Utilizing Aging Public Sector Information

Per Myrseth[1], Jon Atle Gulla[2], Veronika Haderlein[1], Geir Solskinnsbakk[2] and Olga Cerrato[1]
[1] Det Norske Veritas (DNV), N-1322 Høvik, Oslo, Norway
{per.myrseth, veronika.haderlein, olga.cerrato}@dnv.com

[2] Department of Computer and Information Science, Norwegian University of Science and Technology, Trondheim, Norway
{jon.atle.gulla, geir.sols}@idi.ntnu.no

Abstract. The Public Sector Information EU directive (2003/98/EF)[1] about opening up public information is implemented in Norwegian law[2] and the Brønnøysund Registers Center (BR) offers the Register for Legal Entities as national master data. The register has been operated electronically since 1988. Understanding and interpreting the increasing volume of aging data is challenging. Important reasons for this are changes in terminology, immature semantic annotation, lack of definitions, increased distance between data owner and data user, changes in work processes and usage of data. The set of reasons are seldom explicitly or formally documented. Few studies and little tool development related to visualizing a merge of temporal data, records and ontologies in a long term perspective have been performed so far. In this paper we describe a semantic solution helping users to understand and interpret a primary set of time-stamped data (the Norwegian Business Register) in a concrete context. The changes in data are visually distributed on a timeline-oriented GUI. Secondary information, such as events governing the related organizations, classes from temporal ontologies and changes in regulations, are depicted on the timeline in order to facilitate users' comprehension of data even further. The set of secondary data is chosen based on the use of context-aware indicators, user input, semantic methods and ontologies. By simple usage of semantic technologies we demonstrate how to open up national master data as Public Sector Information, and how important the supplement of secondary data is when interpreting aging information.

Keywords: ontology evolution, temporal data, long term records management, long term information governance, public sector information.

1 Introduction

The volume of stored data has been steadily increasing over the past decades and the amount of aging and obsolete data has been rising as well. Part of this aging data is still in frequent use in ongoing business work processes, data warehouses, various system integrations and in interchange between organizations. This challenge initiated several best practice initiatives and research studies in records management. The LongRec [18] project is one such research initiative, looking into the aspects of reading, finding, trusting and understanding data in a long term perspective.

The Public Sector Information EU directive (2003/98/EF) about opening up public information is implemented in Norwegian law, and the Brønnøysund Registers Center (BR) offers the Business Register as national master data [16]. In EU the Public Sector Information initiative is seen as a major contributor to an open information society that may result in large economic impact [8]. Semicolon [25] is a project in which the goal is to develop and test ICT-based methods, tools and metrics to obtain faster and cheaper semantic and organisational interoperability both with and within the public sector. This paper is based on results from both the LongRec and the Semicolon projects.

There are two different time perspectives involved in the challenge mentioned above. First, we need to asure that the already existing data can be used also in the future. Second, we need to prepare ourselves for the long-term usage of data yet to be produced. The meaning of records and their intended use will change and evolve over time. Records produced at a specific point in time often

[1] http://epsiplus.net/ Public Sector Information (PSI) Re-use
[2] Offentleglova, http://www.lovdata.no/all/hl-20060519-016.html

describe some aspects of the lifespan of real objects/ referents. As time passes we will need to use the aging data and, hence, cope with a number of challenges with semantic implications.

To make records useful in a long-term perspective, we need to be able to align temporal versions of terms, concepts, instance data / records and context at large. Some illustrative examples of this challenge is given in [13]

A general statement of the problem can be formulated as follows: When you understand a set of data then you can evaluate if they are fit for purpose.

An important goal of this work is user empowerment, i.e. enabling knowledge workers to interpret aging primary data by relating it to and presenting it together with the relevant secondary data. Below, we describe how we have approached this goal in our pilot solution.

1.1 Our Approach

In this paper we describe a solution that provides interpretation support for a set of aging records, i.e. the primary data, by viewing secondary time-stamped data which are relevant. The challenges targeted in this solution are applicable to most Public Sector Information having a validity life span of some years or if data are relevant to be include in analyses of time series.

The solution is based on semantic technologies being used to structure the primary time-stamped data within the context of a given set of related secondary data. The primary set of data and the registered changes to it over time are visually arranged on a timeline-oriented GUI. Based on filters set by the users and the timestamp of the data, specific attributes of the primary data, such as the legal type of a business enterprise relevant secondary information is added to the timeline. Examples of such secondary information are changes in laws and regulations affecting the way enterprises can or might act in the national economy, general changes in language use, changes in the way public registry data is collected, changes in accounting rules, or changes in the responsibility between the board and company CEO. The selection of secondary data to be viewed is done by using context-aware indicators, user input, semantic annotations in the data, merging methods and ontologies.

The purpose of primary data and secondary data differs fundamentally.

- Primary data is the main object of interest in a given context, it is the information source to be interpreted and understood.
- The secondary data shall help the user to interpret and understand the primary data.

This combined view of primary data and secondary information on a timeline is enabled by a simple time-aware ontology. This ontology both models and holds corresponding instance data on secondary information which can be linked to the primary data through the common attribute of a timestamp.

The pilot is an information mesh-up with time as the primary GUI focus. In the pilot implementation we demonstrate how to assist the end user's navigation through a sea of information. The records are visually represented as items scattered on a timeline. Part of the solution is based on the open source software solution Timeline, which is part of the MIT project SIMILE [10].

2 Related Work

The related work for our research comes mainly from two main research areas: visualization of ontological data, and visualization of temporal data. We start this section with a brief description of different ontology visualization techniques, followed by a section on visualization of temporal data.

There are many different ways of displaying ontological data, such as tree-based display (e.g. [1, 2, 3]), graph-based visualization (e.g. [4, 5, 6]), and 3D visualization (e.g. [7]). In a tree-based visualization the data is viewed as a hierarchy (such as the class-browser widget of Protégé [2]).

Graph-based visualization uses graph-structures to visualize the ontology data. Jambalaya (see [6]) is a tool for Protégé, which uses a combination of tree visualization (from Protégé) and graphical visualization. OntoViz [4] is a plug-in for Protégé which uses the GraphViz [9] visualization toolkit to visualize the ontology in terms of a graph.

Visualization of temporal data is not new. The Simile Project [10] is an open source project aiming at giving the user tools to utilize their data more effectively. Two of the widgets from the Simile project are especially interesting, Timeplot and Timeline. Timeplot enables the user to plot temporal data and overlay temporal events (e.g. WW1) on the plot, letting the user understand the data in the context of events. Timeline lets the user create an interactive timeline with temporal events. Another interesting application is GapMinder [11] which is also used to visualize temporal data. Finally we would also like to mention the Google News Timeline [12]. The user can search for news articles, and Timeline organizes the search results according to a timeline. With our focus on Public Sector Information we are interested in solutions like Exhibit, Fresnel and Anzo on the Web. These solutions are compared in [23]. This comparison is done by the vendor of Anzo on the web, but it illustrates a progression in tool support.

Even though there is some interesting development; there is a lack of proper tool support for visualizing data from temporal ontologies.

Whereas Google Earth builds its mesh-up on geographical data and themes, and articles like [21] combines spatial and temporal focus on themes, we limit our focus to time and themes.

3 Temporal Ontologies

In this research a temporal ontology is used to enable the visualization of temporal data associated with instances of an ontology. The temporal aspects of the instances may be split in two: (i) a single point in time; (ii) an interval of time. A temporal ontology is an ontology that lets the user model temporal aspects (points in time and time intervals) of the underlying data of the domain.

In our project, the temporal ontology is based on a regular OWL ontology created by following standard design guidelines. Since our approach from a semantic technology point of view is a light weight solution we choose not to use Owl Time [24].

Converting the ontology into a temporal ontology is done by additionally modeling time into the ontology. This is done by creating two data type properties, *startDate* and *endDate*. These two properties are applicable to all instances of all concepts of the ontology, so that we can model the time aspect properly. Further we note that the properties of the ontology do not have any temporal aspects related to them. Our main purpose is to use this temporal information found in the ontology to enhance the users understanding of the underlying data. The purpose is not to do advanced reasoning or ontology evolution analysis.

4 The Pilot Use Case

4.1 The Norwegian Register of Business Enterprises at Brønnøysund Registers Center

In this paper we describe a solution that provides interpretation support for users working with public registry records on Norwegian enterprises. The records have been collected by Norwegian public authorities over decades. For the last 20 years, the records have been maintained and time-stamped according to the information governance rules at the Brønnøysund Register Center (BR). They can be regarded as examples of national master-data and form the primary set of records for the pilot. The case study is based on the needs of long-term understanding and interpretation of these national master data and making them available as Public Sector Information.

Many private companies and public bodies use these master data on a daily basis. The most common type of request is to get a copy of the latest version of "Certificate of Registration". This certificate contains documentation of e.g. organization number (national identifier), type of enterprise, date of incorporation, name, business address, general manager, members of the board, auditor, power of procurement, etc.

The types of reports offered can be grouped in two different ways, (i) reporting the status of the registered data at a certain point in time and (ii) reporting the development of records for a given enterprise in a certain time period.

The intended user of the pilot is a knowledge worker who is either employed at the Brønnøysund Register Center or by some external organization that makes use of the national master data.

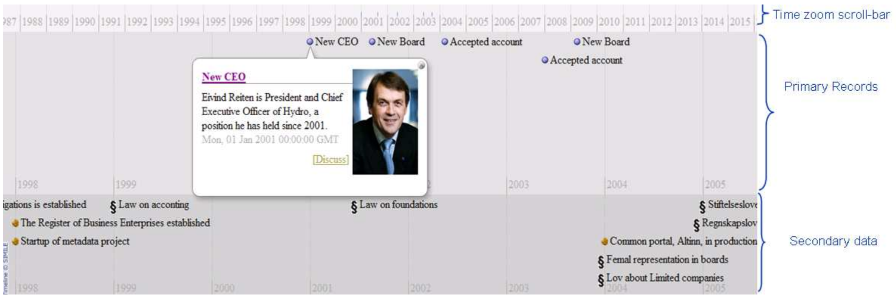
The *as-is* situation related to the interpretation of aging records has been made: (i) There is no tool support for interpreting aging enterprise data and its historic context; (ii) There is no structured overview nor any lists of relevant secondary information; (iii) There is lack of implemented information governance policy for leveraging the implicit semantics of business enterprise records in the future, but there is ongoing work with an ontology repository called SERES [15] that may address this issue.

Currently, the semantics of business enterprises records are implicitly captured in (i) regulations and juridical practice; (ii) data base models; (iii) tools and systems for registering data in the Business Register; (iv) operational procedures; (v) the implicit knowledge of the employees at BR; (vi) code tables; and (vii) import and export formats.

4.2 Functionality

The pilot application provides an information service, in which information from the public registry is related to other external sources of data and presented along a temporal dimension. After the user has chosen a company, the resulting graphical user interface is split into several parts. These are (Figure 1):

1. A resulting web page, with title, explanations, links to related pages and the configuration of the timeline GUI, etc.
2. the timeline GUI, which is split into the following three panes (often called swim lanes):
   1. A scroll bar we configured with a zoom scale of approx 1:4 related to the two lower panes.
   2. The pane showing the primary records.
   3. The pane showing secondary data.
3. A pop-up window-functionality is trigged when the user clicks on one of the bullets in the pane. The pop-up window contains further information, pictures and links to external web-resources. See the example below where "New CEO" is represented as a hyperlink, leading to further information given on the company's web-site.


New CEO

Eivind Reiten is President and Chief Executive Officer of Hydro, a position he has held since 2001.

Mon, 01 Jan 2001 00:00:00 GMT

[Discuss]

Figure 1. The components in the user interface.

The example above is related to Norsk Hydro, a major Norwegian business enterprise. Our example with persons and his/hers roles in different companies, would place the records on changes of roles in the primary record pane, and a chosen set of secondary data in the secondary pane.

To build this solution we used Protégé 3.4 as an editor, Jena to manipulate the ontology and appending instances to the ontology and Sparql for querying the ontology. We developed some Java code , and used the Similie.mit.edu Timeline Ajax application.

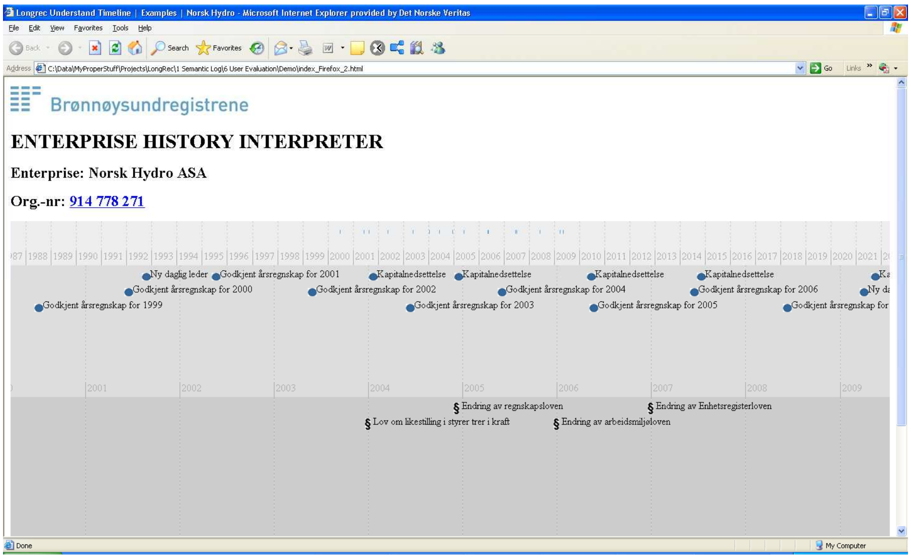
The use of the ontology has in several ways been highly useful in the pilot. One interesting effect is the ability to relate a chosen business enterprise to the set of relevant laws. In the enterprise-db there are several types of enterprises. Some of the types are limited companies (AS), public limited company (ASA), foundations etc. Another effect of the ontology is the ability to choose icons like the "law symbol" to visualize changes in laws, see Figure 1.

5 Pilot Evaluation

To complete the pilot process, we ran a qualitative, statistically non-relevant user study to assess the overall usefulness of the pilot and to detect functional holes and conceptual weaknesses. The user study consisted of structured in-depth user interviews with five representatives from first and second line customer support at BR.

In order to be able to evaluate the users' replies within the context of their pre-knowledge, we conducted a basic categorization of the users by three aspects: (i) the depth of their overall computer knowledge and mastery; (ii) their overall knowledge of the subject matter area covered in enterprise-db, i.e. legal aspects of Norwegian enterprises; (iii) and their familiarity with the dragging metaphor indicated by the hand symbol which the Timeline application depends on heavily. To achieve this kind of categorization we asked a set of test questions at the beginning of each interview.[3]

During the interviews, we used a version of the pilot that showed a limited set of historic information about the business enterprise Norsk Hydro ASA along with a limited set of information about changes in Norwegian laws and regulations[4]. Figure 2 shows a screenshot of this:



Longrec Understand Timeline | Examples | Norsk Hydro - Microsoft Internet Explorer provided by Det Norske Veritas

Address C:\Data\MyProperStuff\Projects\LongRec\1 Semantic Log\6 User Evaluation\Demo\index_Firefox_2.html

Brønnøysundregistrene

ENTERPRISE HISTORY INTERPRETER

Enterprise: Norsk Hydro ASA

Org.-nr: 914 778 271

Figure 2. Pilot setup used during user interviews.

The interview results can be summarized as follows:

(i) The fact that the metaphor of organizing information horizontally from left to right to indicate a timeline was immediately clear to the users, regardless of their familiarity with using computers and graphical interfaces, indicates that this is a useful way of organizing historic information.

(ii) Arranging primary data on a timeline is useful for a broad group of users at BR and probably also be useful to BR's customers directly. However, the importance of secondary data related to the information in enterprise-db depends heavily on the type of tasks a user has to perform.

(iii) In addition to the purely horizontal presentation of a sequence of events in both primary and secondary data, the presentation of primary and secondary information that was valid at a certain point

[3] To assess the degree to which the individual user is familiar with using computers, we asked whether the users had computers at home and whether they previously had installed any kind of software on them. If they had so, we assumed they had good knowledge of computers. To assess the users' knowledge of Norwegian enterprise legislation, we asked the users to give us a description of the difference between enterprises of the type "AS" (Limited Company) and enterprises of the type "ASA" (Public Limited Company). To assess the users' familiarity with the drag metaphor indicated by the hand symbol, we asked the users whether they had used Google Maps before.

[4] In this user study, we concentrated on records on one enterprise over a certain period of time rather than testing examples of both enterprise records and records about a certain person, as we think that the main flaws can be discovered with only one of these two record sets as well.

in time (snapshot-presentation) marks a clear need and requires further investigations as to how to prepare the underlying data so that this is possible. I the pilot, as it is today, the users have to summarize the primary and secondary events to the left of a certain point in time to get a current version of valid primary and secondary data.
(iv) In terms of pure user interaction, the time zoom bar and the hand symbol of the mouse pointer need a more thorough re-design in order to express their purpose and functioning more clearly.

6 Discussion and Future Research

Understanding and interpreting aging data that has been collected over a long period of time is challenging due to a number of reasons. Our approach to this challenge is centered around the focus on primary and secondary data by use of a temporal ontology with time stamps on all instances. This gives us the possibility to establish a graphical view of data using commonly metaphors that are familiar to most users.

We have identified some challenges we believe will influence the success of our solution. The most important ones are: (i) lack of APIs and access to open sources; (ii) lack of common identifiers across sources; (iii) no common ontology across sources; (iv) lack of common representation formats for data; (v) Limitations in GUI design and navigation capabilities handling large number of events in the Timeline GUI.

In spite of the challenges we are optimistic about the opportunities provided by design principles for Linked Open Data [19]. These principles are recommended to be used by Public Sector Initiatives [17]. Mesh-ups of a variety of sources of the user's choice are feasible and help us relate information in an intuitive manner. Furthermore, such a mesh-up can view primary and secondary data from different views like an N-dimensional information cube.

The feedback from the evaluation done at BR and ad-hoc feedback from colleagues within the computer science field are encouraging so far. In the long run we hope we can have a situation where the temporal ontology can be one component in an architecture where e.g. agents are used to interpret primary records with the help of secondary data.

The response of the pilot in the Norwegian market has been positive and information providers like Proff [20] has shown interest in the approach and solution. In November 2009 BR decided to allow the Semicolon project to open up the Norwegian Central Coordinating Register for Legal Entities as a Linked Open Data source. This work is now in progress.

Future development of the pilot would include the search for Linked Open Data to our primary or secondary data to improve the usability of the pilot. This is a challenging task and introduces several interesting questions. Is the secondary data trustworthy (security), is the data complete enough, is the quality satisfactory, what risk and liabilities would BR take by using data from sources outside their control, as a service provider: what liabilities is BR exposed to, etc. This also involves organizational and legal issues that are outside the scope of the pilot itself.

7 Conclusions

This paper suggests a solution for Public Sector Information improving the way users can understand and interpret aging data by the use of a temporal ontology and timestamps on instances. Based on a preliminary user evaluation, tests of a technical solution and related models we describe a set of opportunities and obstacles. The pilot's temporal view of primary data and secondary data is new and the user evaluation indicates that users do understand this way of merging and aligning data. Success criteria mentioned are: access to relevant secondary data and the ability of the semantic solution to filter and view a relevant set of secondary data. Other secondary data will often be other sources of Public Sector Information. In a world with increasing focus on public sector information, Semantic Web activities like Linked Open Data, the time seems mature to follow up the ideas described in this paper.

Acknowledgements

This research was carried out as part of both the LongRec project (project no. 176818/I40) and the Semicolon project (project no. 183260/S10), both funded by the Norwegian Research Council of Norway. The participants from the Brønnøysund Register Center have been Jostein Dyrkorn, Dr. ing. Even Thorbergsen, Alyass Muhammad, Per Fjelde and Tor Skjørdal.

References

1. S. Bechhofer, I. Horrocks, C. Goble, R. Stevens. OilEd: A Reason-able Ontology Editor for the Semantic Web. Joint German/Austrian Conference on AI: Advances in Artificial Intelligence, LNCS 2147, 2001.
2. Protege Ontology Editor. http://protege.stanford.edu/
3. N. Noy, M. Musen. PROMPTDIFF: A fixed-point algorithm for comparing ontology versions. Proceedings of National Conference on Artificial Intelligence (AAAI), 2002.
4. OntoViz. http://protegewiki.stanford.edu/index.php/OntoViz
5. RDFGravity. http://semweb.salzburgresearch.at/apps/rdf-gravity/
6. M. Storey, M. Musen, J. Silva, C. Best, N. Ernst, R. Fergerson, N. Noy. Jambalaya: Interactive visualization to enhance ontology authoring and knowledge acquisition in Protégé. Proceedings of the International Workshop on Interactive Tools for Knowledge Capture, 2001.
7. A. Bosca, D. Bonino, P. Pellegrino. OntoSphere: more than a 3D ontology visualization tool. SWAP 2005, the 2nd Italian Semantic Web Workshop, 2005, CEUR Workshop Proceedings, online http://ceur-ws.org/Vol-166/8.pdf
8. COMMISSION STAFF WORKING DOCUMENT, Accompanying document to the COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS, on the re-use of Public Sector Information, – Review of Directive 2003/98/EC http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=SEC:2009:0597:FIN:EN:PDF
9. Gansner, E. R., North, S. C.: An open graph visualization system and its applications to software engineering. Software – Practice and Experience, 30 (11), 2000.
10. Simile Project, Semantic Interoperability of Metadata and Information in unLike Environments, MIT. http://simile.mit.edu/
11. GapMinder. http://www.gapminder.org/
12. Google News Timeline. http://newstimeline.googlelabs.com/
13. T. Mestl, O. Cerrato, J. Ølnes, P. Myrseth, I. Gustavsen.: Time Challenges – Challenging Times for Future Information Search. D-Lib Magazine May 2009. http://www.dlib.org/dlib/may09/mestl/05mestl.html
14. P. Myrseth. T. R. Christiansen, H. Gayorfar.: Preservation of semantic value - A study of the state of the art. DNV Report No 2008-0123. http://www.longrec.com/ResearchResults/Pages/StateOfTheArt.aspx
15. SERES project. http://www.brreg.no/samordning/semantikk/
16. Central Coordinating Register for Legal Entities, http://www.brreg.no/english/registers/entities/
17. Linking Open Data, a W3C community project. Started as an activity under the Semantic Web Education and Outreach (SWEO) Interest Group. http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData
18. LongRec project. www.LongRec.com
19. Tim B. Lee, Design issues for linked data. 2009/06/18 http://www.w3.org/DesignIssues/LinkedData.html
20. www.Proff.no supplying value added information to Norwegian enterprises.
21. Sheth, A., Perry, M.: Traveling the Semantic Web through Space, Time , and Theme. IEEE internet computing March/ April 2008. http://knoesis.wright.edu/library/download/Sheth-Perry-IEEE08.pdf
22. Berners Lee, T.: The next Web of open, linked data, TED, http://www.ted.com/index.php/talks/tim_berners_lee_on_the_next_web.html
23. Three Approaches to Lenses for Web 3.0 Applications: A Survey. Presentation at Semantic Technologies San Jose, 16. June 2009. Jordi Albornoz Mulligan. Cambridge Semantics.
24. Owl Time Ontology. http://www.w3.org/TR/2006/WD-owl-time-20060927
25. www.semicolon.no