

Schema Translation and Semantics in Data Interoperability

1 Abstract

Part of the CIPI 1.2 (Critical Infrastructure Protection Initiative phase 1 stage 2) initiative run by the Open Geospatial Consortium (OGC) in 2003 developed prototype implementations of Translating Web Feature Servers (WFS-X). These feature servers served geographical data in a single common GML application schema with the source data being provided from a number of sources each with a different GML application schema. The implementations were therefore required to translate the data from the various source schemas into the common target schema.

The initiative showed that schema translation is a feasible technical process but the process was hampered by a poor definition of semantics. Data was served from a number of sources to a common schema but use of the common schema was not consistent in all translations. Data structures in the common schema were sometimes populated with different information depending on the source of the data. The common schema was not associated with a definition of required content or quality thus allowing one translation to provide data that might be regarded as incomplete by some applications. (One data source contained only major roads whilst another contained minor streets and un-surfaced tracks).

Many GI interface standards emphasise system interoperability i.e. enabling different software packages to be used interchangeably. However, a major requirement for cross-border working is data interoperability i.e. allowing datasets from different geographical regions to be used interchangeably. This requires data to conform not only to a common structure, but to common semantics, levels of completeness and quality.

There are currently a wide range of initiatives under way to define standard schemas for different sectors. (For example, eXploration and Mining Markup Language (XMML), City GML for 3D city models, TransXML for transport, ALKIS and ATKIS for topography and cadastre.) The important factor in these developments is that these schemas are associated with well known or well defined semantics. Because of their semantic content the creation of these schemas represents significant progress towards creating interoperable data sets.

Provision of data in this new generation of schemas requires schema translation. No organisation is currently capturing data which conforms to these schemas; nor should they. Standard schemas represent a common model for exchanging data. They do not contain elements required to support the business processes of the organisations which capture the data. Furthermore, many data capturing organisations will need to publish data to more than one standard schema, and so their data storage schemas will need to be a superset of the schemas they need to export.

Schema translation tools such as WFS-X are now available as commercial products. The main area for work towards data interoperability is therefore not on delivery technology but on standard schemas and the harmonisation of data to be able to fulfil those schemas.

The formal definition of semantics is an active area of research both for the internet in general and for geospatial information in particular. Formal semantics hold the promise of on-the-fly discovery and translation of data. However, this is a long term goal, and in the short term the use of formal semantics is hampered by the mathematical nature of the formal languages which make them inaccessible to all but expert users and by the limited number of tools available for manipulating and interpreting semantics. At present no tools exist that make use of formal semantics in data translation.

2 Introduction

This paper examines some of the issues currently affecting interoperability for geospatial information. The translation of data between different schemas is considered and the related issues of specifying semantics for those schemas.

3 Schema Translation Experience

Part of the CIPI 1.2 (Critical Infrastructure Protection Initiative phase 1 stage 2) initiative run by the Open Geospatial Consortium (OGC) in 2003 developed prototype implementations of Translating Web Feature Servers (WFS-X). The aim of the exercise was to demonstrate the exchange of information from heterogeneous data stores.

The data stores each had a different data model (schema), were implemented using different database technologies and the data used for the exercise was supplied in a variety of formats and schema. The aim was to eliminate the differences between the different source data in the data served to the client and to serve data in a single schema from different servers. The client application should therefore have been able to transparently switch between servers and data from different sources without any change of schema.

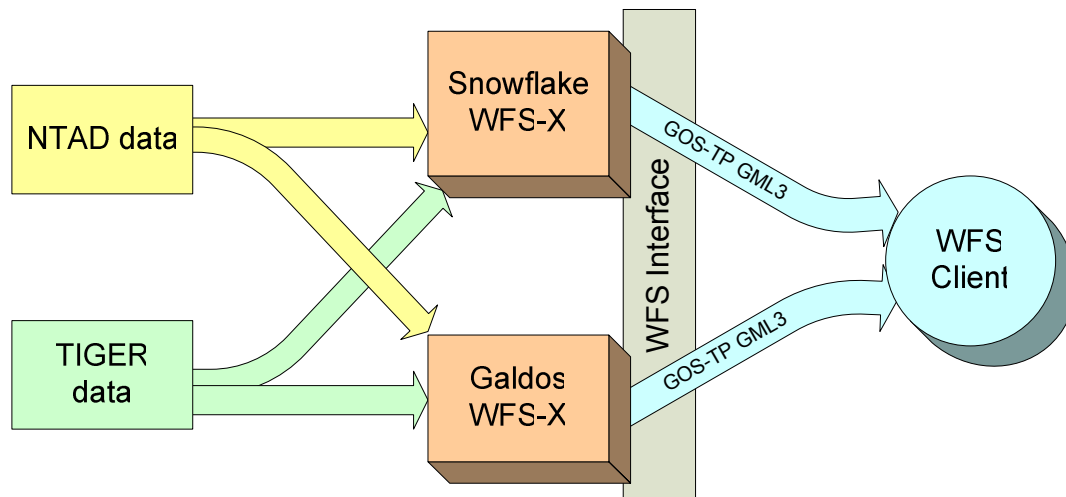


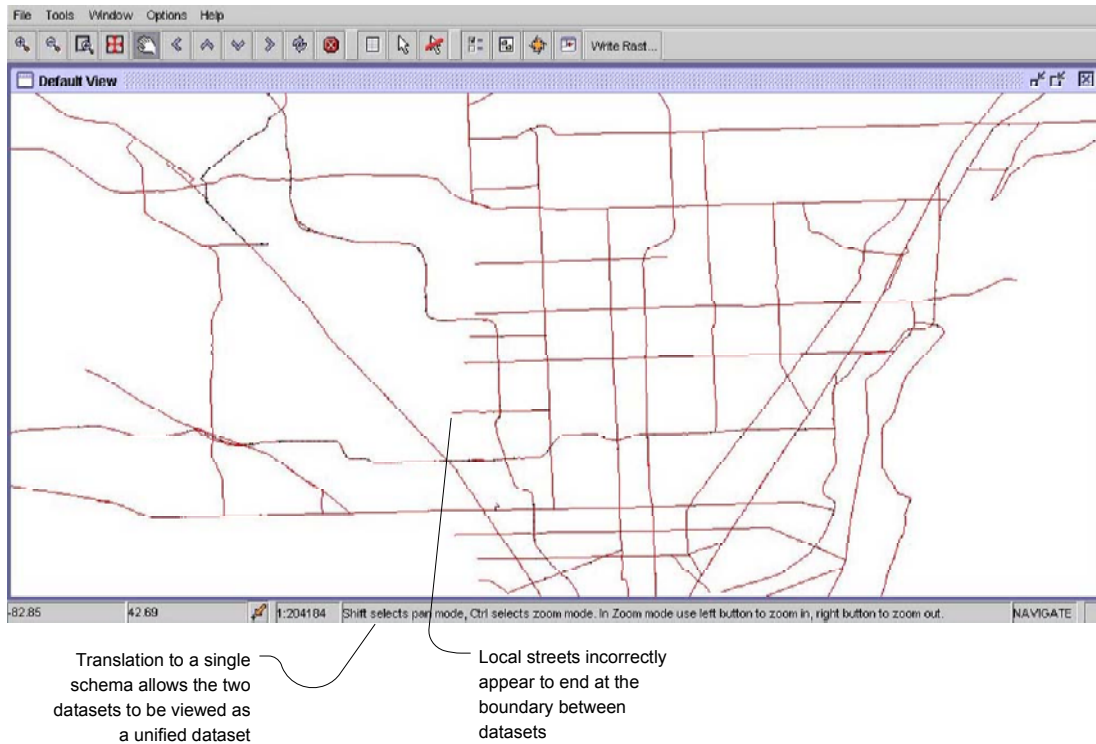
Figure 1 - Components of the CIPI 1.2 test bed

Communication between the data stores and client applications was carried out through an interface conforming to the OpenGIS Web Feature Server (WFS) standard. The transport format for the data was Geography Markup Language (GML) 3.0. Road data was used for the exercise and data was sourced from the US Census TIGER database and from the US NTAD (National Transportation Atlas Database). The common schema used to communicate between the servers and client was the GOS-TP (Geospatial One-Stop – Transport Pilot) - Roads schema developed during a previous OpenGIS test bed.

3.1 Lessons Learned

The initiative showed that schema translation is a feasible technical process but the process was hampered by a poor definition of semantics. Data was served from both the NTAD and TIGER sources to the GOS-TP schema but use of the common schema was not consistent in all translations. Data structures in the common schema were sometimes populated with different information depending on the source of the data. Different interpretations by the different development teams also lead to differences in the way that data was translated from the TIGER and NTAD sources into the GOS-TP schema. Consequently the same data was occasionally placed in different parts of the GOS-TP structure by the different implementations.

Whilst the two source data sets superficially appeared to contain the same kind of information (a road network) in detail there were significant differences in information content. The NTAD data contained only major roads whilst the TIGER data contained all streets. This gave the appearance of the local street network simply ending at the boundary of the two data sets. There were also many differences in the attribution of features in the different data sets.



The differences were not readily apparent until after the datasets had been integrated because the GOS-TP schema was not associated with a definition of required content or quality. This allowed both TIGER and NTAD datasets to be translated into valid GOS-TP data despite the discrepancy in content.

4 Existing Standards for Interoperability

Many GI interface standards emphasise system interoperability i.e. enabling different software packages to be used interchangeably. One of the most commonly used interface standards is ISO 19128 – Web Map Service (WMS). This protocol allows a client application to request a map from a server and the server to return a map image. The semantics of the request are well defined, with the interpretation of the geographic area requested, coordinate system and projections etc. being well specified. However, the service returns an image which is essentially only suitable for human interpretation as the returned data structure (a raster image) contains no features. The semantic definition returned is limited to a map legend.

The WFS interface allows client applications to request feature data from a server. The feature data is returned as a stream of GML data. Like the WMS interface the semantics of the request are well specified. For the returned data the GML specification defines a level of semantics for the basic elements of the returned data such as the general idea of a geographical feature, geometry, coordinate system etc. GML must be extended to provide an application schema for each GML dataset so GML provides a framework and components for building the application schema but the application schema must define the actual feature types and attributes used in instance documents e.g. roads, buildings, rivers. Conformance to the GML standard is achieved if the data structure of the application schema extends the GML schema within the rules defined in the GML specification. Conformance can therefore be achieved without stating the semantics associated with the feature types of the application schema. This is sufficient for system interoperability since it provides data structures which can be accessed by any system.

The application schemas used in WFS are typically generated from software based on the data model of the underlying data repository. The semantics of the underlying models usually reside in natural language documents or CASE tool models. The semantics are not available to the WFS software and so this automated process is limited to creating data structure and does not define semantics.

By enabling systems to exchange data these standards clearly make a major contribution to interoperability but the success they have achieved in creating system interoperability has exposed the issue of data interoperability. Where systems are able to work across borders the differences in the data content on each side of the border becomes apparent. This raises the need for data to conform not only to a common structure, but to common semantics, levels of completeness and quality.

5 Current Initiatives

There are currently a wide range of initiatives under way to define standard schemas for different sectors. For example, eXploration and Mining Markup Language (XMML), City GML for 3D city models, TransXML for transport, ALKIS and ATKIS for topography and cadastre. These initiatives are focussing on defining common data models for particular communities of users rather than focussing on systems. They are seeking to provide an encoding for concepts which are well understood within those communities and consequently are either implicitly or explicitly considering the semantics associated with the data structure. Because of their semantic content the creation of these schemas represents significant progress towards creating interoperable data sets.

These schema are designed for data exchange, not for data storage. Many data capturing organisations need to exchange data amongst more than one community of users. For example, a mapping agency dealing with topographic data has information which is valuable for navigation, asset management, cadastre, planning, agriculture and many other applications. Additionally requirements for data may vary between local, regional, national and international levels. Each of these communities of users will require a subset of the information available and each community would be best served if the data were presented in a form which is easily interpreted by that community.

Data providers must store data in a form which supports their own business processes as well as meeting the requirements of many data users. Data can then be provided in an appropriate form to each data user community through a process of translation from the storage schema to the various exchange schemas.

6 Future Work

The issue of system interoperability has been successfully addressed by a range of existing standards, albeit that these standards and technologies are still maturing. With schema translation tools such as WFS-X are now becoming available as commercial products the main area for work is therefore data interoperability i.e. the creation of standard schemas and the harmonisation of data to be able to fulfil those schemas in a consistent manner.

The informal, natural language definition of semantics is currently the norm. The formal definition of semantics is an active area of research both for the internet in general and for geospatial information in particular. Formal semantics hold the promise of on-the-fly discovery and translation of data through machine interpretation of data. However, this is a long term goal, and in the short term the use of formal semantics is hampered by the mathematical nature of the formal languages which make them inaccessible to all but expert users and by the limited number of tools available for manipulating and interpreting semantics. At present no tools exist that make use of formal semantics in data translation.