GRIAN - Quality Assessment Framework for Linked Data on the Web

(a proposal for SW project)

Supervisors: Tomáš Knap (tomas.knap@mff.cuni.cz, ICQ# 315102977)
Irena Mlýnková
Department of Software Engineering

Number of team members: 4 - 5
Language: arbitrary (Java is preferred)
OS: arbitrary

Motivation:
The term *Linked Data* [2] refers to a set of best practices for exposing, sharing, and connecting structured data on the Web; it was introduced in 2006 by Tim Berners-Lee, inventor of the Web, who outlined several basic principles for linking data on the Web [9]. Due to key technologies that support Linked Data - URIs (a unique identification of pieces of data), HTTP (a simple and universal mechanism for retrieving data), and RDF (a generic data model for structuring and linking data together) – the concept of Linked Data enables in a large scale the idea of Semantic Web as a web of interlinked data understandable by humans and machines.

A significant effort in adoption and application of the Linked Data principles has been done in the Linked Open Data W3C SWEO Community Project (LOD) [1], supported by the W3C Semantic Web Education and Outreach Group [4]. The original and ongoing aim of the project is to extract data available under open licenses from wide variety of data sources (relational databases, XML native stores, RDF silos, XHTML pages, RSS feeds etc.), link them together, and publish them on the Web. In the early stage of the Project LOD, project participants were primary university researchers and small companies; however, very soon, companies like the BBC, Thomson Reuters and the Library of Congress revealed the power of Linked Data and joined this effort. In 2009, the government of the United States [5] and of the United Kingdom [6] realized the importance of publishing government data on the Web using open standards and committed towards this direction.

The advent of Linked Data in the recent years accelerates the evolution of the Web into a giant information space where the unprecedented volume of resources will offer to the information consumer a level of information integration and aggregation that has up to now not been possible. Consumers can now 'mashup' and readily integrate information for use in a myriad of alternative end uses. Indiscriminate addition of information can, however, come with inherent problems such as (1) the provision of poor quality, (2) inaccurate, (3) irrelevant, or (4) fraudulent information. All will come with an associate cost which will ultimately affect decision making, system usage and uptake.

The ability to assess the quality of information on the Web, thus, presents one of the most important aspects of the information integration on the Web and will play a fundamental role in the continued adoption of Linked Data principles [10].

Goals of the project:
The goal of the project is to design and implement Grian – a quality assessment framework for Linked Data, which will involve the quality assessment (QA) process helping the information consumer to discover and avoid the problems (1 – 4) above.

The QA process should be based on the set of analyses (such as an analysis of data provenance, analysis of resources' reputation) and collection of policies (e.g. “Prioritize the most up-to-date data resources“, “Trust resources hosted by http://nyse.com“) customizable according to the information consumer's needs. Grian should filter, rank, and sort data resources (data) according to the quality values assessed to them in the QA process and, thus, help the information consumer to pick the right resource from a myriad of resources.

From an information consumer's perspective, the basic workflow of the QA process should work as follows. Firstly, the consumer defines his QA profile holding the set of QA analyses and QA policies participating in the QA process and, then, queries the Web. As a result, the set of resources relevant to the query is returned, based on the results of the conventional search engines, such as Google, or semantic search engines, such as Sindice [11]. In this moment, the QA process is entered - the QA analyses specified in the consumer's QA profile are fetched, launched, and the QA policies belonging to these QA analyses are tried to be applied to the resources. The successfully applied QA policies are reflected in the QA score and QA color of the processed resource. The QA score is the main indicator of the quality of the resource; the QA color of the resource serves rather as a first signal to the information consumer whether the resource is recommended (green color), recommended with warnings (yellow), unknown, in the sense that no QA policies were successfully applied to this resource (grey), or not recommended at all (red). Afterwards, the resources are ranked according to their QA score and returned to the information consumer together with the information about their QA colors, the sets of QA policies successfully applied to them, and the explanations of the steps done by the QA process.

The main components of Grian are:
–a component executing the QA process enabling to: ***(1-2 team members)***
–incorporate QA analyses to the QA process according to the consumer's needs
–apply QA policies of the selected QA analyses to the incoming resources and compute QA scores and QA colors of the resoruces
–audit the steps of the QA process and provide explanations to the information consumers why the particular QA score and QA color was associated with the particular resource
–a component ranking the resources and visualizing the QA scores and QA colors of the resources assessed during the QA process; the component can be either a plugin to Firefox or a standalone application ***(1 team member)***
–a component enabling information consumers to define and maintain their QA profiles; the component should have a user-friendly frontend ***(1 team member)***

Since the provenance or lineage of information provides the information consumer the necessary contextualization of the provided information [7, 8, 3] and, thus, represents the cornerstone element which will enable information consumers to assess the quality of the information under their perspective, the Provenance analysis usable in the QA process should be created as part of the project. Such Provenance analysis must obey the requirements on the interfaces of the QA analyses, so that it is pluggable as a QA analysis to the main component executing the QA process ***(1 team member).*** Further QA analyses will be created by the information consumers themselves.

Requirements on the team members:
Some experience with RDF and Semantic Web technologies is preferred, however, the basic necessary knowledge can be obtained before the start of the project in 1-2 weeks.

Other requirements on the project:

–Grian will be publicly available
–The documentation of the project will be in English

Notes:
The students working on this project will get a medium to strong knowledge of Linked Data and Semantic Web technologies (depending on the position in the team). The problems solved as part of the SW project can be extended in master theses, mainly in the area of provenance, trust, data quality, and ranking of data on the Web.

During the works on the project, we will stay in touch with Digital Enterprise Research Institute (DERI) in Ireland [13], one of the biggest Semantic Web Research institutes in the world with a specialized Linked Data Research Centre [14]. Tens of projects regarding Linked Data are currently running there, including development and maintenance of the Semantic web search engine Sindice [11] and the Semantic information mashup application Sig.ma [12].

References:
[1] Linking Open Data W3C Community Project.
http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

[2] Bizer, CH., Heath, T. and Berners-Lee, T. Linked Data – The Story So Far. To appear in Special Issue on Linked Data, International Journal on Semantic Web and Information Systems, 2009

[3] A. Freitas, T. Knap, S. O'Riain, and E. Curry. W3P: Building an OPM based Provenance Model for the Web. Submitted to Future Generation Computer Systems, The International Journal of Grid Computing and eScience. (http://www.ksi.m.cuni.cz/~knap/qa/prov10.pdf)

[4] Semantic Web Education and Outreach Group. http://www.w3.org/2001/sw/sweo/

[5] http://www.whitehouse.gov/the_press_office/Transparency_and_Open_Government/

[6] http://blogs.cabinetoffice.gov.uk/digitalengagement/post/2009/06/09/Data-So-what-happens-now.aspx

[7] L. Moreau. The Foundations for Provenance on the Web. Foundations and Trends in Web Science, November 2009.

[8] Hartig, O. Provenance Information in the Web of Data. In Proceedings of the Linked Data on the Web (LDOW) Workshop at the World Wide Web Conference (WWW), Madrid, Spain, 2009

[9] Berners-Lee, T. Linked Data - Design Issues.
http://www.w3.org/DesignIssues/LinkedData.html [2] DBLP (http://www.informatik.uni-trier.de/~ley/db/)

[10] F. Naumann and C. Rolker. Assessment Methods for Information Quality Criteria. In Proceedings of the International Conference on Information Quality, 2000.

[11] Tummarello, G. et al. Sindice, The Semantic Web Search Engine. http://sindice.com/

[12] Catasta, M. et al. Sig.Ma, Semantic Information Mashup. http://sig.ma/

[13] Digital Enterprise Research Institute, National University of Ireland, Galway. http://www.deri.ie/

[14] Linked Data Research Centre, DERI. http://linkeddata.deri.ie/