Community Schemas

Making sense of disparate datasets

CAMERON SHORTER and STEFAN HANSEN

Non-proprietary or open source information processing is very much the flavour of the month. Paradoxically, open source computing is also having a profound effect on communities of interest, where highly specialised people, systems and data interact.

A couple of recent events point to what this might mean. Funding from the National Collaborative Research Infrastructure Strategy, the CSIRO and the Department of Primary Industries in Victoria has been used to add support for a community schema to an open source GIS called Geoserver. The same group is investigating Deegree, another open source project.

A number of other developments are also on the go. The CSIRO is managing an open source product called FullMoon, which supports the transformation of data models. The Australia Water Data Infrastructure Project has developed a tool called DuckHawk, which also tests data models.

Several themes stand out: the specialised nature of these projects; the number of participants; and the open source nature of these tools. There is a good reason for this. On any given project of reasonable complexity, the most valuable dataset usually belongs to an agency over which you have minimal influence. When many agencies from different departments, states and countries are working to varying timelines and budgets, it can get really messy. One way forward is to sponsor organisations to develop open source tools. This will greatly reduce the financial barriers to getting data online.

> One way forward is to sponsor organisations to develop open source tools...

But concentrating on these applications should not blind one to the real challenge, which is better re-use of data. That means the creation of generally accepted data models, in which different terms and attributes mean the same thing.

Groups like the hydrology community, which is backing the Australian Water Data Infrastructure, are solving these data integration issues by defining a community schema for their domain. They are then better placed to ensure that all the agencies in the community publish data using that same schema.

Community schemas are used to describe a set of semantics for a domain. The first was Geographic Markup Language (GML), which treated the whole of geography as a special domain.

The new trend is to go far beyond that. It allows communities of interest to define schemas appropriate for their data. These can then be referenced to ensure consistent structure and taxonomy between related datasets.

For example, a hydrology schema may define a class called Water Use, with acceptable terms defined as *irrigation*, *domestic* and *industrial*. If a dataset that describes Water Use as, say, storm water is published against this schema, the data will not validate and the user will have to correct the mistake.

A standardised community schema means that large amounts of quality data become more accessible. This in turn leads to the development of new applications and data analysis projects. It also delivers clear data definitions, which reduce data misinterpretation.

Defining a community schema for a domain is no trivial task. It requires

A river flows somewhere in the Murray Darling Basin. The water management community will be one of the spurs to the creation of special schema in Australia.

participating parties to create data architecture, vocabularies and an interchange protocol. Luckily, the first community schema projects have left a trail of re-usable building blocks and processes that can be used by future efforts.

For instance, specifications for Observation and Measurement (O&M) were developed as part of the Sensor Web Enablement (SWE) specification. They have since been used as a basis for Geoscience Markup Language (GeoSciML), Water Markup Language (WaterML) and others. Other building blocks include GML, SensorML, CityGML and the ANZLIC profile of ISO19115 for metadata.

The other critical component in the development of a schema is involvement of the user community through buy-in, governance and testing. This is often an international effort. GeoSciML has participants from BGS (UK); BRGM (France); CSIRO; GA and GSV (Australia); GSC (Canada); APAT (Italy); JGS (Japan); SGU (Sweden); USGS (USA) and the OGC (International).

Communities need to organise a governance structure to resolve inevitable disagreements over technical details. The GeoSciML community started in 2003 and is now into its third schema. It has organised working groups for the development of information and computational models, the definition of vocabulary and testing. It also has an outreach working group that promotes the schema to the user community.

Data is collected by numerous agencies, following different collection guidelines…

Most agencies first encounter community schemas when they are asked to deploy datasets that use them. Spatial data is collected by numerous agencies for various purposes, following different collection guidelines. Storage models tend to reflect the original data use and are rarely designed for data exchange. When data is published through web services, the schema usually reflects the storage model. This works for the original application but is an integration nightmare when trying to share data between agencies.

Changing the storage format usually isn't desirable if it breaks legacy applications or introduces sub-optimal performance. Hence it is necessary to differentiate between the storage and exchange formats. Storage format can be defined by the custodian who generates and maintains the data. The challenge is to map the storage model to a community schema. Again, prior projects have left a suite of tools that can help solve many of the problems.

Australia, like the rest of the world, has a huge variety of datasets, all fulfilling their own purpose. Data integration is no easy task, but we have the knowledge, tools and processes to integrate disparate datasets.

This will enable more powerful analysis and create new business opportunities for participating parties.

Cameron Shorter and Stefan Hansen are both with Lisasoft. Shorter is a geospatial systems architect and Hansen is a software developer. Both were involved in the recent development of DuckHawk and the testing and validation of AWDIP's community schema-enabled WFS Spatial Data Infrastructure. Lisasoft builds open geospatial solutions.