Linking Data for Industrial Knowledge Management—A Case Study

Katariina Nyberg, Matias Frosterus, and Eero Hyvönen

Semantic Computing Research Group (SeCo)
Aalto University, Dept. of Media Technology, and
University of Helsinki, Dept. of Computer Science
http://www.seco.tkk.fi/
firstname.lastname@tkk.fi

Abstract. Manufacturing companies face the challenge of maintaining documentation and knowledge about their projects and products, scattered in heterogenous, distributed databases, represented in different formats and languages, and hosted in mutually incompatible systems. At the same time, the knowledge needs to be accessed on a global level from different perspectives and user groups, such as project planners, designers, and maintenance personnel. This paper presents a case study, based on real datasets of a major international diesel engine and power plant manufacturer, where these problems are addressed simultaneuosly by harmonizing the datasets from different sources using RDF, and by linking them together into a global repository using shared resources. Based on the global RDF store, services for both human and machine users, such as a faceted search engine and a SPARQL end-point, can be provided to support access from different perspectives to the company knowledge base.

1 Introduction

The Semantic Web[1] and Linked Data [3] provide an RDF-based[2] framework for global linking of data on the web, using heterogeneous datasets produced by independent actors in a distributed environment. On a company and an intranet scale, the semantic web provides a solution approach to problems of managing scattered, hard-to-find heterogenoues data, too. Using the methods and tools of the semantic web one can provide structure and meaning for the data, and facilitate the creation of intelligent user-interfaces and visualizations for it. The linked data principles allow also for integration between the company data and external data stores, such as the Linked Open Data cloud[3].

In the following, we first show a general publication pipeline, through which heterogeneous data originating from different datasets can be harmonized using the RDF data model, validated and corrected in a metadata editor if needed, and published instantly as a faceted semantic portal, with interfaces for both human users and the machine. After this, application of the pipeline to the contents of large international diesel engine and power plant manufacturer is described. In conclusion, contributions of the work are summarized, some related work discussed, and future research proposed.

[1] http://www.w3.org/standards/semanticweb/
[2] http://www.w3.org/RDF/
[3] http://linkeddata.org

2 Publication Process

An overview of the process of utilizing the Semantic Web and linked data approach in an industrial knowledge management context is shown in Figure 1. On the left, there are different kinds of datasets from different parts of the organization in different formats and for different needs. The first step is to transform the heterogeneous data into RDF form, and give URIs to the different resources, whose descriptions involve properties. Some resources, such as projects, shipments, and installations, often already have some form of ID numbers to be used as a basis for the local names in URIs. However, ID formats can differ and uniqueness of ID numbers between different datasets is not necessarily guaranteed. Furthermore, resources such as employees and buildings, may lack any unambigous ID numbers.

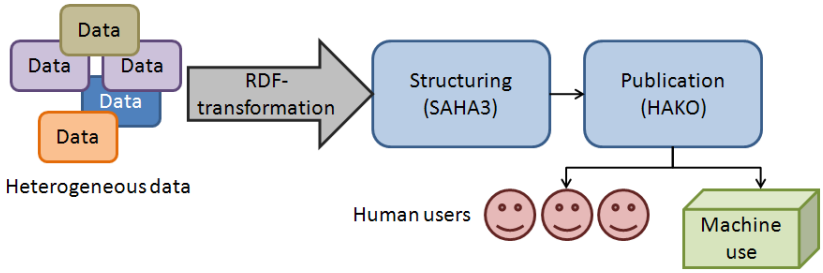

Fig. 1. Overview of the process of utilizing the linked data paradigm in an industrial knowledge management context

Once the data of each dataset is in RDF format, they can be merged together into a coherent whole, which in the case of RDF means technically simply taking the union of the datasets. However, when linking data in this way, two semantic data alignment problems must be solved before this is feasible. First, the schemas used in different datasets must be aligned (e.g. Dublin Core and in-house metadata schemas). Second, the vocabularies used in filling out the metadata schema values (e.g., persons, motor types, locations, etc.) must be aligned. In the linked data paradigm, standard properties such as owl:sameAs, rdfs:subclassOf, rdfs:subpropertyOf, and skos:narrowMatch are typically used for this.

In our process model, the aligned and merged data is imported into the SAHA 3 [8] metadata editor, which can be used to validate the data and make (manual) corrections to it, if needed.

For the publication of the data we used HAKO [8], a faceted search engine generator for publishing a SAHA3 project as a readily usable, faceted portal with machine usable APIs. The RDF data in SAHA3 is instantly available in HAKO, which is then configured to produce a portal matching the needs of the end-user in a few seconds. The publisher simply specifies 1) the classes whose instances are to be searched, and 2) what properties form the search facets for these instances. An example of the faceted

search portal can be seen in Figure 3. The facets are on the left and the search results, correspoding to the current selection of facet categories, on the right. There is also the possibility of using traditional, text-based search.

For machine use, SAHA3 and HAKO have two machine APIs: one for using the content as an ONKI ontology service [11] for annotation work, and one for using the content via a SPARQL end-point, used by other applications and HAKO itself.

3 Case Study: Making Data Available for Humans and Machines

Our case study concerns knowledge management of project ja product documentation in a major international manufacturer that produces diesel engines and power plant solutions for its customers. The idea is to apply the publicaton process of the previous section to large amounts of heterogenous company datasets. This case study aims at making it easier for the company's employees with different responsibilities and perspectives, such as project management or plant maintenance, to find and browse through the data in multiple ways and gain efficient access to the desired information.

There are three aspects to consider, when making data available in RDF form: 1) the data itself, 2) the metadata (including schemas), and 3) shared vocabularies for the domain of the (meta)data, i.e. ontologies [4]. These three aspects are considered in the following sections 3.1– 3.3.

3.1 Converting the Datasets into RDF

We used the company's Enterprise Resource Planning (ERP) system for obtaining the following datasets containing information about the power plant stations.

Plant Operation Manuals This data was in the form of numerous XML files containing parts of power plant manuals, such as plant operation manuals. They were grouped according to different power plant projects. The schemas for these XML files were splintered into numerous different DTD files, which made it difficult to find a consistent way to parse the files into other formats. Most of the XML files contained text fragments of the manuals for plant operations. The materials were in zip files that were named with a global unique identifier (GUID) corresponding to the project that the XML files described. We were also provided with the actual plant manuals for each plant project in PDF form. By analyzing the PDF files, we were able to identify the corresponding manual's page number for each XML file.

Power Plant Project Information This dataset was a large spreadsheet file that contained information about the personnel in plant projects, i.e. the people that had worked in a given project as project managers, controllers, or engineers. In addition, the data contained technical information about the power plant projects, such as diesel engines used and fuel types needed. The data also contained information about the country of registration of the plant and important project dates.

The dataset was converted into RDF form in such a way, that each row in the spreadsheet was turned into a resource representing a plant project, and each column corresponded to a property of that resource. If the information in a column represented an entity, such as a person or a place, then the column content was turned into a resource, and the resulting property was an object property of the resources corresponding to the rows. If the information in the column was a numerical value, such as the engine quantity or a date, then the column contents were represented as a literal property.

Block Diagrams of Plant Systems A PDF file containing 14 different block diagrams was received from the company. These diagrams are an abstract and general representation of the inner workings of a plant. Each diagram represents a subsystem of the plant, such as the lubrication oil system or the steam generation system. The blocks in a diagram are connected to each other, labeled, and most of them are marked with a three-lettered code. Figure 2 shows as an example an extract from a block diagram that describes the lube oil system. The information contained in the diagrams, such as the different connections between individual blocks, are intended for visual use by human readers, and therefore an RDF representation of them was created by hand using the SAHA 3 editor.

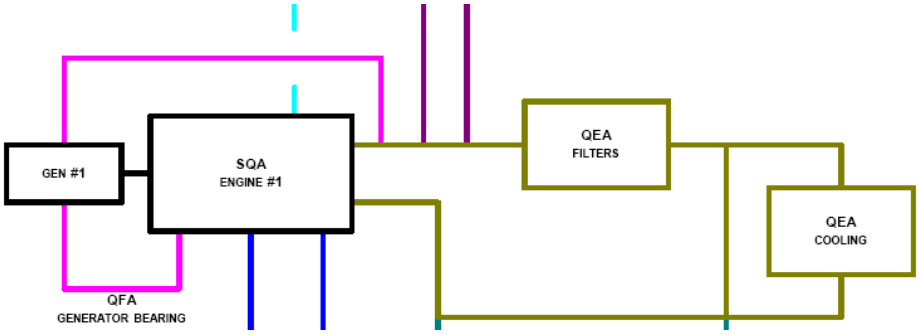

Fig. 2. A part of one of the block diagrams. The concepts in this block diagram detail the lube oil system in a power plant.

Other Data The links between the XML files and the projects were established by yet another spreadsheet file. Each row contained the project ID used in the power plant project information spreadsheet file (described above), and the project GUID that identified the respective XML files.

3.2 Creating Metadata Schemas

We created an RDF metadata schema for representing the XML files, where an RDF instance of a document class was created for each XML file. The properties of the

document class included the text contents of the XML file, its file name, and the power plant project it belonged to. In addition to this, a link to the manual and the respective page number were recorded as literal properties of the resource.

The schema for describing the contents of the power plant project information spreadsheet file followed mostly its structure. Some of the columns were turned into literal properties and others into object properties. In the case of an object property column, a resource of the type corresponding to the column was created, and the column content was represented as its label.

A schema for representing the 14 block diagrams was created, too. Here the individual blocks are represented as resources corresponding to the underlying ontological concepts. Information about a block, such as the three letter code, its label, and its connection to other blocks, were represented by the properties of the resource.

3.3 Shared Vocabularies

The ontologies that described people, places and mechanical information on power plants, such as fuel type, were created and populated from the spreadsheet file, when transforming it into RDF form.

A power plant system ontology was manually created based on the concepts presented in the block diagrams. The relations between the concepts express a consequential and other connections. For example, in Figure 2 the concept of the filters (QEA) is connected to the concept of the engine (SQA), indicating that a text passage that mentioning the filters might possibly also more generally apply to the concept of the engine without expressing it explicitly. There is part of relation from each concept to the block diagram (subsystem) it belongs to.

The manual text fragments corresponding to the XML files were analyzed, the concepts mentioned in the block diagrams were extracted from the text, and the links between the XML file resources and the concepts were established. Since the XML files belonged to a certain plant project, the projects and the concepts were linked, too.

3.4 Making the Data Searchable

In its original form, the underlying data is sorted and searched for according to a hierarchical classification of the individual power plant projects. As a result, it is not easily accessible to human users with complex access needs, such as, searching for a project documentation based on the personnel involvement. Browsing through a spreadsheet file with a lot of columns, makes it hard for a user to see the possible connections that exist between the projects. By transforming the data dumps into RDF form and making them compatible with each other, we managed to create a network of data, which allows for the data's information to be used to its full potential. Because of this, complex and perchance unexpected connections in the data can be found.

The idea was to create a system that would make it easy to answer questions, such as: "Which power plants are in the execution state of their life cycle and use dual fuel for power production?" or "Do I know anyone who has worked with John Smith on a power plant?" For this purpose, we published all of the resulting RDF in SAHA3. It was

easy to view the results in it and see if all the connections between different data were done correctly. The same RDF could be used with HAKO for configuring a faceted search for the data.

We configured HAKO in such a way, that the plant projects were the instances of the search. There was a total of 93 projects shown. As facets for the plant projects we chose all the relevant properties of the plant projects. Figure 3 shows the HAKO portal and how the amount of projects is narrowed down to ten according to a plant's fuel type and state. It gives an answer to the first of the above mentioned questions. To answer the second question, a user could clear the search by removing the selections, and narrow it by clicking on the name "John Smith" in one of the property lists that mention persons.

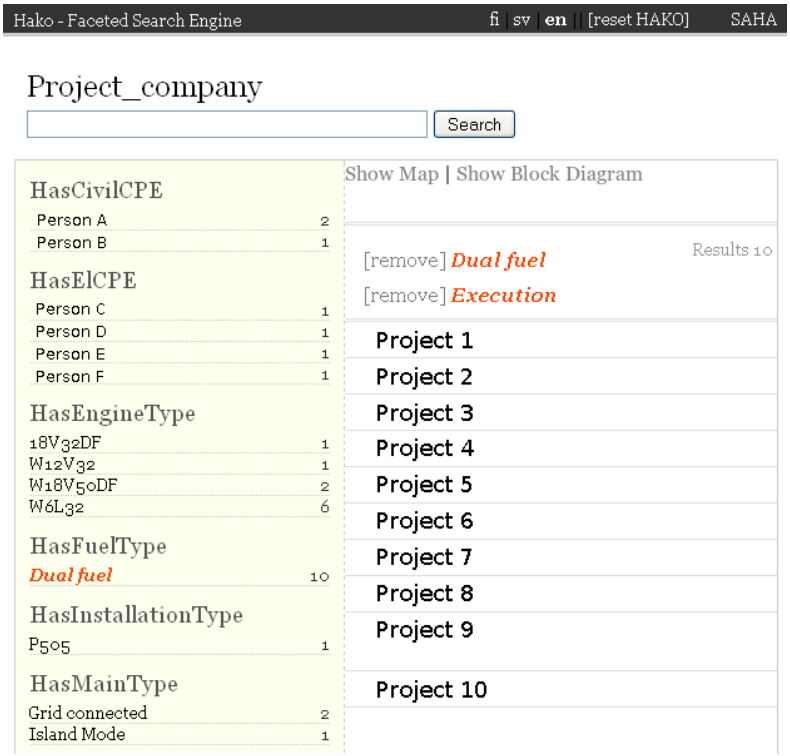

Fig. 3. HAKO faceted search portal (note that the project and person names have been edited out)

3.5 Block Diagram and Map Facets

Aside from the text-based facets already provided by HAKO, the nature of the block diagrams lends itself intuitively to a graphical, block based facet interface. This allows for easy access to the relevant manuals for maintenance and engineering personnel based

on the actual structure of a certain power plant system, such as the lube oil system in Figure 2. The block diagrams being general means that no customization is needed for different installations regardless of the specific decisions made in the construction.

The graphical block diagram facet was built into HAKO and placed above the results view into a tab system, which allows for easy integration of multiple graphical facets, as needed by a given system. We also implemented a map view, which is a worthwhile graphical facet for an international company that has projects in different parts of the world. It allows for an easy way of restricting the search results to arbitrary geographical areas and is useful in gaining an overview of a certain region's projects, especially when combined with the status information. "Show Map" and "Show Block Diagram" in Figure 3 open the map and block diagram view respectively when they are clicked on.

4 Discussion and Related Work

Contributions This paper presented a Semantic Web and Linked Data -based model and tools for publishing project and product documentation in an industrial company. From a human viewpoint, the idea is to aggregate and link related heterogeneous datasets, and make the whole data cloud more easily accessible from different perspectives using faceted search and browsing. At the same time, the content can be published for other services to use based on Linked Data principles, i.e. as a SPARQL endpoint, as a dereferenceable URI service, or as an RDF dump.

To evaluate the approach, a case study with real data from a manufacturing company was carried out with promising first results. The data dumps provided by the company contained information about persons and their roles in the projects. This is useful information for knowledge management, when special know-how needs to be found inside the company. Using HAKO it would be easy for the management of the company to see who has worked with whom and in what projects, and what experience they therefore now have. In a similar vein, also other additional perpectives to the project and product documentation datasets can be provided, using a single RDF-based knowledge repository. At the moment, end-user tests are being planned by the company in order to test the usefulness of the case study system in practice.

Related Work Kobilarov et al. [7] describe how the heterogeneous data from various sources in the BBC was made accessible using the tools of the Semantic Web, and linking to datasets in the LOD cloud, such as the DBPedia. They argued that interlinking data is beneficial to the users of the company's web page and the company itself at large. Antezena et al. [2] summarize different ways for supporting knowledge management in biology, discuss the emerging role of the Semantic Web, and introduce projects for knowledge management, such as BioGateway. It intergrates diverse datasets and offers a graphical interface and a SPARQL endpoint for its users [1].

Pollit [10] argues that the knowledge structures, on which the search target depends on, provide the facets with which the search can be narrowed down. This approach has been adapted for semantic faceted search, where the key design criteria have been to create a search interface for arbitrary RDF data [6, 5, 9].

Acknowledgements This work is part of the National Semantic Web Ontology project in Finland[4] FinnONTO (2003–2012), funded currently by the National Technology and Innovation Agency (Tekes) and a consortium of 35 public organizations and companies.

References

1. Erick Antezana, Ward Blondé, et al. Structuring the life sciences resourceome for semantic systems biology: lessons from the biogateway project. In A. Burger, A. Paschke, P. Romano, and A. Spendiani, editors, *Semantic Web Applications and Tools for Life Sciences, SWAT4LS 2008*, volume 435. CEUR Workshop Proceedings, http://CEUR-WS.org, 2008.
2. Erick Antezana, Martin Kuiper, and Vladimir Mironov. Biological knowledge management: the emerging role of the semantic web technologies. *Briefings in Bioinformatics*, 10(4):392–407, 2009.
3. Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data—the story so far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 2009.
4. Tom Gruber. Ontology. In Ling Liu and M. Tamer Özsu, editors, *Encyclopedia of Database Systems*. Springer–Verlag, 2009.
5. Michiel Hildebrand, Jacco van Ossenbruggen, and Lynda Hardman. /facet: A browser for heterogeneous semantic web repositories. In Isabel Cruz, Stefan Decker, Dean Allemang, Chris Preist, Daniel Schwabe, Peter Mika, Mike Uschold, and Lora Aroyo, editors, *The Semantic Web - ISWC 2006*, volume 4273 of *Lecture Notes in Computer Science*, pages 272–285. Springer–Verlag, 2006.
6. E. Hyvönen, S. Saarela, and K. Viljanen. Application of ontology-based techniques to view-based semantic search and browsing. In *Proceedings of the First European Semantic Web Symposium*. Springer–Verlag, 2004.
7. Georgi Kobilarov, Tom Scott, et al. Media meets semantic web — how the BBC uses DBpedia and linked data to make connections. In *Proceedings of the 6th European Semantic Web Conference (ESWC 2009)*, pages 723–737. Springer–Verlag, 2009.
8. Jussi Kurki and Eero Hyvönen. Collaborative metadata editor integrated with ontology services and faceted portals. In *Workshop on Ontology Repositories and Editors for the Semantic Web (ORES 2010), the Extended Semantic Web Conference ESWC 2010, Heraklion, Greece*. CEUR Workshop Proceedings, http://CEUR-WS.org, 2010.
9. Eyal Oren, Renaud Delbru, and Stefan Decker. Extending faceted navigation for RDF data. In Isabel Cruz, Stefan Decker, Dean Allemang, Chris Preist, Daniel Schwabe, Peter Mika, Mike Uschold, and Lora Aroyo, editors, *The Semantic Web - ISWC 2006*, volume 4273 of *Lecture Notes in Computer Science*, pages 559–572. Springer–Verlag, 2006.
10. A Steven Pollit. The key role of classification and indexing in view-based searching. Technical report, University of Huddersfield, UK, 1998. http://www.ifla.org/IV/ifla63/63polst.pdf.
11. Jouni Tuominen, Matias Frosterus, Kim Viljanen, and Eero Hyvönen. ONKI SKOS server for publishing and utilizing SKOS vocabularies and ontologies as services. In *Proceedings of the 6th European Semantic Web Conference (ESWC 2009)*, 2009. Springer–Verlag.

[4] http://www.seco.tkk.fi/projects/finnonto/