On-the-fly automated federation of spatial data using semantics - Foundation Spatial Data Framework (FSDF)

Dr Zaffar Sadiq Mohamed-Ghouse
Director – NSW Business Development, Research & International Relations
Australia New Zealand Cooperative Research Centre for Spatial Information

ISO TC 211 Workshop
9th December 2015
Sydney
What is the CRCSI?

- A CRC is a ‘Cooperative Research Centre’ setup through the Business Cooperative Research Centres Programme

- Funded through Government and Industry (investing over $180 million*) to conduct user-driven research in spatial information that address issues of national importance.

- Our partners include federal and state government agencies, universities and over 50 companies.

*Cash and In-kind Support
Research Programs
Research partners...

- Curtin University
- University of New England
- University of Canterbury
- Telethon Kids Institute
- QUT
- UNSW
- The University of Melbourne
- Swinburne
- Murray-Darling Basin Authority

Government partners...

- Landgate
- Queensland Government
- State Government of Victoria
- Department of Environment and Primary Industries
- NSW Government
- Land & Property Information
- Australian Government Geoscience Australia
- Ergon Energy
- Government of Western Australia Department of Health
- Department of Agriculture and Food
- Land Information New Zealand
International Collaborations
Our PhD Program

Since 2003:

- 32 active
- 46 completions

(Employed equally by industry, government, universities)
Current Problem

• Spatial data are gathered by various sources
  – Each source store the data as they see fit
  – Leading to heterogeneous, disparate databases
  – Differing in syntax, and schema
  – User has to find, understand, process them and unify the data themselves

• Data have no semantics
  – Making dealing with heterogeneity hard
Current Solution (Aggregator – PSMA)

Data warehousing

– Bulk-downloading of data – Conflation & Duplication Issues
– Major manual work involved
– Heavy in human resources
– Data ownership – to whom does the duplicated data belong to
– Lag in data updates
Possible Solutions

• A federated schema (e.g. INSPIRE)
  – Unlikely – Data is owned by the different authorities

• Data warehousing (PSMA)
  – Current Solution

• Brokering approach [Chosen one]
  – Stand-alone approach not requiring any change from the stakeholders
Our Approach - Automation

[Diagram showing the flow of data and interactions between interfaces and services, with labels such as Interface 1, Interface 2, Web Service, Language Processor, Main Process, BROKER, Natural Language, Data, Query/Data, Ontologies, SPARQL, Serviced Data, Landgate, Western Australia, DELWP, Victoria, Private Portal/API, Private Servers, Public Portal/API, Public Earth.]
Current Requirements

• Web Services (WFS) -
• Web Services Description (OWL-S)
• Federated Schema (FSDF)
• Intermediary System (Broker approach in progress – OWL2)
Web Features Services

- Current use case:
  - LGATE-069, Landgate
  - Vmadmin.state_council_2005, DELWP
- Currently matching the “name” and “REGION_LABEL” eg: Electoral Boundaries Western Australia
  - Victoria

```xml
<xsd:element name="ogc_fid" minOccur="0" nillable="true" type="xsd:int"/>
<xsd:element name="land_id_number" minOccur="0" nillable="true" type="xsd:int"/>
<xsd:element name="usage_code" minOccur="0" nillable="true" type="xsd:int"/>
<xsd:element name="postcode" minOccur="0" nillable="true" type="xsd:string"></xsd:element>
</xsd:simpleType>
</xsd:restriction>
</xsd:simpleType>
</xsd:element>
<xsd:element name="name" minOccur="0" nillable="true" type="xsd:string"></xsd:element>
</xsd:restriction>
</xsd:simpleType>
</xsd:element>
<xsd:element name="row_id" minOccur="0" nillable="true" type="xsd:int"/>
<xsd:element name="the_geom" minOccur="0" nillable="true" type="gml:MultiPolygonPropertyType"/>
```
WFS Descriptions

• Using OWL-S (W3C Ontology)
  – Designed to be able to describe any web services
Federated Schema

• Using an ontology adaptation of the FSDF
  – Required as a role model
  – All other spatial data from the States should map into the FSDF either directly or as an abstraction
Ontologies Used & How they link [Simplified]
The Automated Process

- WFS’ URIs are known
- User query is processed into SPARQL and ontology queried
- Simple SPARQL filtering queries can be matched to WFS filters

If request not found:
- GetCapabilities is called and the FeatureTypes stored in the ontology
- DescribeFeatureType of specific Feature is called and metadata about that feature is stored in ontology
  - Modification from XML Schema to OWL is done automatically
- GetFeature is called, and the data is returned to the user (without duplication)
- Only duplicated data are the schemas onto ontological equivalent
Error Handling

- Either HTML error output
  - Error Code indicates direct cause of error
  - Usually on the server side
  - Not much can be done – Simply notify the user about the error (e.g. Server down)

- Or JSON error output
  - Can be parsed to find the root cause (e.g. Feature not found)
  - Solved by matching the ontology to GetCapabilities of WFS, and looking for discrepancies
Web Demo – Landgate & DELWP

FSDF (UML)
Summary

• To provide a federated, and unified view of the various data sources
• To allow on-the-fly data retrieval – hence undelayed updates
• To provide automation of the data sources retrieval process – The user should not need to know about them
• To provide an answer to database scalability issues
What does it bring?

• Automated ways to semantically consolidate differing datasets
  – Less human resources / efforts
  – Less cost
• On-the-fly techniques would allow access to most recent data (instead of outdated ones, as data do not need to be modified, just their schemas adjusted)
• Data stays at their original source, removing duplication issues
• Semantics would promote the jump to the next cyber generation (Web 3.0)
• Ontology used can be continuously built and improved
• Caters for both current and future data providers (stakeholders need not change their workflow)
Future Potential

- Expansion to other domains
- Including a self-learning algorithm to expand the ontological capabilities, and features
- Expansion to Local Government Agencies
- Automatic integration of other data sources
- Expansion to upper ontologies
Thank You !!!

zsadiq@crcsi.com.au

Acknowledgement:
This is the work of our PhD Researcher Mr Jeremy Fa