Interlinking Data and Knowledge in Enterprises, Research and Society with Linked Data

Sören AUER [a,1], Christoph LANGE [b]
[a] *University of Bonn and Fraunhofer IAIS, Germany*
[b] *University of Bonn and Fraunhofer IAIS, Germany*

Abstract. The Linked Data paradigm has emerged as a powerful enabler for data and knowledge interlinking and exchange using standardised Web technologies. In this article, we discuss our vision how the Linked Data paradigm can be employed to evolve the intranets of large organisations – be it enterprises, research organisations or governmental and public administrations – into networks of internal data and knowledge. In particular for large enterprises data integration is still a key challenge. The Linked Data paradigm seems a promising approach for integrating enterprise data. Like the Web of Data, which now complements the original document-centred Web, data intranets may help to enhance and flexibilise the intranets and service-oriented architectures that exist in large organisations. Furthermore, using Linked Data gives enterprises access to 50+ billion facts from the growing Linked Open Data (LOD) cloud. As a result, a data intranet can help to bridge the gap between structured data management (in ERP, CRM or SCM systems) and semi-structured or unstructured information in documents, wikis or web portals, and make all of these sources searchable in a coherent way.

Keywords. Linked Data, Enterprise Data Integration, Data Portals

1. Introduction

Integrating and analysing large amounts of data plays an increasingly important role in today's society. Often, however, new discoveries and insights can only be attained by integrating information from dispersed sources. Despite recent advances in structured data publishing on the Web (such as RDFa and schema.org) the question arises how larger datasets can be published, described in order to make them easily discoverable and to facilitate their integration and analysis.

One approach to this problem is Linked Data, which enables organisations to publish and describe data using comprehensive metadata schemes. Similar to digital libraries, networks of such data endpoints can support the description, archiving and discovery of datasets on the Web and within organisations' intranets.

[1]Corresponding Author: Sören Auer, Fraunhofer IAIS, 53754 Sankt Augustin, Germany; E-mail: soeren.auer@iais.fraunhofer.de.

Recently, we have seen a rapid growth of Linked Open Data and data catalogues being made available on the Web. The data catalog registry *datacatalogs.org*, for example, already lists 285 data catalogues worldwide. Examples for the increasing popularity of data catalogues include Open Government Data portals, data portals of international organisations and NGOs as well as scientific data portals.

Governments and public administrations started to publish large amounts of structured data on the Web, mostly in the form of tabular data such as CSV files or Excel sheets. Examples include the US' data portal *data.gov*, the UK's *data.gov.uk*, the European Commission's *open-data.europa.eu* as well as numerous other local, regional and national data portal initiatives. Also in the research domain data portals can play an important role, as almost every researcher produces or analyses data. However, quite often only the results of analysing the data are published and archived. The original data, that is ground truth, is often not publicly available thus hindering repeatability, reuse as well as repurposing and consequently preventing science to be as efficient, transparent and effective as it could be. Some examples of popular scientific open data portals are the Global Biodiversity Information Facility Data Portal[2]. Also many international and non-governmental organisations operate data portals such as the World Bank Data portal or the data portal of the World Health Organization.

This article discusses our vision of using Linked Data to evolve the intranets of large organisations – enterprises, research organisations or governmental and public administrations – into networks of internal data and knowledge. Section 2 states the problem, section 3 establishes principles for Linked Data in organisational intranets; the following sections explain them in more detail.

2. Data Integration in Large Organisations

Data integration is a key challenge in large organisations. We will discuss it in detail for large enterprises but note that other sectors are facing similar challenges. While production-critical information is often maintained in dedicated information systems already, such as Enterprise Resource Planning (ERP), Customer Relationship Management (CRM) or Supply Chain Management (SCM), the integration of these systems as well as their integration with the wealth of information from other sources is still challenging. Large enterprises often use hundreds of different information systems or databases. Just to mention one example we are familiar with, Daimler is still running around 3,000 independent IT systems, even after a decade of consolidation efforts.

With IT becoming increasingly widespread in enterprises, numerous data integration approaches, techniques and methods have been introduced. Over the previous ten years, these approaches were mainly based on XML, web services and service-oriented architectures (SOA). XML is a standard for syntactic data representation, web services provide data exchange protocols, and SOA is a holistic approach for distributed system architecture and communication. Still, these technologies have proved insufficient for a complete enterprise data integration. Particularly SOA requires a high integration effort.

[2] `http://data.gbif.org`

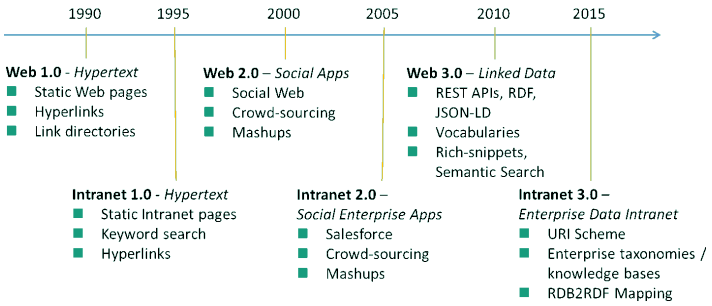

Figure 1. The evolution of the Web and enterprise intranets.

We argue that classical SOAs are well suited for transaction processing, whereas the Linked Data paradigm is more efficient for networking and integrating data. Like there is now a Web of Data complementing the original documented-centred Web, data intranets may help to enhance and flexibilise the intranets and SOAs that exist in large enterprises. Furthermore, using Linked Data boosts enterprises' information integration potential by giving them access to 50+ billion facts from the growing cloud of Linked *Open* Data (LOD). Finally, Linked Data is a lightweight, flexible information integration approach. It therefore suits well the increasing requirement to integrate data, e.g. about production, in real time, and to quickly respond to IT restructurings entailed by restructurings of enterprises such as mergers. Overall, using Linked Data for information integration improves an enterprise's competitivity and innovative capacity.

3. Evolution of the Web of Documents to the Web of (Linked) Data

The early World Wide Web (WWW) mainly consisted of static documents, but only a few years later dynamic applications such as content management or e-commerce systems emerged, which accessed data from structured sources and presented them in user- and context-specific views. Around the mid-2000s there was a sharp increase in user-generated content, e.g. in social networks or crowd-sourced knowledge bases. At the same time, mashups facilitated the aggregation of data from different sources. Both developments were subsumed under the term "Web 2.0". Finally, the late 2000s brought a turn to a Web of Data, which gained broad popularity with the *schema.org* initiative driven by the major search engine operators Google, Bing, Yandex and Yahoo, with Google's Knowledge Graph, an extension of the search engine by networked knowledge, and Facebook's Open Graph, which links external web resources to Facebook's social network.

With some delay web technologies were also introduced in intranets, starting from the mid-1990s (cf. figure 1). Meanwhile, portals, wikis and mashups are well established components of intranets. It is plausible that the next evolution step to a Web of Data will also soon take place, as intranets of large organisations and the WWW have many properties in common:

Decentral organisation: Information is organised in a decentral and distributed way both on the WWW and in large intranets. In large enterprises this

decentrality is owed to the fact that they comprise many organisational units, national sections and subsidiaries, and that therefore centralisation is either not completely necessary or not desired. Too much centralisation would, e.g., hamper the integration of units acquired from other companies, or the sale of units. Organisations therefore tend to accept a certain degree of decentrality, as long as common standards ensure interoperability.

Self-dependent units: Despite central management, often comprising a CIO (Chief Information Officer), organisational units and subsidiaries are self-dependent to some extent and have their own local key performance indicators measured. Thus, organisational units are often free to choose their own information architectures, independently of, and possibly incompatibly to, other units. This is similar to the WWW, where every operator of a sub(domain) is free to choose their own technology.

Heterogeneous information: Domain-specific applications, knowledge bases, document templates and data formats vary across organisational units.

In response to these characteristics, the following Linked Data design principles were conceived:[3] • identifying information units by HTTP URIs, • returning uniformly structured data when these URIs are dereferenced, i.e. treated as URLs, and • linking to other, related data. We propose to extend these principles as follows in organisational settings: • evolving existing thesauri, taxonomies, wikis and master data management systems into corporate knowledge bases and knowledge hubs, • establishing an organisation-wide URI naming scheme, • extending existing information systems in the intranet by Linked Data interfaces, and • establishing links between sources of related information.

4. Establishing Organisational Knowledge Bases

The availability of domain-specific knowledge is a key prerequisite to the success of an organisation. Such knowledge can be found in cross-organisation sources such as textbooks or standards, but mainly in organisation-specific sources such as glossaries, taxonomies, internal documents or data schemas. Large organisations are increasingly standardising their internal terminology in, often multilingual, thesauri, which comprise terms, definitions, synonyms, and other relations among terms, like WordNet[4] or Wiktionary[5]. The latter thesaurus is a successful example of crowdsourcing. Organisations can pursue a similar approach by enabling their staff to capture terms and their relations in a wiki. The choice of a wiki furthermore has the potential of turning such a thesaurus into the nucleus of an organisational knowledge base, which can, furthermore, be expanded by links to other data and information sources in the organisation. Despite being partly accessible publicly, Google's Knowledge Graph serves as an example for such an organisational knowledge base: it originates from Freebase, a Wikipedia knowledge extraction effort similar to DBpedia, which was acquired by Google and

[3] `http://www.w3.org/DesignIssues/LinkedData.html`
[4] `http://wordnet.princeton.edu`
[5] `http://wiktionary.org`

evolved into Google's central enterprise knowledge base. By now, it comprises, in addition to general-purpose data, extended information about products and creative works.

5. Identification by an Organisation-wide URI Schema

The key prerequisite for networking and integrating information is to establish unique identifiers for things such as persons, places, organisations, but also terms, data sources, products and contracts. On the Web, Uniform Resource Identifiers (URIs) and their internationalised extension (IRIs) have been established for identifying things, whereas URLs (Uniform Resource Locators), a subcase of URIs, make information resources accessible. URIs have the following properties:

Decentral maintenance: Everyone owner of an Internet domain or some (web-)space under a domain can define URIs within this namespace. Any organisation can thus define URIs independently from other organisations. Organisational units can proceed similarly, but an organisation may opt for a central URI management strategy, based on the same considerations as for (de)centrality in general, as discussed above. Table 1 reviews the respective advantages and disadvantages of different URI management strategies. In practice, they can be combined: e.g., organisational units could maintain their own URI registries, which are part of a federated organisation-wide registry. Decentral URI management can also be made manageable by central crawling and indexing, comparable to intranet-wide document search.

Dereferenceability: The Linked Data principles advise URIs that are dereferenceable as HTTP URLs and thus make information accessible.

Provenance: As Domain names are issued by registrars, one can determine their owners. URIs thus reveal their author, at least down to the level of an organisation or organisational unit; retrieving information about things by deferencing their URIs furthermore allows, in principle, to validate the authenticity and correctness of the information.

6. Managing Linked Data Throughout their Life Cycle

As data are becoming increasingly central to organisations, one has to consider all stages of their life cycle, as shown in the figure on the right. The following ones are of particular interest in organisational settings:

RDF data management: RDF is the native data model of Linked Data. RDF data is stored in dedicated databases, called triple stores, or relational databases with RDF interfaces. Techniques such as column stores, dynamic query optimisation, adaptive query caching, optimised graph processing, and

Management strategy	Pros	Cons
Central service, issuing uniform URIs on request	• easy overview of all identified resources • uniform structure of identifiers	• single point of failure • low flexibility • hard to centrally ensure dereferenceability
Identifiers issued decentrally, but registered centrally	• easy overview of all identified resources • resilient against technical failure and organisational changes	• requires synchronisation
Identifiers managed completely decentrally	• highly flexible • highly resilient against technical failure and organisational changes	• lack of central overview • lack of structural uniformity

Table 1. Properties of different identifier management strategies

clustering and cloud technologies have made the performance of RDF databases comparable to relational databases, with a higher modelling flexibility.

Authoring: Semantic wikis and WYSIWYM interfaces (what you see is what you mean) help users to create and maintain data in semantic knowledge bases.

Linking: (Semi-)automated detection and maintenance of links between different datasets helps to integrate large sets of organisational data.

Classification and Enrichment: Linked Data, on the Web or in intranets, largely consists of instance data. Data integration, fusion, search and applications require integration of raw data with shared taxonomies and vocabularies.

Quality Assessment: The quality of data, be they from the Web or from organisation-internal sources, varies. Decisions on whether to reuse certain data therefore need to be based on an assessment of their quality w.r.t. criteria such as provenance, context, coverage or structure.

Evolution and Repair: Data sources often evolve dynamically. In this situation it must remain possible to detect changes and modifications to datasets and vocabularies. Automated methods can help to identify problems in knowledge bases and to suggest repair strategies.

Search and Exploration: Different techniques for data search, browsing, exploration and visualisation, by spatial, temporal or statistical dimensions, help to present data and their interrelations to users.

These stages influence each other. Creation of mappings on the schema level influences mappings on the instance level, and vice versa. Schema mismatches across knowledge bases can be compensated by learning equivalences between concepts. End user feedback, e.g. about instance- or schema-level links, can serve as training data for machine learning techniques for inductive linking of larger knowledge bases, whose results can once more be assessed by end users for iterative refinement. Semantically enriching knowledge bases improves the detection of inconsistencies and modelling problems, which in turn improves linking, fusion and classification. The query power of the RDF data management component influences the software components corresponding to all other life cycle stages.

Being aware of these interdependences one should not aim at supporting individual stages in isolation, but at a cross-fertilisation of multiple stages and thus

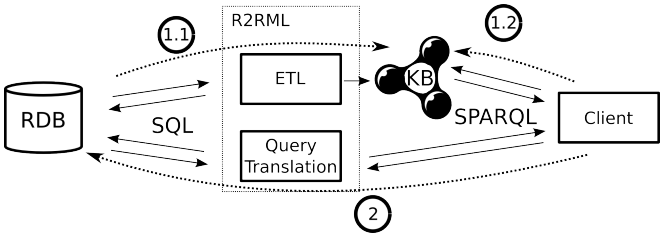

Figure 2. Different approaches to relational-to-RDF mapping: extraction vs. query translation

the life cycle as a whole. Improving a knowledge base w.r.t. one aspect (such as enriching it with links to a new knowledge hub) entails a number of possible improvements (e.g. additional instances matched by an auto-linker). Possibilities include sharing vocabularies among algorithms for data publishing, merging, repair and enrichment, and services that notify other services, or the original publisher of a dataset, of improvement possibilities. Provenance vocabularies help to transparently record such subsequent improvements of data. In the EU FP7 project LOD2 we have developed the LOD2 stack, which provides tool support for all stages of the life cycle [1].

7. Linked Data Interfaces

Adding Linked Data interfaces to existing information systems, so as to establish new crystallisation points for data-driven applications in organisational intranets, in most cases means translating relational data to Linked Data.

When publishing a relational database as Linked Data, a fundamental choice is between materialising the database into RDF, or exposing it as a virtual RDF graph on demand by query translation (cf. Figure 2). The first option means executing an extract-transform-load (ETL) process, where the relational-to-RDF mapping is applied in the *transform* step, and the result is *loaded* into an RDF database, whose resource URIs are then made dereferenceable over HTTP, or which is exposed using a SPARQL endpoint, i.e. a standard interface for querying Linked Data over HTTP. Creating a virtual graph means translating HTTP accesses or SPARQL queries into one or more SQL queries using a mapping; the relational database then answers these queries, the result set is transformed into RDF and sent back to the client. Materialisation benefits from indexing in the RDF database. The advantage of query translation lies in direct access to the master data, without the need for another ETL detour when they get updated. The choice of materialisation vs. query translation is also influenced by the access permissions to the original data.

Lifecycle-aware relational-to-RDF mapping requires further considerations. The identifier management strategies discussed in section 5 need to be considered for instance as well as schema URIs. The choice of RDF schema depends on the application; the Simple Knowledge Organization System (SKOS[6]) is suitable for a

[6] `http://www.w3.org/2004/02/skos/`

thesaurus. Responding to updates of the original data has already been addressed above. A Linked Data version of a company's thesaurus certainly does not have to be as up to date as a database that captures the current state of a production machine; however, communication with business partners may benefit from the thesaurus' revision history being exposed in a transparent and queryable way. Finally, feedback helps to fix or improve the mapping definition and the relational master data. As relational-to-RDF mappings usually access the relational database read-only, any feedback about the master data has to be indirect.

Relational databases are not only important sources for Linked Data in organisations, particularly in enterprises, but also for most existing Linked Open Data. Therefore, the W3C standardised the RDB to RDF Mapping Language (R2RML[7]). R2RML helps to describe transformations from a specific database schema to a specific RDF vocabulary. Different tools accommodate the diverse relational-to-RDF mapping requirements. Triplify[8] is a lightweight PHP library that exposes relational databases of web applications as Linked Data using a pattern matching approach simpler than R2RML. D2R[9] exposes relational databases as a SPARQL endpoint. SPARQL queries are translated to one or more SQL queries using a predecessor of R2RML. SparqlMap[10] supports R2RML; it maps a given SPARQL query to exactly one SQL query, which facilitates exposing large relational databases as Linked Data without materialising them. The R2RML package for the Virtuoso triple store[11] merges RDF extracted from relational databases into RDF databases already available inside the triple store, giving convenient SPARQL access to both.

8. Discovering Datasets using Data Portals and Link Discovery Tools

Once individual relational databases have been turned into Linked Data, their utility can be increased by linking them to other Linked Datasets within or outside of the organisation. The main challenge within the organisation is to make existing data available as Linked Data, as explained above for relational data, whereas the main challenge outside the organisation is to discover suitable open datasets.

The largest, most widely known, and most useful set of general-purpose data is DBpedia[12], which is extracted from Wikipedia. Similarly, there is LinkedGeoData[13] with information about spatial entities extracted from OpenStreetMap. Open datasets of specific interest to, e.g., enterprises include OpenCorporates[14] with information about 50,000+ companies world-wide and the Product Types Ontology[15], which extends DBpedia with precise definitions of around 300,000 types of products or services. As these datasets are community-maintained, it is,

[7] http://www.w3.org/TR/r2rml/
[8] http://triplify.org
[9] http://d2rq.org/d2r-server
[10] http://aksw.org/Projects/SparqlMap.html
[11] http://virtuoso.openlinksw.com
[12] http://dbpedia.org
[13] http://linkedgeodata.org
[14] http://opencorporates.com
[15] http://www.productontology.org

however, necessary to assess their quality. The quality requirements depend on the scenario. For example, when DBpedia is used to recommend related movies, missing information about some actor is negligible. On the other hand, one would rather not build an expert system suggesting medical treatments on DBpedia.

Governments, public administrations but also research and other non-governmental institutions increasingly use data portals to publish datasets – not always *Linked* Datasets, but often. Data portals usually feature a *catalogue* that lists the datasets available, each with a dataset-level metadata record and information of how to access the actual data. Datasets of interest can be found, e.g., by a search for keywords in the metadata records. High-quality datasets among them can be identified if the metadata records include quality metrics; we have developed a vocabulary for this [2] and are working on an extensible library that implements quality metrics for Linked Datasets. To enable quality-based browsing of data portals, we will develop an extension for the widely used CKAN data portal software[16], which will allow for filtering and ranking datasets according to different quality metrics, e.g. to select those datasets that carry provenance information and rank them by the well-typedness of their data entries.

Once suitable Linked Datasets have been found, one can employ tools to (semi-)automatically link them to existing organisational datasets. As a result of this link discovery process, e.g., resources with similar properties (such as the same name and prices equal within some tolerance) may be identified as equivalent and connected with an *owl:sameAs* link, which makes each property of one resource also apply to the other resource. The Silk and LIMES tools facilitate link discovery according to user-defined rules.[17] Linking organisation-internal data sources such as product taxonomies or domain-specific knowledge bases to big Linked Open Datasets pays off particularly well. The latter serve well as common referencing targets; e.g., Wikipedia and OpenStreetMap cover almost any concept or location of interest. Once two organisation-internal datasets share common links to open datasets, it becomes once more easier to automatically link them.

9. Conclusion and Outlook

Despite great efforts to consolidate and integrate organisational data, e.g. using SOA, ERP and CRM technology in enterprises, there is still a gap between business-critical, usually structured data and semi-structured or unstructured information available in documents, wikis or web portals. Information exchange between organisations and along value chains also still needs to improve, while at the same time organisations are increasingly specialising on their core businesses, which leads to increasingly complex supply and value chains. Information to be exchanged includes addresses, contact persons, consumption and delivery estimates, quality assurance and logistics information. Mastering such information exchange within one organisation requires flexible information structures that adapt to dynamic changes of corporate settings. Information exchange across organisations needs to be organised in a distributed way to ensure data protection and privacy.

[16] `http://ckan.org`

[17] See `http://silk.wbsg.de/` and `http://aksw.org/Projects/LIMES.html`.

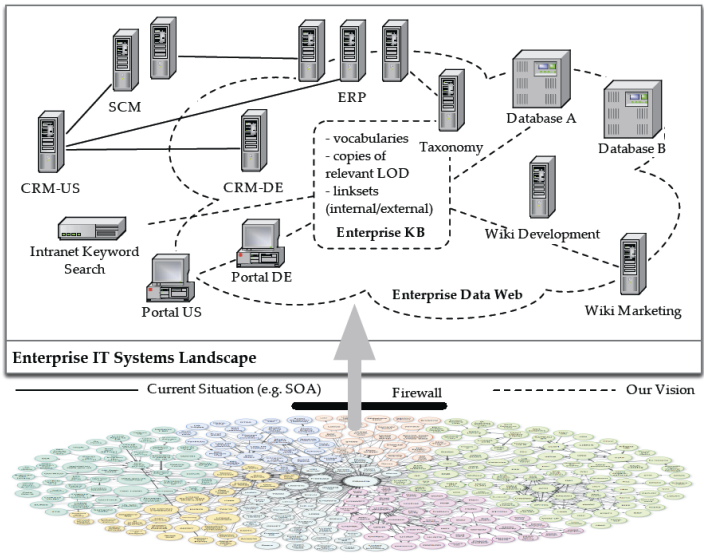

Figure 3. Towards an Enterprise Knowledge Hub [3].

The Linked Data paradigm, which is already widely in use on the Web, provides methods, standards and techniques for effective and efficient organisational information integration. Figure 3 shows our vision of an organisational knowledge hub in the concrete case of an enterprise (cf. [3]). Solid lines show connections between enterprise information systems that have existed before the introduction of Linked Data. Dashed lines show potential further connections enabled by Linked Data. The Enterprise Data Web comprises a central knowledge base, comprising a shared terminology, local copies of relevant Linked Open Datasets, and links among internal and to external datasets. Data from the public LOD Cloud can be reused inside the intranet, whereas internal data are not accessible from outside. Establishing such an Enterprise Data Web is possible with existing tools, which already cover the whole Linked Data life cycle, and are therefore suitable for incremental improvement and integration of data in organisational intranets, coming from both structured and unstructured sources.

References

[1] S. Auer, L. Bühmann, C. Dirschl, O. Erling, M. Hausenblas, R. Isele, J. Lehmann, M. Martin, P. Mendes, B. van Nuffelen, C. Stadler, S. Tramp, H. Williams. Managing the life-cycle of linked data with the LOD2 stack. *International Semantic Web Conference (ISWC)*, 2012.

[2] J. Debattista, C. Lange, and S. Auer. daQ, an ontology for dataset quality information. *Linked Data on the Web (LDOW)*, 2014.

[3] P. Frischmuth, S. Auer, S. Tramp, J. Unbehauen, K. Holzweißig, and C. Marquardt. Towards linked data based enterprise information integration. *Workshop on Semantic Web Enterprise Adoption and Best Practice (WASABI)*, 2013.