A Case Study of Linked Enterprise Data

Bo Hu[1], Glenn Svensson[2]

[1]SAP Research
[2]BTS EMEA, SAP AG
{bo01.hu, glenn.svensson}@sap.com

Abstract. Even though its adoption in the enterprise environment lags behind the public domain, semantic (web) technologies, more recently the linked data initiative, started to penetrate into business domain with more and more people recognising the benefit of such technologies. An evident advantage of leveraging semantic technologies is the integration of distributed data sets that benefit companies with a great return of value. Enterprise data, however, present significantly different characteristics from public data on the Internet. These differences are evident in both technical and managerial perspectives. This paper reports a pilot study, carried out in an international organisation, aiming to provide a collaborative workspace for fast and low-overhead data sharing and integration. We believe that the design considerations, study outcomes, and learnt lessons can help making decisions of whether and how one should adopt semantic technologies in similar contexts.

1 Introduction

Thus far, the Linked Data (LD) initiative has demonstrated its value through a variety of projects aiming at improving data accessibility for primarily public and academic users [2]. The success stories certainly have not slipped the attention of large enterprises. Cautious attempts were made to experiment the LD principles and to evaluate the benefits, leading to the so-called "linked enterprise data" paradigm, the counterpart of LD in the business domain [14].

The motivation behind linking enterprise data is evident. Nowadays, with the deepening of globalisation, more and more non-mission-critical businesses are outsourced away from the home countries to for example design teams in Europe, manufacturers in China and service support in India. Fluctuation and risk in local markets, especially volatile ones, therefore becomes more manifested at the global level. This phenomenon has drawn more attention to business agility and continuity, a common ingredient of both being the easy access to data facilitating coordination and collaboration across different geographical locations. Businesses must be able to optimise their internal enterprise data landscape and explore such a landscape at the speed of thought so as to react to the rapidly changing market. Executives must be timely and comprehensively informed so that they can make decisions to counteract the threats to business revenue. More importantly, everyone needs to have ready and immediate access to information/data that enable her to carry out the allocated tasks.

Accessing data in an enterprise context, though not a new research area, is not a topic that we can comfortably mark as "solved" [10]. Enterprise data management has become a prevalent challenge with the rapidly plummeted storage and digitising cost resulting in an unprecedent amount of artefacts available in electronic form[1]. Linked Data initiative was proposed to deal with exactly this problem in the public domain, i.e. removing the barriers to data access and sharing. Intuitively, it seemed that we can just borrow the concepts having been so successfully implemented and recreate the stories in the enterprise environment. Our experience, however, prove otherwise. Indeed, enterprise data has many characteristics that resemble the data from public domains [6]. It, at the same time, presents unique requirements that put into test the principles and assumptions that are widely enjoyed when linking public data sets. The differences are demonstrated in the following aspects. Firstly, enterprise data is normally tied closely with the business processes. Peeling off the contexts wherein the use of such data takes place might render the data linking effort less fruitful. Secondly, it becomes increasingly important to link to data sets outside organisational boundaries. This is evident in use cases such as supply chain management and pre-sale where data from public domain significantly enrich internal data sets. We, therefore, see a mixture of public, partner, and proprietary data complicating data transparency and accessibility. To our best knowledge, none of the existing efforts have addressed the process driven requirement unique to enterprise data.

Inspired by the misalignment between the requirements of enterprise data and existing LD efforts, we carry out studies with real users to investigate how the linked data principles and concepts can assist customer account executives and team members when they need comprehensive and real-time access to internal and external data sets. We first elaborate on the differences (Section 2) between enterprise data and public data. Bearing these differences in mind, we discuss certain design considerations and the system architecture in the context of a customer information portal (CIP) project (Section 3). This is followed by three real-life use cases demonstrating the value of CIP (Section 4). We then discuss the lessons learnt (Section 5) and conclude the paper in Section 6.

2 Why corporate data is different?

As a collaborative and international effort, Linked Data has gained good publicity in the academic and to some extent the public sector communities [2]. With all the exciting success stories of massive development effort in linked data projects, we now face the question regarding the applicability of "linked data" principles in the corporate sub-domain.

It is evident that enterprise data lend themselves as both an opportunity and a challenge. On the one hand, enterprise data have well-defined boundaries with rigid protocols regulating the transition across the boundaries. They present

[1] http://www.thegoldensource.com/component/attachments/download/36

less heterogeneity and diversity comparing with public data from the Internet. Furthermore, even though divided into different departments focusing on different areas, modern enterprise normally reinforces a common corporate culture, which fosters a common, shared corporate "language", i.e. domain vocabulary. In many cases, this vocabulary may even impinge on communities beyond corporate boundaries. A good example is the jargons and acronyms used by the global SAP customer network. Finally, enterprise data are normally well documented and preserved either formally as white-papers, official publications, etc. or informally in e-mails, task log data, wiki pages, etc. Different from public data from the Internet, enterprise data are normally subject to internal review, for the purpose of auditing and quality control, or, at minimum, created with good intentions to fellow workers. We can therefore enjoy a much smaller amount of noise compared to general public data.

On the other hand, enterprise data still present significant research challenges. Simply connecting different islands together in an archipelagic data landscape will not be convincing enough. "Process-driven" is a unique feature that one has to bear in mind when migrating the LD concepts into the enterprise environment. Meanwhile, the relatively small size and homogeneous nature of enterprise data suggest that superficial connection of in-house data may not generate a good enough business value. In many cases, internal data alone is not sufficiently rich to satisfy diverse business requirements and thus incorporating external data sources is inevitable. How and what data should be exploited, however, can only stem from real-life scenarios. It is, therefore, salient to align with end users to understand and demonstrate the "return of value" of linking enterprise data. We will discuss these points further in this section.

2.1 Process-driven

Currently, there are roughly two approaches to fulfill the LD vision, namely data-driven and community-driven. Data-driven starts with a set of core data and tries to establish connections with as many relevant data sets as possible to emerge patterns not possible to individual data sets alone, while community-driven tries to fulfill the data request of a community, e.g. movie fans, gene researchers, etc. Both approaches may find themselves struggling in the enterprise environments.

Data management in an enterprise environment always has one ultimate purpose: improving the efficiency of a company's core business. However, linking data together does not necessarily mean that the implications, with which data are generated and leveraged, automatically become explicit to those linked in. The business implications can only be understood when we situate data into their original business processes. Therefore, different from the dominant data-driven nature when linking data from the Internet, linking enterprise data demonstrates a strong process-driven characteristic. That is the connections among data can and should only be revealed within the context of business processes where such data are consumed. Similarly, links among data should not be arbitrarily created independent from business processes. Aligned with companies' mission-critical

businesses, linking enterprise data from both inside and outside a company can be rightfully leveraged in decision making.

We would also argue that the successful community-driven approach (c.f. [16]) is not strictly applicable in the enterprise environments. Such communities are normally self-organised by common interests and loosely regulated, mainly self-disciplined. Misconduct and inappropriate behaviours do not result in the same consequence as in enterprise environments. Meanwhile, members of the community are organised in a rather flat structure with equal access to resources, which is a freedom that is not valid in companies. To the best interests of employees, taking a process-driven approach to data linking therefore can guarantee the alignment between personal interests and organisational policies.

2.2 External data

At the beginning of the CIP project, our intuition was that in an international organisation, the internal data alone should present enough challenges and offer sufficient business value for the LD paradigm. This was proved partially wrong during the discussion with end users. Internal data, although distributed across a large geographical region, are well-regulated and to some extent aligned attribute to common corporate cultures and operational regulations. Making internal data compliant with LD principles is more an organisational and motivational effort than a technical challenge.

The real challenge comes from defining good scenarios that can meaningfully link data together to satisfy needs of everyday businesses. For such a purpose, internal data can only tell part of the story. Very frequently, employees refer to external data sets for essential information that is not available from within the corporate boundaries. For example, the latest volcanical ash disturbance resulted in changes of project execution, project management decisions, and customer relationship management; natural disasters (e.g. the earthquake in SiChuan Province, China) can lead to major changes in supply chain management. The importance of such external data will not be fully demonstrated if they are not combined with internal enterprise data and consumed in real-time business decision making. The linking of public, partner and proprietary data should conform to the following guidance. External data should not interfere with internal ones. Where conflict observed, organisational protocols should be consulted to resolve the inconsistences. Meanwhile, it should follow existing organisational policies: this again points back to the process-driven aspect.

2.3 Personal space

The most controversial argument that we would like to put forward, which can be deemed against the total "openness" of the LD initiative, is that when linking enterprise data, the personal comfortable zone in data sharing should be respected. For organisations of different sizes, cultures, and structures, there is a long standing tendency of information *disintegration* attribute to a lack of trust in fellow workers, feeling of insecurity, and fear of disgrace [12]. We did not plan

to deal with such motivational issues. Rather we acknowledge the existence of such barriers and try to accommodate user requirements that stem therefrom. Obsering such a requirement allows users to more comfortably position themselves in data sharing initiatives. This is, however, done at the price of sacrificing fundamental LD principles to a certain extent.

3 Customer Information Portal

The concept of linked enterprise data is materialised in a pilot study that is meant to facilitate data integration and data sharing in a geographically distributed international organisation. When a company operates in more than one locations, it is not surprised to find different regional representatives approaching the same customer with different stories. The representatives sometimes are caught totally unprepared with questions regarding latest business and technical development and, even worse, regarding technical proposals and sales offers made from other units or even within the same units. A simple and effective remedy to such a problem is to create a portal for all the data concerning a customer. It can serve as a briefing tool for any one working on a customer so as to avoid the aforementioned embarrassment. We take advantage of the CIP project as a platform for understanding benefits and constraints of applying LD concepts in the enterprise environment.

3.1 Design decisions

During the definition of this pilot project, we try to address the unique characteristics of enterprise data (as discussed in previous sections).

Process-driven is given particular emphasis. Projected on design decisions, this implies the ability to answer "what data should be accessed by whom at what stage?". Based on business processes, one is prescribed to navigate the internal resources, employee profiles, and external data only specified in the business processes. Doing so will ensure that enterprise data are linked in line with organisational policies and strategies. Business processes can be standard ones or created for personal needs using predefined building blocks. We provide a list of exemplary business processes that are modelled and executed using in-house software (e.g. SAP Netweaver BPM) due to practical considerations. The in-house software is well understood by all the end users that reduces the learning curve. Meanwhile, in order to ensure a smooth integration with internal data sets, we try to avoid unnecessary disturbance to the platforms wherein such data sets are used. The in-house business process management system offers adapters compliant with J2EE Connector Architecture[2] and thus can seamlessly integrate with Java-based semantic systems.

The privacy concerns are addressed by maintaining a clear separation between data sets that are available to everyone and those that are only visible

[2] http://java.sun.com/j2ee/connector/

to the selected few. When creating an online article, a new business process, or uploading a document, people can opt-in to share or not share such resources. Effectively this is tantamount to linking private data space with the public one. Regardless of whether or not the resources are made public, an excerpt is produced to inform others of the contents.

We try to accommodate the general LD principles as followings. Using URI for resource identification can be easily satisfied—all internal resources (including documents and people) are uniquely identifiable through URIs. When this is not the case, we annotate data sets with uniquely identifiable labels based on RDF-coded ontologies. Links among internal resources are implemented as ontology properties among annotated resources. For internal data, syntactic and semantic mismatch does not present as a problem due to the existence of well-defined common vocabularies normally exercised by large organisations. Semantic interoperability becomes more of an issue when linking to external data sets. We adopt a simple but effective solution: embedding a Wikipedia link in concept definition. For instance, the "Course" Wikipedia article (URI) is introduced as a super-concept of concept Training_Course. The benefit is seen in two aspects: explanation and alignment. With links pointing to Wikipedia articles, we can easily extract the natural language based explanation of a concept. This is, in many cases, the first paragraph of the article. This explanation can be displayed to human readers for better understanding of the concept. Nearly all end users find this helpful. On the other hand, Wikipedia serves as a good reference point for aligning external resources with internal ones, for instance, through DBPedia. For those that are not currently covered by DBPedia, we leverage existing ontology mapping tools [7].

RDF representation is used exclusively in the background. We would argue that any efforts to make the underlying RDF representation transparent to the end users are likely to create more questions than answers in an enterprise environment. The following observations underpin our contentions. A majority of the corporate users are not semantic-web minded. More precisely, they do not care whether the data provision is facilitated by traditional technologies or semantic technologies, as long as data are provided in a timely and accurate manner. Such end users are for instance executives, sales and pre-sale personnel, service support and human resource. Understanding semantic technologies is certainly not a competence that they intent to develop. Ironically, the end users who will benefit from linked enterprise data is likely to enjoy such benefits only when the semantic technologies totally disappear from the user interface. A direct design consequence is that we had to improve user experience through good visualisation techniques (*c.f.* [4]) and RDF adaptors for conventional programming languages (*c.f.* [15]) for intuitive RDF data manipulation.

3.2 System architecture

CIP is a multiple-layered data/information integration platform (see Figure 1). At the bottom, there is the Data Layer. We clearly distinguish data sources that are only available to internal users and those in the public domain due to

data safety and privacy concerns. We also differentiate data that are properly structured (e.g. databases), semi-structured (e.g. wiki pages, calendars, to-do lists, etc.), and un-structured (e.g. e-mails, blogs, and legacy documents).

Structured data from internal sources are mapped directly to the ontologies via for example manually/semi-automatically crafted D2RQ scripts[3]. Note that semi-automatically identifying correspondences between database schemata and ontologies is not a disadvantage. In our case, the internal databases are specialised for managing certain types of mission-critical data where consistence and stability is observed. We do not expect the schemata to be frequently updated/upgraded. Therefore, the DB2RDF mapping, once defined, has a knock-on effect on data migration. On the other hand, data stored in such databases capture critical information of the company's core business. In order to support sensible and accurate decision making, such data have to be faithfully presented. We evaluated several automatic database to ontology mapping toolkits and none of them produced satisfactory results. Human intervention and verification is inevitable and, we believe, is more cost-effective if introduced in the early stage of mapping. String similarity was leveraged to produce recommendations and based on our experience string similarity or a combination of its variants is by far the most effective method. We leverage DBpedia to align structured data from public domain. At this stage, structured public data exploited in CIP is mainly Wikipedia infobox presenting basic facts of key customers, the partners and competitors. Wikipedia can also provide semantic enhanced applications (*c.f.* [5]). We plan to investigate the applicability of such technologies in the next phase of this pilot.

Semi-structured data from both internal and external sources are processed in two stages. First, the structured part is extracted. For instance, the dates, locations and priority levels in Calenders are used to populate the ontology. The free-text contents of such semi-structured data sets are feed into a keyword extractor for shallow natural language processing. We use Gate [3] to create such extractors.

Processed data are stored in a semantic repository and are consumed by a business process management system residing in the integration layer. End users of the CIP do not assume equal privilege of internal as well as external data sets. What data sets should be linked is entirely decided by use cases and thus essentially driven by the business processes associated with the use cases. For instance, if the use case is to establish new sales opportunities, one needs to access potentially full customer engagement history in the Customer Relationship Management (CRM) data. On the other hand, if the use case is cost reduction, one focuses on product life-cycle management data, supply-chain management data, etc. Process driven is facilitated by providing predefined use cases at the personal landing page (Figure 2) of the CIP tuned against one's profile (role, area of working, professional responsibilities, etc.).

[3] http://www4.wiwiss.fu-berlin.de/bizer/D2RQ/spec/

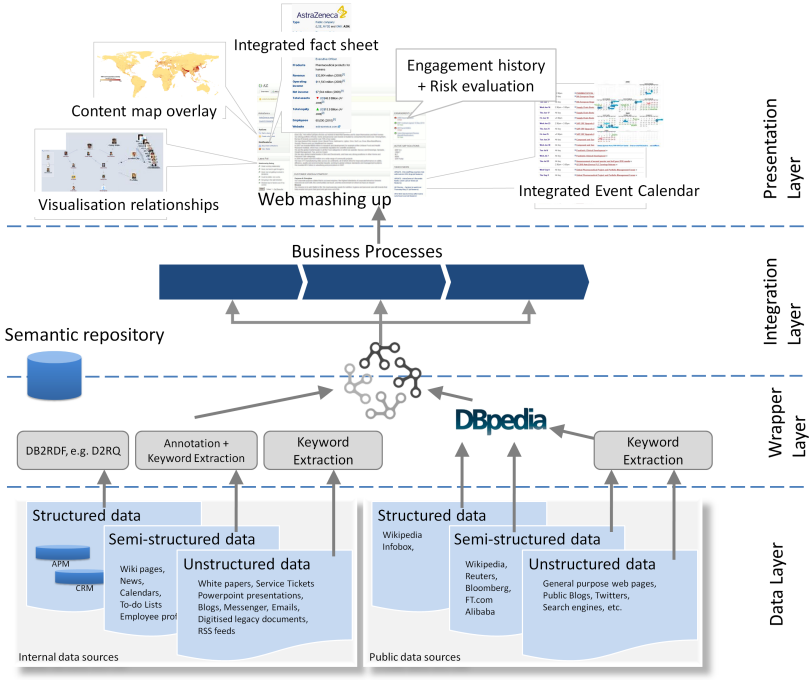

Fig. 1. Customer Information Portal Architecture

4 Use cases

The value of linked enterprise data can only be fully appreciated if it supports the real needs from end users. In the context of the CIP project, we carried out workshops with different stake-holders to elicit their requests. Out of the discussion with end users, we identify a list of interesting web mashing up scenarios. In this section, we elaborate on three exemplary ones.

4.1 Meeting the customers

Nearly all the modern sales 101 courses emphasise on "focusing on the prospect's point of view". Meeting with the prospects is always the best way to establish mutual trust and to understand their needs. The information portal facilitates this through linking external and internal data showing major events that the prospect is likely to participate and how events overlay with internal events (from e.g. internal event calendars).

Finding the prospect's events presents a technical challenge. We tackled this in the following steps. Firstly, we extract event information from the prospect's home page. Such pages can be easily found since almost all large enterprises maintain event calendars of various details. With little variants, entries in the

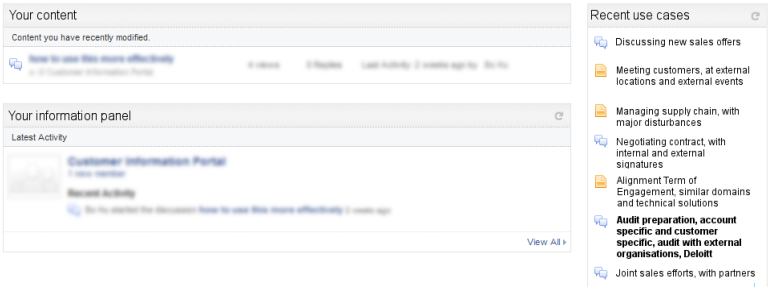

Fig. 2. User landing page

event calendars are normally in the form of ⟨Date, Type, (Location), Description⟩ and can be easily processed with text analysis tools. The second data source is the recurrent past events identifiable in the internal customer engagement record. This shows where positive contacts were established before and are likely to happen in the future. Keywords from the past events (e.g. titles) are used to search and retrieve the date and location of the next event in the series from the Internet. We also identified several event portals as auxiliary data sources. Such portals are domain specific and can only be identified on a per customer basis. For pharmaceutical industries, exemplary web portals include pharmiweb.com, pharmaceutical-int.com, etc.

Data from the above three sources are used to create instances of the CIP domain ontology. We define seven different event types, namely conferences, exhibitions, trade fairs/expos, media/press events, training courses, unconferences, and the unspecified, while Unspecified is used to collect those of unknown or unconcerned types. Equation 1 is fragments of event type Training_Course: where $\text{Course}_{\text{wpd}}$ refers to the corresponding Wikipedia article via its URL. Denoted in Turtle notation[4], an event instance is as follows:

```
<http://www.***.com/EventsCalendar.mvc/EventDetail/32831>
      rdf:type #Training_Course ;
      rdfs:label "GCP"^^xsd:string ;
      #starts "07/06/2010"^^xsd:string ;
      #ends "07/09/2010"^^xsd:string ;
      #location "Costa Mesa"^^xsd:string ;
      #participants #NovoNordisk , ... , #Pfizer ;
      ...
```

$$\begin{aligned} \mathsf{Training\_Course} \equiv\ & \mathsf{Event} \sqcap \text{Course}_{\text{wpd}} \sqcap =_1 \mathsf{starts.xs{:}date} \sqcap =_1 \mathsf{ends.xs{:}date} \\ & \sqcap =_1 \mathsf{location.xs{:}string} \\ & \sqcap \forall\, \mathsf{participants.Organisation} \sqcap \ldots \end{aligned} \tag{1}$$

[4] http://www.w3.org/TeamSubmission/turtle/

We used simple domain heuristics to recognise types of events. In majority of the cases, types of events are either explicitly specified (e.g. in AstraZeneca event page), indicated in the titles (e.g. names of conferences), or given in event description. For instance, in the above example, we looked for keywords such as "course", "educational", "learning", etc. Such keywords are manually compiled and so far have produced good results: an F-measure value of 64.77% with respect to the six named event types. This value is obtained by comparing to the classification from human experts.

We use web crawlers to regularly harvest events from the Internet and populate the RDF repository accordingly. End users can then choose from several visualisation options: a list of next events, map overlay of event locations, and conventional calendars. A typical map overlay (implemented with GoogleMap) is illustrated in Figure 3, showing the location of events and one's current location (marked with "L" and retrieved from the employee's directory).

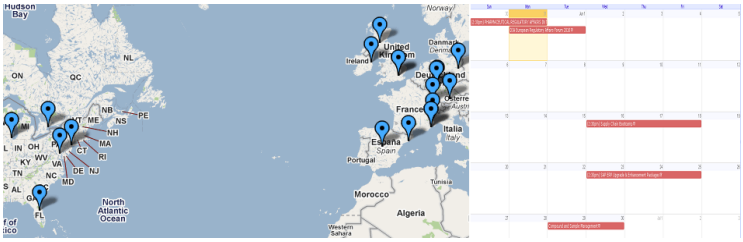

Fig. 3. Visualisation of events

4.2 What has happened to the project?

Public news can lead to major decisions on the customer relations and thus impinges on account activities. In the CIP project, we compile multiple news sources and present to the end users in a coherent story.

We source news from mainly the following categories: i) internal news bulletin, ii) press releases from targeted customer, and iii) public news websites, e.g. FT.com and Bloomberg.com. Harvesting from the first two categories is straightforward and is constrained by the role of the requestor in the organisation. The third requires fine-tuning. News from public domain can be easily retrieved with the current capacity of general web search engines. We, however, would like to go one step beyond simply retrieving and presenting the news to the end users. We have done this by combining customer specific news together with other major events coinciding at the same location. In many cases, apparently irrelevant events happening in the same geographic area might significantly influence on the decision making regarding sales and long-term customer relations. Therefore, keeping end users up-to-date is crucial. This is done as follows:

1. identify customer's headquarters and important branch offices through internal customer profile,

2. use extracted locations to search in public data sets for major events (denoted as $\mathcal{E}$), e.g. festivals, natural disasters, urban uprising, etc.
3. use the geographic scales of $\mathcal{E}$ to analysis whether known partners or competitors with respect to the aforementioned customer are affected.
4. federalise $\mathcal{E}$ with news stories from internal and external sources.

Strategic locations of an organisation can be found from the organisation's homepage with shallow text analysis and simple domain heuristics. In some cases, this will require manual extraction for new customers. Deciding the scale of major events is simply done with shallow text analysis to extract location names. The connections between events and organisation locations is done via ontology properties. In CIP, news is introduced as a sub-concept of Event and is linked to organisations through location property.

Summarisation of collected new stories are presented to the end users ranked by significance. So far the best news summarisation technique is simply extracting the first paragraph of the news article based on the observation that the baseline algorithm, extracting the first n sentences, has outperformed most of the "smarter" algorithms [11].

4.3 Where are we with the customer?

Customer accounts are in different stage of maturity. Moreover, one customer can be of different maturity with respect to different technical solutions. The content of the customer information pages should reflect such a diversity and put emphasis on different aspects accordingly by way of page layout, highlighting, etc. For instance, for a potential customer in pharmaceutical industries, the information page can focus on the solutions of competitors, key facts from similar customers, rules of engagement, etc. that will facilitate smooth initial contact of the account team. The emphasis will for example dynamically change to presale, sales, and supports according to the status of the account. This is guided by high level business processes of general customer engagement.

Meanwhile, we support linking data based on more specific business processes for real-time decision making. For instance, a customer-facing project, P, may be divided into several tasks each having milestones and checkpoints. Team members working on P use the dedicated customer page for keeping up with the progress of the project. It could be the case that the news of recent volcanic ash cloud raise concerns of potential disturbance to air-travel that can in turn impinge on project execution coinciding with the affected areas. Task managers can then use widgets on the CIP to adjust progress indicator (Figure 4(a)). The impact of such a change is two-fold: the disturbance can be propagated along task dependency links (through ontology properties) and cause the status of other tasks to change accordingly; management will be informed if the effects reach a certain level. Semantic technologies can facilitate such a scenario through modelling and reasoning of task dependency and the alignment between tasks and (news) events (as illustrated in Figure 4(b)). It is evident that the connections are established by extracting locations from external news stories, which

are then mapped to the locations of customer organisations. At this moment, connecting news with projects cannot be fully automated. News that has potential to impinge projects are first crawled from selected news agencies (normally as regional headline stories). Harvested news stories are processed and presented as potential threats to tasks/projects that can be affected. Project managers or task owners will be summoned to confirm or reject such connections. If he/she opts for accepting, the page content is then updated accordingly.

(a) Project monitoring (b) Linking project and news data

5 Discussions

Even though semantic technologies have been around for many years, the introduction of them into a well-established, high-tech organisation is not entirely hassle-free. In this section, we report some of the findings acquired when carrying out the pilot study. We believe our experience can be beneficial to those projects in similar settings.

5.1 Motivational barriers

One of the major barriers to successfully exploiting the LD concepts in an enterprise environment is the lack of incentive. In the past, we witnessed the ups and downs of similar initiatives (e.g. Enterprise 2.0 [9]). The initial excitement slowly fades off when the attention from management has been deviated to other businesses and when "try-out" has become work routine. Unless such tools become an integral part of one's daily working environment, it is not likely to maintain the same level of enthusiasm in the long term. In order to convince the end users, we reckon it is important to demonstrate the benefit from two aspects: showing added value in the business context and showing improved work efficiency.

Business context: "Providing better access to data" has been a cliche when persuading end users with semantic technologies. In an enterprise environment, this will have to be made tangible in terms of business applications. Our experience indicates that the presentation is as important as, if not more important

than, the underlying technologies. The merit of new technologies can only be delivered and well-accepted if they are presented in the end users' language. In our case, this is achieved by situating semantic technologies in the core business of an organisation. Linking enterprise data is then guided with the use cases derived from everyday work activities to avoid over exposure of data.

Improved work efficiency: We worked closely with the end users, customer-facing teams, to concretise the benefit of CIP in terms of saving on capital expenses and operational expenses. More specifically, we observe how employees work with the current technologies and how many short cuts they can enjoy with the help of semantic technologies. We estimated the time saved as per employee per customer with respect to mission-critical businesses and then summed up across the entire department. We also take a practical approach restricting "short cuts" to those that will cause minimum disturbance to employees' work routine and those that request only minimum investment in terms of labour and monetary resources. Able to demonstrate the improvement through financial gain increases the chance of obtaining management endorsement—this is a unique characteristic differentiate us from public social web-sites.

Individual participation: The barrier commonly seen in Enterprise 2.0 applications [9] were not observed in the CIP pilot. Collaborative and social network platforms have gained popularity in both public and corporate domains. The failure of certain initiatives does not deny their values but emphasises how the contents are organised, presented, and delivered. Again, we situate linked enterprise data in everyday work routine and bind it closely with a company's core businesses. We, therefore, experienced a very low level of reluctance from individuals and management.

5.2 Lightweight ontologies and incremental approaches

Introducing ontologies was proved to be a more difficult task than we originally had expected, even though alternative names e.g. "vocabularies" and "taxonomies" were used. The hesitancy towards ontologies is seen from the availability and cost of domain expertise, the threshold of comprehending representation formalisms, and the misunderstanding of ontology commitment. In practice, we took an incremental and application-driven approach. Instead of constructing the ontology once for all, we started with a selected application (news integration for example) for a small subset of the end users. The ontology is made modular with only the most essential entities. For each entity, we did not make effort to cover every aspect that defines a concept for conceptual perfectness, but only those that are necessary to enable the application.

We adopted the "Scrum" agile development principle with end users' involvement throughout the project (with different intensities at different stages). Our experience is that through small and manageable projects, we can demonstrate the merit of semantic technologies and thus establish the initial trust among end users. The "teaser" applications can then be gradually extended with one concrete and tangible improvement at a time. By doing this user commitment

and involvement are kept to minimum. In the CIP project, this approach was proved to be effective.

5.3 Minimum disturbance

The importance of minimum disturbance to existing infrastructure was address already (c.f. [1]). Our experience underpins such a conclusion and further extend it with two other principles. Firstly, introducing semantic technologies should not manifest at the user interface level. Secondly, new technologies should not alter established protocols.

Semantic technologies worked better when they have totally disappeared from the user interface, blending into everyday work environments. The value of linking enterprise data is best shown in areas where timely delivery of data is deemed important. It is exactly such areas where concerns were raised regarding the potential risk of not meeting key performance indicators while staff are trying to gain proficiency of new technologies. We confined the semantic technologies to the background and worked closely with end users on the foreground (interface). Meanwhile, we based our development on platforms (e.g. confluence wiki[5] and Jive[6]) currently in use to ensure a smooth learning curve. We observe the integrity and access control of all legacy data. For instance, even though we maintain a link to existing CRM database, what are shown to the end users depends entirely on his/her access right granted based on business processes. Meanwhile, we did not migrate legacy data. Instead, semantic annotation and mapping are established on-the-fly with the help of tools such as CROSI [7,13]. In general, it is impractical (if not impossible) to remodel all the legacy data. It is equally difficult to abandon existing relational database (RDB) implementations to switch entirely to RDF repositories. In fact, there was a major discussion regarding the benefit and disadvantage of RDF. The end users' main concern appears not on the change of mindset, but at the programming cost and extra learning efforts.

Obeying the Linked Data principles, however, should not compromise the intact of existing data repositories and established work processes of a company. Customer information is confidential with multiple levels of clearance. As an international organisation, majority of the data is continuously accessed by different departments across the globe in different time zones. Applying semantic technology should not alter existing data models causing disruption to normal business. Linking different data sets should not break existing access restrictions. In order to maximise the value of linked enterprise data, data sets with restricted access are handled as follows. We introduced a transition layer on top of raw data. Based on users' access privilege, the transition layer either populates ontology with data from such databases or presents a demilitarised summary of what data could have been accessed. Maintain a transition layer on top of existing data seems redundant. The extra cost, however, is marginal comparing to interrupted businesses.

[5] http://www.atlassian.com/software/confluence/
[6] http://www.jivesoftware.com/

Considering individual motivation, the same minimum disturbance requirements exist. The true value of semantic technologies can only emerge when a large amount of data is ready to be consumed which, in turn, relies on the willingness of involvement. Less disturbance to people's everyday work routine is likely to encourage their participation.

The requirement of minimum disruption also suggested that any development should be based on existing platform instead of introducing new ones. Fortunately, we were aiming at employee-only communities, for which modern organisations tend to impose less strict regulations, encourage employees to experiment the benefits of new technologies, and deploy at minimum some collaboration platforms [8]. Many of such platforms can be easily extended with web widgets encapsulating extended functionalities.

6 Conclusions

Linked Data has demonstrated noticeable value in the public domain [2]. Whether the same principles and the same outcomes are true in a "semi-closed" and proprietary domain has not been properly investigated. In this paper, we report on a pilot study carried out in an international high-tech company. Even though the project was not initiated as a proof-of-the-concept for semantic technologies, we discovered the real value of giving it a semantic touch. The advantages of doing this are not much different from the ones reported previously [1]. However, due to the characteristics of our domain, we made the following arguments that are unique to the studied domain. We would argue that linking enterprise data should be derived from processes faithfully reflecting the core business of a company. Deviation from this principle may lead to "yet another Enterprise 2.0" toy that does not bring real values to the company as well as its employees. The second principle emphasises on the interplay between external and internal data. Thus far, reaching a semantic consensus across different departments has not been a real challenge due to well-defined organisational boundaries and a common corporate culture, leading to well-understood common vocabulary. Counterintuitively, linking internal data sometimes can be made easier through references to external ones than solely based on internal links. Such a phenomenon might eventually encourage more companies to make their non-mission critical data sets available to the public to savor the payback. Finally, in many cases, linking enterprise data presents as less a technical challenge than a psychological one. How to motivate corporate employees to consume as well as actively contribute worth further investigation. We briefly discussed some of the motivational issues, for which we only gleaned the tip of the iceberg.

The crux of our further work is on performing larger scale evaluation with users invited from different departments and different geographic regions. It is also important to identify more application scenarios that can show values to a diversity of users including pre-sales, sales, education, technical support, etc.

Acknowledgements

This work is partially supported by the European Union IST fund through the EU FP7 MATURE Integrating Project (Grant No. 216356).

References

1. Harith Alani, David Dupplaw, John Sheridan, Kieron O'Hara, John Darlington, Nigel Shadbolt, and Carol Tullo. Unlocking the potential of public sector information with semantic web technology. In *Proceedings of the 6th International Semantic Web Conference (ISWC/ASWC2007)*, pages 701–714. Springer, 2007.
2. Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data - the story so far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22, 2009.
3. H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. GATE: A framework and graphical development environment for robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the ACL*, 2002.
4. Leonidas Deligiannidis, Krys J. Kochut, and Amit P. Sheth. Rdf data exploration and visualization. In *Proceedings of the ACM first workshop on CyberInfrastructure: information management in eScience*, pages 39–46. ACM, 2007.
5. E. Gabrilovich and S. Markovitch. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of the 20th IJCAI*, pages 1606–1611, 2007.
6. Rayid Ghani. Research challenges in enterprise information retrieval, 2008. available at http://videolectures.net/active09_ghani_rdekm/.
7. Y. Kalfoglou, B. Hu, D. Reynolds, and N. Shadbolt. Semantic integration technologies. 6th month deliverable, University of Southampton and HP Labs, 2005.
8. Gary Matuszak. Enterprise 2.0 - The Benefits and Challenges of Adoption, 2007. Available from http://www.kpmg.com/Global/en/IssuesAndInsights/ArticlesPublications/Enterprise-fad-future/Documents/Enterprise-2.0-The-benefits-and-challenges-of-adoption.pdf.
9. Andrew P. McAfee. Enterprise 2.0: The dawn of emergent collaboration. *MIT Sloan Management Review*, 47(3):21–28, 2006.
10. Bjørn Erik Munkvold, Tero Päivärinta, Anne Kristine Hodne, and Elin Stangeland. Contemporary issues of enterprise content management: the case of statoil. *Scand. J. Inf. Syst.*, 18(2):69–100, 2006.
11. Ani Nenkova. Automatic text summarization of newswire: lessons learned from the document understanding conference. In *AAAI'05: Proceedings of the 20th national conference on Artificial intelligence*, pages 1436–1441. AAAI Press, 2005.
12. Andreas Riege. Three-dozen knowledge-sharing barriers managers must consider. *Journal of Knowledge Management*, (3):18–35, 2005.
13. Roman Roset, Miguel Lurgi, Madalina Croitoru, Bo Hu, Magí Lluch i Ariet, and Paul Lewis. A visual mapping tool for database interoperability: the healthagents case. In *Proceeding of the 3rd CS-TIW Workshop*, 2008.
14. François-Paul Servant. Linking enterprise data. In *Linked Data on the Web Workshop at the 17th International World Wide Web Conference*, 2008.
15. Max Völkel and York Sure. Rdfreactor - from ontologies to programmatic data access. In *Poster session at the International Semantic Web Conference*, 2005.
16. Jun Zhao, Alistair Miles, Graham Klyne, and David Shotton. Linked data and provenance in biological data webs. *Briefings in Bioinformatics*, (2):139–152, 2009.