White Paper

Mar 2007

GeoSpatial Knowledge Discovery

- Brad Spencer, CubeWerx Australia

Background

As we collect mountains of geospatial data from various sources such as government mapping programs, from private aerial photography and satellite acquisition programs we migrate into an era of deployment that demands efficient storage and access to these resources. A data dissemination environment that has its own peculiar issues that the GIS data production community typically have limited experience with. Notwithstanding data quality and currency, we recognise that the cost of acquiring spatial data is not insignificant and therefore we must squeeze as many benefits as possible from these assets and that implies the reuse and or the sharing of geospatial resources via the Internet.

In a web search engine analogy; in the early days of the Internet there were thousands of web sites out there with many more being created every hour but how did we know where to find them? Well along came the likes of AltaVista, Yahoo, Google and others. They had new technologies that could harvest snippets of information from these many websites, stuff these into a huge registry or database and allow people to search these by keywords, inspect the summary information returned and then hyperlink to any of the resources they had found and that considered were of any interest to the user. This process has helped us navigate around the myriad of available web services out there. So important has this become that there is a whole strategy around Search Engine Optimisation and major suppliers of registry services attract not-insignificant advertising revenues via this registry mechanism to make it big business.

But what of geospatial resources? With the proliferation of web-enabled geospatial resources its not hard to see that the challenge of finding the most appropriate geospatial data for any given requirement could also become a difficult and frustrating task – there is nothing worse than having to wade through irrelevant data. Users not only have to be selective regarding the linage or integrity of the data but also its currency. So the concept of geospatial metadata has evolved to help us in this quest to discover appropriate resources for the job at hand.

The secret to all this is of course metadata, or data that describes the geospatial resources being published and how to access them. We have of course standardised the content of these metadata entries (records) to fit into structured matadata schemas or definitions. We have created ISO 19115, ANZLIC, FGDC, etc and even a standard OGC GetCapabilities document is also being used as metadata content for discovering OGC services. These standards (ontologies) are important in that they provide a level of consistency across catalogued resources. Predictability is always a good thing when we want automatic or semi-automatic access to any web resource. But once created they need to reside in a Registry where they can be searched by users and applications alike.

Registries

Registries provide a database of metadata structured via these ontologies that provides the infrastructure of geospatial discovery. Ontologies can also be applied in a hierarchical way that facilitates geospatial knowledge discovery to assist the users or computer programs to locate and process the most suitable resource for any given application. However, its very important that those who build metadata select the appropriate industry-specific standards so that inferences and automatic geo-processing can be reliably performed against these published resources. This is of course based on the assumption that the meanings of metadata within these industries are consistent across organisations.

For example, each industry may have its own jargon or vocabulary that forms the basis of the metadata they create for the data resources they catalogue and which are ultimately harvested into an SDI Registry. Yet a truly interoperable Registry should permit a user to discover the resources they need via their own vocabularies which may be yet again different. This semantic interoperability is essential when you consider that the basic premise of SDIs are that they are a collection of distributed resources. The custodians building the metadata that describe these resources are typically working independently of each other and focused in different industries. This means that there will be differences in metadata content meaning. Take this to a global scale and the issue deepens. So a Registry must be able to store not only different industry specific metadata models but must be extensible to capture and exploit these different industry specific vocabularies.

Registries should be able to link or associate data resources with desired processes such that computer programs can automatically bind the process and the data resources together. A data resource may be accessible by different geo processes. Both resources (data and process) can be registered in a Registry and the binding of these together will be dependent on the actual query. So the data resource will need to have some link to the relevant geo processes. For example, assuming that cadastral boundaries are stored as a vector based layer and given that this published resource is associated with a WMS process it's a simple encoding exercise to bind the service type with the data resource to get the web map programmatically and on the fly. This same data resource may also be associated with a WFS service that is required to return some geometry to satisfy a query for the vector data. Add to this other high level services such as thematic mapping, cartographic transformations, geocoding, terrain analysis, routing, etc and its not difficult to see that by providing the capability within the Registry to store any number of associations between registered resources provides a powerful data mining approach to query requests from either people and or computers.

Service Chaining

The combination of predictable metadata and associations between registered resources allows applications such as geospatial portals to not only provide a discovery capability but to also (with or without user interaction) traverse these associations and to construct a work flow of logical and predictable processing steps to deliver the end user solution. This service

chaining can be completely automatic and hidden from the user, the user may be prompted to make interim decisions in the chain of services or the user may have all the knowledge to specify exactly what they need. Whilst the associations provide the data mining features it's the ontologies that provide the various levels of knowledge required to infer which service is the most likely candidate for any given service chain step. Geospatial knowledge discovery is based on this service chaining paradigm and can also be used to provide e-business interfaces where access fees are applicable.

Solutions

There is a lot of R&D and pilot work going on around the world establishing geospatial knowledge discovery solutions. Where interoperability is important, these projects adopt the OGC Service Framework (OSF) publish-find-bind model (see Fig 1 below) which outlines an operational architecture for SDIs. This model includes the Registry service that contains the catalogued metadata harvested from the service providers via some formal registration process. The Requestor provides the application that queries the registry, builds and executes the service chaining.

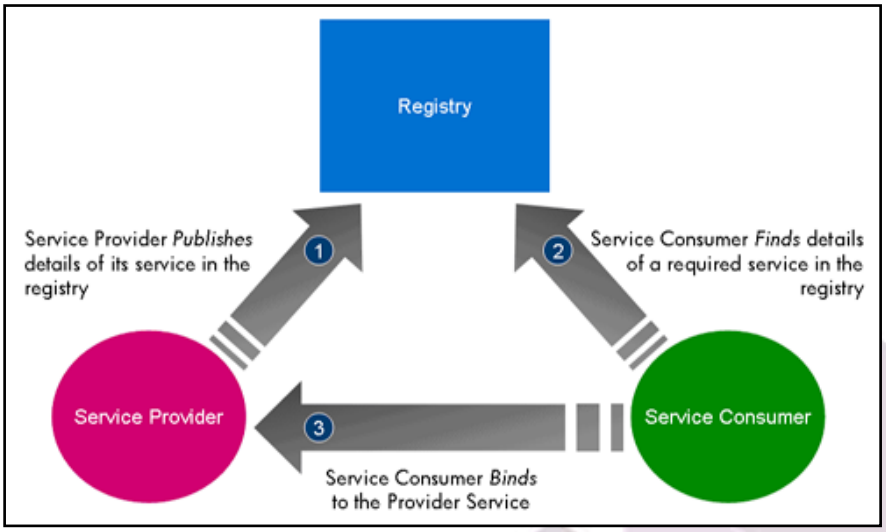

Fig 1: OGC Service Framework

OGC has also approved a standard API specification that developers can follow to develop application interfaces between the Requestor and the Registry called Catalogue Services (CSW). This enables an interoperable interface between heterogeneous and distributed service providers and users. This specification also supports the ebRIM data model within the Regsitry. The ebRIM database model is in fact a meta-metadata model in that it can ingest, store and search through many different ontologies and supports associations between different metadata models – a Registry need not have just one standard metadata model. These two features are fundamentally important to support service chaining.

There are also very interesting developments in portal frameworks that facilitate the building of applications that can chain services together via the application of standards-based portlets. A portlet is a functional component that is written as a web service and can be used and re-used within a portal application framework of the enterprise. Interoperable versions can also be registered, discovered and accessed remotely.

Conclusion

Setting up a robust reliable and flexible SDI is not a simple process due to the complexities associated with the discovery and merging of distributed geospatial resources. Not only is it important to provide interoperability across platforms and systems but there is also a major requirement to support semantic interoperability so that we and computer processes can confidently find and bind the appropriate published resources quickly and reliably. Fortunately, these issues have been and continue to be addressed in many different projects around the world. Also, standards are maturing to the level that we can now confidently start to build fully interoperable SDIs.

Of particular interest is the advancements made in the geospatial knowledge discovery area with the development of basic and domain specific ontologies that standardise the meaning of metadata. Add to that the flexibility in the ebRIM meta-metadata based Registries, portal development tools and its clear that interoperable service chaining is now deliverable. These and other advancements means that the publish-find-bind vision outlined by OGC® is now a reality and the only thing that is required is leadership to seize the day from those responsible for establishing regional, national and global SDIs.

CubeWerx®

15 Rue Gamelin, Suite 506
Gatineau, Quebec J8Y 6N5
Canada

North American sales
sales@cubewerx.com
Telephone: (819) 771-8303
Facsimile: (819) 771-8388

Australia and New Zealand sales
CubeWerx Australia Pty Ltd
Brad Spencer, General Manager
ABN: 37 115 163 285
Mob: +61 (0)404 841 131
Tel/Fax: +61 (0)2 9481 7024
brad.spencer@cubewerx.com.au
www.cubewerx.com.au