n2Mate: Exploiting social capital to create a standards-rich semantic network

David Peterson
BoaB interactive
2/84 Denham St.
Townsville, QLD Australia 4810
+61 7 4724 2933
david@boabinteractive.com.au

Anne Cregan
National ICT Australia
223 Anzac Parade
Kensington NSW Australia 2052
+61 2 8306 0458
anne.cregan@nicta.com.au

Rob Atkinson
CSIRO Land & Water
Lucas Heights Research Laboratories
Private Mail Bag 7, Bangor
NSW 2234, Australia
rob.atkinson@csiro.au

John Brisbin
BoaB interactive
2/84 Denham St.
Townsville, QLD Australia 4810
+61 7 4724 2933
john@boabinteractive.com.au

ABSTRACT

A significant boost on the path towards a web of linked, open data is the establishment and promotion of common semantic resources including ontologies and other operationalised vocabularies, and their instance data. Without consensus on these, we are hamstrung by the famous "n-squared" mapping problem. In addition, each vocabulary has its own associated attributes to do with why it was developed, what purposes it is best suited for, and how accurate and reliable it is at both a content and technical level, but most of this information is opaque to the general community.

Our theory is that it is the lack of socially-sensitised processes highlighting who is using what and why, that have led to the current unmanageable plethora of vocabularies, where it is far easier to build your own vocabulary than try to find a suitable, reliable existing one.

We therefore suggest that there is considerable value in the development of an online facility that performs the function of providing a space listing vocabulary and ontology resources with their associated authority, governance and quality of service attributes. Presenting this in a visual form and providing pivotable search facilities enhances recognition and comprehension.

Additionally, and critically, the facility provides a focal point where discourse communities can make authority claims, rate vocabularies on various parameters, register their commitment to or usage of particular vocabularies, and provide feedback on their experiences. Through social interaction, we expect the most solid and useful vocabularies to emerge and form a stable semantic platform for content representation and interlinked knowledge.

Our strategy is to become sufficiently enmeshed in the native information habits of people and their derivative institutions to reveal and collect their standards-seeking needs and activities with a minimum of effort on their part.

This paper describes a pilot facility testing the theory above. Dubbed "n2Mate", it is a novel exploitation of social networking software to provide a lightweight and flexible platform for testing the efficacy of leveraging social networks to link existing registers and 'seed' an information space focussing on the use of standards in online information management.

The paper uses examples from the Australian context to provide clear illustration of the central arguments.

Keywords

Registers, vocabularies, standards, linking density, rdf graph, social networking, knowledge re-use, n2Mate, n-squared

1. SOCIAL AND TECHNICAL CONTEXT

The current emergence of a data web has re-focussed our attention on standards. To be truly effective, the semantic web needs to evolve towards a minimum number of ontologies, highly re-used, and densely interlinked, rather than a sparse network with minimal interoperability.

1.1 The standard problem with standards

The project to link open data can be realised through explicit declarations by one data source in relation to another. These "hard" linkages provide a high degree of certainty, but make data maintenance exponentially difficult as the number of hard linkages grows.

Standards, understood as nodes of agreed meaning, provide a more scalable approach to data linking. By agreeing to use the same term to describe similar ideas in our different data, we establish an implicit (semantic) linkage between our data. The project to conceive, negotiate, and promote standards, however, has proven to be even more difficult than the maintenance of hard linkages.

It is often noted, with some irony, that the great thing about standards is that there are so many to choose from...and if you can't find one you like, you can always create your own. .

While these sentiments provide excellent platforms for pub-based oratory, the realities are not so easily dismissed. Application designers, knowledge seekers, and agencies with a mandate to interoperate are all too familiar with the significant resource drains that occur when standards are hard to locate, difficult to apply, or confusing to distinguish between.

Standard vocabularies and data definitions have been quietly multiplying in traditional media since ancient Sumer (ca. Wikipedia, Cuneiform) but in more recent times the Semantic Web has inspired a hyperbolic growth in contributions to the standards project. For instance, a search in Swoogle on the word "address" returns 12,834 semantic web documents; on "book" it returns 19,601 (at 2008-01-24). For someone seeking to exercise the efficiencies of knowledge reuse, this wealth of choice is simply overwhelming and self-defeating. The current state of affairs reveals semantic fragmentation, not semantic integration and knowledge creation.

Even within a narrow domain like the Australian government, there are a wealth of terminologies and metadata "standards" available for government agencies to consider. It is not clear if a whole of government survey of standards has ever been undertaken, but informal observation suggests that there are hundreds of attempts to describe very similar concept spaces.

1.2 Does anyone have a wheel like mine?

People have been trying to standardise themselves in one way or another for quite some time. The most obvious benefit of this instinct toward standardisation is communication efficiency, a direct input to the rate of knowledge creation. By speaking the same language, we can communicate and collaborate far more effectively. Yet the barriers to standardisation appear to take on new forms as fast as we evolve knowledge.

In our present age the benefits of information interoperability are now well understood, if only through their absence. Most people and institutions involved in project scoping, information product development, and online service provision clearly grasp the power of knowledge re-use and the cost efficiencies of standards-based interoperation. This assertion is supported by the existence of an entire government department whose mandate is to promote effective and efficient information sharing, governance structures, tools, methods and re-usable technical components across the Australian Government.

The Australian Government Information Management Office (AGIMO) published a Government Architecture Reference Model [1]that discusses "...a repository of architectural artefacts (including standards, guidelines, designs and solutions) that may be utilised by agencies to deliver an increasing range of Whole of Government services."

In practice, however, we find that the task of identifying and verifying the suitability of existing artefacts is simply too time-consuming. As a consequence, there are a great many ontologies and informal vocabularies used by a very limited number of organisations or agencies, with a great sparsity of intermappings between them, even though there is a very large amount of crossover in terms of content.

More globally, the Linking Open Data (LOD) project [1] holds datasets that currently comprise over 2 billion triples but reveal only about 3 million links (SWEO, 2007), so overall the graph is very sparsely interconnected [2].

In many ways the current situation is akin to a train network that has millions of stations (nodes) covering the same area (knowledge domains) but with a great sparsity of tracks (mappings) between stations, and hardly any trains and passengers (services, publishers, agents, users) running on the vast majority of them.

Our experience with efficient rail networks shows that we want to reach a necessary minimum of stations interconnected with an optimised number of tracks, and attract a maximum number of trains to utilise the infrastructure. This obviously gives us a far more robust and useful semantic network to traverse.

In related research, it should be possible to show how the density of interconnectedness in the RDF graph improves the efficiency of machine process operation without producing a debilitating level of ambiguity. We would argue that the degree of interconnectedness implemented between ontologies can be taken as a proxy indicator of interoperability across the knowledge domain.

[1] http://www.agimo.gov.au/services/GovDex

1.3 Scalable register networks

As we have argued, there are many technical standards and common policies in use across a wide range of government activities, but the very number of such activities and standards is in itself posing a significant challenge.

AGIMO and others have a role in promoting the use of common approaches, but it is increasingly difficult to track which standards apply to which set of problems.

In general, there is an issue about the scalability of any approach for improving interconnectedness. We believe that the most promising strategy is to utilise registers to hold metadata about standards and their implementation, including records of organisations, projects, standards, controlled vocabularies (and associated people and roles). A network of such registers, coupled through normal web services mechanisms, has the potential to form a semantic fabric that addresses the business-level needs of people and institutions. Whilst this is potentially a vast undertaking, the bulk of target information already exists, and there are already a great many people actively tasked with identifying, using and promoting standards. These people are likely to be receptive to an effort such as n2Mate.

A network of registers, supported by a "register of registers" addresses the most important questions: who is doing what, which standards are relevant, who can I talk to, what is the governance model for these artefacts, and how trustworthy is the source. Through a richly populated network of registers, these become questions any organisation can rapidly address, and in doing so can promote commonality of approach within and amongst various discourse communities.

1.4 Socially-sensitive metadata

One of the dark secrets of the machine-based knowledge project is the enormous loss of content as we move from people's minds to their documents and datasets. David Snowden, amongst many others, has pointed to the impossibility of "collecting" knowledge from people without providing a meaningful context:

"Human knowledge is deeply contextual, it is triggered by circumstance and need, and is revealed in action. .... to ask someone what he or she knows is to ask a meaningless question in a meaningless context. Tacit knowledge ... comes about when our skilled performance is punctuated in new ways through social interaction" [3].

A socially-sensitised strategy provides the meaningful context and familiar atmosphere that people require before they can (or will) reveal their knowledge in a useful way.

We suggest there is a cluster of persistent problems in complex information spaces that can be socially characterised as follows:

Who and what:

- Owner: Who owns it?
- Creation: Who created it?
- Maintenance: Who is responsible for maintaining it?
- Domain: which domains is it relevant to? This will include a number of different ways of considering domains.
- Usage: Who uses it?

- Endorsement: Who endorses it? This will include various parameters and a rating system.
- Processes: What Business, Government or other processes is it used in? What role does it play?
- Governance: Who is in charge of it? Who has formally agreed to support, maintain, and implement it?

Quality of Service Parameters:

- Provenance: What guarantees are there that the information is accurate and verified?
- Currency: How often is it updated? What guarantees are there that it is up to date?
- Availability: What guarantees are there regarding the availability of the vocab, dereferencing considerations

Other Considerations:

- How does it relate to other standards in the space?
- User experiences

2. SOCIAL ARCHITECTURES AND SEMANTIC NETWORKS

The principle social platform techniques we seek to exploit include:

- **Popularity Rankings**: number of times a standards artefact is referenced (implemented).
- **Authority Badges**: mechanism to advertise an authority claim over a standards artefact.
- **Related to ("Friends of a Standard (FOAS)")**: linkages from standards artefacts to their cohort of implementers.
- **Trust ratings**: showing satisfaction with the custodian of a standards artefact.
- **Hero worship**: most interlinked, most trusted, most useful

Each of these techniques have corresponding interface strategies that provide a powerful social platform in which people (and institutional roles) can operate quite naturally.

Each of these techniques also forms a search facet that can be traversed with high efficiency faceted search and browsing tools.

2.1 Use Case

A simple use case will help us set the stage for describing the technical architecture proposed.

A researcher is preparing her research plan on a section of the Great Barrier Reef. Although she is an experienced marine scientist, she is new to the GBR and to her host research facility. She suspects she should be using:

- standard naming conventions for the GBR regions;
- standard identifications for the particular reefs;
- standard data sampling techniques appropriate to the Australian tropics;
- standard data formats, enumerators, and vocabularies in her datasets;
- standard citations of agencies, programmes, and people referenced in her work;
- standard metadata fields and vocabularies to describe her research output;
- standard project management practice in reporting on her project's progress.

In the absence of a useful standards locator, it's not likely that she will achieve a high standard of conformance to the norms of her discourse community.

In the absence of a socially-sensitised register space, it is not likely her discourse community is actively sharing their experience and wisdom with standards.

2.2 Instance Data

The facility needs to be designed around a sufficient minimum of predicates that embody the "business logic" of the facility and establish the semantic armature we require for inferencing.

We propose the following [shows predicate] as a starting point:

- Organisations are [responsible for] people, projects, standards, and vocabularies
- People are [associated with] Projects
- Projects are [implemented by] Standards
- Standards are [expressed with] vocabularies
- Trust or utility of Standards are [ranked by] People

Using these indicative predicates as a starting point, we can answer a matrix of discovery questions through faceted visualisation. In each search operation, the user can rotate to a facet of interest to continue the discovery process.

- I know someone like me [PersonName] > What projects are they associated with?
- Those projects are like mine [ProjectName] > What standards are used in them?
- Those standards are of interest [StandardName] > How can I decide which one is most appropriate for me?

The logic described here is possible because we have imposed a limited set of predicate types. These types are native to the n2Mate facility. To take advantage of existing social networks that utilise other predicate types, Semantic Web vocabularies such as SIOC [4] and FOAF [5] will be used.

The facility will also consider structured lists of resources, like a list of country names available from the same address, to itself be a kind of register. For instance, many applications need a list of every valid country name for users to input their address information. The ability to reference an external source that is authoritative, accurate, up-to-date and reliably available and derefenceable reduces the need for application maintenance.

The metadata held in these registers can be typed according to existing conceptualisations. For example, the National Data Network [2] draws on ideas from the Metadata Open Forum [3] to classify their metadata as: Discovery metadata; Quality metadata; and Definitional metadata.

We note that the semantic register network can also list web services in addition to typical standards artefacts such as ontologies and vocabularies.

[2] http://www.nationaldatanetwork.org/

[3] http://metadataopenforum.org/

We intend to specifically tune this facility to the needs of government and community agencies that have a mandate to participate in the creation and maintenance of highly effective approaches to service improvement.

3. IMPLEMENTATION OPTIONS

A demonstrator version of n2Mate can be established using readily available tools and datasets so that a more detailed critique can be pursued with a minimum of upfront overhead. In this section we discuss some of the more promising approaches.

3.1 Key components

The registration process, and maintaining a network of linked objects, is the function of traditional registry technologies, such as ebXML Registry. Navigating and efficiently querying the contents and relationships is not well supported by this environment.

It is proposed to automate the harvesting of object relationships from the "Register of Registers" into a triple-store. This is the same pattern found in data-mining, where transactional database content is restructured into generalised query-oriented structures. For our purposes, automated discovery of patterns is not the focus: fast, efficient visual presentation is essential.

Users will be parsing through extensive data structures, and may need to propose and refine their discovery logic in quick, exploratory sorties.

Visualisation and facet search: Gnizr + Solr

We want a tool that thinks natively in URIs and triples. Gnizr [4] is an open source front end that handles user account management, bookmarking, tagging, and semantic search

Every object stored by gnizr is a bookmark (URI), and the folksonomy tag interface is SKOS [6] enabled.

Solr [5] is an open source enterprise search server based on the Lucene Java search library, with XML/HTTP and JSON APIs, hit highlighting, faceted search, caching, replication, and a web administration interface.

Solr could be used to facet the data into searchable and browseable components. For example, if users are interested in what ontologies Sun Microsystems is using, they select Sun from the 'Who is Using' facet. The other facets instantly re-order and re-number themselves and the user is free to further refine the results by selecting additional facets.

Faceted search visualisation can be negotiated through cluster maps (eg, Aduna [6]) with a high degree of efficiency.

Semantic interpretation: MOAT

MOAT [7] (Meaning of a Tag) could serve as the basis for giving extended quality of information to free form folksonomy tagging.

This will allow users of the bookmarking system to have the flexibility of folksonomy and the interlinked structure of the Semantic Web. The added benefit is that MOAT is a distributed system and can tap into other servers to give extended meaning to free-form tags.

Triple-store: Sesame

Sesame [8] could provide backend triple store, graph manipulation, RDF inferencing, and remote SPARQL [7] endpoint access.

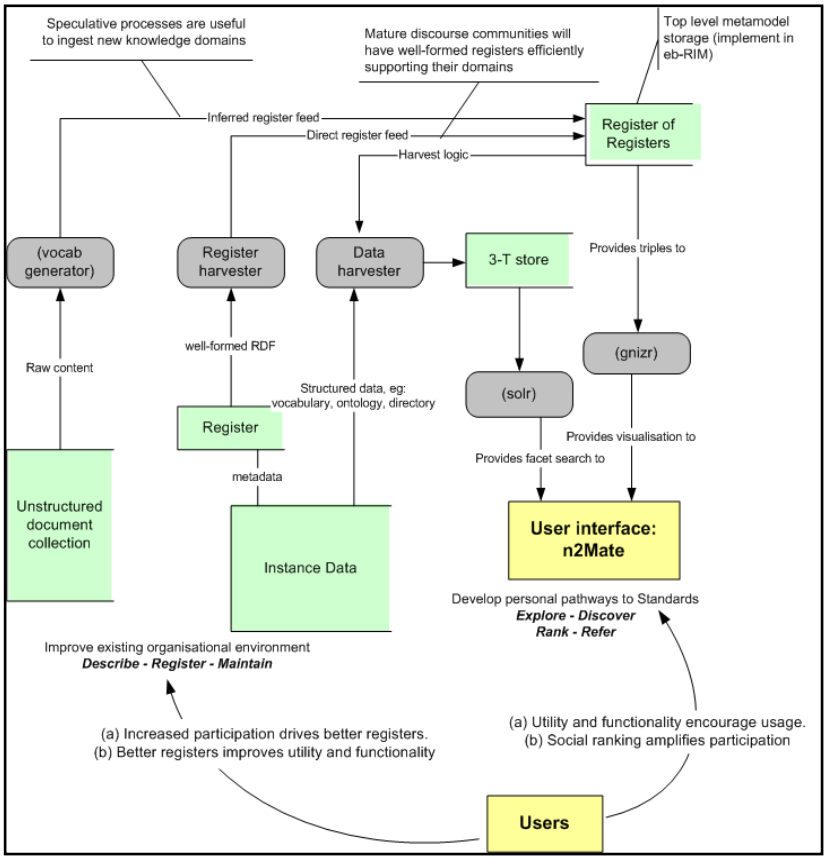

Figure 1: n2Mate Conceptual architecture

Policy layer: PLING

The development of robust approaches to policy negotiation is being driven by a W3C Interest Group [9]. The n2Mate project could field test various strategies for handling issues of personal privacy, information reuse, and access control.

[4] http://code.google.com/p/gnizr/

[5] http://lucene.apache.org/solr/

[6] http://www.aduna-software.org

[7] http://moat-project.org/

[8] http://sourceforge.net/projects/sesame/

[9] http://www.w3.org/Policy/pling/

Trust and Governance: POWDER

POWDER [10] is the W3C's Protocol for Web Description Resources, currently in development.

Governance: is related to the idea of trust. In the context of this project, we want to explore two aspects of governance:

1. How to make it easy for agencies who have a mandate to be an authority for some asset to discharge their duty in an efficient and useful way.

2. How to provide users with a suite of trust measures that will allow them to evaluate the qualities of a particular asset in relation to their needs.

POWDER seeks to develop a mechanism through which structured metadata can be authenticated and applied to groups of web resources.

POWDER provides us with a means to both retrieve information about a block of Web Resources and authenticate that this information may be attributed to the owners of the information.

3.2 Testing the system with existing resources

There are already many semantically rich registers implicit in the operations of government, including the identifier of government agencies, registers of company names, standards recognised by Standards Australia, legislation and regulations, management areas for land, water, soils, health etc. This represents a wealth of entities about which assertions can be made, to create a semantically rich environment.

Semantic Web data can be roughly broken down into 3 levels: [2]

1. Vocabulary / Ontology
2. Individual occurrence of those terms and actual instances of non-information resources
3. The links that tie the vocabularies to their occurrences

All three of these need to be captured with adequate provenance data to bootstrap n2Mate.

The following web services can be utilised to populate/update information as well as add important metadata to the Register of Registers component of n2Mate.

- **Watson [11]**: A gateway to the Semantic Web, focusing on: semantic data quality; relations between ontologies; access to semantic data
- **Talis Schema Cache** [12]: Cross-linked and navigable index of ontologies and vocabularies.
- **Swoogle [13]**: Search engine for Semantic Web artefacts
- **Sindice [14]**: Indexes the RDF web and pulls out the triples. From there it essentially creates a reverse lookup.
- **Falcons [15]**: Currently indexing 34,566,728 objects (2008-02-01), Provides bi-directional resource linking.
- **Ping the Semantic Web [16]:** archives the location of recently created/updated, web-accessible RDF

3.3 Data harvesting and processing

n2Mate can leverage existing search engine services, such as those listed above, to collect data instances from target registers and sources. Many of these have or are developing APIs that facilitate direct access to their collections and service points.

Where well-formed registers and artefact collections exist already, n2Mate could establish harvesting relationships (presumably through appropriate API arrangements). OWL files, RDF data dumps, and SPARQL endpoints could be pointed to the n2Mate system for automated data fetching and processing.

Additionally, trust algorithms would be created from graph inferencing, metadata and social data to further guide the prospective n2Mate user, allowing them to more quickly determine what is the best artefact to use in their situation. This will be an evolving process that will occur over time as the quality of data and user interactions flows back and forth.

4. CONCLUSION

The unique aspect of this proposal is that it leverages the hidden formal and informal knowledge networks created by existing business processes, and marries this information with social networking models to provide a useful way of organising and navigating the wealth of available information. It uses the community of people using vocabularies to empower others, starting with the places where agreements already exist.

The n2Mate provides a tool that encourages use of standardised artefacts by exposing existing registers, leveraging social networks and building a central reference point for users that will assist them to identify relevant semantic assets for their needs, choose amongst them, and feel confident about their utilisation.

Further, research into the strategy proposed should provide contributions to related projects, such as the development of:

- A lightweight mechanism revealing the state of interconnectedness in and between discourse communities.
- A bridging space between government, business, community, academia and science knowledge assets to enhance broadscale interoperability.
- A genetic algorithm to breed, select, and hybridise various standards artefacts such as ontologies, services, and trust authorities.

In conclusion, we suggest that there is currently a significant level of inefficiency in the applied domain of project scoping, information product development, and online service provision due to the inadequacy and irrelevance of existing knowledge registers.

We further suggest that a promising solution strategy involves using the power of social networks, coupled with semantic discovery and visualisation tools, to create a socially-sensitised semantic network of standards registers.

[10] http://www.w3.org/2007/powder/

[11] http://watson.kmi.open.ac.uk/Overview.html

[12] http://schemacache.test.talis.com/

[13] http://swoogle.umbc.edu

[14] http://sindice.com

[15] http://iws.seu.edu.cn/services/falcons/

[16] http://pingthesemanticweb.com/

5. REFERENCES

5.1 Citations

[1] C. Bizer, T. Heath, D. Ayers, and Y. Raimond. Interlinking Open Data on the Web (Poster). In 4th European Semantic Web Conference (ESWC2007), pages 802–815, 2007. http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

[2] M. Hausenblas, W. Halb, Y. Raimond, and T. Heath. What is the Size of the Semantic Web? - Metrics for Measuring the Giant Global Graph, 2007.

[3] Snowden, Dave. Information vs Knowledge. http://www.rkrk.net.au/index.php/Information_Vs_Knowledge

[4] J. Breslin, A. Harth, U. Bojars, and S. Decker. Towards Semantically-Interlinked Online Communities. In Second European Semantic Web Conference, ESWC 2005, Heraklion, Crete, Greece, May 29-June 1, 2005. Proceedings, 2005. http://sioc-project.org/

[5] D. Brickley and L. Miller. FOAF Vocabulary Specication. Namespace Document 2 Sept 2004, FOAF Project, 2004. http://xmlns.com/foaf/0.1/.

[6] D. Brickley and A. Miles. SKOS core vocabulary specication 2005-11-02. W3C working draft, W3C, November 2005. updated version under http://www.w3.org/TR/swbp-skos-core-spec.

[7] E. Prud'hommeaux, A. Seaborne, eds, SPARQL Query Language for RDF http://www.w3.org/TR/rdf-sparql-query/.

5.2 Special thanks….

Renato Iannella provided background thinking on the Policy Aware Web. Alan Ruttenberg of the Science Commons and Tom Heath of the Linking Open Data project provided encouragement and wisdom from their broad experience. Steve Matheson from the Australian Bureau of Statistics corroborated our intuition that social platforms could play an important role in standards adoption.