Research Strategy
Spatial Infrastructures (Program 3)

Geoff West

CRCSI P3 Science Director

Document Attributes

File name	File owner	File Location
P3 Strategy – final	Geoff West	

Document Control

Version	Status & revision notes	Author	Date
1.0	Endorsed by CEO, CRCSI and Chair, P3 Board	Geoff West	29/6/2012
1.0	Endorsed as an Exposure Draft for circulation and feedback from key stakeholder bodies	Geoff West	16/7/2012
1.01	Addition of document version control table and Figure 1 on SDI development	T.O. Chan	18/7/2012
2.0	Formatting	Jane Inall	19/7/2012
2.1	Incorporating minor edits and changes in response to recent comments	T.O. Chan	20/7/2012
2.2	Nature of draft and version control details	Peter Woodgate	27/7/2012
2.3	Remove “commercial-in-confidence” footer	T.O. Chan	3/8/2012
Final	Remove “Exposure Draft” and replace footer “Exposure draft to CRCSI participants” by “Final version for release on P3 website”	T.O. Chan	4/3/2013

Contents

EXECUTIVE SUMMARY 4

1.0 Introduction 4

2.0 Representing SDIs as Supply Chains 7

3.0 Research Drivers 8

4.0 Research Issues 14

5.0 Proposed Research Program 18

6.0 Use Cases 21

7.0 Use of the Established Software 22

References 24

Berners-Lee, T., (2000), Semantic Web - XML2000, Presentation available from: http://www.w3.org/2000/Talks/1206-xml2k-tbl/ [last accessed 12th June 2012]. 24

Appendix 1 27

1.0 The Semantic Web 27

Appendix 2 30

EXECUTIVE SUMMARY

This strategy sets out to significantly improve the organisation, access and use of spatial data in Australia and New Zealand by supporting the development of the Australia and New Zealand Spatial Marketplace and its underpinning National Spatial Data Infrastructure. Further development of the Marketplace will allow access to the wide range of spatial resources including datasets, apps and services. Priority will be given to improving the efficiency of the integration of key datasets identified by ANZLIC. Issues holding back the development of the Marketplace and the greater use of its resources include lack of automation, the complexity of the federated data models, rapidly increasing data volumes, massive uptake of spatial activities by private organisations and individuals, and the challenges of web-based processing.

These issues will be addressed through research into techniques focussed on advances in the Semantic Web, Federated Models and Web Processing. Important research drivers and research areas are identified and justified through reference to the latest research advances and recognised gaps. Four areas of research are proposed along with relevant use cases. The areas are (1) enhancing the Australia and New Zealand Spatial Marketplace, (2) data integration at a national level, (3) data integration at a jurisdiction level, and (4) licensing. It is proposed to run the program as one large project overseen by the Science Director, Program 3 with a management committee (the P3 board), and an industry advisory board consisting of 43pl members and relevant external parties.

1.0 Introduction

Of increasing importance to Australia and New Zealand is the ability of all organisations and individuals from all sectors to gain improved access to, and use of, the significant holdings of spatial data and other resources. For this to be achieved there are several key challenges to be addressed: improving discoverability, improving interoperability, simplifying access arrangements, streamlining the efficiency of processing, creating a fertile environment for value-adding data, apps and other services, and simplifying publishing arrangements.

There are two initiatives already underway that this strategy is designed to support: the Australia New Zealand Spatial Marketplace (ANZSM), and the National Spatial Data Infrastructure (NSDI).The ANZSM is sponsored by ANZLIC and aims to produce an environment for finding spatial data, products and services that can be free or purchased, downloaded, or processed on line. In addition data and processes from various organisations and individuals, public and private, will be publishable and offered for sale. ANZSM is conceived not to duplicate and compete with existing SDIs, both public and private, including global spatial resources offered by international players, such as, Google. It is conceived that free or paid spatial resources offered by Google, Bing, Yahoo! and Amazon etc., would be discoverable via ANZSM while the resources on the latter would be made searchable through Google and Bing etc. It is envisaged the marketplace will become self-sustaining.

In this context, the thinking behind the ANZSM has extended the scope of SDI to cover spatial resources that include not just mapping data but also data collected by remote-

sensing and ground-based sensors, derived information products, web-based data services, online processing services, mobile “apps”, in/volunteered geographic information and on-demand customisable spatial solutions and products.

The NSDI is coordinated by ANZLIC and comprises about a dozen fundamental datasets (including cadastre, roads, hydrology, and digital terrain models) that will be integrated from jurisdictions and other sources to form national datasets. Both of these initiatives as well as the general area of Spatial Data Infrastructures (SDIs) require research to reach their objectives.

SDIs are reaching maturity around the world in terms of making data available in standard forms (those from the Open Geospatial Consortium (OGC) for example) and having simple search methods. Research challenges include the need to go to the next level of capability i.e. towards more intelligent infrastructures that might be called Spatial Cyber Infrastructures or Spatial Knowledge Infrastructures. These will significantly improve the functionality, usability and automation of SDIs and assist the move towards more server-sided processing to cope with the massive and increasing data volumes. Research is needed into higher levels of representation and reasoning covered by what is termed the Semantic Web (using the World Wide Web Consortium (W3C) standards) as well as other higher-level aspects (including the latest OGC standards).

Semantic Web is also called Web 3.0 and represents the current stage of development of the World Wide Web (WWW) that serves as the platform for SDI development over the years. The relationship between WWW and SDI development in Australia and New Zealand is illustrated in Figure 1. It highlights a significant gap between the SDIs now and the SDIs CRCSI participants aspire to have in five years time.

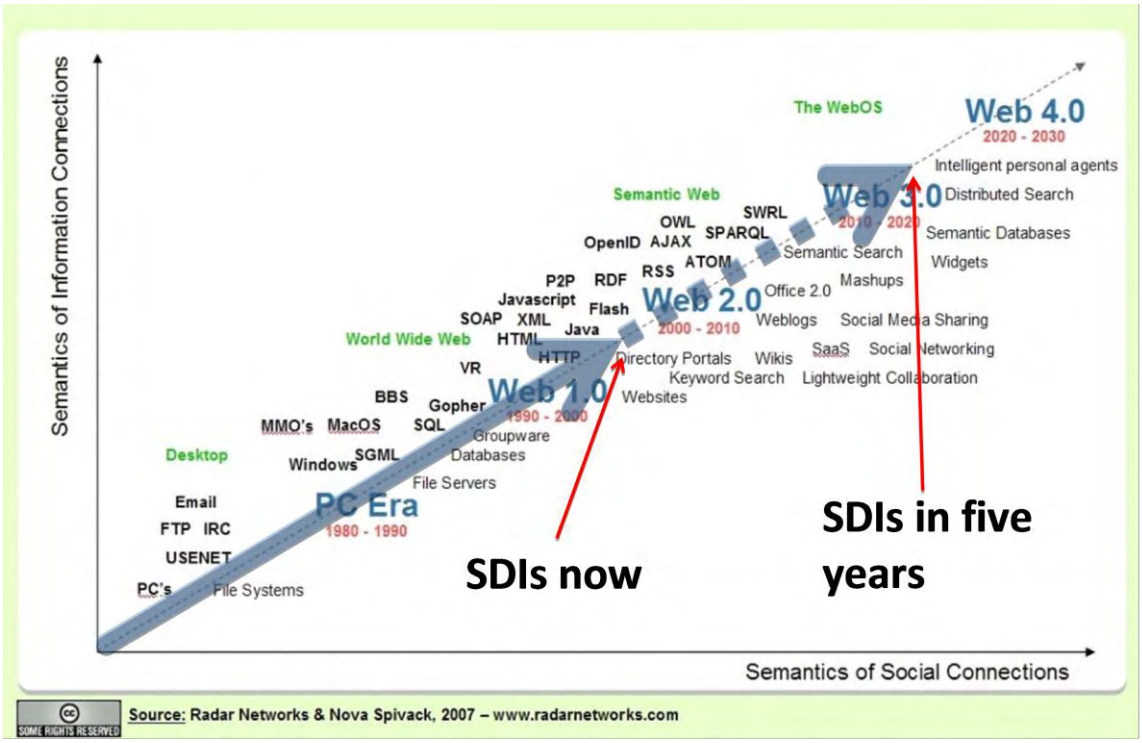

Figure 1. The development of SDIs in Australia and New Zealand in relation to WWW.

This strategy is the culmination of work carried out since 1 January, 2012 by a team consisting of the Program 3 Science Director Geoff West, Program Managers Kylie Armstrong and T.O. Chan, and Research Fellow David McMeekin.
A living presentation was developed that evolved during the development period as more information was acquired. Literature surveys were carried out to determine the current state of the art in terms of research. Discussions via face to face and teleconference meetings were held with many stakeholders in the spatial community (see Appendix 2).

CRCSI associated jurisdictions, agencies, and members of 43pl were targeted for their input to the process. Particular guidance was provided by ANZLIC and the ANZSM Steering Committee. Advice and assistance was also provided by CSIRO, the Spatial Industries Business Association (SIBA), the Office of Spatial Policy (OSP), Landgate and DSE amongst others. The participation of all these organisations in the consultation process was valuable in helping to determine the scope of the research. The culmination of the consultation was a workshop at the CRCSI conference in Brisbane in May 2012 at which broad agreement was obtained regarding the directions and content of the proposed strategy. Immediately following the CRCSI Conference the opportunity was taken to attend the Global Spatial Data Infrastructure (GSDI) Conference held in May 2012 in Quebec. GSDI is the world's largest conference devoted to SDI. It confirmed that the research directions set out in this strategy are cutting edge, relevant and likely to result in high impact benefits for Australasia.

Three main drivers have influenced the development of the research strategy, namely the need for excellent usability, automation and web-based processing. These drivers have led the research strategy to be focussed on the Semantic Web, Federated Models and Web Processing Services.

A brief overview of the Semantic Web is provided because much of the research proposed is dependent on this technology (a fuller description is included in Appendix 1 and in particular the concepts that underpin much of what follows in this strategy document). The Semantic Web is still somewhat in its infancy although some of its related technologies have been investigated in some form over the past 20 years or so. However all indications are that it is undergoing rapid development and now is the time for Australia and New Zealand to systematically prepare for users in the two nations to capitalise on the opportunities that it will bring.
The Semantic Web is becoming increasingly important for its power to develop federated models for combining disparate datasets requiring intelligent guidance to reconcile the different structures. Web processing is becoming *de rigueur* for the latest generations of spatial applications with the move towards mobile lightweight devices and cloud processing.

Specific research topics are described that are focussing on issues concerning spatial data and processes, and, importantly, are concentrating on the future development of the ANZSM and NSDI. These specific research topics will be addressed in the proposed research projects that will be developed from this strategy. To demonstrate the need for the research projects, some use cases are presented that relate back to the specific projects and the three main research areas.

The research required to advance the development of SDIs is wide ranging and it is recognised that the CRCSI has limited resources so careful thought has been given in this strategy to ensure investments are made in research that will yield the highest benefits.

This strategy is designed to be entirely consistent with and support the 2012 (revised) milestones of the CRCSI's Commonwealth Agreement.

2.0 Representing SDIs as Supply Chains

SDIs can be represented as supply chain architectures that involve processes that run from data collection through to production based on the data. A CRCSI commissioned investigation (van der Vlugt, 2012) into supply chain capabilities proposed the model shown in Figure 2 in which a number of stages are identified. Figure 2 shows the variability in data from suppliers and products used by consumers. Of importance are the feedback loops that are needed for quality assurance and performance monitoring. This model applies at various levels including the jurisdictional level as well as the national level. A modification that shows how the ANZSM and NSDI may be represented as a hierarchy of supply chains is shown in Figure 3. At the national level, the supply chain makes provision for various data stores as well as distributed datasets at a jurisdictional level using a federated model. Importantly it is foreshadowed that this arrangement operating through the Semantic Web will satisfy the needs of the ANZSM and the NSDI.

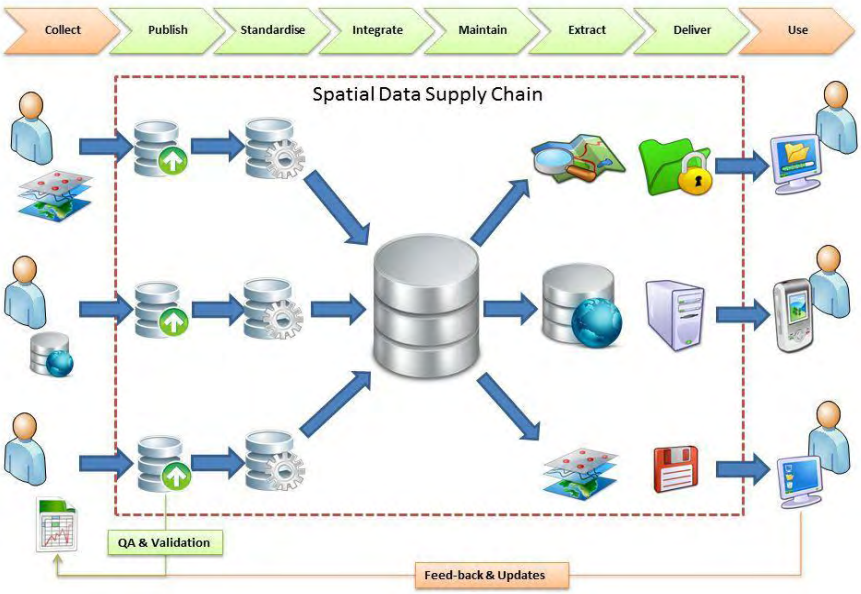

Figure 2. Spatial Data Supply Chain (van der Vlugt, 2012)

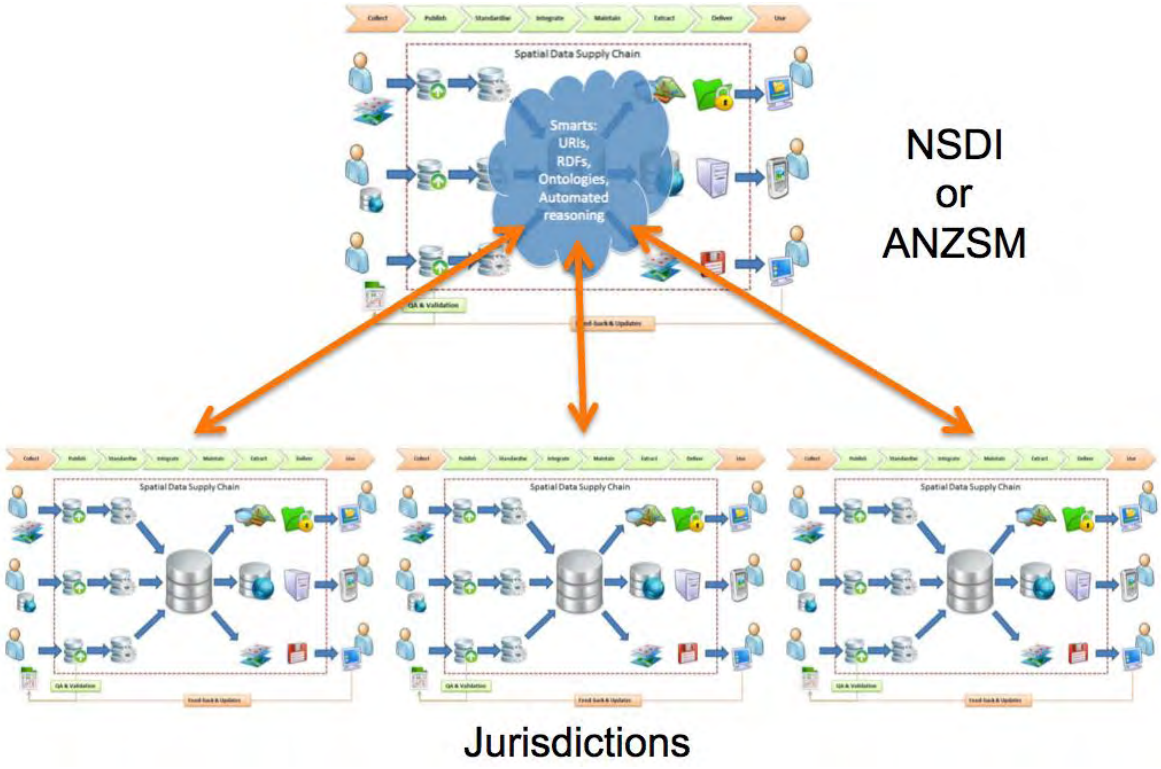

Figure 3. Hierarchical Spatial Data Supply Chain (after van der Vlugt, 2012)

3.0 Research Drivers

This strategy is responding to the following drivers:

Usability: It is widely held in computing circles that the two best user interfaces produced over the past few years have been the Google home page and the Apple iPod clickwheel (generation 4 and after). The Google home page has few words and a text window. It hides enormous complexity from the user including Google's private network, its complex and evolving search engine, its index and its copy of the Web.
The iPod clickwheel allows very fast manual searching for media such as songs and movies using one finger. Both now have hundreds of millions of users. The key to their success has been their great usability.
Apple, Google and others are still striving to build better user interfaces using new approaches (ACM[1]). Usability success is very much a product of a more user focussed approach rather than the more supplier driven approach to current SDIs and is very much seen as critical to the success and popularity of devices and software services. Excellent usability is required for many aspects of spatial infrastructures and in particular the ANZSM to attract users and to enable fast searching for data processes and processing chains, as well as in publishing e.g. through metadata creation. A processing chain is needed when the required spatial product is not available but can be generated from existing data using one or more processes. Improved search and discovery capabilities for data are seen by many stakeholders as a priority.

[1] Proceedings of recent Annual ACM symposia on user interface software and technology (latest occurring in 2012).

Automation: Spatial infrastructures, and specifically SDIs, are still very much manually configured. Data has to be acquired and manually linked or more likely copied to the infrastructure. Data integration, required when dealing with data from numerous sources, is, in the main, a manual process. There are rule engines available that help with this process although these have to be manually configured. Searching uses simple techniques and cannot help a user find data that they don't know exists. The automatic configuration of intelligent search will make it far easier to find data and to find other relevant attributes within the data. There is an increasing move towards server-sided and cloud based processing that allows data to be processed *in situ*.

This offers substantial efficiency gains as it avoids the difficulties of moving around the huge volumes of data being generated. It also offers improved security and protection of intellectual property. For this to occur, processes are required that enable a user to build and invoke process chains to generate derived products. The proposition is that increasing the automation of the procedures and workflow around SDIs is critical to realising increased access and use of spatial data.

Web-based processing: The world is moving away from the traditional personal workstation approach to computing, to one where very lightweight clients (phones, tablets) are becoming prominent. These rely on web-based services running on servers somewhere on the Internet. This is having increasing impact on the spatial discipline because of the very large amounts of data that are available and needed to power spatial applications and fulfil ever more demanding end-user needs. A well-known example is Google Maps that makes use of tiling and other methods to deliver to the client just what is required to be displayed thereby keeping network bandwidth requirements to a minimum. Google Earth (although not really a thin client application) does the same for many sources of data covering 2D vectors (roads etc.), imagery (satellite etc.), 2.5D data (surface and terrain), and 3D models (buildings etc.). Note that if we can achieve a good implementation of the Web Mapping Service (WMS[2]) OGC standard and a judicious design of the associated tiling and caching regime, we can achieve very substantial efficiency gain despite bandwidth and online processing capacity limitations.

SDIs have to move beyond just hosting and allowing downloading of data, to linking to data sources, providing access to web processes (hosted or via links) and orchestration of processing chains. This implies the need to include intelligence in some form to build what is increasingly becoming known as Spatial Cyber Infrastructures or Spatial Knowledge Infrastructures. In addition to web services for access to data, web processing service standards are being proposed including the OGC's Web Processing Service standard (WPS[3]). WPS will allow the execution of processing chains. We can help shape these standards and ensure they operate to maximum utility in the context of the ANZSM.

[2] Web Mapping Service (WMS): an OGC standard that allows map boundaries to be moved and displayed across the Web.
[3] Web Processing Service (WPS): an OGC standard that allows processing to occur over the Web.

Search for web-based services: Search and discovery of web services has been identified as an important issue to be addressed given the relatively poor search mechanisms used in SDIs.
It is a requirement for any user to find data (based on developing standards such as WMS, WFS[4], WCS[5]), processes (using WPS), as well as their capabilities. Other resources should be discoverable as required, even if they are not spatial but can be combined with spatial data for the required task. Catalogues can be used to store information about the location on the Web of the resources (these are known as Universal Resource Identifiers (URIs)) if such resources are not available on your local computer.

The software platform of ANZSM, Geonode, currently uses a full text search engine called Lucene[6] and elastic search[7]. This is scalable and can index large collections of text-based documents. Maué (2008) proposed multiple options for search and support schema and ontology based solutions because of the diversity of representation of web services and the lack of performance when using just one method. Of interest is the identification of the need to repeatedly check to make sure the catalogue is up to date. Lopez-Pellicer et al (2011) carried out research into searching for OGC compliant web services across Europe using search terms such as "getcapabilities wms soil Italy", They found 6,544 SDIs showing that Google, Yahoo! and Bing searches can find useful spatial information. They argue for more research into more intelligent search techniques as the SDI research world has not paid much attention to the automatic discovery of web services. Searching will be enhanced by the linking of data and processes to each other to allow related resources and properties to be identified.

Web service orchestration: Orchestration is where data and processes on the Web can be linked together to satisfy a user's requirement or query. Orchestration is required when a suitable dataset is not available but could be generated from available datasets and processes. The development of intelligent reasoning and inference is needed to find the required data and processes, and generate the required processing chain to satisfy the user query or request. The goal is to generate new and accurate knowledge via this orchestration or "mash-up" process. This orchestration of web services and data is currently mainly a manual process once data and processes have been identified. There is a need to eventually automate orchestration. Orchestration requires knowledge of the location of data and services which requires better searching methods and the processing of ontologies to find the compatible links between data and processes.

An example of the use of manually configured web services is given by Wiemann et al (2012) for a simple processing chain. They describe how WPS is used for visualisation and schema transformation, WFS is used to provide data, WCS is used to provide ortho-images, and an OGC Catalogue service is used for publication and search. These can run on more than one server. They identify the need for more research into schema mapping and transformation, and in particular, into semantic schema transformation using ontologies. In

[4] Web Feature Service (WFS): an OGC standard that allows map attributes to moved and used over the Web.
[5] Web Coverage Service (WCS): an OGC standard that allows operations across the Web to multiple different platforms of the user and the source of data and services.
[6] http://lucene.apache.org.
[7] http://elasticsearch.org.

addition, they recommend research is needed into the Rule Interchange Format (RIF), and Simple Object Access Protocol (SOAP) and Web Service Definition Language (WSDL) for transformational services. Jung, C.-T. & Sun, C.-H.,(2010) present an example of automatic orchestration in a limited application using ontologies that describe geo-processes, data types and sequences of tasks for emergency management.
The Business Process Execution Language (BPEL) has been proposed to orchestrate processes. Gone and Schade (2008) identify the limitations of BPEL for orchestrating web services and proposes a Web Service Modelling Ontology (WSMO) approach and identifies a need for research into “on the fly” orchestration.

Searching, integrating and publishing data and orchestration has been investigated in other disciplines, for example, Semantic Automated Discovery and Integration (SADI) (Wilkinson et al, 2011) has been investigated for the discipline of bioinformatics. The data is not spatial but there are similar needs to the spatial community.

When considering the processing of data, it is necessary to consider the access to the data and any derived products. Many datasets are large and it is costly and slow to transfer the data between machines. Consider a WPS that the user wants to run over a dataset. It may be better to move the code associated with the process to where the data is stored. Information that decides this needs to be identified (may be in the metadata for the WPS or obtained via a *getcapabilities* query to the WPS). This and other knowledge about the orchestration needs to be encoded in some form to allow automatic determination of the best approach. This may be extended to cover the use of cloud based processing when dealing with big data.

Use of crowd sourced data and its integration with authoritative data: It is becoming more desirable for a number of reasons (speed, data currency, cost, commercial gain) to consider crowd sourced spatial data, either voluntary (VGI) or involuntary. Crowd sourced data from Twitter feeds has been shown to be useful in emergency situations (20-30 minutes quicker than more traditional intelligence (Mark Wallace, per. comm., 2012)). It can also be used to correct or augment authoritative data. Of course VGI needs good quality control (Goodchild, 2007). An example of the use of relatively simple ontologies and Semantic Web technologies operates at present to identify and resolve errors between Open Street Map and the UK’s Ordinance Survey road data (Du et al 2011). Diaz et al (2012) propose a technique for searching for VGI data, including real time data, using OGC’s Open Search Geospatial and Time specification. Examples include Twitter, Flickr and Wikipedia. An example is described for queries concerning forest fire monitoring. Flickr produced geo-located pictures of the fire. Another recent development is the OGC’s Open GeoSMS that allows the encoding of location content in a short message (the GeoSMS app available for the iPhone) and allows the location to be shown on Google Maps. Overall there are a number of avenues to explore the use of the Semantic Web, W3C and OGC standards to make more use of crowd sourced data.

The Semantic Web: Semantic Web architectures and its components are seen as the next stage in the development of SDIs, in particular the NSDI and the ANZSM. We’re moving away from just data provision and simple search mechanisms to intelligent, easy and simple to use infrastructures that can be termed Spatial Cyber Infrastructures. The Semantic Web approach essentially allows much desired functionality to be realised.

For example, the linking of data and processes through metadata and ontologies[8] will enable data to be discovered using more natural language queries. If the right product isn't available to satisfy the query, it will generate, suggest and invoke processing chains using web processing services and available data. Ontologies will allow more abstract knowledge of datasets, processes and other components to be described and allow intelligent processes to automatically and semi-automatically integrate datasets and services.

Research is required into many aspects of the Semantic Web in the context of spatial data because of the Web's increasing reliance on standards including OGC standards, their complexity, and the need to consider the orchestration of data and processes. Spatial data supply chains require the use of OGC standards to enable data to be extracted from a data store e.g. using a WFS, and then processed by a WPS that must have compatible output and input data formats. It is observed that the spatial community has no rigorous way of describing functions, and hence no effective means of searching for them (Goodchild 2012). If we want to make better use of WPS to deliver spatial functions in future, we have to facilitate their search through suitable standardisation or use of ontologies. The research will need to consider the lower levels of the Semantic Web stack (ontologies, reference data frameworks) as well as the higher levels (rules) to build in intelligence to aid the user. These rules may cover legal rules for licensing, copyright, pricing etc.

Data and other resources can be uniquely located on the Web (via Unicode and URI). They can be operated through the use of machine language processes (such as XML). The resources can be more readily discovered because their metadata are described through the use of the Resource Description Framework (RDF) which also facilitates machine-language interfaces. Ontologies facilitate the organisation of the data for their use with applications and processes. Importantly, when set up in this way the resources are capable of being accessed through the use of plain English or natural language commands that are translated into machine language for the purpose of analyses, and lend themselves to much intelligent manipulation by machine. This enables intelligence to be built into systems that greatly facilitates accessibility for a very wide and much larger range of end users. The plain English interface opens up the Web to far more users.

The main components of the Semantic Web are outlined in Appendix 1.

Currently within the Semantic Web much of the development of the semantically enabled data is generated manually. It is this manual generation of data that is slowing down the development of the Semantic Web and especially within the spatial domain. Data sets within the spatial domain tend to be extremely large and hence require much annotation if they are to be useful beyond the specific purpose for which they were collected. This annotation requires considerable human resources. Automation is emerging as a promising and high priority research area for building descriptions of the data into the metadata.

The use of ontologies to enable automated resource generation is another high priority area of research. Reasoning and/or query processing can be applied to these new resources to generate new knowledge. Importantly links can be generated between this new knowledge and knowledge already in the system to generate even more knowledge. The linking of data and knowledge is an important advantage of the Semantic Web. Understanding and acting

[8] Ontologies, in the context of information science, describe relationships and concepts between features common to the topic of interest. They typically produce a taxonomy (and hierarchy of relationships) that can be readily interpreted by all users. They are now a fundamental requirement of any architecture in any enterprise providing a framework for handling information (known as an enterprise architecture framework).

on the knowledge and data links requires machine learning and pattern recognition algorithms.

Multiple information systems need to be seamlessly working together in order for the Semantic Web to work. A useful way to achieve this is through schemas. Schemas need to be openly available and clearly defined. It will then be possible for them to be processed automatically and the linking between different information systems to occur.

Semantic Web search will place much higher demands on systems because of the way people will expect data to be made available and useable. Software struggles to process natural language because of ambiguities in meaning that can occur in sentences or statements. Natural language search engines should not just return web pages. They should return meaningful results by understanding what has been searched for, and then from the semantic understanding of linked data, return results that are far more meaningful than simply a list of web pages containing information on them.

“Ontologies are essential to the Semantic Web as they provide interpretation to its data contents” (Qin & Atluri, 2006). Over time ontologies change both semantically and structurally. If they are not evolving then they are not reflective of the dynamic state of most data. Several issues are important including the detection of the changes in the ontologies as well as the evolution of the ontologies. The evolution of ontologies is important because changes in one area of the Web have consequences elsewhere on the Web. An intelligent, Semantic Web anticipates these changes and automatically makes adjustments. A challenge is to structure the data, and other resources to enable the power of the Semantic Web to be available to users. The manual changing and evolving of ontologies and RDF triples (a technique of linking two resources together) is simply an overwhelming and daunting concept. Automation is the only long term option.

Research into both the natural evolution of ontologies and RDF as well as the automated detection of the evolving ontologies and RDF is needed. An ontology may need to change for the simple reason that the world has evolved (Stojanovic et al, 2003). Another reason for the change in an ontology is that there is now a different view or interpretation of a domain e.g. a road network. Components of a dataset or the database model may have changed, and hence a new perspective is needed (Noy & Klein, 2004).

A powerful technique that can be used to analyse and make sense of knowledge represented as ontologies and RDFs is machine learning. Machine learning can find patterns in the knowledge that can be used to aid the user e.g. decision trees (if…then…else rules) can be used in a question/answer session to help design mash-ups. Machine learning is typically carried out on static data sets and there are challenges when considering changing data, ontologies and RDFs.

Federated Data: The reality of contemporary data management is that data sets of common heritage (e.g. roads, cadastre, DEMs etc) are stored in many agencies and organisations, the component parts of which when stitched together make up the contiguous whole.

These are federated data. Federated models are therefore necessary when considering the combining of common data from a number of agencies in a jurisdiction as well as nationally across jurisdictional boundaries or their equivalent in the private sector. The different schema used for the data by different organisations means that complex data models have to be built to aid in building mapping systems to unify the data in higher level schema. Such data models are currently built manually. Dealing with changes in the schema at different

organisational levels is problematic, costly and, in many cases, performed manually. Research is needed around the maintenance of federated schemas to quickly enable changes to propagate without any apparent changes to the users' view of the highest level schema. Every one of the fundamental data sets in the NSDI requires a federated data model in order to unlock its full benefit to users across the nation (in the case of Australia).

4.0 Research Issues

Following on from the discussions above, a number of important research issues of high priority for us have been identified for consideration:

Automatic Metadata and Ontology Generation and Evolution: Most of the metadata and ontologies available have been hand crafted, in many cases using software tools to improve the workflow. Much metadata contains errors that can render it useless. There is a need to consider semi-automatic and automatic generation from the data and other information.

Florczyk et al (2012) describe a prototype system for the automatic generation of metadata for the web pages of providers of geospatial web resources even when these resources are not present in SDIs. This is potentially of great assistance to us and other users of the ANZSM because it could save very substantial amounts of time in not having to manually generate these metadata and also helps ensure that the metadata are actually created, a real problem at present. The aim is to automatically make such data available for spatial analysis. A web crawler is used and a search made using heuristics for data in web pages that has spatial information. Results are deemed to be acceptable for more than 80% of the tested web resources. Research into improving such methods is justified because of the difficulty of the current approach to manually generating metadata.

Bedini and Nguyen (2007) review automatic ontology generation techniques e.g. from XML files and show there is much research needed. For specific domains there has been some success. For example Gantayat (2011) demonstrated the production of ontologies from lecture materials.

Semi-automated and automated data integration: When considering data integration, it is necessary to identify the different ways this can be achieved:

- Federation: Data is combined at say, the national level. The data custodians are still responsible for the underlying data and are allowed to change the data model to suit their circumstances. Data should not be copied to a central repository but accessed as needed from the data custodian. Federated models are also used between LGAs and land agencies in which LGAs own and are responsible for some data types such as property footprints.
- Harmonisation: A common data model is agreed upon by all the data custodians. Examples of initiatives for this are CityGML and GeoSciML.
  For data harmonisation, Atkinson et al (2007) proposed a number of design patterns to construct data models across domains using UML to represent ontologies.
- Aggregation: Data is copied from data custodians and combined (by transformation) into a new representation. This is the approach used by PSMA. PSMA uses 1Spatial's Radius Studio to manually create a large number of rules to allow

automated integration. An example of a basic rule that always applies would be for the topology of boundaries.

- Brokering: This recognises that data custodians and potential data users do not have the resources or ability to generate the right data models to allow compatibility between them. A 3rd party brokering technique is used to do the transformation and allow communication to occur. This is the technique proposed for the EuroGEOSS project (Nativi et al, 2012). Brokering is argued to be better for linking cross-disciplinary data. The argument is that mutual respect and mediation between the various data and service providers and users is better than standardization or federation.

Although federated models have been identified as the main focus, awareness of the other possible models and how aspects might contribute to the research into federated models is important.

In terms of research, Cruz et al (2004) propose the use of ontologies to enable the generation of mappings and schema to combine datasets from different organisations. This is a semi-automated process requiring identified differences in the data revealed in the ontologies to be resolved. The objective is to do as much automatically but also to have methods to improve workflow for manual interventions. Other research communities have also addressed the use of ontologies for data integration. For example, in the bio-medical area, the Open Biomedical Ontologies (OBO) consortium is pursuing the generation of a number of standard ontologies (60 currently) to allow the sharing of data. Along the same lines, Jones et al (2006) identified the integration of heterogeneous data as a new area in bioinformatics.

Where data is sitting on some central repository, techniques are required to keep the various copies in sync. There is potentially benefit to be gained in the exploration of OGC transaction services e.g. WFS-T for this, as well as the new OGC synchronisation standard.

Automated Schema Evolution: With federated models and querying, there is a need to be able to allow the modification of local schema and automatically change federated views to keep legacy queries built on the local schema working. Waters et al (2011) propose schema transformation based on web services. They explore different aspects of the problem and recommend the use of GML from OGC for data and schema descriptions, and RIF (Rules Interchange Format) from W3C for schema mapping definitions. Note they propose a hybrid approach using standards from different standards bodies, and not just from the W3C. Research is ongoing on schema evolution in which dependencies across federated datasets are held centrally, schema changes are proposed, and changes to federated queries made to accommodate the local changes (Xiaoying et al, 2012). Manual and semi-automated process needs to be modified to enable full automation.

Open standards: Many projects in Program 3 (and other research themes) will rely on the use of open standards. It is proposed to consciously introduce and use standards including those from OGC, ISO, W3C and OASIS[9] throughout the duration of the program and as a

[9] OASIS: Organization for the Advancement of Structured Information Standards: https://www.oasis-open.org/

matter of principle for the ANZSM and NSDI. There is a deluge of standards discussed in the many papers cited in this report and the choice of the most appropriate ones to use is a significant issue. The latest standards should be investigated and contributions made to current and new standards where there is benefit in doing so. An example is a proposed OGC standard to keep data consistent. In the scenario in which copies of a dataset are kept in other locations, it is necessary to keep one or more copies of the dataset updated with the latest changes even when unreliable or non-existent electronic communication channels exist e.g. access to the Internet is sporadic. Implementing and using the standards are some of the best ways to discover limitations and identify opportunities for improvements to be suggested to the various standards organisations. Apart from the standards already mentioned, standards of interest include OGC security and privacy standards, the OGC geo-synchronisation standard (that allows local copies to be kept up to date from authoritative sources), and OGC emergency standards.

Integration of 3D and 4D data: Most SDI development has been developed for 2D data and, in some cases 2.5D data (DEMs for example). There is increasing demand for the inclusion of 3D and 4D (3D and time) data. The representation of 3D and 4D data is a mature discipline with such data processed and visualised in CAD packages. There are opportunities for research into the integration of 3D and 4D data with more traditional 2D and 2.5D data by integrating GIS and CAD systems. Integrating indoors and outdoors through GIS and BIM (Building Infrastructure Models) is a hot topic and much exploration is being carried out on BIM and its usefulness. 3D and 4D data complicate SDIs and the ANZSM because of metadata, visualisation and processing issues. For example there is a need for WPS that works with 3D and 4D models using CityGML, KML etc.

User interface issues and usability: He et al (2012) investigate quantitative and qualitative methods to assess the usability of an SDI, specifically the Swedish national portal Geodataportalen, and target effectiveness, efficiency and satisfaction based on ISO 9241-11. Although quite successful, a number of research questions are identified including the need to provide better feedback to users during trials, and extending the method to deal with user interfaces more targeted at the general population and not just domain experts, and dealing with different age groups and types of user.

SDIs, the NSDI, and the ANZSM will benefit from learning how users interact with the systems. This is well known for search algorithms, many of which change the order of the returned results based on the popularity of learned links between resources. This relies on logging user interactions.
There is much research possible into the analysis of such logs in a complex web-based system such as an SDI or the ANZSM.
The logs can aid user interaction through case based reasoning (CBR) in which user interactions are stored and used to prompt and guide future users e.g. by suggesting datasets others have used given a similar query. This builds more intelligence into the systems.

There is potential for server-sided visualisation tools to be developed for various aspects of an SDI or the ANZSM, moving away from the traditional client-sided approach. The objective would be to remove the dependence on client-sided plugins and applications, and allow the use of lightweight clients such as tablets and smart phones. Visualisation of orchestration, mash-ups, 3D and 4D datasets as well as various analyses is possible. Some progress on complex and relevant server-sided visualisation tools has been carried out in the CRCSI project P4.4.1 (Moncrieff and West, 2012).

Licensing, Copyright and Terms of Use: There has been much research into the issue of licensing for geo-spatial data including prior research by the CRCSI. Pressure has been brought to organisations to free up their data and the associated licensing arrangements by organisations such as Open Street Map. Creative Commons (Fitzgerald, 2010) has become popular for many datasets. Data Commons[10] is a more liberal standard that allows the combination of data products to form derived products. Dealing with different licensing agreements is a crucial part of orchestration because there is a need to deal with combining data to produce derived products. At GSDI 12[11] in Singapore in 2010, it was agreed to set up a committee to explore the issue of global licensing harmonisation. At GSDI 13[12] in Quebec, a workshop discussed progress (van Loenen, 2012) towards harmonisation using the various licensing schemes available. This is a potential research area through being involved in the GSDI working group.

Given that there are a number of licensing schemes in use, an issue is then around "terms of use". Each data set used to generate derived products may have specific conditions covering how it can be exploited. A mechanism is needed to enable terms of use to be encoded in a form suitable for machine examination so that allowable combinations of data in derived products can be determined. It is also necessary to consider making such terms of use available for human inspection along with protocols to allow a user to agree to the terms of use before access to the data and derived products. The terms of use need to be formulated in such a way as to allow a user to easily understand them. Once the terms of use are agreed to, then the data suppliers may, or should not be, liable for misuse etc.

There is potential for technical solutions for dealing with licensing. Automation when considering combinations of data in process chains has been investigated by Onsrud (2010). The scenario painted is a person finding a number of datasets of relevance, each of which has a different license. A user interface allows the user to decide on the terms of use given the licenses. There is potential research into the automation of the process using Semantic Web technologies to advise the user on how to reconcile conditions over all the licenses and obtain a solution, and to advise if one is not possible.

In terms of policy, Janssen and Kuczerawy (2012) propose a code of practice for the European Union that recognises that multiple license models are needed for public sector data. Two license models are proposed that cover free and paid access. Other license models are discouraged. Issues such as terms of use, charging and metadata are considered in the formulation of license models and the realisation that existing practices by

[10] opendatacommons.org
[11] Global SDI Conference: Conference of the Global SDI Association, Singapore, October 2010.
[12] Global SDI Conference: Conference of the Global SDI Association, Quebec, May 2012.

data custodians may need to be preserved. Research identified includes the need to address questions such as how can licenses be harmonised and comply with legal requirements, and how can the culture and approach of public bodies to data dissemination be changed so that only a limited number of general policies are used. Mäkelä (2011) proposes a systematic approach to the development of a licensing and pricing model for the DEM across Europe that is very much focussed on customer needs.

5.0 Proposed Research Program

The integrated nature of the research challenges leads to the proposal of a single overarching research project. It is proposed that it should be made up of sub-projects that reflect the need to schedule the timing of activities with respect to developments in the Semantic Web happening elsewhere, the availability of researchers with the requisite skills and the needs and priorities of key stakeholders. Use cases are proposed to demonstrate the timely delivery of outputs. The sub-projects address the key drivers or usability, automation and the adoption of web-based processing. They also address the outcomes of the joint ANZLIC/CRCSI workshop of April 2012 that identified the Semantic Web, Federated Models and Web-based Processing as the main foci for the research.
It is proposed that the Science Director, Program 3 will be the overall Project Leader. The Project will report to the Board of Program 3. Sub-project advisory groups will be established as needed. Because the P3 Board sits with the ANZSM Steering Committee, governance structured in this way ensures tight control over the research and good linkages with the key bodies and stakeholders.

Sub-Project 3.01: Semantic Web Technologies for the Spatial Marketplace

This project is concerned with research that adds intelligent functionality to the ANZSM. This will cover Semantic Web research to enable a spectrum of users to easily access and use the marketplace. Specific aspects of this project will cover activities concerning data and processes, namely search and discovery, publishing, and their orchestration to satisfy user queries. This will involve the following specific research topics

- Search for web-based services
- Web service orchestration
- Automatic metadata and ontology generation and evolution
- User interface issues and usability

The overall objective is to advance Semantic Web technologies to enhance the ANZSM. Much research has been carried out into the lower levels of the Semantic Web (RFDs and Ontologies) but there is a need to increase the sophistication of Semantic Web based systems through researching into the higher levels including rules and how ontologies and metadata are used for decision making and support. The proposed next version of the ANZSM will follow a supply chain model. A supply chain represents the SDI process from data collection through to production of products based on the data.

Sub-Project 3.02: Semantic Web Technologies for Federated Data Integration

This project is concerned with research to solve issues involving the generation of national fundamental datasets as proposed by ANZLIC and lead to satisfying the needs for a NSDI. The fundamental data themes are (ANZLIC 2011[13]):

- Geodetic network (Positioning Services)
- Property Boundaries
- Address – Physical/Allocated/Postal
- Transportation
- Geographic Names (Gazetteer)
- Elevation/Relief
- Imagery
- Administration Boundaries
- Hydrography/ Bathymetry
- Hydrology – Surface Water Features
- Land cover (built environment; and vegetation)

Data is owned by a number of different jurisdictions and agencies and are represented using different schemas, structures, vocabularies and concepts. This vertical integration requirement from jurisdictions to a national level needs techniques and tools developed to seamlessly combine the data to present a consistent view to the user. This vertical integration fits in with the hierarchical supply chain architecture of Figure 3. The project will involve the following specific research topics

- Automatic metadata and ontology generation and evolution
- Semi-automated and automated data integration
- Automated schema evolution
- User interfaces issues and usability

The overarching objective of this sub-project is to automate the data integration as much as possible and at the least, provide better tools to aid in the process. Although targeted at Australia, the project will encompass the needs of New Zealand.

Sub-Project P3.03: Jurisdictional Level Data Integration

Research into the efficient operation of supply chains both vertically and horizontally is important at the jurisdictional level. Vertically, there are issues with the integration of various datasets from organisations such as LGAs into state wide datasets. Horizontally there are issues with respect to combining datasets from various state agencies e.g. linking parcel data to road or environmental health data. At the vertical level, it will also include research into the issues of crowd sourced data for enhancing and augmenting the authoritative data. The alignment study (van der Vlugt, 2012) and Figure 2 show the supply chain model at a jurisdictional level, the need to consider different data sources and the need for efficient quality assurance and feedback tools and methods. The research will involve the following specific research topics

- Use of crowd sourced data and its integration with authoritative data
- Automatic metadata and ontology generation and evolution
- Semi-automated and automated data integration

[13] As accessed on the ANZLIC website 28 June 2012

- User interfaces issues and usability

A number of issues arise when considering integration that will be addressed in this sub-project including: identifying manual methods that can be automated; identifying where feedback mechanisms are needed; modelling data flows and how the data changes (geometrically, relationally, graphically, behaviourally and semantically) along the supply chain; determining at what stage along the supply chain should data be integrated; and addressing the question of where should the point of product differentiation occur to ensure supply chains are cost effective and provide most value to consumers in the form of a knowledge-base for derived information products.

Sub-Project P3.04: Automated Determination of Licensing Models for Spatial Information

The widespread use of datasets and processes is being held back by complexities to do with licensing, costing and terms of use issues. This is especially important for the ANZSM that will have datasets and processes that have the same or similar functionality. Currently much spatial data is released covered by a number of different licensing schemes e.g. one of a number of different Creative Commons versions. Some licensing schemes are more liberal than others. Understanding licensing terms is reasonably easy when considering one dataset for use in one application but becomes increasingly complicated when combining different datasets and processes from different organisations to produce derived products.

Research is needed into tools to help a data producer or consumer understand and deal with issues to do with licensing, terms of use and costing to enable them to make the right choices for their requirements. This will involve the following specific research topics:

- User interfaces issues and usability
- Licensing, Copyright and Terms of Use

Organisation of the sub-projects

For each of these sub-projects, standards such as those from the OGC will be used as much as possible. It is not proposed to research and develop new standards but to investigate their usefulness and identify any shortcomings that can be addressed through modifications to the standards. There is commonality across the projects in terms of some of the specific research topics. These topics are seen as fundamental to each project and cannot easily be isolated and investigated in their own right. There is much modularity in the projects and there are opportunities for cross fertilisation to occur between them. It is therefore proposed to generate teams of researchers across the participating universities, government agencies and commercial organisations coordinated to tackle each project. This will be a planned activity and follow identification of the most appropriate participants and their invitation to contribute. The first project proposal will be generated for Project P3.01, followed by Project P3.02. Project P3.01 will have completed the CRCSI's approval processes by the end of September 2012 for commencement in January 2013. Project P3.02 will follow for commencement April 2013. Projects P3.03 and P3.04 are likely to commence July 2013.

6.0 Use Cases

The following use cases are examples of how the above projects will satisfy the needs of a number of users and applications. The list here is by no means complete but gives an indication of the relevance and importance of the proposed projects.

Disaster and Emergency Management

Emergency and disaster management requires the agile access and interpretation of data, much of it of a spatial nature. Examples of traditional data include static data such as street, waterways, terrain information, as well as real time temporal data such as flood levels and wind speed. Non-traditional data include Twitter feeds and Facebook blogs. There is a need to be able to identify the resources available, how they can be processed and combined to satisfy the objectives at high speed and ideally in real time. It is recognised that before an emergency, depending on the emergency, the actual data sets required may not be envisaged. **Project P3.01** is relevant to this use case because the many datasets available need information extracted to enable them to be discovered and incorporated into processing chains and mash-ups, given an emergency specific query.

Real time social network data streams will need filtering techniques to identify the right "tweets" etc. for the particular emergency. For example, if the emergency is concerned with bushfires, then bushfire specific words need to be identified such as smoke, flames and specific phrases such as "fire is close to the road". **Project P3.02** is relevant to this project because of the need to access national datasets because, in many cases, emergencies and disasters do not obey jurisdictional boundaries. Some aspects of **Project P3.03** are relevant, especially concerning the need to deal with crowd sourced data. In an emergency situation, the issue of licensing, terms of use and pricing will be expected to have less importance that for normal operation but there are aspects of **Project P3.04** that are relevant.

Federated Data Integration

The Department of Climate Change and Energy Efficiency, following a request by COAG initiated a project to develop high resolution DEMs for most of coastal Australia. The CRCSI was commission to manage the project. It is currently well progressed and involves many of the partners of the CRCSI. The task is to integrate the many DEM data sources of different resolutions, different classifications, and dispersed locations to allow integrated access for particular areas and uses.
Project P3.02 is relevant because of the need to automate the mainly manual integration and maintenance processes (currently scripts etc. are being generated for integration). All the data will be hosted at GA (many terrabytes) because some jurisdictions and other data suppliers do not have data online.

Much of the data can be hosted by the original acquirer (jurisdictions for example) and linked through the Web. Updates to the source data can be propagated through to GA when they occur. This objective is relevant to **Project P3.02** as part of a federated model. In addition, using ontologies and schema evolution methods can accommodate changes in data format. Licensing of the data may be irrelevant because the availability of a national dataset means

licensing has been successfully negotiated and a national license is available. However **Project P3.04** may be relevant if difference licenses are used for the different datasets.

Data Integration at the Jurisdictional Level

At the jurisdictional level data integration is multifaceted and complex in nature. This has led to convoluted data supply chains, data inconsistency, inefficient data usage and duplication of effort across the sector. State land authorities generally acquire data from LGAs, as well as, other state agencies and commercial providers. The management of this large stakeholder group is unwieldy, especially when one considers the assumption of a land agency owning and controlling authoritative data and the legal ramifications. For example in Western Australia, there are 141 LGAs feeding location information into Landgate.

In addition, the point of differentiation of end user information products often occurs at intermediary points along the value chain leading to aspects of information loss and the spatial industry's inability to build highly developed geographic knowledge-bases for complex decision making. In some cases, there are multiple agencies acquiring the same or similar data e.g. there are 22 agencies that collect points of interest data, five of which collect fire hydrants at varying degrees of accuracy and semantics, and several agencies that define their own high water mark line for infrastructure management and planning. This raises the question 'which agency holds the authoritative source?' Versioning of data is another major issue with different agencies either holding datasets that are not up-to-date or updating their own copies and not propagating the changes back to the authoritative data custodian.

To overcome these issues research needs to generate methods that can be integrated into a sustainable canonical supply chain allowing for multiple data inputs, quality assurance and versioning methods so that the whole 'supplier to user' process can be managed in a more controlled way. **Project P3.01** is relevant to this use case because of its focus on the marketplace being a supply chain and the use of automation to gain efficiencies in supply chain control. **Project P3.02** is relevant because many of the authoritative datasets at a state or territory level need to feed into the national datasets. **Project P3.03** is particularly relevant because it concentrates at the jurisdictional level and deals with vertical and state-based horizontal integration of data. **Project P3.04** is relevant because of the different licensing schemes that will be applicable to different data sources and derived products e.g. database schemas built across databases from different agencies.

7.0 Use of the Established Software

There is much open source software available that this research can use and build upon. In particular, the ANZSM demonstrator has produced code (mainly based on OpenGeo and Geonode software) that can serve as a platform on which to build.

This is particularly attractive as a model because it can be maintained and improved upon incrementally by researchers e.g. new search techniques can be relatively easily included. This is because the software is modular.

There are a large number of spatial infrastructure projects occurring in Australia, some of which are concerned with spatial data and hence relevant to the proposed strategy. Examples include AURIN[14], TERN[15], ANDS[16] and NeCTAR[17].
There is much commonality across these programs as well as a lack of coordination. The research activities proposed here need to be conscious of these activities and explore potential benefits through collaboration at various levels.

An example of research infrastructure of interest is the development in Australia by the CSIRO of Semantic Web technologies for numerous purposes including the building of SDIs. In particular, the Spatial Information Services Stack (SISS) originally constructed for AUSCOPE[18]. It is planned to investigate the SISS and incorporate already developed Semantic Web tools including ontologies, vocabularies etc.

[14] Australian Urban Research Infrastructure Network: http://aurin.org.au
[15] Terrestrial Ecosystem Research Network: http://www.tern.org.au
[16] Australian National Data Service: http://www.ands.org.au
[17] National eResearch Collaboration Tools and Resources: http://nectar.org.au
[18] An organization for a national earth science infrastructure program: http://www.auscope.org.au/

References

ANZLIC, (2011), NATIONAL DATA THEME WORKGROUP, in National Dataset Framework - ToR's website version.pdf, Available at: http://www.anzlic.org.au/Latest+News/default.aspx, [accessed 28th June 2012].

Atkinson, R., Millard, K. & Arctur, D., (2007), Standards Based Approaches for Cross-Domain Data Integration, *Int. Journ. Spatial Data Infrastructures*, Vol. 2, 74-89

Bassiliades, N., Antoniou, G. & Vlahavas, I., (2006), A Defeasible Logic Reasoner for the Semantic Web, *International Journal on Semantic Web and Information Systems (IJSWIS)* 2, no. 1.

Bedini, I & Nguyen, B., (2007), Automatic ontology generation: State of the art. In *PRiSM Laboratory Technical Report.* University of Versailles.

Berners-Lee, T., (2000), Semantic Web - XML2000, Presentation available from: http://www.w3.org/2000/Talks/1206-xml2k-tbl/ [last accessed 12th June 2012].

Berners-Lee, T. & Fischetti, T., (1999), *Weaving the Web : The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor,* San Francisco: Harper San Francisco.

Chebotko, A., Lu, S., Atay, M. & Fotouhi, F., (2008), Efficient processing of RDF queries with nested optional graph patterns in an RDBMS, *International Journal on Semantic Web and Information Systems (IJSWIS)* 4 (4): 1-30.

Cristani, M. & Cuel, R., (2005), A survey on ontology creation methodologies. *International Journal on Semantic Web and Information Systems (IJSWIS),* 1 (2): 49-69.

Cruz, I.F., Sunna, W. & Chaudhry, A., (2004), Semi-automatic Ontology Alignment for Geospatial Data Integration, *GEOGRAPHIC INFORMATION SCIENCE, Lecture Notes in Computer Science*, Volume 3234/2004, 51-66, DOI: 10.1007/978-3-540-30231-5_4.

Daconta, M.C., Obrst, L.J. & Smith, K.T., (2003), *The Semantic Web : A Guide to the Future of XML, Web Services, and Knowledge Management,* illustrated ed. Indianapolis, Ind.: Wiley Pub.

Diaz, L., Nunez-Redo, M., Gonzalez, D., Gil, J., Arago, P., Pultar, E. & Huerta, J., (2012), Alternative search mechanism for web 2.0 resources, *Int. Journ. Spatial Data Infrastructures*, under review.

Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., Kanungo, T., Rajagopalan, S., Tomkins, A., Tomlin, J.A., & Zien, J.Y., (2003), Semtag and seeker: Bootstrapping the Semantic Web via automated semantic annotation, Proceedings of the 12th International Conference on World Wide Web (pp. 178-186). ACM Press.

Du H., Jiang W., Anand S., Morley J., Hart G., Leibovici D. & Jackson M., (2011), An Ontology Based Approach for Geospatial Data Integration of Authoritative and Crowd Sourced Datasets, *Proceedings of the 25th International Cartographic Conference*, Paris: 3-8 July.

Dubois, D., Lang, J., & Prade, H., (1994), Possibilistic logic. In D.M. Gabbay et al. (Eds.), Handbook of logic in artificial intelligence and logic programming, (Vol. 3, pp. 439-514), Oxford: Oxford University Press.

Fitzgerald, A., (2010), Legal Issues and Experiences from an Australia Perspective, *GSDI-12 Conference: Realising Spatially Enabled Societies*, Singapore, 19-22 October.

Florczyk, A., Lopez-Pellicer, F., Nogueras-Iso, J. & Zarazaga-Soria, J., (2012), Automatic generation of geospatial metadata for web resources, *Int. Journ. Spatial Data Infrastructures*, Vol. 7, pp 151—172.

Gantayat, N., (2011), Automated Construction Of Domain Ontologies From Lecture Notes, M.Tech Project Dissertation, Department of Computer Science and Engineering Indian Institute of Technology, Bombay, June.

Gone, M. & Schade, S., (2008), Towards semantic composition of geospatial web services – using WSMO instead of BPEL, *Int. Journ. Spatial Data Infrastructures*, Vol. 3, pp. 192—214.

Goodchild, M.J., (2007), Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0, *Int. Journ. Spatial Data Infrastructures*, Vol. 2, pp 24—32.

Goodchild, M.J., (2012), Personal communication.

Gruber, T., (1995), Toward Principles for the Design of Ontologies Used for Knowledge Sharing, Int'l J. Human-Computer Studies, vol. 43, nos. 5-6, pp. 907—928.

Gruber, T., (2007), Ontology of folksonomy: A mash-up of apples and oranges. *International Journal on Semantic Web and Information Systems (IJSWIS),* 3 (1): 1-11.

Hammond, B., Sheth, A., & Kochut, K., (2002), Semantic enhancement engine: A modular document enhancement plat-form for semantic applications over heterogeneous content, In V. Kashyap & L. Shklar (Eds.), Real-world Semantic Web applications (pp. 29-49). IOS Press.

Handschuh, S., Staab, S., & Ciravegna, F., (2002, October 1-4), S-CREAM—semi-automatic CREAtion of metadata, Proceedings of the European Conference on Knowledge Acquisition and Management (EKAW-2002), Madrid, Spain, Berlin: Springer-Verlag.

He, X., Persson, H. & Östman, A., (2012), Geoportal usability evaluation, *Int. Journ. Spatial Data Infrastructures*, Vol. 7, pp. 88—106.

Janssen, K., & Kusczerawy, A., (2012), Increasing the availability of spatial data held by public sector bodies: some experiences and guidelines from the OneGeology-Europe project, *Int. Journ. Spatial Data Infrastructures*, Vol. 7, pp. 249—276.

Jepsen, T.C., (2009), Just what is an ontology, anyway? *IT Professional DOI - 10.1109/MITP.2009.105* 11 (5): 22-27.

Jones, M., Schildhauer, M., Reichman, O. & Bowers, S., (2006) The new bioinformatics: Integrating ecological data from the gene to the biosphere. *Annual Review of Ecology, Evolution, and Systematics* 37: 519–44

Jung, C.-T. & Sun, C.-H.,(2010), Ontology-driven Problem Solving Framework for Spatial Decision Support Systems, *Proc. Global Geospatial Conference (GSDI 12)*, Singapore, Oct.

Koffina, I., Serfiotis, G., Christophides, V. & Tannen, V., (2006), Mediating RDF/S queries to relational and XML sources, *International Journal on Semantic Web and Information Systems (IJSWIS)* 2 (4): 68-91.

Li, N., Raskin, R., Goodchild, M. & Janowicz, K., (2012), An ontology-driven framework and web portal for spatial decision support, Trans. in GIS, Vol. 13, No. 2, pp 313—329. SDS website at http://www.spatial.redlands.edu/sds/

Lopez-Pellicer, F.J., Béjar, R., Florczyk, A.J., Muro-Medrano, P.R. & Zarazaga-Soria, F.J., (2011), A review of the implementation of OGC web services across Europe, *Int. Journ. Spatial Data Infrastructures*, Vol. 6, pp. 168—186.

Maedche, A., & Staab, A., (2001). Ontology learning for the Semantic Web, IEEE Intelligent Systems, 16(2), 72-79.

Mäkelä, J., (2011), Aspects of a licensing and pricing model for a multi-producer pan-European data product, *Int. Journ. Spatial Data Infrastructures*, Vol. 6, pp. 344—364.

Maué, P., (2008), An extensible semantic web catalogue for geospatial web services, *Int. Journ. Spatial Data Infrastructures*, Vo. 3, pp. 168—191.

Moncrieff, S. & West, G., (2012), An open source, server sided framework for analytical web mapping and its application to health data, Int. Journ. Digital Earth, Submitted February.

Nativi, S., Craglia, M. & Pearlman, J., (2012), The Brokering Approach for Multidisciplinary Interoperability: A Position Paper, *Int. Journ. Spatial Data Infrastructures*, Vol.7, 1-15.

Noy, N.F., & Klien, M.C.A., (2004), Ontology evolution: not the same as schema evolution, Journal of Knowledge Information Systems, Vol. 6, No. 4, pp. 428-440.

OGC, (2010), OGC Identifiers – the case for http URIs, Open Geospatial Consortium Inc., Ed. Simon Cox, 2010, http://www.opengis.net/doc/WhitePaper/Identifiers/1.0

Omelayenko, B., (2001), Learning of ontologies for the Web: The analysis of existent approaches. Proceedings of the International Workshop on Web Dynamics.

Onsrud, H.J., (2010), Legal Interoperability in Support of Spatially Enabling Society, *Spatially Enabling Society: Research, Emerging Trends and Critical Assessment*. Ed. Rajabifard, Abbas, Joep Crompvoets, Mohsen Kalantari, and Bas Kok. Leuven: Leuven University Press, 163-172. Available at: http://works.bepress.com/harlan_onsrud/1 [accessed 25/05/2012].

Patil, A., Oundhakar, S., Sheth, A., & Verma, K., (2004, May), METEOR-S Web service annotation framework, Pro-ceedings of the World Wide Web Conference (pp. 553-562). New York.

Qin, L. & Atluri, V., (2006), SemDiff: An approach to detecting semantic changes to ontologies,

International Journal on Semantic Web and Information Systems (IJSWIS) 2 (4): 1-32.

Salvadores, M., Correndo, G., Omitola, T., Gibbins, N., Harris, S. & Shadbolt, N.,(2010), 4s-reasoner: RDFS backward chained reasoning support in 4store, In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT).*

Sheth, A., Ramakrishnan, C. & Thomas, C., (2005), Semantics for the semantic web: The implicit, the formal and the powerful, *International Journal on Semantic Web and Information Systems (IJSWIS)* 1 (1): 1-18.

Stojanovic, L., Maedche, A., Stojanovic, N., Studer, R., (2003), Ontology Evolution as Reconfiguration-Design Problem Solving. In Proceedings of the 2nd International Conference on Knowledge Capture, pp. 162-171.

Tai, W., Keeney, J. & O'Sullivan, D., (2011), RESP: A computer aided OWL reasoner selection process, In Proceedings of Fifth IEEE International Conference on Semantic Computing (ICSC).

van der Vlugt, M., (2012) Alignment Study of Spatial Data Supply Chains, Phase 1: Alignment Study of Spatial Data Supply Chains, CRCSI.

van Loenen, B., Janssen, K. and Welle Donker, F., (2012), Quest for a Global Standard for Geo-data Licenses, This chapter is based on and contains parts of Van Loenen, Janssen, B.K., & Welle Donker, F., (2012), Towards true interoperable geographic data: developing a global standard for geo-data licenses, in: K. Janssen and J. Crompvoets, *Geographic data and the law. Defining new challenges,* Leuven: Leuven University Press, forthcoming, http://www.gsdi.org/gsdiconf/gsdi13/papers/45.pdf [accessed 25/05/2012].

Wang, Q. & Yu, X., (2011), Improving reasoning capability for ontology-based geometric product model, In IECON 2011 - 37th Annual Conference on IEEE Industrial Electronics Society.

Waters, R., Beare, M., Walker, R. & Millot, M., (2011), Schema transformations for INSPIRE, *Int. Journ. Spatial Data Infrastructures*, Vol. 6, pp. 1—22.

Wiemann, S., Bernard, L., Wojda, R., Milenov, P., Sagris, V. & Devos, W., (2011), Wen services for spatial data exchange, schema transformation and validation as a prototypical implementation for the LPIS quality assurance, *Int. Journ. Spatial Data Infrastructures*, Vol. 7, pp. 66—87.

Wilkinson, M.D., Vandervalk, B. & McCarthy, L., (2011),The Semantic Automated Discovery and Integration (SADI) Web service Design-Pattern, API and Reference Implementation, *Journal of Biomedical Semantics*, 2:8 doi:10.1186/2041-1480-2-8.

Xioaying, W., Xia, J., West, G., Arnold, L. & Veenendaal, B., (2012), Managing schema evolution in a federated database system, In: Discovery of geospatial resources: methodologies, technologies, and emergent applications, Laura Diaz, Carlos Granell and Joaquin Huerta, editors.

Zadeh, L.A., (2003), From search engines to question-answering systems—the need for new tools, Proceedings of the 1st Atlantic Web Intelligence Conference.

Appendix 1

The main components of the Semantic Web are outlined below.

1.0 The Semantic Web

The Semantic Web is “a web of data that can be processed directly or indirectly by machines” (Berners-Lee, 1999). The Semantic Web is the next step in the evolving World Wide Web (WWW). Berners-Lee, the creator of the WWW, sees the Semantic Web as a web of data in which the data contains enough structure that software can process and derive meaning from it.
The meaning that can be derived by the software from the data is due to the fact that data from diverse locations is linked together. It is this linking between the data in a structured manner that facilitates the software to understand and derive the meaning that is found in the data through the linking. It is in the semantics built into the data that allows for web resources to be automatically harnessed (Sheth et al, 2005). This requires large scale interoperability between data sources that are scattered throughout the world (Koffina et al., 2006). For this to occur there are at least two essential objectives that need to be satisfied: the generation of ontologies (Maedche & Staab, 2001; Omelayenko, 2001) and the automatic annotation of resources (Hammond et al., 2002; Dill et al., 2003; Handschuh et al., 2002; Patil et al., 2004).
The Semantic Web is built upon the Semantic Web technology stack, shown in Figure A.1. It is through the application of different technologies within this stack that the power of the Semantic Web will be realised.

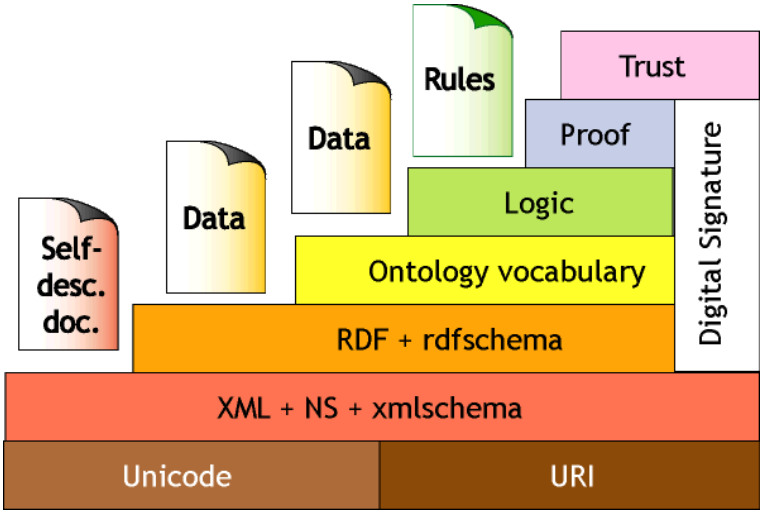

Figure A.1. The Semantic Web Technology Stack (Berners-Lee, 2000)

1.1 Universal Resource Identifier

A Universal Resource Identifier (URI) is an identifier that is used to uniquely identify, reference and interact with a resource (typically data) located on a network somewhere. A http URI is used so that these resources can be looked up and/or referred to. The http URI usually consists of two parts, the http reference (Universal Resource Location: URL) and the resource label (Universal Resource Name: URN) used to uniquely identify the required resource. Currently within the global web community there is a push for the increased use of http URIs in order to resolve resource deployment challenges and make resources persistent. Persistent http URI support is actually built into all modern web browsers and “...

so the adoption of http URIs would allow documents referring to resources with OGC identifiers to be used more effectively in the generic web context" and "...everyone who encounters an OGC document to know where the OGC resolver is" (OGC, 2010).

The importance of persistent http URIs is that developers of web services are guaranteed to be able to find the needed resource, containing OGC identifiers, by using the URI and then to use that resource within their web process.

1.2 Resource Description Framework

A Resource Description Framework (RDF) is "an XML-based language to describe resources." (Daconta et al., 2003). A RDF is a type of meta-data but actually is a standalone document that describes the data within the document. Hence, instead of marking up a document's internals e.g. tags that state whether a word is in bold or not, the RDF captures many of the document's externals e.g. author, date created. "RDF data are a collection of statements, called triples, of the form <s,p,o>, where s is a subject, p is a predicate, and o is an object, and each triple states the relation between the subject and the object" (Chebotko et al., 2008). RDF has matured and evolved into a web resource description modelling language. It is currently used for creating descriptions of resources that are machine readable. Defining the relationship between the subject and object, in the triples set up, enables machines to read and process what resources may be available from different sources on the Web in order to start to build knowledge and understanding about how and where different data can be processed. Importantly RDF triples can be linked together by matching objects to formed linked data to allow the derivation of extra information. Linked data ideas including RDFs and URIs are being explored by the linked data community (http://linkeddata.org).

1.3 Ontologies

The term ontology stems from ancient philosophy and refers to the study of the nature of reality. Parmenides was one of the earliest philosophers to be credited with using the term. The artificial intelligence community was one of the earliest adopters and promoters of ontology use (Gruber, 1995). Ontology is defined as "... an explicit specification of a conceptualisation" (Gruber, 1995) or "... an explicit specification of a shared conceptualization that holds in a particular context." (Cristani & Cuel, 2005). Ontologies are extremely useful when groups can clearly define concepts and the relationships that exist between different types of knowledge. These groups have data, software or services available for use. It is when these clearly defined concepts and relationships are defined in a way that allows them to be implemented within software that ontologies become a very powerful tool (Gruber, 2007; Jepsen, 2009). In the Semantic Web, ontologies can be built using RDF statements to form directed graph representations.

1.4 Reasoning and Rules

With the ontology layer being the highest layer within the Semantic Web Technology Stack to be developed to a certain level of maturity (Bassiliades et al, 2006) the next level that needs to be addressed is the logic or reasoning level. From the ontologies and automated resources generated, reasoning and querying must then be applied in order for the newly generated information to be interpreted and acted upon (Sheth et al., 2005). Wang and Yu (2011) have started to look at reasoners and ontologies within the area of Geometric Product

Models. Their reasoner is built using explicit rules found within a Semantic Rule Web Language (SWRL) and the implicit rules found within ontologies. Some researchers are looking at “a computer aided reasoner selection process designed to perform reasoner selection for different applications and so reduce the effort and communication overhead required to select the most appropriate reasoner” (Tai et al., p. 27, 2011). Also a start has been made in using reasoners in Resource Data Framework Schema (RDFS) in specific software solutions (Salvadores et al., 2010).

That which separates human reasoning from computational reasoning is the ability to reason and make decisions based on uncertainty and missing data (Sheth et al, 2005). There have been attempts at this however: Dubois et al (1994) examined probabilistic and possibilistic reasoning, and fuzzy reasoning. Zadeh (2002) combined fuzzy logic with probabilistic reasoning through a formalism to take advantage of the positives from both areas.

The descriptive language of a Web Ontology Language (OWL) is designed to clearly represent knowledge as well as be used to allow implicit knowledge to be deductively derived (Sheth et al., 2005).
Some of the challenges in this area is that algorithms that have been developed for use within the machine learning community traditionally work on flat data categories rather than hierarchies of data. Hence, within the realm of ontologies and RDF, research is needed into developing algorithms that will function on hierarchical data sets and knowledge representations.

1.5 Controlled Vocabularies

The search for required data sets and compatible web services requires a common set of terms and words to allow the matching and linking to occur. This is especially important when it is expected for software to carry out the search and orchestration of data and services automatically. In the Semantic Web, controlled vocabularies are used to aid in this process. Many controlled vocabularies have been produced for many domains and many are available for use.

Appendix 2

Thanks are given to the following people and organisations who were consulted about the research strategy and projects. They cover academia, the Commonwealth and State organisations and 43pl companies.

Lesley Arnold – Landgate
Arthur Berrill – Tecterra, Canada
Jace Carson – Uni. Canterbury
Drew Clarke – ANZLIC
Simon Cox – CSIRO
Cathy Crooks – DSE, Victoria (ANZSM Demonstrator)
Michael Dixon – PSMA
Paul Farrell – NGIS
Ryan Fraser – CSIRO
Chris Gentle – Mercury
Simon Jellie – e-Spatial
Scott Kennedy – Critchlow
Peter Loughrey – 43pl
Denise McKenzie – DSE, Victoria (ANZSM Demonstrator)
Peter Newton – Swinburne
Dan Paull – PSMA
Phil Poole – NGIS
Abbas Rajabifard – Uni. Melb
Yasser Robi – Fugro
Brad Spencer – SIBA
Bruce Thompson – DSE, Victoria
Maurits van der Vlugt - Mercury
John Weaver – OSP
Andre Zerger – BOM

Chris Bellman – RMIT
Mike Bradford – Landgate, WA
George Curran – CRCSI
Simon Cope – freelance
Ralph Croker – SKM
Jack De Lange – SIBA
Franz Eilert – QSIF
Alan Forghani – MDBA
Tom Gardner – ESRI
Matt Higgins – QLD DNRM
Mark Judd – Geomatic Technology
Manu King – e-Spatial
Elizabeth McDonald – BOM
Richard Murcott – LINZ, NZ
Helen Owens – OSP
Chris Pettit – AURIN
David Purnell – Whelans
Femke Reitsma – Uni. Canterbury
Mary Sue Severn – NZ CRCSI
Sonny Tham – Amristar
Phil Tickle – CRCSI
Mark Wallace – DCS, Qld
Stephan Winter – Uni. Melb