LINKED DATA PARADIGM FOR ENTERPRISES: INFORMATION INTEGRATION AND VALUE CHAIN

D. Hladky,
National Research University Higher School of Economics, Business Informatics Faculty
E-mail: dhladky@hse.ru

S.V. Maltseva,
National Research University Higher School of Economics, Business Informatics Faculty
E-mail: smaltseva@hse.ru

Address: 33 Kirpichnaya Str., Moscow, Russia

Data integration in enterprises is a crucial but at the same time costly and challenging problem. While business-critical information is stored in ERP, CRM, SCM and in Content Management systems the integration of such becomes even more critical when integrating with the growing information space on the Web. The IT industry has developed over the last decade integration solutions based on Master Data Management, Business Intelligence and the Service Oriented Architecture. However, we become increasingly aware that such technologies are not sufficient to ultimately solve all data integration challenges. Under the vision of context-aware services and integration we propose to apply the technology of the Linked Data paradigm. This approach seems to be promising, as scientists in the evolution of the Semantic Web have used it. We discuss Linked Data approaches in relation to the value chain and information integration of heterogeneous content and present an example of a CRM business process applying the Linked Data principles.

Key words: Linked data, tnterprise data integration, data WEB, linked open data, RDF, value chain.

Introduction

Data integration in enterprises is a crucial but at the same time costly, long lasting and challenging problem. Legacy data and business-critical information is often gathered and integrated in such systems like Enterprise Resource Planning (ERP), Customer Relationship Management (CRM) and Supply Chain Management (SCM). Large enterprises apply technologies like Master Data Management (MDM) and Business Intelligence (BI) in order to analyse and integrate data from the abovementioned systems. The landscape of those enterprises typically consists of several information systems, databases and content management systems. With the growing information on the Web, especially from News, Blogs, Social Networks (SN) and the growing Linked Open

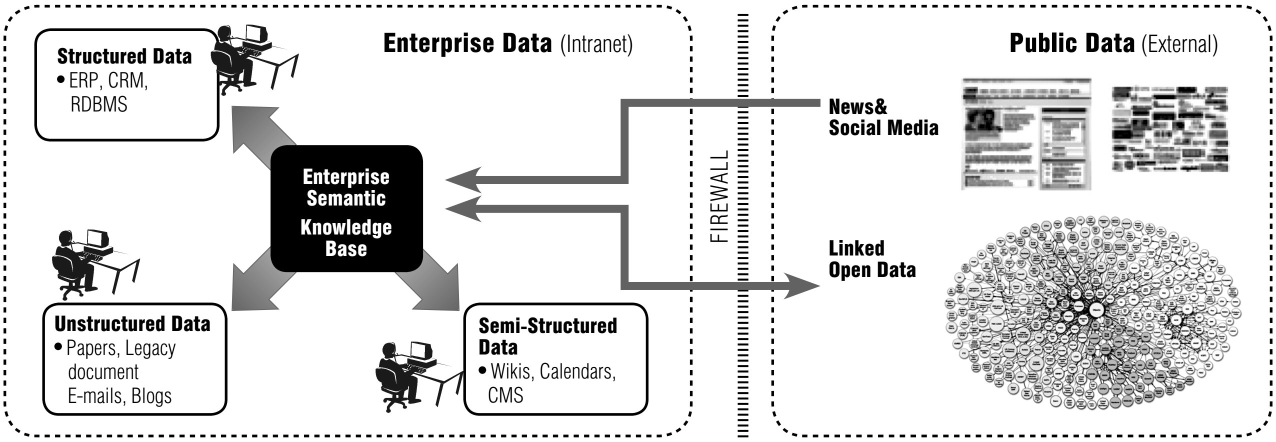

Fig. 1. Enterprise Data landscape

Data[1] (LOD) cloud the challenge of information integration has increased.

After the arrival and proliferation of IT in enterprises, various approaches, techniques and methods have been introduced in order to solve the information (data) integration challenge. In the last decade those data integration approaches were primarily based on XML, Web Services, Service Oriented Architecture (SOA) and the implementation of Master Data Management (MDM) and Business Intelligence (BI) [1]. XML defines standard syntax for data representation, Web Services provide data exchange protocols, SOA is a holistic approach for distributed systems architecture and communication, MDM can be used to support the maintenance of master data within an organisation which might be distributed to various systems and BI is a methodology and tool to analyse transactional data from different databases.

In this paper we explore the challenges of information integration and the link to the value chain, especially based on the example of a CRM process. These include but are not limited to, enterprise taxonomies (ontology), management of data fusion and interlinking, information extraction approaches from unstructured text and a model of the Linked Data value chain. We discuss Linked Data approaches in these areas, highlight benefits and present an example of a successful prototype based on CRM using the Linked Data principles.

In the remaining paper we first elaborate the landscape of enterprise data and public available data (Section 1). Bearing these in minds, we discuss the integration challenges (Section 2) followed by the CRM use case (Section 3) linked to the value chain using the Linked Data model. We then conclude the paper in Section 4 and highlight some future research challenges.

1. Enterprise data landscape

An enterprise data landscape (*Fig. 1*) consists of various data sources and corresponding applications. We clearly distinguish data sources that are only available to internal users and those in the public domain. We also differentiate data that are properly structured (e.g. databases, Linked Open Datasets), semi-structured (e.g. wiki pages, content management systems, etc.), and unstructured (e.g. Online News, Blogs, Tweets, and legacy documents). The enterprise landscape is a multiple-layer data/information integration platform.

We find that each system has an own structure and taxonomy describing the underlying data. In order to integrate the relevant enterprise datasets into an enterprise knowledge base, additional datasets from the public Web have to be integrated and interlinked with the enterprise taxonomy.

Some structured data from internal sources can be accessed directly by using for example SQL and others are only available via Web Services or specific API calls. Public structured data such as the LOD datasets can be processed using SPARQL[2] whereas other datasets are only available via predefined API calls (e.g. https://dev.twitter.com/docs/api). Semi-structured data from both internal and external sources are processed typically in two stages. First, the structured part is extracted and secondly the free-text part is processed using Natural Language Processing (NLP) us-

[1] http://linkeddata.org
[2] http://www.w3.org/TR/rdf-sparql-query/

ing named entity extraction. Similar to the semi-structured data process unstructured data (e.g. Online News, Blogs) is processed using only NLP and named entity extraction. Social Networks (e.g. Twitter) typically provide certain functions via their API calls but when processing the plain text we have to rely again on named entity or keyword extraction using the NLP engine.

In order to achieve a vision of applying Linked Data for Enterprises resulting in semantically interlinked enterprise we identified several integration challenges that will be discussed in the next section.

2. Data integration challenges

We identified 4 crucial areas (*Table 1*) where data integration challenges arise in enterprises. In the following section we investigate those challenges in more detail, 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.

Table 1.

Overview of date integration challenges

Information integration challenge	Current state of the art	Linked Data benefit
Enterprise taxonomy	Proprietary, centralized, no relationship between terms, multiplane independet terminologies	Open standarts based on RDF, OWL vocabularies (e.g. SKOS), multilingual, re-usable
Web Portal and Intranet	Keyword search over textual content	Facedet search and semantic contend extraction using NLP
Database Integration	Data Warehouse (MDM, BI), schema mediation, query federation	Lightweight data integration through RDF layer
Data fusion and linking for Knowledge Base	Mainly manual process of merging	Several research projects like LOD2, SILK developed semiautomated processes that support linking and fusion based on RDF

2.1. Enterprise taxonomy

Nowadays many larger enterprises uses taxonomies to provide a shared linguistic model aiming at structuring the large quantities of documents, emails, product descriptions, enterprise directives, etc. which are produced on a daily basis.

2.1.1. State of the Art

It is widely agreed that taxonomies are usable, however, there are multiple challenges that must be addressed in order for taxonomies to work [2]. A common and obvious problem is the additional manual work involved in the creation of the metadata for each digital object. The creator often doesn't see the direct benefit arising from the proper classification through the taxonomy but can be solved by providing internal guidelines and explanations. A larger problem is related to the fact that different metadata creators use different terminologies for the same digital object and hence various metadata descriptions arise [3]. Another challenge is on avoiding duplicates and maintaining a list of synonyms. Larger enterprises started to deploy solutions such as Microsoft SharePoint Term Store whereas smaller enterprises rely on centralised file systems where the enterprise taxonomy is stored for example in an Microsoft Excel Sheet.

2.1.2. Linked Data Approach and Benefit

Following the W3C guidelines it is proposed to present the enterprise taxonomies in RDF using the SKOS vocabulary [4], [5] and publishing terms definitions via the Linked Data principles. The benefits of using this approach are:

1. As terms are attached to globally unique identifiers (URI) they can be dereferenced using HTTP using just a Web browser. No additional software or application is needed.
2. Term management in a distributed environment can be easily achieved as the terms can be interlinked.
3. Applying SKOS vocabulary metadata granularity can be solved.
4. Terms can have multiple labels and hence can be used to solve the problem of multilingualism.

In order to build and maintain the enterprise taxonomy using RDF and SKOS several approaches are available. A first starting point is to use the service DBpedia Spotlight [6] integrated with an existing environment allowing to annotate text resources from DBpedia. Such a service will not satisfy all enterprise requirements, especially the one related to the enterprise specific taxonomy. Hence it is possible to use Computer Linguistic services (e.g. OntosMiner[3], OpenCalais[4], Alchemy[5],etc.) to further annotate the data sources. Such services can be customized in a way that enterprise vocabularies can be integrated. Various taxonomy management systems ex-

[3] http://www.ontos.com/?page_id=630
[4] http://www.opencalais.com/
[5] http://www.alchemyapi.com/

ist whereas the enterprise can define and maintain the taxonomy using the W3C SKOS standard [7]. Possible tools include Protégé, OntoWiki and TopQuadrant's Enterprise Vocabulary Net (EVN).

2.2. Web Portal and Intranet

One of the biggest problems with enterprise intranets is the huge difference in user experience in comparison to the de facto standards like Google or Facebook [8]. The user is spoiled with intuitive front ends, precise search, autocomplete text boxes etc. The challenge within enterprises is to offer the same kind of experience.

2.2.1. State of the Art

Independent of the budget various commercial solutions (e.g. Microsoft FAST Search[6], SAP NetWeaver Search[7], Autonomy IDOL[8]) are kind of state-of-the-art available for the deployment within enterprises. They typically support key word based search, are supporting multiple languages, allow federation across different sources and have basic support for taxonomies. The search engine solutions are quite sophisticated but have room for improvement, especially applying Linked Data technologies [2; 9].

2.2.2. Linked Data Approach and Benefit

In enterprises two distinct searches for information are required. On the one hand a user needs to find the data inside the enterprise relevant to his task but on the other hand very often the user needs to enrich or enhance the result with information and data from public Web portals. Let's consider the example of a CRM user that wants to understand all the facts about his customer. Customer master data and activities are typically stored in the CRM system. Other data like open invoices, support issues, status of the product production and product documentation is spread across various data silos inside the enterprise. Latest news, tweets or social network contacts are only available on external sources. The classical approach of finding information can be achieved by using the previously mentioned solutions. Applying the Linked Data paradigm it is argued that the search, linking and enhancing of information can be improved. Results are more precise and meaningful due to the domain ontology (an apple could be the fruit or the company) where results are matched against the enterprise ontology and concepts. Due to the RDF linking concept (e.g. owl:sameAs) concepts from different sources are interlinked and can be easily retrieved using the SPARQL query. Using the same method the internal terms can be connected to the growing Linked Open Data cloud and enhance the result by automatically enriching the content via Linked Data mashup [10]. In chapter 3 the use is further examined and connected to the value chain. *Fig. 8* shows the prototype where various data sources are connected and visualized inside the CRM system using the Linked Data paradigm approach.

2.3. Database Integration

Relational Database Management Systems (RDBMS) are the predominant mode of data storage in the enterprise context. RDBMS are used practically everywhere in the enterprise, serving, for example, enterprise resource planning (ERP), supply chain management (SCM), customer relationship management (CRM) and content management systems (CMS). We therefore deem the integration of relation data into Linked Data a crucial Enterprise Data Integration technique.

2.3.1. State of the Art

For providing a unified view over different databases multiple methods like data warehousing, schema mediation and query federation have been devised and successfully used. Many enterprises also use master data management (MDM) and business intelligence (BI) in order to extract and load subsets of data into a new environment allowing analytics (e.g. using BI) or to remove duplicates, standardized data or to author the data (e.g. using MDM). Integration of the heterogeneous data landscape requires a costly transformation process and is very often limited to a subset of the available information.

2.3.2. Linked Data Approach and Benefit

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

[6] http://sharepoint.microsoft.com/en-us/product/capabilities/search/Pages/Fast-Search.aspx
[7] http://www54.sap.com/solutions/tech/collaboration-content-management/software/enterprise-search/index.html
[8] http://www.autonomy.com/content/Products/products-idol-server/index.en.html

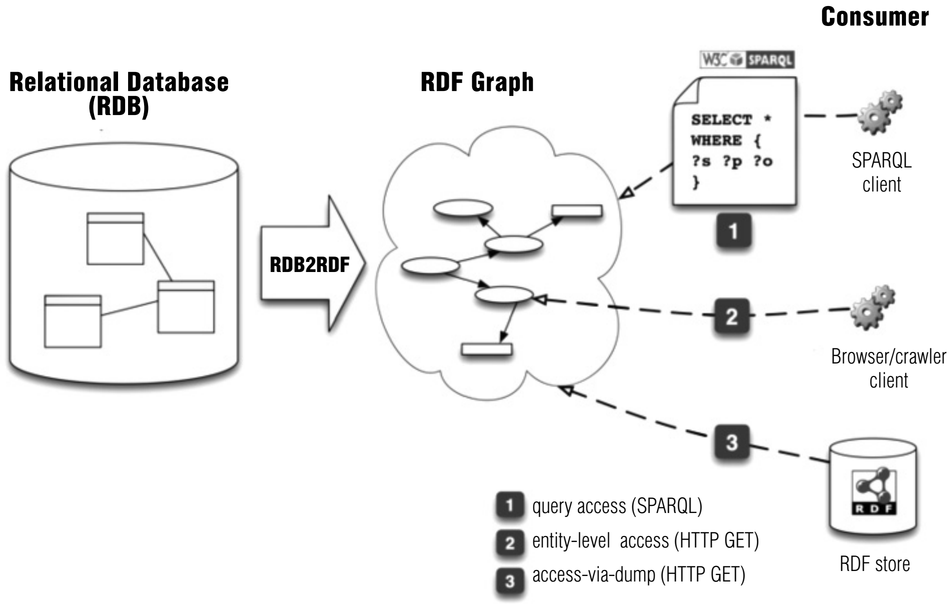

Fig. 2. RDB to RDF conversion. Source: http://www.w3.org/2001/sw/rdb2rdf/use-cases/

resources, integration with non-relational and external data is facilitated. The general approach of mapping a relational database into RDF using the RDB to RDF Mapping Language (R2RML) is specified by the W3C RDB2RDF Working Group[9]. The results of the transformation are terms and triples that are stored as a RDF knowledge base in a triple store. The triple store can be accessed using the SPARQL language. The triples can be easily linked with the enterprise taxonomy and enhanced in a similar way as described under chapter 2.2.2.

2.4. Data Fusion and Linking

One of the key challenges in the Linked Data process is related to data fusion and linking. Data fusion is known from the database literature where the integration process starts with a schema mapping task, followed by duplicate detection and data fusion steps in order to have clean data. In our Linked Data approach for enterprises we process various data sources and hence the fusion part is very relevant in order to have a unique identifier for a given instance of a term (concept or named entity such as organization or person). As the goal is to connect the term to other relevant data sets we need to enable a process to discover and link such terms using the Linked Data paradigm.

2.4.1. State of the Art

Once data has been retrieved from distributed sources, it must be integrated in a meaningful way before it is displayed to the user or is further processed. Today, most Linked Data applications display data from different sources alongside each other but do little to integrate it further. To do so does require mapping of terms from different vocabularies to the applications target schema, as well as fusing data about the same entity from different sources, by resolving data conflicts. A lot of this work is today based on a manual approach. More research work is need to support schema mapping and data fusion, especially on the resolution of data conflicts. A large body of work on data fusion [11] in the database community can be used as a foundation. Recently a tool that is being used is google-refine[10] with the RDF extension. The tool allows data cleansing, fusion and an export to RDF. The downside is that it is still very much manual work and doesn't support high volume and automatic processes.

2.4.2. Linked Data Approach and Benefit

Current EU FP7 funded projects like LOD2[11] and LATC[12] incorporate new techniques based on the Linked Data paradigm to improve the instance matching in order to generate links between data sets. As stated before this area needs more research work in or-

[9] http://www.w3.org/2001/sw/rdb2rdf/
[10] http://code.google.com/p/google-refine/
[11] http://lod2.eu/
[12] http://latc-project.eu/

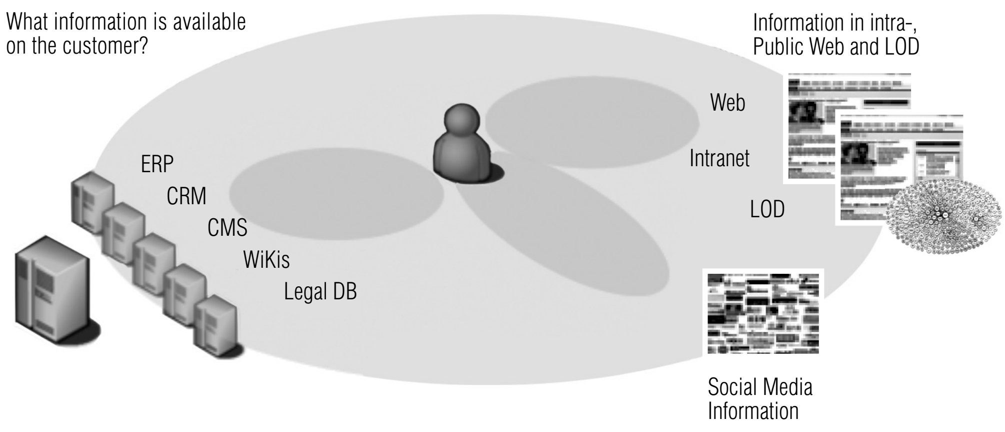

Fig. 3. Know Your Customer

der to improve the automation process and fusion and link quality. Another interesting project is related to the cleansing part of the data allow to resolve conflicts of content. The Sieve[13] project aims to improve the data quality and hence complements the tools such as SILK and LIMES from the LOD2 and LATC project.

3. Linked Data Value Chain based on CRM

As explored in the previous chapter Linked Data is evolving and creating a lot of potential for enterprises [12] but is still very much driven by the research community. The essentials of how to use the technology and how to publish Linked Data is known but not widely adopted in the commercial sector [13], [14]. A reason for this is caused by the lack of conceptual work supporting the business case. In order to bridge the gap we suggest linking the Value Chain and the Linked Data paradigm by developing a model. Such a conceptual model assists the design of a business case, supports the previously described challenge of information integration along the value chain and sets the foundation for future assessment of the investment. In chapter 2.2.2 we briefly introduced the use case for which we will develop the model.

In a simplified view the user is interested in knowing all relevant data about his customer (*Fi. 3*) that is spread across many information sources. By collecting data, linking information together and visualising them in a suitable user interface the user gains knowledge about his customer [15]. Applying new technologies such as Linked Data along the value chain can support the process and automate many tasks of the knowledge generation.

[13] http://sieve.wbsg.de/

3.1. Value Chain in the Internet Age

The original Value Chain was used by M. Porter [16] and describes the activities the organization performs and links them to the organisation competitive position. Porter distinguishes between primary activities and support activities. Primary activities are directly concerned with the creation or delivery of a product or service. They can be grouped into five main areas: inbound logistics, operations, outbound logistics, marketing and sales, and service. Each of these primary activities is linked to support activities, which help to improve their effectiveness or efficiency. There are four main areas of support activities: procurement, technology development (including R&D), human resource management, and infrastructure (systems for planning, finance, quality, information management etc.). The term «Margin» implies that organizations realise a profit margin that depends on their ability to manage the linkages between all activities in the value chain. In a simplified view the better the management of the activities is the greater the margin will be. With the advent of the Internet new technologies are influencing the information age within an organisation. Integration in real-time and across platforms, bidirectional communication, and ease of connectivity at much lower costs are part of the enterprise IT landscape. Those prominent applications from the Internet required an adaptation of the value chain (*Fig. 4*). Multiple activities are being linked together through such tools like Enterprise Resource Planning (ERP), Customer Relationship Management (CRM) and Supply Chain Management (SCM).

A suggested next stage in the improvement is to optimise the connectivity and exchange of context in real-time. The power of the Internet in the value chain can influence the

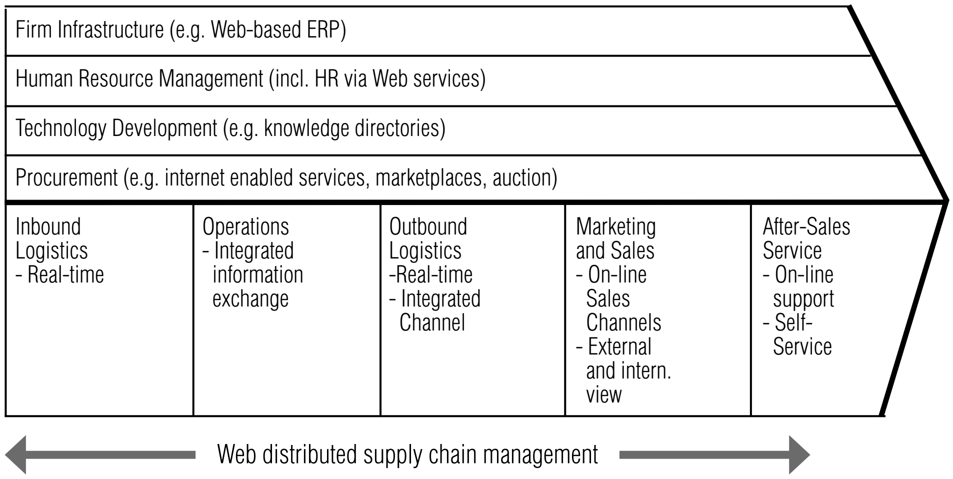

Fig. 4. Value Chain in the Internet Age

cost and quality of activities and leading to better margins. Within the next section we will depict the CRM Value Chain highlighting areas where new tools from the Internet Age, especially from the Linked Data and Semantic Web can play an important role for cost savings.

3.2. CRM Value Chain

Based on the previous discussed value chain model from Porter we can adapt the model to the specific process of a CRM system (*Fig. 5*). The focus on connecting new technologies is on the primary activities such as customer portfolio analysis, customer intimacy, network development, value proposition and the management of the relationship. Understanding customer needs, markets, competition and trends can influence the success of the sales person and hence the profitability/margin of the organisation. In a generalised form the goal is to reduce manual work of retrieving relevant content and to increase the knowledge about the customer and the environment and therefore increase the competitive advantage [17]. Reducing manual work results in cost savings and increased competitive advantage should lead to more revenue.

Linked Data and information integration along the value chain assists the user in gaining knowledge by providing human readable data. For a successful business case we argue that a conceptual model is needed in order to describe the involved entities, the various roles and data allowing a proper assessment.

3.3. Linked Data Conceptual Model Value Chain

The gathering of data and the transformation of such using the Linked Data approach (*Fig. 6*) can be simplified by introducing the participating entities, assigning the Linked Data roles and processed type of data. The valuable output is human-readable data for the user creating consumable knowledge. The model (*Fig. 7*) makes the interdependencies of entities, roles and different types of data explicit.

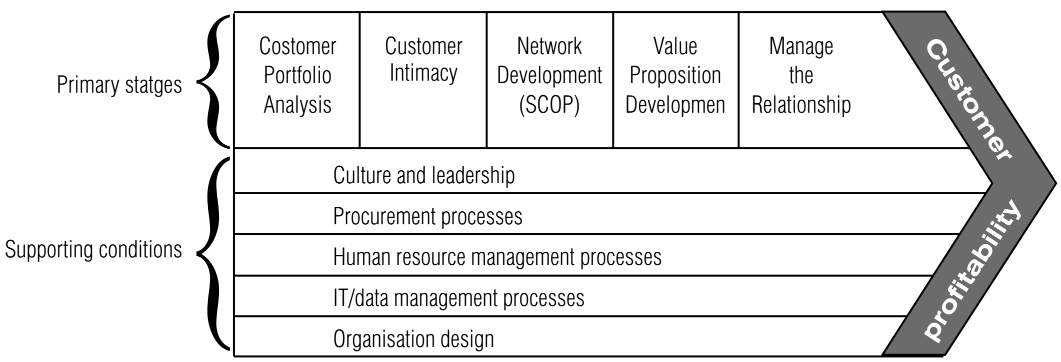

Fig. 5. CRM Value Chain

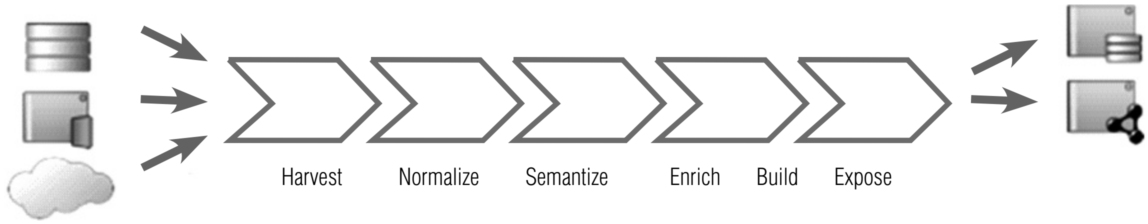

Fig. 6. Linked Data process chain

In the context of Linked Data, participating entities (corporate and non-corporate, e.g. persons, enterprises) can occupy one or more of the following roles:

✦ Raw Data Provider is a role that provides any kind of data in any non-RDF format.

✦ Linked Data Provider is a role that provides any kind of data in a machine-readable (e.g. RDF) format.

✦ Linked Data Application Provider is a role that processes Linked Data and creates human-readable output.

✦ End User is a human consuming the data within his preferred user interface (e.g. CRM UI).

The Linked Data Value Chain allows flexible assignment of roles to entities. For example an enterprise could own the role of Raw Data and Linked Data Provider.

The Linked Data Value Chain also supports multiple sources of data and those can be linked to more than one Linked Data Application Provider. The Data Types within the process are:

✦ Raw Data is any kind of data (structured or unstructured) that has not been converted yet to Linked Data (e.g. RDF).

✦ Linked Data is data in a RDF format that uses dereferenceable HTTP URIs[14] to identify resources and is linked with other RDF data.

✦ Human-Readable data is any kind of data that is intended, arranged and formatted for consumption by humans using a suitable user interface.

Within the following proof of concept we explain how the Linked Data Value Chain is used and describe how the user is consuming the human-readable data.

3.3.1. Proof of Concept

Using Linked Data within enterprises is still new and the following example has been developed together with Ontos[15] a pioneer in semantic Web technologies. The motivation of the use case was to provide additional information deriving from various sources inside the CRM system. Applying the Linked Data Value Chain model we can specify the roles and interactions along the transformation process.

Table 2.

Linked Data Roles and Participating Entities use case

Linked Data Roles	Participating Entities
Raw Data Provider	Ontos
Linked Data Provider	Ontos DBpedia, Freebase
Linked Data Application Provider	Ontos SILK LIMES

As summarized in *Table 2*, Ontos acts as Raw Data Provider, Linked Data Provider and Linked Data Application Provider. Within the process DBpedia and Freebase act as Linked Data Provider where information about companies and persons are consumed. On the application layer open source modules like SILK and LIMES are used in order to establish links between Linked Data Providers. The human-readable data for the end user is visualised inside the SugarCRM[16] system via the connector[17] (*Fig. 8*). The benefit for the user derived from the Linked Data Value Chain is the seamless information integration from heterogeneous sources.

4. Conclusion

In this work we identified several data integration challenges that arise in enterprise environments. In particular we examined taxonomies (ontology), structure and unstructured data integration as well as the integration and usage of Linked Open Data. Furthermore, we discussed the use of the Linked Data technologies and present some insight gained through the development of

[14] http://www.w3.org/2001/tag/doc/httpRange-14/2007-05-31/HttpRange-14
[15] http://www.ontos.com
[16] http://www.sugarcrm.com/
[17] similar to iframe http://en.wikipedia.org/wiki/HTML_element#Frames

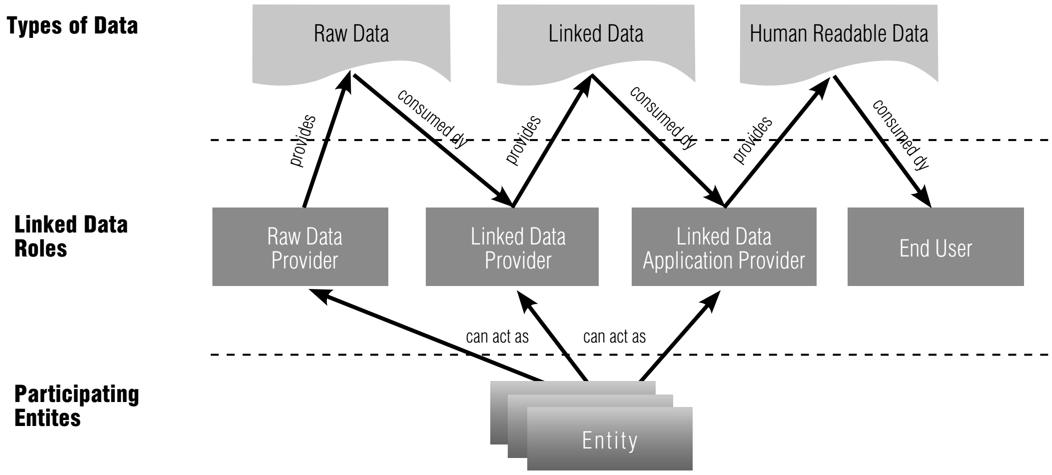

Fig. 7. Linked Data Value Chain Model

the CRM prototype. We conclude from our prototype that the Linked Data approach has huge potential and can result in extensive benefits. However, we are aware that not all of the aforementioned Linked Data technologies currently provide fully automated processes.

The future work of the on-going project is to identify more application scenarios that can show the benefit of the information integration and linkage to the value chain and to develop a measuring model (Return on Investment «ROI») based on the Linked Data model presented in this paper. We foresee especially more research work on data fusion and linkage solving Unique ID's (UID) and the improved connection between available datasets. In order to deploy such a technology on a large scale we need to make further test with high performance and scalable RDF stores.

Acknowledgements

The work was performed with financial support from the Ministry of Education and Science of the Russian Federation within the scope of Government Contract No. 07.524.11.4005 as of October 20, 2011 concluded between the Ministry of Education and Science of the

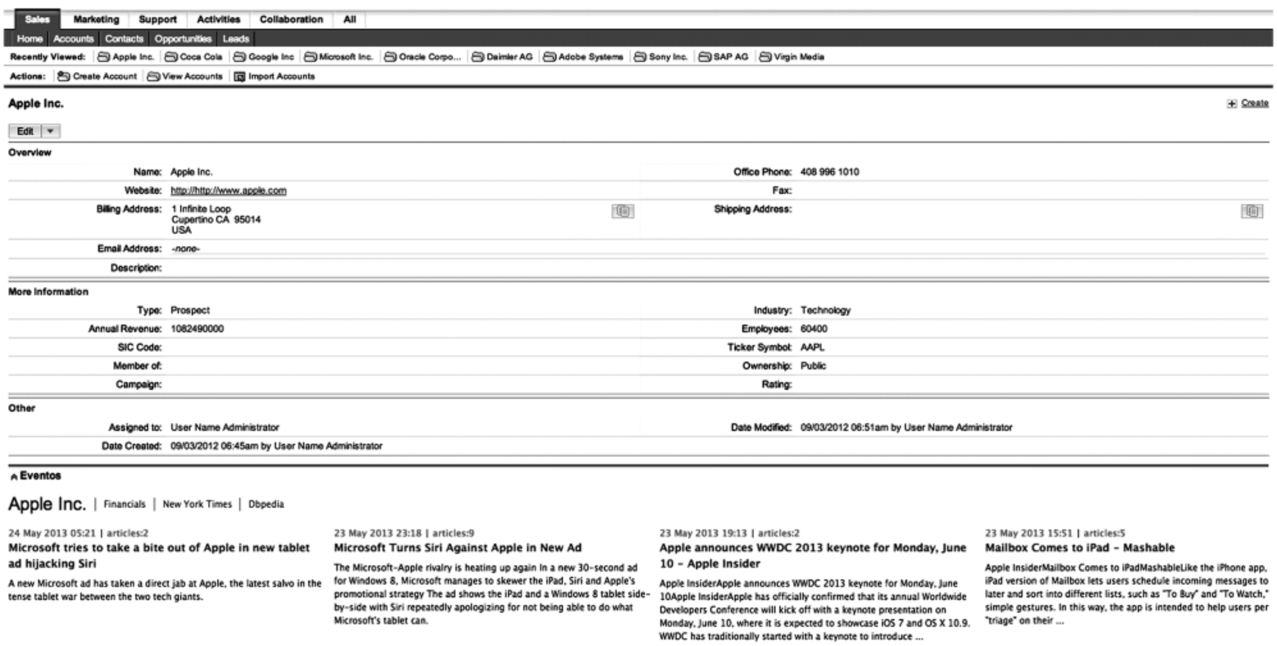

Fig. 8. SugarCRM with integrated information from Knowledge Base

Russian Federation and CJSC «Eventos».

Further we like to express our gratitude to Ontos AG for supplying modules from the Semantic Framework and for implementing the CRM use case on their SugarCRM system. We also like to express our appreciation to the EU FP7 funded research project Geoknow[18] (Grant Agreement No. 318159) providing modules of the technology stack. ■

References

1. Halevy A., Rajaraman A., Ordille J. Data integration: the teenage years. – Sep. 2006. – P. 9-16.
2. Grudin J. Enterprise knowledge management and emerging technologies // Proceedings of the 39th Annual Hawaii International Conference on System Sciences (HICSS'06), 2006. – P. 57.
3. Furnas G.W., Landauer T.K., Gomez L.M., Dumais S.T. The vocabulary problem in humansystem communication // Communications of the ACM. – Nov. 1987. – Vol. 30, No. 11. – P. 964-971.
4. Miles A., Bechhofer S. SKOS Simple Knowledge Organization System Reference. – W3C Recomendation, 2009. – [Online]. Available: http://www.w3.org/TR/skos-reference/.
5. Isaac A., Summers E. SKOS Simple Knowledge Organization System Primer. – Knowledge Organization, 2009. – [Online]. Available: http://www.w3.org/TR/skos-primer/.
6. Mendes P., Jakob M. Dbpedia spotlight: Shedding light on the Web of documents // Proceedings of the 7th ..., 2011 – P. 1–8.
7. DuCharme B. Improve your taxonomy management using the W3C SKOS standard, 2011. – [Online]. Available: http://www.ibm.com/developerworks/library/x-skostaxonomy/index.html.
8. Mukherjee R., Mao J. Enterprise Search. Tough stuff // Queue. – No. April 2004.
9. Schmidt K., Oberle D., Deissner K. Taking enterprise search to the next level // Relation. – 2009. – P. 2.
10. Bizer C., Heath T. Linked Data – Evolving the Web into a Global Data Space. – Morgan & Claypool Publishers, 2011 – P. 137.
11. Bleiholder J., Naumann F. Data fusion // ACM Computing Surveys. – Dec. 2008. – Vol. 41, No. 1. – P. 1-41.
12. Servant F. Linking Enterprise Data // Development. – 2008. - Vol. V, No. 369. – P. 1-5.
13. Groza T., Handschuh S., Decker S. Bridging the Gap Between Linked Data and the Semantic Desktop // Scenario. – 2009. – P. 815-830.
14. Bizer C., Heath T., Idehen K., Berners-Lee T. Linked data on the Web (LDOW2008) // Proceeding of the 17th international conference on World Wide Web, WWW '08. – 2008. – P. 1265.
15. Wu S.-C. The CRM of Tomorrow with Semantic Technology // Knowledge Technology, Vol. 295. – D.Lukose, A. Ahmad and A. Suliman, Eds. – Springer Berlin Heidelberg, 2012. – P. 46-51.
16. Porter M.E. Competitive Advantage: Creating and Sustaining Superior Performance. – Free Press, February 1985. – Vol. 1. – P. xviii, 557.
17. Hladky D. Sustainable Advantage for the Investor Relations Team through Semantic Content // Semantic Web, InTech, 2010.

[18] http://geoknow.eu/