SEARCH AND ORCHESTRATION OF DATA AND PROCESSES IN A FEDERATED ENVIRONMENT

Jeremy Siao Him Fa, Tristan W. Reed, Chet Tan, Geoff West, David A. McMeekin, Simon Moncrieff, Simon Cox*

Department of Spatial Sciences, Curtin University, GPO Box U1987, Perth 6845, Western Australia, Australia
Cooperative Research Centre for Spatial Information
*Commonwealth Scientific and Industrial Research Organisation, Australia

Commission VI, WG VI/5

KEY WORDS: Search, Federation, Orchestration, OGC, W3C, Semantic Web, Artificial Intelligence

ABSTRACT:

This paper describes on-going research on streamlining the access and use of spatial data and processes in Australia. Spatial data in Australia is available on-line at many levels of government from local authorities, state and territories (jurisdictions), and nationally from the Commonwealth and other sources. Much of this data is available via Open Geospatial Consortium and World Wide Web Consortium standard web services. This abstract discusses three related research topics that have been identified by a wide range of stakeholders through a comprehensive consultation process. These are search and discovery, federation and orchestration of data and processes. The commonality across the three research topics is that they all require Semantic Web and Artificial Intelligence methods and embrace the various standards, and if needed, propose modifications to such standards.

1. INTRODUCTION

This paper presents progress into research on improving the efficiency of access to and use of spatial data and processes in Australia. Australia has much spatial data available on-line from many levels of government ranging from local authorities, state and territories (jurisdictions) to nationally available data from the Commonwealth and other sources. Much of this data is available via web services using Open Geospatial Consortium (OGC) and World Wide Web Consortium (W3C) standards. It is important to note that Australia is a federation and it has to be recognised that the eight States and Territories own much of the data that makes up the national datasets. Data pertaining to each State or Territory can be obtained directly through a number of mechanisms. For example Landgate in Western Australia has the SLIP portal that allows access to many datasets generated by Landgate as well as from many of the other state agencies in Western Australia. Other web sites are available from States including New South Wales, Queensland and Victoria. Different pricing policies apply for different datasets, different levels of detail and for different States and Territories. Queensland and New South Wales are promoting a free open data policy whereas Western Australia has some free data and some you have to pay for. Much data at a national scale is available from the Public Sector Mapping Agency (PSMA), a company owned by the States, Territories and Commonwealth that integrates data to produce a number of commercially viable datasets such as addresses, the road network and the cadastre. Recently there have been moves by the Commonwealth to generate and provide much spatial data via a National Map initiative but problems exist because of the need to fund the integration of data and deal with the various formats and pricing policies.

This paper discusses three related research topics to address some of the problems of access and use of spatial data that have been identified by a wide range of stakeholders through a comprehensive consultation process. These are search and discovery, federation and orchestration of data and processes. The commonality across the three research topics is that they all require Semantic Web and Artificial Intelligence methods as well as embracing the various standards from the W3C and OGC, and if needed, propose modifications to such standards. The Semantic Web enables data, information and knowledge to be represented in an open way using data representations such as the Resource Description Framework (RDF) that can be accessed and queried using SPARQL (the equivalent of SQL for relational databases), as well as at a higher level using first order logic rules (Description Logics or DL). An important aspect of the Semantic Web is the ability to use knowledge from other sources such as schema definitions and controlled vocabularies as long as these are published on the Web.

Search is important as evidenced by the success of Google and others. However spatial data searching is very poor-and mainly relies on querying associated text and not spatial attributes. We are developing a natural language interface that understands spatial semantic concepts such as “Parks in Perth” in which “Parks” is the subject and “in” implies that “Perth” is a location. Hence data concerning parks can be searched for which are in a bounding box or shape file for Perth or simply near Perth.

The Australia and New Zealand Land Information Council (ANZLIC) have recently proposed the Foundation Spatial Data Framework (FSDF) (ANZLIC, 2014) that defines a number of national data themes (e.g. transport, water). Each theme consists of a number of datasets, each of which has its own model. Research is being carried out to federate the different representations of data from the States and Territories “on the fly” through matching ontologies describing the data models for each jurisdiction, and matching them to the ontology for the desired data model at the national level. The Web Feature Service (WFS) is used to acquire the data models, OWL-2 used to describe the models, and SPARQL rules used for the federation.

In many situations, a user query cannot be satisfied by the spatial data available, and the amount of data to be processed is too large to be downloaded and processed to produce the required information. Web Processing Services (WPS) have

been proposed by the OGC to allow a user to run algorithms on data remotely e.g. for determining the probability of flooding from the large Landsat dataset acquired over the last few decades. The orchestration of a number of WPS processes is being investigated in the context of emergency management that requires "on the fly" production of appropriate information. The functionality of the current WPS 2.0 standard is being explored to determine how the metadata about capabilities can be used to orchestrate a number of WPSs. Orchestration uses Semantic Web methods to determine what processes need to be performed given a user query, what data is required and how a particular process and data can be brought together on an appropriate architecture for processing. Deciding whether to move the data to the process or visa versa depends on bandwidth, processing power, processing speed and other factors.

The three research topics are now described in more detail.

2. SEARCH AND DISCOVERY

The proposed search and discovery tool: GeoMeta aims to overcome the inadequacies of existing geospatial search tools. It was estimated in 2008 that AUD$500 million of productivity was lost due to difficulties in obtaining relevant spatial data (ACIL Tasman, 2008). A large portion of this may be attributed to the lack of proliferation of advanced search tools. Although search tools for the World Wide Web (WWW) have improved with new technologies, search tools for geospatial data have not. For example, although Google's web search can correct misspellings and find pages containing similar words to the user's query (Nickel et. al., 2015), Google's map search tools do not contain these features.

Although Google appears to have a natural language interface that returns what appear to be meaningful results for spatial queries such as "Parks in Perth", the word "in" is taken as just another (very popular) search term instead of a constraint. In fact the query "Parks Perth" returns exactly the same results as "Parks in Perth".

Enterprise geospatial search tools are no different, and largely consist of either Google web-based or ESRI-backed applications (Golhani et. al., 2015) and similar open-source tools. These operate largely on keyword-search, attempting to find matches of each word in a user's query within a metadata record that describes a data set. Some, such as GeoNetwork, provide a primitive map interface that allows the user to set a bounding box to further narrow results (Ožana & Horáková, 2008).

GeoMeta aims to address these inadequacies by using semantic technologies to bring today's technologies for web search into geospatial search. The system primarily consists of two parts. Firstly, the user-facing front-end interface used to retrieve the user's query and display the results, and secondly the back-end processing interface used to determine the results.

The front-end interface is written using HTML5 and jQuery for a rich, application-like user experience that provides instant feedback such as loading screens compared to fixed pages in systems such as GeoNetwork. This is achieved through AJAX requests that re-load the results when the query is modified by the user. The back-end system is written in Python using the Django framework, which allows a wide range of extensibility and modularity for components to the system.

A traditional user story would start with the user entering a text query in the web interface. Currently a query can be in one of the simple four possible formats shown below. It is planned to extend the complexity of possible queries in future work allowing queries that are combinations including operations such as "and" and "or". Considering "Parks in Perth", "Parks" would be an <OBJECT>, "in" would be an <OPERATION>, and "Perth" would be a <LOCATION> and, hence, would match format (1). The back-end attempts to identify the appropriate format in the following order of precedence:

(1) <OBJECT> <OPERATION> <LOCATION>
(2) <LOCATION> <OBJECT>
(3) <OBJECT> <LOCATION>
(4) <OBJECT>

The back-end's next step is contingent on the type of query determined at the above step. In the cases of (1) to (3), where a location is found, a database lookup is completed to find a polygon representing the location, which is stored as part of the back-end. If the query is of type (1) where an operation is specified, further information may be requested from the user, depending on the type of spatial operation.

In all cases, the back-end then uses the WordNet graph interface to find similar words to the ones specified in the object (Budanitsky & Hirst 2006). The back-end then attempts to find the words within the database of metadata records, only looking at results that fall within the location's polygon (or that satisfy the spatial operation on the location for type (1) queries) and then ranks them based on the following equation:

$$RW_{Query} = \frac{\sum_0^W D}{WC}$$

where RW is the ranking weight of the query, W is the set of all words in the query, WC is the count of all words in the query and D is the distance from the original word. If there is a spatial operation involving distance, this ranking is then multiplied by the inverse of the spatial distance.

The user is then presented with the matching data sets and is free to visit the data set of their choosing. The results are displayed including a small diagram of the bounding box to supply context to the result, alongside all the metadata information stored in the backend database. Figure 1 shows the result for "Parks in Perth". In this example, the metadata used has been generated through the Google Map Engine API that accesses Landgate spatial data held in the Google cloud. For the query, the top two results are shown in full. For each, the extent of the data for the query is displayed on the map and the description from the metadata is displayed alongside.

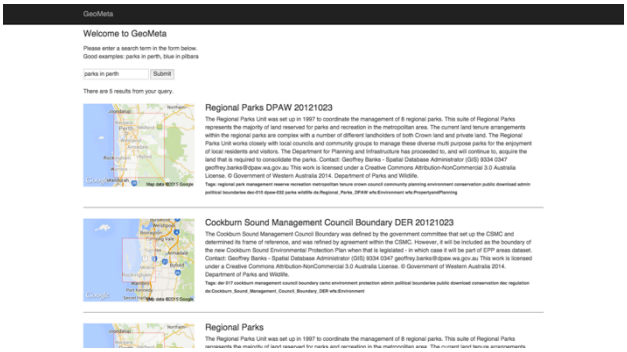

Figure 1. An example query using GeoMeta.

Future improvements to GeoMeta include a web crawler and metadata interface system to provide more context to the results and interfacing to existing Geonetwork instances. The metadata interface will allow existing metadata crawlers to supply related metadata in an interoperable and extendable RDF format, which will then be included in the search algorithm. The web crawler will discover more data sets on the Web, and metadata relating to them from the pages that contain the links to said data sets. It is planned to use a recently proposed RDF schema that is ISO 11915 compatible to represent the metadata and to use software such as Omniscient, Voyager and Sintelix that automatically searches file systems for spatial data and spatial data descriptions to populate the RDF representations. Each organisation is expected to publish their metadata in the RDF schema such that GeoMeta will be able to access it "on the fly". Ontologies, vocabularies and rules in SPARQL and DL using OWL-2 will be used to search the RDF metadata files and enhance the natural language interface.

3. FEDERATING DATA

ACIL Tasman (2008) states that spatial information is crucial to most business choices and better access to accurate and consistent spatial data is of increasing importance (West, 2014; Janakiraman et al., 2010).

Currently, Australia is using data integration to accommodate the need of sharing spatial data across different jurisdictions and generating national datasets (PSMA, 2009). Data integration is a method that duplicates data from different sources to a unified data warehouse, where the different schemas and syntax are modified to match the warehouse's ones. Due to the ever increasing volumes of data and changes, a more automatic and efficient method needs to be adopted (West, 2014). As it stands, a vast array of rules and regulations are used by PSMA to deal with differing representation of data models, coordinate systems, accuracies, and technologies (PSMA, 2009). The rules and regulations are all designed manually and hardcoded meaning that if any major change takes place at the data source level, cascading alterations must be done manually – leading to update delays of up to six months (ANZLIC, 2014).

The main challenge is the authoritative spatial data being managed by Australia's jurisdictions (States and Territories). The data owned by the different jurisdictions are represented with different schemas, vocabularies, and concepts (West, 2014). Although a global schema, the Foundation Spatial Data Framework (FSDF), is currently being developed to solve database heterogeneities, there are reasons why such a schema may not be adopted (Halevy, 2005). As such, federation techniques and tools to seamlessly unify the different datasets from the jurisdictions to a national level are needed (West, 2014).

An important issue regarding database heterogeneities is due to their distributed nature. As databases are developed independently and remotely from each other, the same or similar concepts can be represented differently (Halevy, 2005). Even though two or more entities might be semantically equivalent, their representations can be vastly different both syntactically and schematically. A possible approach is to unify the distributed spatial concepts together while allowing spatial entities to dynamically relate with these concepts. In other words, what is required is to find a method for each digital entity with the same semantics to reflect back to their real world concept. This is being explored through the use of Semantic Web techniques.

Two major semantic elements are ontologies and rules. Ontologies allow the matching of data and concepts, enabling better data semantics. Additionally, rules and logics are needed so that the data at the sources are properly mapped to their semantic equivalent in the FSDF global schema. Each data source can be accessed using Web Feature Services (WFS). A WFS is an OGC standard for external entities to interact with a database system over the Web, offering an easy portal for a federated system. As such, semantic heterogeneity can be addressed by the use of the ontologies (Jung, 2008), schematic heterogeneity by using a unified schema (FSDF), and syntactic heterogeneity with the use of thesauri (Bishr, 1998).

A working proof of concept called Aexon has been developed for administration boundaries. The FSDF schema represented in UML is translated to an equivalent ontology. Schema from two States: Victoria and Western Australia are acquired on the fly using WFS and mapped to the FSDF ontology using SPARQL operations and rules. A web interface has been generated that allows a user to explore the FSDF ontology, choose a particular attribute of interest e.g. name, and retrieve, via WFS calls, the corresponding attributes from the two States. Word matching is used to identify the same data and research is currently exploring more complex rules and vocabularies to match data that is described by different words. The resulting administration boundaries are then displayed on a map in the interface. Although not implemented in the web interface, the data about the administration boundaries can be downloaded resulting in one dataset. There are two parts to the web interface. The top part shown in Figure 2(a) shows the FSDF structure with StateElectoralUnit selected (a sub-class of UnitDefinition -> HierarchicalUnit ->ElectoralUnit. Attributes specific to this class are shown under Properties. Selecting fsdfName and inputting "Albany" in the search window will cause Aexon to search for the boundary in the two jurisdictions, returning the details for Albany in Western Australia.

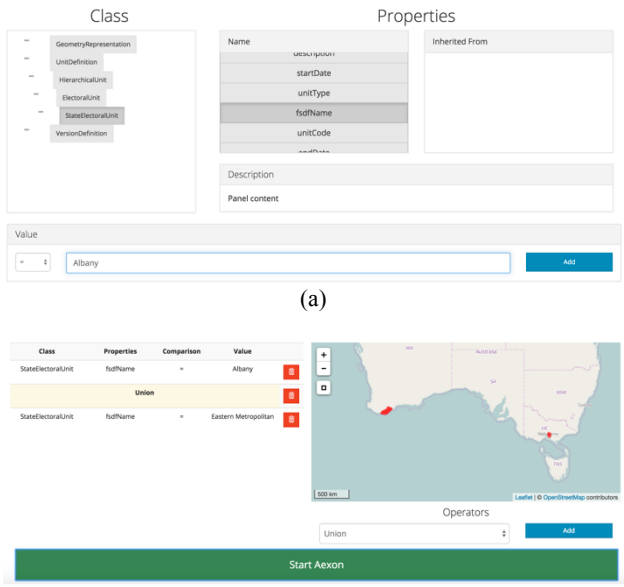

Figure 2. An example query and results for finding admin boundaries.

The lower part of Figure 2(b) shows a more complex query that combines Albany for Western Australia and Eastern Metropolitan for Victoria. The resulting boundaries are indicated on the map.

This process is not expected to completely match all data as each jurisdiction will have data that is specific to that jurisdiction, and there will be ambiguities that only a human can deal with. However it is expected that the majority of user requirements will be satisfied.

By federating Australia's spatial information, a more efficient method at consolidating and harmonizing uncontrollable heterogeneous databases is offered. Thus, this will solve issues regarding data duplication, delayed updates, and lifting the semantic burdens off users, leading to a more efficient and automatic way of accessing disparate spatial data in Australia.

4. ORCHESTRATION

The Web has been moving towards a Service Oriented Computer (SOC) architecture that supports automated use (Huhns, 2005). This architecture aims to build a network of interoperable and collaborative applications, independent of platform, called web services (Pugliese & Tiezzi, 2011). The geospatial world is also moving away from the traditional desktop application paradigm to processing and accessing data on-the-fly from the Web using web processing services, as outlined by ell et al. (2014). As web services technology has matured in recent years, an increasing amount of geospatial content and processing capabilities are available online as web services. These web services enable interoperable, distributed and collaborative geo-processing to significantly enhance the abilities of users to collect, analyse and derive geospatial data, information and knowledge over the internet (Zhao et al., 2013).

Geospatial organizations and business currently utilize workflows that rely on the manual chaining and integration of geospatial web services and datasets. These workflows also require human analysis of the output at each stage of processing, and manual determination of which web processing services to run on the data to achieve the final output. This introduces bias - a human user will use datasets or web services they are familiar with, regardless of the currency of the data, the frequency of updating of the data, or the validity of the process, a phenomenon known as the Mere-repeated-exposure paradigm (Zajonc, 2001). This, in turn, also increases the possibility of generating output based on out-dated and irrelevant data and processes. The problem is also amplified by current text search capabilities not being sufficient in finding more appropriate geospatial datasets and web services.

Automatically and intelligently orchestrated multiple geospatial web services and datasets is needed to provide the desired output from a complex user query. Yu and Liu (2015) have documented the need for automation in their attempts to implement a new system that republishes real-world data as linked geo-sensor data. Utilizing Semantic Web concepts, geospatial datasets and web services are linked via functionality, the inputs required and the outputs produced. The use of ontologies and rules then allows for the intelligent determination of which web services and datasets to use, and the order to use them to achieve the desired final output.

The research on orchestration (the process of linking and executing web processes in the correct order) builds on one of the aims of the Semantic Web to integrate semantic content into web pages that helps describe the contents and context of the data in the form of metadata (data about data) (Handschuh & Staab, 2003). This idea is adapted for use in geospatial datasets and web services. This will greatly improve the quality of the datasets and web services as a machine is able to understand what the data is for, what it can be used for and what other things are linked to it (Harth, 2004). This allows machines to process and chain data automatically (by utilizing ontologies available to intelligently deduce the next step to take along the process chain).

Crucial to the orchestration of web processes and data, is the use of standards given that the services and data are published and need to be easily integrated. The OGC has established standards for storing, discovering and processing geospatial information (Janowicz, 2010) including the well know WMS and WFS standards. The OGC has also published a standard for Web Processing Services (WPS) currently in its second revision (WPS 2.0). This standards lays out the groundwork for exposing features, inputs and outputs, as well as processing of geospatial web services (Lopez-Pellicer et al,. 2012). However much of WPS 2.0 is geared towards manual examination of the various processes via GetCapabilities queries and human readable descriptions. Examination of the standard reveals that to use WPS to orchestrate processes, extensions will be needed to include extra metadata such as algorithmic complexity, expected runtime, cost to run, as well as whether the process should be migrated to the location of the data and visa versa.

Current research is concerned with (1) finding WPS occurrences, (2) quantifying information from GetCapability requests, (3) building ontologies that can represent the WPS processes, and (4) developing rules to orchestrate sequences of data processes to satisfy the user query. An example is from the emergency management domain with queries such as "Where do I evacuate to escape the flood in my area?" This requires a user's location, a polygon of the region that is being flooded or will potentially flood, ontologies describing how to evacuate to a safe area, and a route finding process. The polygon of the region being flooded can come from a number of sources and use many processes. For example, the Landsat archive can be used to generate a probabilistic map of potential flooding areas from the historical record.

5. SUMMARY

This paper has reported on on-going research into a number of issues identified as important in the Australian and New Zealand context for easy access to spatial data and processes. Specifically search and discovery, federation of data and orchestration of data and processes are covered. The federated nature of Australia means it is difficult to generate spatial datasets at the national level because the States and Territories own their own data and use different schema, storage and access methods. Instead of mandating for all spatial data to be in the same format, the proposed solution is to exploit Semantic Web and Artificial Intelligence advances, especially RDF, ontologies and rules. The research addressing the three issues reported on in this paper are at a preliminary stage and are very much concerned with automating the various mainly manual methods used now and recognising that the spatial data should remain and be stored by the organisations responsible for its collection and custodianship. The methods deal with this dynamically accessing the information as and when needed and reducing the need to download the typically vast quantities of data to run on the client's machine. This fits well with the increasingly popular

idea of relying on server-sided processing allowing thin-clients to be used such as mobile devices.

ACKNOWLEDGEMENTS

The work has been supported by the Cooperative Research Centre for Spatial Information (CRCSI), whose activities are funded by the Australian Commonwealth's Cooperative Research Centres Programme.

REFERENCES

ACIL Tasman (2008). The Value of Spatial Information. Retrieved from http://www.crcsi.com.au/assets/Resources/7d60411d-0ab9-45be-8d48-ef8dab5abd4a.pdf [last accessed 29-04-2015].

ANZLIC (2014).The Australian and New Zealand Foundation Spatial Data Framework. Retrieved from http://www.anzlic.gov.au/__data/assets/pdf_file/0017/47321/FSDF_Booklet_edition_2_web.pdf [last accessed 29-05-2015]

Bishr, Y. (1998, June). Overcoming the semantic and other barriers to GIS interoperability. International Journal of Geographical Information Science, 12(4), 299–314.

Budanitsky, A., & Hirst, G. (2006). Evaluating Wordnet-based measures of lexical semantic relatedness. Computational Linguistics, 32(1), 13-47.

Golhani, K., Rao, A. S., & Dagar, J. C. (2015). Utilization of Open-Source Web GIS to Strengthen Climate Change Informatics for Agriculture. In Climate Change Modelling, Planning and Policy for Agriculture (pp. 87-91). Springer India.

Granell, C., Diaz, L., Tamayo, A., and Huerta, Joaquin. (2014). Assessment of OGC Web Processing Services for REST Principles. *International journal of Data Mining, Modelling and Management,* Special Issue, 6(4):391-412.

Halevy, A. (2005, October). Why your data won't mix. ACM Queue Magazine, 3(8):50-58..

Handschuh, S. and Staab, S. (2003). CREAM: CREAting Metadata for the Semantic Web. *Computer Networks,* 42(5):579-598.

Harth, A. (2004). An integration site for Semantic Web metadata. *Web Semantics: Science, Services and Agents on the World Wide Web,* 1(2):229-234.

Huhns, M. (2005). Service-Oriented Computing: Key Concepts and Principles. *IEEE Internet Computing*, 9(1):75-81.

Janakiraman, K. K., Orgun, M. A., & Nayak, A. (2010). Geospatial editing over a federated cloud geodatabase for the state of NSW. In Proc. of the 18th SIGSpatial International Conference on Advances in Geographic Information Systems - GIS'10 (p. 144). New York, USA: ACM Press.

Janowicz, K. (2010). *Semantic Enablement for Spatial Data Infrastructures.* Trans. in GIS, 14(2):111-129..

Jung, J. J. (2008). Query Transformation Based on Semantic Centrality in Semantic Social Network 1. 14(7):1031–1047.

Lopez-Pellicer, F.J., Renteria-Agualimpia, W., Bejar, R., Muro-Medrano, P.R., and Zrazaga Soria, F.J. (2012). Availability of the OGC geoprocessing standard. *Computers & Geosciences,* 47(3):13-19.

Nickel, M., Murphy, L., Tresp, V., & Gabrilovich, E. (2015). A Review of Relational Machine Learning for Knowledge Graphs: From Multi-Relational Link Prediction to Automated Knowledge Graph Construction. Proc. IEEE (to appear).

Ožana, R., & Horáková, B. (2008). Actual State in developing GeoNetwork opensource and metadata network standardization. Proc. 15th Int. Symp. GIS Ostrava 2008, Czech Republic.

PSMA. (2009). A concise history of PSMA Australia Limited (Tech. Rep.). Retrieved from http://www.psma.com.au/psma/wp-content/uploads/ACONCISEHISTORYOFPSMAAUSTRALIALIMITED.pdf [last accessed 29-05-2015].

Pugliese, R. and Tiezzi, F. (2011). A calculus for orchestration of web services. *Journal of Applied Logic,* 10(1):2-31.

West, G. (2014). Research Strategy Spatial Infrastructure. , 1–29. Retrieved from http://www.crcsi.com.au/assets/Resources/e0d480e5-b6c9-48a8-b6b5-f6ff5cc8b169.pdf [last accessed 29-05-2015].

World Wide Consortium. (2007). *Web 3.0 Emerging.* Retrieved from http://www.w3.org/2007/Talks/0123-sb-W3CEmergingTech/Overviewp.pdf [last accessed 29-04-2015].

Yu, L. and Liu, Y. (2015). Using Linked Data in a heterogeneous Sensor Web: challenges, experiments and lessons learned. *International Journal of Digital Earth,* 8(1):17-37.

Zajonc, R. (2001). Mere Exposure: A Gateway to the Subliminal. *Current Directions in Psychological Science*, 10(6):224-228.

Zhao, P., Lu, F., and Foerster, T. (2012). Towards a Geoprocessing Web. *Computers and Geosciences*, 47:1-2.