Utilizing Ageing Information

Per Myrseth[1], Jon Atle Gulla[2], Veronika Haderlein[1], Geir Solskinnsbakk[2] and Olga Cerrato[1]
[1]Det Norske Veritas (DNV), Oslo, Norway
[2]Norwegian University of Science and Technology, Trondheim, Norway
per.myrseth@dnv.com
jon.atle.gulla@idi.ntnu.no
veronika.haderlein@dnv.com
geir.sols@idi.ntnu.no
olga.cerrato@dnv.com

Abstract: In public administrations, the volume of stored data has been increasing over the past decades, leading to a growing amount of ageing information. Part of this ageing information is still in frequent use in ongoing business processes and data warehouses. However, one concern is the meaningful and trustworthy use of this ageing data in present applications. It is challenging to interpret and use this data correctly over time. For eGovernments, this challenge is overwhelming. Important reasons for this are: an increasing pace of changes in terminology, work processes and use of data; immature semantic annotation of data, lack of temporal ontologies, and lack of automation. In this paper we describe and evaluate a solution for visualizing data from multiple data sources on a timeline-oriented user interface. The solution is made to cope with the challenges of understanding ageing enterprise data. To improve users' ability to properly understand the enterprise data in its historic context we supplement the enterprise data with historic context data. The solution is based on a case study on Norwegian national master data on business enterprises. These data have been collected by Norwegian public authorities over decades. For the last 20 years these data have been stored electronically and time-stamped. To our knowledge the pilot's temporal view of primary data and secondary data is new. The user evaluation and other feedback indicate that users do understand this way of merging and aligning data, and find it useful.

Keywords: temporal mash-up, temporal ontology, long-term information governance, public sector information, visualization

1. Introduction

1.1 The overall challenge

The volume of stored data has been steadily increasing over the past decades in both the private and public sector; so has the amount of ageing information. Part of this ageing data is still in frequent use in ongoing business work processes, data warehouses, and various system integrations between organisations.

Within the framework of the pilot study described here, we define the term *ageing information* as data that will be used for 20 years or more. There are two different time perspectives involved in utilizing ageing information: First, we need to assure 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. It is the first perspective on time that is in focus in this study. The meaning of data and their intended use change and evolve over time. Records or data produced at a specific point in time often describe some aspects of the lifespan of real objects/referents. The context knowledge about these referents might be present in the users' minds at the point in time in which the information is created, but weakens as time passes by. Thus information that is available to us today, but that was created some 20 years ago is not as readily understandable as information that was created yesterday, since we do not have immediate access to the necessary context knowledge. Some illustrative examples of this challenge are given in Mestl 2009. Based on this, a solution that helps users understand ageing information will have to provide some support to reconstruct its context knowledge, and hence, enrich the primary information.

Within the public sector, this problem becomes particularly relevant when seen within the context of the Public Sector Information EU directive, 2003/98/EF. This directive requires public authorities to open up public information for reuse by both public and private agents. With multiple users re-using public information across service domains and across long time-spans, the problem of interpreting ageing information is multiplied. Thus, solutions that allow the correct interpretation of information independent of the time or space it was created in, become increasingly important.

In Norway, the Public Sector Information EU directive is implemented in Norwegian law. Accordingly, the data held by the Brønnøysund Registers Center (BRREG), a national public authority whose mandate it is to collect several sorts of publicly relevant information, is subject to this directive. Thus, the Register of Business Enterprises (Business Register) collected by BRREG is offered as national master data. The challenge that BRREG is experiencing today is the fact that this Business Register information is evolving over time. Change in information could be e.g. that an enterprise changes its CEO, but also the responsibility and duties of being a CEO evolves.

BRREG has an obligation to provide information support to the public in understanding the information held in the Business Register. However, as of today, case workers at BRREG's first line support do not have any useful support tools to help them view the Business Register data and its changes in a time perspective. Currently, case workers have to collect this type of information from heterogeneous external sources, without proper support to arrange the various pieces of information in a meaningful manner. It is this specific problem of merging distributed information sources that the pilot described here is addressing.

1.2 Our overall approach

At the core of our approach lies the idea of providing interpretation support for a set of ageing records by aligning it with relevant, time-stamped context information on a time-line. The main target group for this solution is caseworkers at BRREG, but we believe that the challenge targeted in this solution is relevant to many types of data in both the public and private sector.

In our approach, we distinguish between 2 types of data:

- *Primary data* – this is the information that the user is trying to interpret correctly in its historic context. It is the information that is in the user's primary focus and is thus called *primary*.
- *Secondary data* – this is the context information that helps the user interpret the primary data in its historic context. It is supporting information and is thus called *secondary*.

In the specific case of our pilot study, the Norwegian Business Register at BRREG represents the primary data. This data could be viewed as national master data, and is used in many different processes and IT-systems. As the supporting secondary data, we have chosen (i) changes in Norwegian laws and regulations for how to run certain types of businesses and (ii) eGovernment milestones including changes in the way BRREG has been registering enterprise information over time.

The *as-is* situation related to the interpretation of ageing records is in short as follows: (i) There is no tool support for interpreting ageing Business Register data and its historic context; (ii) There is no structured overview nor any lists of relevant secondary data; (iii) There is a lack of implemented information governance policy for leveraging the implicit semantics of the Business Register.

Currently, the semantics of the Business Register 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 BRREG; (vi) code tables; and (vii) import and export formats.

2. Related work

Related work for our research comes mainly from two 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. Bechhofer 2001, Protege, Noy 2002), graph-based visualization (e.g. Ontoviz, RDFGravity, Storey 2001), or 3D visualization (Bosca 2005). In a tree-based visualization the data is viewed as a hierarchy such as the class-browser widget of Protégé. Graph-based visualization uses graph-structures to visualize the ontology data. Jambalaya (Storey 2001) is a tool for Protégé, which uses a combination of tree visualization (from Protégé) and graphical visualization. OntoViz is a plug-in for Protégé which uses the GraphViz (Gansner 2000) visualization toolkit to visualize the ontology in terms of a graph.

Visualization of temporal data is not new. However, the Simile Project 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 in a bar-chart manner, letting the user understand the data in the flow of events. Timeline lets the user create an interactive timeline with temporal events. Another interesting application is GapMinder (GapMinder) which is also used to visualize temporal data. Finally we would also like to mention the Google News Timeline (Google News Timeline). The user can search for news articles, and the search results are arranged according to a timeline. Similarly interesting, particularly with focus on public sector information, are solutions like Exhibit, Fresnel and Anzo on the Web. These solutions are compared in Mulligan 2009. Even though this comparison is provided by a vendor of Anzo on the web, it is a useful illustration of the progression in tool support.

For the domain of geophysical data visualization, a research group at Statoil and the University of Bergen are developing methods and software tools supporting geologists to rapidly create 3-dimensional, animated illustrations of geologic structures and geologic evolution (Christian Michelsen Research 2010).

However, even though there is some interesting development in the fields of both ontology visualization and data visualization in general, there is a lack of proper tool support for visualizing data from temporal ontologies.

Whereas Google Earth builds its mash-up on geographical data and given topics, and initiatives like Sheth 2008 combine a spatial and temporal focus on given topics, there is little research on the combined visualization of time and topics alone. Our temporal mash-up is aiming at helping to close this gap.

3. The pilot solution

3.1 The data sets

As mentioned above, the data of the Norwegian Business Register represents the primary information in our pilot solution. Many private companies and public bodies use these master data on a daily basis. The most common type of request is to ask for a copy of the latest version of the "Certificate of Registration". A certificate for the enterprise Norsk Hydro contains documentation such as:

Organisation number: 914 778 271

Type of enterprise: Public Limited enterprise

Established: 02.12.1905

Name: Norsk Hydro ASA

Municipality: OSLO

Registered: 19.02.1995

General Manager: Eivind Reiten

Members of the board: nn, nn, nn

Procuration: nn, nn, nn

Auditor: nn

Domain of business: 24.421 Production of aluminium

We chose to use legal information as one set of secondary information. Persons from BRREG with historical and juridical knowledge made a spreadsheet containing important events related to laws and regulations relevant to the primary data. Information about these laws where captured manually

from the Norwegian information portal Lovdata. The purpose of Lovdata is to establish and operate legal information systems on a non-profit basis.

A supplementary set of secondary information is historical events at BRREG related to eGovernment milestones and the way Business Register data is collected and handled. These events were listed in a spreadsheet by persons from BRREG with knowledge of the history of the Business Register and the BRREG organisation.

One prerequisite for both the primary and secondary data is that they need to be time-stamped. Our chosen primary and secondary data have sufficient dates.

This selection of secondary information was based on the assumption that changes in laws and regulations will affect the way enterprises can or might act in the national economy, and that changes in eGovernment services may change the meaning of dates, roles, code lists, classifications etc.

3.2 Functionality in the pilot

The user performs the following steps when using the timeline visualization:

- Open the webpage with an input form asking for Organisational number
- Input of an organisation number, e.g. “914 778 271”
- Choose what kind of secondary data is of interest (drop down menu)
- Reads the resulting visualization

The main user interface is shown in Figure 1. The user interface is characterized as follows:

- It is a web-based user interface with title, explanations, links to related pages and the hidden configuration of the timeline GUI.
- It contains the timeline GUI, which is split into the following two panes (often called swim lanes):
- *One pane showing the primary information.*
- *One pane showing secondary information.*
- The panes can be scrolled horizontally.
- 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.

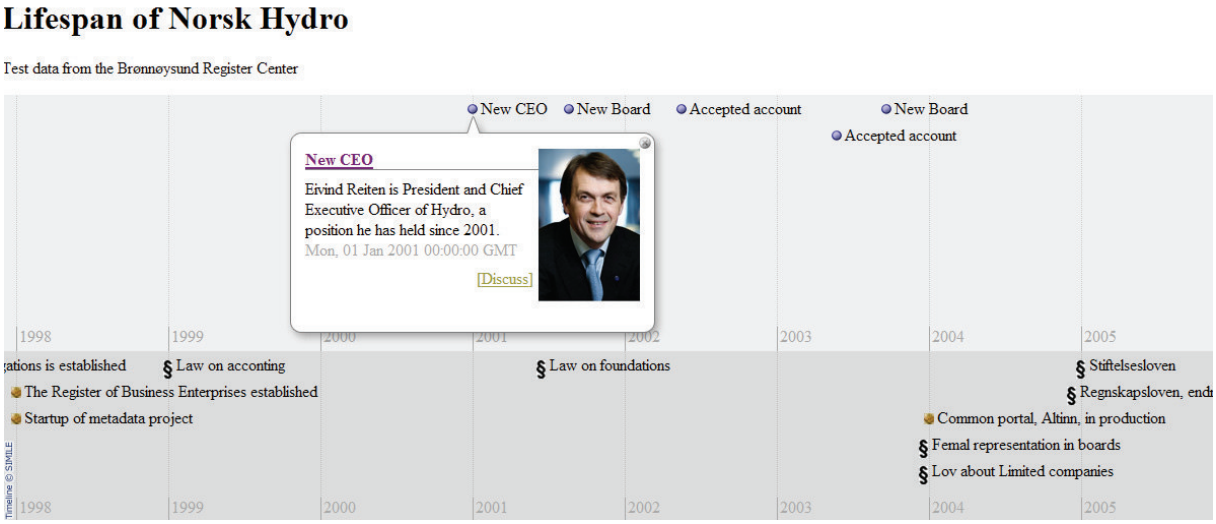

Figure 1: The components in the user interface

The example above is related to Norsk Hydro, a major Norwegian business enterprise. Based on the user’s needs, the user can choose which secondary data to study. In the example in figure 1, the user

has chosen to view all changes in laws in the given time-span (no ontology based filtering used). In addition, changes in eGovernment milestones are displayed as well.

An alternative view could be to switch the focus of the primary pane: Instead of displaying information about changes in the history of the company Norsk Hydro, one could also use this to focus on a given person and to view changes in his or her involvement in various boards of directors in various companies, accompanied by relevant information about changes in laws or economic figures in the secondary pane. Furthermore, a future solution could visualize relationships between entries in primary and secondary pane. But such a solution needs mechanisms to manually or automatically generate relationships between primary and secondary data at a more advanced level than what is demonstrated here.

3.3 The ontologies used

In our pilot, we use an ontology for two purposes. First, to enable visualization of heterogeneous temporal data along a timeline we implemented a *startDate* and *endDate* on every object in the ontology as data type properties. Second, we made a hierarchical ontology structure used for categorization. This categorization uses the OWL inheritance as hierarchical structure.

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. To limit number of visual elements in the user study we did not use endDate. This resulted in only use of point in time visualizations, and no intervals/ continuum was visualized.

In addition to modeling time, our ontology also has the role of a categorization ontology. We manually engineered a hierarchical categorization structure as an ontology in the ontology editor Protégé 3.4. It provides the basis for annotating instances of both primary and secondary data, and thus to link primary and secondary data based on a categorization attribute. This gives us the possibility to sort information about events along the time axis and also filter the secondary information based on categorization attributes. A simplified example is illustrated below:

Company types listed in the ontology:

- "Norwegian enterprises"
- *"Not for profit enterprises"*
- *"Public Limited enterprises"*

New laws and regulations:

- Annotated laws:
- *"Law for Accounting regulation" annotated as relevant to: "Norwegian enterprises"*
- *"Law for not for profit enterprises" annotated as relevant to: "not for profit enterprises"*
- *"Law for Limited enterprises" annotated as relevant to: "Public Limited enterprises"*

By using the primary data knowledge that Norsk Hydro is a "Public Limited Company", the ontology and the annotated laws, we are able to filter the relevant laws as follows:

- "Law for Public Limited enterprises" = true (based on direct match)
- "Law for not for profit enterprises" = false (based on no-match)
- "Law for Accounting regulation" = true (based on match by ontology inheritance)

In the specific case of our pilot, however, it has to be noted that since most public data is not semantically annotated, the annotation in our pilot had to be hand-coded. This situation raises a series of quality questions regarding how to annotate secondary data, especially if the consequence of errors in annotation is severe. This paper does not discuss this risk further.

The purpose of our temporal ontology is not to do advanced reasoning or ontology evolution analysis. Since from a semantic technology point of view, our approach is a light weight solution, we choose not to use OWL Time (Owl Time Ontology).

In our work we did not study temporal aspects of classes or temporal relations in our categorization ontology. Further our primary and secondary data have fixed annotations towards the categorization ontology. An elaboration on this topic can be found in Artale 2009.

To build this solution we used Protégé 3.4 as an ontology editor and MS Excel to store instances. We made some Java code to extract instances stored in excel files. Jena was used to merge the ontology and the instances. Java and Jena was used to perform SPARQL queries, filtering and to generate xml-documents in the Timeline format. Further we used the Similie.mit.edu Timeline Ajax application to perform the visualization along the timeline.

4. Methodology and pilot evaluation

4.1 User study

In concluding our pilot project, 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 BRREG. In order to evaluate the users' replies within the context of their pre-knowledge, we conducted a basic categorization of the users according to three aspects: (i) the depth of their overall computer knowledge and mastery; (ii) their overall knowledge of the subject matter covered in the Business Register, i.e. legal aspects of Norwegian enterprises; (iii) and their familiarity with the dragging metaphor indicated by the hand symbol on which the Timeline application heavily depends. To achieve this kind of categorization we asked a set of test questions at the beginning of each interview. 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 had previously installed any kind of software on them. If they had, we assumed that 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.

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. 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, since we think that the main flaws can be discovered with only one of these two record sets.

The interview results can be summarized as follows:

(i) The metaphor of organising 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. This indicates that it is a useful way of organising historic information

(ii) Arranging primary data on a timeline is useful for a broad group of users at BRREG and would probably also be useful for BRREG customers directly. However, the importance of secondary data related to the information in the Business Register 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, there is a clear need for a presentation of primary and secondary data that were valid at a certain point in time (snapshot-presentation). Further investigations are required as to how to prepare the underlying data so that this is possible. In 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.

4.2 Conclusions on practical feasibility

We also ran a structured workshop to conclude on the practical implications we had learned from the project. The conclusions were as follows:

- Access to, and formats of, secondary data from public sources like juridical corpora were not straightforward. We chose to copy and paste the links and data we needed to perform the demos and user evaluation. As listed earlier in this paper the types of data relevant as secondary data for our case could typically be data defining and describing changes in: terminology, code lists, laws and regulations etc. However, the identifiers used to link between sources need to be very strong and resistant to change over time, otherwise the trust in the links will be reduced or the links could be broken. This is why we focused on changes in laws and eGovernment milestones including changes in the way BRREG has been registering enterprise information over time.
- Few of the data sources we looked at existed in machine processable formats like XML, RDF etc. File catalogues on internal file-servers and RDBMS based information management systems seemed to be in frequent use. The secondary data that existed was governed by systems that handled annotations of start-date / end-date differently, and different ways of describing the meaning of their data (information models or ontologies) were used. The granularity of start and end-date could differ between days, month and year. Access to data was often not open and not for free. It was difficult to get a sufficient overview of who had what kind of information. We had to lower our ambition concerning what data we could use in our pilot, and we chose to manually annotate the secondary data we used in the pilot.

5. Discussion and future research

Understanding and interpreting ageing data that has been collected over a long period of time is challenging for a number of reasons. Our approach to this challenge is centred 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 a common horizontal timeline metaphor that is familiar to most users. We believe that this can be a useful way to support the correct understanding of information that has been created at a different point in time than that at which it is consumed.

However, if more initiatives start using the principles of primary and secondary data, there will be a need for coordination between the initiatives. Here, a common or coordinated distributed ontology for categorization of data will be very useful. Coordination is also needed for establishing interoperability between data sources that have time-stamps, provenance data, metadata, quality attributes and different levels of granularity. Furthermore, differences in how data sources use information modeling methodologies, representation formats and access control are obstacles to common and easy establishment of mash-ups for utilizing ageing information.

Mash-ups involve a certain degree of distance between information creator and information consumer, especially when we focus on reuse of information in the public sector. In many cases the creator is not aware of whom the consumer is. This could typically be the situation with open public sector information and will demand a series of different types of metadata and usage of semantic annotation. The consequence of potential errors in semantic annotation has to be discussed and decided upon.

We have identified some challenges we believe will limit the success of a mash-up like ours if they are not dealt with. The most important ones are: (i) little support for open APIs and limited access to free data sources; (ii) lack of common and stable 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 to handle large numbers of events in the Timeline GUI.

In spite of these challenges we are optimistic about the opportunities provided by the concept of mash-ups and the design principles for Linked Open Data (Lee 2009). These principles are recommended to be used by public sector initiatives (Linking Open Data). Mash-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 mash-up can view primary and secondary data from different perspectives like an N-dimensional information cube.

The feedback from the evaluation done at the BRREG and ad-hoc feedback from workshops and colleagues within the computer science field are encouraging. In the long run we hope to 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 to the pilot in the Norwegian market has been positive, and information providers have shown interest in the approach and solution. In November 2009 BRREG decided to allow the Semicolon project to open up the Business Register as a Linked Open Data source. This work is now in progress.

Future development of the pilot would include search for existing Linked Open Data sources as a supplement to our chosen primary or secondary data. This is a challenging task and introduces several interesting questions. Is the secondary data trustworthy (security), is it complete enough? Is the quality satisfactory, what risks and liabilities would BRREG as a service provider take by using data from sources outside their control, and what licenses should be used when publishing open public data? This also involves organisational and legal issues that are outside the scope of the pilot itself.

6. Conclusions

This paper suggests a solution for utilizing ageing data by improving the way users can understand and interpret multiple data sources through 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. To our knowledge the pilot's temporal view and merging of primary data and secondary data is new. The user evaluation and other feedback indicate that users do understand this way of merging and aligning data, and find it useful.

Success criteria mentioned are: Categorizing the secondary data, getting adequate access to the secondary data, interoperability related to quality attributes of the secondary data and provenance data. Further there is a need for collaboration between information providers of potential secondary data to establish a categorization ontology, principles for use of timestamps, identifiers and information modeling regimes etc.

Based on our case, secondary data will probably often be public sector information. In a world with increasing focus on public sector information and Semantic Web activities like Linked Open Data, the time seems ripe 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 Research Council of Norway. The participants from the Brønnøysund Register Center were Jostein Dyrkorn, Dr. ing. Even Thorbergsen, Alyass Muhammad, Per Fjelde, and Tor Skjørdal.

References

Artale, A. Franconi, E. Fondations of Temporal Conseptual Data Models (p 10-35). SpringerLink, Conceptual Modeling: Foundations and Applications. ISBN 978-3-642-02462-7. July 06, 2009.

Bechhofer, S. Horrocks, I. Goble, C. Stevens, R. OilEd: A Reason-able Ontology Editor for the Semantic Web. Joint German/Austrian Conference on AI: Advances in Artificial Intelligence, LNCS 2147, 2001.

Bosca, A. Bonino, D. Pellegrino, P. 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

BRREG, the Brønnøysund Registers Centre. http://www.brreg.no/english/registers/

Business Register, Central Coordinating Register for Legal Entities, http://www.brreg.no/english/registers/entities/

Christian Michelsen Research. Intuitive Seismic Visualization, http://www.cmr.no/index.cfm?id=274149 accessed january 2010.

Gansner, E. R., North, S. C.: An open graph visualization system and its applications to software engineering. Software – Practice and Experience, 30 (11), 2000.

GapMinder. http://www.gapminder.org/

Google News Timeline. http://newstimeline.googlelabs.com/

Lee, T.B. Design issues for linked data. 2009/06/18 http://www.w3.org/DesignIssues/LinkedData.html

Linking Open Data, a W3C community project. http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

LongRec project. www.LongRec.com

Mestl, T. Cerrato, O. Ølnes, J. Myrseth, P. Gustavsen, I,M.: Time Challenges – Challenging Times for Future Information Search. D-Lib Magazine May 2009. http://www.dlib.org/dlib/may09/mestl/05mestl.html

Mulligan, J. A. Three Approaches to Lenses for Web 3.0 Applications: A Survey. Presentation at Semantic Technologies conference, San Jose, 16. June 2009. Cambridge Semantics.