3.4.2
Exploitation and dissemination plan

Coordinator: Elena Simperl
With contributions from: John Domingue, Peter Haase, Maria Maleshkova, Marin Dimitrov

Quality Assessor: Maria Maleshkova

Editor:	Elena Simperl, STI-R
Deliverable nature:	R
Dissemination level:	PU
Contractual delivery date:	30.04.2014
Actual delivery date:	16.05.2014
Version:	1.0
Total number of pages:	29
Keywords:	Exploitation, commercialization, dissemination, communication, community building, sustainability

Executive summary

Following the strategy laid out in D3.4.1 this document analyses the impact of the dissemination and exploitation activities performed throughout the duration of the project and gives an outline of the plans the consortium has defined to continue such activities after the project's end in order to ensure the sustainability and full uptake of EUCLID's major achievements. Most importantly, this deliverable is concerned with the way the consortium will make use of the results of the project, be that for educational purposes in academia, or as part of commercial products and services offered by the core and associate industry partners.

In EUCLID dissemination and community engagement activities are implemented in WP2. As such, three deliverables in that work package (D2.2.3, D2.3.3, and D2.1.5), all due M24, provide a comprehensive overview and analysis of the activities the project has pursued online and offline to ensure that its outcomes, most importantly the suite of learning materials produced in the last two years, are promoted to the right audiences and widely used throughout Europe and worldwide in higher education and commercial training offerings. From a dissemination and community engagement point of view the focus of this deliverable will be on providing an executive summary of the strategy and plans of the project vs. the actual implementation, and a set of lessons learned which we hope will prove useful for future research projects of this kind. In particular, we will discuss our approach to use social media to reach out to developers and data practitioners interested in Linked (Open) Data. A second component of this deliverable is concerned with the exploitation of the results of the project, most importantly with the question of how the curriculum will be used in different kinds of training activities, and the effects it will have on the development of the professional training market in this field. We have reviewed the training market and potential competitors and engaged in a SWOT analysis to identify those areas in which EUCLID outcomes offer a competitive advantage to commercial companies involved in the project.

Document information

IST Project Number	FP7 - 296229	**Acronym**	EUCLID
Full Title	Educational curriculum for the usage of Linked Data		
Project URL	http://www.euclid-project.eu/		
Document URL			
EU Project Officer			

Deliverable	**Number**	3.4.2	**Title**	Exploitation and dissemination plan
Work Package	**Number**	3	**Title**	Management

Date of Delivery	**Contractual**	M24	**Actual**	M24
Status	2.0		final X	
Nature	prototype □ report X dissemination □			
Dissemination level	public X consortium □			

Authors (Partner)	STI-R			
Responsible Author	**Name**	Elena Simperl	**E-mail**	elena.simperl@gmail.com
	Partner	STI-R	**Phone**	

Abstract (for dissemination)	Following the strategy laid out in D3.4.1 this document analyses the impact of the dissemination and exploitation activities performed throughout the duration of the project and gives an outline of the plans the consortium has defined to continue such activities after the project's end in order to ensure the sustainability and full uptake of EUCLID's major achievements. Most importantly, this deliverable is concerned with the way the consortium will make use of the results of the project, be that for educational purposes in academia, or as part of commercial products and services offered by the core and associate industry partners.
Keywords	Exploitation, commercialization, dissemination, communication, community building, sustainability

Table of contents

List of tables and figures

Abbreviations

DL – Description Logic
FOAF – Friend of a Friend
HTTP – Hypertext Transfer Protocol
KIT – Karlsruhe Institute of Technology
KMi – Knowledge Media Institute
LD - Linked Data
OA – Ontotext AD
ONTO – Ontotext
OU – Open University
OWL – Ontology Web Language
OWL-S – OWL for Services/ OWL-based Web Service Ontology (formerly DAML-S)
RDF/S – Resource Description Framework / Schema
SPARQL – SPARQL Protocol and RDF Query Language
URI – Uniform Resource Identifier
URL – Uniform Resource Locator
WP – Work Package
XML - Extensible Mark-up Language

1 Introduction

Following the strategy laid out in D3.4.1 this document analyses the impact of the dissemination and exploitation activities performed throughout the duration of the project and gives an outline of the plans the consortium has defined to continue such activities after the project's end in order to ensure the sustainability and full uptake of EUCLID's major achievements. Most importantly, this deliverable is concerned with the way the consortium will make use of the results of the project, be that for educational purposes in academia, or as part of commercial products and services offered by the core and associate industry partners.

In EUCLID dissemination and community engagement activities are implemented in WP2. As such, three deliverables in that work package (D2.2.3, D2.3.3, and D2.1.5), all due M24, provide a comprehensive overview and analysis of the activities the project has pursued online and offline to ensure that its outcomes, most importantly the suite of learning materials produced in the last two years, are promoted to the right audiences and widely used throughout Europe and worldwide in higher education and commercial training offerings. From a dissemination and community engagement point of view the focus of this deliverable will be on providing an executive summary of the strategy and plans of the project vs. the actual implementation, and a set of lessons learned which we hope will prove useful for future research projects of this kind. In particular, we will discuss our approach to use social media to reach out to developers and data practitioners interested in Linked (Open) Data. A second component of this deliverable is concerned with the exploitation of the results of the project, most importantly with the question of how the curriculum will be used in different kinds of training activities, and the effects it will have on the development of the professional training market in this field. We have reviewed the training market and potential competitors and engaged in a SWOT analysis to identify those areas in which EUCLID outcomes offer a competitive advantage to commercial companies involved in the project.

EUCLID's dissemination and community engagement efforts have been continuous and wide-ranging. We have used a variety of broadcast and responsive channels to make sure the main outcomes of the project, in particular its learning materials, reach the right audiences. As a result, a large part of the Linked Data community, in particular educators, but also practitioners associated Linked Data training with EUCLID. We will exploit this advantageous position to update and expand the curriculum by aligning with other projects such as FORGE at the OU, and the Data Science MSc to be launched at Southampton in 2015. In addition, the training resources are used and will be maintained by industrial partners as part of their training offers (Ontotext, STI via certification and summer school, and fluidOps).

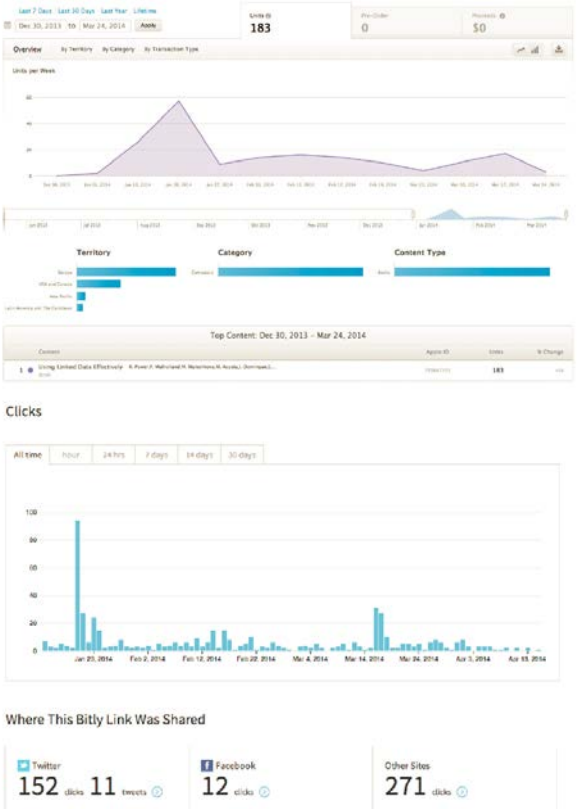
Commercial core and associate partners have outlined a clear and convincing pathway for commercial exploitation early in the project. The curriculum has been defined with an industrial audience in mind and learning materials have been used in corporate training by Ontotext and others from the outset of the project. An analysis of the corporate training market and existing professional programs on Linked Data and semantic technologies has revealed the strengths and weaknesses of EUCLID from an exploitation point of view. Extrapolating from the positive resonance received during the project paired with the vested interest of all partners in education and Linked Data technologies we are confident that the competitive advantage and reputation that EUCLID has established will be maintained and amplified through the consortium partners and their future collaborations.

2 Dissemination and community building

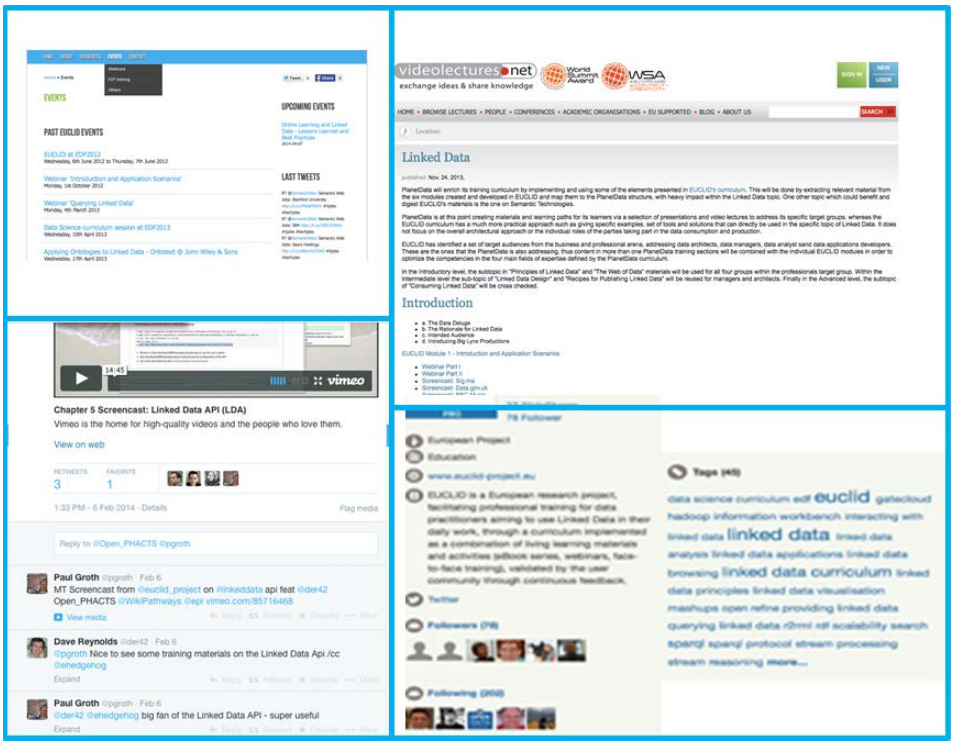
This section will give a high-level, unified view of the dissemination and community building strategy of the project. The associated instruments and communication and engagement channels are part of WPs 1 and 2. Channels such as vimeo, SlideShare, and iTunes U are used for hosting and promoting specific types of learning materials. The online presence of the project, including project logo, Web site, project fact sheet, slide deck, flyers, a Twitter account, and a discussion group on LinkedIn are used to advertise the materials, as well as training and other types of events in which members of EUCLID are participating. Most of these activities have been reported and discussed in detail in deliverables in the other work packages as explained in Section 2.1. In the following we will give an overview of the main focus of each of these deliverables and how they are interrelated, followed by a summary of general project communication instruments and marketing materials, and a discussion of the most important lessons we have learned in the process. These insights have been documented in several publications targeting the eLearning and the data science education communities (see Appendix I).

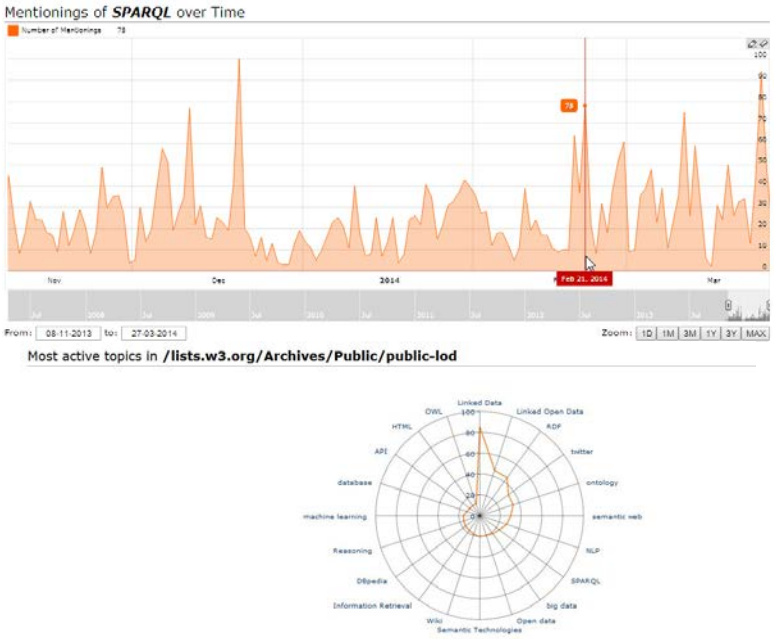
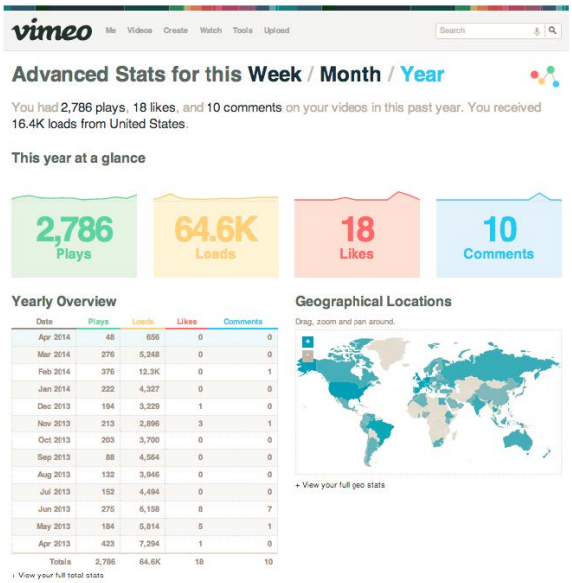
2.1 Summary of WP1 and WP2 deliverables due M24

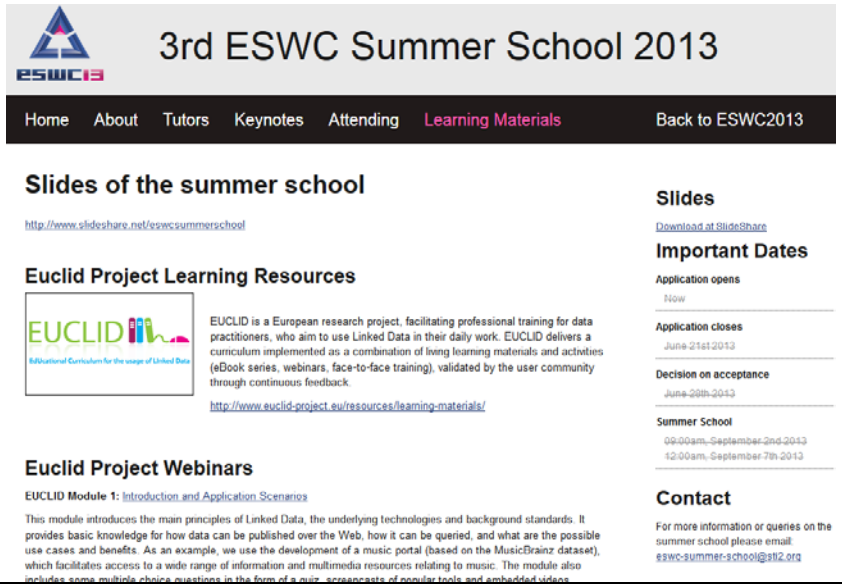
Table 1 Report and analysis of dissemination and community building activities in other M24 deliverables

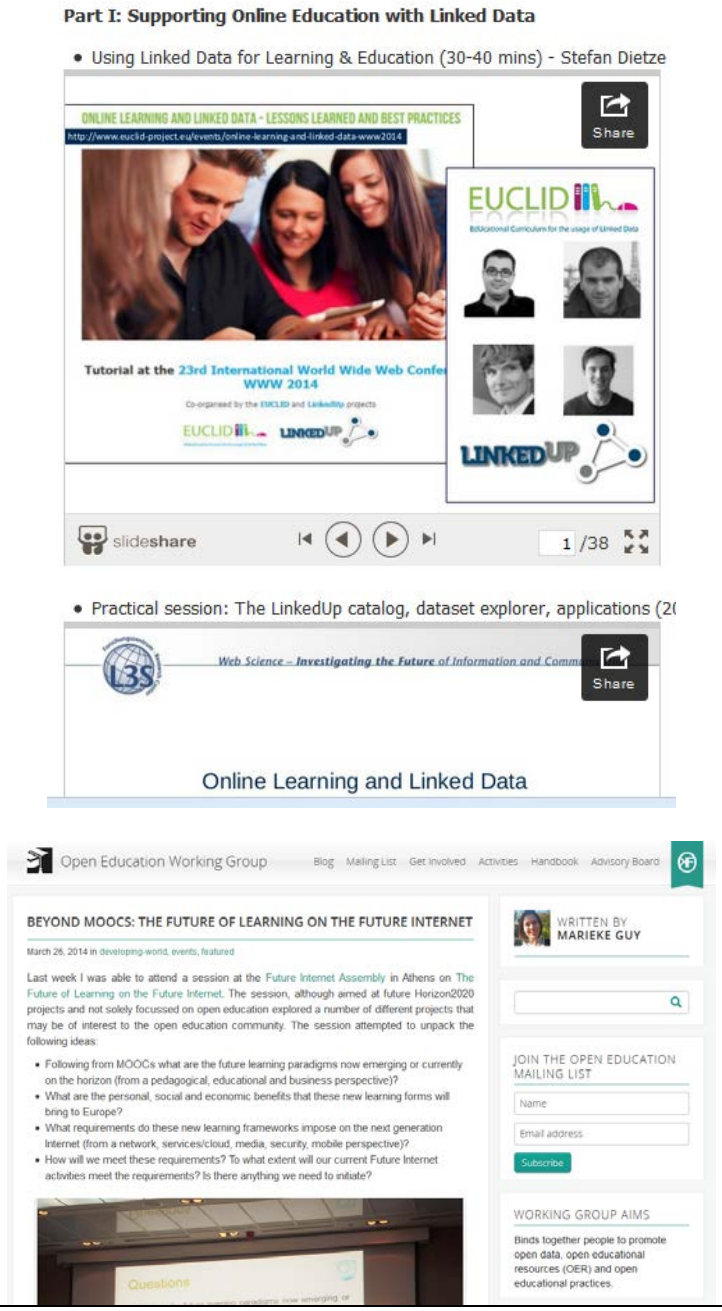
Deliverable	Summary	Dissemination and community building
D1.2.3 Final educational eBook delivery report	*Gives an overview of the learning materials produced in the second half of the project, which have resulted in an eBook consisting of six chapters, one for each module of the curriculum, which has been published on iTunes U.*	Measure the impact of the eBook on the community via download statistics and the bitly online service[1] which has been used to bookmark the download page of the EUCLID eBook in order to monitor its visitor statistics. This covers the sites and social platforms on which the link was shared together with a geographic distribution of the visitors. 

[1] https://bitly.com/L6eOdv+

D2.1.5 Final online community engagement report	*Reports on all online activities undertaken by the project, including broadcast and responsive channels used to host and promote the learning materials, and the monitoring platform built on top of technologies from Ontotext and fluidOps*	We have analysed the impact of the following channels and outlined the cornerstones of our sustainability plan. This includes: • The project Web site • Twitter and LinkedIn • SlideShare, vimeo, Videolectures.net, and iTunes U in their promotional capacity • Dissemination via mailing lists For each of these channels we present key performance indicators that measure the impact of the results of the project on the Linked Data community, including number of downloads, number of tweets, number of followers, as well as some information about the location of the users.  **Monitoring platform** The monitoring platform uses the same technologies to gain insight into the topics the community is interested in. It draws upon information from all channels used by EUCLID and a representative sample of the community that uses these channels to identify themes and trends the project should take into account in its learning materials and training. The tool can be configured for similar purposes and used by community managers and impact work packages in similar research projects. The technology in itself forms the basis for a semantically enabled social media analytics dashboard that may have a much wider range of applications that engagement in research projects.

		
D2.2.3 Final webinar report	*The deliverable gives an overview of the webinars held by the project in the second year and analyses their impact.*	The deliverable follows up on a similar report produced in the first year of the project and presents the 4th, 5th, and 6th webinars organized by EUCLID. In addition, we discuss key performance indicators measured via the platforms that we use to broadcast the live webinar and the recordings, including statistics provided by vimeo, SlideShare, OU Stadium and LiveStream.  *Figure 10: Vimeo general statistics*

		<table><tr><th>Webinar</th><th>Title</th><th>Broadcasting Date</th><th>Live Broadcasting Platform</th><th>Presenter(s)</th><th>Live Audience</th></tr><tr><td>1</td><td>"Linked Data: Introduction and Application Scenarios"</td><td>01st October 2012</td><td>OU Stadium</td><td>Barry Norton (OA)</td><td>35</td></tr><tr><td>2</td><td>"Querying Linked Data"</td><td>04th March 2013</td><td>OU Stadium</td><td>Barry Norton (OA)</td><td>91</td></tr><tr><td>3</td><td>"Providing Linked Data"</td><td>22nd April 2013</td><td>OU Stadium</td><td>Barry Norton (OA)</td><td>114</td></tr><tr><td>4</td><td>"Interaction with Linked Data"</td><td>10th June 2013</td><td>OU Stadium and LiveStream</td><td>Barry Norton (OA) and Michael Meier (fluidOps)</td><td>579</td></tr><tr><td>5</td><td>"Building Linked Data Applications"</td><td>14th October 2013</td><td>OU Stadium and LiveStream</td><td>Christoph Pinkel (fluidOps)</td><td>40</td></tr><tr><td>6</td><td>"Scaling up Linked Data"</td><td>19th December 2013</td><td>OU Stadium and LiveStream</td><td>Marin Dimitrov (OA)</td><td>46</td></tr></table>
D2.3.3 Final real-world community engagement report	*This deliverable refers to activities such as f2f training, tutorials, publications, and community building events.*	The deliverable gives a summary of all f2f training sessions undertaken by project partners or by other parties using EUCLID materials. A major activity to which the project contributed was the ESWC summer school, which was largely based on EUCLID materials. In addition, we report on tutorials and community sessions at ESWC2014 and FIA 2014. 



2.2 Project communication and marketing materials

This section gives a high-level overview of the communication strategy of the project. As noted earlier, some PR channels such as the project Web site and our Twitter account have been extensively used for dissemination and community engagement and as such their development has been analysed in deliverables in WP2. Here we will hence focus on the basic marketing instruments we have created, which have not been described in other deliverables.

Part of the marketing toolkit are the project logo, a slide deck, a document template, a flyer, the project factsheet, and a poster. They accompanied every single dissemination and engagement activity undertaken throughout the project. The slide deck was used both for project-specific presentations and for the learning materials. It features all core and associate partners, as well as all channels used for dissemination and communication strategy and the types of learning materials produced. The document template was used for internal communication and for the contractual deliverables. The poster was used at the Open CourseWare Conference (OCWC, see also Appendix I) and at the regular project networking events organized yearly at the ESWC Conference. Flyers were generously

distributed at all events attended by project members to advertise the project and its curriculum. The design of the flyer and of the poster has been updated continuously throughout the project in order to give an accurate account of its achievements.

Figures 1 to 7 should give an impression of the content, and the look-and-feel of the marketing toolkit.

Figure 1 Project logo

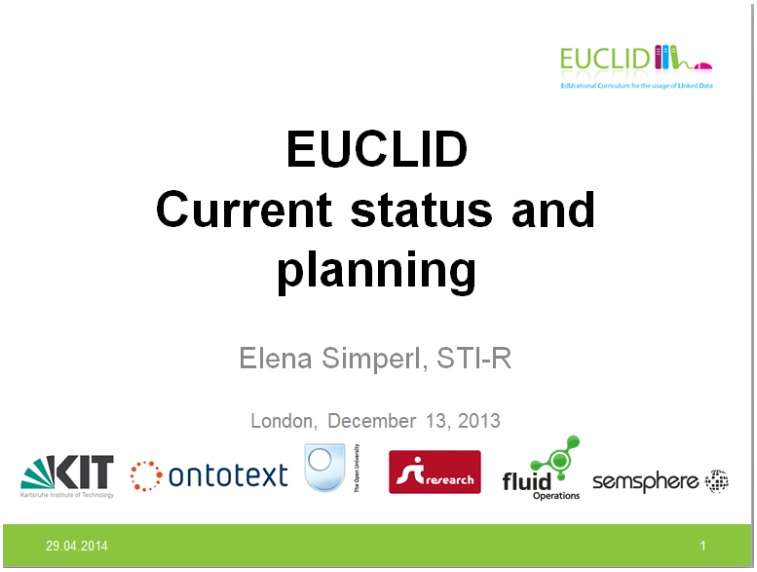
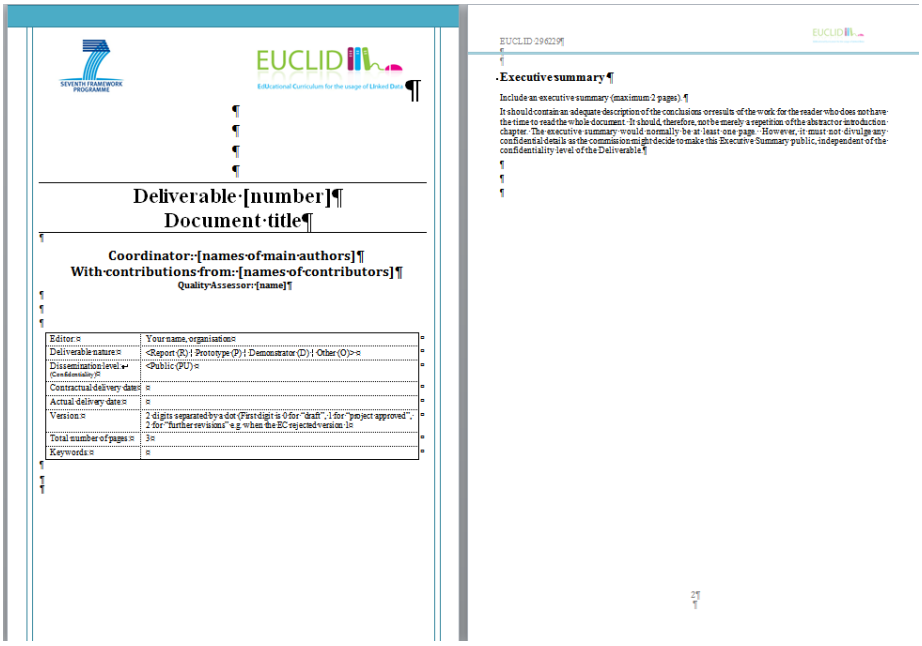


Figure 2 Front cover of the project slide deck

Figure 3 Back cover of the project slide deck



Deliverable·[number]¶
Document·title¶

Coordinator:·[names·of·main·authors]¶
With·contributions·from:·[names·of·contributors]¶
Quality·Assessor:·[name]¶

Editor:¤	Your name, organisation¤
Deliverable nature:¤	<Report (R) | Prototype (P) | Demonstrator (D) | Other (O)>¤
Dissemination level:↵ Confidentiality (P	<Public (PU)¤
Contractual delivery date:¤	¤
Actual delivery date:¤	¤
Version:¤	2 digits separated by a dot (First digit is 0 for "draft", 1 for "project approved", 2 for "further revisions" e.g. when the EC rejected version 1)¤
Total number of pages:¤	3¤
Keywords:¤	¤

EUCLID 296229¶

· Executive summary ¶

Include an executive summary (maximum 2 pages).¶

It should contain an adequate description of the conclusions or results of the work for the reader who does not have the time to read the whole document. It should, therefore, not be merely a repetition of the abstract or introduction chapter. The executive summary would normally be at least one page. However, it must not divulge any confidential details as the commission might decide to make this Executive Summary public, independent of the confidentiality level of the Deliverable.¶

Figure 4 Document template

Figure 5 Project poster

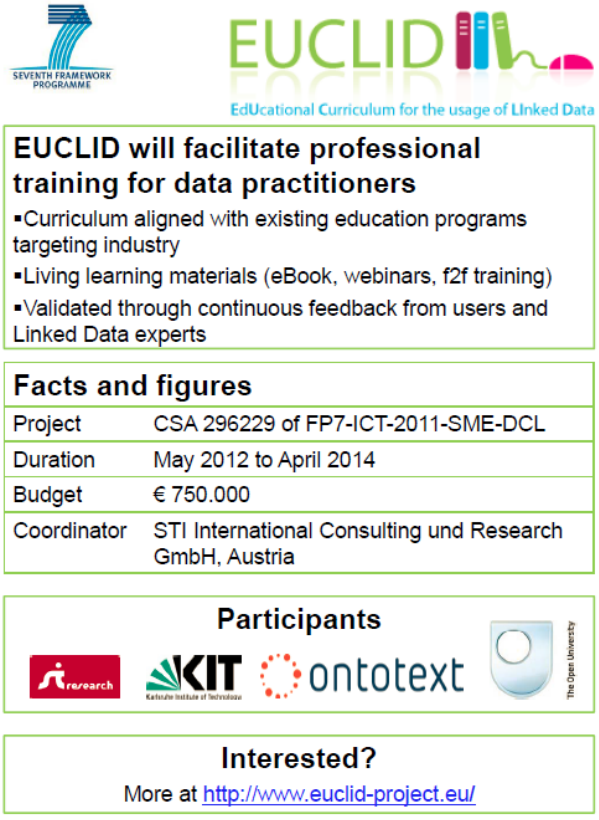


Figure 6 Project factsheet

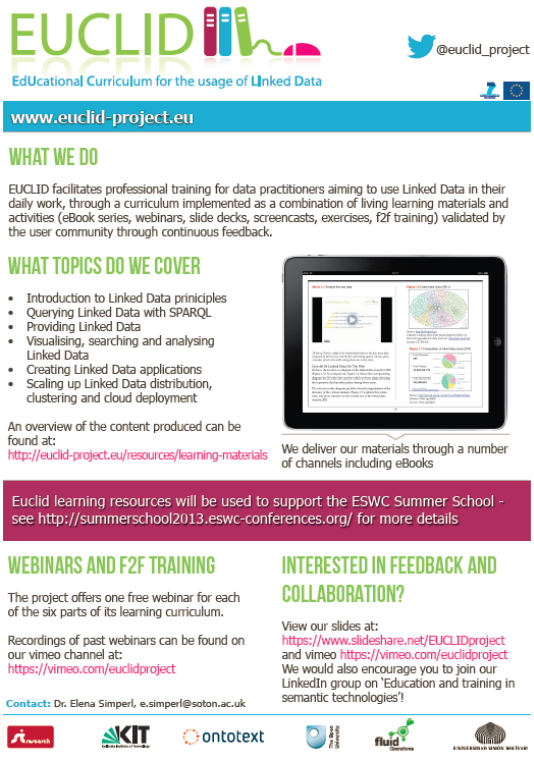


Figure 7 Project flyer

2.3 Lessons learned

EUCLID gave us the opportunity to gain comprehensive insight into the use of online and social media channels for scientific dissemination, communication, and community engagement. The project recognized early on the need for a sustained effort in this regard already at proposal writing time, as stated in the Technical Annex of the Grant Agreement.

> *"The work plan of EUCLID consists of two complementary parts: one focusing on the production of educational and training content and its delivery over the most effective and relevant channels, and a second, equally important part, through which the learning materials are presented, adjusted and improved* ***through feedback and engagement with their intended audience, and the Linked Data community in general****. […]* ***the entire Linking (Open) Data movement stands as proof of how adjustments in focus with the technology can lead to huge shifts in interest****. Furthermore,* ***the impact and potential of changes in focus can be seen to be most readily judged by monitoring applications; rather than, by simply listening to academics and pure technologists****. For this reason it is imperative to* ***develop visibility for the project within the community****; this is necessary also to ensure the successful delivery of the project results. In this community* ***attempting to replace ongoing activities by building overly-ambitious portals and hosting (academic-style) workshops will not suffice****. What is needed is* ***genuine engagement****.* ***Existing mailing lists, IRC chats, hack days and workshops constitute the very fabric of the movement****. Only by becoming a beneficial presence, and making a positive contribution, will this aim of engagement be realized.* ***Project efforts, both online and in the real world, must be carefully aligned with existing activities and their drivers****. This also requires appeal at the level of the developer, the IT manager and even the hacker."*

The project has successfully implemented this vision. Two years after the start of the project, EUCLID is the default venue for high-quality, application-oriented training materials.

This has been achieved through sustained community engagement via multiple channels, most importantly Twitter, SlideShare, vimeo, live webinars, mailing lists, iTunes U, and project Web site. Our learning materials, though concentrating on a type of technologies, have not been biased towards specific product lines, but have tried to offer a well-balanced overview of those tools that can be considered established. We have actively sought collaboration with several other research projects (OpenPhacts, TrendMiner, PlanetData) and tool developers as a source for learning resources (slides, screencasts etc.). We have embraced different training delivery channels and invested substantial effort into testing the materials through carefully planned webinars. Rehearsing the webinar proved to be extremely beneficial for the learning experience. Advertising the curriculum followed a similar pattern for each module: each webinar was announced several months in advance and reminders were sent twice after the initial announcement (one week before and the day before the event). The webinar recordings were made available closely after the live event to use the positive momentum. We used primarily three channels: mailing lists, LinkedIn, and Twitter for this purpose. The same channels were used to advertise the release of each module as slide decks or book chapter. In addition, we tweeted several times a week, focusing on topics related to eLearning, data management, Linked Data, MOOCs in order to keep the interest of the followers alive and build a community around Linked Data education.

A monitoring platform was developed to ensure the project stays at the pulse of the topical demands of the community. It is, however, fair to point out that the initial ambitions of the project to engage with Q&A sites and mailing lists and offer basic Linked Data literacy services at scale remained largely unrealized. The reason for this shift in priorities was the interest which we encountered from the community in our webinars and other forms of training delivery. Originally conceived as means to test-and-trial the learning materials, they soon developed into a core training delivery activity which required significant resources; given the extremely positive resonance we had with this activity, we are confident this change had a beneficial effect on the impact of the project.

Our materials are used and referenced worldwide in academia and professional training in several countries (see also D1.2.3 and D2.2.3). A combination of personal commitment of the project team, high-quality materials, and very effective promotion channels (iTunes U, SlideShare) was the recipe to achieve this.

We have documented our efforts in three publications, which are attached to this deliverable in Appendix I:

- Alexander Mikroyannidis et al. Developing a Curriculum of Open Educational Resources for Linked Data. 10th annual OpenCourseWare Consortium Global Conference (OCWC 2014), Ljubljana, Slovenia.
    - This paper summarizes the process model followed by the project to develop its curriculum. It is targeted at the eLearning community interested in delivering open course materials.
- Alexander Mikroyannidis, John Domingues, and Elena Simperl. Raising the Stakes in Linked Data Education. In ERCIM News, Special Theme Linked Open Data, January 2014.
    - This short paper is meant as an executive summary of the project for the Linked Data practitioners community.
- John Domingue, Mathieu d'Aquin, Elena Simperl, and Alexander Mikroyannidis. The Web of Data: Bridging the Skills Gap. IEEE Intelligent Systems, 29(1):70-74. 2014.
    - This article was published in the Web Science column of IEEE Intelligent Systems. Hence, it's audience are mainly educators in the Linked Data, semantic technologies, and Artificial Intelligence fields interested in adopting a state-of-the-art approach to curriculum development and promotion. Among other topics, the article features a list of best practices for the design and development of educational curricula for professionals, as well as a proposal for using open access and Linked Data as basic principles and support technologies to collect and learn from student feedback.

To conclude we believe that a combination of highly effective platforms for content hosting and training delivery paired with sustained engagement via Twitter and LinkedIn, monitoring technology, and personal commitment have been instrumental for the popularity of EUCLID. However, as every marketing expert would happily confirm, the most important ingredient of our success was the product itself, a collection of rich, multi-modal learning resources targeting the very real needs of the Linked Data practitioners' community.

3 Commercial exploitation

3.1 Market analysis

Corporate training market

According to the "2013 Training Industry Report" by one of the leading market watch organizations, Training Magazine, the total spending on corporate training related products and services in the US only is estimated to $55 billion (Figure 8).

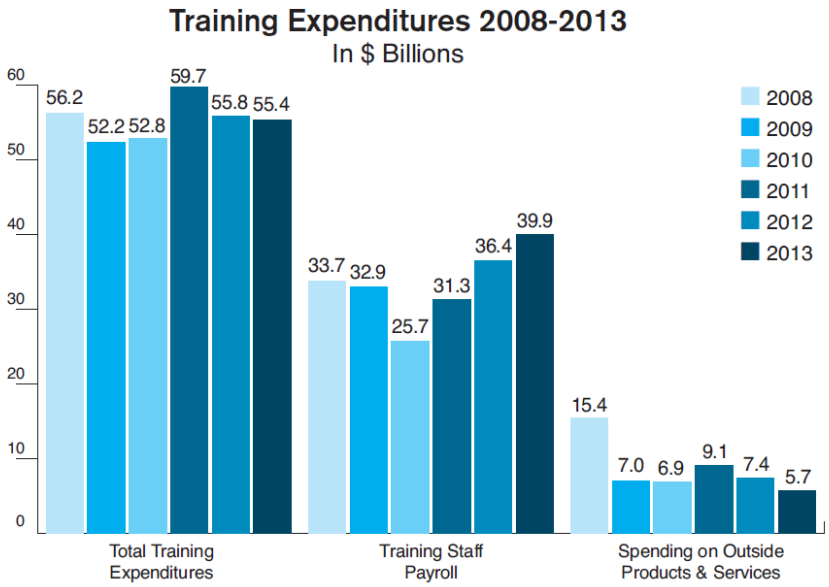


Figure 8 Total corporate training expenditure in the US, 2013 (by Training Magazine)

The average spending on training services and products per company employee in the US is estimated to be $881 in 2013, with SMEs spending a significant share of this budget on training (Figure 9)

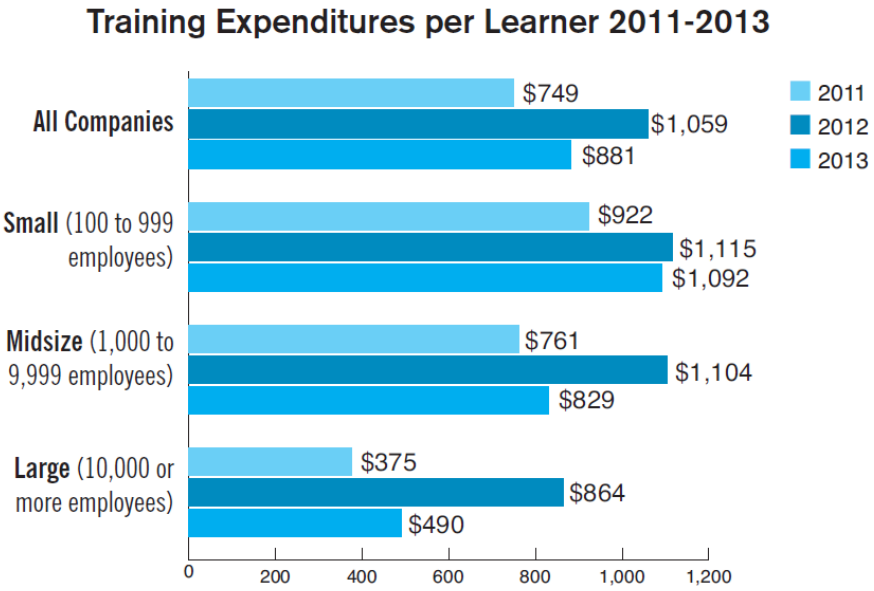


Figure 9 Training per US employee, 2013 (by Training Magazine)

These figures are indicative for the European market as well; cumulatively the EU economy has a similar size (by GDP) as the US, with an even more pronounced prevalence towards SMEs.

Linked Data and semantic technologies training

Reliably estimating the relative size of Linked Data and semantic technologies-related training and services market remains a challenging task. This is due to the fact that this class of IT solutions is not yet a core part of the annual market reports by major analyst companies such as Gartner and Forrester. In the absence of such external source, our analysis extrapolates from the experience of the commercial partners and collaborators of EUCLID.
Training is still a relatively small revenue stream, accounting for less than 5% of the total company revenue of a typical technology provider in the Linked Data area. However, it is a very important activity due to two reasons:

- Training is required by big customers who are just buying into semantic and Linked Data technologies and lack in-house expertise.
- Training provides a new channel to promote company technology and products to new potential customers and adopters.

3.2 Competitive landscape

In this section we give an overview of other training providers in the space. We focus on those organizations whose services target professionals, and less on higher-education institutions or collaborative research projects.

Cambridge Semantics / Semantic University

Cambridge Semantics[2] is a US company providing products and solutions for semantic and smart data management. Its *Anzo Enterprise*[3] suite features capabilities for unified data access, management, and analysis.
In addition to its commercial Anzo product line, Cambridge Semantics started an activity called *Semantic University*[4], which aims to produce educational materials related to Semantic Web and Linked Data. The educational content is currently structured in 34 very short lessons grouped in 7 main tracks:

- Understanding Semantic Technologies[5]
- Semantic Technologies Compared[6]
- Semantic Technologies Applied[7]
- Learn RDF[8]
- Learn OWL and RDFS[9]
- Learn SPARQL[10]
- Semantic Web Design Patterns[11]

Open Data Institute (ODI)

The Open Data Institute[12] is a UK non-profit independent organization aiming at promoting Open Data adoption and innovation. At present ODI provides various paid courses focusing on Open Data, but somewhat related to Linked Data issues as well:

- Open Data in a Day[13]
- Open Data in Practice[14]

---

[2] http://www.cambridgesemantics.com/
[3] http://www.cambridgesemantics.com/products/anzo-enterprise
[4] http://www.cambridgesemantics.com/bg/semantic-university
[5] http://www.cambridgesemantics.com/semantic-university/getting-started
[6] http://www.cambridgesemantics.com/semantic-university/comparing-semantic-technologies
[7] http://www.cambridgesemantics.com/semantic-university/semantic-technologies-applied
[8] http://www.cambridgesemantics.com/semantic-university/learn-rdf
[9] http://www.cambridgesemantics.com/semantic-university/learn-owl-and-rdfs
[10] http://www.cambridgesemantics.com/semantic-university/learn-sparql
[11] http://www.cambridgesemantics.com/semantic-university/learn-semantic-web-design-patterns
[12] http://theodi.org/
[13] http://theodi.org/courses/open-data-day

- Open Data Technologies[15]
- Open Data, Law & Licensing[16]
- Finding Stories in Open Data[17]

Additionally, the ODI provides series of *Friday Lunchtimes Lectures*[18] covering Open & Linked Data aspects. The lectures are free to attend and are also available on iTunes.

LOD2

LOD2[19] is a large EC-funded FP7 project that aims to provide an integrated software stack for managing the Linked Data lifecycle. While educational materials were not the main focus of the project, it has managed to deliver 22 free webinars[20] so far covering various Linked Data related topics, from very general to LOD software stack specific ones.

Open Data Support

Open Data Support[21] is a three years European project targeting potential Open Data publishers. Besides IT and data provisioning and enrichment services, they offer training in the areas of Linked and Open Data. Open Data Support reuses EUCLID materials.

TopQuadrant

TopQuadrant[22] is a US supplier of products and solutions for Enterprise Information Integration with semantic technologies. In addition to its *TopBraid EVN* and *TopBraid Composer* product lines the company offers an "Introduction to Semantic Web Technology" training course[23] covering various aspects related to ontology modelling, RDF/S and OWL, and SPARQL.

TenForce

TenForce[24] is a Belgian company providing consulting services in the area of Semantic technologies. The company also provides semantic technologies-related training, organized in three modules:

- Introduction to Semantic Technology[25]
- Modelling for Semantic Technology Applications[26]
- Programming for Semantic Technology Applications[27]

Franz

Franz[28] is a US company providing products for semantic data management. In addition to that the company offers a 3-day training course[29] which is a mixture of general Semantic Web topics and product specific (AllegroGraph database) training.

---

[14] http://theodi.org/courses/open-data-practice
[15] http://theodi.org/courses/applying-linked-open-data
[16] http://theodi.org/courses/open-data-law-and-licensing
[17] http://theodi.org/courses/finding-stories-in-open-data
[18] http://theodi.org/lunchtime-lectures
[19] http://lod2.eu/
[20] http://lod2.eu/BlogPost/webinar-series
[21] https://joinup.ec.europa.eu/community/ods/description
[22] http://www.topquadrant.com/
[23] http://www.topquadrant.com/training/
[24] http://tenforce.com/
[25] http://www.tenforce.com/training-course1
[26] http://www.tenforce.com/training-course2
[27] http://www.tenforce.com/training-course3
[28] http://franz.com/
[29] http://franz.com/ps/services/classes/

OpenLink

OpenLink[30] is a US company providing products for RDF and Linked Data management. The company also provides a training course[31] covering topics such as Linked Data & Semantic Web, Data- and Knowledge Base Technologies.

Amtera

Amtera[32] is a Brazilian company providing products for Knowledge Management and consulting for Semantic Technologies and Big Data. In addition to its product line, Amtera is offering Sematic Web related training[33] covering the following topics:

- Web 3.0 paradigm
- Ontology engineering
- Natural language processing
- Developing Enterprise Intelligent Systems

3.3 SWOT Analysis for EUCLID

Table 2 provides a summary of the SWOT[34] analysis for EUCLID, based on the competitive market landscape analysis from Sections 3.1 and 3.2.

Table 2 SWOT analysis for EUCLID

	Positive	Negative
Internal	STRENGTHS • Strong consortium comprised of leading experts in the area of Semantic Web and Linked Data • Open content (CC license) which is free for anyone to use and extend • Good industry contacts of some consortium members with leading media & publishing organisations in Europe & US	WEAKNESSES • Consortium partners may not be able to easily collaborate on extending the EUCLID content after the project end (due to differences in priorities) • Lack of US partners who can establish links with US enterprises
External	OPPORTUNITIES • Growing demand for Linked Data related training & consulting services • Growing popularity of MOOCs	THREATS • Growing competition from companies & non-profit organisations offering Linked Data related training & consulting

[30] http://www.openlinksw.com/
[31] http://ps.openlinksw.com/training/
[32] http://www.amtera.com.br
[33] http://www.amtera.com.br/?page_id=48&lang=en
[34] Strengths, Weaknesses, Opportunities, and Threats: http://en.wikipedia.org/wiki/SWOT_analysis

3.4 EUCLID exploitation

Exploitable results

EUCLID has delivered three main classes of exploitable results, which provide various exploitation & monetization options for consortium members after the project end.

A proven methodology for multi-channel delivery of high-quality educational content

The established process and lessons learned while building and delivering the Linked Data educational content via different channels (website, eBooks, webinars) can be reused and adapted for building and delivering educational content in other emerging domains such as Open Data, Big Data or Data Science.

High-quality training content for data practitioners

Both commercial partners of the consortium already offer training services related to Linked Data and Semantic Web and the educational materials developed within EUCLID can be (and in some cases, have already been) reused to extend the commercial training content with new topics.

Semantically enabled topic and community monitoring technology

The EUCLID monitoring platform uses existing Linked Data sets, as well as technology by Ontotext and fluidOps to collect, store, enrich, analyze, and visualize activities of a community of interest on specific channels. The technology is configurable to arbitrary communities and topics, and will be exploited by commercial partners in industry projects which focus on data analysis, information extraction, and social media.

Ontotext

Ontotext is a leading provider of Semantic Technology products and solutions. In addition to its product lines, the company offers a 3-day training course[35] related to Linked Data and Semantic Web, covering the following topics:

- Introduction to the Semantic Web
- RDF, RDFS and OWL knowledge representation languages
- Reasoning
- Ontology design
- Querying RDF data with SPARQL
- Linked Data publishing
- Introduction to OWLIM
- Advanced OWLIM features
- OWLIM optimisation and performance tuning
- OWLIM Enterprise Cluster
- OWLIM administration

Ontotext's training services target two main customer segments:

- Existing Ontotext customers who need on-boarding with Linked Data and Semantic Web topics
- Potential customers / general public, which need general Linked Data and Semantic Web introduction but may as well be interested in Ontotext products

[35] http://www.ontotext.com/training

Ontotext delivers its training via three main channels: on-premise trainings or webinars for customers; and open training sessions co-located with big events: Semantic Technology Conference (USA & UK), Semantic Days (Norway), Semantic Web meetups (USA & UK).

The directly exploitable artefact of the EUCLID project for Ontotext is the high-quality training content which has already been partially adapted and incorporated within various trainings by Ontotext in the last 12 months. In particular, the following parts of the Ontotext training curriculum have already partially benefitted from EUCLID created content:

- RDF, RDFS and OWL knowledge representation languages
- Reasoning
- Querying RDF data with SPARQL
- Linked Data publishing

This content has already been presented to big customers from the media & publishing or cultural heritage and digital libraries domains, in particular[36]:

- BBC (UK)
- British Museum (UK)
- UK Parliament
- One of the world's leading financial news organizations (UK)
- One of the leading monthly magazines focused on business & finance (UK)
- A global academic publishing company (US)
- One of the biggest & most visited museums in the US

fluidOps

fluidOps develops the Information Workbench, a platform for enterprise applications building on Linked Data technologies. Comprehensive, high-quality training for their customers, partners and application developers is essential for the commercial success of our business. Therefore fluidOps participates in the EUCLID project to

- Contribute to the creation Linked Data learning material; and
- Reuse and align the material with our own training material.

The commercial value of EUCLID is thus two-fold, by

- Increasing the visibility of fluidOps' products and services, and
- By being able to offer high-quality trainings based on EUCLID material.

As part of its professional training offerings fluidOps offers a series of training modules, two of which are closely related to EUCLID:

- **Basic Training in Semantic Technologies**

  Training for Consultants and Administrators

  This module provides insights into how semantic technologies can be applied for data management, integration, analysis and visualization:

  - Understand the concept and benefits of Linked Data

[36] Note that due to Non-Disclosure Agreements, Ontotext is not allowed to provide details on some of its big training services customers.

- Learn how to model information using the RDF data model
- Learn how to formulate SPARQL queries
- Understand basic reasoning mechanisms
- Learn how the fluidOps Platform can be used for semantic data management

- **fluidOps Information Workbench Platform: In-Depth Training**

  Training for Consultants and Administrators

  This module highlights the concepts, benefits and use cases of the fluidOps Platform and gives detailed insights into customization opportunities.

  - Understand the concepts and benefits of the fluidOps Platform
  - Learn how to install and configure the Information Workbench
  - Learn how to create and modify custom visualizations
  - Learn how to create custom dashboards
  - Learn how to configure and administrate the system, both using the UI and the CLI
  - Learn how to load data into the system and configure data providers
  - Learn how to build and deploy custom solutions

In both modules, fluidOps reuses large parts the EUCLID curriculum. In particular, "Basic Training in Semantic Technologies" builds on and reuses parts of modules 1-4 of this curriculum."fluidOps Information Workbench Platform: In-Depth Training" is closely aligned with modules 4 and 5 of EUCLID.

Besides onsite training, fluidOps regularly points customers and partners to the EUCLID material for self-learning.

The prominent appearance of the Information Workbench platform in the EUCLID training material is an excellent vehicle for marketing. The same applies to the community monitoring platform.

STI Research

As spin-off of STI International, STI Research and Consulting offers a variety of R&D, knowledge transfer, dissemination, training, and community building services targeted at different stakeholders in the semantic technologies and Linked Data community. These services include, most relevantly for EUCLID, the semantic technologies training and certification program run under the Semsphere[37] brand. The training is organized in 10 modules grouped in three levels:

- Specialist level
  - Introduction to Semantic Technologies
  - Linked Data
  - Ontologies
- Professional level
  - OWL and RIF
  - Semantic Annotation
  - APIs and Services
- Academy Trainer level
  - Advanced Ontologies
  - Advanced Reasoning
  - Semantic Application Development
  - Training & Certification Methodology

[37] https://www.semsphere.com/

Semsphere has contributed to the creation of EUCLID materials (providing slides, exercises, and literature reviews) and has incorporated parts of the curriculum into the three levels of its certification program. Modules 1 and 2 are relevant for the Specialist level of the program. Module 4 will replace the current content of the APIs and Services course at the Professional level, while Module 5 will be used to cover application development topics at the Academy Trainer level.
In addition, STI International will continue to rely on EUCLID materials and trainers for the organization of its yearly summer school, which is a collaboration of several PlanetData and EUCLID partners (besides STI International, KIT, OU, and Southampton).

Open University

The Open University has over the last 40 years played a leading role in distance education with now over 250,000 students of which the majority is online. As such, since its beginnings it has always sought to make use of the newest technologies to support teaching and learning. For example, it was nearly named the "University of the Air" to reflect the use of radio for course delivery the dominant mass communication technology at the time of its launch in the 1960s.
The OU is thus very interested in making use of EUCLID results and also is well placed to leverage these within its core business. We see the following routes for EUCLID exploitation:

1. **EUCLID materials in iTunes** – since 2008 when the OU entered iTunes U the OU has seen over 60 million downloads for its materials. All of these downloads provide advertising for the OU in general and lead users and readers to OU courses. The OU is proud to have been the first higher education institution to create a comprehensive Linked Data portal (data.open.ac.uk) and this features in many documents outlining our research impact. Thus, we will use the EUCLID iBook and iTunes U course as a way of highlighting both OU teaching and research.
2. **ESWC Summer School** – through its involvement with STI International, the OU was instrumental in setting up the ESWC Summer School and provides one of its directors, past keynotes and extensive admin support. Since 2013 EUCLID materials form a core part of the School allowing students to get up to speed with key Linked Data concepts before attending.
3. **FORGE** – is a new project coordinated by the OU within the Future Internet Research and Experimentation (FIRE) unit. FIRE supports European research in designing new internets through facilities comprised of a large set of big machines and high speed networks. As such FIRE facilities have an associated significant cost and maximising widespread take-up and use is an issue.
   Forging Online Education through FIRE (FORGE) aims to solve the above by bringing the FIRE and eLearning worlds together. FORGE will align FIRE with the new and emerging educational technologies for mutual benefit. In particular, this project is concerned with specifying development methodologies and best practices for offering FIRE experimentation facilities to learners and to the learning community in general, related both to communications and IT but also to other disciplines including the sciences and social sciences, leading to a strong connection between the learning community and existing FIRE platforms and supporting tools. The learning community will benefit from the use of the very high performance facilities. FIRE will benefit through the addition of an ever-growing set of FIRE-specific learning materials for a rising number of FIRE-based students, leading to increased awareness and use. The eLearning element of FORGE is completely founded on the EUCLID results.
   Within FORGE we are currently talking to Cisco about the Cisco Networking Academy[38] a body setup to train the next generation of Cisco network engineers and which annually trains over 1 million network engineers. The main software platform for training is Packet Tracer[39] a comprehensive networking simulation tool. In January 2014 we had a face-to-face meeting with the academy's senior management to outline our collaboration which will include a port of their tool to the IoS platform to support the interactive iBooks that we will create.
4. **Webcasting** – In 2014 the Open University will be deploying technologies we have tested in the EUCLID webcasting context into a range of high profile experimental live events. One new series, which is due to

[38] https://www.netacad.com/
[39] https://www.netacad.com/web/about-us/cisco-packet-tracer

start February 3rd 2014, will be from our second level science course S288[40] to a body of around 1,000 students. Students will be virtually invited, via the live interfaces, into a series of mentored sessions located in the labs and working spaces of the Open University to explore measurement and scientific inquiry via various complex experimental equipment. In another example, a week long online student conference event will be running from June 30th through to July 4th 2014 for the Faculty of Social Science at the OU. The conference will be open to an estimated 40,000 Open University Social Science students who will be submitting work that will be presented through this week. This 'online conference' will webcast a live stream of student and staff presentations to a large remote audience, and will aim to bring in remote presenters from their home and work to the live 'conference and workshop' model. We anticipate using a range of interaction devices to extend from large scale text chat and social network sharing to online social widgets for live interaction.

Karlsruhe Institute of Technology

As a teaching and research institution, the contributions of KIT within the EUCLID project are targeted more towards academia. Therefore, the exploitation of the developed resources and training materials has no direct commercial focus. Still all the content produced in collaboration with KIT can be reused by other researchers and industrial partners alike, in order to gain knowledge and skills in different areas of expertise related to Linked Data.

The main line of exploitation activities is the presentation and use of the materials as part of teaching activities and lectures. The work conducted within EUCLID is used as part of lectures, including Service Oriented Computing II, Wissensmanagement and Semantic Web Technologies I. Furthermore, KIT plans to reuse the produce modules as part of tutorials, within the scope of university external treating activities such as the Hector School,[41] and as part of summer schools for PhD student, such as the ESWC Summer School.

KIT will continue the exploitation of EUCLID results, beyond the end of the projects through reusing the contributions within further projects, through academic activities and individual research, e.g. as part of Ph.D. theses.

University of Southampton

As one of the largest Computer Science departments in the UK, the University of Southampton will offer an excellent environment for the further development of EUCLID materials. While commercial exploitation of the results is not at the core of our agenda, we are actively collaborating with professional trainers in delivering applied courses for a variety of IT domains. For EUCLID the most relevant is probably our involvement with the Open Data Institute, whose training programs are accredited via the University and are delivered in collaboration with academic and research staff at Southampton. In addition, EUCLID materials will form a core part of our Semantic Technologies lecture (MSc level) and will be integrated into a new program on Data Science, to be launched in 2015. Finally, the Web and Internet Science group at Southampton is one of the main organizers of a yearly Web Science summer school, to be held this year in Southampton in July.[42] We will reuse EUCLID materials for a course on data publishing and interlinking to be delivered by former EUCLID lead trainer, Dr. Barry Norton.

External organizations

As noted earlier, our learning materials have found a wide range of adopters, including research projects (FORGE, Open Data Support, LinkedUp), commercial companies (fluidOps), and other organizations (ODI). EUCLID has established itself as the default location for Linked Data training and EUCLID partners have a vested interest in capitalizing on this advantageous position to ensure that the ultimate goal of the project, that is, to contribute to the global spread of Linked Data principles and technologies, will be attained. The ESWC Summer School, which has been recently established as a collaboration of several higher-education institutions in the UK, Germany, Greece, and Slovenia, will be key to the further development of the learning materials.

---

[40] http://www.open.ac.uk/science/main/studying-science/s288-practical-science
[41] http://hector.idschools.kit.edu
[42] http://www.summerschool.websci.net

4 Concluding remarks

In this deliverable we presented and discussed the cornerstones of our dissemination and community building strategy, and provided a comprehensive exploitation plan based on an analysis of the professional training market and potential competitors. The key tangible outcomes of the project (methodology, learning materials, monitoring platform) will be reused and further developed by core and associate partners as part of other collaborative projects and as a commercial training service. Through the ESWC Summer School and the academic partners we will ensure that materials remain up-to-date and the main engagement channels continue their activity.

References

[1] http://www.euclid-project.eu/

[2] http://conference.ocwconsortium.org/2014/wp-content/uploads/2014/04/Paper_15-Curriculum.pdf

[3] http://ercim-news.ercim.eu/en96/special/raising-the-stakes-in-linked-data-education

Appendix: EUCLID publications

Alexander Mikroyannidis et al. Developing a Curriculum of Open Educational Resources for Linked Data. 10th annual OpenCourseWare Consortium Global Conference (OCWC 2014), Ljubljana, Slovenia. http://conference.ocwconsortium.org/2014/wp-content/uploads/2014/04/Paper_15-Curriculum.pdf

Alexander Mikroyannidis, John Domingues, and Elena Simperl. Raising the Stakes in Linked Data Education. In ERCIM News, Special Theme Linked Open Data, January 2014. http://ercim-news.ercim.eu/en96/special/raising-the-stakes-in-linked-data-education

John Domingue, Mathieu d'Aquin, Elena Simperl, and Alexander Mikroyannidis. The Web of Data: Bridging the Skills Gap. IEEE Intelligent Systems, 29(1):70-74. 2014.

Developing a Curriculum of Open Educational Resources for Linked Data

Alexander Mikroyannidis and John Domingue, Knowledge Media Institute, The Open University
{Alexander.Mikroyannidis, John.Domingue}@open.ac.uk
Maria Maleshkova, Karlsruhe Institute of Technology (KIT)
maria.maleshkova@kit.edu
Barry Norton, British Museum
BNorton@britishmuseum.org
Elena Simperl, University of Southampton
E.Simperl@soton.ac.uk

Abstract
The EUCLID project is developing an educational curriculum about Linked Data, supported by multimodal Open Educational Resources (OERs) tailored to the real needs of data practitioners. The EUCLID OERs facilitate professional training for data practitioners, who aim to use Linked Data in their daily work. The EUCLID OERs are implemented as a combination of living learning materials and activities (eBook, online courses, webinars, face-to-face training), produced via a rigorous process and validated by the user community through continuous feedback.

Keywords
Open Educational Resources, Linked Data, Massive Open Online Courses

Introduction

There is a revolution occurring now in higher education, largely driven by the availability of high quality online materials, also known as Open Educational Resources (OERs). OERs can be described as "teaching, learning and research resources that reside in the public domain or have been released under an intellectual property license that permits their free use or repurposing by others depending on which Creative Commons license is used" (Atkins, Brown, & Hammond, 2007). The emergence of OERs has greatly facilitated online education through the use and sharing of open and reusable learning resources on the Web. Learners and educators can now access, download, remix, and republish a wide variety of quality learning materials available through open services provided in the cloud.

The OER initiative has recently culminated in MOOCs (Massive Open Online Courses), which offer large numbers of students the opportunity to study high quality courses with prestigious universities. These initiatives have led to widespread publicity and also strategic dialogue in the higher education sector. The consensus within higher education is that after the Internet-induced revolutions in communication, business, entertainment, media, amongst others, it is now the turn of universities. Exactly where this revolution will lead is not yet known but some radical predictions have been made including the end of the need for university campuses[1].

Linked Data (Berners-Lee, 2006) has established itself as the de facto means for the publication of structured data over the Web, enjoying amazing growth in terms of the number of organizations committing to use its core principles for exposing and interlinking Big Data for

[1] http://www.theguardian.com/education/2012/nov/11/online-free-learning-end-of-university

seamless exchange, integration, and reuse (Bizer, Heath, & Berners-Lee, 2009). More and more ICT ventures offer innovative data management services on top of Linked Data, creating a demand for Data Scientists possessing skills and detailed knowledge in this area. Ensuring the availability of such expertise will prove crucial if businesses are to reap the full benefits of these advanced data management technologies, and the know-how accumulated over the past years by researchers, technology enthusiasts and early adopters.

The European project EUCLID[2] contributes to this goal by developing a comprehensive educational curriculum, supported by multimodal OERs and highly visible eLearning distribution channels, tailored to the real needs of data practitioners. The EUCLID curriculum focuses on techniques and software to integrate, query, and visualize Linked Data, as core areas in which practitioners state to require most assistance. A significant part of the learning materials produced in the project consists of examples referring to real-world data sets and application scenarios, code snippets and demos that developers can run on their machines, as well as best practices and how-tos.

The EUCLID approach

The EUCLID educational curriculum consists of a series of modules, each containing multi-format OERs, such as presentations, webinars, screencasts, exercises, eBook chapters, and online courses. These learning materials complement each other and are connected to deliver a comprehensive and concise training programme to the community. Learners are guided through these materials by following learning pathways, which are sequences of learning resources structured appropriately for achieving specific learning goals. Different types of eLearning distribution channels are targeted by each type of learning materials, including Apple and Android tablets, Amazon Kindles, as well as standard web browsers (see Figure 1). The EUCLID learning materials are available for free on the project web site, as well as on Apple's iBook Store[3] as an interactive iBook for use on the iPad and MacOS. All the materials are made available under a Creative Commons Attribution 3.0 Unported License[4].

Instead of mock Linked Data examples, we use in our learning materials and exercises a collection of datasets and tools that are deployed and used in real life. In particular, we use a number of large datasets including, for example, the MusicBrainz dataset, which contains 100Ms of triples. Our collection of tools includes Seevl, Sesame, Open Refine and GateCloud, all of which are used in real-life contexts. We also showcase scalable solutions, based upon industrial-strength repositories and automatic translations, e.g. by using the W3C standard R2RML for generating RDF from large data contained in standard databases.

Additionally, EUCLID has a strong focus on the community and encourages community engagement in the production of OERs through, for example, collecting user feedback via our webinars, Twitter, LinkedIn, and more. EUCLID combines online and real-world presence, and attempts to integrate with on-going activities in each sphere such as mailing lists and wikis. The project engages with the Linked Data community, both practitioners and academics, by

[2] http://www.euclid-project.eu

[3] http://bit.ly/using-linked-data-effectively

[4] http://creativecommons.org/licenses/by/3.0

collecting user requirements as well as feedback to the OERs so that they can be tailored to what the learner really needs.

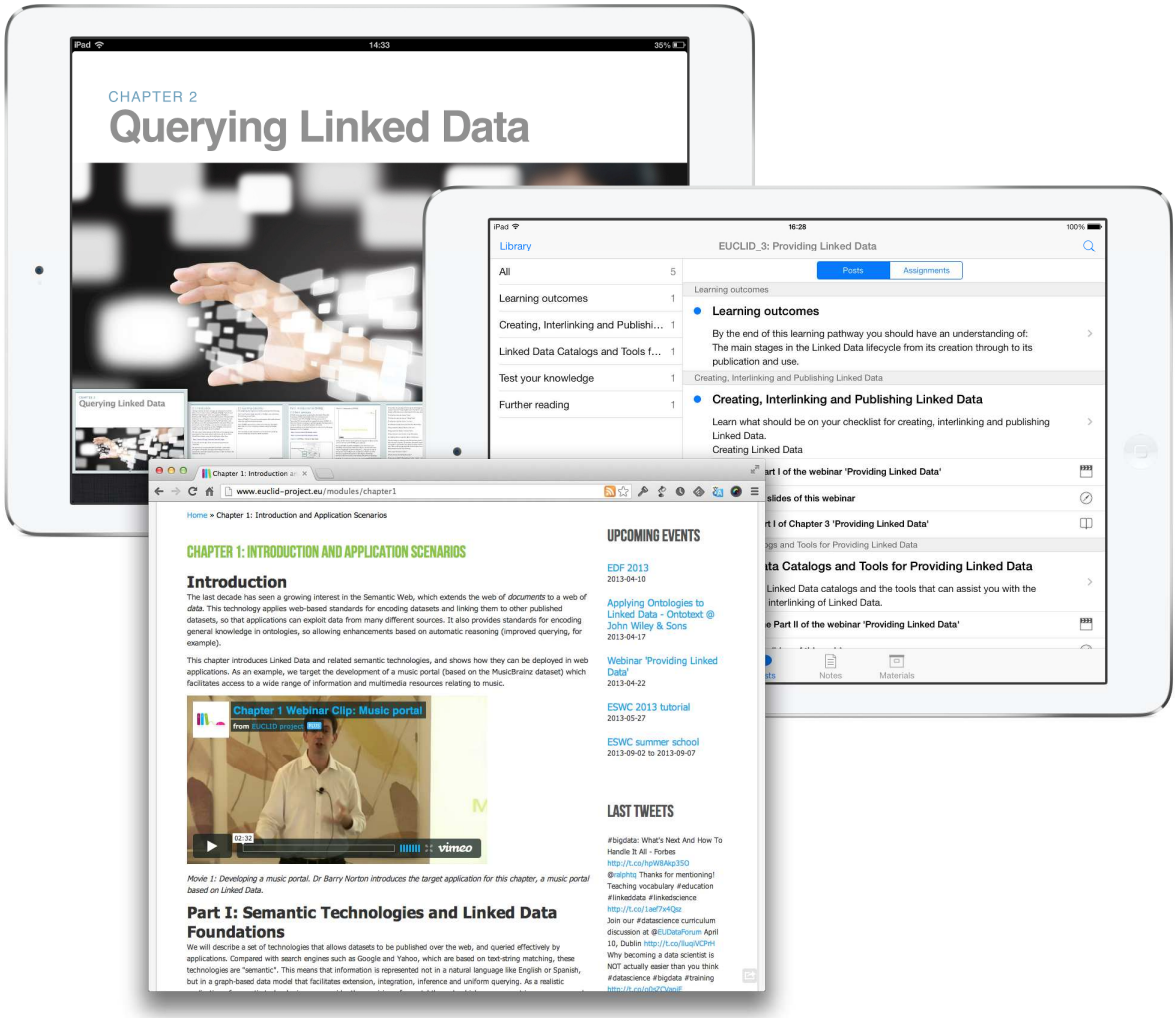


Figure 1. A selection of EUCLID learning materials in different formats and platforms, i.e. eBooks and online courses for the web and the iPad.

The EUCLID curriculum

The EUCLID curriculum has been designed to gradually build up the learner's knowledge. It enables learners with previous knowledge on a specific area of interest to only briefly go over the introductory materials and directly dig into one of the more advanced modules. As shown in Figure 2, the EUCLID curriculum is organized in the order of 3 expertise levels (top: introductory, middle: advanced, bottom: expertise). It is composed of 6 modules that cover all the major aspects of the Linked Data consumption lifecycle.

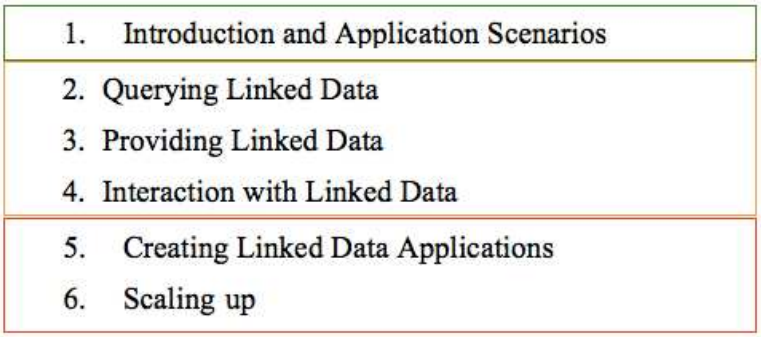


Figure 2. The EUCLID curriculum

The 6 EUCLID modules have been structured to cover the following range of topics:

- *Module 1: Introduction and Application Scenarios.* This module introduces the main principles of Linked Data, the underlying technologies and background standards. It provides basic knowledge for how data can be published over the Web, how it can be queried, and what are the possible use cases and benefits. As an example, we use the development of a music portal (based on the MusicBrainz dataset), which facilitates access to a wide range of information and multimedia resources relating to music. The module also includes some multiple-choice questions in the form of a quiz, screencasts of popular tools and embedded videos.
- *Module 2: Querying Linked Data*. This module looks in detail at SPARQL (SPARQL Protocol and RDF Query Language) and introduces approaches for querying and updating semantic data. It covers the SPARQL algebra, the SPARQL protocol, and provides examples for reasoning over Linked Data. The module uses examples from the music domain, which can be directly tried out and ran over the MusicBrainz dataset. This includes gaining some familiarity with the RDFS and OWL languages, which allow developers to formulate generic and conceptual knowledge that can be exploited by automatic reasoning services in order to enhance the power of querying.
- *Module 3: Providing Linked Data*. This module covers the whole spectrum of Linked Data production and exposure. After a grounding in the Linked Data principles and best practices, with special emphasis on the VoID vocabulary, we cover R2RML, operating on relational databases, Open Refine, operating on spreadsheets, and GATECloud, operating on natural language. Finally, we describe the means to increase interlinkage between datasets, especially the use of tools like Silk.
- *Module 4: Interaction with Linked Data*. This module focuses on providing means for exploring Linked Data. In particular, it gives an overview of current visualization tools and techniques, looking at semantic browsers and applications for presenting the data to the end used. We also describe existing search options, including faceted search, concept-based search and hybrid search, based on a mix of using semantic information and text processing. Finally, we conclude with approaches for Linked Data analysis, describing how available data can be synthesized and processed in order to draw conclusions. The module includes a number of practical examples with available tools, as well as an extensive demo based on analysing, visualizing and searching data from the music domain.

- *Module 5: Creating Linked Data Applications*. This module gives details on technologies and approaches towards exploiting Linked Data by building bespoke applications. In particular, it gives an overview of popular existing applications and introduces the main technologies that support implementation and development. Furthermore, it illustrates how data exposed through common Web APIs can be integrated with Linked Data in order to create mash-ups.
- *Module 6: Scaling up*. This module addresses the main issues of Linked Data and scalability. In particular, it provides gives details on approaches and technologies for clustering, distributing, sharing, and caching data. Furthermore, it addresses the means for publishing data trough could deployment and the relationship between Big Data and Linked Data, exploring how some of the solutions can be transferred in the context of Linked Data.

In an effort to provide high-quality training, suitable for the data practitioner's needs, the EUCLID curriculum has been through several revisions on structure, arrangement and content after presenting it to a number of experts and gathering their feedback. As a result of these revisions, the curriculum was refined and developed in more detail in order to include a number of expected outcome competencies, as well as a variety of exercises and examples. The content of the EUCLID modules has been redesigned to be better aligned and support a smoother process of skills built-up and development. While having an individual objective, each module contributes to further developing the skills and knowledge gained by the previous one thus aiding to acquiring an overall understanding and expertise in the field. As mentioned before, the curriculum is constantly updated based on feedback from the community.

The EUCLID OER production process

The OER production process defines the sequence of steps for the production of the EUCLID learning materials. Initially, 3 basic steps were planned in order to create each module and its exercises (see Figure 3). Firstly, following the curriculum, the draft of the training material would be created, which includes slides for a webinar, as well as HTML content for online distribution. Secondly, feedback on the drafts would be gathered and analysed. Finally, based on the comments and feedback, each module would be refined before delivering an eBook encompassing all the training materials, which include written documents, examples, presentation slides, as well as the video recording of the webinar.

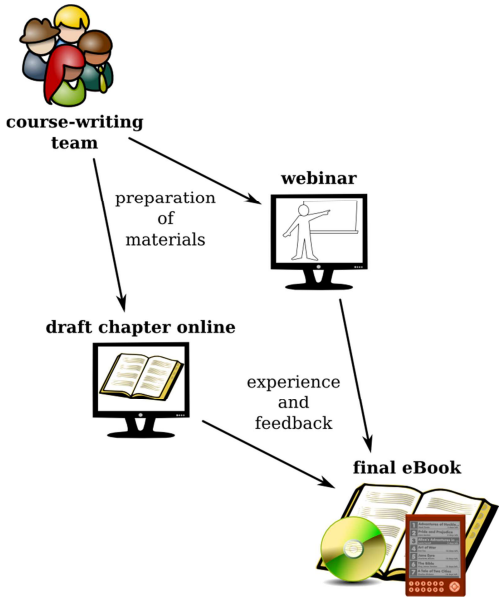


Figure 3. The initial OER production process

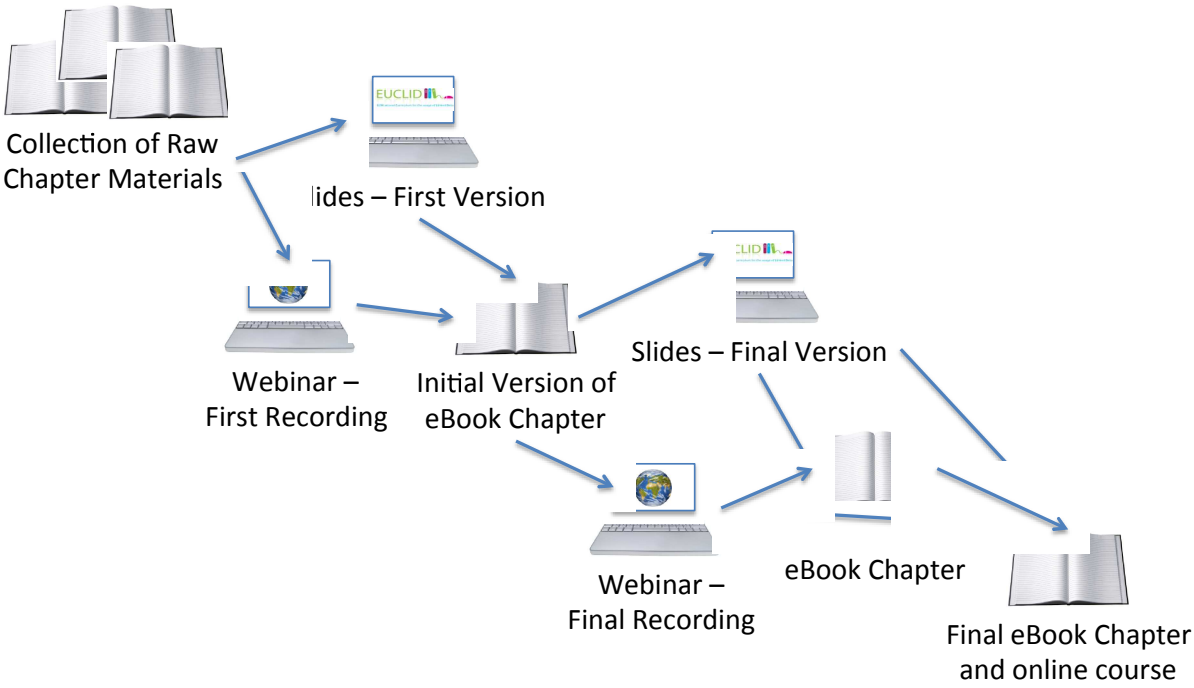


Figure 4. The revised OER production process

During the production of the first EUCLID module, this process was further refined and elaborated to include some intermediate steps (see Figure 4). One thing that became obvious was the instrumental role of the preparation and delivery of the webinar in the production process. The webinar was therefore produced in two stages. First, an internal webinar was held in order to collect feedback from project partners about its content and structure. The learning materials were revised through collecting comments and feedback from the internal broadcasting of the first webinar and the publication of the first version of the eBook chapter.

Subsequently, a second version of the webinar was produced, this time publically broadcasted. Based on the community feedback received from the broadcasting of the second webinar, the structure and content of the module were finalised and the eBook chapter was produced from all the finalised content. It was also decided that additional material in the form of an online course would accompany the final eBook chapter and would be part of the training programme offered to the community. This process has been applied for the production of all EUCLID modules.

Best practices for the design and delivery of Linked Data OERs

The design and implementation of the OER production process has provided us with a valuable insight into the various challenges associated with the design and delivery of learning materials specifically for Linked Data. We have thus distilled our experiences and lessons learned into a set of best practices, which is outlined in the next 2 sections.

Best practices for the design of Linked Data OERs

1. **Industrial Relevance** – our curriculum takes into account the needs of industry related to Open Data and Linked Data. Future work aims to automatically mine and analyse relevant job adverts to gain desired competencies for the sector. This is supported by the following best practice.
2. **Team Curriculum Design** – where the team is composed of a number of roles to fully capture industrial, academic and pedagogical requirements. Our team comprises of industrial partners (Ontotext, FluidOps), who have extensive experience with professional training, industrial requirements and scalable tools, academic partners (KIT, STI International), who have research expertise in Linked Data and pedagogical experts (The Open University).
3. **External Collaboration** – to gain world-class curriculum expertise where necessary and to facilitate course delivery and dissemination.
4. **Explicit learning goals** – to which all learning materials (slides, webinars, eBook chapters) are developed. Learners are guided through the learning goals by learning pathways – a sequence of learning resources to achieve a learning goal.
5. **Show realistic solutions** – rather than mock examples we utilize systems that are deployed and used for real.
6. **Use real data** – we use a number of large datasets including for example, the MusicBrainz dataset that contains 100Ms of triples.
7. **Use real tools** – our collection of tools are used in real life, including for example Seevl, Sesame, Open Refine and GateCloud.
8. **Show scalable solutions** – based upon industrial-strength repositories and automatic translations, for example using the W3C standard R2RML for generating RDF from large data contained in standard databases.
9. **Eating our own dog food** – we monitor communication and engagement with the Linked Data community through W3C email lists, in the social network channels LinkedIn and Twitter, as well as content dissemination channels such as Vimeo and SlideShare. We transform the monitoring results into RDF and make these available at a SPARQL endpoint. In this respect we use Linked Data to support Learning Analytics.

Best practices for the delivery of Linked Data OERs

1. **Open to Format** – our learning materials are available in a variety of formats including: HTML, iBook (iPad and MacOS), ePUB (Android tablets), MOBI (Amazon Kindle).
2. **Addressability** – every concept in our curriculum is URI-identified so that HTML and RDF(a) machine-readable content is available.
3. **Integrated** – to ease navigation for learners the main textual content, relevant webinar clips, screencasts and interactive components are placed into one coherent space.
4. **High Quality** – we have a formalised process where all materials go through several iterations to ensure quality. For example, for each module we run both a practice and a full webinars facilitating critique and commentary.
5. **Self-testing and reflection** – in every module we include inline quizzes and exercises formulated against learning goals enabling students to self-monitor their progress.

Conclusion

The EUCLID project has established a rigorous process for the production and delivery of OERs about Linked Data. This process defines a series of iterations in the production of learning materials, with multiple revisions from internal and external stakeholders, in order to ensure high quality in the produced materials. Based on our experiences and lessons learned in designing and implementing the production process, we have also established a set of best practices for the design and delivery of OERs specifically for Linked Data.

One of our main goals is to reach out to the community in as many ways as possible, in order to engage it and acquire its feedback. For this purpose, considerable effort has been put to delivering the learning materials in a variety of formats and for different purposes. The EUCLID learning materials can be accessed from a wide range of platforms, both from desktop/laptop computers, as well as from different mobile devices. With the learning materials reaching a constantly growing community, it is expected that there will be more comments and feedback received, which we will continuously monitor in order to improve the quality of the EUCLID training programme.

References

Atkins, Daniel E., Brown, John Seely, & Hammond, Allen L. (2007). A Review of the Open Educational Resources (OER) Movement: Achievements, Challenges, and New Opportunities (pp. 4): The William and Flora Hewlett Foundation.

Berners-Lee, T. (2006). Linked Data - Design Issues. from http://www.w3.org/DesignIssues/LinkedData.html

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked Data—The Story So Far. *International Journal on Semantic Web and Information Systems, 5*(3), 1–22.

Number 96 January 2014

ERCIM NEWS

www.ercim.eu

Special theme:

Linked Open Data

Also in this issue:

ERCIM News is the magazine of ERCIM. Published quarterly, it reports on joint actions of the ERCIM partners, and aims to reflect the contribution made by ERCIM to the European Community in Information Technology and Applied Mathematics. Through short articles and news items, it provides a forum for the exchange of information between the institutes and also with the wider scientific community. This issue has a circulation of about 6,500 printed copies and is also available online.

ERCIM News is published by ERCIM EEIG
BP 93, F-06902 Sophia Antipolis Cedex, France
Tel: +33 4 9238 5010, E-mail: contact@ercim.eu
Director: Jérôme Chailloux
ISSN 0926-4981

KEYNOTE

JOINT ERCIM ACTIONS

SPECIAL THEME

The special theme section "Linked Open Data" has been coordinated by Irini Fundulaki: Institute of Computer Science, FORTH, and Sören Auer, University of Bonn and Fraunhofer IAIS

RESEARCH AND INNOVATION

This section features news about research activities and innovative developments from European research institutes

EVENTS

Announcements

Linked Data: The Quiet Revolution

Wendy Hall, Professor of Computer Science at the University of Southampton and Dean of the Faculty of Physical Science and Engineering.

In 2014 we celebrate the twenty-fifth anniversary of the birth of the idea that became the World Wide Web. In March 1989, while working at CERN, Tim Berners-Lee wrote a document entitled "Information Management – a Proposal" which described his ideas for creating a hypertext system that worked over the internet to enable physicists and other researchers around the world to easily exchange documents [1]. I first heard about his ideas at the European Hypertext Conference (ECHT'90) in November 1990, where we presented our first paper on the Microcosm "open hypermedia system" that we had been developing at Southampton since 1989 [2]. I first saw Tim and his colleague Robert Cailliau demonstrate the system they had by then called the World Wide Web at the ACM Hypertext conference (HT'91) in December 1991. This was the conference that famously rejected their paper about the WWW but it also marked the beginning of the revolution that the Web has brought to all our lives – the killer application for the Internet, the disruptive technology that has changed the way we communicate and manage information forever.

The Web in its purest form is a set of open standards and protocols for downloading documents from the Internet via hypertext links which are embedded in the documents. In the early 1990's, explaining the concept of a global hypertext system to people was incredibly difficult. Ted Nelson had been trying for decades. As more and more content was put on the Web, it became possible to at least show people a vision of a future in which linked documents became the norm for information exchange on the Internet. To me the Web lacked the richness that we had been exploring in the Microcosm system where the links existed as entities in their own right and could describe relationships between information objects. However, this was also part of Tim's original vision of a web that not only linked documents but also linked data as he articulated at the first WWW conference in Geneva in May 1994.

"The web is a set of nodes and links. To a user, this has become an exciting world, but there is very little machine-readable information there … To a computer, then, the web is a flat, boring world devoid of meaning. This is a pity, as in fact documents on the web describe real objects and imaginary concepts, and give particular relationships between them. Adding semantics to the web involves two things: allowing documents which have information in machine-readable forms, and allowing links to be created with relationship values. Only when we have this extra level of semantics will we be able to use computer power to help us exploit the information to a greater extent than our own reading." [3]

This was too complicated for a world that was just discovering linked documents to grasp. Over the next few years the Web exploded. Search engines like Google emerged to help us find information on it and as the browsers became more interactive the social networks that seem to define the Web today began to evolve. Tim was still talking about a web of linked data, or the Semantic Web as he called it, and wrote about it in his book Weaving the Web [4] but it didn't seem to be emerging. The people taking it most seriously were the artificial intelligence community who were keen to theorise about it and started their own conference series on the Semantic Web in 2002.

But the linked data revolution was creeping up on us quietly rather than with the big bang by which the first Web seemed to appear. The Semantic Web community tracked its initially slow emergence through the linked data cloud, and the technical requirements were simplified [5]. But in essence, it has emerged as a necessary part of the data revolution that has been a natural consequence of the development of the Web and the Internet. Driven initially by the scientific community and further fuelled by the movement towards open data in the public sector, businesses are rapidly realizing the advantages of harnessing the power of big data and the absolute need for linked data technology as part of these developments.

"… linked data is the next step for the Web. There are multiple benefits over and above traditional Business Intelligence. It's cheaper and faster to implement. If knowledge is power, with linked data's capability to compare internal with external data, there is the potential to pull information to provide unrivalled context on business strategy." Anwen Robinson, Managing Director, Unit 4, Business Software [6]

The behemoths of the Internet, such as Google and Microsoft, having denied it for so long, are finally making linked data an integral part of their offerings. This quiet revolution is actually going to have a bigger and more far reaching impact that the initial Web revolution. Linked data will become an integral part of the development of data-driven systems architectures that will revolutionize the way we build and maintain information management systems over the next few years. They will supersede relational databases and unify the worlds of hypertext, document management and databases to create rich interlinked knowledge-based systems as envisaged by the pioneers such as Bush, Nelson and Englebart over fifty years ago.

The excellent set of articles about linked data in this special issue of the ERCIM news is timely indeed. Let the revolution begin.

Links/References:

[1] http://www.w3.org/History/1989/proposal.html
[2] A. Fountain, et al.: "Microcosm: An Open Model for Hypermedia with Dynamic Linking", proc. of ECHT'90, Paris, 1990, Cambridge University Press, 298-311
[3] http://www.w3.org/Talks/WWW94Tim/
[4] T. Berners-Lee: "Weaving the Web", Harper Collins, 1999
[5] N. Shadbolt, T. Berners-Lee, W. Hall: "The Semantic Web Revisited", IEEE Intelligent Systems, 21(3), 96-101
[6] A. Robinson: "Think Link" http://www.bigdatarepublic.com 9/4/2013

Future Plans and Strategy for ERCIM

A Conversation with Domenico Laforenza, New President of ERCIM AISBL

In a conversation with our Editor, Domenico Laforenza outlines his vision for the role to be played by ERCIM in a rapidly evolving global scenario. "Although many things are changing worldwide, and in particular in Europe" – Domenico says – "I am confident that ERCIM will remain at the forefront of developments into the future. Plans have been put in place for this to happen effectively and efficiently, and the right people - young dynamic, excellent researchers - are positioned to take over from the current senior generation. I believe that ERCIM has a great future: my efforts will be devoted to contributing to its preparation!"

In 2014, ERCIM will celebrate twenty-five years of cooperation for excellence in research in the Information and Communication Sciences and Technologies (ICST) domain. Over these years, ERCIM has developed from the initial three-member consortium to an open and inclusive European organization of currently twenty-two members. Originally based on the "one member per country" model, each acting as a node linking academia and industry in that country and - via the fellow ERCIM members - to other countries, I am happy to see that ERCIM has now changed its structure, in 2012 opening its doors to new members head-quartered in Europe.

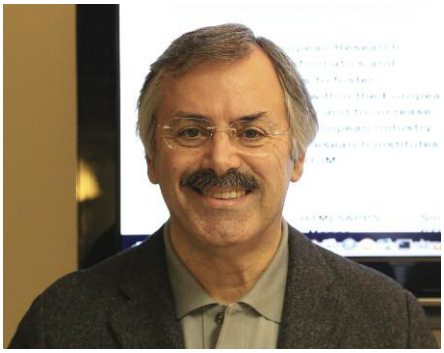
ERCIM has a long history of activity, participating and coordinating European projects through its member institutes and working groups, and has been recognized by the European Commission as a significant player. In fact, the network of researchers working in ERCIM institutes is a unique asset in the European research area, fostering collaboration and excellence in research. The widening of the ERCIM network will further strengthen this impact. It is my conviction that the start of the Horizon 2020 framework programme presents new challenges and opportunities for ERCIM-led actions that will stimulate innovation across Europe and beyond.

For this reason, the ERCIM AISBL Board in close collaboration with the ERCIM EEIG Board of Directors is already undertaking initiatives aimed at mobilising ERCIM in the direction of proposal preparation ready for the first calls of H2020. We are currently planning a line of action aimed at contributing to the future landscape in ICST. The objective is to first identify the emerging grand challenges and then to define the strategic roadmaps and research topics needed to meet these challenges. At the same time, ERCIM will establish new working groups and refocus existing groups in areas that are identified as priority for both basic and applied research.

In order to define and implement future strategy, synergy with other international institutions will play a crucial role. ERCIM is not alone in the European public ICST research ecosystem. Other ICST scientific and professional societies have come into being in recent years. In view of its long history and tradition, I believe that ERCIM has a key responsibility to cooperate with these organizations in understanding and shaping Europe's digital future. Our common target must be to collectively identify the vision, needs and priorities, and offer our expertise to society as a whole. In particular, European public ICST research organizations must contribute to shaping the H2020 work programmes. In the past, each organisation interacted independently with the EC; however, the lack of a single voice has seriously impacted on their ability to be heard by EC decision makers. The development of common viewpoints and strategies for ICST in Europe and, whenever appropriate or needed, a common representation of this vision at the international level are the foundational principles of the European Forum for Information and Communication Sciences and Technologies.

EFICST (http://www.eficst.eu/) was established in November 2011 by the joint action of seven leading organizations and societies in ICST in Europe: ACM Europe (ACM Europe Council), European Association for Programming Languages and Systems (EAPLS), European Association of Software Science and Technology (EASST), European Association for Theoretical Computer Science (EATCS),

New President for ERCIM AISBL



The General Assembly of the ERCIM AISBL, held in Athens in November 2013, unanimously elected Domenico Laforenza, Director of the Institute for Informatics and Telematics (IIT) of the Italian National Research Council (CNR), as its new President. Domenico succeeds Keith Jeffery who served as the President of ERCIM for almost ten years.

On behalf of the Assembly, Domenico thanked Keith greatly for his outstanding commitment to ERCIM over many years. Keith began his involvement in ERCIM in 1990 as coordinator of the Database Working Group, becoming subsequently member of the Executive Committee and then the Board of Directors. He was elected President in 2004. During his presidency, ERCIM has grown to become a widely recognized key player in European ICST. In particular, together with Michel Cosnard, INRIA, President of ERCIM EEIG, Keith has been responsible for steering the Consortium through an important set of structural changes, including the decision to open up membership to allow the participation of multiple members per country, thus encouraging excellent ICST research institutions to be involved in ERCIM activities directly rather than via a national intermediary.

Domenico began his activity in ERCIM in 1993 contributing to the creation of the ERCIM Parallel Processing Network (PPN), serving on the PPN Steering Committee (1993-1998), and as chairman in the period 1994-1996. He has represented CNR on the Board of Directors since 2006. His three-year term of office as ERCIM AISBL President began on 1 January 2014.

European Coordinating Committee for Artificial Intelligence (ECCAI), ERCIM, and INFORIE (Informatics Europe). The Forum is intended to be an open platform for cooperation among the scientific ICT societies in Europe. Under my Presidency, ERCIM intends to play a leading role in this Forum.

Another important area where I intend for ERCIM to be active in the near future will be the strengthening of relationships with the Knowledge and Innovation Communities (KICs) of the European Institute for Innovation and Technology (EIT), and in particular with the EIT KIC ICT Labs. This is essential if we are to stimulate innovation through a more rigorous and dynamic link with higher education, research and business.

Summing up, my main vision for ERCIM during my Presidency is that we continue to play our role as leaders of research and innovation in the European ICST domain, defining strategy, encouraging synergy, and promoting research which reaches far beyond formal disciplinary barriers in order to meet the new demands and challenges of the continually evolving global community. ERCIM has been considered as a key institution for ICST in Europe for many years now, acting as a focal point of expertise and vision. We intend to maintain this position for many years to come.

Please contact:
Domenico Laforenza, ERCIM President
IIT-CNR, Italy
E-mail: domenico.laforenza@iit.cnr.it

New Swiss ERCIM Member: the University of Geneva

The University Center for Informatics of the University of Geneva has just joined ERCIM as the Swiss member institution.

The Centre Universitaire d'Informatique (CUI) is an interdisciplinary center devoted to computer science teaching and research. Its members belong to three of the eight faculties of the University of Geneva: the Faculty of Sciences, the Faculty of Humanities and the Faculty of Economic and Social Sciences. At the end of 2012, the CUI community counted 118 collaborators, including 20 at the professor level.

The CUI has nine research centers:

- CLCL - Computational Learning and Computational Linguistics
- CVML - Computer Vision and Multimedia Laboratory
- GAIL - Geneva Artificial Intelligence Laboratory
- ICLE - Institute of Knowledge Integration and Access to Knowledge Repositories
- ISS - Institute of Services Science
- LATL - Laboratory for the Analysis and Technology of Language
- SMV - Software Modeling and Verification
- SPC - Scientific and Parallel Computing and
- TCS - Theoretical Computer Science

The CUI, one of the oldest computer research institutions in the world, has strong international ties with substantial financial support both from within Switzerland itself and from the EU.

Link: http://cui.unige.ch

Please contact:
Jose Rolim, University of Geneva,
E-mail: Jose.Rolim@unige.ch

HORIZON 2020 Project Management

A European project can be a richly rewarding tool for pushing your research or innovation activities to the state-of-the-art and beyond. Through ERCIM, our member institutes have participated in more than 70 projects funded by the European Commission in the ICT domain, by carrying out joint research activities while the ERCIM Office successfully manages the complexity of the project administration, finances and outreach.

Horizon 2020: How can you get involved?
The ERCIM Office has recognized expertise in a full range of services, including:

- Identification of funding opportunities
- Recruitment of project partners (within ERCIM and through a strategic partnership with Ideal-IST)
- Proposal writing and project negotiation
- Contractual and consortium management
- Communications and systems support
- Organization of attractive events, from team meetings to large-scale workshops and conferences
- Support for the dissemination of results.

How does it work in practice?

Contact the ERCIM Office to present your project idea and a panel of experts within the ERCIM Science Task Group will review your idea and provide recommendations. Based on this feedback, the ERCIM Office will decide whether to commit to help producing your proposal. Note that having at least one ERCIM member involved is mandatory for the ERCIM Office to engage in a project.

If the ERCIM Office expresses its interest to participate, it will assist the project consortium as described above, either as project coordinator or project partner.

For more information, please contact:
Philippe Rohou, Project Group Manager
Tel: +33 492 385 010
E-mail: philippe.rohou@ercim.eu

Small Data

by Steven Pemberton

It is now more than 20 years since the World Wide Web started. In the early days of the web, the rallying cry was "If you have information it should be on the web!" There are – possibly apocryphal – tales of companies not understanding this, and refusing initially to put their information online. One mail-order company is reported to have refused to put their catalogue online, because "they didn't want just anybody getting hold of it"! Luckily that sort of attitude has now mostly disappeared, and people are eager to make their information available on the web, not least of all because it increases the potential readership, and is largely cheaper than other methods of promulgation.

However, nowadays, with the coming of the semantic web and linked open data, the call has become "if you have *data* it should be on the web!": if you have information it should be machine-readable as well as human-readable, in order to make it useful in more contexts. And, alas, once again, there seems to be a reluctance amongst those who own the data to make it available. One good, and definitely not apocryphal, example is the report that the Amsterdam Fire Service

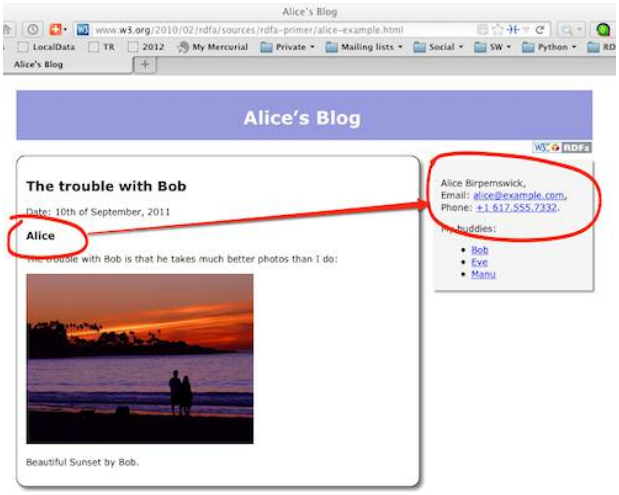


Figure 1: Example for using RDFa: individual blog items on the left, personal data, linked from the blog using RDFa terms, in a sidebar

needed access to the data of where roadworks were being carried out, in order to allow fire engines to get to the scene of fires as quickly as possible by avoiding routes that were blocked. You would think that that the owners of such data would release it as fast as possible, but in fact the Fire Service had the greatest of trouble getting hold of it. [1]

But releasing data is not just about saving lives. In fact, one of the fascinating aspects of open data is that it enables applications that weren't thought of until the data was released, such as improving maps [2], mapping crime, or visualising where public transport is in real time [3].

The term "Open Data" often goes hand in hand with the term "Big Data", where large data sets get released allowing for analysis, but the Cinderella of the Open Data ball is Small Data, small amounts of data, nonetheless possibly essential, that are too small to be put in some database or online dataset to be put to use.

Typically such data is on websites. It may be the announcement of a conference, or the review of a film, a restaurant, or some merchandise in a blog. Considered alone, not super-useful data, but as the British say "Many a mickle makes a muckle": many small things brought together make something large. If all the reviews of the film could be aggregated, together they would be much more useful.

The fairy godmother that can bring this small data to the ball is a W3C standard called RDFa [4] (ERCIM is the European host of W3C), the second edition of which (version 1.1) was recently released. RDFa is a light-weight layer of attributes that can be used on HTML, XHTML, or any other XML format, such as Open Document Format (ODF) files [5], that makes text that is otherwise intended to be human-readable also machine-readable. For instance, the markup

```
<div typeof="event:Vevent">
    <h3 property="event:summary">WWW 2014</h3>
   <p property="event:description">23rd International
World Wide Web Conference</p>
    <p>To be held from
        <span property="event:dtstart" content="2014-
04-07">7th April 2014</span>
         until <span property="event:dtend" con-
tent="2014-04-11">11th April</span>,
         in <span property="event:location">Seoul,
Korea</span>.</p>
</div>
```

marks up an event, and allows an RDFa-aware processor or scraper to extract the essential information about the event in the form of RDF triples, thus making them available for wider use. It could also allow a smart browser, or a suitable plugin for a browser, to extract the data and make it easier for the user to deal with it (for instance by adding it to an agenda).

RDFa is already widely used. For example, Google and Yahoo already accept certain RDFa markup for aggregating reviews, online retailers such as Best Buy and Tesco use it for marking up products, and the London Gazette, the daily publication of the British Government uses it to deliver data [6].

Links:

More information about RDF: http://rdfa.info/

[1] http://blog.okfn.org/2010/10/25/getting-started-with-governmental-linked-open-data/

[2] http://vimeo.com/78423677

[3] http://traintimes.org.uk/map/tube/

[4] http://www.w3.org/TR/xhtml-rdfa/

[5] http://docs.oasis-open.org/office/v1.2/OpenDocument-v1.2.html

[6] http://www.london-gazette.co.uk/reuse

Please contact:
Steven Pemberton, CWI, The Netherlands
E-mail: steven.pemberton@cwi.nl

Introduction to the Special Theme

Linked Open Data

by Irini Fundulaki and Sören Auer

The Linked Data paradigm has emerged as a powerful enabler for publishing, enriching and sharing data, information and knowledge in the Web. It offers a set of best practices that promote the publication of data on the Web using semantic web technologies such as URIs and RDF, support the exchange of structured data to be done as easily as the sharing of documents, allow the creation of typed links between Web resources and offer a single, standardized access mechanism. In particular, the Linked Data shift is based on (1) using Universal Resource Identifiers (URIs) for identifying all kinds of "things", (2) making these URIs accessible via the HTTP protocol and (3) providing a description of these things in the Resource Description Format (RDF) along with (4) URI links to related information (see Tim Berners-Lee's Linked Data design principles http://www.w3.org/ DesignIssues/LinkedData.html).

The RDF format is a relatively simple but a powerful formalism for representing information (very close to natural language) in triple statements consisting of a subject, predicate and object, where again each of them can be a URI (or an atomic value in the case of object). As a result, the World Wide Web of documents (and Intranets) is complemented with a Web of Linked Data, where everybody can publish, interlink and enrich data. This Linked Data Web has some interesting characteristics:

- URIs serve two purposes: identifying "things" and serving as locators and access paths for information about these "things".
- Since everybody can coin his own URIs by simply using the address of some webspace under his control, the Web of Linked Data is as distributed and democratic as the Web itself.
- Identifiers defined by different people or organizations can be mixed and meshed.
- Linked Data published in various locations can be easily integrated by merging the sets of RDF triple statements thus dramatically simplifying data integration using special purpose links as defined in semantic web languages such as OWL.
- The same triple statement formalism is used for defining structure and data, thus overcoming the strict separation found in relational or XML databases.

Linked Data management goes beyond the classic data management approaches that assume complete control over schema and data. In the Web that is distributed and open, users and applications do not have control over the data. For this reason, the Linked Data lifecycle was introduced that involves issues such as (a) data extraction from unstructured or semi-structured sources and representing them in the RDF data model (b) storage and querying of data (c) manual revision and authoring (d) interlinking and fusing of related data in order to enable data integration across highly heterogeneous sources (e) classification and enrichment with additional upper level structures such as ontologies and rich vocabularies (f) quality analysis (g) evolution and repair and (h) search/browsing and exploration [1]. The LOD2 project (http://lod2.eu) has created a comprehensive stack of tools for supporting these different aspects of Linked Data management [2].

Linked Data deployments benefit society, research and enterprises and support a great variety of different application areas. The Linked Open Data (LOD) movement is a growing trend for a variety of organizations and in particular governmental ones, to make their data accessible in a machine-readable form. The result of this effort is the creation of the Linked Open Data Cloud that during its last inventory in 2011 consisted of already 31 billion RDF triples from 295 datasets. These datasets contain user-generated content as well as content covering a variety of different domains, such as media, geographic, government, bibliographic data, and life sciences. A number of datasets have sprung from this effort, the most prominent one being DBpedia that is a community effort to extract structured information from Wikipedia, widely used in data integration efforts. DBpedia data are published in a consistent ontology and are accessible through multiple SPARQL endpoints.

The European Commission is one of the main evangelists of the adoption of Linked Data practices for opening government data to Europe's people and institutions. Through its ISA Programme the Commission provides "good practices and helpful examples to help public administration apply Linked Data technologies to eGovernment" (https://joinup.ec.europa.eu/sites/default/files/D4.3.2_Case_Study_Linked_Data_eGov.pdf). Moreover, the European Commission has been funding a number of projects related to Linked Data in the context of the 7th Framework Programme (FP7).

This special theme on Linked Data comprises 22 articles presenting some of the diversity of Linked Data research, technology and applications in Europe. The variety of topics include:

- foundational issues such as benchmarks of Linked Data infrastructure (Angles et al.: "Benchmarking Linked Open Data Management Systems"),
- experimental evaluation (Ferro & Silvello: "Making it easier to Discover, Re-Use and Understand Search Engine Experimental Evaluation Data"),
- Linked Data metadata catalogs (Vandenbussche & Vatant: "Linked Open Vocabularies", Stegmaier et al.: "Lost in Semantics? Ballooning the Web of Data") and
- registries (Vandenbussche et al: "SPARQL: A Gateway to Open Data on the Web?") as well as
- archiving and evolution (Papastefanatos & Stavrakas: "Diachronic Linked Data: Capturing the Evolution of Structured Interrelated Information on the Web").

Two articles tackle the management of spatial (Athanasiou et al.: "GeoKnow: Making the Web an Exploratory for Geospatial Knowledge") and statistical (Petrou & Papastefanatos: "Publishing Greek Census Data as Linked Open Data") linked data.

The heterogeneity of linked data demands for:

- approaches for querying, browsing and visualization (Pham & Boncz: "MonetDB/RDF: Discovering and Exploiting the Emergent Schema of RDF Data", Hoefler & Mutlu: "CODE Query Wizard and Vis Wizard: Supporting Exploration and Analysis of Linked Data", Micsik et al.: "Browsing and Traversing Linked Data with LODmilla", Sack & Plank: "AV-Portal - The German National Library and Technology's Semantic Video Portal") as well as
- analytics (Roatiș:"Analyzing RDF Data: A Realm of New Possibilities", Hall et al.: "The Web Science Observatory - The Challenges of Analytics over Distributed Linked Data Infrastructures).

An important application area of Linked Data are:

- research infrastructures, for example, for marine research (Fugazza et al: "RITMARE: Linked Open Data for Italian Marine Research", Tzitzikas et al.: "Ontology-based Integration of Heterogeneous and Distributed Information of the Marine Domain"),
- virtual earth observatories (Kyzirakos et al.: "Building Virtual Earth Observatories Using Scientific Database and Semantic Web Technologies") or
- history (Marx: "Linking Historical Entities to the Linked Open Data Cloud) and
- employing Linked Data in education (d' Aquin & Dietze: "Open Education: A Growing, High Impact Area for Linked Open Data", Mikroyannidis et al.: "Raising the Stakes in Linked Data Education").

ERCIM Open Data Working Group

Open Data is about knowledge that can be used, reused and redistributed. The last years have seen an enormous effort in publishing open data from scientific and government sources and this explosion of information has raised new and interesting data management challenges. The objective of the Open Data Working Group is to build and maintain a network of participants from academia, data producers and consumers who are involved in different aspects of Open Data Life Cycle. The group will promote the organisation of events such as workshops and schools and the preparation of common project proposals on Open Data.

ERCIM Working Groups are open to any researcher in the specific scientific field. Scientists interested in participating in the ERCIM Open Data Working Group should contact the coordinator Irini Fundulaki.

Further application areas represented through articles in this special issue are:

- real-time data (Martínez-Prieto et al: "A SOLID Architecture to Weather the Storm of Real-Time Linked Data") and the
- publishing domain (Dirschl et. al., "Supporting the Data Lifecycle at a Global Publisher using the Linked Data Stack").

References:

[1] S. Auer et al.: "Introduction to Linked Data and its Lifecycle on the Web", in Reasoning Web, Semantic Technologies for Intelligent Data Access - 9th International Summer School 2013, Mannheim, Germany, Springer LNCS. ISBN 978-3-642-39783-7, http://dx.doi.org/10.1007%2F978-3-642-39784-4_1

[2] S. Auer et al.: "Managing the life-cycle of Linked Data with the LOD2 Stack" in proc. of ISWC 2012, http://iswc2012.semanticweb.org/sites/default/files/76500001.pdf

Please contact:

Irini Fundulaki
ICS-FORTH, Greece
E-mail: fundul@ics.forth.gr

Sören Auer
University of Bonn and Fraunhofer IAIS, Germany
E-mail: auer@cs.uni-bonn.de

Building Virtual Earth Observatories Using Scientific Database and Semantic Web Technologies

by Kostis Kyzirakos, Stefan Manegold, Charalampos Nikolaou and Manolis Koubarakis

TELEIOS is a recent European project that addresses the need for scalable access to petabytes of Earth Observation (EO) data and the identification of hidden knowledge that can be used in applications. To achieve this, TELEIOS builds on scientific databases, linked geospatial data and ontologies. TELEIOS was the first project internationally that introduced the Linked Data paradigm to the EO domain, and developed prototype services such as the real-time fire monitoring service that has been used for the last two years by decision makers and emergency response managers in Greece.

Linked Data is a new research area which studies how one can make data available on the Web, and interconnect it with other data with the aim of making the value of the resulting "Web of data" greater than the sum of its parts. The Web of data has recently started being populated with geospatial data. Great Britain's national mapping agency, Ordnance Survey, has been the first national mapping agency that has made available various kinds of geospatial data from Great Britain as Linked Open Data. With the recent emphasis on open government data in many countries,

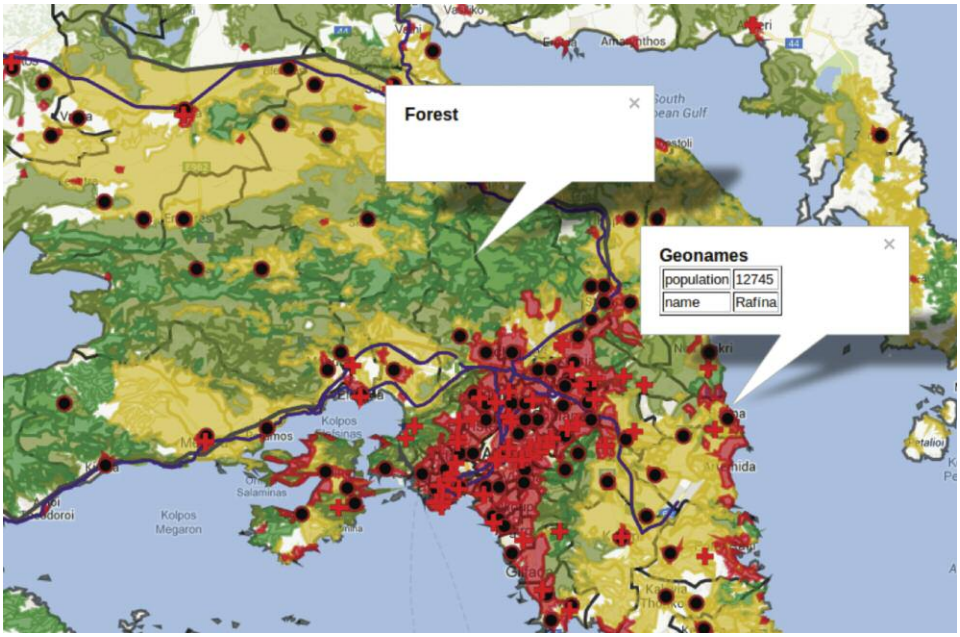


Figure 1: Fire product of 2009 in the prefecture Attica, Greece using linked earth observation data and Sextant.

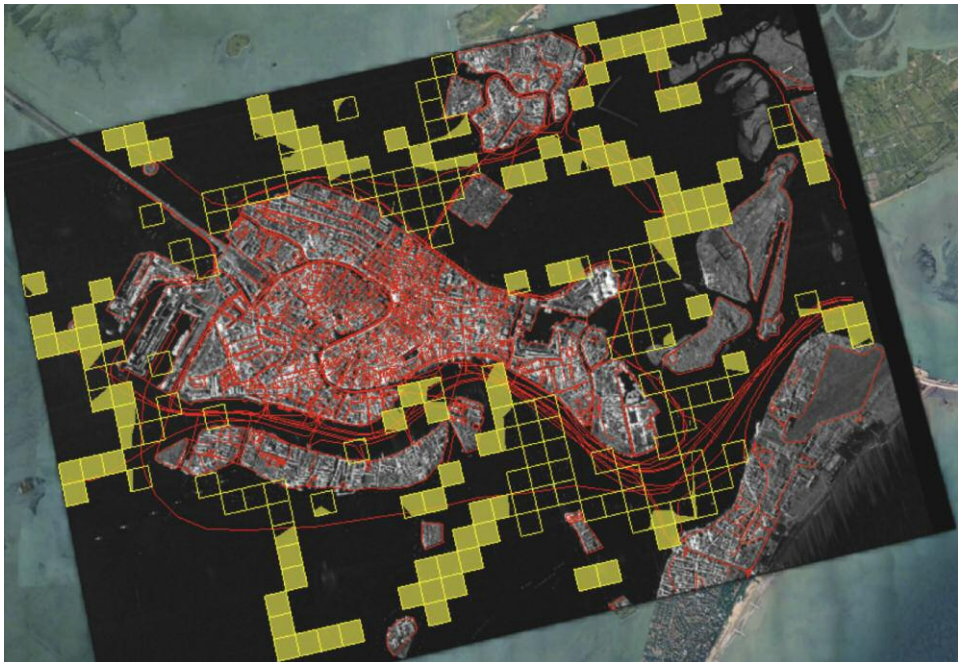


Figure 2: Buoys and water ways overlayed with a TerraSAR-X image using Sextant.

some of which is already encoded as Linked Data, the development of useful Web applications utilizing geospatial data is just a few SPARQL queries away.

In the European research project TELEIOS (1), we dealt with linked geospatial data in the domain of Earth Observation. TELEIOS addresses the need for scalable access to petabytes of EO data and the discovery of knowledge hidden within them. To achieve this, TELEIOS builds on scientific databases, linked geospatial data, ontologies and techniques for discovering knowledge from satellite images and auxiliary data sets. TELEIOS follows a database approach for the development of EO applications and pioneers the use of the following state-of-the-art results:

- The query language SciQL [1], an SQL-based query language for scientific applications with arrays as first class citizens (2).

- The data vault [2], a mechanism that provides a true symbiosis between a DBMS and existing (remote) file-based repositories such as the ones used in EO applications. The data vault keeps the data in its original format and place, while at the same time enabling transparent data and metadata access and analysis using the SciQL query language. SciQL and the data vault mechanism are implemented in the well-known column store MonetDB (3).

- Publicly available Linked Data, especially geospatial data such as OpenStreetMap and GeoNames. In TELEIOS, we published more than 100 GB of Linked Data with rich geospatial information. We published as Linked Data the CORINE Land Use/Land Cover and the Urban Atlas datasets produced by the European Environmental Agency, which provide information about the land use of Europe and the land cover of highly populated cities respectively.

- The model stRDF, an extension of the W3C standard RDF that allows the representation of geospatial data that changes over time [3]. stRDF is accompanied by stSPARQL, an extension of the query language SPARQL 1.1 for querying and updating stRDF data. stRDF and stSPARQL use OGC standards (Well-Known Text and Geography Markup Language) for the representation of temporal and geospatial data and are implemented in the open source geospatial RDF store Strabon (4).

- The model RDFi, an extension of stRDF for representing spatial regions about which the known information is incomplete or indefinite (e.g., we do not know the exact geographic location of a region, but we know neighbouring regions) [4].

- The novel visualization tool Sextant (5) for the visualization and exploration of time-evolving linked geospatial data and the creation, sharing, and collaborative editing of "temporally-enriched" thematic maps which are produced by combining different sources of such data and other geospatial information available in vector or raster file formats.

TELEIOS was the first project internationally that introduced the Linked Data paradigm to the EO domain, and developed prototype applications that are based on transforming EO products into RDF, and combining them with linked geospatial data. Examples of such applications include wildfire monitoring and burnt scar mapping, semantic catalogues for EO archives, as well as rapid mapping.

More precisely, we developed a wildfire monitoring service (6) using only the Scientific Database and Linked Data technologies presented above [5]. The service is currently operational at the National Observatory of Athens and has been used during the last two fire seasons by decision makers and emergency response managers monitoring wildfires in Greece.

We developed a Virtual Earth Observatory for synthetic aperture radar images obtained by the satellite TerraSAR-X of the German Aerospace Center, that goes beyond existing EO portals and EO data management systems by allowing a user to express such complex queries as "Find all satellite images with patches containing water limited on the north by a port" that combine both satellite data and linked EO data. The abundance of linked EO data can prove useful to the new space missions (e.g., Sentinels) as a means to increase the usability of the millions of images and EO products that are expected to be produced by these missions.

Links:

(1) http://www.earthobservatory.eu/
(2) http://www.sciql.org/
(3) http://www.monetdb.org/
(4) http://www.strabon.di.uoa.gr/
(5) http://sextant.di.uoa.gr/
(6) http://bit.ly/FiresInGreece

References:

[1] M.L. Kersten, Y. Zhang, M. Ivanova, N.J. Nes: "SciQL, A Query Language for Science", Workshop on Array Databases 2011, EDBT, http://oai.cwi.nl/oai/asset/18678/18678A.pdf
[2] M. Ivanova, M.L. Kersten, S. Manegold, Y. Kargin: "Data Vaults: Database Technology for Scientific File Repositories", Computing in Science and Engineering, 2013, dx.doi.org/10.1109/MCSE.2013.17
[3] K. Kyzirakos, M. Karpathiotakis, M. Koubarakis: "Strabon: A Semantic Geospatial DBMS", ISWC, 2012, dx.doi.org/10.1007/978-3-642-35176-1_19
[4] C. Nikolaou and M. Koubarakis: "Incomplete Information in RDF", in RR, 2013, dx.doi.org/10.1007/978-3-642-39666-3_11
[5] M. Koubarakis, C. Kontoes, S. Manegold: "Real-time wildfire monitoring using scientific database and linked data technologies", EDBT, 2013, dx.doi.org/10.1145/2452376.2452452.

Please contact:

Kostis Kyzirakos, Stefan Manegold
CWI, The Netherlands
E-mail: Kostis.Kyzirakos@cwi.nl, Stefan.Manegold@cwi.nl

Charalambos Nikolaou, Manolis Koubarakis,
National and Kapodistrian University of Athens, Greece
E-mail: charnik@di.uoa.gr, koubarak@di.uoa.gr

GeoKnow: Making the Web an Exploratory Place for Geospatial Knowledge

by Spiros Athanasiou, Daniel Hladky, Giorgos Giannopoulos, Alejandra Garcia Rojas and Jens Lehmann

The GeoKnow project aims to make geospatial data accessible on the Web of Data, transforming the Web into a place where geospatial data can be published, queried, reasoned, and interlinked, according to Linked Data principles.

In recent years, Semantic Web methodologies and technologies have strengthened their position in the areas of data and knowledge management. Standards for organizing and querying semantic information, such as RDF(S) and SPARQL are adopted by large academic communities, while corporate vendors adopt semantic technologies to organize, expose, exchange and retrieve their data as Linked Data [1]. RDF stores have become robust enough to support volumes of billions of records (RDF triples), and also offer data management and querying functionalities very similar to those of traditional relational database systems. Currently, there are three major sources of open geospatial data in the Web: Spatial Data Infrastructures (SDI), open data catalogues, and crowdsourced initiatives. Crowdsourced geospatial data are emerging as a potentially valuable source of geospatial knowledge. Among various efforts we highlight OpenStreetMap, GeoNames, and Wikipedia as the most significant. Recently, GeoSPARQL [2] has emerged as a promising standard from W3C for geospatial RDF, with the aim of standardizing geospatial RDF data modelling and querying. Integrating Semantic Web with geospatial data management requires the scientific community to address two challenges: (i) the definition of proper standards and vocabularies that describe geospatial information according to RDF(S) and SPARQL protocols, that also conform to the principles of established geospatial standards, (e.g. OGC), (ii) the development of technologies for efficient storage, robust indexing, and native processing of semantically organized geospatial data.

Geoknow is an EU funded, three-year project that started in December 2012. While several research projects, such as LOD2[4], are supporting the Linked Data LifeCycle, Geoknow addresses the key issues of integrating geographically related information on the Web, scalable reasoning over billions of geographic features within the Linked Data Web, as well as efficient crowd-sourcing and collaborative authoring of geographic information. In particular, GeoKnow will apply the RDF model and the GeoSPARQL standard as the basis for representing and querying geospatial data and will contribute to the following areas:

- Efficient geospatial RDF querying: Existing RDF stores lack performance and geospatial analysis capabilities compared to geospatially-enabled relational DBMS. We will focus on introducing query optimization techniques for accelerating geospatial querying by at least an order of magnitude.
- Fusion and aggregation of geospatial RDF data: Given a number of different RDF geospatial data for a given region containing similar knowledge

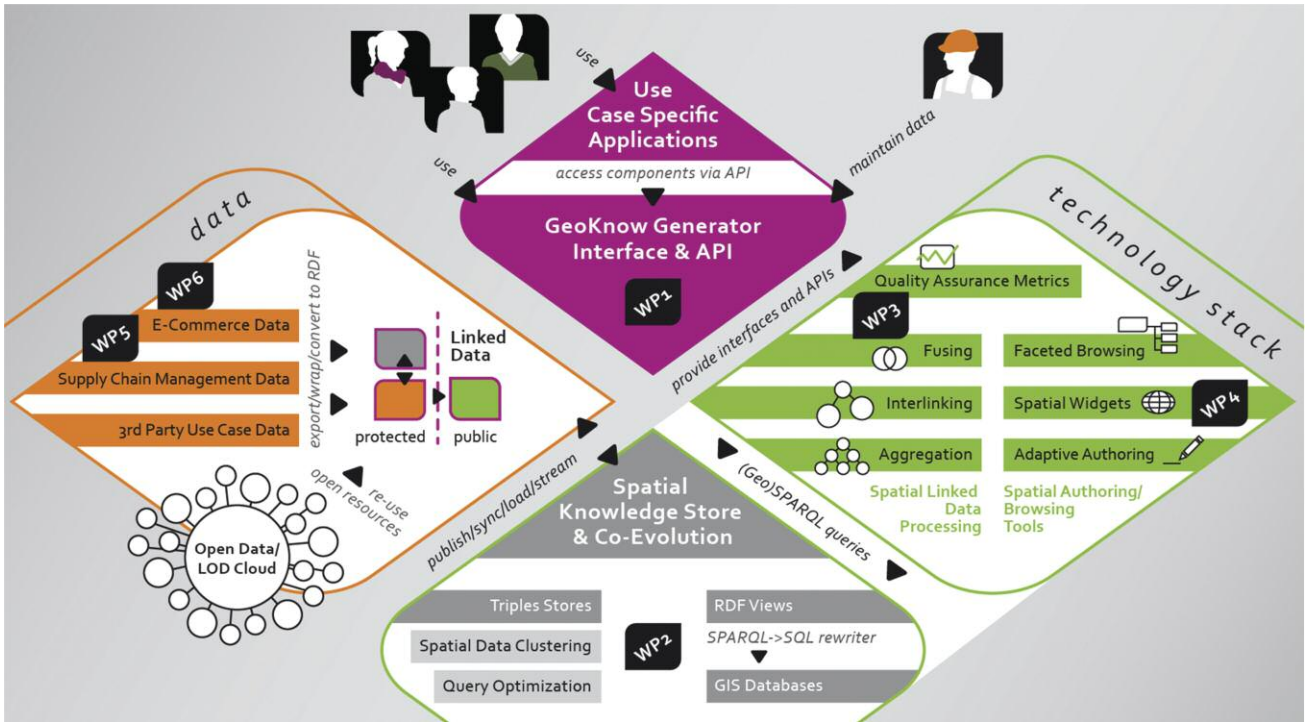


Figure 1: Overview of GeoKnow technologies

(e.g. OSM, PSI and closed data) we will devise automatic fusion and aggregation techniques in order to consolidate them and provide a dataset of increased value and quantitative quality metrics of this new data resource.

- Visualization and authoring: We will develop reusable mapping components, enabling the integration of geospatial RDF data as an additional data resource in web map publishing. Further, we will support expert and community-based authoring of RDF geospatial data within interactive maps, fully embracing crowdsourcing.
- Public-private geospatial data: To support value added services on top of open geospatial data, we will develop enterprise RDF data synchronization workflows that can integrate open geospatial RDF with closed, proprietary data.
- GeoKnow Generator: This will consist of a full suite of tools supporting the complete lifecycle of geospatial Linked Data. The GeoKnow Generator will enable publishers to triplify geospatial data, interlink them with geospatial and non-geospatial data sources, fuse and aggregate linked geospatial data to provide new data of increased quality, and finally, visualize and author link geospatial data in the Web.

The GeoKnow Generator

A prototype of the GeoKnow Generator is already available at http://generator.geoknow.eu. It allows the user to triplify geospatial data, such as ESRI shapefiles and spatial tables hosted in major DBMSs using the GeoSPARQL, WGS84 or Virtuoso RDF vocabulary for point features geospatial representations (TripleGeo). Non-geospatial data in RDF (local and online RDF files or SPARQL endpoints) or data from relational databases (via Sparqlify) can also be entered into the Generator's triple store. With these two sources of data it is possible to link (via LIMES), to enrich (via GeoLift), to query (via Virtuoso) , to visualize (via Facete) and to generate light-weight applications as JavaScript snippets (via Mappify) for specific geospatial applications. Most steps in the Linked Data lifecycle [1] have been integrated in the Generator as a graph-based workflow, which allows the user to easily manage new generated data. The components comprising it are available in the Linked Data Stack (http://stack.linkeddata.org)

Achievements and Future Work

Geoknow is concluding its first year and has already achieved important advancements. The first step was to perform a thorough evaluation of the current standards and technologies for managing geospatial RDF data and identify major shortcomings and challenges [3]. The next step has already produced significant output in the form of ready-for-use tools comprising the GeoKnow Generator. These components are being further enhanced and enriched. For example, Virtuoso RDF store is being extended in order to fully support OGC geometries and the GeoSPARQL standard and FAGI is being developed to support fusion of thematic and geospatial metadata of resources, either manually or automatically. Also, within 2014 the consortium will start testing the use cases and evaluating the performance and scalability of the GeoKnow Generator. Finally, future activities include, among others, the enhancement of the already developed frameworks, as well as the development of sophisticated tools for (a) aggregation of crowdsourced geospatial information and (b) exploration and visualization of spatio-temporal RDF data.

Acknowledgement

The research leading to these results has received funding under the European Commission's Seventh Framework Programme from ICT grant agreement (no. 318159) for GeoKnow. The consortium consists of the following partners: Institute of Applied Computer Science / University of Leipzig (Germany), Institute for the Management of Information Systems/Athena Research and Innovation Center (Greece), Open Link Software Ltd (United Kingdom), Unister GmbH (Germany), Brox (Germany), Ontos AG (Switzerland), and Institute Mihailo Pupin (Serbia).

Links:

GeoKnow project: http://geoknow.eu
LOD2 project: http://lod2.eu
Linked Data Stack:
http://stack.linkeddata.org
GeoKnow Github:
https://github.com/GeoKnow
Open Geospatial Consortium:
http://www.opengeospatial.org/

References:

[1] S. Auer, J. Lehmann: "Making the web a data washing machine - creating knowledge out of interlinked data", Semantic Web Journal, volume 1, number 12, p. 97-104, IOS Press, 2010, http://www.semantic-web-journal.net/sites/default/files/swj24_0.pdf
[2] M. Perry, J. Herring (eds): "OGC GeoSPARQL standard - A geographic query language for RDF data", Open Geospatial Consortium Inc, v.1.0, 27/04/2012, https://portal.opengeospatial.org/files/?artifact_id=47664
[3] K. Patroumpas et al.: "Market and Research Overview", GeoKnow EU/FP7 project deliverable 2.1.1, 2013, http://svn.aksw.org/projects/GeoKnow/Public/D2.1.1_Market_and_Research_Overview.pdf
[4] S. Auer et al.: "Managing the lifecycleof Linked Data with the LOD2 Stack", in proc. of ISWC'11, Springer, 2012
[5] A. G. Rojas, et al.: "GeoKnow: Leveraging Geospatial Data in the Web of Data", in Open Data Workshop, W3C, London, 2013.

Please contact:

Spiros Athanasiou
Institute for the Management of Information Systems
Athena Research Center, Greece
E-mail: spathan@imis.athena-innovation.gr

Daniel Hladky
Ontos AG, Switzerland
E-mail: daniel.hladky@ontos.com

Jens Lehmann
InfAI, University of Leipzig
Email: lehmann@informatik.uni-leipzig.de

Giorgos Giannopoulos
Institute for the Management of Information Systems
Athena Research Center, Greece
E-mail: giann@imis.athena-innovation.gr

Open Education: A Growing, High Impact Area for Linked Open Data

by Mathieu d'Aquin and Stefan Dietze

Education is now entering a revolution in the form of open education, where Linked Open Data has the potential to play a vital role. The Web of Linked Educational Data is growing with information about courses and resources, and emerging as a collective information backbone for open education.

Education has often been a keen adopter of new information and communication technologies. This is not surprising given that education is all about informing and communicating, and it is currently entering a revolution in the form of open education. This requires the use of state-of-the-art technologies for sharing, publishing and connecting information globally, free from technological barriers and cultural frontiers: namely, Linked Data [1].

Traditionally, educational institutions produce large volumes of data, much of which is publicly available, either because it is useful to communicate (e.g., the course catalogue) or because of external policies (e.g., reports to funding bodies). In this context, open data has an important role to play. Implementing open data through Linked Data technologies can be summarized as using the web both as a channel to access data (through URIs supporting the delivery of structured information) and as a platform for the representation and integration of data (through creating a graph of links between these data URIs). Considering the distribution and variety of providers (universities, schools, governments), topics (disciplines and types of educational data) and users (students, teachers, parents), education also represents a perfect use case for Linked Open Data.

One of the earliest examples of such a use of Linked Open Data in education can be found, unsurprisingly, at the Open University in the UK. The Open University is a 40 year old institution entirely dedicated to open and distance learning. It is also the largest University in the UK with more than 250,000 students per year. The amount of public information made available by such an organization is impressive, but even more striking is the way this information is siloed in different systems, with different channels of access to different parts for both the users and the developers of information systems within the University. In 2010, data.open.ac.uk was created, providing up-to-date Linked Data and a query endpoint (using the SPARQL protocol) about the courses on offer at the university, the available educational material and multimedia resources, the research outputs of its members and many other aspects of its academic activities. This system made it easier for both internal developers and users, and others, to use the information, which originates from different internal systems. It enabled innovative developments of new types of applications for the university, and made it possible to integrate, re-purpose and re-mix such information with others, from other educational institutions, reference data (e.g. geographical places) and government data (through linked data at data.gov.uk).

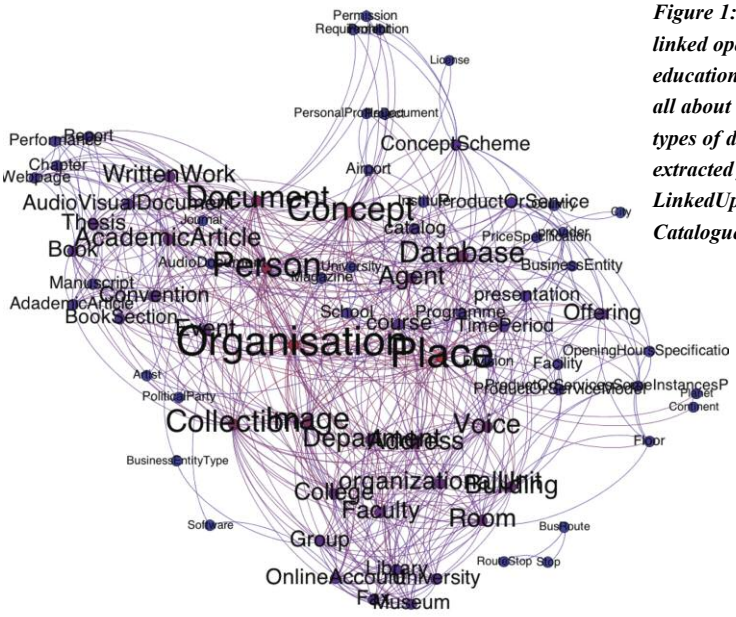


Figure 1: What linked open educational data is all about - graph of types of data entities extracted from the LinkedUp Data Catalogue.

What used to be an exception is now becoming increasingly common [2]. Universities around the world are joining the movement by deploying their own Linked Data platforms, and many other organizations and initiatives are also creating new Linked Data resources that will be of value to education.

Although beneficial to each institution, however, such bottom-up contributions to the landscape of educational Linked Data do not necessarily converge to a cohesive and truly connected global view, and ignore a fundamental question: What can be done with all of these data, combined and integrated from all of these contributions? This is the question the LinkedUp project aims to answer. This European project (led by the Leibniz Universität Hannover, and including The Open University, UK, The Open University of the Netherlands, Elsevier and Lattanzio Learning Spa, Italy) is collecting all of these educational Linked Data in a global data catalogue (see [3] for an initial view into the content of this catalogue), providing not only a unique access point to these isolated initiatives, but also aligning their vocabularies so that they can be used and queried jointly. The main activity of the project is then to organize a series of competitions. The LinkedUp Challenge, now running for the second time, asks developers of educational services to demonstrate what new, innovative applications can be built

out of all these newly available, linked and integrated data. The third and final challenge will take place in the second half of 2014.

Of course, it does not stop there, as new scenarios are continually emerging within open education, including both the new models for delivering education (open educational resources, MOOCs, etc.) and the increased mobility of students, leading traditional institutions to adopt more online and distance learning models. Here lies the real potential of Linked Data in education – reconciling these largely distributed and heterogeneous information spaces, all of value to students and prospective students, within one, global, easily accessible information space.

This new global and distributed environment for education does not fit very well the traditional, top down (government mandated) way of informing students about their options, and even less to engage with the providers, creators and developers of contents and services for education. As the results of the first LinkedUp competition demonstrate, using Linked Open Data both to support such contributions to open educational data and to coordinate them into a global, jointly usable data landscape, is showing strong promise as the information backbone for the new open education economy.

Links:
Linked data platform of the Open University: http://data.open.ac.uk
LinkedUp project: http://linkedup-project.eu
LinkedUp Challenge: http://linkedup-challenge.org
LinkedUp Data Catalog: http://data.linkededucation.org/linkedup/catalog/
Linked Universities portal: http://linkeduniversities.org
Linked Education portal: http://linkededucation.org
Open Education working group: http://education.okfn.org/

References:
[1] S. Dietze, et al.: "Interlinking educational resources and the web of data: A survey of challenges and approaches" Program: electronic library and information systems 47.1 (2013): 60-91, dx.doi.org/10.1108/00330331211296312
[2] M. d'Aquin: "Linked data for open and distance learning", Commonwealth of Learning report, 2012, http://www.col.org/resources/publications/Pages/detail.aspx?PID=420
[3] M. d'Aquin, et al. "Assessing the educational linked data landscape", in proc of ACM Web Science, 2013, dx.doi.org/10.1145/2464464.2464487

Please contact:
Mathieu d'Aquin
KMi, The Open University
E-mail: mathieu.daquin@open.ac.uk

Stefan Dietze
L3S Research Center, Leibniz University Hannover, Germany
E-mail: dietze@l3s.de

Raising the Stakes in Linked Data Education

by Alexander Mikroyannidis, John Domingue and Elena Simperl

There is currently a revolution going on in education generally, but nowhere more so than in the ICT field, owing to the availability of high quality online learning resources and MOOCs (Massive Open Online Courses). The EUCLID project is at the forefront of this initiative by developing a comprehensive educational curriculum, supported by multimodal learning materials and highly visible eLearning distribution channels, tailored to the real needs of data practitioners.

MOOCs (Massive Open Online Courses) offer large numbers of students the opportunity to study high quality courses with prestigious universities. These initiatives have led to widespread publicity as well as strategic dialogue in the higher education sector. The consensus within higher education is that after the Internet-induced revolutions in communication, business, entertainment and the media, it is now the turn of universities. Exactly where this revolution will lead is not yet known but some radical predictions have been made, including the end of the need for university campuses (http://www.theguardian.com/education/2012/nov/11/online-free-learning-end-of-university).

Linked Data [1] has established itself as the de facto means for the publication of structured data over the Web, enjoying amazing growth in terms of the number of organizations committing to use its core principles for exposing and interlinking Big Data for seamless exchange, integration, and reuse [2]. More and more ICT ventures offer innovative data management services on top of Linked Data, creating a demand for data scientists possessing skills and detailed knowledge in this area. Ensuring the availability of such expertise will prove crucial if businesses are to reap the full benefits of these advanced data management technologies, and the know-how accumulated over the past years by researchers, technology enthusiasts and early adopters.

The European project EUCLID contributes to this goal by developing a comprehensive educational curriculum, supported by multimodal learning materials and highly visible eLearning distribution channels, tailored to the real needs of data practitioners. The EUCLID curriculum focuses on techniques and software to integrate, query, and visualize Linked Data within core areas in which practitioners indicate that they require the most assistance. A significant part of the learning material comprises examples referring to real-world datasets and application scenarios, code snippets and demos that developers can run on their machines, as well as best practices and how-tos.

The EUCLID educational curriculum consists of a series of modules, each containing multi-format learning materials, such as presentations, webinars, screencasts, exercises, eBook chapters, and online courses. These learning materials complement each other and

are connected to deliver a comprehensive and concise training programme to the community. Learners are guided through these materials by following learning pathways, which are sequences of learning resources structured appropriately for achieving specific learning goals. Different types of eLearning distribution channels are targeted by each type of learning material, including Apple and Android tablets, Amazon Kindles, as well as standard web browsers.

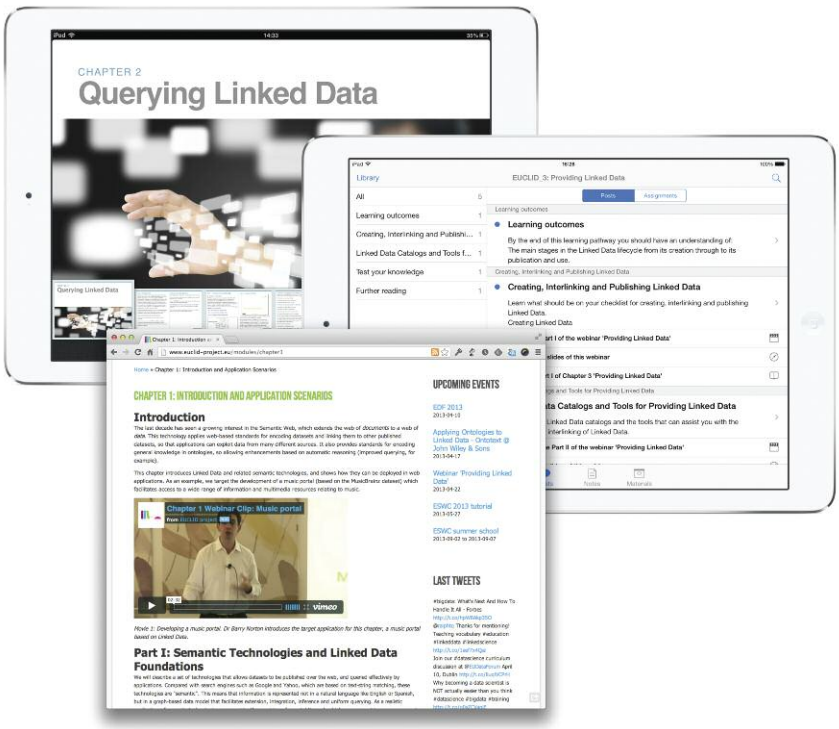


Figure 1: A selection of EUCLID learning materials in different formats and platforms, i.e. eBooks and online courses for the web and the iPad.

Instead of mock Linked Data examples, we use in our learning materials and exercises a collection of datasets and tools that are deployed and used in real life. In particular, we use a number of large datasets, including the MusicBrainz dataset, which contains 100Ms of triples. Our collection of tools includes Seevl, Sesame, Open Refine and GateCloud, all of which are used in real life contexts. We also showcase scalable solutions, based upon industrial-strength repositories and automatic translations, e.g. by using the W3C standard R2RML for generating RDF from large data contained in standard databases.

Additionally, EUCLID has a strong focus on the community and encourages community engagement in the production of learning materials through, for example, collecting user feedback via our webinars, Twitter, LinkedIn, and more. EUCLID combines online and real-world presence, and attempts to integrate with on-going activities in each sphere such as mailing lists and wikis. The project engages with the Linked Data community, both practitioners and academics, by collecting user requirements as well as feedback about the materials so that they can be tailored to what the learner really needs.

The EUCLID project consortium brings together renowned institutions in eLearning and Linked Data research. The project consortium is coordinated by STI Research (Austria) and consists of the Karlsruhe Institute of Technology (KIT - Germany), Ontotext (UK) and The Open University (UK). Additionally, the project has a number of associate partners, including Fluid Operations AG (fluidOps - Germany), University Simón Bolívar (Venezuela) and the University of Southampton (UK). EUCLID is supported by the Seventh Framework Programme. The project started in May 2012 and has a duration of two years.

Links:

The EUCLID web site, showcasing the project's learning materials:
http://www.euclid-project.eu/

The EUCLID channel on Vimeo, including all webinars and screencasts:
https://vimeo.com/euclidproject

The EUCLID channel on SlideShare, including all presentations:
http://www.slideshare.net/EUCLIDproject

The EUCLID community page on LinkedIn:
http://www.linkedin.com/groups/Education-Training-on-Semantic-Technologies-4917016

The EUCLID Twitter channel:
http://www.twitter.com/euclid_project

References:

[1] T. Berners-Lee: "Linked Data - Design Issues", 2006, http://www.w3.org/DesignIssues/LinkedData.html.

[2] C. Bizer, T. Heath, T. Berners-Lee, "Linked Data—The Story So Far," International Journal on Semantic Web and Information Systems, vol. 5, pp. 1–22, 2009, http://dx.doi.org/10.4018/jswis.2009081901

Please contact:

Alexander Mikroyannidis
Knowledge Media Institute, The Open University, UK
E-mail: Alexander.Mikroyannidis@open.ac.uk
http://alexmikro.net/about/

John Domingue
Knowledge Media Institute, The Open University, UK
E-mail: John.Domingue@open.ac.uk
http://people.kmi.open.ac.uk/domingue

Elena Simperl
Web and Internet Science group, University of Southampton, UK
E-mail: E.Simperl@soton.ac.uk
https://sites.google.com/site/elenasimperl/

RITMARE: Linked Open Data for Italian Marine Research

by Cristiano Fugazza, Alessandro Oggioni and Paola Carrara

The RITMARE (la Ricerca ITaliana per il MARE – Italian Research for the sea) Flagship Project is one of the National Research Programmes funded by the Italian Ministry of University and Research. Its goal is the interdisciplinary integration of national marine research. In order to design a flexible Spatial Data Infrastructure (SDI) that adapts to the audience's specificities, the necessary context information is drawn from existing RDF-based schemata and sources. This enables semantics-aware profiling of end-users and resources, thus allowing their provision as Linked Open Data.

RITMARE includes public research bodies, inter-university consortia and private companies involved in marine research. The project's objectives are:

- To support integrated policies for environmental protection (the health of the sea);
- To facilitate sustainable use of resources (the sea as a system of production);
- To implement a strategy of prevention and mitigation of natural impacts (the sea as a risk factor).

RITMARE is organized into seven sub-projects (SPs); SP7 is building an interoperable infrastructure for the Observation Network and Marine Data [1]. The infrastructure enables coordination and sharing of data, processes and information among all sub-projects. Harmonizing the practices and technologies already in use by the thousands of researchers within the RITMARE scientific community calls for the integration of heterogeneous data, metadata, workflows, and requirements.

Traditional online access with generic tools for all users does little to support friendly collaboration among researchers or their interaction needs. The RITMARE infrastructure is improving the interoperability, interdisciplinarity, and usability features via semantic tools tailored to the habits, skills and needs of researchers.

Expressing context information as RDF

In accordance with Italian legislation regarding Liked Open Data (LOD) [2], we have built an RDF knowledge base to ground the aforementioned semantic tools. It is composed of:

- Categorization of researchers, research institutes and the project's internal organization as Friend Of A Friend (FOAF) data structures (around 1,800 entities);
- Features relating to Italy or the Mediterranean Sea from the GeoNames ontology (around 40,000 entities);

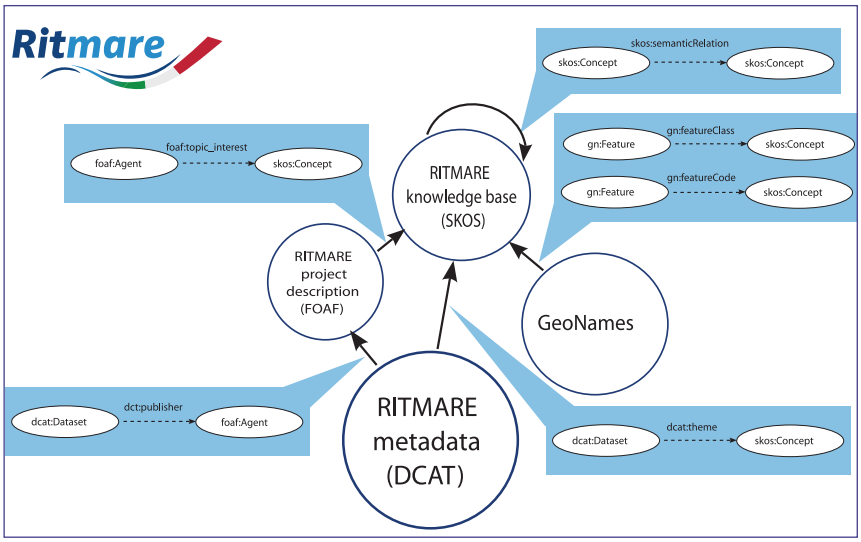


Figure 1: Context information in the RITMARE infrastructure.

- Observation parameters and measuring units from the British Oceanographic Data Centre (BODC) vocabulary repository, expressed as Simple Knowledge Organisation System (SKOS) concepts (around 30,000 entities);
- The research domains associated with the project, code lists and thesauri expressed as SKOS data structures (around 300 entities).

These data structures allow for semanatics-aware metadata descriptions in the Data Catalog Vocabulary (DCAT) format. In order to build a coherent structure from context information, these domains have been interconnected (see Figure 1):

- FOAF entities are related to SKOS concepts;
- Heterogeneous SKOS thesauri are aligned;
- DCAT metadata items reference entities from the above data structures and from the gazetteer.

The intended outcome is a repository of metadata from the heterogeneous, peripheral nodes of the RITMARE research network and the capacity for RDF-based metadata discovery.

Building a novel SDI user experience

One of the goals in the development of the RITMARE infrastructure is to provide the end-user with advanced functionalities on both the front- and back-end sides. Another main goal, building operating capacity by the providers of data and services, is beyond the scope of this article.

Front-end functionality includes user profiling, which allows for the creation of a customized user interface by relating users' FOAF descriptors to the SKOS concepts expressing research areas (top-left corner of Figure 1). The latter are associated with widgets in the interface by means of a matrix describing the appropriateness of a specific widget to a given research area. The layout of the interface can then be arranged according to a list of widgets ordered by relevance to the specific user.

By relating research areas to the entities referred to in metadata items, such as key-

words, observation parameters and measuring units (bottom part of the picture), it is then possible to populate widgets with content that is tailored to the user's profile (e.g., datasets and services applicable to the user's research area). Also, recourse to a gazetteer for expressing geographic locations simplifies discovery by reducing the need for a map.

On the back-end side, the architecture comprises collaboration tools for supporting activities that are not directly related to the discovery of resources but that, nevertheless, constitute essential phases in data production (e.g., the management of field work). Once users start enriching their profile data (even by simply using the infrastructure), the user experience should slowly but steadily converge to the user's expectations.

Acknowledgement
The activities described in this paper have been funded by the Italian Flagship Project RITMARE.

Links:
RITMARE Flagship Project: http://www.ritmare.it
Data Catalog Vocabulary (DCAT): http://www.w3.org/TR/vocab-dcat/
BODC webservices: http://seadatanet.maris2.nl/v_bodc_vocab/welcome.aspx

References:
[1] P. Carrara et al.: "An interoperable infrastructure for the Italian Marine Research", IMDIS 2013
[2] Commissione di Coordinamento SPC: "Linee guida per l'interoperabilità semantica attraverso i Linked Open Data", 2013

Please contact:
Cristiano Fugazza, CNR-IREA, Italy
E-mail: fugazza.c@irea.cnr.it

Lost in Semantics? Ballooning the Web of Data

by Florian Stegmaier, Kai Schlegel and Michael Granitzer

Although Linked Open Data has increased enormously in volume over recent years, there is still no single point of access for querying the over 200 SPARQL repositories. The Balloon project aims to create a Meta Web of Data focusing on structural information by crawling co-reference relationships in all registered and reachable Linked Data SPARQL endpoints. The current Linked Open Data cloud, although huge in size, offers poor service quality and is inadequately maintained, thus complicating access via SPARQL endpoints. This issue needs to be resolved before the Linked Open Data cloud can achieve its full potential.

Today's vision of a common Web of Data is largely attributable to the Linked Open Data movement. The first wave of the movement transformed silo-based portions of data into a plethora of open accessible and interlinked data sets. The community itself provided guidelines (e.g., 5 stars Open Data) as well as open source tools to foster interactions with the Web of data. Harmonization between those data sets has been established at the modelling level, with unified description schemes characterizing a formal syntax and common data semantic.

Without doubt, Linked Open Data is the de-facto standard to publish and interlink distributed datasets within the Web commonly exposed in SPARQL endpoints. However, a request considering the globally described data set is only possible with strong limitations:

- The distributed nature of the Linked Open Data cloud in combination with the large amount of reachable endpoints hinders novice users from interacting with the data.
- Following the Linked Data principle, specific URIs are in use to describe specific entities in the endpoints and are further resolvable to get further information on the given entity. The problem arises since each endpoint uses its own URI to describe the single semantic entities leading to semantic ambiguities.

One outcome of the EU FP7 CODE project is the Balloon framework. It tackles exactly this situation and aims to create a Meta Web of Data focusing on structural information. The basement for this is a crawled subset of the Linked Data cloud, resulting in a co-reference index as well as structural information. The main idea behind this index is to resolve the aforementioned semantic ambiguities by creating sets of semantically equivalent URIs to ease consumption of Linked Open Data. This is enabled by crawling information expressing the links between the endpoints. For this purpose, we consider a specific set of predicates, e.g., sameAs or exactMatch, to be relevant. The complete crawling process relies on SPARQL queries and considers each LOD endpoint registered at the CKAN platform. Here, RDF dumps are explicitly excluded. During the crawling, a clustering approach creates the co-reference clusters leading to a bi-directional view on the co-reference relationships and is the result of a continuous indexing process of SPARQL endpoints. In addition to properties defining the equality of URIs, the indexing service also takes into account properties that enable structural analysis on the data corpus, e.g., rdfs:subclass. On the basis of this data corpus, interesting modules and application scenarios can be defined. For instance, on-going research is focusing on the creation of the following two modules as starting point:

- Intelligent and on the fly query rewriting by utilizing co-reference clusters and SPARQL 1.1 Federated Query.
- Data analysis, e.g., retrieving common properties or super types for a given set of entities.

These modules are integrated in the overall Balloon platform and serve as a starting point for further applications. To foster community uptake and to increase available modules in the platform, the Balloon project along with the data corpus will soon be made available as open source project.

The idea of leveraging co-reference information is nothing new: The Silk

framework [1], SchemEX [2] and the well-known sameAs.org project proposed similar techniques. Nevertheless, the Balloon co-reference approach further considers consistent data provenance chains and the possibilities of cluster manipulations to enhance the overall quality and correctness. Further, the explicit limitation to LOD endpoints sets a clear focus on the data that is (in principle) retrievable, in contrast to RDF dumps that are not searchable out of the box.

While creating the co-reference index, we encountered several issues in the current Linked Open Data cloud. Missing maintenance of endpoints over years as well as a lack of quality of service hinders the Linked Open Data cloud from reaching its potential. Our findings gathered during the crawling process are in keeping with the current statistics provided by the LOD2 project of the Linked Open Data cloud: From a total of 700 official data sets, only approximately 210 are enclosed in a SPARQL endpoint and registered at the CKAN platform. Further, more than half of the available endpoints had to be excluded due to insufficient support of SPARQL as well as unattainability. Finally, only 112 endpoints have been actively crawled for co-reference information leading to a total of 22.4M distinct URIs (approx. 8.4M synonym groups). During the crawling phase we also encountered the need for a SPARQL feature lookup service. The main intention is to describe the actually supported retrieval abilities of an endpoint in a standardized way. Currently, discussions on this topic are observable at community mailing lists.

Links:
Code Project: http://code-research.eu/
Overview of Balloon: http://schlegel.github.io/balloon/
Crawled data: ftp://moldau.dimis.fim.uni-passau.de/data/ (on-going research, frequently/live updated)
5 stars Open Data: http://5stardata.info/
CKAN platform: http://ckan.org/
LOD2 project: http://stats.lod2.eu/
Community mailing lists: http://lists.w3.org/Archives/Public/public-lod/2013Oct/0077.html

References:
[1] J. Volz et al: "Silk–a link discovery framework for the web of data," in proc. of the 2nd Linked Data on the Web Workshop, 2009, pp. 559–572, http://events.linkeddata.org/ldow2009/papers/ldow2009_paper13.pdf

[2] M. Konrath, et al: "Schemex efficient construction of a data catalogue by stream-based indexing of linked data," Web Semantics: Science, Services and Agents on the World Wide Web, vol. 16, no. 5, 2012, http://www.websemanticsjournal.org/index.php/ps/article/view/296/297

Please contact:
Florian Stegmaier, Kai Schlegel, Michael Granitzer
University of Passau, Germany
Tel: +49 851 509 3063
E-mail: <forename.surname>@uni-passau.de

Publishing Greek Census Data as Linked Open Data

by Irene Petrou and George Papastefanatos

Linked Open Data technology is an emerging way of making structured data available on the Web. This project aims to develop a generic methodology for publishing statistical datasets, mainly stored in tabular formats (e.g., csv and excel files) and relational databases, as LOD. We build statistical vocabularies and LOD storage technologies on top of existing publishing tools to ease the process of publishing these data. Our efforts focus on census data collected during Greece's 2011 Census Survey and provided by the Hellenic Statistical Authority. We develop a platform through which the Greek Census Data are converted, interlinked and published.

Statistical or fact-based data about observations of socioeconomic indicators are maintained by statistical agencies and organizations, and are harvested via surveys or aggregated from other sources. Census data include demographic, economic, housing and household information, as well as a set of indices concerning the population over time, such as mortality, dependency rate, total fertility rate, life expectancy at birth, etc.

The main objective in publishing socio-demographic data, such as census data, as LOD is to make these data available in an easier-to-process format (they can be crawled or queried via SPARQL), to be identifiable at the record level through their assignment with URIs and finally to be citable, ie, make it possible for other sources to link and connect with them. Being available in LOD format will make them easier to access and use by third parties, facilitating data exploration and the development of novel applications. Furthermore, publishing Greek census data as LOD will facilitate their comparison and linkage with datasets derived from other administrative resources (e.g. public bodies, Eurostat, etc.), and deliver consistency and uniformity between current and future census results.

Best practices for publishing Linked Data encourage the reuse of vocabularies for describing common concepts in a specific domain. In this way, interoperability and interlinking between published datasets is achieved. In the statistics field, a number of statistical vocabularies and interoperability standards have been proposed, such as the SDMX (Statistical data and metadata exchange) standard, the Data Cube Vocabulary, and SCOVO. In our approach, we employ the Data Cube Vocabulary for representing census results. The Data Cube Vocabulary relies on the multidimensional (or cube) model. Its main components are the

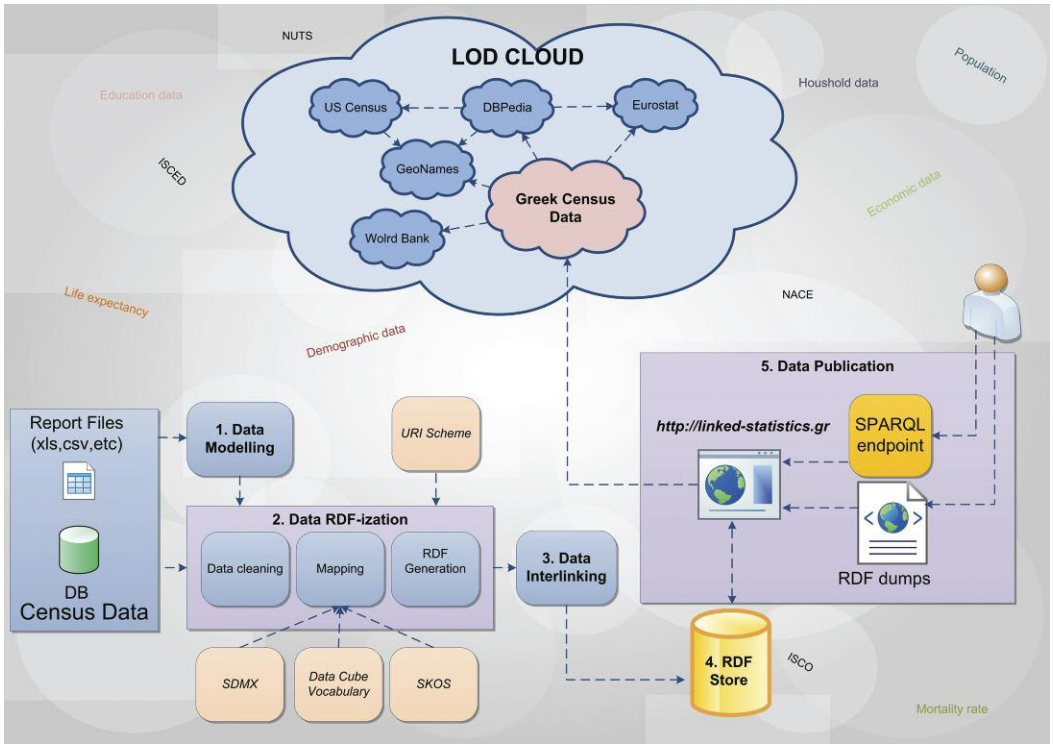


Figure 1: Publishing Statistical Data as LOD

dimensions, the attributes and the measures. Dimension components capture common characteristics across datasets, such as the reference period or the reference location; an attribute component captures attributes of the observed value(s), such as the unit of measure. Finally, a measure component represents the phenomenon being observed, such as the number of inhabitants. Data Cube Vocabulary, furthermore, uses SKOS and SDMX concepts for defining classifications, hierarchies and common statistical concepts.

To publish the data we apply the proposed methodology in Figure 1, comprising the following steps:

- Data modelling: This step involves identifying and modelling custom ontologies for all census-specific concepts and indices, which are not defined by other sources. An important part of this step is to tackle problems related to the evolution of the concepts (both in terms of structure and data values) over time. A typical example concerns the structure of administrative divisions in Greece: in the 2001 Census Survey the divisions were defined according to "KAPODISTRIAS" Plan containing six hierarchy levels of divisions, whereas in 2011, restructuring according to "KALLIKRATIS" Plan resulted in eight levels.
- Data RDF-ization: This step involves cleaning up the data, the selection of the appropriate URI Scheme for each type of resource (e.g., datasets, dimensions, observations, etc.) and the mapping of each concept within the source file (e.g., a column in case of xls files) either to the appropriate component of the Data Cube Vocabulary or to a concept of other related vocabulary (SDMX, SKOS). The data are then exported to RDF. The mapping along with the RDF generation is done within the custom platform developed for data transformation in real-time.
- Data interlinking: The transformed data are interlinked with other resources. For example, indices are linked with datasets from World Bank and economic activities, occupational and educational data are linked with Eurostat's datasets via the NACE, ISCO and ISCED classifications, respectively.
- Data storage: The produced RDF data are uploaded, stored and maintained in a LOD triple store. OpenLink Virtuoso is used for storing and dereferencing data.
- Data publication: Finally, the data become available for dereferencing and further exploration through a SPARQL endpoint service and for downloading as RDF dumps.

The current work is implemented in the context of a national large scale project regarding the management of socio-demographic data in Greece. The project is co-financed by the European Union (European Regional Development Fund - ERDF) and Greek national funds through the Operational Program "Competitiveness and Entrepreneurship" (OPCE II) of the National Strategic Reference Framework (NSRF) - Research Funding Program: KRIPIS. The project started at the beginning of 2013 and will run over the next two years. It involves the Institute for the Management of Information Systems at the Research Center "Athena", and the Institute of Social Research of the National Centre for Social Research.

Links:

http://linked-statistics.gr
http://www.statistics.gr/portal/page/portal/ESYE
http://www.w3.org/TR/2013/CR-vocab-data-cube-20130625/

Reference:

[1] I. Petrou, G. Papastefanatos, T. Dalamagas: "Publishing Census as Linked Open Data. A Case Study", in proc. of the 2nd Int. Workshop on Open Data (WOD'13), Paris, France, 2013.

Please contact:

George Papastefanatos,
Athena Research Centre, Greece
E-mail:
gpapas@imis.athena-innovation.gr

Irene Petrou,
Athena Research Centre, Greece
E-mail:
irene.p@imis.athena-innovation.gr

Linked Open Vocabularies

by Pierre-Yves Vandenbussche and Bernard Vatant

The "Web of Data" has recently undergone rapid growth with the publication of large datasets – often as Linked Data - by public institutions around the world. One of the major barriers to the deployment of Linked Data is the difficulty data publishers have in determining which vocabularies to use to describe the semantics of data. The Linked Open Vocabularies (LOV) initiative stands as an innovative observatory for the re-usable linked vocabularies ecosystem. The initiative goes beyond collecting and highlighting vocabulary metadata. It now plays a major social role in promoting good practice and improving overall ecosystem quality.

The last few years have seen the emergence of a "Web of Data". Open government transparency initiatives, such as data.gov (US) and data.gov.uk (UK), have played a key role in its emergence, together with a diverse range of players including: crowd sourcing projects (e.g. DBpedia), heritage organizations (e.g. Europeana, Library of Congress) and Web companies (e.g. schema.org). This development has been facilitated by Semantic Web technologies and standards for exposing, sharing and connecting data. In particular, the adoption of Linked Data best practices has bridged the gap between separately maintained data silos describing people, places, music, movies, books, companies, etc. Publishing data on the Web as Linked Data makes it easy for other organizations and data providers to create detailed links to your data (and vice-versa) and to make your data interoperable in other contexts, resulting in your data being more visible and reusable.

Initiated in March 2011, within the framework of the DataLift research project [1] hosted by the Open Knowledge Foundation, the Linked Open Vocabularies (LOV) initiative is now standing as an innovative observatory of the vocabulary ecosystem. It gathers and makes visible indicators that have not previously been harvested, such as interconnection between vocabularies, versioning history and maintenance policy, and, where relevant, past and current referents (individual or organization).

LOV's features include:

- *Documentation:* the best way to publish information about a vocabulary is to formally declare the metadata in the vocabulary itself [2]. The documentation assists any user in the task of understanding the semantics of each vocabulary term and therefore of the data using it. For instance, information about the creator and publisher is a key indication for a vocabulary user in case help or clarification is required from the author, or to assess the stability of that artefact. About 55% of vocabularies specify at least one creator, contributor or editor. We augmented this information using not formally defined and manually gathered information, leading to inclusion of data about the creator in over 85% of vocabularies in LOV.
- *Versions:* the LOV database stores every different version of a vocabulary over time since its first issue. For each version, a user can access the file (even though the original online file is no longer available) and a log of modifications since the previous version.
- *Dependencies:* the very nature of the Web is distributed and uncontrolled. To embrace the complexity of the vocabulary ecosystem and assess the impact of a modification, one needs to know in which vocabularies and datasets a particular vocabulary term

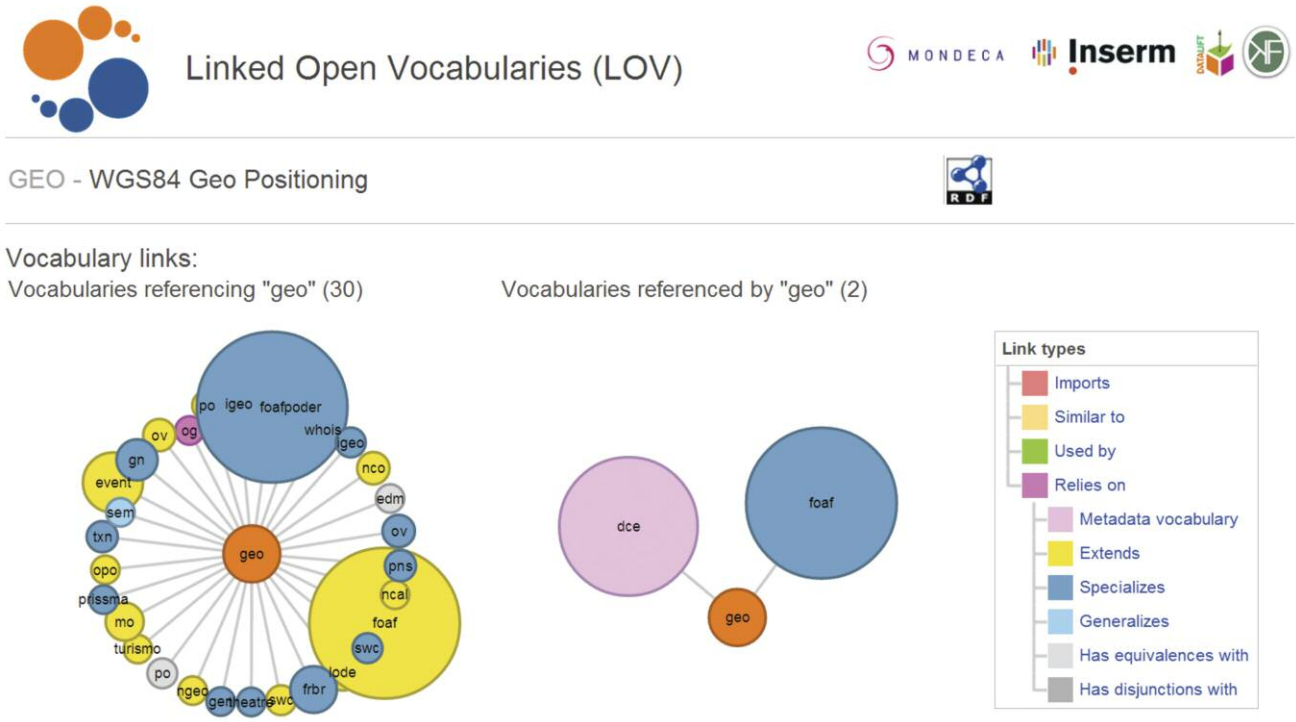


Figure 1: Linked Open Vocabularies

is referenced. For the first time LOV provides such a vision.
- *Search:* the LOV search feature queries a repository which contains the entire vocabulary ecosystem along with LOV metadata and metrics of vocabulary terms used in the Linked Open Data cloud. To help users in the selection of a vocabulary term, the results are ordered by a ranking algorithm based on the term popularity in the LOD datasets and in the LOV ecosystem.

All data produced within the LOV initiative are published and openly available for the community. LOV has opened new paths in using vocabularies for Linked Open Data representation by offering new search features based on rich metadata, social support and by fostering the "long tail" of vocabularies that have thus far remained unknown despite their high quality and potential usefulness. Our results reveal the high diversity of practices, both technical and social, taking place in the life cycle of vocabularies [3]. They highlight both the healthy interconnectivity of organic growth, and a certain number of pitfalls and potential points of failure in the ecosystem. Furthermore, results show encouraging signs of a growing awareness in the community of the importance of keeping the whole ecosystem alive and sustainable, through ways of governance which, for the most part, are yet to be invented

Links:
http://lov.okfn.org/dataset/lov/
http://datalift.org/en/
http://okfn.org/

References:
[1] F. Scharffe et al.: "Enabling linked-data publication with the datalift platform" in proc. AAAI workshop on semantic cities, 2012, http://www.aaai.org/ocs/index.php/WS/AAAIW12/paper/view/5349/5678
[2] P.-Y. Vandenbussche, B. Vatant: "Metadata recommendations for linked open data vocabularies", white paper v1.1., 2012, http://lov.okfn.org/dataset/lov/Recommendations_Vocabulary_Design.pdf
[3] M. C. Suárez-Figueroa, A. Gómez-Pérez: "NeOn methodology for building ontology networks: a scenario-based methodology", in proc. of the International Conference on Software, Services & Semantic Technologies, 2009, http://oa.upm.es/5475/1/INVE_MEM_2009_64399.pdf

Please contact:
Pierre-Yves Vandenbussche
Fujitsu Limited, Ireland
E-mail: pierre-yves.vandenbussche@ie.fujitsu.com

Linking Historical Entities to the Linked Open Data Cloud

by Maarten Marx

We investigate the coverage of Wikipedia for historical public figures. Unsurprisingly, the probability of a figure having a Wikipedia entry declines with time since the person was active. Nevertheless, two thirds of the Dutch members of parliament that have been active in the last 140 years have a Wikipedia page. The need to link historical figures to existing knowledge bases like Wikipedia/DBpedia comes from current large scale efforts to digitize primary data sources, including proceedings of parliament and historical newspapers. Linking entries to knowledge bases can provide values of key background variables, such as gender, age, and (party) affiliation.

The term "wikification" [1] refers to the process of automatically creating links from words or phrases in free text to their appropriate Wikipedia page. The common motivation for wikification is that a reader may want to consult additional (background) information about the phrase while reading the text. Typical candidates for wikification are named entities and rare terms. Another motivation for wikification comes from information retrieval: linking named entities in texts to external knowledge bases can improve both precision (by disambiguating names) and recall (by including spelling variants obtained from the knowledge page). A third motivation comes from the new fields of Computational Humanities and Computational Social Science [2] and is driven by current large-scale efforts to digitize primary historical sources and archives. Primary sources typically do not contain common background information about the entities (persons, organizations) mentioned in the sources. For instance, for each word spoken in the proceedings of the British parliament (Hansards), the name of the speaker and the speaker's constituency are recorded. But key variables such as age, gender, and even party affiliation are not recorded.

Wikification of Historical Texts: Research Challenges

Clearly, if wikification is desirable for texts from our own age, it is even more so for digitized historical archives. Performing wikification on historical scanned documents has several challenges that are not present in modern digital material: named entity recognition (NER) must deal with OCR-errors, spelling reforms and mismatches in language use in modern text material (on which NER taggers are trained) and historical texts. Disambiguating named entities (a key part in linking to external knowledge bases) may be harder because less background information is available. Finally, it may be that there is simply no data to link to. This latter aspect is addressed in greater detail below.

Coverage of Historical Persons in Wikipedia

We investigated Wikipedia's coverage of historical public figures using the

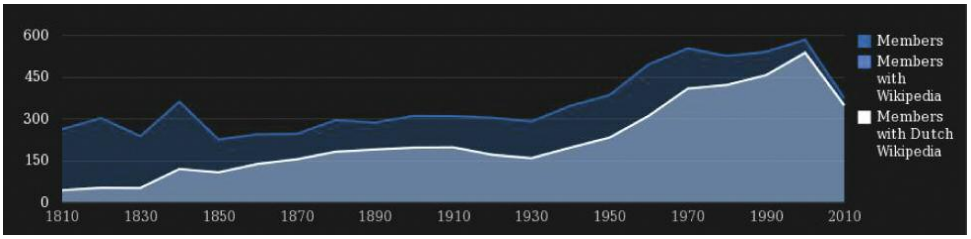


Figure 1: Coverage of Wikipedia for Dutch politicians in the period 1810-2013.

Politician	Party	# languages
Drs. M. (Mark) Rutte	(VVD)	40
G. (Geert) Wilders	(VVD)	30
Drs. A. (Ayaan) Hirsi Ali	(VVD)	23
Prof.Mr. J.G. (Jaap) de Hoop Scheffer	(CDA)	21
Dr. W. (Willem) Drees	(SDAP)	15
Drs. N. (Neelie) Kroes	(VVD)	15
Dr. A. (Abraham) Kuyper	(antirevolutionair)	14
Dr. J. (Jelle) Zijlstra	(ARP)	14
Mr. W. de Sitter	(liberaal)	13
Prof. Dr. Mr. F. (Frits) Bolkestein	(VVD)	13

Table 1: Number of Wikipedia pages in different languages per politician. Top 10 of the Dutch politicians.

Dutch parliamentary proceedings, which are available as scanned and OCRed PDF files from 1814 (http://www.statengeneraaldigitaal. nl/). Speakers in the proceedings are linked to a biographical database (http://www.parlement.com), which has an entry for each MP in this period. We linked the full names of the MPs taken from this database to the Dutch Wikipedia using the linking methodology of [3]. This software yields a ranked list of Wikipedia page candidates for each input. We filtered this list using a "political biography page" filter, which effectively removed false positives. Automatic checking of correctness of the links was facilitated by the fact that the majority of political biography pages on Wikipedia linked back to the database we used. Manual inspection and search on Wikipedia showed that with this automatic process we discovered virtually all existing biographical Wikipages.

Figure 1 presents the results. We grouped MPs in ten year periods. The top line in the graph shows the number of MPs that were active for at least one day during each period. The line below shows the number of them having a Wikipedia page. We have a coverage of over 90% for the period after 2000, at least 66% for the period 1850-2000 and a minimum of 16% for the period 1810-1820. The dip in 2010 is caused by the fact that this period consists only of three years.

The Most International Dutch Politicians

To illustrate the benefits of having the links we show the top 10 Dutch politicians with Wikipedia pages in most languages, an indicator of how well-known these people are internationally. This top 10 is made up of a rather varied club of politicians. First place is occupied by the current prime minister; second and third by leading anti-Muslim politicians. Fourth is the former secretary general of NATO. Place five shows the first person who is no longer alive: the Dutch PM in the period 1948-1958. Positions seven, eight and nine also have historical rather than modern politicians.

Conclusion

Linking entities occurring in historical material to present day knowledge bases like Wikipedia is an exciting and rewarding research field with great potential benefits. Wikipedia is rich enough to cover at least two thirds of the Dutch MPs active in the last 140 years. The linking process has the additional benefit of showing gaps in existing knowledge bases.

References:

[1] R. Mihalcea, A. Csomai: "Wikify!: Linking documents to encyclopedic knowledge", in proc. of CIKM '07, pp 233-242, 2007, http://dx.doi.org/10.1145/1321440.1321475

[2] D. Lazer, et al.: "Computational social science", Science, 323(5915):721-723, 2009, http://dx.doi.org/10.1126/science.1167742

[3] E. Meij, W. Weerkamp, M. de Rijke: "Adding semantics to microblog posts", in proc. of WSDM 2012, http://dx.doi.org/10.1145/2124295.2124364

Please contact:

Maarten Marx
University of Amsterdam, The Netherlands
E-mail: maartenmarx@uva.nl

Benchmarking Linked Open Data Management Systems

by Renzo Angles, Minh-Duc Pham and Peter Boncz

With inherent support for storing and analysing highly interconnected data, graph and RDF databases appear as natural solutions for developing Linked Open Data applications. However, current benchmarks for these database technologies do not fully attain the desirable characteristics in industrial-strength benchmarks [1] (e.g. relevance, verifiability, etc.) and typically do not model scenarios characterized by complex queries over skewed and highly correlated data [2]. The Linked Data Benchmark Council (LDBC) is an EU FP7 ICT project that brings together a community of academic researchers and industry, whose main objective is the development of industrial-strength benchmarks for graph and RDF databases.

Objective, well-designed and good quality benchmarks are important to fairly compare the performance of software products and uncover useful insights related to their strengths as well as their limitations. They encourage the advancement of technology by providing both academy and industry with clear targets for performance and functionality.

The Linked Data Benchmark Council (LDBC) aims to create benchmarks following principles including relevance, simplicity, fairness and sustainability. In particular, a goal of LDBC is to develop benchmarks that test critical usability features more thoroughly than the benchmarks so far produced by academia, including provision of a mechanism that ensures benchmarking results are reviewed for conformance by independent auditing. To this end, LDBC will provide open source benchmarks, developed by task forces integrated by expert architects who know the critical functionalities inside data management engines ("choke points"), and supported by a Technical User Community (TUC) that provides use-cases and feedback.

The Social Network Benchmark (SNB) is an LDBC benchmark intended to test various functionalities of systems used for graph data management. The scenario of this benchmark, a social network, is chosen with the following goals in mind: it should be understandable to a large audience, and this audience should also understand the relevance of managing such data; the scenario in the benchmark should cover a complete range of interesting challenges, according to the benchmark scope; and the query challenges in it should be realistic in the sense that, though synthetic, similar data and workloads are encountered in practice.

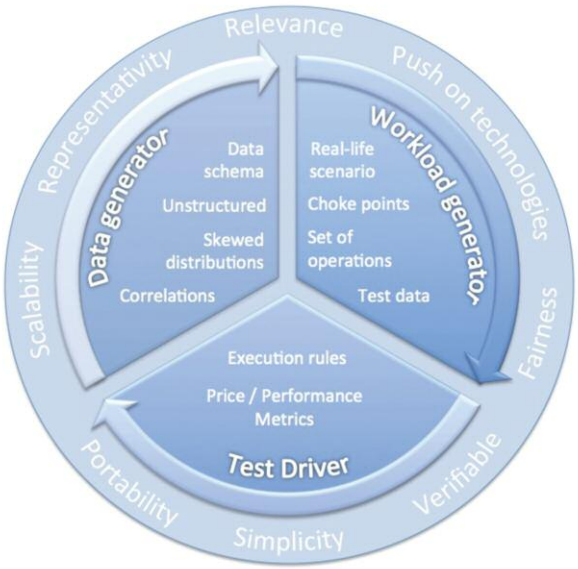


Figure 1: Elements and features of the LDBC Social Network Benchmark.

The SNB is composed of three main elements: a data generator, which allows creation of data according to a given data schema; a workload generator, which defines the set of operations that the System Under Test (SUT) has to perform; and a test driver, which is used to execute the workload over the SUT, and measure its performance according to well-defined performance metrics and execution rules. The features of these elements are summarized in Figure 1.

The SNB data generator is being designed to create synthetic data with the following characteristics: the schema must be representative of a real social network; the generation method must consider properties of real-life data, including data correlations and distributions; and the software generator must be easy-to-use, configurable and scalable. By leveraging parallelism through Hadoop, the current version of the data generator (based on the S3G2 generator [3]) ensures fast and scalable generation of huge datasets, allowing a social network structure with millions of user profiles, enriched with interests/tags, posts, and comments (see an example in Figure 2). Additionally, the generated data exhibits interesting realistic value correlations (e.g. German people having predominantly German names), structural correlations (e.g. friends being mostly people living close to one another), and statistical distributions (e.g. the friendship relationship between people follows a power-law distribution).

Aiming at covering all the main aspects of social network data management, the SNB provides three different workloads: an interactive workload, oriented to test the throughput of the systems with relatively simple queries and concurrent updates; a business intelligence workload, consisting of complex structured queries for analysing online behaviour of users for marketing purposes; and a graph analytics workload, thought to test the functionality and scalability of the systems for graph analytics that typically cannot be expressed in a query language. Each workload will be designed based on well-identified key technical challenges called "choke points". The objective is to ensure that the workload stresses important technical functionalities of actual systems.

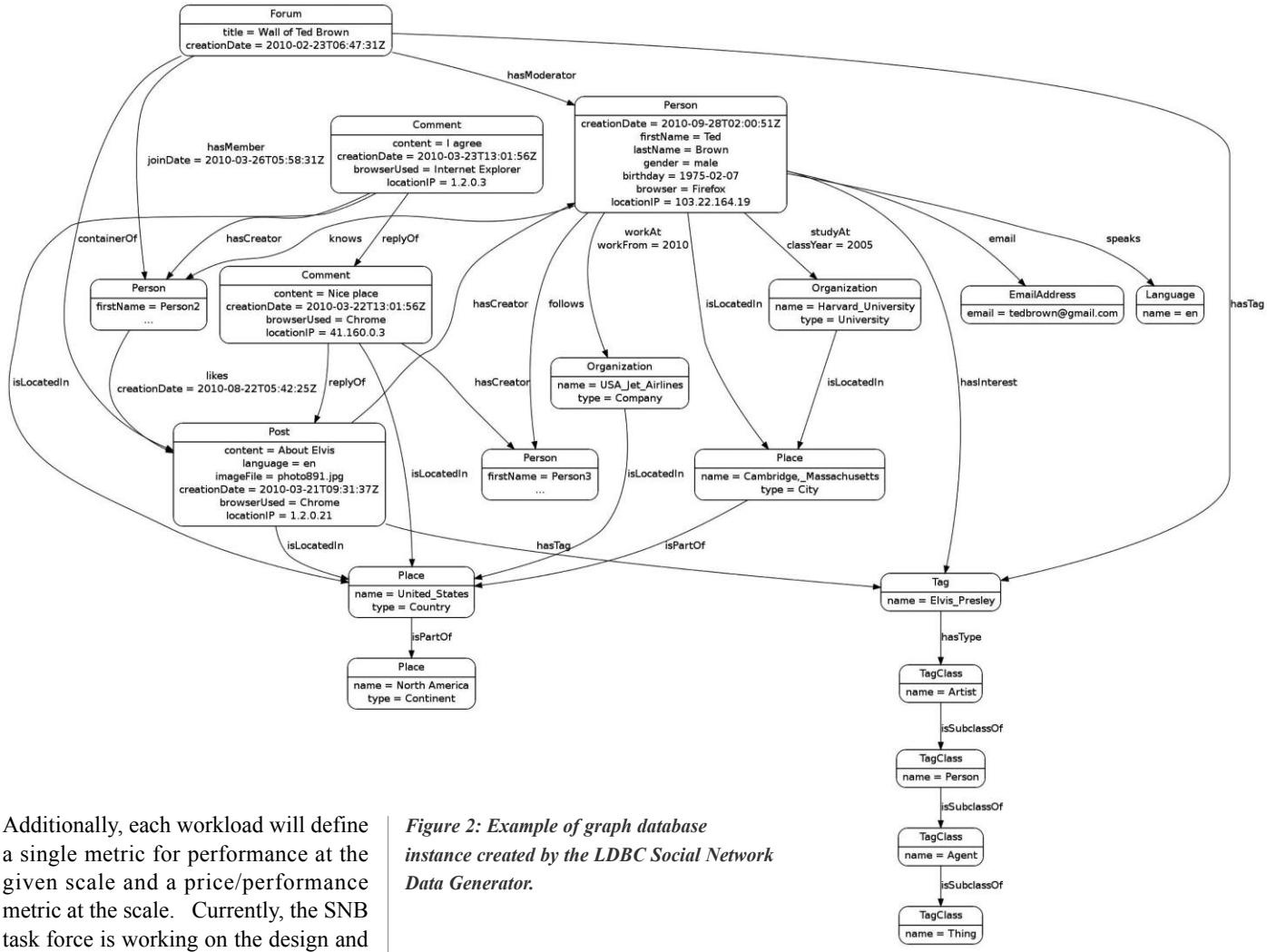


Figure 2: Example of graph database instance created by the LDBC Social Network Data Generator.

Additionally, each workload will define a single metric for performance at the given scale and a price/performance metric at the scale. Currently, the SNB task force is working on the design and implementation of the interactive workload generator. The first version of the interactive workload specification consists of twelve queries, inspired by a collection of choke points, which include "classical" complex operations (e.g. aggregated queries) and "non-traditional" complex operations (e.g. graph traversals). Besides the operations, it is also relevant to define smart methods for test data selection (ie data specifically selected to be used as substitution parameters for the operations). Test data must be carefully selected to obtain comparable results, and hence ensure the repeatability and fairness of the benchmark.

The SNB can be downloaded from GitHub and more information about its development is available in the TUC Wiki (see Links). We would like to invite readers to join the LDBC community initiative by sharing their user experience, testing their systems and participating in the LDBC-related events.

Links:
http://www.ldbc.eu
ldbc.eu:8090/display/TUC
github.com/ldbc

References:
[1] K. Huppler: "The Art of Building a Good Benchmark", in: TPCTC, 2009.
[2] S. Duan et al.: "Apples and Oranges: A Comparison of RDF Benchmarks and Real RDF Datasets", in: ACM SIGMOD, 2011.
[3] M.-D. Pham et al.: "A Scalable Structure-Correlated Social Graph Generator", in TPCTC, 2012.

Please contact:
Renzo Angles
VU University Amsterdam (Netherlands) / Universidad de Talca (Chile)
E-mail: r.anglesrojas@vu.nl

Making it Easier to Discover, Re-Use and Understand Search Engine Experimental Evaluation Data

by Nicola Ferro and Gianmaria Silvello

Experimental evaluation of search engines produces scientific data that are highly valuable from both a research and financial point of view. They need to be interpreted and exploited over a large time-frame, and a crucial goal is to ensure their curation and enrichment via inter-linking with relevant resources in order to harness their full potential. To this end, we exploit the LOD paradigm for increasing experimental data discoverability, understandability and re-usability.

Experimental evaluation of multilingual and multimedia information access systems is a demanding activity that benefits from shared infrastructures and datasets that favour the adoption of common resources, allow for replication of the experiments, and foster comparison among state-of-the-art approaches. Therefore, experimental evaluation is carried out in large-scale evaluation campaigns at an international level, such as the Text REtrieval Conference (TREC) in the United States and the Conference and Labs of the Evaluation Forum (CLEF) in Europe.

Figure 1 shows the main phases of the experimental evaluation workflow, where the information space entailed by evaluation campaigns can be considered in the light of the "Data Information Knowledge and Wisdom" (DIKW) hierarchy, used as a model to organize the produced information resources [1]. Each phase is carried out by people with different roles and produces scientific data that need to be managed, curated, accessed and re-used.

As a consequence, experimental evaluation has a big scientific and financial impact. From a scientific point of view, it has provided sizable improvements to key technologies, such as indexing, ranking, multilingual search, enterprise search, expert finding, and so on. From a financial point of view, it has been estimated that for every $1.00 invested in TREC, at least $3.35 to $5.07 in benefits accrued to researchers and industry, meaning that, for an overall investment in TREC of around 30 million dollars over 20 years, between 90 and 150 million dollars of benefits have been produced.

A crucial goal, therefore, is to ensure the best possible exploitation and interpretation of such valuable scientific data, possibly over large time spans. To this end,

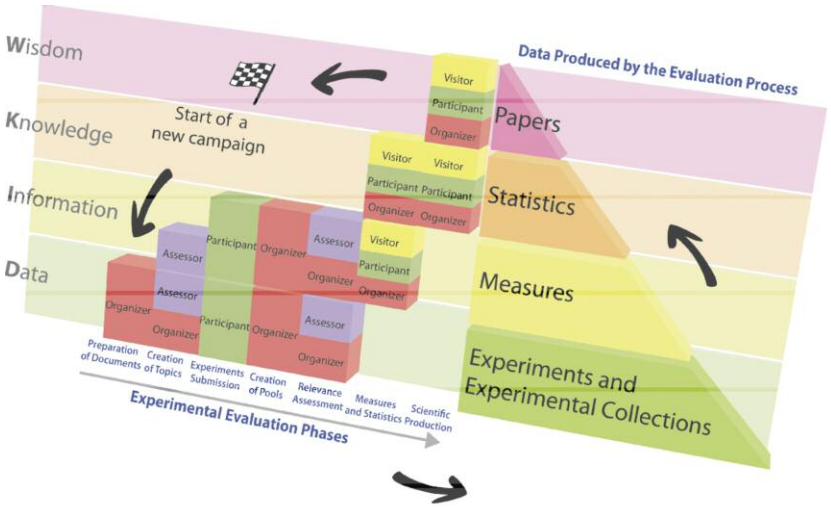


Figure 1: The main phases of the experimental evaluation workflow, the roles involved and the scientific data produced.

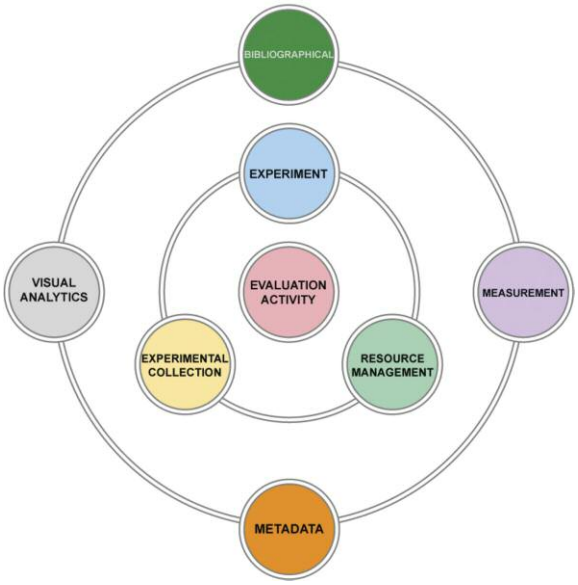


Figure 2: The eight functional areas of the DIRECT conceptual schema, which allows for representing, managing and accessing the scientific data produced by experimental evaluation.

we have been modelling the experimental data and designing a software infrastructure to manage and curate them. This has led to the development of the DIRECT system [2]. Figure 2 reports the eight functional areas of the DIRECT conceptual schema [3], which allows the representation and management of the evaluation workflow along with the data produced, as reported in Figure 1. Not only do they cover the scientific data in a strict sense but they also address follow-up activities, such as data visualization and interaction and scientific and bibliographical production.

Some items of information - in particular, scientific data and literature - that are built upon these data grow and evolve over time. In order to make the information as easy as possible for users

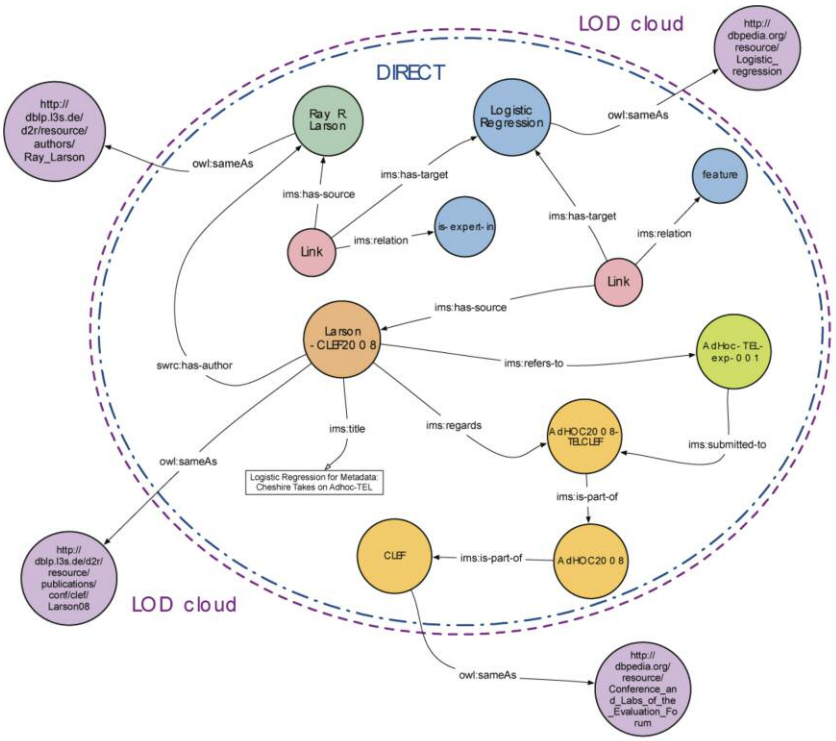


Figure 3: Discovering, understanding and re-using enriched experimental evaluation data as LOD exposed on the Web: a use-case..

to discover, re-use and understand, thereby maximizing its potential, we need to ensure that the information is promptly curated, enriched, and updated with links to other relevant information resources.

In order to tackle these issues, we mapped the DIRECT conceptual model into an RDF schema for representing experimental data and exposing them as LOD on the Web. This enables a seamless integration of datasets produced by different international campaigns as well as the standardization of terms and concepts used to label data across research groups. Furthermore, we adopted automatic data enrichment methodologies focused on finding experts and topics in the produced scientific literature. These techniques also allow us to keep the experimental evaluation data continuously updated and linked in a timely manner to the LOD cloud.

Figure 3 shows a use-case of an RDF graph representing part of the experimental data enriched with the publications connected to them, the automatically defined expert profiles and the relationships with external concepts in the LOD cloud. The experimental data is shown in the lower part of Figure 3, where the CLEF campaign is connected to a track (AdHOC2008) composed by a task (AdHoc2008-TELCLEF); furthermore, we report an experiment (AdHoc-TEL-exp-001) submitted to that task. The relationships between a publication (Larson-CLEF2008), the experiment and the author (Ray R. Larson) are enriched by expertise topics (Logic Regression), expert profiles and connections to the LOD cloud.

The connections between experiments and publications enable an explicit binding between the presentation of scientific results and the data actually used to achieve them. Furthermore, publications provide a description of the data, which increases their understandability and the potential for re-usability.

The author is also enriched with information about his or her expertise, and the publication is similarly enriched with information about its topic – logical regression in this case. Identifying, measuring, and representing expertise has the potential to encourage interaction and collaboration by constructing a web of connections between experts. Additionally, this information provides valuable insights to outsiders and novice members of a community.

Finally, the LOD approach allows new access points to the data to be defined; indeed, the expertise topics are connected to external resources belonging to DBPedia, and authors and contributions are connected to the DBLP linked open dataset allowing the experimental data to be easily discovered on the Web.

Future work will focus on the application of these semantic modelling and automatic enrichment techniques to other areas of the evaluation workflow. For example, the expert profiling and the topic extraction could be used to automatically improve and enhance the descriptions of the single experiments submitted to an evaluation campaign.

Links:
CLEF: http://www.clef-initiative.eu/
DIRECT: http://direct.dei.unipd.it/
PROMISE: http://www.promise-noe.eu/
TREC: http://trec.nist.gov/

References:
[1] M. Dussin and N. Ferro: "Managing the Knowledge Creation Process of Large-Scale Evaluation Campaigns", in proc. of the 13th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2009), Springer LNCS 5714, 2009, dx.doi.org/10.1007/978-3-642-04346-8_8
[2] M. Agosti, G. M. Di Nunzio, N. Ferro: "The Importance of Scientific Data Curation for Evaluation Campaigns", in proc. of the First Intl. DELOS Conference, Revised Selected Papers, Springer LNCS 4877, 2007, dx.doi.org/10.1007/978-3-540-77088-6_15
[3] M. Agosti, E. Di Buccio, N Ferro et al.: "DIRECTions: Design and Specifiation of an IR Evaluation Infrastructure", in proc. of CLEF 2012, Springer LNCS 7488, dx.doi.org/10.1007/978-3-642-33247-0_11

Please contact:
Nicola Ferro, University of Padua, Italy
Tel: +39 049 827 7939
E-mail: ferro@dei.unipd.it

Analysing RDF Data: A Realm of New Possibilities

by Alexandra Roatiș

The WaRG framework brings flexibility and semantics to data warehousing. The development of Semantic Web data represented within W3C's Resource Description Framework [1] (RDF), and the associated standardization of the SPARQL query language now at v1.1 has lead to the emergence of many systems storing, querying, and updating RDF. However, as more and more RDF datasets (graphs) are made available, in particular Linked Open Data, application requirements also evolve.

DW4RDF (Data Warehousing for RDF) is a three year R&D project funded by the Digiteo foundation, an important player in the French IT R&D environment in the greater Parisian area. Within this project, we have developed the Warehousing RDF Graphs (WaRG) framework, a joint project involving Dario Colazzo (U. Paris Dauphine, France), François Goasdoué (U. Rennes 1, France) and Ioana Manolescu (Inria & U. Paris Sud, France). Within this framework, we have developed new models and tools for analytics and OLAP-style analysis of RDF data, taking into account data heterogeneity, its lack of a strict structure, and its rich semantics.

Heterogeneity

Data heterogeneity significantly complicates Big Data analytics. For example, although we can reasonably expect to find information about restaurants online, we cannot always find the menu, opening hours or closing days. Existing data warehousing tools tackle such issues by cleaning the data in the Extract Transform Load process, or allow nulls and nested columns in tables. In contrast, we view heterogeneity not as a problem, but as a desired feature. Instead of trying to eliminate or hide it, we propose ways of incorporating heterogeneity within the data warehousing model, and tools to build meaningful aggregates over heterogeneous data.

Warehouses need not revolve around a single concept

Generally, data warehouses follow a star (or snowflake) schema, where facts of a single kind can be analysed based on certain dimensions and measures. To analyze different concepts, such as restaurants, shops, museums, each must be modelled by a different schema and put into a distinct data warehouse. WaRG models a type of data warehouse where several core concepts coexist, interlinked by meaningful queries.

Just a matter of semantics?

RDF Schema is a valuable feature of RDF that allows the descriptions in

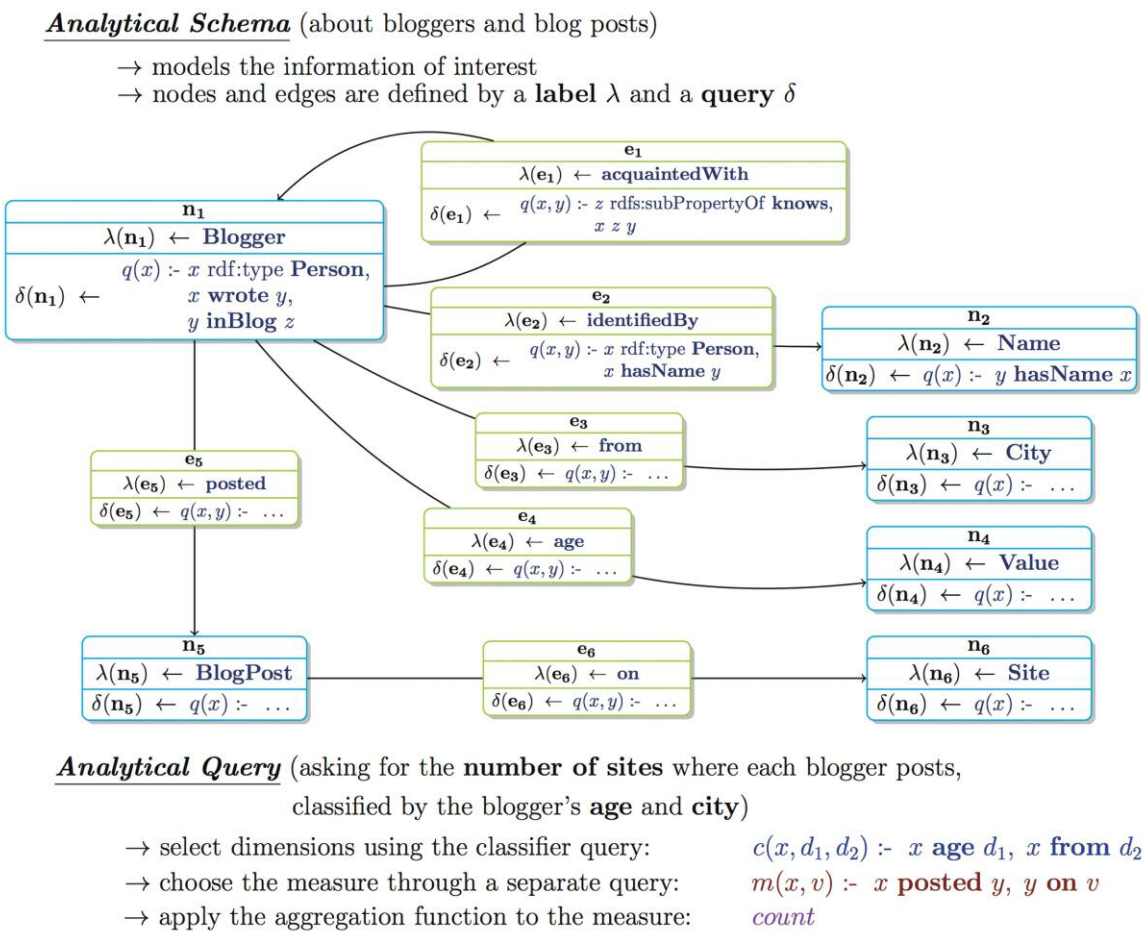


Figure 1: Sample analytical schema and query

RDF graphs to be enhanced by declaring semantic constraints between classes and properties. Such constraints are interpreted under the open-world assumption [2], propagating instances from one relationship to another. For example, in a database where "Océan is a pancake house" and "a pancake house is a restaurant", we can infer that "Océan is a restaurant". Querying the database for restaurants should also return all the pancake houses! Our framework is centreed on RDF, thus it natively supports RDF semantics when querying.

Explore data though analytics

Even for an experienced database developer, understanding a new dataset is always a challenge, as each brings its own set of features, which may be particularly subtle in the case of semantic-endowed data such as RDF. To facilitate their understanding, datasets are generally published along with a schema. But how does one understand the schema? Contemporary published schemas tend to be complex and can be seen as datasets in their own rights. While still small compared to the data, they can be real puzzles for the analyst. Working with RDF, one can seamlessly query the schema and the data, for example ask for all the relationships linking people to other entities. Our model allows analytics not only over the data, but over the schema as well.

Data cubes, no longer a dictatorship

In order to perform data warehouse analysis, one must first establish the dimensions and measures according to which to analyze the facts. Data cubes are built as a result of aggregating the measures along the dimensions. For instance, when asking "what are the total sales for region Lorraine in autumn 2013?", the sales are a measure, while region and period represent dimensions. However, such a warehouse cannot answer the query "how many regions registered sales in autumn 2013?", since region is a dimension, and relational data cubes do not allow aggregating over the dimensions. In contrast, our framework is very flexible, allowing a choice of dimensions and measures at data cube (query) time, not at data warehouse design time.

WaRG models the analytical schema of an RDF warehouse as a graph. Each node represents a set of facts, modelling a new RDF class. The edges connecting these nodes are defined independently and correspond to new RDF properties. The instances of these classes and properties, modelling the data warehouse contents to be further analyzed, are intentionally defined in the schema, following the well-know "Global As View" approach for data integration. For more details we refer the interested reader to [3].

Our ongoing work includes RDF analytical schema recommendation and efficient algorithms for massively parallel RDF analytics.

Link: https://team.inria.fr/oak/warg/

References:

[1] W3C, Resource Description Framework, http://www.w3.org/RDF/
[2] S. Abiteboul, R. Hull, V. Vianu: "Foundations of Databases", Addison-Wesley, 1995
[3] D. Colazzo, F. Goasdoué, I. Manolescu, A. Roatiș: "Warehousing RDF Graphs", in "Bases de Données Avancées", 2013, http://hal.inria.fr/docs/00/86/86/16/PDF/paper.pdf.

Please contact:

Alexandra Roatiș
Inria Saclay and LRI, Université Paris-Sud
http://www.lri.fr/~roatis/
E-mail: alexandra.roatis@lri.fr

The Web Science Observatory - The Challenges of Analytics over Distributed Linked Data Infrastructures

by Wendy Hall, Thanassis Tiropanis, Ramine Tinati, Xin Wang, Markus Luczak-Rösch and Elena Simperl

Linked data technologies provide advantages in terms of interoperability and integration, which, in certain cases, come at the cost of performance. The Web Observatory, a global Web Science research project, is providing a benchmark infrastructure to understand and address the challenges of analytics on distributed Linked Data infrastructures.

The evolution from the Web of documents to the Web of data, social networks, and crowdsourcing has opened up new opportunities for innovation driven by analytics on public datasets, on online social network activity, as well as on corporate or private datasets [1]. These opportunities are evidenced by the emergence of a number of Web Observatories [2] that not only collate and archive, but attempt to provide analytics on such datasets (http://www.nextcenter.org:8080/ugcp/live/observer, http://www.truthy.indiana.edu). These developments have been accompanied by an evolution of data literacy that has been taking place in parallel; people are no longer just consumers of documents on the Web but also contributors of content, data, and applications. The open data movement has shown the potential of crowdsourcing data and applications, while linked data technologies have established their potential for dataset interoperability and integration. The basic associative structure of linked data has advanced the idea of regarding the Web as a global dataspace that, in the future, may be queried as if it were one giant database. Dataspaces are a generic data management abstraction, which helps to derive and maintain relationships between a large number of heterogeneous but interrelated data sources. Relationships are regarded as integration hints until they are reviewed and approved. The overall data management and integration effort is distributed among various

stakeholders and performed in an evolutionary fashion [3].

But is that enough? Analytics on the Web of data and social networks present developers with a dilemma between data warehouse architectures with high performance, big data analytics on the one hand, and analytics on distributed, diverse, separately maintained datasets with potentially lower performance on the other hand. This is where certain challenges for Linked Data emerge; they concern both the representation of diverse datasets as Linked Data and the performance analytics over, potentially distributed, Linked Data infrastructures.

Some of those challenges are explored as part of the Web Observatory (WO) project, which was initiated under the auspices of the Web Science Trust (WST, http://webscience.org). The WO aims to provide a distributed global resource in which datasets, analytic tools and cross-disciplinary methodologies can be shared and combined to foster interdisciplinary research in the context of Web Science [2]. This effort involves many different communities including the Web Science Trust network of laboratories (WSTnet, http://webscience.org/wstnet-laboratories/), other major research groups in this area, government agencies, public sector institutions, and the industry. The WO project has grown as a bottom-up effort, and aims to enable interoperability among datasets and analytics on a large, distributed scale. It involves the crowdsourcing, publication and sharing of both datasets, and of querying, analytic or visualisation tools. At the same time, it involves the development of appropriate standards to enable the discovery, use, combination and persistence of those resources; effort in the direction of standards is already underway in the W3C Web Observatory community group (http://www.w3.org/community/webobservatory/).

The Web Observatory infrastructure that is deployed at the University of Southampton includes different datastore technologies, however, a significant part of it are Linked Data stores. To that end, the infrastructural deployment involves (i) converting large datasets into 5-star linked data formats (http://www.w3.org/DesignIssues/LinkedData.html), (ii) supporting distributed queries over Linked Data stores, and (iii) supporting analytic and visualisation tools over distributed data stores and datasets. Essentially, this infrastructure (shown in Figure 1) will not only provide for valuable analytics over distributed resources but it will also provide for benchmarking analytics on Linked Data.

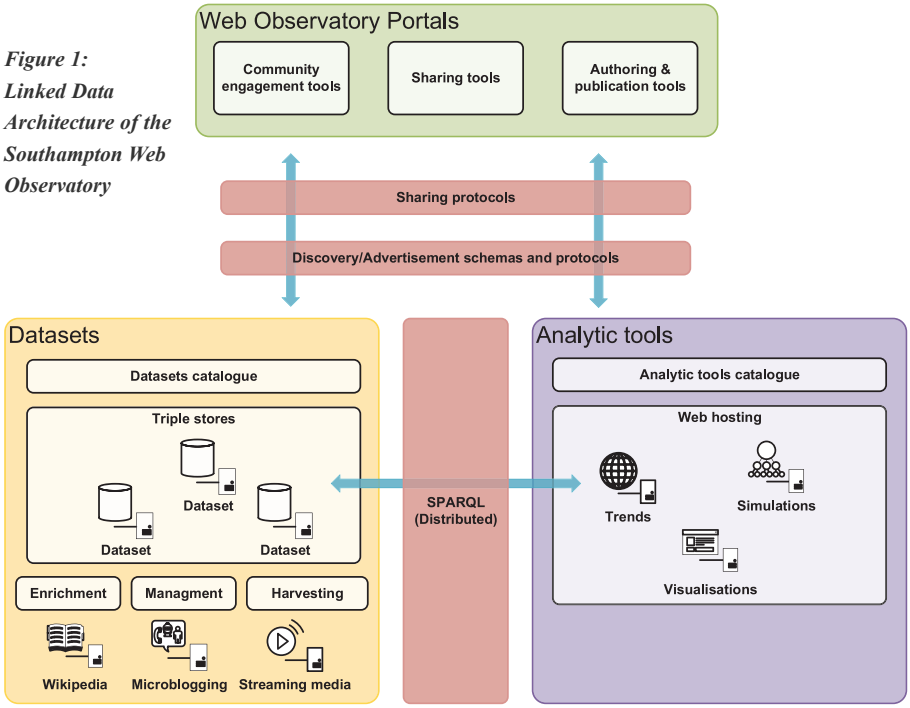


Figure 1: Linked Data Architecture of the Southampton Web Observatory

The datasets that are made available in Linked Data formats for this infrastructure include open data, licensed data and private data that can be accessed only by authorised parties. They include microblogging activity, access to Web 2.0 services (such as Wikipedia), Wellbeing data, and Web of data services (such as USEWOD). Given the volume of some of these datasets (e.g. microblogging data) and the use of diverse and distributed data stores, different approaches to optimisation for the performance of analytics are explored. These involve:

- Representation of datasets in Linked Data formats (e.g. representing microblogging activity, weighted graphs)
- Representation of qualitative data in Linked Data formats
- Enrichment of Linked Data stores for analytics
- Distributed Linked Data query optimisation
- Dataset management and security in Linked Data stores
- Dataset provenance and preservation.

As well as creating a platform for Web Science research, the Web Observatory project will provide insights into performance and optimisation for analytics over distributed Linked Data infrastructures on varying scales of data store size and distribution. It will propose appropriate approaches to data representation, distributed queries and integration in such environments. Contributions to standardization for Web Observatories and Linked Data can also be anticipated. The biggest test for this activity is the extent to which Linked Data infrastructures can be used not only in an efficient manner but also to support researchers across disciplines engaging in interdisciplinary work.

Link:
http://webscience.org/web-observatory/

References:
[1] W. Hall, T. Tiropanis: "Web evolution and Web Science", Computer Networks, 56(18), 3859–3865, 2012, dx.doi.org/10.1016/j.comnet.2012.10.004

[2] T. Tiropanis et al: "The Web Science Observatory", Intelligent Systems, IEEE, 28(2), 100–104, 2013, http://dx.doi.org/10.1109/MIS.2013.50

[3] T. Heath, C. Bizer: "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, http://linkeddatabook.com/editions/1.0/

Please contact:
Thanassis Tiropanis
University of Southampton
E-mail: tt2@ecs.soton.ac.uk

SPARQL: A Gateway to Open Data on the Web?

by Pierre-Yves Vandenbussche, Aidan Hogan, Jürgen Umbrich and Carlos Buil Aranda

Hundreds of datasets on the Web can now be queried through public, freely-available SPARQL services. These datasets contain billions of facts spanning a plethora of diverse topics hosted by a variety of publishers, including some household names, such as the UK and US governments, the BBC and the Nobel Prize Committee. A Web client using SPARQL could, for example, query about the winners of Nobel Prizes from Iceland, or about national electric power consumption per capita in Taiwan, or about homologs found in eukaryotic genomes, or about Pokémon that are particularly susceptible to water attacks. But are these novel SPARQL services ready for use in mainstream Web applications? We investigate further.

With a wealth of Linked Data now available on the Web, and more on the way, the robustness of SPARQL technology (a W3C Standard [1]) to access and query these data is of key importance. While SPARQL has undeniable advantages when it comes to query expressivity and interoperability, publishing data using this technology comes at a price. SPARQL services are usually offered free-of-charge to arbitrary clients over the Web, and quality-of-service often suffers. Endpoints may go offline or only return partial results or take longer to return answers than a user is willing to wait. As a result, these endpoints may not be usable for mainstream applications.

The SPARQLES (SPARQL Endpoint Status) project aims to clarify the current state of public SPARQL endpoints deployed on the Web by monitoring their health and performance. The project is an ongoing collaboration between Fujitsu Labs; DCC, Universidad de Chile; PUC, Chile (Grant NC120004); and INSIGHT@NUI Galway. Hosting of the project is provided by the not-for-profit Open Knowledge Foundation (OKFN).

The SPARQLES project currently monitors 442 public SPARQL endpoints registered in Datahub (a community-based data catalogue). The results from the monitoring system are continuously updated on a public website (see links) that provides information along the following dimensions:

- *Discoverability – how can an agent discover a SPARQL endpoint and what data/metadata is stored?* Among the two methods available to describe an endpoint, SPARQL 1.1 Service Descriptions are used in only 10% of the endpoints and VoID descriptions are used in 30% of the endpoints [2]. SPARQLES indicates if these descriptions are provided for monitored endpoints, offering direct access where available.
- *Interoperability - which SPARQL functionalities are supported?* As for any database, the implementation of SPARQL standards (versions 1.0 and 1.1) can vary from one vendor/tool to another. The SPARQLES system assesses the compliance of each endpoint with respect to the SPARQL standard and presents any exceptions that occur to the user. We generally find good compliance with the SPARQL 1.0 standard but support for recently-standardised SPARQL 1.1 features is still sparse [2].
- *Performance – what general query performance can be expected?* Is the endpoint's performance good enough for a particular application? SPARQLES runs timed experiments over the Web against each endpoint, testing the speed of various operations, such as simple lookups, streaming results and performing joins. A detailed breakdown of the performance of the endpoint is then published on the SPARQLES website. Across all endpoints, the median time for answering a simple lookup is 0.25 seconds, for streaming 100,000 results is 72 seconds, and for running a join with 1,000 intermediate results is 1 second [2]. However, the performance of individual endpoints can vary by up to three orders of magnitude for comparable tasks.
- *Availability – what is the average uptime based on hourly pings?* Which SPARQL endpoints can we trust to be online when we need to query them? For the past three years, SPARQLES has been issuing hourly queries to each public SPARQL endpoint to test if they are online. From these data, the system computes the availability of an endpoint for a given period as the ratio of the total requests that succeed vs. the total number of requests made. Looking at monthly availability, we found that 14.4% of endpoints are available 95% to 99% of the time, 32.2% of endpoints are available 99% to 100% of the time, while the remainder are available less than 95% of the time [2]. The SPARQLES website shows each endpoint's availability during the last 24 hours and during the last seven days, so application developers can make a more informed decision about whether or not they can rely on an endpoint.

As a whole, SPARQLES contributes to the adoption of SPARQL technology by being seminal in providing the community a complete view on the health of available endpoints [3]. Furthermore, for the first time, this project provides a tool to monitor the service provided by data publishers, creating an incentive for publishers to maintain a high quality service. Future work will include the packaging of the tool (already openly available in github) in a standalone version, which will make it easy for anyone to monitor their endpoint locally. This next step will include an alerts feature in case errors occur.

Links:

http://sparqles.okfn.org/
https://github.com/pyvandenbussche/sparqles
http://datahub.io/
http://okfn.org/

References:

[1] E. Prud'hommeaux, A. Seaborne: "SPARQL query language for RDF", W3C Recommendation, 2008, http://www.w3.org/TR/rdf-sparql-query

[2] C. Buil-Aranda et al.: "SPARQL Web-Querying Infrastructure: Ready for Action?", in The Semantic Web–ISWC, 2013, http://vmwebsrv01.deri.ie/sites/default/files/publications/paperiswc.pdf

Please contact:

Pierre-Yves Vandenbussche,
Fujitsu (Ireland) Limited
pierre-yves.vandenbussche@ie.fujitsu.com

CODE Query Wizard and Vis Wizard: Supporting Exploration and Analysis of Linked Data

by Patrick Hoefler and Belgin Mutlu

Although the concept of Linked Data has been increasing in popularity, easy-to-use interfaces to access and make sense of the actual data are still few and far between. The CODE project's Query Wizard and Vis Wizard aim to fill this gap.

The amount of Linked Data available on the Web is growing continually, due largely to an influx of new data from research and open government activities. However, it is still quite difficult to directly access this wealth of semantically enriched data without having in-depth knowledge of semantic technologies.

Therefore, one of the goals of the EU-funded CODE project has been to develop a web-based visual analytics platform that enables non-expert users to easily perform exploration and analysis tasks on Linked Data. CODE's vision is to establish a toolchain for the extraction of knowledge encapsulated in scientific research papers along with its release as Linked Data [1]. A web-based visual analytics interface should empower the end user to analyse, integrate, and organize the data. The CODE Query Wizard and the CODE Vis Wizard fulfill this role.

When it comes to working with data, many people know how to use spreadsheet applications, such as Microsoft Excel. In comparison, very few people know SPARQL, the W3C standard language to query Linked Data. The CODE Query Wizard [2] provides a web-based interface that dramatically simplifies the process of displaying, accessing, filtering, exploring, and navigating the Linked Data that's available through a SPARQL endpoint. The main innovation of the interface is that it turns the graph structure of Linked Data into tabular form and provides easy-to-use interaction possibilities by using metaphors and techniques that the end user is already familiar with.

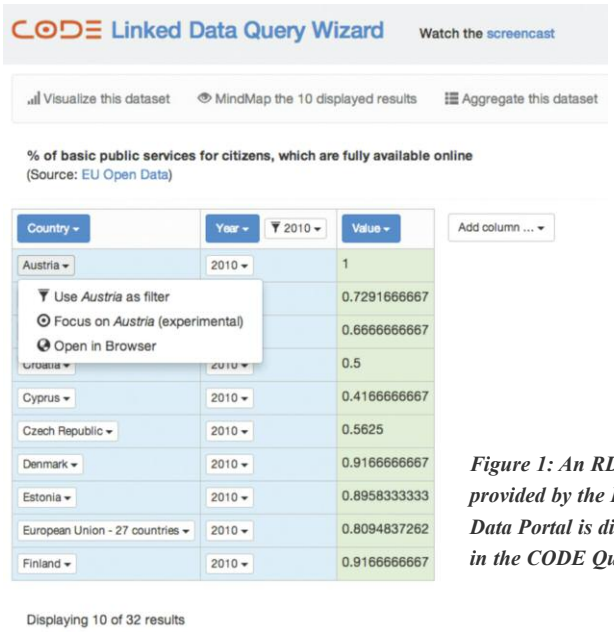


Figure 1: An RDF Data Cube provided by the European Open Data Portal is displayed and filtered in the CODE Query Wizard.

The CODE Query Wizard offers two entry points: A user can either initiate a keyword search over a Linked Data repository, or select any of the already available datasets, represented as RDF Data Cubes. In both cases, the CODE Query Wizard presents a table containing the results. The user can then select columns of interest and set filters to narrow down the displayed data. Additionally, the user can explore the

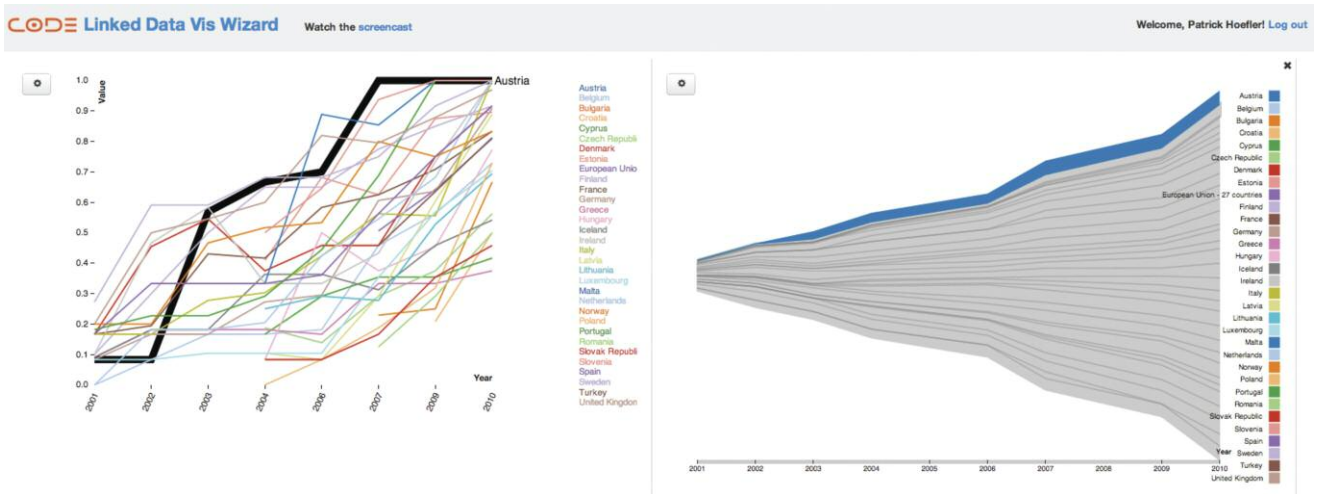


Figure 2: The CODE Vis Wizard displays an interactive visual representation of the percentage of public services available online. Austria is selected in the left chart by the user and automatically highlighted in the right chart by the system.

data by "focusing" on an entity, or can aggregate a dataset to obtain a summary of the data.

Once a user is happy with the selected data, it can be visualized using the CODE Vis Wizard [3]. This tool enables visual analysis of Linked Data, and supports the user by automating the visualization process. This means that after analyzing the structural and semantic characteristics of the provided Linked Data, the CODE Vis Wizard automatically suggests any of the 10 currently available visualizations that are suitable for the provided data. Furthermore, the Vis Wizard automatically maps the data on the available visual channels of the chosen visualization. If the user wishes to adjust the mapping, this can be achieved with a few simple clicks.

Usually more than one visualization is suitable for any given dataset. In this case, all visualizations can be displayed side by side. When certain parts of the data are selected in one of the visualizations, they are automatically highlighted in the others as well. This can provide quick insights into complicated data, taking advantage of the powerful human visual perception system.

The CODE Query Wizard and Vis Wizard are purely web-based systems. They currently support Virtuoso, OWLIM and Bigdata SPARQL endpoints, since these also provide integrated full-text search. However, since the prototypes have been designed to use Semantic Web standards, such as SPARQL, wherever possible, support for other suitable endpoints could be added at a later point with minimal effort.

Both prototypes have been developed within the CODE project at the Know-Center in Graz, Austria, with support by their project partners University of Passau, Mendeley (London) and MeisterLabs (Vienna). The project started in May 2012 and will finish in April 2014.

Links:
CODE project: http://code-research.eu/
CODE Query Wizard & Vis Wizard: http://code.know-center.tugraz.at/

References:
[1] C. Seifert et al.: "Crowdsourcing Fact Extraction from Scientific Literature", in A. Holzinger & G. Pasi (Eds.), Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured Big Data, Springer, 2013. dx.doi.org/10.1007/978-3-642-39146-0_15
[2] P. Hoefler et al.: "Linked Data Query Wizard: A Tabular Interface for the Semantic Web", in P. Cimiano (Ed.), The Semantic Web: ESWC 2013 Satellite Events, Springer, 2013, dx.doi.org/10.1007/978-3-642-41242-4_19
[3] B. Mutlu et al.: "Suggesting Visualisations for Published Data", in proc. of IVAPP 2014, SCITEPRESS, 2014.

Please contact:
Patrick Hoefler
Know-Center GmbH, Austria
E-mail: phoefler@know-center.at

AV-Portal - The German National Library of Science and Technology's Semantic Video Portal

by Harald Sack and Margret Plank

In addition to specialized literature, 3D objects and research data, the German National Library of Science and Technology also collects scientific video clips, which will now be opened up via semantic media analysis, and published on the web with the help of Linked Open Data.

The German National Library of Science and Technology (TIB) is one of the largest specialized libraries worldwide. The TIB's task is to comprehensively acquire and archive literature from around the world pertaining to all areas of engineering, as well as architecture, chemistry, information technology, mathematics and physics. The TIB´s information portal GetInfo provides access to more than 150 million datasets from specialized databases, publishers and library catalogues. The Competence Centre for non-textual Materials at the TIB aims to improve ease of access and use of non-textual material, such as audiovisual media, 3D objects and research data. To address these challenges, tools and infrastructure are being developed that actively support users in scientific work processes, enabling the easy publication, finding and long-term availability of non-textual objects.

TIB collects digital audiovisual media (AV-media) in the form of computer visualizations, explanatory images, simulations, experiments, interviews and recordings of lectures and conferences. TIB also holds a historical film collection of almost 11,500 research films, university teaching films and documentaries, some of which date back to the 1910s. Given the rapidly increasing volume of AV-media and the need to index even individual film sequences and media fractions, an intellectual, "manual" indexing is too expensive: an efficient automated content-based metadata extraction is required. In cooperation with the Hasso-Plattner-Institut for IT Systems Engineering (HPI) a web-based platform for TIB's audiovisual media, the AV-Portal has been developed, combining state-of-the art multimedia analysis with Linked Open Data based semantic analysis and retrieval.

The processing workflow of the AV-Portal starts with the media ingest, where additional authoritative (textual) metadata can be provided by the user. After structural analysis based on shot detection, representative keyframes for a visual index are extracted, followed by optical character recognition, automated speech-to-text transcription, and visual concept detection. The visual concept detection classifies video image content according to predefined context related visual categories, for example, "landscape", "drawing", or "animation" [1]. All metadata are further processed via linguistic and subse-

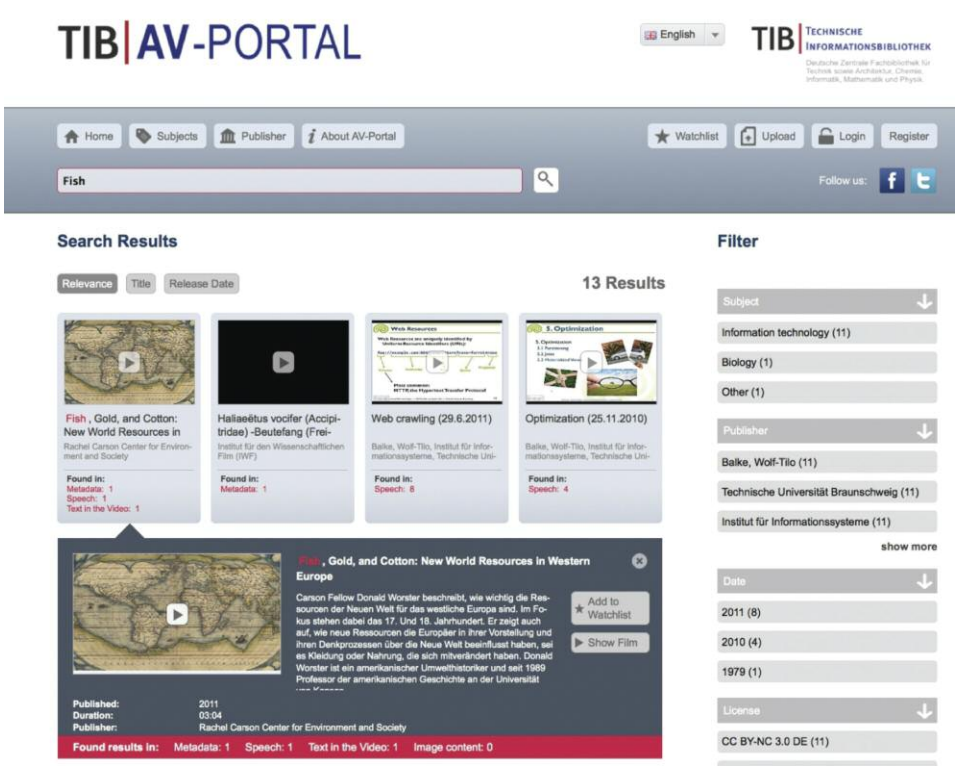


Figure 1: The AV-Portal of the German National Library of Science and Technology

quent semantic analysis, ie, named entities are identified, disambiguated and mapped to an authoritative knowledge base. The underlying knowledge base consists of parts of the Integrated German authority files relevant for TIB's subject areas, which is available as Linked Data Services of the German National Library.

Because of the heterogeneous origin of the available metadata, ranging from authoritative metadata provided by reliable experts to unreliable and faulty metadata from automated analysis, a context-aware approach for named entity mapping has been developed that considers data provenance, source reliability, source ambiguity, as well as dynamic context boundaries [2]. For disambiguation of named entities, the Integrated German Authority files do not provide sufficient information, ie, they only provide taxonomic structures with hypernyms, hyponyms, and synonyms, while cross-references are completely missing. Thus, named entities must also be aligned to Dbpedia entities as well as to Wikipedia articles. DBpedia graph structure, in combination with Wikipedia textual resources and link graphs, enables reliable disambiguation based on property graph analysis, link graph analysis, as well as text-based coreference and cooccurrence analysis.

Semantic video search at the TIB AV-Portal is supported by content-based filter facets for search results that enable the exploration of the ever-increasing number of video assets in order to make searching for AV-media as easy as it already is for textual information (Figure 1). In 2011, a low fidelity prototype of the AV-Portal was developed, in 2012-2013 the beta operation of the system followed and, for 2014, the full operation of the portal is scheduled.

In addition to improving the quality of automated analysis, multilinguality is a main focus of future work. Scientific AV-media collected by TIB often comprise content and metadata in different languages. Whilst taking into account the preferred language of the user, text in different languages has to be aligned with a language dependent knowledge base (Integrated German Authority files) and potentially displayed in a second language in the GUI of the AV-Portal. Therefore, the integration and mapping of international authority files, such as the Library of Congress Subject Headings or the Virtual International Authority File is the subject of future and ongoing work.

Links:
TIB: http://www.tib-hannover.de/en/
GetInfo:
http://www.tib-hannover.de/en/getinfo/
Competence Centre for non-textual Materials: http://www.tib-hannover.de/en/ services/competence-centre-for-non-textual-materials/
Linked Data Services of the German National Library:
http://www.dnb.de/EN/lds
Dbpedia: http://dbpedia.org/

References:
[1] Ch. Hentschel, I. Blümel, H. Sack: "Automatic Annotation of Scientific Video Material based on Visual Concept Detection", in proc. of i-KNOW 2013, ACM, 2013, article 16, dx.doi.org/10.1145/2494188.2494213

[2] N. Steinmetz, H. Sack: "Semantic Multimedia Information Retrieval Based on Contextual Descriptions", in proc. of ESWC 2013, Semantics and Big Data, Springer LNCS 7882, 2013, pp. 283-396, dx.doi.org/10.1007/978-3-642-38288-8_26

Please contact:
Harald Sack
Hasso-Plattner-Institute for IT Systems Engineering, University of Potsdam, Germany
E-mail: harald.sack@hpi.uni-potsdam.de

Margret Plank
German National Library of Science and Technology (TIB)
E-mail:
Margret.Plank@tib.uni-hannover.de

Browsing and Traversing Linked Data with LODmilla

by András Micsik, Sándor Turbucz and Zoltán Tóth

There are a range of problems associated with current Linked Data visualization tools, including lack of genericity and reliance on non-standard dataset endpoint features. These problems hinder the emergence of generic Linked Data browsers and can thus complicate the process of accessing Linked Data. With LODmilla we aim to overcome common problems of Linked Open data (LOD) browsing and to establish an extensible base platform for further evolution of Linked Data browsers.

The Linked Open Data (LOD) concept is based on the Linked Data principles set by Tim Berners-Lee in 2006 and the open data movement (formalized by the Open Definition). Semantic data is not only machine processable but it is often the primary source for information sought by humans. Recent years have seen the development of many useful semantic data visualizations prepared for insular purposes. There are also several generic approaches to present LOD for humans, such as Graphity, RelFinder and LodLive, which offer quite different solutions in terms of data presentation and navigation capabilities.

The situation can be compared to the WWW paradigm; roughly 20 years after the birth of WWW we have a broad consensus on how a web browser should work and what its common functions are. Recent developments show that a similar convergence has started in the realm of LOD browsing. The LODmilla browser is our attempt to advance this convergence.

LOD has a dual nature; it can be seen as a graph as well as a set of data tables or records, consequently a visualization is needed that combines these two features. In LODmilla, therefore, the entities are shown as graph nodes with predicates drawn as named edges among nodes (Figure 1). On the other hand, both datatype and object properties of the entities can also be looked over in a text list. Figure 1 shows the main parts of the LODmilla screen: the graph view takes the major part of the screen, while the information panel shows details about the selected node on the right side. The palettes on the right contain a customizable and extensible set of tools for navigation and graph expansion. The menu of common actions such as "undo" and "clear" is positioned at the bottom.

LODmilla provides some functionality in each of the following categories, which we identified as desirable functionality groups for LOD browsing:

- Graph management: e.g. open/close/select nodes
- Graph view manipulation: zoom, pan, etc.
- Viewing properties: list/filter properties, open linked nodes, etc.
- Personal use: settings, undo, load/save graph views
- Collaboration: share graph view with others, comment
- Exploration: e.g. search the graph neighbourhood according to various criteria.

Some of the more interesting features of these functionality groups are highlighted below.

Specific screens can be saved by users and shared with other users via secret URLs. The recipient's screen will show exactly the same graph as the one saved by the sender.

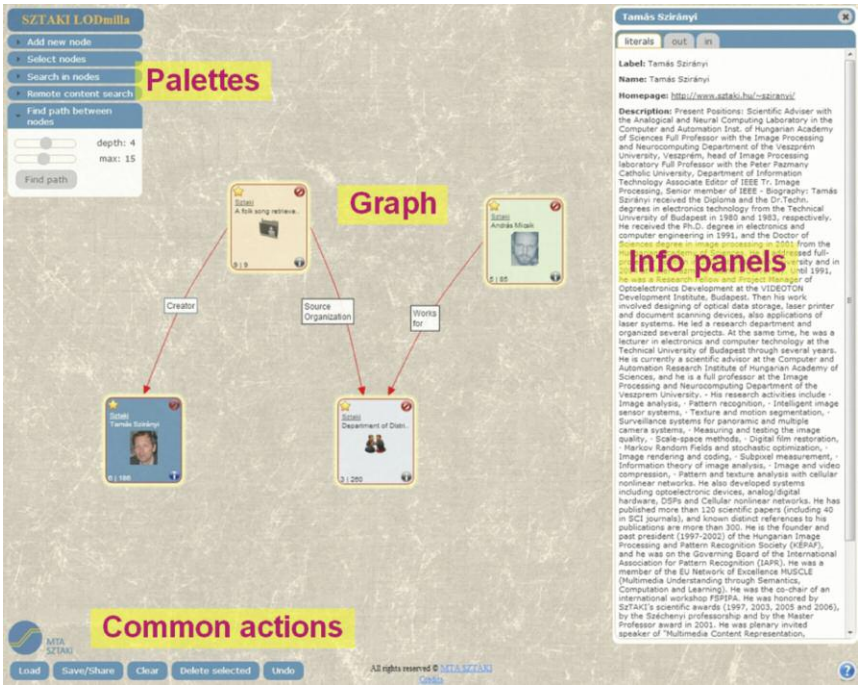


Figure 1: The screen elements of LODmilla

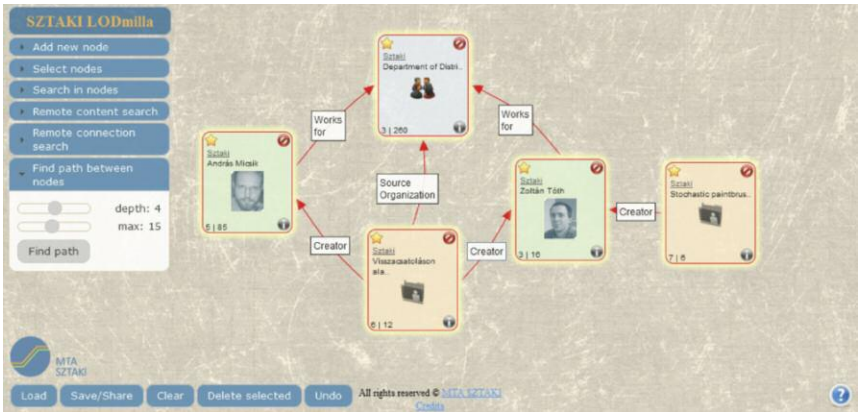


Figure 2: Finding a path with LODmilla

The core difficulty of browsing the LOD graph is the high number of nodes and edges. Providing access to all information but presenting only the selected pieces are key principles in this case. The performance of a graph action may squeeze hundreds of new nodes or connections on the screen. In this case the user can push the "undo" button to end confusion and revert to the previous view. It is also common for nodes to have hundreds of data properties, which is unmanageable in a traditional list. We provide grouping of properties by type and progressive search in the contents of data properties, in a similar manner to "Ctrl-F" in web browsers.

The most exciting aspect is the ability to expand the graph in intelligently selective ways: LODmilla has three experimental expansion functions implemented, and many more can be plugged in. First, a given type of link can be expanded starting from selected nodes. This will show, for example, the creator network surrounding a paper. Second, a text search can be run in neighbouring nodes, and nodes containing the text pattern will appear on the screen. The third option is path finding between nodes: the search starts from all selected nodes and runs until they link up or the maximum search depth is reached (see Figure 2 for an example, where new connections are drawn with red).

Finally, LODmilla aims to be usable with the broadest range of LOD datasets. For this, we avoid the need for per dataset configuration in the browser, and provide a browsing solution that works both via dereferencable URIs and SPARQL endpoints.

The LODmilla browser is open both to try out and to extend via the links below. Thus we hope to progress the community towards common understanding and implementation of intelligent human browsing for Linked Data.

Links:
LODmilla web application: http://munkapad.sztaki.hu/lodmilla
Source code: https://github.com/dsd-sztaki-hu/LODmilla-frontend

References:
[1] A. Micsik et al.: "LODmilla: shared visualization of Linked Open Data" in proc of First Workshop on Linking and Contextualizing Publications and Datasets, La Valletta, Malta, 2013
[2] A-S. Dadzie, M. Rowe: "Approaches to visualising linked data: a survey. Semantic web 2, 2 (April 2011), 89-124, dx.doi.org/10.3233/SW-2011-0037

Please contact:
András Micsik
SZTAKI, Hungary
Tel: +36 1 279 6248
E-mail: andras.micsik@sztaki.mta.hu

Diachronic Linked Data: Capturing the Evolution of Structured Interrelated Information on the Web

by George Papastefanatos and Yannis Stavrakas

The recent development of Linked Open Data technologies has enabled large scale exploitation of previously isolated, public, scientific or enterprise data silos. Given its wide availability and value, a fundamental issue arises regarding the long-term accessibility of these knowledge bases; how do we record their evolution and how do we preserve them for future use? Until now, traditional preservation techniques keep information in fixed data sets, "pickled" and "locked away" for future use. Given the complexity, the interlinking and the dynamic nature of current data, especially Linked Open Data, radically new methods are needed.

In this respect, several challenges arise when preserving Linked Open Data:

- How can we monitor changes in third-party LOD datasets released in the past (the evolution tracking problem), and how can ongoing data analysis processes consider newly released versions (the change synchronization problem)?
- How can we understand the evolution of LOD datasets with respect to the real world entities they describe (the provenance problem), and how can we repair various data imperfections, e.g., granularity inconsistencies (the curation problem)?
- How can we assess the quality of harvested LOD datasets in order to decide which and how many versions of them deserve to be further preserved (the appraisal problem)?
- How can we cite a particular revision of a LOD dataset (the citation problem), and how will we be able to retrieve them when looking up a reference in the form in which we saw it – not the most recently available version (the archiving problem)?
- How can we distribute preservation costs to ensure long-term access even when the initial motivation for publishing has changed (the sustainability problem)?

The DIACHRON project aims at tackling these problems with an innovative approach that tries to integrate the preservation processes in the traditional lifecycle of production-processing-consumption of LOD data. LOD should be preserved by keeping them constantly accessible and integrated into a larger framework of open evolving data on the Web. This approach calls for effective and efficient techniques to manage the full lifecycle of LOD. It requires enriching LOD with temporal and provenance annotations, which are produced while tracking LOD re-use in complex value making chains. According to this vision both the data and metadata become diachronic, and the need for third-party preservation (e.g., by memory institutions) is greatly reduced. We expect that this paradigm will contribute towards a self-preserving Data Web or Data Intranets.

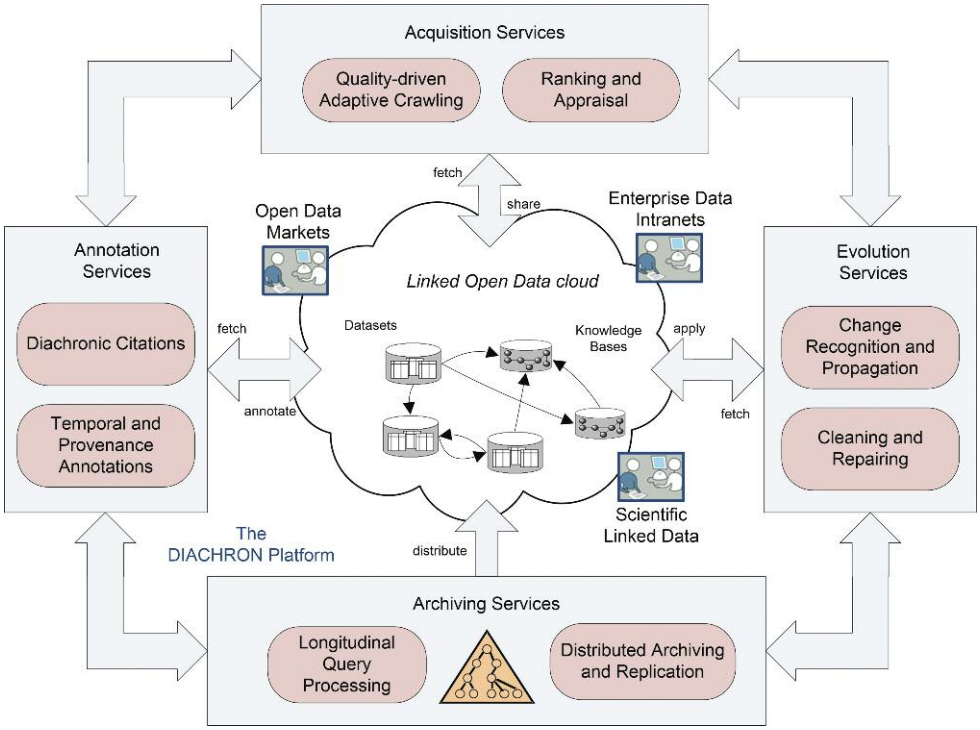


Figure 1: DIACHRON Platform Architecture

To this end, DIACHRON's main artefact will be a platform for diachronic linked data. The platform is not intended to replace existing standards and tools, but rather to complement, integrate, and co-exist with them, as shown in Figure 1. Notably, we foresee four groups of services for long-term LOD accessibility and usability: acquisition, annotation, evolution, and archiving services.

The acquisition module is responsible for harvesting LOD datasets published on the Data Web and assessing their quality with regard to critical dimensions such as accuracy, completeness, temporal consistency or coverage. It includes services for:

- Ranking LOD datasets according to various quality dimensions.
- Crawling datasets on the Web or Intranets based on their quality criteria.

The annotation module is responsible for enriching LOD with superimposed information regarding temporal validity and provenance of the acquired datasets. It consists of services for:

- Diachronic citations based on persistent URIs of LOD datasets, ie, references to data and their metadata that do not "break" in case those data are modified or removed over time.
- Temporal and provenance annotations. Given that LOD datasets change without any notification while they get freely replicated on the Data Web, understanding where a piece of data (or metadata) came from and why and how it has obtained its current form is also crucial for appraisal.

The evolution module is responsible for detecting, managing and propagating changes in LOD datasets monitored on the Data Web. It provides services for:

- Cleaning and repairing LOD datasets. DIACHRON intends to deal with LOD inconsistencies arising due to evolving information (e.g., changes in scientific knowledge), revisions to their intended usage or simply errors posed by data replication between repositories.
- Change recognition and propagation by monitoring and comparing snapshots of LOD datasets. DIACHRON will pay particular attention to the LOD change language used to produce deltas that can be interpreted both by humans and machines.

The archiving module is responsible for storing and accessing multiple versions of annotated LOD datasets as presented in the previous modules and services. It comprises services for:

- Multi-version Archiving of LOD datasets that is amenable to compression of inherently redundant information, as well as to querying of the evolution history of LOD. The archived data will be replicated in several nodes in order to enable community-based preservation of LODs.
- Longitudinal querying featuring complex conditions on the recorded provenance and change information of archived LOD datasets.

To this end, the results of DIACHRON will be evaluated in three large-scale use cases focusing on open governmental data, enterprise data and scientific data ecosystems. The DIACHRON Project is an FP7 – IP project that started in April 2013 and will run for 36 months. The consortium comprises academic institutions (Institute for the Management of Information Systems/"ATHENA" Research Center, Greece; FORTH, Greece; University of Bonn, Germany, and University of Edinburgh, UK), companies (INTRASOFT, Belgium; DATA PUBLICA, France; DATA MARKET, Iceland; HANZO ARCHIVES, UK; BROX IT SOLUTIONS, Germany) as well as user communities (European Bioinformatics Institute, EMBL ,UK).

Link: http://www.diachron-fp7.eu

Reference:

[1] S. Auer et al.: "Diachronic linked data: towards long-term preservation of structured interrelated information", in proc. of WOD '12, Nantes, France, 2012, dx.doi.org/10.1145/2422604.2422610

Please contact:

George Papastefanatos,
Athena Research Centre, Greece
E-mail:
gpapas@imis.athena-innovation.gr

Yannis Stavrakas
Athena Research Centre, Greece
E-mail:
yannis@imis.athena-innovation.gr

Supporting the Data Lifecycle at a Global Publisher using the Linked Data Stack

by Christian Dirschl, Katja Eck and Jens Lehmann

The Linked Data Stack is an integrated distribution of aligned tools that support the whole lifecycle of Linked Data from extraction, authoring/creation via enrichment, interlinking and fusing through to maintenance. A global publishing company represents an ideal recent real-world usage scenario, illustrating the Linked Data Stack and the underlying lifecycle of Linked Data (including data-flows and usage scenarios).

In these times of omnipresent electronic devices, the ways of consuming information are changing and so too are the expectations of customers. Documentation and processing of publishers' data, however, are lagging behind.

Let's assume an accounting professional is working for a leading consultancy and is responsible for certifying tax returns for an international customer. In the future, the publisher aims to deliver more personalized and context specific information precisely fulfilling his need to track changes in information provided by a multitude of sources.

The Linked Data Stack provides specialized tools for each Linked Data lifecycle stage (e.g. data enrichment, management of knowledge bases, reasoning techniques and semantic search support) and can consequently be used to facilitate the semantic content processing workflows.

The Linked Data Stack and Life-Cycle

The description of the Linked Data stack and the Linked Data lifecycle are based on earlier work in [1] and [2]. The Linked Data Stack is an integrated distribution of aligned tools that support the whole lifecycle of Linked Data from extraction, authoring/creation via enrichment, interlinking and fusing through to maintenance. The major components of the Linked Data Stack are open-source in order to facilitate wide deployment. The stack is designed to be versatile; for all functionalities there are clear interfaces, which enable the plugging in of alternative third-party implementations.

In order to fulfill these requirements, the architecture of the Linked Data Stack is based on the following basic principles:

- Software integration and deployment using the Debian packaging system: The Debian packaging system is one of the most widely used packaging and deployment infrastructures, and facilitates packaging and integration as well as maintenance of dependencies between the various Linked Data Stack components. Using the Debian system also facilitates the deployment of the Linked Data Stack on individual servers, cloud or virtualization infrastructures.
- Use of a central SPARQL endpoint and standardized vocabularies for knowledge base access and integration between different tools: All components of the Linked Data Stack access this central knowledge base repository and write their findings back to it. In order for other tools to make sense out of the output of a certain component, it is important to define vocabularies for each stage of the Linked Data lifecycle.

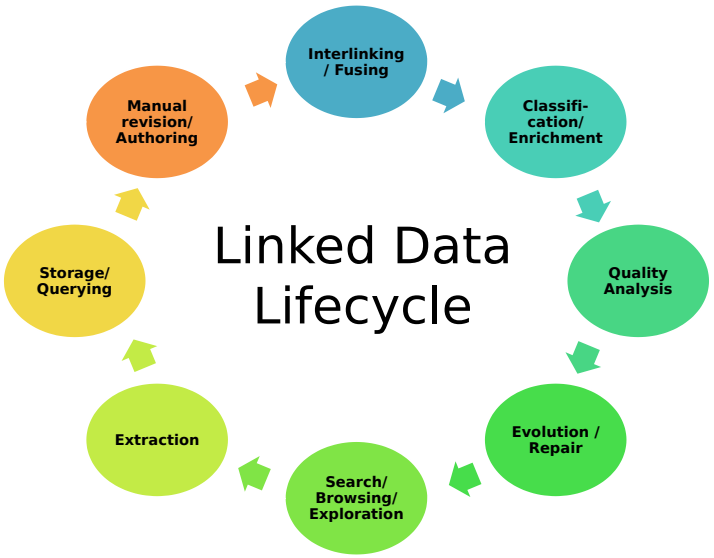


Figure 1: Overview of the stages of the Linked Data lifecycle [3].

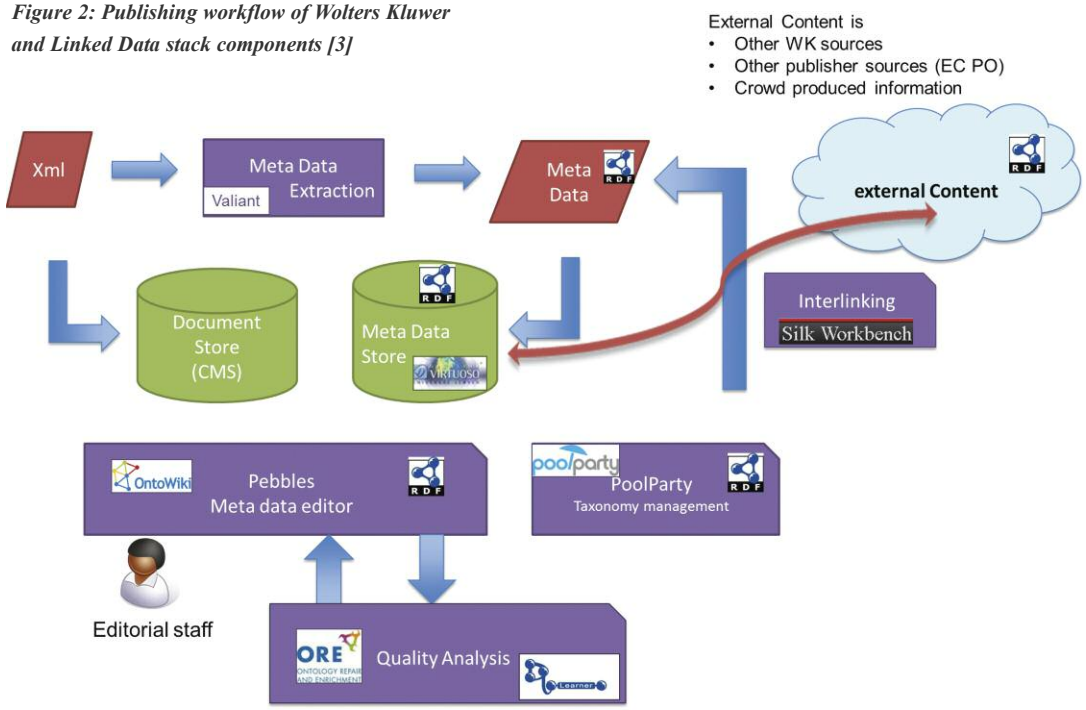
Usage of the Linked Data Stack at WKD

Wolters Kluwer Germany (WKD) is an information service provider in the legal, business and tax domain. WKD is part of global Wolters Kluwer n.v. In 2012, the company had an annual revenue of 3.6 billion Euro,19,000 employees worldwide and customers in over 150 countries across Europe, North America, Asia Pacific and Latin America.

The paradigm of Linked Data and its lifecycle is highly compatible with the existing workflows at Wolters Kluwer as an information provider; consequently the Linked Data stack can offer functionality and technology that is relevant and complementary to the existing content management and production environment.

The main aim of implementing tools from the Linked Data stack into WKD's operational systems was to make the internal content processes more flexible and efficient, but feature requirements of the company's electronic and software products also had to be taken into consideration. Once the technological basis was laid, opportunities for further

Figure 2: Publishing workflow of Wolters Kluwer and Linked Data stack components [3]



enhancements were immediately revealed; thus the Linked Data stack proved its value from early on, and there is no doubt that its importance will only continue to grow. The tools currently used from the Linked Data stack are well integrated with one other, which enables an efficient workflow and processing of information. URIs in PoolParty based on controlled vocabulary are used by Valiant for the content transformation process, and stored in Virtuoso, for easy querying via SPARQL and display in OntoWiki.

Using a Linked Data stack has the major advantage that installation is easy and issues associated with different versions not working smoothly together are avoided. These represent major advantages compared to the separate implementation of individual tools. Figure 2 shows the interplay of partially operational Linked Data stack components in the processes of WKD.

The major challenge, however, is not the new technology per se, but a smooth integration of this new paradigm into WKD's existing infrastructure and a stepwise replacement of old processes with the new and enhanced ones.

The lack of public machine-readable legal resources in many European countries led to the decision to publish legal resources ourselves in order to initiate discussions within the publishing industry, but also within the Linked Data community and public bodies. These resources are available via the SEMIC Semantic Interoperability Community platform as semantic assets, but also directly at vocabulary.wolterskluwer.de.

Future Work

In the future, we will concentrate on adding further tools from the tool stack into our internal content processing engine as well as adding additional external sources to our knowledge base. These steps will be accompanied by detailed user and usability tests as well as documentation of business impact.

Acknowledgement: The research leading to these results has received funding under the European Commission's Seventh Framework Programme from ICT grant agreements LOD2 (no. 257943) and GeoKnow (no. 318159).

Links:

Linked Data Stack Website: http://stack.linkeddata.org
Wolters Kluwer Germany: http://www.wolterskluwer.de
LOD2 project: http://lod2.eu
GeoKnow project: http://geoknow.eu
Thesauri: http://vocabulary.wolterskluwer.de/court.html
http://vocabulary.wolterskluwer.de/arbeitsrecht.html

References:

[1] S. Auer et al.: "Managing the life-cycle of Linked Data with the LOD2 Stack" in proc. of the 11th International Semantic Web Conference (ISWC), Springer, 2012, dx.doi.org/10.1007/978-3-642-35173-0_1
[2] S. Auer, J. Lehmann: "Making the Web a Data Washing Machine - Creating Knowledge out of Interlinked Data", Semantic Web Journal, 1,1-2, pp 97-104, IOS Press, 2010, http://www.semantic-web-journal.net/sites/default/files/swj24_0.pdf
[3] C. Dirschl et al.: "Facilitating Data-Flows at a Global Publisher using the LOD2 Stack"; submitted to the Semantic Web journal, http://www.semantic-web-journal.net/content/facilitating-data-flows-global-publisher-using-lod2-stack

Please contact:

Christian Dirschl, Katja Eck
Wolters Kluwer Deutschland GmbH, Germany
E-mail: CDirschl@wolterskluwer.de, KEck@wolterskluwer.de

Jens Lehmann
University of Leipzig, Germany
E-mail: lehmann@informatik.uni-leipzig.de

A SOLID Architecture to Weather the Storm of Real-Time Linked Data

by Miguel A. Martínez-Prieto, Carlos E. Cuesta, Javier D. Fernández and Mario Arias

Linked Open Data has increased the availability of semantic data, including huge flows of real-time information from many sources. Processing systems must be able to cope with such incoming data, while simultaneously providing efficient access to a live data store including both this growing information and pre-existing data. The SOLID architecture has been designed to handle such workflows, managing big semantic data in real-time.

The successful Linked Data initiative has driven the publication of big, heterogeneous semantic datasets. Data from the governments, social networks and relating to bioinformatics, are publicly exposed and interlinked within the Web of Data. This can be seen as part of the more general emerging trend of Big Data, where the actual value of our data depends on the knowledge which can be inferred from it, in order to support the decision making process. The term "Big Semantic Data" refers to those RDF datasets for which volume, velocity, and variety demand more computation resources than are provided by traditional management systems. Whilst this can usually mean terabytes or petabytes, a few gigabytes may be enough to collapse an application running on limited devices. Both high-performance and small devices must be considered at the confluence of the Web of Data and the Internet of Things (IoT).

The IoT hosts diverse potential real-time sources of RDF, such as RFID labels, Web processes, smartphones and sensors. All of these can be exploited as a whole in the emergent "smart-cities". With the cohabitation of diverse devices, this scenario results in a variety of data flows; a smart-city comprises sources about the daily life of the city (weather sensors, distributions of people at points of interest, traffic congestion, public transport), but it generally requires a less-dynamic background knowledge, such as road maps, infrastructure and organization relationships. The most interesting applications are those integrating different sources to provide services, such as predictions of traffic patterns depending on the weather or certain specific days, or other decision support for city management.

Traditionally, platforms managing Big Semantic Data and those dealing with real-time information follow completely different design principles. Big Semantic Data management involves, in general, heavyweight batch processes focused on generating suitable indexes to enable efficient SPARQL resolution. Real-time management, on the contrary, focuses on continuous queries executed directly against the incoming stream. In practice, some data need to be discarded, or in the best case, are stored in non-optimized data structures. We argue that combining the two philosophies provides the benefits of both worlds, but how to do it correctly is still an open challenge.

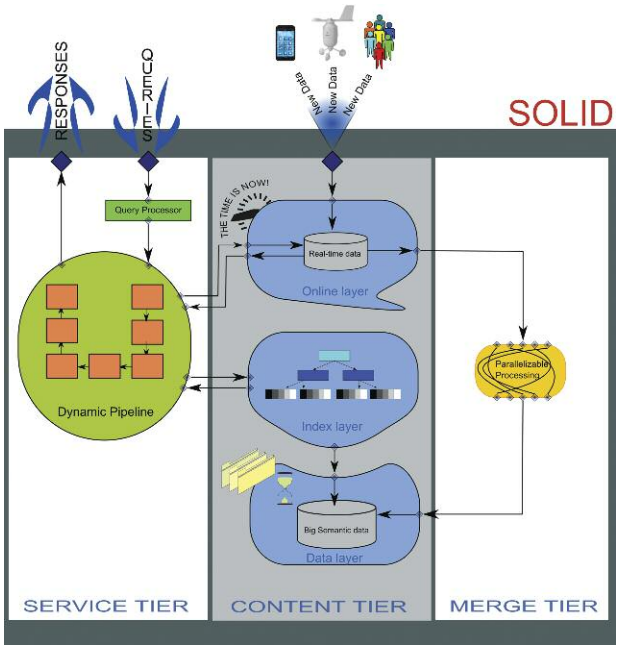


Figure 1: The SOLID architecture.

To this end, we propose Service-OnLine-Index-Data (SOLID) [1] as a scalable architecture that addresses separately the complexities of Big Data and real-time data management. SOLID proposes a three-tiered architecture (see Figure 1). The Content tier contains the repository of information. It comprises three layers optimized to provide efficient data retrieval depending on the stage of the information (historic versus live data). The Data layer stores raw RDF in compressed but accessible binary form, leveraging compactness and guaranteeing data immutability. The Index layer deploys a lightweight configuration of data structures on top of the Data layer to allow efficient querying capabilities with a minimal extra memory footprint. The Online layer is designed to store incoming RDF triples at a very high throughput. Since its performance degrades progressively, it must be dumped to the historic store eventually. The Merge tier implements this responsibility. It runs a batch process which integrates data from the Online layer into the Data Layer, updating the Index layer when the process finishes. Note that this integration process does not stop the overall operation. The Service Tier provides a SPARQL-based API to user applications. Its query processor instantiates a dynamic pipeline which delegates the retrieval to the Online and the Index layers, and binds their answers to obtain the final result.

We have implemented the SOLID components using state-of-the-art technology. On the one hand, the Data and Index layers are deployed using the open

RDF/HDT format [2]. It serializes the Big Semantic Data in highly compressed space, while providing SPARQL resolution. The Online layer uses a traditional RDF Store. Here, we make sure a competitive insertion throughput by keeping small to medium size collections; when the size is too big, data are sent to the Data Layer to be stored as historical information. Finally, the Merge and the Service tiers are implemented natively to ensure their efficiency. Our preliminary results show that the different layers excel at their tasks, and given that the Online layer contains few data, the integration is very lightweight, leading to competitive query performance in the latest SPARQL benchmarks, comparable to state-of-the-art RDF Stores but with a better write throughput.

Our future work includes tuning these layers and improving the algorithms that communicate them, providing SOLID as a general architecture, able to adapt to special needs of potential applications.

Links:

DataWeb Research Group:
http://dataweb.infor.uva.es
RDF/HDT Project:
http://www.rdfhdt.org
HDT W3C Member Submission:
http://www.w3.org/Submission/2011/03/

References:

[1] C.E. Cuesta, M.A. Martínez-Prieto, J.D. Fernández: "Towards an Architecture for Managing Big Semantic Data in Real-Time", in proc. of ECSA 2013, dx.doi.org/10.1007/978-3-642-39031-9_5

[2] J.D. Fernández, M.A. Martínez-Prieto, C. Gutiérrez et al.: "Binary RDF representation for publication and exchange", JWS vol. 19, 2013, dx.doi.org/10.1016/j.websem.2013.01.002

Please contact:

Miguel A. Martínez-Prieto, Javier D. Fernández
University of Valladolid, Spain
E-mail: migumar2@infor.uva.es, jfergar@infor.uva.es

Carlos E. Cuesta
Rey Juan Carlos University, Spain
E-mail: carlos.cuesta@urjc.es

Mario Arias
INSIGHT @ National University of Ireland Galway
E-mail: mario.arias@deri.org

MonetDB/RDF: Discovering and Exploiting the Emergent Schema of RDF Data

by Minh-Duc Pham and Peter Boncz

The Resource Description Framework (RDF) has been used as the main data model for the semantic web and Linked Open Data, providing great flexibility for users to represent and evolve data without need for a prior schema. This flexibility, however, poses challenges in implementing efficient RDF stores. It i) leads to query plan with many self-joins in triple tables, ii) blocks the use of advanced relational physical storage optimization such as clustered indexes and data partitioning, and iii) the lack of a schema sometimes makes it problematic for users to comprehend the data and formulate queries [1]. In the Database Architecture group at CWI, Amsterdam, we tackle these RDF data management problems by automatically recovering the structure present in RDF data, leveraging this structure both internally inside the database systems (in storage, optimization, and execution), and externally as an emergent schema towards the users who pose queries.

This research project is inspired by the work on "characteristic sets" (CS's) [2] showing that it is relatively easy to recover a large part of the implicit structure underlying RDF data. Here, a characteristic set is a set of properties that occurs frequently with the same subject. We have observed that structure typically surfaces with certain kinds of subjects having the same set of properties (belonging to the same CS), and certain CS's being connected over the same kind of property paths ("foreign key" relationships). The preliminary evaluation (see Figure 1) shows that even in the most dirty and messy RDF datasets, such as web-crawl data, 85% of the triples can be covered by a few hundred CS's. The key idea in this research is to discover a rough "emergent" relational schema which covers most input RDF triples (e.g., 85% of the dataset). This schema consists of a set of CS's and foreign key relationships between them, where tables and columns have logical names. This idea is being realized and experimentally evaluated inside MonetDB, an open-source column-store. By exploiting this schema, we provide better data locality and lower query processing cost. Our research focuses on three topics:

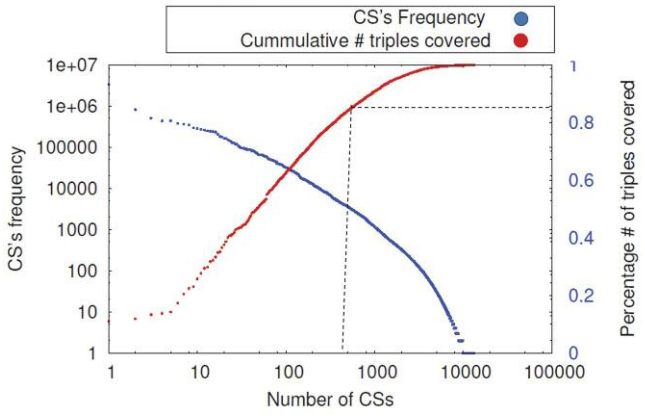


Figure 1: Number of CS's to cover 100 millions web-crawl RDF triples

- *Emergent Schema Recognition:* Our goal is to represent a large part of the RDF data using a compact and under-

standable relational schema, derived from its emergent structure. To achieve this, after using the basic algorithm to detect CS's based on common property sets, we enrich the schema by analysing the frequency distributions of Object-Subject references between CS's and analysing the literal type frequency distributions of CS's properties. To compact the schema, we merge CS's that are semantically and structurally similar, and further optimize the schema by filtering out dirty data such as infrequent Properties or Subjects that miss most properties. To allow users to understand and easily browse the schema via SQL-based tools, we develop methods to label each CS and its properties using understandable names exploiting (when available) ontologies, special Properties (e.g., rdf:type), the content of URI values, and labels of Properties that often reference the Subjects belonging to a CS.

- *Data Storage:* The emergent schema is exploited to build a physical storage scheme that clusters the triples in a particular order, such that relational clustered indexing and data partitioning become possible. From the RDF store point of view this comes down to Subject clustering where loaded triples get a renumbered Subject: for each CS, the Subjects that belong to it get densely ascending object identifiers. The PSO triple representation (ie, triples in lexicographic Predicate–Subject–Object order) thus becomes highly efficient, as P and S compress away because they are repetitive resp. densely increasing. It can alternatively be viewed as columnar table storage of only Objects, and, in fact, gets stored as regular MonetDB table columns. The resulting "CS" tables are directly usable in SQL queries. Note that these tables do not represent the irregular triples (those that do not belong to a CS). These are stored separately as a triple table, and only play a role in SPARQL query answering.
- *Query Execution:* To leverage this emergent schema in SPARQL queries, we propose to extend the database kernel with new algorithms (RDFscan and RDFjoin). These operators accelerate typical "star join" patterns in SPARQL queries, avoiding self-joins. They stitch together CS columns in one sequential scan, like column stores do, but simultaneously merge-join the data with the irregular triple table to provide correct SPARQL query answers.

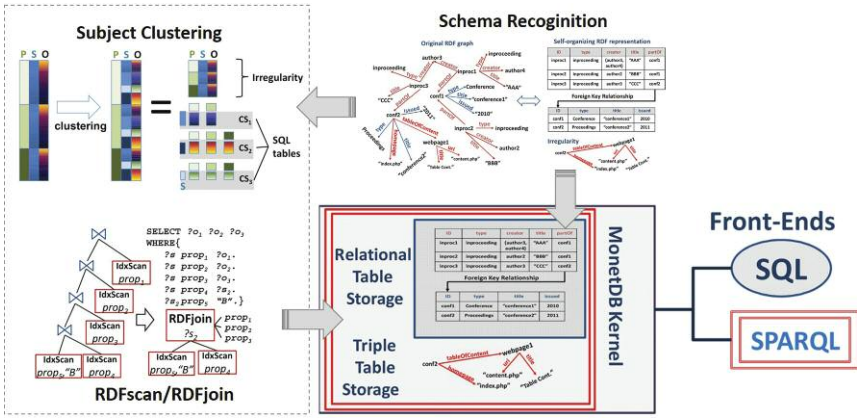


Figure 2: Overview of MonetDB/RDF

We believe this architecture will benefit RDF data management. It solves the performance issues of RDF stores, adds semantic web features to the relational model, helps SPARQL users better comprehend RDF data, and allows them to enjoy the benefits of the relational tool-chain.

Link:
homepages.cwi.nl/~duc/monetrdf/

References:
[1] P. Minh-Duc, P.A. Boncz: "Self-organizing Structured RDF in MonetDB," PhD Symposium, ICDE, 2013, http://oai.cwi.nl/oai/asset/20747/20747B.pdf
[2] T. Neumann, G. Moerkotte: "Characteristic sets: Accurate cardinality estimation for RDF queries with multiple joins", in ICDE, IEEE, 2011, dx.doi.org/10.1109/ICDE.2011.5767868

Please contact:
Minh-Duc Pham
CWI, The Netherlands
E-mail: duc@cwi.nl

Ontology-based Integration of Heterogeneous and Distributed Information of the Marine Domain

by Yannis Tzitzikas, Carlo Allocca, Chryssoula Bekiari, Yannis Marketakis, Pavlos Fafalios and Nikos Minadakis

The European iMarine project has defined a core ontology for publishing marine data, which is suitable for setting up warehouses that can answer complex queries.

One of the main characteristics of biodiversity data is its cross-disciplinary nature and its extremely broad range of data types, structures and semantic concepts. Biodiversity data, especially in the marine domain, is widely distributed, with few well-established repositories or standard protocols for archiving, access and retrieval. Currently, the various laboratories have in place databases for keeping their raw data (e.g. observations), while ontologies are primarily used for metadata that describe their raw data.

iMarine (funded by the European Commission under Framework Programme 7, Nov 2011-April 2014) is an open and collaborative initiative for establishing a data infrastructure to support the "ecosystem approach" to fisheries management and conservation of marine living resources. It is coordinated by Consiglio Nazionale delle Ricerche (IT) and GEIE ERCIM (FR)

(for the complete list of partners please refer to the website).

One of the challenges of the iMarine project is to enable users to access a coherent source of facts about marine entities, rather than a bag of contributed contents. Queries like "Given the scientific name of a species, find its predators with the related taxon-rank classification and with the different codes that the organizations use to refer to them", could not be formulated (and consequently nor answered) by any individual source. To formulate such queries we need an expressive conceptual model, while to answer them we also have to assemble pieces of information stored in different sources.

For this reason we have designed and implemented a top level ontology: MarineTLO[1]. MarineTLO is generic enough to provide consistent abstractions or specifications of concepts included in all data models or ontologies of marine data sources and provide the necessary properties to make this distributed knowledge base a coherent source of facts relating observational data with the respective spatiotemporal context and categorical (systematic) domain knowledge. It can be used as the core schema for publishing Linked Data, as well as for setting up integration systems for the marine domain. It can be extended to any level of detail on demand, while preserving monotonicity. For its development and evolution we have adopted an iterative and incremental methodology where a new version is released every two months. For the implementation we use OWL 2, and to evaluate it we use a set of query requirements provided by the related communities.

To answer complex queries, we have to assemble pieces of information stored in different sources. For this reason, we have established a process (supported by a tool that we have developed for this purpose) for creating MarineTLO-based warehouses that integrate information derived from various sources. To fetch the data we have to use a plethora of access methods (SPARQL endpoints, HTTP accessible files, JDBC), while to connect the fetched data we have to define schema mappings, transformations, as well as rules for instance matching. The current version of the warehouse integrates information coming from WoRMS, ECOSCOPE, FLOD, FishBase and DBPedia, contains around three million triples, and provides harmonized and integrated information for about 37,000 distinct marine species. The warehouse is currently used for generating fact sheets (e.g. TunaAtlas from IRD), and for enhancing the search services offered by the iMarine infrastructure (specifically the semantic post-processing of search results).

Figure 1 shows how information from three different sources about the same species can be assembled, while Figure 2 describes the contents of the current MarineTLO-based warehouse. We plan to continue these activities until the end of iMarine and beyond. Currently we focus on methods for quantifying the quality and value of such warehouses.

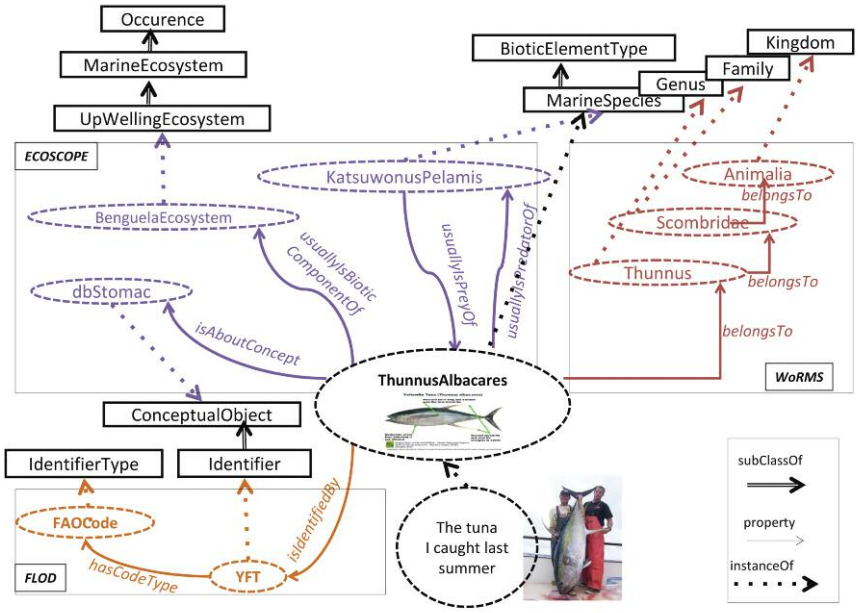


Figure 1: Assembling complementary information about "Thunnus Albacares"

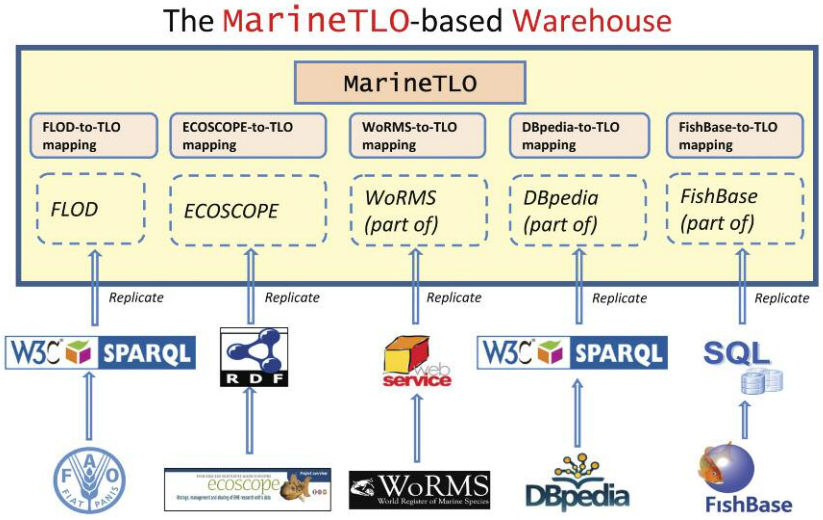


Figure 2: The current MarineTLO-based Warehouse

Links:
Website of iMarine Project
http://www.i-marine.eu/
Documentation and OWL version of the ontology MarineTLO:
http://www.ics.forth.gr/isl/MarineTLO/

Reference:
[1] Y. Tzitzikas et al.: "Integrating Heterogeneous and Distributed Information about Marine Species through a Top Level Ontology", in proc. of MTSR'13, 2013, dx.doi.org/10.1007/978-3-319-03437-9_29

Please contact:
Yannis Tzitzikas
FORTH-ICS and University of Crete
Tel: +30 2810 391621
E-mail: tzitzik@ics.forth.gr

Research activities and innovative developments in European research institutes

The D4Science Research-Oriented Social Networking Facilities

by Massimiliano Assante, Leonardo Candela, Donatella Castelli and Pasquale Pagano

Modern science calls for innovative practices to facilitate research collaborations spanning institutions, disciplines, and countries. Paradigms such as cloud computing and social computing represent a new opportunity for individuals with scant resources, to participate in science. The D4Science.org Hybrid Data Infrastructure combines these two paradigms with Virtual Research Environments in order to offer a large array of collaboration-oriented facilities as-a-Service.

Scientists are expected to produce enhanced forms of scientific communication based on publication of "comprehensive scientific theories" – including the data and algorithms on which they are based –to make it possible for "others to identify errors, to support, reject or refine theories and to reuse data for further understanding and knowledge" [1]. This is a pressing requirement not only in the context of "big sciences" (e.g. physics, astronomy, earth observation) but also in the long tail of science, ie the large number of relatively small laboratories and individual researchers that have the potential to produce a bulk of scientific knowledge but have no access to large-scale dedicated IT.

To effectively serve such scenarios, D4Science.org is operating a Hybrid Data Infrastructure (HDI), ie an IT infrastructure built as a "system of systems" that integrates other infrastructures including grid and cloud, services and information systems. The HDI can thus offer its users a disparate set of technologies, computing and storage resources made dynamically available via the elastic acquisition model offered by Cloud technologies. It provides Virtual Research Environments (VREs) "as-a-Service", ie, web-based working environments where groups of scientists can transparently and seamlessly access shared sets of resources (data, tools and computing capabilities). The VRE services include a social networking area that promotes innovative scientific collaboration patterns inspired by social computing and supported by the underlying infrastructure facilities.

This social networking facility provides an environment to foster large scale collaborations, ie scenarios where many - potentially geographically distributed - co-workers can access and process large amounts of data. It offers: (a) a continuously updated list of events/news produced by users and applications (Home Social) that becomes an easy-to-use trigger of a continuously updated timeline of user-centric facts, (b) a folder-based file system to manage complex information objects, including files, datasets, workflows and maps, in a seamless way (Workspace), (c) an email-like facility for exchanging "large" messages with co-workers, ie messages with attachments in the form of the complex information objects described above (Messages), (d) a list of events organized by date, e.g. publication of, or comments on

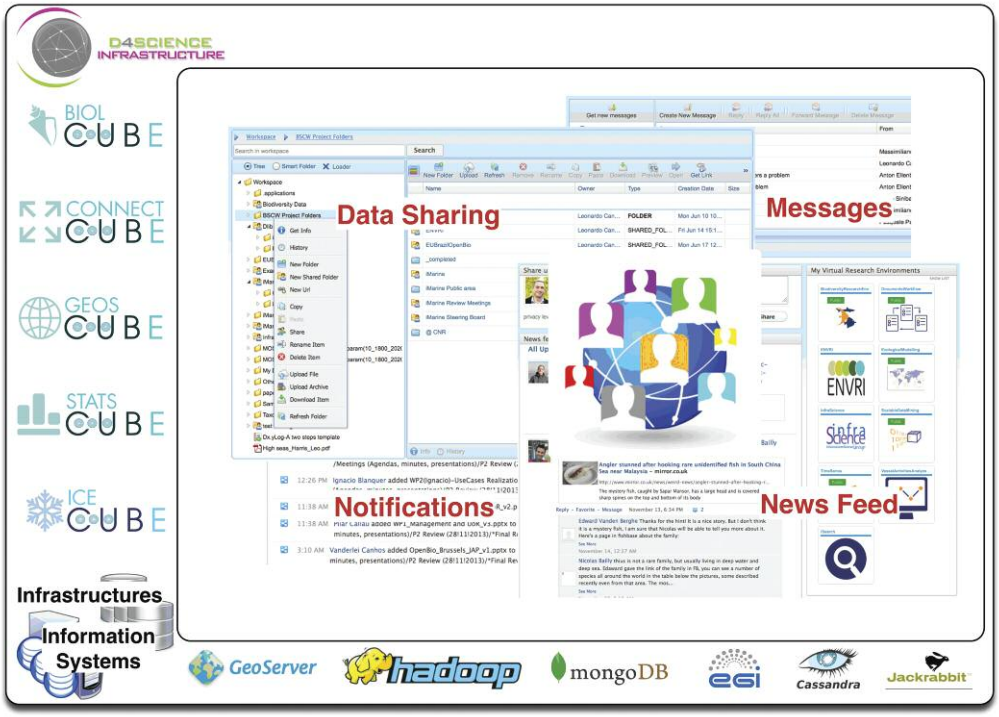


Figure 1: D4 Science infrastructure.

a research product (Notifications), (e) a settings area where the user can configure various aspects of the system including his/her data and notification preferences (Personalization).

The Home Social consists of two facilities. The News Feed makes users' and applications' updates available to every user according to his/her preferences and enables users to comment, subscribe or re-share these updates. These updates are "actionable", e.g. contain a link to the actual product or facility. The Share Updates enables users to post updates or interesting links to others and applications to post possible updates of a new product or facility.

The Workspace resembles a folder-based file system, where the added value is represented by the type of information objects it can manage in a seamless way. It supports items ranging from binary files to information objects representing tabular data, workflows, species distribution maps, time series, and comprehensive research products. Through it, data sharing is fostered, making results, workflows, annotations and documents etc immediately available.

The Messages realise a common email environment as-a-Service whose distinguishing feature is its integration with the rest, e.g. it is possible to send as an attachment any information object residing in the workspace, however big and complex, without consuming bandwidth.

The Notifications alert users on an as-it-happens basis. These notifications offer a sense of anticipation and create a productivity boost. Users receive an alert (through a priori selected channels, e.g. email, web portal, Twitter) notifying them when something of interest has happened in their VRE(s).

The Personalization provides users with facilities to customize the overall behaviour of the "social area". It enables information to be specified, including biographic data, interests and skills, and notification settings.

The D4Science social networking facilities will reshape the modern approach to communication, largely implemented by LinkedIn, Twitter, and Facebook, by porting it to research communities. Sharing of large datasets, quite common in the big data era, and workflows, have become as easy as sharing an image. Scientific collaboration and communication have become immediate and smart by sending a short message or posting a tweet. The virtualization of the working environments through the on-demand and timely creation of dedicated VREs, the virtualization of the resources offered as-a-Service through the HDI, and now the support for collaboration and communication make D4Science a unique service for the effective production of comprehensive scientific theories.

Link:
D4Science website: www.d4science.org

Reference:
[1] G. Boulton et al.: "Science as an open enterprise," The Royal Society, Final report, 2012, http://royalsociety.org/policy/projects/science-public-enterprise/report/

Please contact:
Massimiliano Assante
ISTI-CNR, Italy
E-mail: massimiliano.assante@isti.cnr.it

ShapeForge: Modeling by Examples for 3D Printing

by Sylvain Lefebvre

Consumer level 3D printing holds great promise for the future [1]. However, few people possess the required skills and time to create elegant designs that conform to precise technical specifications. "By–example" shape synthesis methods are promising to address this problem: New shapes are automatically synthesized from existing ones, by deforming or assembling parts cutout of examples.

The ShapeForge project, funded by the European Research Council ("Starting Grant Project"), aims at helping users design new, complex objects from examples. The goal is to preserve the appearance of the examples while ensuring that the newly generated objects enforce a specific purpose, such as supporting weight distributed in space, standing balanced in a specific position, or providing storage spaces. Such constraints are crucial for designing many of the common objects surrounding us: containers, enclosures, various parts with mechanical functions; but also for artists and enthusiasts wishing to fabricate interesting, surprising objects. We also consider fabrication constraints: There are several limitations to the additive manufacturing processes, such as minimal thickness or maximal overhang angles. We therefore investigate algorithms that can turn a virtual object into a design that can be fabricated.

As an example of our methodology, we recently developed, in collaboration with ETH Zurich, a method for balancing 3D models [2]. Many artistic 3D objects designed on a computer are created without any special consideration for the laws of physics. As long as the models remain in a computer this is of no consequence. However, fabrication through 3D printing breaks the illusion: printed models topple instead of standing as initially intended. We aimed to assist users in producing novel, properly balanced designs. Balance optimization is formulated as an energy minimization, improving stability by modifying the volume of the object, while preserving its surface details. Our optimizer combines two operations changing the mass distribution of the object: Carving and deformation. This takes place during interactive

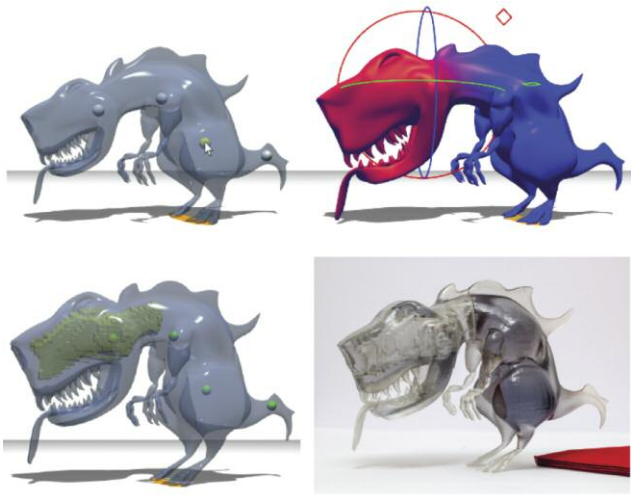


Figure1: A T-Rex model (top left) is deformed by the user to have a bigger head (top right). Our algorithm automatically carves the model (bottom left, cavity visible in yellow) and slightly deforms it to achieve a balanced configuration. After 3D printing (bottom right), the physical model stands in the intended position. (3D model from www.turbosquid.com: T-Rex by csirkeFrs).

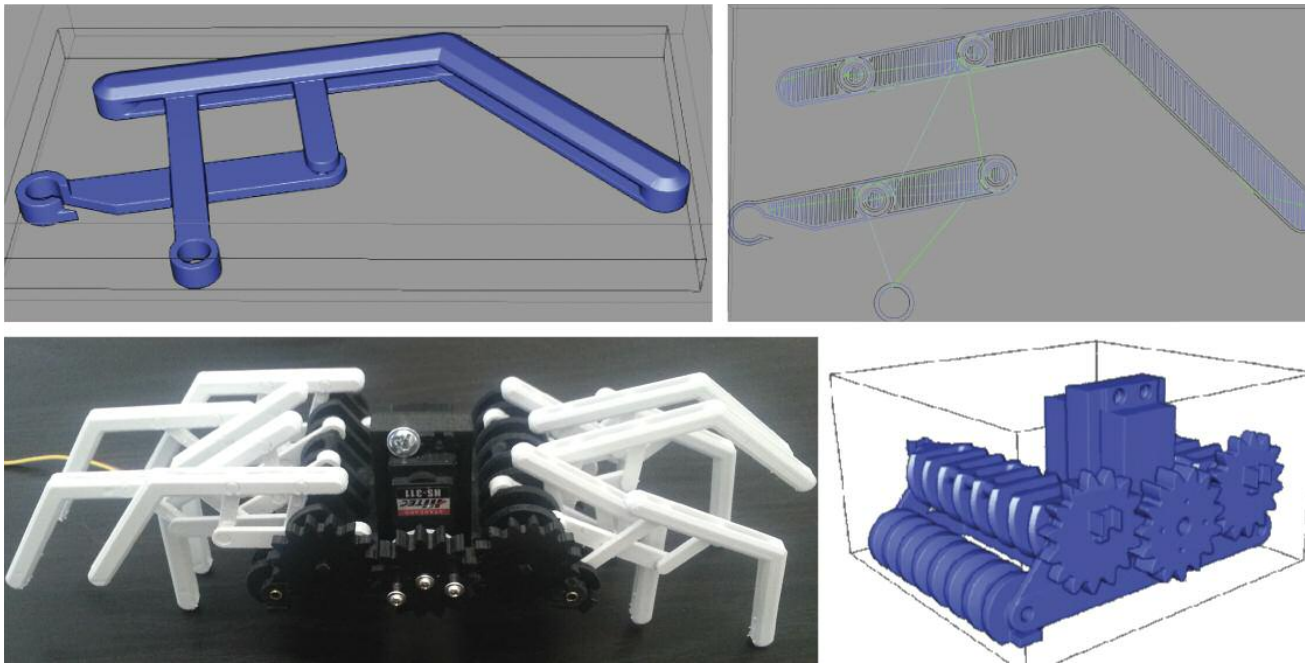


Figure2: A spider robot we designed with our software. Top left: An articulated leg designed to print as a single piece: no further assembly is required. Top right: The toolpath for plastic deposition on one layer, computed by our software. Bottom left: The final robot, entirely printed but for the servo motor and visible screws. Bottom right: The model of the body of the robot.

editing: the user cooperates with our optimizer to achieve the end result. With our technique, users can produce fabricated objects that stand in one or more surprising poses without requiring glue or heavy pedestals, as illustrated Figure 1.

The ShapeForge project requires a full understanding of the creation pipeline, from the modelling of the object to the fabrication on the 3D printer. We are developing an innovative CAD software for additive manufacturing [3]. It starts from a description of an object in terms of Boolean operations between solids (union, difference, intersection) and directly generates the instructions driving the printer. Our software is especially well-suited for combining parts from existing designs, as well as dealing with shapes with complex, intricate geometries. Most importantly, our software erases the boundaries between modelling and the printing process, letting us develop novel approaches considering both issues simultaneously. Figure 2 illustrates a spider robot designed with our publicly available.

With ShapeForge we hope to bring novel ways to model complex shapes, developing algorithms that handle the difficult task of finding a compromise between the intention of the designer, the technical requirements of the fabrication process, and the function of the final, real object.

Links:
IceSL: http://webloria.loria.fr/~slefebvr/icesl/
Walking robot: http://www.thingiverse.com/thing:103765

References:
[1] H. Lipson, M. Kurman: "Factory@home: The emerging economy of personal fabrication", Report Commissioned by the Whitehouse Office of Science & Technology Policy, 2010, http://diyhpl.us/~bryan/papers2/open-source/The%20emerging%20economy%20of%20home%20and%20desktop%20manufacturing%20-%20Hod%20Lipson.pdf

[2] R. Prévost et al.: "Make it Stand", in proc. of SIGGRAPH (2013), http://dx.doi.org/10.1145/2461912.2461957

[3] S. Lefebvre: "IceSL: A GPU Accelerated modeler and slicer", in proc. of AEFA 2013, http://webloria.loria.fr/~slefebvr/icesl/icesl-whitepaper.pdf

Please contact:
Sylvain Lefebvre,
Inria Nancy - Grand Est, France
E-mail: sylvain.lefebvre@inria.fr

Discriminating Between the Wheat and the Chaff in Online Recommendation Systems

by Roberto Di Pietro, Marinella Petrocchi, Angelo Spognardi and Maurizio Tesconi

MyChoice is an ambitious research project that aims to model, detect, and isolate outliers (aka fake) in online recommendation systems, as well as in online social networks. The final outcome of the project will be the prototype of an automated engine able to recognize fake information, such as reviews, and fake friends/followers, and able to filter out malicious material, in order to return reliable and genuine content to the user .

MyChoice aims to provide novel models and tools to search genuine and unbiased content on Web platforms, while filtering out partial and fake information. The expected outcome of the project is twofold: firstly, focusing on real online recommendation systems, MyChoice intends to tackle the (malicious) bias that may influence a high percentage of users. Secondly, the project pays attention to fake accounts on social networks and provides automatic fake detection techniques. As an example on Twitter, "fake followers" are those accounts created to inflate the number of followers of a target account, to make it more trustworthy and influential, in order to stand out from the crowd and attract other genuine followers.

Fake reviews are currently so widespread that this phenomenon has captured the attention of academia and the mass media. Fake reviews can influence the opinions of users, having the effect of either promoting or damaging a particular target, thus a strong incentive exists for opinion spamming. Markets are strongly influenced by review scores: a recent survey by Cornell University on Internet travel booking revealed that online advice had a strong impact on bookings, occupancy rates, and revenue of commercial accommodation establishments. Defining efficient methodologies and tools to mitigate proliferation of fake reviews has become a compelling issue that MyChoice is addressing. One main outcome of the project will be a prototype for an unbiased ranking system, which identifies and evicts malicious reviews and provides a more appropriate choice of services and products, based on the fusion of their objective characteristics and subjective tastes and interests of the individual user.

The project started its activity in 2012, monitoring some of the most popular websites providing online advice for hotels, such as TripAdvisor, and online services for e-booking, such as Booking. A crawler was used to collect several million reviews relating to thousands of different hotels all around the world. Starting from the state-of-the-art in the field, the researchers involved in the project quantified the robustness of the rating aggregators used in such systems, against the malicious injection of fake reviews. The current experi-

mental outcomes, for example, enrich past results attesting that a simple arithmetic mean of the ratings by the hotel guests (which is the usual way to provide aggregated information to users) is not the most robust aggregator, since it can be severely affected by even a small number of outliers. Experiments have been carried out considering different kinds of attack, such as batch injections, hotel-chain injections, and local competitor injections. To improve the robustness of the ranking, the project is defining new aggregators to more effectively tackle the activity of malicious reviewers.

To enhance the comprehension of the fake phenomenon, the project is also looking at other instances of the concept of "fake". In particular, a research effort is focusing on the proliferation of fake Twitter followers, which has also aroused a great deal of interest in the mainstream media, such as New York Times and Financial Times. We have created a "gold standard", namely a collection of both truly genuine (human) and truly fake accounts. In December, 2012, MyChoice launched a Twitter campaign called "The Fake Project", with the creation of the Twitter account @TheFakeProject, whose profile claims "Follow me only if you are NOT a fake". To obtain the status of "certified human", each account that adheres to the initiative was the target of further checks to attest its credibility. The "certified fake" set was collected by purchasing fake accounts, which are easily accessible to the general public. Based on experiments over the gold standard, we are determining to what extent the problem of detecting fakes can be considered similar to the problem of detecting spam. To this end we are leveraging machine-learning classifiers trained on the gold standard to evaluate how the state-of-the art proposals for spammer Twitter account detection [1-2] perform on fake account detection [3]. This has been achieved by crawling 600,000 tweets and around one million Twitter accounts. Classifiers instrumented with the best features are able to obtain highly accurate fake detections (higher than 95%); hence, they can be used to estimate the number of fake followers of any Twitter account.

The classifiers highlight a behavioural difference between humans, spammers and fake accounts. In our opinion, this is the basis for a more thorough comprehension of the fake phenomenon, which can lead to formal modelling that can help discriminate between an anomalous (possibly fake) account and a standard (possibly legitimate) one. Using this formalization as a reference model, the definition of fakes could be exported into different contexts, even to online reviews and reviewers.

In conclusion, combining a robust metrics with the formalization of "fakeness" leads to the final goal of MyChoice: to develop a prototype for a ranking system able to discriminate between genuine and unbiased information, getting rid of malicious content, and providing the user with reliable search results.

MyChoice is a regional project funded by the Tuscany region under the "Programma operativo regionale Competitività e Occupazione (Por Cro)" Program, within the European Union Social Fund framework 2007-2013, the Institute for Informatics and Telematics of the Italian National Research Council (IIT-CNR), and the start up company Bay31. It is a two-year project, which started in November 2012.

Figure 1: Screenshot of the Twitter account @TheFakeProject, used to launch the campaign for recruiting the real humans in a training dataset of Twitter accounts

Links:
http://twitter.com/TheFakeProject
http://wafi.iit.cnr.it/TheFakeProject/

References:
[1] C. Yang et al.: "Die free or live hard? Empirical evaluation and new design for fighting evolving Twitter spammers", in proc. of RAID 2011, Springer, http://dx.doi.org/10.1007/978-3-642-23644-0_17

[2] G. Stringhini et al.: "Detecting spammers on social networks", in proc. of ACM ACSAC '10, http://dl.acm.org/citation.cfm?id=1920263

[3] S. Cresci et al.: "Fake accounts detection on Twitter", Tech Rep IIT-CNR nr. 15-2013, http://www.iit.cnr.it/node/22730

Please contact:
Marinella Petrocchi
IIT-CNR, Italy
Tel: +390503153432
E-mail: marinella.petrocchi@iit.cnr.it

Uncovering Plagiarism - Author Profiling at PAN

by Paolo Rosso and Francisco Rangel

PAN is a yearly workshop and evaluation lab on uncovering plagiarism, authorship, and social software misuse.

Since 2009, PAN has been organizing benchmark activities on uncovering plagiarism, authorship, and social software misuse [1]. An additional task - author profiling - has also recently been proposed. Author profiling, instead of focusing on individual authors, studies how language is shared by a class of people. Author profiling is a problem of growing importance in applications in forensics, security and marketing. For instance, a person working in the area of forensic linguistics may need to know the linguistic profile of a suspected text message (language used by a certain type of person) and identify characteristics (with language as evidence). Similarly, from a marketing viewpoint, companies may be interested in determining, through the analysis of blogs and online product reviews, what types of people like or dislike their products.

Author profiling at PAN [2] has been focusing on gender and age identification in social media, both in English and Spanish. We looked for open and public online repositories with posts labelled with author demographics. For age identification, three classes were considered: 10s (13-17), 20s (23-27) and 30s (33-47). We also incorporated a small number of samples from conversations of sexual predators, together with samples from adult-adult sex conversations, with the aim of unveiling fake profiles of potential sexual predators.

With 21 teams, author profiling was one of the most popular tasks at the CLEF conference in 2013. Participants took diverse approaches to the problem: content-based, stylistic-based, n-gram based, etc. Accuracy for gender and age identification, both in English and Spanish, is shown in Figures 1 and 2. Results show the difficulty of the task in a challenging scenario (with 374,100 authors), in particular for gender identification - although the accuracy was slightly higher in Spanish with it being a gender-marked language. With respect to the texts of sexual predators, correct demographics were identified by majority of the participants.

Apart from PAN@CLEF, another two tasks - WCPR@ICWSM and BEA@NAACL-HLT (see links) - were organized in 2013 on predicting different aspects of an author's demographics: specifically, personality traits and native language. This shows the increasing interest of the research community in author profiling. In 2014, PAN will once again be organizing the task on author profiling in social media, as well as tasks on author identification and plagiarism detection.

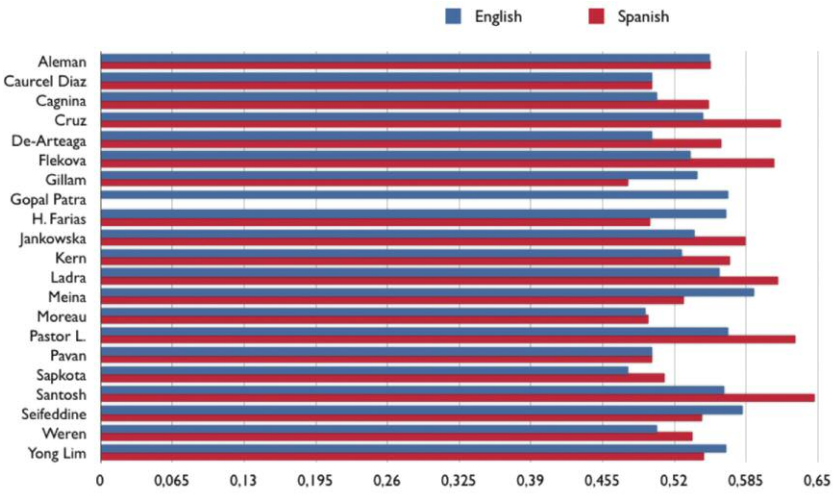


Figure 1: Accuracy for gender identification of social media profiles

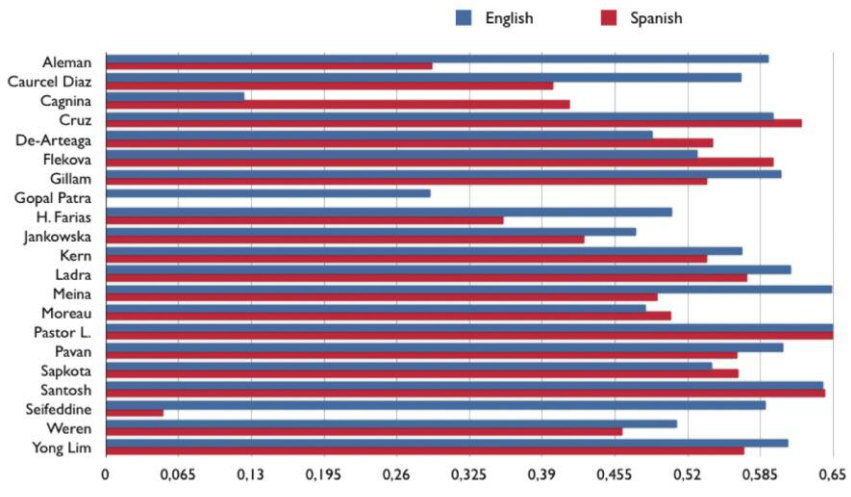


Figure 2: Accuracy for age identification of social media profiles

Links:

PAN: http://pan.webis.de/

http://mypersonality.org/wiki/doku.php?id=wcpr13

http://www.cs.rochester.edu/ ~tetreaul/naacl-bea8.html

References:

[1] T. Gollub, M. Potthast, A. Beyer et al.: "Recent Trends in Digital Text Forensics and its Evaluation: Plagiarism Detection, Author Identification, and Author Profiling", in proc. of the 4th International Conference of the CLEF Initiative (CLEF 13), Springer LNCS 8138, 2013, dx.doi.org/10.1007/978-3-642-40802-1_28

[2] F. Rangel, P. Rosso, M. Koppel et al: "Overview of the Author Profiling Task at PAN 2013. In P. Forner, R. Navigli, D. Tufis, Eds., Working Notes Papers of the CLEF 2013 Evaluation Labs, 2013, http://www.clef-initiative.eu/documents/71612/2e4a4d3a-bae2-47f9-ba3c-552ec66b3e04

Please contact:

Paolo Rosso
Natural Language Engineering Lab, Universitat Politècnica de València, Spain
E-mail: prosso@dsic.upv.es

Call for Participation

22nd Intl. Conference on Pattern Recognition - Unsupervised Image Segmentation Contest and Workshop

Stockholm, 24-28 August 2014

In order to promote evaluation of unsupervised color image segmentation algorithms using publicly available data sets, standard performance assessment methodology and on-line web verification server and database (The Prague Texture Segmentation Datagenerator and Benchmark), a competition for the best segmentation algorithms will take place in conjunction with the ICPR 2014 conference.

Although numerous different methods were already published the ill-defined segmentation problem is still far from being solved. In addition, very little is known about properties and behaviour of already published segmentation methods and their potential user is left to randomly select one due to absence of any counselling.

The aim of the contest is to overcome these problems by suggesting the most promising approaches to unsupervised learning and image segmentation and to unify the verification methodology used in the image segmentation research.

The authors of the best performing algorithm of the competition will receive a prize that will be presented at ICPR 2014. The performance of all submitted algorithms will be summarised in a presentation given at the conference. The best participants will be invited to publish their approach in a special journal issue.

Important Dates:

- 13: June 2014: Deadline for the final results submission
- 24 August 2014: ICPR 2014 contest workshop
- 19 September 2014: Contest special journal issue paper submission

More information:
http://mosaic.utia.cas.cz/icpr2014/
http://www.icpr2014.org/

ERCIM DES Working Group Workshops

As in the previous years, the ERCIM Working Group Dependable Embedded Systems (DES) is organising two workshops at renowned conferences in cooperation with EWICS (European Workshop on Industrial Computer Systems Reliability, Safety and Security) and the ARTEMIS Embedded Computing Systems Initiative:

- A workshop on Dependable Embedded Cyber-physical Systems at SAFECOMP 2014
- a special session on teaching, education and training for Dependable Embedded and Cyberphysical Systems at Euromicro SEAA/DSD, Verona, 27-29 August 2014.

SAFECOMP'14

The 33rd International Conference on Computer Safety, Reliability and Security, takes place this year from 10-12 September 2014 in Florence, Italy. The key theme is "Safety in presence of evolution: design, assessment and certification methods". A separate call for the workshop will be published in February. SAFECOMP '14 is co-located with the EPEW, FMICS, FORMATS and QEST conferences, so there is a rich portfolio available.
http://www.safecomp2014.unifi.it/
http://www.florence2014.org).

EUROMICRO SEAA

The 40th Conference on Software Engineering and Advanced Applications will be held in Verona from 27-29 August 2014. SEAA is a long-standing international forum to present and discuss the latest innovations, trends, experiences, and concerns in the field of Software Engineering and Advanced Applications in information technology for software-intensive systems. The call for papers for the Special TET-DEC Session is already available under SEAA/Call for Papers). Deadline for abstracts is 13 February, for papers 20 Febuary 2014 (may be extended).
http://esd.scienze.un ivr.it/dsd-seaa-2014/

Please contact:
Erwin Schoitsch, AIT Austrian Institute of Technology/ AARIT
E-mail: Erwin.Schoitsch@ait.ac.at

Call for Papers

R&D Management Conference 2014

Stuttgart, 3-6 June 2014

International R&D Management Conference brings together experts from research and industry.

"Management of Applied R&D: Connecting research and development with future markets" is the topic of the international R&D Management Conference 2014, hosted by Fraunhofer IAO in Stuttgart from 3-6 June.

The conference will offer a unique mix of scientific presentations and stimuli from industry to an expected audience of around 200 international participants. Among the high-profile keynote speeches are presentations from Siemens, Bosch and Festo. As one of the fathers of the MP3 codec, Dr. Bernhard Grill from the Fraunhofer Institute for Integrated Circuits IIS will tell the R&D success story behind this groundbreaking technology. The call for papers itself is structured into three broad categories – R&D Organization and Efficiency, Strategic R&D and Technology Management, and Innovative IT Systems in R&D. These are complemented by special sessions on selected topics such as R&D Management in Emerging Economies or Sustainability and R&D Management, each be jointly led by a representative from academia and industry. A separate option for participating only to the keynotes, discussions and the conference dinner on June 5th will be available.

Abstracts for papers to be presented at the conference will be accepted until January 20, 2014.

More information:
http://www.rnd2014.iao.fraunhofer.de

Please contact:
Sven Schimpf
Tel: +49 711 970-2457
E-mail:
sven.schimpf@iao.fraunhofer.de

Call for Contributions

ACM/IEEE 17th International Conference on Model-Driven Engineering Language & Systems

Valencia, Spain
28 September - 3 October 2014

MODELS (formerly the UML series of conferences) is the premier conference series for model-based software and systems engineering which since 1998 has been covering all aspects of modeling, from languages and methods to tools and applications. MODELS in its 17th edition cordially invite contributions related to all aspects of model-based engineering.

MODELS 2014 challenges the modeling community to promote the magic of modeling by solidifying and extending the foundations and successful applications of modeling in areas such as business information and embedded systems, but also by exploring the use of modeling for new and emerging systems, paradigms, and challenges including cyber-physical systems, cloud computing, services, social media, security, and open source. We invite you to join us in Valencia and to help shape the modeling methods and technologies of the future!

Foundations Track Papers

We invite authors to submit original papers in the following categories:

1. Technical papers describing original scientifically rigorous solutions to significant model-based development problems.
2. Exploratory papers describing new, non-conventional model-based development research positions or approaches.
3. Empirical evaluation papers assessing existing problem cases or scientifically validating proposed solutions through, e.g., controlled experiments, case studies, or simulations. Authors are encouraged to make the artifacts used for the evaluation publically accessible by, e.g., uploading them to the Repository for Model-Driven Development (ReMoDD).
4. Modeling Pearls papers describing polished, elegant, instructive, and insightful applications of modeling techniques or approaches. Modeling pearls include demonstration case studies, examples of bad modeling practices that exemplify the problems that can occur if they are not detected and rectified, and models used in the classroom to illustrate modeling concepts and practices.

MDE Practice Track Papers

We invite authors to submit original experience reports and case studies. Each paper should provide clear take-away value by describing the context of a problem of practical, industrial importance and application. The paper should discuss why the solution of the problem is innovative, effective, or efficient and what likely industrial impact it has or will have; it should provide a concise explanation of the approach, techniques, and methodologies employed, and explain the best practices that emerged, tools developed, and/or software processes involved.

Deadlines for Submissions

- 13 March: Conference paper abstracts, workshop and tutorial proposals
- 20 March: Conference full paper submissions
- 11 July: Demonstrations and poster submissions; ACM Student Research Competition paper submissions; Doctoral Symposium and Educator's Symposium paper submissions

More information:
http://www.modelsconference.org

W3C Workshop

Call for Participation

W3C Workshop on Web Payments - How do you want to pay?

Paris 24-25 March 2014

This new W3C workshop, to be held at Palais Brongniart (La Bourse) in Paris on 24-25 March 2014, seeks to make it easier to monetize open Web applications, as an effective alternative to proprietary native app ecosystems. The aim is to improve the end user experience and give users greater freedom in how they pay, to reduce the burden on developers and merchants, and to create a level playing field for competing payment solutions providers large and small.

We are expecting wide participation from financial institutions, governments, mobile network operators, payment solution providers, technology companies, retailers, and content creators. The workshop will seek to establish a broad roadmap for work on open standards for Web payments, along with some concrete proposals for initial steps along the road. Registration is free although a statement of interest or position paper is needed before 8 February 2014.

The W3C Workshop on Web payments is hosted by Ingenico and sponsored by Gemalto. The workshop is funded by the European Union through the Seventh Framework Programme (FP7/2013-2015) under grant agreement n°611327 within the HTML5 Apps EU project

More information:
Web Payments Workshop CfP: http://www.w3.org/2013/10/payments/
HTML5 Apps EU project: http://html5apps-project.eu/

ERCIM
European Research Consortium for Informatics and Mathematics

ERCIM – the European Research Consortium for Informatics and Mathematics is an organisation dedicated to the advancement of European research and development, in information technology and applied mathematics. Its member institutions aim to foster collaborative work within the European research community and to increase co-operation with European industry.

ERCIM is the European Host of the World Wide Web Consortium.

Austrian Association for Research in IT
c/o Österreichische Computer Gesellschaft
Wollzeile 1-3, A-1010 Wien, Austria
http://www.aarit.at/

Portuguese ERCIM Grouping
c/o INESC Porto, Campus da FEUP,
Rua Dr. Roberto Frias, nº 378,
4200-465 Porto, Portugal

Consiglio Nazionale delle Ricerche, ISTI-CNR
Area della Ricerca CNR di Pisa,
Via G. Moruzzi 1, 56124 Pisa, Italy
http://www.isti.cnr.it/

Science and Technology Facilities Council
Rutherford Appleton Laboratory
Chilton, Didcot, Oxfordshire OX11 0QX, United Kingdom
http://www.scitech.ac.uk/

Czech Research Consortium
for Informatics and Mathematics
FI MU, Botanicka 68a, CZ-602 00 Brno, Czech Republic
http://www.utia.cas.cz/CRCIM/home.html

Spanish Research Consortium for Informatics and Mathematics
D3301, Facultad de Informática, Universidad Politécnica de Madrid
28660 Boadilla del Monte, Madrid, Spain,
http://www.sparcim.es/

Centrum Wiskunde & Informatica
Science Park 123,
NL-1098 XG Amsterdam, The Netherlands
http://www.cwi.nl/

SICS Swedish ICT
Box 1263,
SE-164 29 Kista, Sweden
http://www.sics.se/

Fonds National de la Recherche
6, rue Antoine de Saint-Exupéry, B.P. 1777
L-1017 Luxembourg-Kirchberg
http://www.fnr.lu/

Magyar Tudományos Akadémia
Számítástechnikai és Automatizálási Kutató Intézet
P.O. Box 63, H-1518 Budapest, Hungary
http://www.sztaki.hu/

FWO
Egmontstraat 5
B-1000 Brussels, Belgium
http://www.fwo.be/

F.R.S.-FNRS
rue d'Egmont 5
B-1000 Brussels, Belgium
http://www.fnrs.be/

University of Cyprus
P.O. Box 20537
1678 Nicosia, Cyprus
http://www.cs.ucy.ac.cy/

Foundation for Research and Technology – Hellas
Institute of Computer Science
P.O. Box 1385, GR-71110 Heraklion, Crete, Greece
http://www.ics.forth.gr/

University of Geneva
Centre Universitaire d'Informatique
Battelle Bat. A, 7 rte de Drize, CH-1227 Carouge
http://cui.unige.ch

Fraunhofer ICT Group
Anna-Louisa-Karsch-Str. 2
10178 Berlin, Germany
http://www.iuk.fraunhofer.de/

University of Southampton
University Road
Southampton SO17 1BJ, United Kingdom
http://www.southampton.ac.uk/

Institut National de Recherche en Informatique
et en Automatique
B.P. 105, F-78153 Le Chesnay, France
http://www.inria.fr/

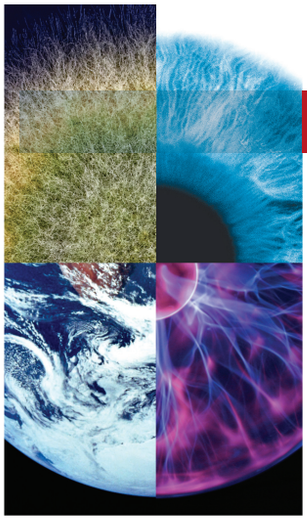
Universty of Warsaw
Faculty of Mathematics, Informatics and Mechanics
Banacha 2, 02-097 Warsaw, Poland
http://www.mimuw.edu.pl/

Norwegian University of Science and Technology
Faculty of Information Technology, Mathematics and Electrical Engineering, N 7491 Trondheim, Norway
http://www.ntnu.no/

Universty of Wroclaw
Institute of Computer Science
Joliot-Curie 15, 50–383 Wroclaw, Poland
http://www.ii.uni.wroc.pl/

I.S.I. – Industrial Systems Institute
Patras Science Park building
Platani, Patras, Greece, GR-26504
http://www.isi.gr/

Technical Research Centre of Finland
PO Box 1000
FIN-02044 VTT, Finland
http://www.vtt.fi/



Editors: **Elena Simperl,** University of Southampton, e.simperl@soton.ac.uk

The Web of Data: Bridging the Skills Gap

John Domingue and Mathieu d'Aquin, *The Open University*
Elena Simperl, *University of Southampton*
Alexander Mikroyannidis, *The Open University*

In 2011 the McKinsey Global Institute published a highly cited report on Big Data, which shows that "data have swept into every industry and business function and are now an important factor of production."[1] Taking only the figures for US health care, EU public sector administration, and the global consumer surplus from using personal location data, the report estimates an annual potential value of more than US$1 trillion. The report concludes, however, that by 2018, "there will be a shortage of talent necessary for organizations to take advantage of big data," estimating that there will be 140,000–190,000 deep analytical talent positions for which no trained personnel will be available (see Figure 1 here, based on Exhibit 4 in the report).

The report focuses on data in general, which naturally includes more specific forms of data sources and data management technologies such as Open and Linked Data, the latter both in its online realization as a freely accessible "Web of Data" and as data integration solutions applied to corporate datasets. Among the ways in which Big Data creates value, the report mentions *transparency*: "simply making big data more easily accessible to relevant stakeholders in a timely manner can create tremendous value."[1] The report further elaborates by identifying several issues that must be addressed to capture the full potential of Big Data, including access to data: "to enable transformative opportunities, companies will increasingly need to integrate information from multiple data sources." Similar arguments can be made for Open Data. For example, a related recent McKinsey report[2] estimates $3 trillion of value generated globally by Open Data from seven domains: education, healthcare, transport, consumer products, oil and gas, electricity, and consumer finance.

Linked Data and related semantic technologies directly address these challenges; while certainly not all of the Big Data and Open Data will be Linked Data, we can nevertheless have confidence that the Big Data and Open Data economies will have a large demand for Linked Data experts.

Linked Data has the potential to lower the barrier of entry into data-intensive industries by providing a lightweight solution, based on standard Web technologies and principles, to decentralized data publication, management, and use for the multitude of datasets that are openly available on the Web. Its potential for business intelligence and enterprise application integration has been reflected, for instance, in PriceWaterhouseCooper's Technology Forecast of Spring 2009: "During the next three to five years, we forecast a transformation of the enterprise data management function driven by explicit engagement with data semantics."[3] Projects and technology in a variety of domains, including news, media, finance, life sciences, and agriculture successfully showcase its benefits. However, all this potential will be lost if the critical mass of professionals proficient in Linked Data aren't present.

So how should we achieve this? In this article, we cover two aspects related to bridging the skills gap: first, what are the design principles that could support the design and delivery of Web of Data learning materials? And second, how can educational establishments lead the way in supporting professional training in this field?

Web of Data Training: Principles

Within the Euclid project, we've been developing Open Educational Resources (OERs) for Linked Data. Based on our experiences to date within the project and also the extensive previous experience by the partners in producing OERs and professional training in general, we've generated the following design principles. They're applicable to any other educational scenario with a strong knowledge

1541-1672/14/$31.00 © 2014 IEEE
Published by the IEEE Computer Society

transfer component and a focus on a technology area with a high demand for skilled practitioners. All of our materials are freely available and can be found at www.euclid-project.eu.

Curriculum Design Principles

Here are the principles we used for designing the project's curriculum.

- *Industrial relevance*—a curriculum must take into account the needs of industry related to managing large amounts of data. One possible mechanism to maintain relevance over time is to mine and analyze relevant job adverts to gain the desired competencies for various industrial sectors.
- *Team curriculum design*—producing the right curriculum relies on a number of heterogeneous factors, and a team approach is best able to fully capture industrial, academic, and pedagogical requirements.
- *External collaboration*—will aid in bringing in any missing expertise, provide an additional layer of quality assurance, and where necessary, can facilitate course delivery and dissemination.
- *Explicit learning goals*—to which all content (slides, webinars, and eBooks) are developed. Learners are guided through the learning goals through learning pathways—a sequence of learning resources to achieve a specific learning goal.
- *Use of realistic solutions*—rather than toy examples, this exposes learners to the data challenges faced by industrial sectors.
- *Use of real data*—similarly, a variety of issues only appear once someone is dealing with large datasets; for example, only automated solutions work with large data.
- *Use of real tools*—although potentially complex, such tools typically ingrain good practice, are more likely to be robust and maintained in the longer term, and will be similar to those used in industry.
- *Show scalable solutions*—based upon industrial-strength repositories and automatic translations—for example, the World Wide Web Consortium (W3C) standard R2RML for generating a Resource Description Framework (RDF) from large data contained in standard databases.
- *Eating your own dog food*—can provide extra exposure of the technology to learners and support a number of useful activities.

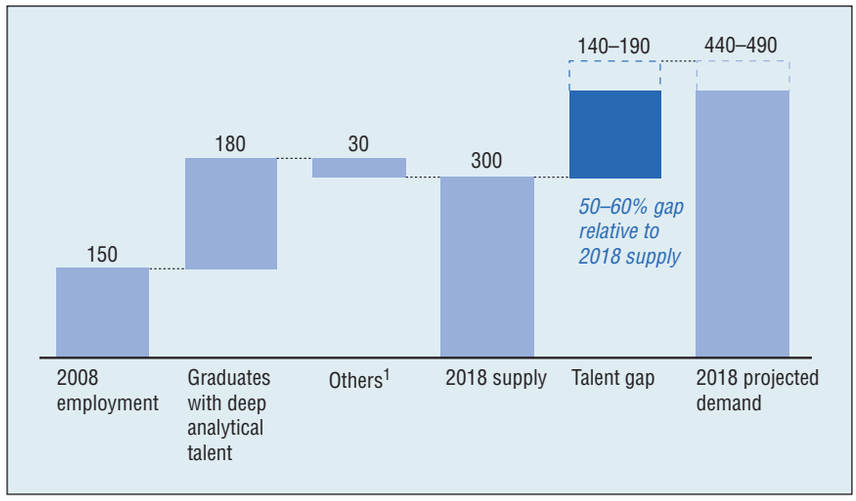


Figure 1. Demand for data analytics expertise according to the McKinsey Global Institute. It's expected that soon a shortage of talent will ensue.

Within the Euclid project, we monitor communication and engagement with the Linked Data community through W3C email lists, in the social network channels LinkedIn and Twitter as well as content dissemination channels such as Vimeo and SlideShare. We transform the monitoring results into RDF and make these available at a SPARQL endpoint. This can be thought of as a first step in using Linked Data to support Learning Analytics.

Curriculum Delivery Principles

We used the following as our guiding principles for delivering the curriculum.

- *Open to format*—content should be available in as many formats as feasible. Within Euclid, we've made our OERs available in a variety of formats, including HTML, and as an eBook (available as an Apple iBook and for Amazon Kindles).
- *Addressability*—every concept in the curriculum should be URI-identified, so that HTML and RDFa (RDF that can be embedded within HTML webpages) machine-readable content is available. This feature allows learners, educators, and the wider community to discuss specific points and to point interested parties to relevant content.
- *Integrated*—to ease navigation for learners, all media (including, for example, textual content, video clips, and interactive components) should be placed in one coherent space.
- *High quality*—having a formalized quality assurance process where all materials go through several iterations will help ensure that the learning experience is positive. For example, within Euclid we used a full dress rehearsal for each public webinar.
- *Self-testing and reflection*—learning materials should allow learners to test what they've learned against

learning goals, enabling the self-monitoring of progress.

These principles have enabled the Euclid consortium to create rich multimedia eBooks, which have proven popular in general and also in particular with the attendees of the European Semantic Web Conference (ESWC) Summer School that we run annually in southern Crete (see http://summerschool2014.eswc-conferences.org). Making these materials available beforehand enables the school attendees to pursue topics related to Open and Linked Data in greater depth.

All of these principles apply to the release of open, online educational materials related to Linked Data. Another aspect of interest, however, is how Linked Data can itself contribute to these principles, and to the education sector in general.

Eating Our Own Dog Food: How Educational Establishments Should Lead the Way

"Eating your own dog food" is obviously a valuable principle in education, whatever the topic being learned and taught. This is even more important when the principles and technologies being considered (Linked Data) can contribute to solving some of the fundamental issues experienced in teaching and learning the topic itself.

Indeed, educational institutions such as universities or specialist training programs like Euclid naturally produce large amounts of data that benefit from being widely shared. In the past, fields such as technology-enhanced learning and eLearning have partly tried to address this challenge, through creating metadata standards and repositories of (especially multimedia) course material.[4] More recently, the idea of open data in education has mostly been related with the explosion of OERs where course material is made available, generally on the Web, through an open license.[5] However, the field of education is now experiencing changes that go beyond merely putting course material on the Web for free:

- Traditional educational establishments are increasingly proposing online courses, which can be flexibly combined with other learning activities and aren't constrained by location.
- As exemplified by Euclid, alternative methods of delivering training outside of these establishments are emerging, including the development of MOOCs (Massive Open Online Courses) such as P2PU.org (peer-to-peer university), where communities of tutors and students "self-organize" around a particular topic.
- All of these trends are being combined with traditional universities partnering in creating MOOCs (see, for example, FutureLearn.com) and delivering open education, while open educational resources are used in traditional teaching. This combination is one of the key elements that's intended to provide more students with more options and more flexibility in their choice of a study path. It's also creating new challenges that can only be addressed with adequate Web-based data management methods, namely: How can we discover learning opportunities? How can we connect them within a common environment? How do we promote them so that they can be assessed, understood, and compared?

As some institutions related to education start realizing the potential implications of these questions, we see a number of initiatives emerging to apply linked data principles to education.

Educational Establishments and Initiatives Embracing Linked Data

One of the earliest examples of the use of Linked Open Data in education can be found, unsurprisingly, at the Open University in the UK. The Open University is a 40-year-old institution entirely dedicated to open and distance learning. It's also the largest university in the UK, with more than 250,000 students per year.

The amount of public information made available by such an organization is extensive, but what's even more striking is the way this information is siloed in different systems, with different channels of access to different parts for both the users and the developers of information systems inside the University. In 2010, we therefore created data.open.ac.uk to provide up-to-date linked data about the courses offered at the university, the educational material and multimedia resources available, the research outputs of its members, and many other aspects of its academic activities. This has led to many applications (both internal and external)[6] addressing, for this university, some of the aforementioned challenges.

Of course, this is only one example of a general trend involving many other universities in the UK (Southampton, Bristol, Oxford, and so on) and elsewhere (such as Aalto in Finland, Munster in Germany, and Aristotle University in Greece). Datasets of particular interest to education were also created out of specific projects (such as the Open Courseware Consortium metadata, SlideWiki, and the Terence reading comprehension dataset). An especially interesting initiative to mention is the one of LRMI.net (Learning Resource Metadata

Initiative, led by Creative Commons), which provides an extension of Schema.org to annotate learning resources on the Web.

Beyond Individual Initiatives: Towards a Web of Educational Data

Of course, while individual initiatives are important, the full potential of linked data resides in the possibility to connect and link them to obtain a global data environment, for an increasingly global education landscape. LinkedUp (see http://linkedup-project.eu) is a European project created to drive this move towards a Web of Educational Data. It's achieving this mainly in two ways: by collecting, putting together, and reconnecting existing data of relevance to educational applications, and by organizing a series of competitions to demonstrate and push forward the use of such Web data in real, concrete learning and teaching situations.

The LinkedUp data catalogue is based on both the *Comprehensive Knowledge Archive Network* (CKAN) and a linked data representation of the metadata and content of datasets that are explicitly identified as related to education. It serves both as a homogeneous access point to the data for application developers, and as an "observatory" to analyze the state, relatedness, and evolution of education within the Web of Data. We can indeed, through this catalogue, understand what type of information is interesting to institutions, what vocabularies are used to represent them, and how datasets relate to each other. Included within the catalogue is also a dedicated curation effort where disparate vocabularies are aligned with each other, so that heterogeneous datasets can be connected and queried jointly (see elsewhere[7] for an analysis of an early version of the catalogue and of the effect of alignments).

One of the biggest questions of linked data, which naturally emerges at this point, is: "Now that we have all these data, what do we do with them?" The main goal of LinkedUp is to answer this question by showing what concrete challenges can be solved when such data are available. This is achieved through the LinkedUp challenge (see http://linkedup-challenge.org): a series of competitions for developers to create new educational services, data management techniques or apps that exploit available Web data for education in an innovative way, and responding to a concrete need from learners and teachers.

Closing the loop, the product of such competitions (and generally of all these initiatives) is not only the applications that are created. It's a demonstration of what can be achieved with the technologies that directly relate to the learner's environment, as well as a way to draw more developers to consider the benefits of linked data for their applications, generating new learning opportunities (transformed into new resources) in the process.

As outlined in the latest McKinsey Global Institute Report,[2] we're now seeing the global economy beginning to operate in real time. We can gather data on consumers, cities, and markets, among others, instantaneously. Big Data analytics will significantly increase efficiency of both commercial and public services and open up new opportunities and markets. The total value generation for the impact of new data technologies will be measured in trillions of dollars globally.

However, the fulfilment of such opportunities requires that competent data science personnel are available, and current forecasts see a six-figure projected gap in the US alone. Here, we've outlined some lessons learned in relation to bridging this gap through the training of data scientists.

Within the Euclid project, which has been developing OERs for Linked Data, we generated a set of principles related to the design and delivery of data-centric curriculum. These principles have kept the project in good stead, with initial feedback on the released resources being very positive. The Euclid book is now freely available within the Apple Book Store (see www.euclid-project.eu for further details).

One of the principles we've outlined is "eating your own dog food," which the LinkedUp project addresses in a general sense, demonstrating how educational organizations can lead the way and also benefit from Linked Data. Along with the current excitement around the possibilities for MOOCs, a Web of educational data will offer a range of very interesting possibilities for educators and learners in the near-to-medium term. The possibilities for value creation are enormous—for example, it's estimated that Open Data will contribute up to $1.2 trillion in education alone. Exciting times are ahead. ⬒

References

1. McKinsey Global Institute Report, *Big Data: The Next Frontier for Innovation, Competition, and Productivity,* tech. report, May 2011; www.mckinsey.com/mgi/publications/big_data/index.asp.
2. McKinsey Global Institute Report, *Open Data: Unlocking Innovation and Performance with Liquid Information*, tech. report, Oct. 2013; www.mckinsey.com/insights/business_technology/open_data_unlocking_innovation_and_performance_with_liquid_information.
3. *Technology Forecast*, PriceWaterhouseCooper, tech. report, Spring 2009; www.pwc.com/en_US/us/

technology-forecast/assets/PwC-Tech-Forecast-Spring-2009.pdf.
4. M. McClelland, "Metadata Standards for Educational Resources," *Computer*, vol. 36, no. 11, 2003, pp. 107–109.
5. D.E. Atkins, J.S. Brown, and A.L. Hammond, *A Review of the Open Educational Resources (OER) Movement: Achievements, Challenges, and New Opportunities*, Creative Common, 2007.
6. M. d'Aquin, "Putting Linked Data to Use in a Large Higher-Education Organisation," *Proc. Interacting with Linked Data*, 2012; http://ceur-ws.org/Vol-913/01_ILD2012.pdf.
7. M. d'Aquin, A. Adamou, and S. Dietze, "Assessing the Educational Linked Data Landscape," *Proc. Web Science*, 2013; http://linkedu.eu/catalogue/publications/websci2013-rn.pdf.

John Domingue is the deputy director of the Knowledge Media Institute at The Open University. Contact him at john.domingue@open.ac.uk.

Mathieu d'Aquin is a research fellow at the Knowledge Media Institute at The Open University. Contact him at mathieu.daquin@open.ac.uk.

Elena Simperl is a senior lecturer at The University of Southampton. Contact her at e.simperl@soton.ac.uk.

Alexander Mikroyannidis is a postdoctoral researcher at the Knowledge Media Institute at The Open University. Contact him at a.mikroyannidis@computer.org.