Towards Linked Data based Enterprise Information Integration

Philipp Frischmuth[1], Sören Auer[1], Sebastian Tramp[1], Jörg Unbehauen[1], Kai Holzweißig[2] and Carl-Martin Marquardt[2]

[1] Universität Leipzig, Institut für Informatik, AKSW,
{lastname}@informatik.uni-leipzig.de
[2] Enterprise Services Portal, CMS & Search, Daimler AG, Plant 096-0191, 70546 Stuttgart, Germany
{firstname.lastname}@daimler.com

Abstract. Data integration in large enterprises is a crucial but at the same time costly, long lasting and challenging problem. In the last decade, the prevalent data integration approaches were primarily based on XML, Web Services and Service Oriented Architectures (SOA). We argue that classic SOA architectures may be well-suited for transaction processing, however more efficient technologies can be employed for enterprise data integration. In particular, the use of the Linked Data paradigm appears to be a very promising approach. In this article we explore challenges large enterprises are still facing with regard to data integration. We discuss Linked Data approaches in these areas and present some examples of successful applications of the Linked Data principles in that context.

1 Introduction

Data integration in large enterprises is a crucial but at the same time costly, long lasting and challenging problem. While business-critical information is often already gathered in integrated information systems such as enterprise resource planning (ERP), customer relationship management (CRM) and supply chain management (SCM) systems, the integration of these systems itself as well as the integration with the abundance of other information sources is still a major challenge. In the last decade, the prevalent data integration approaches were primarily based on the Extensible Markup Language (XML), Web Services and Service Oriented Architectures (SOA) [6]. However, we become increasingly aware that these technologies are not sufficient to ultimately solve the data integration challenge in large enterprises. We argue that classic SOA architectures are well-suited for transaction processing, but more efficient technologies are available that can be deployed for solving the data integration challenge. With the use of the Linked Data paradigm for integrating enterprise information, data intranets can complement the intranets and SOA landscapes currently found in large enterprises.

In this paper, we explore the challenges large enterprises are still facing with regard to data integration. These include, but are not limited to, the development, management and interlinking of enterprise taxonomies, domain databases, wikis and other enterprise information sources. We discuss Linked Data approaches in these areas and present some examples of successful applications of the Linked Data principles in that context.

2 Data Integration Challenges in the Enterprise

We identified crucial areas where data integration challenges arise in large enterprises. In the following section we investigate those challenges, each by considering the current situation first. We then examine the benefits of employing Linked Data technologies in order to tackle the respective challenge. Finally, we describe the challenges that need to be addressed to make the transition from the current state of the art to the Linked Data approach feasible.

2.1 Enterprise Taxonomies

Nowadays, almost every large enterprise uses taxonomies to provide a shared linguistic model. It is widely agreed that taxonomies are usable, however, there are multiple challenges that must be addressed in order for taxonomies to work correctly [5]. A problem that arises is that different metadata creators use different terminologies and therefore the same object may receive different metadata descriptions by different people [4]. Another challenge is that large taxonomies require certain time for the users to get their bearings so that they can start to use the taxonomies correctly and avoid creating duplicities and other errors. In [15], the author discusses whether taxonomies are really necessary and stresses the importance of establishing relations to documents via URLs, which indicates already a clear shift towards the Linked Data vision.

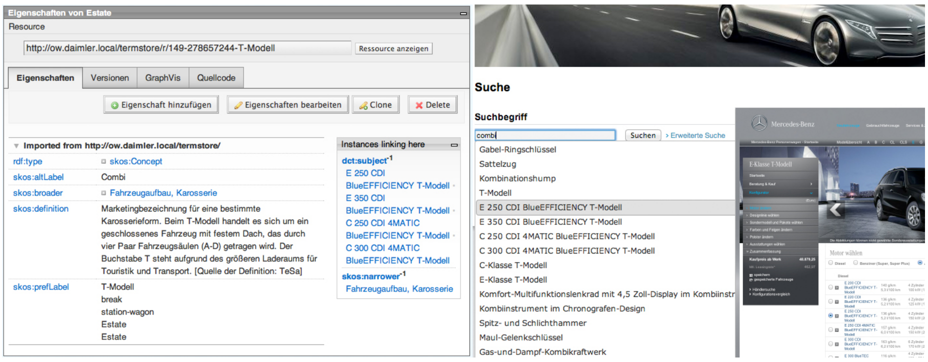
If we take a look at commercial implementations, there is Microsoft SharePoint's[3] Term Store (also referred to as Managed Metadata), which enables enterprises using SharePoint to tag objects stored in SharePoint with terms from a taxonomy. However, there are some strong limitations to this approach. There is very restricted multilingual support – separate SharePoint language packs need to be installed for each language to be used in the taxonomy. Also, the implementation is proprietary, thus hindering the integration with taxonomies or data outside of SharePoint.

Linked Data Approach We propose to represent enterprise taxonomies in RDF employing the standardized and widely used SKOS [12] vocabulary as well as publishing term definitions via the Linked Data principles. This approach entails the following main benefits: (1) Since terms are attached to URIs, which can be dereferenced using HTTP, term definitions can be obtained without the need for additional software (a browser is sufficient). (2) For the same reason, term creation and management can be realized in a distributed scenario, where, for example, certain departments are responsible for different parts of the taxonomy. Terms can then be interlinked and re-used regardless of department boundaries. (3) By employing the SKOS vocabulary, terms can have a hierarchical order and thus the problem of different metadata granularity can be easily solved. (4) Also, since data is represented using RDF, which works with arbitrary vocabularies and properties, overlapping, coinciding or conflicting term definitions (e.g. by different departments) can be interlinked by explicitly stating the relationship between terms via links. (5) Terms can be assigned multiple labels, which are represented as RDF literals. As such they can be furnished with a language tag resulting in

[3] http://sharepoint.microsoft.com/

multilingual taxonomies with very little additional effort. (6) Ultimately the result of employing the Linked Data approach for enterprise taxonomies is, that terms can be easily re-used in other scenarios as the originally intended ones. For example, a detailed term definition (with alternative labels in multiple languages) can be very useful for search applications, where terms are associated with documents which then become reachable via a variety of keywords.

However, the initial costs for building a basic taxonomy may still be quite high. The problem of finding initial terms (and their definitions) can be solved by integrating a service like DBpedia Spotlight [11]. With this service texts can be annotated with resources from DBpedia [1], which already contain a detailed description in multiple languages in many cases. For enterprise-specific terms, where a description is not available via DBpedia, a keyword extraction service like FOX[4] can be used instead. Thus an initial set of term URIs can be gathered, which can then be annotated manually with for example a data wiki like OntoWiki (see subsection 2.2).



http://ow.daimler.local/termstore/r/149-278657244-T-Modell

Imported from http://ow.daimler.local/termstore/

rdf:type	skos:Concept
skos:altLabel	Combi
skos:broader	Fahrzeugaufbau, Karosserie
skos:definition	Marketingbezeichnung für eine bestimmte Karosserieform. Beim T-Modell handelt es sich um ein geschlossenes Fahrzeug mit festem Dach, das durch vier Paar Fahrzeugsäulen (A-D) getragen wird. Der Buchstabe T steht aufgrund des größeren Laderaums für Touristik und Transport. [Quelle der Definition: TeSa]
skos:prefLabel	T-Modell break station-wagon Estate Estate

Instances linking here

dct:subject^{-1}
E 250 CDI BlueEFFICIENCY T-Modell
E 350 CDI BlueEFFICIENCY T-Modell
C 250 CDI 4MATIC BlueEFFICIENCY T-Modell
C 300 CDI 4MATIC BlueEFFICIENCY T-Modell

skos:narrower^{-1}
Fahrzeugaufbau, Karosserie

combi

Gabel-Ringschlüssel
Sattelzug
Kombinationshump
T-Modell
E 250 CDI BlueEFFICIENCY T-Modell
E 350 CDI BlueEFFICIENCY T-Modell
C 250 CDI 4MATIC BlueEFFICIENCY T-Modell
C 300 CDI 4MATIC BlueEFFICIENCY T-Modell
C-Klasse T-Modell
E-Klasse T-Modell
Komfort-Multifunktionslenkrad mit 4,5 Zoll-Display im Kombiinstr
Kombiinstrument im Chronografen-Design
Spitz- und Schlichthammer
Maul-Gelenkschlüssel
Gas-und-Dampf-Kombikraftwerk

Fig. 1. The left side shows OntoWiki, which displays a term definition and resources linking to it. The right side shows a search application, which employs the term metadata for finding and suggesting relevant content.

Figure 1 shows two screenshots, which demonstrate some of the advantages described in this section. The left side shows OntoWiki, which displays the definition of the term *T-Modell* along with some additional information. The location bar on the top of the screen displays the URI used for this very concept, which other resources can link to. It is also possible to directly de-reference this identifier and obtain the description for this resource in a machine-readable format. A dedicated list shows other resources that link to this concept, in this case certain car models. This circumstance is used in a search application, which is shown on the right side of Figure 1. When a user types the keyword *combi*, the knowledge base is used to obtain the fact, that this search term is a synonym for the concept *T-Modell*. Once this is done, all linked car models are retrieved

[4] `http://aksw.org/Projects/FOX`

and shown to the user. The depicted scenario is a good example of an application of a taxonomy outside the scope of the originally intended use. One main reason for the efficient realization of this application is that data from multiple sources (term store and car configuration metadata) was made available via the Linked Data principles.

Challenges Currently terminology in large enterprises is managed in a centralized manner mostly by a dedicated and independently acting department, which is in charge to standardize all corporate terms. As a result they create a variety of dictionary files for different scopes that are not interconnected. An employee that wants to look up a certain term, needs to know which dictionary to use in that very context, as well as where to retrieve the currently approved version of it. The main challenge in the area of enterprise taxonomies is defragmentation of term definitions without centralization of taxonomy management.

Defragmentation. The proposed Linked Data approach can be implemented by keeping the centralized structure and assign the corporate language management (CLM) department the task to create a knowledge base that contains the complete terminology of the company. This solves the fragmentation problem occurring with the dictionary approach, but it also keeps the barrier for participation high, since a single department still is in charge for maintaining the knowledge base.

Decentralization. On the other hand an entire decentralized solution can be implemented, by assigning each department in the enterprise it's own taxonomy namespace. Adding new terms or refactoring existing terms becomes very easy with this approach, due to the reduced communication overhead with other departments. Nevertheless the problem of fragmentation arises again.

2.2 Wikis

Wikis have become increasingly common through the last years reaching from small personal wikis to the largest Internet encyclopedia Wikipedia. Since large companies have special requirements, such as fine-grained access-control, enterprise scalability, security integration and the like, a special type of wikis – enterprise wikis – emerged. A survey [10] about the use of corporate wikis pointed out, that corporate users expect three benefits from using wikis, namely improved reputation, relaxation of work and helping in the advancement of processes. Widely utilized wikis in the enterprise context are Confluence[5] and Jive[6]. Popular open-source wikis include FOS wiki[7] and TWiki[8]. These tools differ in their provided functionality, but they are all mainly centered around textual content, although some wikis provide limited support for managing structured information (e.g. FOS wiki via data forms). Consequently, the knowledge contained in those wikis can in most cases only be extracted by human reading of the documents and not by other applications used within the company.

[5] http://www.atlassian.com/software/confluence/overview
[6] http://www.jivesoftware.com/social-business/platform
[7] http://foswiki.org/
[8] http://twiki.org/

Linked Data Approach In addition to traditional wikis, there is also another category of wikis which are called semantic wikis. Those can again be divided into two categories: semantic text wikis and semantic data wikis. Wikis of this kind are not yet commonly used in enterprises, but crucial for enterprise data integration since they make (at least some of) the information contained in a wiki machine-accessible. Text-based semantic wikis are conventional wikis (where text is still the main content type), which allow users to add some semantic annotations to the texts (e.g. typed links). The semantically enriched content can then be used within the wiki itself (e.g. for dynamically created wiki pages) or can be queried, when the structured data is stored in a separate data store. An example is Semantic MediaWiki [9] and its enterprise counterpart SMW+. It extends the well-known MediaWiki engine (which powers Wikipedia) with syntax for typecasting links and data, classifying articles and creating dynamic pages. The knowledge in a wiki (KiWi) [14] project also developed a semantic wiki, which provides an adaptable platform for building semantic and social tools.

We propose the usage of semantic data wikis such as OntoWiki [2,8] in enterprises for the following main reasons: (1) Data wikis focus on structured information, which is kept as such and thus can be easily re-used by other applications consuming the data. (2) Since OntoWiki is solely based on RDF, all information is automatically published via the Linked Data principles, making it trivial for other parties to consume the data. (3) Information fragments can be interlinked with other resources within an enterprise (e.g. taxonomies, XML schemas, databases, Web services), which leads to a better reuse of information and thus to better maintained data. (4) Since textual information can also be represented in RDF (via literals), text wikis can be emulated and thus (additional) human-friendly information can be added. Such annotations and the structured information can then be used to create customized views on the data.

Challenges A challenge is to train users of wikis to actually create semantically enriched information. For example, the value of a fact can be either represented as a plain literal, or as a relation to another information resource (eventually already attached with some metadata). The more users are urged to reuse information wherever appropriate, the more all participants can benefit from the data. It should be part of the design of the wiki application (especially the user interface), to make it easy for users to build quality knowledge bases (e.g. through auto-suggestion of URIs within authoring widgets).

Since data in RDF is represented in the form of simple statements, information that naturally is intended to be stored in conjunction (e.g. geographic coordinates) is not visible as such per se. The same applies for information which users are accustomed to edit in a certain order (e.g. address data). A non-rational editing workflow, where the end-users are confronted with a random list of property values may result in invalid or incomplete information. The challenge here is to develop a *choreography of authoring widgets* in order to provide users with a more logical editing workflow.

Another defiance to tackle is to make the deployed wiki systems available to as many stakeholders as possible (i.e. cross department boundaries) to allow for an improved information re-use. Once Linked Data resources and potentially attached information are re-used (e.g. by importing such data), it becomes crucial to keep them in sync with the original source. Therefore mechanisms for *syndication* (i.e. propagation of changes) and *synchronization* need to be developed, both for intra- and extranet semantic wiki resources.

2.3 Web Portal and Intranet Search

Current state-of-the-art intranet data management system with proper full text search and a comfortable user interface include Microsoft's FAST Search[9], SAP's Netweaver Enterprise Search[10] or Autonomy's IDOL Universal Search[11]. These search engines are based on full-text search and offer taxonomy support, custom ranking and context awareness. Even though the search engines are quite sophisticated, there is still a lot of room for improvement that can be tackled by publishing the data as Linked Data and allowing it to be queried as such [5]. In [7], the author identifies several challenges in enterprise search, one of them being internal multi-source search, which is when a user has a precisely formulated question but only a keyword search is available. The author uses an example where a manager in an oil company wants to identify all wells previously drilled by the company in the Black Gold field where the problem known as "stuck pipe" was experienced. The manager searches for "Black Gold stuck pipe", but he must go through all the found documents, identifying the wells and so on. If the company data was stored or internally published as Linked Data using an ontology describing the oil drilling domain, the result could be gained using a single SPARQL query.

Linked Data Approach In an enterprise exist at least two distinct areas where search technology needs to be applied. On the one hand, there is corporate internal search, which enables employees to find relevant information required for their work. On the other hand, all large enterprises need at least simple search capabilities on their public web portal(s), since otherwise the huge amounts of information provided may not be reachable for potential customers. Some dedicated companies (e.g. automotive companies) would actually have a need for more sophisticated query capabilities, since the complexity of offered products is very high. Nevertheless, in reality, search, both internal and external, is often solely based on keyword matching. We argue that by employing the Linked Data paradigm in enterprises the classical keyword based search can be enhanced. Additionally, more sophisticated search mechanisms can be easily realized since more information is available in a uniform and machine-processable format.

In cooperation with Daimler a prototype that employs multiple Linked Data sources to provide a uniform search application was developed. By entering simple keywords users can (a) find documents that are attached to terms from the taxonomy that match the given query or (b) find specific car models that match the criteria given by the user (e.g. more than 6 seats). In the first case the advantage is, that documents can be found

[9] http://sharepoint.microsoft.com/en-us/product/capabilities/search/Pages/Fast-Search.aspx

[10] http://www.sap.com/platform/netweaver/components/enterprisesearch/index.epx

[11] http://www.autonomy.com/content/Products/idol-universal-search/index.en.html

even if the content does not mention the keyword. The second case would not be even possible without taking another datasource into account, namely structured information about possible car configurations. Thus, when a user queries for a keyword that matches a term that is linked to a car related property and also provides a value restriction (e.g. less than 10), the system can obtain a list of matching cars (via SPARQL queries) and return them to the user together with some metadata about the models.

Challenges In order to implement search systems that are based on a Linked Data approach and that provide a substantial benefit in comparison with traditional search applications, the challenge of *bootstrapping an initial set of high-quality RDF datasources* needs to be tackled first. Mechanisms then need to be established to automatically create high-quality links between datasets.

Finally, although a search engine that queries RDF data directly works (in fact the prototype described above was implemented using this approach), it results in suboptimal performance. The challenge here is to develop methods for improving performance to match traditional search engines, while keeping the advantages of using SPARQL directly.

2.4 Database Integration

Relational Database Management Systems (RDBMS) are the predominant mode of data storage in the enterprise context. We therefore deem the integration of relation data into Linked Data a crucial Enterprise Data Integration technique. For providing a unified view over different databases multiple methods like data warehousing, schema mediation and query federation have been devised and successfully used. However, problems arise with more heterogeneous data landscapes, where strict schema adherence can not be guaranteed and external data is utilized. The integration of heterogeneous sources requires a costly transformation of the data into the relational model. This has the effect, that only key data sources and thus only a small fraction of the RDBMSes in a typical enterprise are integrated.

Linked Data Approach The mapping of relational data to the RDF data model adopts relational database integration techniques and augments them. By employing a mapping from relational data to RDF, data can be integrated into an internal or external data cloud. By using URIs for identifying resources, integration with non-relational and external data is facilitated. The RDB to RDF Mapping Language (R2RML) standard describes how a relational database can be transformed into RDF by means of term maps and triple maps. In order to avoid a costly materialization step, R2RML implementations can dynamically map an input SPARQL query into a corresponding SQL query, which renders exactly the same results as the SPARQL query being executed against a materialized RDF dump. By avoiding a costly materialization of the relational data into a dedicated triple store, a light-weight integration into existing architectures is possible. Consequently, semantic wikis, query federation tools and interlinking tools can work with the data of relation databases. The usage of SPARQL 1.1 query federation [13] allows relational databases to be integrated into query federation systems with queries spanning over multiple databases.

Challenges A primary concern when integrating relational data is *scalability and query performance*. With our R2RML based tool SparqlMap[12] we show that an efficient query translation is possible, thus avoiding the higher deployment costs associated with the data duplication inherent in ETL approaches. The challenge of closing the gap between triple stores and relational databases is also present in SPARQL-to-SQL mappers and drives research. The standardization of the RDB to RDF Mapping Language (R2RML) by the W3C RDB2RDF Working Group establishes a common ground for an interoperable ecosystem of tools. However, there is a lack of mature tools for the creation and application of R2RML mappings. A challenge lies in the creation of user friendly interfaces and establish best practices for creating, integrating and maintaining those mappings. Finally, for a *read-write integration* updates on the mapped data need to be propagated back into the underlying RDBMS. An initial solution is presented in [3]. In the context of Enterprise Data an integration with *granular access control* mechanisms is of vital importance.

3 Conclusions

In this work we identified several data integration challenges that arise in corporate environments. We discussed the use of Linked Data technologies in those contexts and presented some insights gained during the development of corresponding prototypes for Daimler. We conclude from our experiments, that the deployment of Linked Data approaches in enterprise scenarios has huge potential and can result in extensive benefits. However, we are aware that more challenges than the aforementioned need to be tackled when trying to create sophisticated enterprise knowledge intranets. We consider as future work to investigate XML schema governance (due to the numerous applications that employ XML for information exchange) and enterprise single sign-on from a Linked Data perspective. In order to ultimately establish Linked Data as a strategy for enterprise data integration also many organizational challenges have to be tackled. For example, it is relatively easy to determine the return-on-investment for an integration of two information systems, while it is very difficult to precisely assess the cost savings of the Linked Data approach. Also, the added value of the Linked Data approach might only become visible after a critical mass of Linked Data interfaces and resources are already established in the enterprise.

References

1. S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives. Dbpedia: A nucleus for a web of open data. *The Semantic Web*, pages 722–735, 2007.
2. S. Auer, S. Dietzold, and T. Riechert. OntoWiki - A Tool for Social, Semantic Collaboration. volume 4273, pages 736–749. 2006.
3. V. Eisenberg and Y. Kanza. D2RQ/update: updating relational data via virtual RDF. In *WWW (Companion Volume)*, pages 497–498, 2012.
4. G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais. The vocabulary problem in human-system communication. *ACM*, pages 964–971, Nov. 1987.

[12] `http://askw.org/Projects/SparqlMap`

5. J. Grudin. Enterprise Knowledge Management and Emerging Technologies. In *System Sciences, 2006. HICSS '06*, volume 3, page 57a, 2006.
6. A. Halevy, A. Rajaraman, and J. Ordille. Data integration: the teenage years. In *VLDB '06*, pages 9–16. VLDB Endowment, 2006.
7. D. Hawking. Challenges in enterprise search. In *ADC '04*, ADC '04, pages 15–24, Darlinghurst, Australia, Australia, 2004. Australian Computer Society, Inc.
8. N. Heino, S. Dietzold, M. Martin, and S. Auer. *Developing Semantic Web Applications with the OntoWiki Framework*. Networked Knowledge. Springer, 2009.
9. M. Krötzsch, D. Vrandečić, and M. Völkel. Semantic MediaWiki. *The Semantic Web-ISWC 2006*, pages 935–942, 2006.
10. A. Majchrzak, C. Wagner, and D. Yates. Corporate wiki users: results of a survey. In *WikiSym '06*. ACM, Aug. 2006.
11. P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. DBpedia spotlight: shedding light on the web of documents. In *I-Semantics '11*, pages 1–8. ACM, 2011.
12. A. Miles and S. Bechhofer. SKOS Simple Knowledge Organization System Reference. *W3C Recommendation*, 2008.
13. E. Prud'hommeaux. SPARQL 1.1 Federation Extensions, November 2011. `http://www.w3.org/TR/sparql11-federated-query/`.
14. S. Schaffert, J. Eder, S. Grünwald, T. Kurz, and M. Radulescu. Kiwi–a platform for semantic social software (demo). *The Semantic Web*, pages 888–892, 2009.
15. C. Shirky. Ontology is overrated: Categories, links, and tags, 2005.