NATIONAL MASTER DATA AS 5 STAR LINKED OPEN DATA

Per Myrseth[1], Jens Kilde Mjelva[2], David Norheim,[3] Thom-Kåre Granli[4]

Abstract
The existing and stipulated growth of electronic interoperation in government and between the public sector, citizens and businesses, requires improved means for information management and reuse of data. The cost of establishing and maintaining such interoperation is known to be high. Semantic technology and Linked Open Data can be viewed as building blocks of eGovernmental services and electronic interoperation. This technology enables new ways of managing, publishing, integrating and utilizing information. This paper describes five Linked Open Data pilots using Norwegian national master data. The pilots are evaluated against the Linked Open Data star rating scheme and four supplementing success metrics. The main data sources are the Central Coordinating Register for Legal Entities, the Register of Company Accounts and metadata from a national metadata repository. The need for merging national master data with other data sources is increasing, and the pilots demonstrate a technical and semantic approach to meet this need. To our knowledge this is the first attempt to publish main national master data on Legal Entities as 5 star Linked Open Data. The temporary evaluation of the pilots shows interesting and positive findings.

Keywords: eGovernmental services, Linked Open Data, Semantic Technology, Electronic Interoperation.

1. Introduction

The public sector generates and consumes large amounts of data. The commercial value and the business criticality of internal and external data have increased in recent years. This development is a consequence of more use of information and IT systems and more automation in electronic collaboration in and between organisations. Most data is made or captured to serve a specific purpose and is initially used directly or indirectly by people to perform a task. When it comes to using the data as Open Data, the actor collecting the information does not usually have full knowledge of all the potential usages, which leads to challenges related to accessing, understanding and trusting distributed information. These challenges are crucial to both producer and consumer and are potential obstacles to interoperable operations between organisations.

Norwegian core national and public master data is often described as the triangle of central registers that hold data on (i) citizens, (ii) businesses (Legal Entities) and (iii) property (including map data). The Norwegian authority Brønnøysund Register Centre (BRREG) has

1 Det Norske Veritas, Veritasveien 1, 1322 Høvik, Norway, per.myrseth@dnv.com
2 Computas, Lysaker Torg 45, 1327 Lysaker, Norway, jens.kilde.mjelva@computas.com
3 Computas, Lysaker Torg 45, 1327 Lysaker, Norway, david.norheim@computas.com
4 Norwegian Research Centre for Computers and Law, University of Oslo, Norway, t.k.granli@jus.uio.no

the mandate to collect, maintain and publish content in 19 different registers. The information in these registers is subject to the EU directive on Public Sector Information [1] which is implemented in Norwegian law [13]. BRREG launched The Central Coordinating Register for Legal Entities (RLE) in 1995. For many years BRREG has offered online access to their data based on web-services and traditional web pages. The two main groups of data consumers are government bodies and legal entities.

With multiple users re-using public information across service domains and across geography and time, the challenges of information governance and interpreting information are many, but new opportunities also arise [18]. The need to use and merge RLE-data arises in e.g. eGovernmental services, linked enterprise data, computation journalism [8] [9], statistics etc.

According to BRREG, the most common usage patterns of their RLE data are verifying the existence of a legal entity and listing the CEO, board members or registered changes in the RLE data. But there is an increasing request for interoperability to enable both re-use of RLE data and merging capabilities between RLE data and other public and enterprise data. The pilots demonstrate technical and semantic interoperability and bring data to new groups of users. The pilots are online tools that help medium advanced users to study RLE data or study combinations of data sources. The tools enable data consumers to merge, search, visualise and perform both drilldowns and semantic reasoning from multiple sources.

Section 2 describes the methodology and project progress. In section 3 some related academic work is presented, supplemented with the status of related projects. Section 4 lists the pilots made, and section 5 discusses literature, contributions to research and relevant findings. The conclusion in chapter 6 is short and focuses on the fact that this is on-going work, and that this paper is meant as an overview of the LOD activities in quite a large project in a Norwegian context.

2 Methodology

The broader field of interoperability research and electronic government research presented in Grimstad et al [23] lists technical and semantic interoperability as two important levels of interoperability. Interoperability is viewed as a key enabler for eGovernment services [22] and the high level research question we address is “how can Linked Open Data improve interoperability”.

To measure how the LOD pilots influence interoperability we have chosen evaluation questions related to IT-system engineering and volume of reuse. The evaluation approaches are technical quality and organisational performance suggested by Flak [22]. The evaluation questions used are how LOD influence:

A. Cost reduction in establishing IT-solutions dependent on several external information sources. The solutions could be an enterprise solution or an eGovernmental service.
B. Effort needed in order to understand external data, build trust in them, and risk management of external data.
C. Ability for IT systems to change.
D. Re-use of data.

The score scale is as follows; (i) good, (ii) medium, (iii) same as existing.

A number of workshops and meetings with key personnel at BRREG have been conducted. This has been supplemented with interviews and workshops with several other stakeholders from public sector bodies, data consumers and journalists working within computational journalism [8]. The on-going work has been presented at several local conferences, forums and meeting-points in Norway, and valuable feedback has been taken into account.

Preliminary scores according to the evaluation questions above have been obtained through observations, workshops and tests with relevant but small sample size of users [14]. Based on this methodology the scores give an indication of findings. The pilots are also measured against the Linked Open Data (LOD) 5 star scheme, described in the next section.

3 Related Work

The importance of interoperability in electronic collaboration has been studied by a number of research projects within the EU framework programmes 6 and 7 [23]. The role of levels of interoperability in maturity models for digital government has been investigated by Gottschalk [20]. Sollisaeter's [7] study on maturity in e-government interoperability ends with the statement "It would also be interesting to investigate the effect of increased interoperability on benefits, in the direction of increased efficiency, effectiveness, and user satisfaction." The pilots that have been made have increased interoperability primarily at the technical and semantic level.

The high level design of the pilots is based on the 5-star deployment scheme for Linked Open Data [2]. The 5-star scheme described in [2] is as follows:

★	Available on the web (whatever format) but with an open licence, to be Open Data
★★	Available as machine-readable structured data (e.g. excel instead of image scan of a table)
★★★	As (2) plus non-proprietary format (e.g. CSV instead of excel)
★★★★	All the above plus, Use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff
★★★★★	All the above, plus: Link your data to other people's data to provide context

Table 1: Explanation of the star rating regime for Linked Open Data [2][6]

As far as we know (March 2012), no other Norwegian national public body than BRREG has main national master data that scores more than three stars according to the scheme above. The 5-star deployment scheme [2][6] mentions a possible metadata extension, i.e. that metadata should be available from a major catalogue. Recently data.gov responded to this extension and launched a hosting service for RDF vocabularies/metadata [5].

The Norwegian public sector gives open public data some attention. E.g. the letter of instruction from departments to all public entities instructs them to establish access to potential open data sources [21]. In addition, the quality regime for Norwegian public web

sites [12, criterion 3.8] uses the "three stars" from [2] as a quality metric. The initiative "data.norge.no" lists available data sets and offers an open data hosting service [11].

There are few legal obligations in this area. Section 9 of the Freedom of Information Act of 19 May 2006 [13] relating to the right of access to documents held by public authorities and public undertakings, states that: "Any person may request access to a collation of information that is electronically stored in the databases of an administrative agency insofar as the collation can be done using simple procedures."

4 The pilots

Five pilots are presented in the following section. The linked open data star rating for each pilot is shown by the number of stars in the section heading.

The main data sources used in the pilots are the RLE and the Register of Company Accounts. The RLE coordinates information in various registers on business, industry and public agencies and ensures that all the information is available in one place. The Register of Company Accounts is the most important data source for anyone wishing to obtain information on the financial state of Norwegian businesses and industries. To avoid privacy challenges by open and merged data, limited data sets are used in the pilots.

RLE data in a time line perspective [16] (★★★)

The purpose of this pilot was to test how to improve usage of ageing information on legal entities. This pilot is a solution for visualizing data from multiple data sources on a timeline-oriented user interface. To improve users' ability to properly understand the enterprise data in its historic context enterprise data is supplemented with historic context data. To build this solution some new web-services for RLE were made and several semantic technologies have been used.

RLE data subset (★★★★)

The purpose of this pilot was to test how to encapsulate the existing RLE web-services to act as a LOD service. The LOD encapsulation was made by a Java application turning Rest requests into web-service requests, and web-service responses into RDF and html responses. In addition a thin LOD client based on Excel and macros were used to look up RLE data and fill out a spread sheet with addresses of legal entities.

SERES RDF, Norwegian Semantic Repository of Electronic Services (★★★★)

The Norwegian Semantic Repository of Electronic Services (SERES) is an eGovernmental metadata repository and a framework for establishing semantic interoperability. For each term in SERES, the repository contains definitions and links to relevant/connected terms. According to the 5 star scheme, this repository can solve the additional metadata criteria mentioned in [6]. The LOD design principles were used for this metadata repository as well. In order to make the various terms in SERES linkable, an architecture that provides look up functionality, a SERES URI (GUID), was needed. SERES acts as an online dictionary which responds according to LOD design principles, e.g. with RDF or html.

RLE data linked to SERES metadata (★★★★★)

When the SERES RDF pilot was launched, the RLE pilot was updated with links (URI) to SERES content. The previous RLE pilot was limited to name, organization number and address. In this pilot the full response capability from the RLE server is handled. This resulted in more interesting RLE data becoming available. NACE codes [10], number of employees, organization type etc. was thus added in the new pilot. A more sophisticated part of the response was the visual graph of persons related to each legal entity. The graph contains contact persons, the members of the board of directors of an organization, as well as links to related organizations. Related organizations in this context are mother/daughter organizations and organizations that share one or more members on the board of directors. With this graph the RLE data became navigable and the direct connections between organizations are made explicit. Links to the SERES metadata dictionary were added. E.g. the definition of data in the RLE service like organizational number, dates, role etc. could now be looked up. The vocabulary used in the RLE RDF model was enhanced compared to the previous pilot, not only by providing links to SERES, but also by reusing terms from internationally established vocabularies such as Dublin Core, Friend of a Friend and vCard.

RLE and financial data with metadata (★★★★★)

This pilot builds on the experience from the pilots described above, but is technically different because this pilot holds a copy of RLE data and financial data. The pilot receives a dump of the RLE data every day. Through this approach great flexibility and full Sparql capabilities are gained. The pilot combines the RLE data with data from (i) the Register of Company Accounts, (ii) the hierarchy of NACE codes for statistical classification of economic activity [10], (iii) DBpedia and (iv) geo coordinates. Some historical data (some snapshots from the last three years) are included to facilitate time series and trend analysis. A large variety of search-, merge- and drilldown functionality is made available. Export formats for further use are implemented. Currently visualizations in e.g. charts and maps are available.

5 Discussion

This paper describes how the pilots demonstrate a technical and semantic approach to meet the need of interoperability and the increasing number of usage patterns of RLE data. The project has tested technologically if Linked Open Data design principles can be a building block in future eGovernmental services or distributed information solutions. The discussion below focuses on the four indicators, A-D described in section 2 of this paper, and sums up how the results contribute to improved interoperability.

A. Cost reduction in establishing IT-solutions dependent on several external information sources.

The pilots have shown that it is technically possible to utilize semantic technologies to facilitate open data and linked data. The pilots demonstrate semantic technology as a possible choice for implementing distributed information solutions and as a basis for some types of eGovernment services. Currently the pilots have implemented data lookup and merge functionality. Update functionality based on Linked Open data design principles has only been implemented in the SERES RDF pilot. No comparable financial figures exist to conclude if semantic technology actually reduces costs in establishing IT-solutions. But the

experience from making the pilots is valuable proof of concepts which can be used to argue for improvements. On this basis, the score is set to be medium.

B. Effort needed in order to understand external data, build trust in them, and risk management of external data.

To make sure that RLE data consumers understand the data sources the SERES metadata repository has been used [15]. This is in line with potential extension to the LOD star rating [6]. It has been time consuming to get the data owners to participate in making the definitions and getting the definitions online in parallel with the data. Semantic models for some of our data sources have been prepared in the SERES repository before launching the data. The process of getting the semantic models and definitions authorised is a slow and bureaucratic process. To trust data you should understand them, and metadata is a new means to build this trust. The lack of provenance data [17] and data quality [19] indicators is challenging both from a semantic, a trust and a risk perspective. Traditional security mechanisms such as integrity, confidentiality and authenticity in data come in addition to the arguments mentioned. Full risk assessments of LOD or external data in general have not been performed as part of our project. Two of the pilots have implemented support for LOD from the SERES repository to improve understanding, build trust and to manage the risk of being dependent on distributed information resources. We regard this as an innovative solution encompassing traditional distributed information solutions. On this basis, the score is set to be good even if a relevant business risk assessment has not been performed.

C. Ability for IT systems to change.

We have not yet had major changes in our pilots or semantic models, so we do not have observations to conclude on this indicator. But the ability to describe and maintain semantics in models separated from the software code makes us optimistic. At the same time the separation introduces a new need for configuration management between software and metadata models. Governance of complex systems has a tendency to be change resistant while innovation cries for change. Depending on the type of change, the interoperability properties may change as well, generating positive or negative cascade effects.

D. Re- use of data.

We have seen several projects that use data from our pilots, and they argue that the original access methods to RLE data did not meet their needs. At the time being, several research projects and commercial projects use our pilots. E.g. CyberWatcher uses the service for updating their enterprise data. The Norwegian Research Council has a project archive where the data on applicants are linked to the RLE LOD interface. Asker municipality uses the SERES LOD pilot for a web-solution helping families with disabled kids to find their way through the maze of public services. The research project Planet Data combines maps with our RLE data. This suggests a good score on this indicator.

The LOD 5 star scheme is a kind of maturity model indicating the level of technical and semantic interoperability. A LOD source with a high score should be easier to reuse than one with a low score.

The pilots have received good feedback from national LOD interest groups and international semantic technology communities like e.g. SemanticWeb.com [4]. In discussions on the economic impact of open data and LOD we have learned that new usages pop up when the costs of data drop to zero. In addition to the price issue, important drivers for increased use of

data are (i) reduced effort to make usage technically possible, (ii) reduced effort to understand data and (iii) offerings of free online tools to merge and play with data.

December 12th 2011, as part of the European Commission's open data strategy, the Commission presented a proposal for a revision of the Public Sector Information directive which proposes to further open up the market [3]. This includes introducing independent oversight of re-use rules in the Member States, making machine-readable formats for information held by public authorities the norm and limiting the fees that can be charged by the public authorities to the marginal costs. The positive organisation culture at BRREG is in contrast to the findings in the evaluation of the PSI directive, where a main motivator for amending the directive is that public sector bodies fail to realise the economic potential of public sector information [3].

The maturity of semantic technology has developed substantially from the project started in 2007 until now. In the beginning the technology was immature and little knowledge and experience was in place. After the first pilots were demonstrated, we continued to produce pilots, but we also discussed how to solve semantic interoperability, data quality [19] and trust issues when merging data. During the last year, when several different demos have been available, data collectors, value adding service providers, lawyers, politicians, journalists and data consumers in general start to understand the impact of LOD. New discussions arise on issues like: what are the possible LOD business models, how to measure the national economic impact of LOD, how to regulate intellectual property rights, how to handle privacy issues when anyone can merge open data and so on.

6 Conclusion

This paper describes how five pilots have used semantic technology and Linked Open Data as building blocks to handle the challenges of improving interoperability and merging distributed data sources. These challenges are common when establishing eGovernmental services and managing distributed information in general. National master data governed by the Norwegian authority Brønnøysund Register Centre has been used in the pilots.

The pilots made are measured according to two maturity models. The pilots score from 3 to 5 stars according to the LOD star scheme [2], high score indicates good interoperability. The project indicators A-D are: A) cost reduction in establishing IT-solutions, B) reduced effort to understand and manage data, C) increased ability for IT-systems to change and D) increased use of data. The pilots score medium to good on the A-D indicators. Based on this we argue that LOD pilots contribute to improved interoperability. According to Flak and Solli-Saether [22] interoperability maturity is a prerequisite for good electronic collaboration and eGovernment services. The pilots and the evaluation methodology are under development and the data sources used are all structured data with at least one common identifier. This can influence the results evaluation.

Acknowledgement

This research was carried out as part of the Semicolon project and partially funded by the Research Council of Norway, contract no. 201559. We would like to thank BRREG for giving us the opportunity to demonstrate the usage of open data and semantic technologies on

their national master data. Part of the work has been performed by master students and researchers at the Department of Informatics at the University of Oslo.

References

[1] Public Sector Information EU directive 2003/98/EF

[2] Tim Berners-Lee, Linked open data design issues, 2006. http://www.w3.org/DesignIssues/LinkedData.html, visited April 2012

[3] Proposal for a Directive of the European parliament and of the council. Amending Directive 2003/98/EC on re-use of Public Sector Information published the12.12.2012.

[4] Jennifer Zaino. Norwegian Semantic Web Project Latches On To Linked Open Data's Possibilities. 27.10.2010. http://semanticweb.com/norwegian-semantic-web-project-latches-on-to-linked-open-datas-possibilities_b829, visited April 2012

[5] US Open Government Data Portal, a service for publishing and retrieving RDF Vocabularies. http://vocab.data.gov. Visited 15.03.2012.

[6] Tim Berners-Lee, Is your linked open data 5 star? Amendment to [2], 2010. http://www.w3.org/DesignIssues/LinkedData.html, visited April 2012

[7] Hans Solli-Saether: Maturity in e-government interoperability: an exploratory study of e-services in Norway. Proceedings of the IADIS International Conference e-Society 2010, Porto, Portugal. S. 115-122.

[8] Sarah Cohen, James T. Hamilton, and Fred Turner. Computational Journalism, in: Communication of the ACM October 2011.

[9] Data journalism handbook. http://blogstats.wordpress.com/2012/01/22/data-journalism-a-handbook-guides-and-a-competition/, visited April 2012

[10] Statistical Classification of Economic Activities in the European Community. NACE codes

[11] data.norge.no project by Norwegian Agency for Public Management and eGovernment (Difi). http//data.norge.no , visited April 2012

[12] Quality measurement regime for 700 Norwegian public web sites, metrics on open data are part of the regime. http://kvalitet.difi.no/ , visited April 2012

[13] Freedom of Information Act, 19.05.2006 nr 16, Offentlighetslova. http://www.lovdata.no/cgi-wift/ldles?doc=/all/nl-20060519-016.html, visited April 2012

[14] Martin Schmettow, Sample Size in Usability Studies, Communication of the ACM, April 2012.

[15] SERES, Semantic register for electronic collaboration. http://www.brreg.no/samordning/semantikk/

[16] Per Myrseth, Jon Atle Gulla, Veronika Haderlein, Geir Solskinnsbakk and Olga Cerrato, Utilizing Ageing information. 10th European Conference on eGovernment. June 2010.

[17] Luc Morea et al. The Provenance Of Electronic Data, in: Communications of the ACM. April 2008

[18] Terje Grimstad and and Per Myrseth, Information Governance as a basis for cross-sector e-services in public administration. International Conference on E-Business and E-Government, Shanghai, 2011.

[19] ISO 8000 Data Quality.

[20] Petter Gottschalk, Maturity levels for interoperability in digital government. Government Information Quarterly 26 (2009) 75–81.

[21] Amendment to letter of instruction to all Norwegian public entities in 2010.

[22] Leif Flak, Hans Solli-Saether, The Shape of Interoperability: Reviewing and Characterizing a Central Area within eGovernment Research. 45th Hawaii International Conference on System Sciences (HICSS 2012), 4-7 Jan. 2012. Pp. 2643-2652.

[23] Terje Grimstad et al. Semicolon, Interoperability in Public Sector, State of the Art. Report No. 2008-0996, Det Norske Veritas, December 2008.