Australian Bureau of Statistics Implementation of Semantic Web Technology

Arupa Sarkar[1], Michael Mecham[1] and Peter Meadows[1]

1 Australian Bureau of Statistics, Australia
{arupa.sakar,michael.mecham,peter.meadows}@abs.gov.au

Abstract. The ABS is Australia's official national statistical agency.

The requirements for a Web Ontology tool were built from a wide consultation with stakeholders, internal communication experts, technical experts and external clients. The interest was driven from a desire to avoid the rapid decay of explanatory material as statistics and methods updated regularly. The primary focus of this project became a cost effective tool that could deliver a flexible product which could be updated regularly by operational staff. This meant something that could be derived from a complex statistical conceptual metadata such as the 2008 System of National Accounts and be progressively updated.

The primary cycle of research and development work was undertaken in 2010. Significant changes in technologies and strategies mean some options exist in 2013 that were not available to the development team in 2010.

Keywords. Semantic, ontology, official statistics, documentation

1 Introduction

This paper outlines the Australian Bureau of Statistics interest in harnessing the data aggregation and organizational features offered by ontology technologies. ABS business, goals, practical application and technical implementation are discussed. ABS understands an ontology to be a *formal, explicit specification of a shared conceptualisation[1]* for a domain of interest. An ontology provides a shared vocabulary, which can be used to model a domain, that is, the type of objects and/or concepts that exist, and their properties and relations

2 Discussion of the Business Requirement and Ontological Benefits

2.1 Business of the Australian Bureau of Statistics

The ABS is Australia's official national statistical agency. It was established over 100 years ago as the Commonwealth Bureau of Census and Statistics, following enactment of the Census and Statistics Act 1905. The agency became the Australian Bureau of Statistics in 1975 with the passing of the Australian Bureau of Statistics Act 1975. This Act also established the role of the Australian Statistician and defined the functions of the ABS.

The ABS provides statistics on a wide range of economic, social, population and environmental matters, covering government, business and the community. It also has an important coordination function with respect to the statistical activities of other official bodies, both in Australia and overseas.

Economic statistics are produced predominantly from the ABS business survey program. They include an extensive range of statistical outputs relating to the structure and performance of the Australian economy. Population and social statistics are produced mainly through the ABS household survey program. They include statistical information relating to the Australian population, including census and demographic statistics, as well as information relating to the social and economic wellbeing of the population.

The ABS develops national statistical standards, frameworks and methodologies, which are applied, as appropriate, to all ABS statistical collections, including business and household surveys. The ABS takes a leading role by encouraging other Australian state and territory government agencies to adopt these standards, frameworks and methodologies in their statistical activities. The ABS also works closely with other agencies involved in the development of standards and frameworks. These standards are developed and implemented on the basis of consultation and input from a range of stakeholders and interest groups in the statistical and user community.

One of the key methods for communicating with the user community is through the use of Concepts, Sources and Methods (CSM) documentation. These documents, usually large complex pieces of work, are produced irregularly and are intensive to write and re-write. In 2008, the economic statistics part of the organisation decided on the need for an improved means of creating these documents, plus open up the ability for users to comment or discuss methods.

2.2 International conceptual frameworks

An important consideration when preparing CSM documentation is that in order to support comparability and assure quality of statistics internationally there are a range of conceptual frameworks and standards which have been agreed globally and which

are updated over time as information needs evolve. Frameworks and standards for various statistical domains (eg Labour, Economic Accounts) are registered in the Global Inventory of Statistical Standards which is maintained by the UN[2]. The System of National Accounts 2008 (2008 SNA)[3] is one example.

Each nation needs to make decisions when implementing the global framework. There exists, an Australian System of National Accounts (ASNA)[4] which is based on SNA. ASNA identifies, for example, cases where Australian Accounting Standards make it impractical to collect data for a concept on exactly the same basis as defined in SNA. Instead data is compiled for a closely related concept as defined in ASNA.

The European system of national and regional accounts (ESA95) provides an example of a transnational framework which is also consistent with (an earlier version of) SNA[5].

2.3 Rationale for a web ontology tool

The rationale or requirements for a tool were built from a wide consultation process. The main area of interest was driven from a desire to avoid the rapid decay of explanatory material as statistics and methods are updated regularly. The primary focus of the project became a cost effective tool that could deliver a flexible product which could be updated regularly by operational staff. This meant something that could be derived from a complex statistical conceptual metadata such as the 2008 System of National Accounts and be progressively updated as methods and data changed.

Some other external demands placed on a documentation tool included the ability to browse, search and discover concepts. As the 2008 SNA is a linked set of economic accounts, then the conceptual metadata needed to manage overlapping concepts and should only be defined once and linked into the explanation of the account with the ability to see where the concept was used in other parts of the National Accounts.

A pragmatic decision was made in the early stages of developing requirements around release of new material. It was considered not feasible to quality control new explanatory information in each quarter so an annual process was suggested. This added further requirements to have a staging and sign off environment prior to release to the web.

Internal demands for a systematic documentation tool added some requirements such as the ability to have a corporate approach to a tool and looking for ways to link the documentation in the future so there could be further sharing of metadata between reporting areas. There was emphasis on a tool being compatible with the future enterprise architecture of the ABS to avoid re-engineering in a future environment. Another requirement was to increase usability by staff who would be responsible for maintaining and updating documentation. Documentation is considered to be a burden

under current Notes/Office[6] product based environments. The Notes/Office environment is considered to be poorly structured despite the versatility of Notes for controlling documents.

The impact of social media and web 2.0 was seen as a new playing field for the recording of explanatory material. This generated an expectation of being able to hold discussions and add comments to entries, though resources need for moderation were of concern. A need to maintain a history of discussion and methods changes was considered to be a part of the requirements, noting the need for transparency and the risk of duplicating discussion or methods if not recorded.

The ABS works with a secure environment policy for documentation that is released to the web. There was a need for a secure environment to host, stage and to edit documentation. There is also a need to have some governance arrangements over the quality control and final sign off on components. It was recognised at the initial stages that the first set of documentation would mean intensive sign off, but would be minimal for updates once the tool was up and running.

3 Objectives

The primary objective was to allow complete traversal of a networked node structure of Concepts, Sources and Methods for ABS purposes. This would permit a more open and transparent view of ABS statistical techniques. From a technical perspective, a semantic wiki would deliver numerous benefits and the potential for extensibility. Secondary to this, was the need for pragmatic re-use of ABS metadata. Namely, the underlying CSM ontological linkages underpinning the ASNA or economic model. Achieving this would allow for smarter extensibility across numerous ABS systems and their conceptual models. An example of extending systems reach driven by CSM ontology reuse would be opening ABS documentation to public comment through dynamic content driven web forums.

4 Overview of business requirement

The purpose of this project was to formulate, detail and define the process of producing a new Australian System of National Accounts: Concepts, Sources and Methods (CSM) to replace the current CSM published in 2000 by the Australian Bureau of Statistics. The final output of this project was a new and improved CSM. Additionally a new CSM was needed because the existing CSM was out-dated and needed to comply with a new set of international standards released in 2008. In summary, a new CSM was needed due to address a number of factors.

The factors are:

- changes to the system of national accounts and balance of payments;
- user demand for a more interactive CSM; and
- desire for new technology being available to maintain contemporary CSM documentation by contributors;

This project also sought to create a new CSM with richer feature set offering improvements over the existing CSM. Initial analysis of client requirements was followed by an Agile approach to object modeling of CSM, ensuring all-way linkages were maintainable and traversable. OWL fitted this requirement and offered the ability to be extended or modified. Semantic Works Media Wiki plugin was chosen and integrated well with pre-existing wiki templates and again offered a reasonably simple upgrade path for future needs.

5 Establishing the project team

In addressing the business objective "to allow complete traversal of a networked node structure of Concepts, Sources and Methods", use of ontologies, and related tooling, appeared logical as the cornerstone for the solution. Suggested re-write: The key business objective was to produce the publication "Australian National Accounts: Concepts, Sources and Methods" (ABS cat. no. 5216.0). The logical solution was in the use of ontologies and related tooling.

ABS understands an ontology to be a *formal, explicit specification of a shared conceptualization[7]* for a domain of interest. An ontology provides a shared vocabulary, which can be used to model a domain, that is, the type of objects and/or concepts that exist, and their properties and relations.

The project, therefore, included a major component of R&D (research and development). As well as needing to address a specific set of business needs, the project was seen as a potential forerunner to wider use of ontologies and related technologies within the ABS. The project team consisted of a number of developers working closely with subject matter experts in regard to economic statistics. ABS experts in R&D worked closely with the project team.The team undertook a range of research and consultation in regard to product suites and international practices before commencing their main development work.It worked to clarify and formalise the business requirements. It then commenced cycles of object modeling followed by testing, including business review.

6 Technical constraints

The original vision for the project, based on the business requirements, was that

Subject matter experts would be able to draft content within the ABS environment, including relationships and versioning control.

Appropriate content testing, review and approval processes would then occur within the ABS with authorized processes.

Content would then be released to the website in accordance with ABS dissemination processes, standards and protocols.

The aim, therefore, was a technical solution that integrated well with workflow processes and common desktop applications within the ABS[8]. IBM Notes[9] (Client) and IBM Domino[10] (Server) is used as groupware within the ABS. This includes corporate content management roles fulfilled by products such as Microsoft SharePoint in other organisations. IBM Connections[11] is used to support wikis and related collaboration spaces within the ABS. In 2010, this suite of products from IBM did not include plug-ins for defining, managing, publishing and using ontologies.[1] The aim, therefore, became to establish a solution that would interoperate effectively with the ABS groupware environment rather than being able to develop a solution that was "native" to that environment. In regard to the architecture of the solution, there would need to be an interface for definition and maintenance of ontologies; selection, browsing and exporting features; and a canonical store for ontology and versions.

7 Technical evaluations and selections

7.1 Ontology plugins

A range of ontology plugins, beyond the specific wiki functionality associated with IBM Connections, were evaluated. These included Onto Wiki, Media Wiki and Semantic MediaWiki.

These plugins were found to perform adequately in their own right but did not integrate in a viable manner with ABS' internal authoring, approval and web publishing environments.

After initial evaluations, Semantic MediaWiki was chosen by the project team. This selection followed several attempts at integrating semantic plugins within the ABS IT environment, pragmatically Semantic MediaWiki was the easiest to configure. Factors influencing this choice included: useability; search pattern result sets being better than others; and, availability of expert resources.

[1] The authors are unaware of ontology plugins having been added in the past three years but have not undertaken detailed product research to confirm this remains the case.

7.2 Conceptual Modelling

Both Protégé and Altova SemanticWorks were used by the Project Team. Some "hand editing" of RDF was also required at times.

Protégé is a free, open source ontology editor and knowledge-base framework. The platform supports modelling ontologies via a web client or a desktop client. Protégé ontologies can be developed in a variety of formats including OWL, RDF(S), and XML Schema.

The project team used Protégé via desktop client to develop the ontology in OWL format.

SemanticWorks is the visual RDF/OWL editor from the creators of XMLSpy. (The latter is used widely in the ABS for working with XML). It allows you to graphically create and edit RDF instance documents, RDFS vocabularies, and OWL ontologies with full syntax checking.

SemanticWorks provided powerful, easy-to-use functionality for:

- Visual creation and editing of RDF, RDF Schema (RDFS), OWL Lite, OWL DL, and OWL Full documents using an intuitive, visual interface and drag-and-drop functionality
- Syntax checking to ensure conformance with the RDF/XML specifications
- Auto-generation and editing of RDF/XML or N-triples formats based on visual RDF/OWL design
- Printing the graphical RDF and OWL representations to create documentation

SemanticWorks was found to be a relatively user friendly tool for working with ontologies. It tended to be preferred to Protégé by subject matter experts but Protégé was, in general, preferred by developers.

Using OWL as the common representation, it proved viable to move content between SemanticWorks and Protégé in both directions.

Overall Protégé was used more widely by the project team. If ontologies were to be developed and maintained more broadly by subject matter experts from across the ABS in future, however, many of the more advanced capabilities of Protégé may not be required (or understood) by many of the users. For such users they would represent added complexity rather than added capabilities.

No overall decision was made in regard to ongoing use of Protégé, SemanticWorks or a third alternative for future work on developing ontologies within the ABS. The longer term choice of tool would depend somewhat on

- who was responsible for developing ontologies (subject matter experts, or developers advised by subject matter experts)
- requirements for content approval processes
- requirements for integration with other tools in the ABS desktop environment

7.3 Infrastructure

The tools being used by the project team could not be supported quickly and easily within mainstream ABS development environments[13] in regard to repositories (eg Oracle RDBMS, SQLServer) and coding (eg Java).

Ultimately, a MySQL database was used to store the content and PHP was used for scripting. This allowed a flexible approach to defining a large number (10,000's) of classes and their properties. It also allowed for better parent – child linkages and the ability to deal with circular references.

While the pressure to maintain momentum required that the team work outside the mainstream development environment, this added to the challenges in than having the content disseminated to the ABS web environment.

8 The dissemination challenge

Section 6 refers to approval processes for releasing content to the web, together with standardisation of web architecture and of processes for transferring content from internal repositories to the web.

These processes and protocols are important elements in guaranteeing the statistical quality, reliability and consistency of statistical products and services made available to users by the ABS.

These processes and protocols are, however, under increasing pressure to evolve more quickly as dissemination needs, expectations and opportunities change. For example, less than 20 years ago printed publications were the main focus for ABS dissemination processes, now they are a negligible element of our output.

While there are increasingly extensive user driven data services available from the ABS currently, the majority of content on the ABS website in 2013 consists of "static" products such as PDFs, Spreadsheets, Data Cubes and authored (rather than dynamically generated) HTML pages.

A major update to the dissemination strategy released in 2013 aims for a major change in the balance between static products and data services in the next few years. More extensive metadata services (eg to help external users code/define their own

data based on standard classifications/vocabularies published by the ABS) are also expected.

In 2013, the project might have been selected as an "early adopter" for the updated dissemination strategy.

In 2010 the fact that the architecture used for working with ontologies did not (and could not) align with preferred ABS web architecture at that time proved to be a critical barrier to the project becoming a model for further development, and dissemination, of ontologies by the ABS.

9 The outcomes

This project confirmed that there are significant benefits from investing in an in-house semantic authoring tool for explanatory documentation that have a statistical framework. The ability to transverse the linkages in the environment mean improvements in access, development and sharing by staff. One of the barriers that still needs to be reconciled and is worth considering for any National Statistical Office is the existing process and maturity of web publishing systems. This is especially true in the case of strict branding and desire for control of content.

The project succeeded in a proof of concept for managing complex relationships and was a start in the long journey of a consolidated metadata store for concepts. The benefits of using an ontology demonstrate the ease in which a network of relationships could be developed and overlain with significant amounts of documentation. Staff satisfaction with the authoring tool was quite high, with some reservations about security and the ability to migrate from an authoring to a web environment.

The inability to readily integrate the content and the workflows with ABS internally used groupware and with web architecture proved insurmountable barriers to quickly and extensively "productionising" the outputs from the project.

Some of the content developed during the project has been released to the web. For example, the updated Concepts, Sources and Methods for Australian System of National Accounts (ASNA) was released last year as a PDF[14] which comprises 722 pages. Content for the PDF was not extracted automatically from the ontology but migrated manually.

One of the key objectives of stakeholder interaction was not met due to the barrier to publish in a semantic web format. The project may be renewed as the ABS reviews its operational environment and develops new pathways for publishing information.

10 Future Directions

10.1 Vision for expanding the original specifications

As the project developed, there was wider recognition of the potential that the solution based on Semantic MediaWiki and the ontology toolset could provide to clients.

The business customer developed the following scenario as the project was being completed:

Imagine an operator on the economic accounts that could have a button to the conceptual framework in which they were responsible for. Imagine that they could examine the concept they were working on, any internal or external discussion or unforeseen complexity they may not have been aware and the ability to see what impacts will occur from changes they intend to make on their component as well as recognising the changes that others will make on their source information. Now imagine if they could seamlessly move into the detailed documentation need to exactly operate the economic accounts system. Once finished with their component, then seamlessly move onto reading intelligence reports on that component based on machine to machine process, RSS feeds, newspaper feeds or reports back from the providers of the original data. In this way we can see a fully integrated operational environment for complex economic (or any other statistics) information.

They envisaged options for extending the environment to have parallel structures for operational documentation, and setting up mash-up environments that can mine for information from various database and web resources.

While integration issues meant it was not feasible to move forward and realise the wider vision at that time, strong support from the subject matter experts for continuing this direction as soon as it is feasible continues remains a strong business driver for ABS interest in pursuing practical application of semantic technologies.

The vision from the business customer extended well beyond ASNA to a linked network of CSMs for related fields of statistics.

A network could (with less intense connections) then be extended to encompass CSM information for all of the subject matter domains for which the ABS produces statistics. For example, within economic and social statistics there are overlapping concepts that are slightly different but should be by definition the same. One of the main challenges is around the definition of income for households. In the social framework, income for households is defined as the inflow of cash and benefits; in the economic frameworks it is defined as an expenditure of business and government. Once these are reconciled, metadata linkages will be much easier to manage.

10.2 International Connections do we need this section?

Since 2010 there has emerged a much stronger and more active focus among National Statistical Institutes (NSIs) and international agencies on working together in practice to achieve standards based modernisation based on collaboration, sharing and reuse. Much of the leadership has been provided by the High Level Group for the Modernisation of Statistical Production and Services (HLG) which comprises ten heads of national and international statistical organisations, including the head of the ABS.

The strategy to implement the vision of the HLG includes positioning the "official statistics industry" within the wider information industry and recognising a very prominent connection with Open Data.

This prominent connection is, in part, symptomatic of increasing interest over the past three years among producers of official statistics in the opportunities that ontologies and related semantic technologies offer.

For example, NSIs have been active in initiatives such as "Open Government Vocabulary" (US) and "Controlled Vocabulary Service" (Australia). They have also been active in facilitating various "data.gov" initiatives around the world (eg UK, New Zealand) including links to conceptual metadata structured on a semantic basis.

The ABS is actively liaising with a number of NSIs who are currently undertaking planning and development of "Conceptual Content Management Systems" that are characterised by use of ontologies and related technologies.

While the possibility remains under active consideration, a decision has not yet been made by the ABS to seek to engage more closely with one of these projects as a formal international collaboration.

It appears certain in any case, however, that standards based, potentially sharable, developments for managing statistical frameworks as ontologies will emerge within the next few years. This includes the possibility of HLG focusing the attention of the "official statistics industry" on this topic during 2014 or 2015.

11 References

1. Gruber, T: A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition, 5(2),199-220, (1993) http://tomgruber.org/writing/ontolingua-kaj-1993.pdf
2. Global inventory of Statistical Standards http://unstats.un.org/unsd/iiss/Classification-of-International-Statistical-Activities.ashx

3. 2008 System of National Accounts, http://unstats.un.org/unsd/nationalaccount/sna2008.asp
4. Australian National Accounts: Concepts, Sources and Methods, ABS Cat.no.5216.0, 2012, http://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/5216.0Main+Features1Edition%203?OpenDocument
5. Glossary from European System of national and regional accounts 1995, http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Glossary:ESA95
6. IBM Notes/Microsoft Office
7. Gruber, T: A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition, 5(2),199-220, (1993) http://tomgruber.org/writing/ontolingua-kaj-1993.pdf
8. IBM Product Guide http://www-03.ibm.com/software/products/us/en/ibmnotes/
9. IBM Product Guide http://www-03.ibm.com/software/products/us/en/ibmdomino/
10. IBM Product Guide http://www-03.ibm.com/software/products/us/en/conn/
11. ABS IT environment http://www.abs.gov.au/websitedbs/d3310114.nsf/0/b9043642361d7a66ca256b59007bdae7/$FILE/Introduction%20to%20the%20ABS%20ICT%20Environment.pdf
12. Australian National Accounts: Concepts, Sources and Methods, ABS Cat.no.5216.0, http://www.abs.gov.au/AUSSTATS/abs@.nsf/Lookup/5216.0Main+Features1Edition%203?OpenDocument