## EuroSDR/EuroGeographics Workshop

# 'Ontologies and Schema Translation Services'

## Marne-la-Vallée, France, 15-16 April 2004.

34 delegates from Belgium, Denmark, Finland, France, Germany, Great Britain, Ireland, Italy, Luxembourg, Netherlands, Spain, and USA attended a Workshop on 'Ontologies and Schema Translation on 15-16 April 2004. The Workshop was organised by EuroSDR Commission 5 (Delivery and Integration of Data and Services) and EuroGeographics, with support from the Joint Research Centre of the EC. The objectives and agenda of the Workshop were informed in part by the outcomes of the **Eurospec Workshops** held in 2003. It was hosted by IGN France at "l'École Nationale de Sciences Géographiques" (ENSG), at Marne-la-Vallée, near Paris, France.
The Workshop was opened by Claude Luzet, who welcomed the delegates to the EuroGeographics Headquarters at ENSG and described the role of the organisation (**EuroGeographics**). Peter Woodsford then briefly outlined the history and structure of EuroSDR and its current mission statement and research agenda (**EuroSDR**). He gave a definition of 'Ontology' as:

> *'an explicit formal specification of how to represent the objects, concepts and other entities that are assumed to exist in some area of interest and the relationships that hold among them'*

then summarised the Objectives of the Workshop and the issues to be addressed in the Breakout Sessions before handing over to Claude Luzet to chair the opening session.

### *Session 1: Theory and Research*

Heiner Stuckenschmidt, from the Vrije Universiteit Amsterdam, Netherlands, set the wider scene with a presentation on **'Towards a Semantic Web - From Vision to Technology'**. The impressive success of the WWW is creating a problem of too much information. Navigation is only supported via hard-wired links. The real power will come with the next generation - machine accessible semantics. Rich metadata is the key and the bottom line is 'Ontology Alignment'. A framework has to be developed to combine terminological and spatial reasoning ('concept@location'). Heiner's pointers to XML, RDF, RDF Schema and OWL as key technologies from W3C for standardising syntax, and his thinking on the roles of intelligent thumbnails and spatial reasoning were picked up in subsequent presentations.

Werner Kuhn, from the University of Munster, Germany used his presentation on **'Approaching Semantic Problems in a Web Service World'** to ask some fundamental questions about what business(es) are national mapping agencies in. The key interfaces are now data-to-service, service-to-service and service-to-user. Ontologies describe contents for customers; they are not inventories or newspeak for schemas. Users of geographic information should be able to refer thematic data to semantic reference systems, just as they refer geometric data to spatial reference systems. The architecture to achieve this would need three levels, with two threads across them - design and test (using functional language) and expose and visualise (using markup languages and UML).

Joseph Greenwood in a presentation on **'Exploring Knowledge Representation for Geographic Information Exchange - an Ordnance Survey Perspective',** described the Iridium project (Paper-Greenwood) This set out to investigate how to manage and use complementary definitions of features across organisations through the disaggregation of the feature into identity and exchangeable information packages, with examples such as water quality from the hydrological domain. Relationships are established between ontologies (expressed in OWL) and used with a reasoner (Cerebra) to support cross-domain queries. They concluded that OWL has some limitations (e.g. lack of self-referencing) but lets us formalise and validate the semantics of the domain and data. The relationships established between the ontologies and the data supports query expressed in terms of multiple domain or users concepts.

OWL also featured prominently in the presentation on **'Creating translation rules for integration of geo-data sets using OWL (Ontology Web Language)'** by Marian de Vries, TU Delft, Netherlands. She described experience

arising from the IMKICH project on cultural heritage, which sets out to support domain experts such as archaeologists, civil servants and GIS professionals. UML models of the different datasets had to be reverse-engineered and the used to create OWL ontologies which were then matched. OWL was good for matching and for thesaurus construction, less so for conceptual modelling (lack of constraints) and conversion scripts. Syntactic interoperability is within reach but much research is needed to achieve semantic interoperability. The IMKICH project was finished using SQL.

Femke Reitsma, who had just flown in from the University of Maryland, USA, described work from the Mindswap project there in her presentation on **'GeoReferencing the Semantic Web ontology based markup of geographically referenced information'.** Ontology development is based on geographic features and spatial relations (topological, direction, distance, mereological). XSLT is used to translate GML data to OWL. An experimental plug-in provides validation of topological and direction relations as a web service. She concluded that using ontologies for marking up geographically referenced information on the Semantic Web has advantages for knowledge discovery, but that there are limitations in OWL(paper-Reitsma).

## *Breakout Session 1*

There were two groups, facilitated by Peter Woodsford (rapporteur Monika Sester) and Lars Bernard (rapporteur Hardo Mueller). The full outcomes are at (**Sester**) and (**Mueller**, **Graphic**) - key points can be summarised as follows:

- *What are the problems?*

We must avoid being too data-centric. Balance with services viewpoint. It is difficult to communicate with other (non-GI) specialists, but we have to learn to do so. Much knowledge is hidden or implicit. Understanding of ontologies (and of the differing roles they can play, e.g. in semantic web and in databases) is limited. Standards have not yet developed to the point of trust. Modelled knowledge needs to be made visible and hence provable. (How will the metadata for the third generation Web be generated? How do we reason about data and about services?

- *What is the state-of-the-art*
  - *In the short-term (2-3 years)*

  Likely progress (hype to reality) will be similar to that of HTML. Trustable standards will emerge (maybe). Semi-automatic production of ontologies. Watch out for Google - they may have it solved.
  - *In the longer term (~8 years)*

  Agreed semantic reference systems in place. Significant progress on reasoning. Semantic Web.
- *What tools are available/emerging? How do they relate to established data modelling tools, e.g. XML and XMLschema tools*

RDF, RDF-Schema. RuleML is emerging. Georeference online. Unclear as to if/how limitations of OWL can be removed. Limitations of first order logic. Consistency checkers are available and are important. Reasoning engines presently come to differing solutions. All practitioners present and using OWL/DLs concluded that the language was too mathematical to be easily understood and it is therefore difficult to express domains in an unambiguous manner (to humans)

The day concluded with a very enjoyable dinner at 'La Ferme du Buisson'.

## *Session 2: Requirements and Current Projects*

This session, chaired by Tapani Sarjakoski, began with a presentation by Claude Luzet of EuroGeographics on **'A Survey of the Requirements of EuroRoadS, EuroSpec and other pan-European projects'**. He outlined the current timetable, priorities and revised scope for INSPIRE (European Spatial Data Infrastructure). The EuroSpec programme sets out to deliver interoperable reference data is support of ESDI. Active programmes address small scales and boundaries data, roads, metadata and coordinate systems. Proposals exist for addressing elevation and hydrological data and geographical names. Web mapping standards allows mixing of mapping (acceptable for some applications) but does not achieve semantic interoperability. Investments in current data holdings have to be utilised, so the way ahead involves agreement on content and catalogues, a new generation of data models (starting from expression in UML) and probably customisation by scale range. Multilingual aspects need to be addressed early in this iterative process.

Lars Bernard of the EC Joint Research Centre focused on implementation issues in his presentation on **'Requirements on (Semantic) Interoperability in the ESDI'**. To make data interoperate requires functionality (Services), contracts (Standards) and service cooperation (Chaining). ESDI Technical Guidelines are needed to resolve ambiguities in current standards, to address the INSPIRE requirements and to be dynamic across increasing levels of interoperability. The first step is publishing and finding digital GI resources – semantic issues include multi-lingual support, the sufficiency of current metadata standards and how to describe the link between catalogued GI-data and GI-services. The next step – visual exploring and simple query – requires harmonised portrayal. Can EuroSpec be the semantic reference system needed to support the third stage - complex query?

Alia I. Abdelmoty from Cardiff University, Wales described the EU-funded collaborative SPIRIT project for spatially-aware information retrieval on the Internet in her presentation on **'Maintaining Ontologies for Geographical Information Retrieval on the Web'**. A geo-ontology encodes knowledge of the terminology and structure of geographic space and plays key roles in query interpretation, formulation, metadata extraction and relevance ranking. Design considerations include geo-ontology components, footprint granularity and integrity maintenance. Precomputation vs online computation is a key efficiency factor. Future research includes the representation of imprecise places and data integration and distributed architectures for global geo-ontologies.

The presentation on **'Describing and Customising the Content of Geographical Data sets: Approaches based on Data Schemas'** by Sandrine Balley covered two related projects of the IGN France COGIT Laboratory, both arising from the fact that IGN produces different datasets for different uses. The goal of the project on ontology-based formalisation of data specifications (Nils Gesbert) is to formalise descriptions for easier consultation by users. The goal of Sandrine's project - an interface to select and restructure available datasets - is to enable the expression of the user requirements and to perform customisations. Research has been based on the hydrographical theme of IGN databases and has demonstrated that the ontology of geographic entities can assist the process of schema customisation.

In the first of two presentations on the EU-funded IST GiMoDig project, Andreas Illert of BKG, Germany described **'The GiMoDig Global Schema'**. The project prototypes delivery of geospatial data to the mobile user, using real-time data-integration and generalization, based on a seamless data service providing access to topographic geo-databases maintained by NMA's (Germany, Denmark, Sweden, Finland). The global schema is central to this. Conceptually it is based on the DIGEST FACC and the ERM. An inventory of contributing data holdings, revealed differences in classification systems, feature definitions, attribution, levels of collection detail and generalization. On-the-fly harmonisation is possible for 1:1 and n:1 relationships, 0:1 requires extra data and n:m requires interactive editing to an adopted map style.

## *Breakout Session 2*

Glen Hart presented two use cases to help crystallise thinking:

- Crime and Deprivation. Many organisations involved (police, social services, probation services, office of statistics, etc), with different vocabularies and classifications and often non-matching collection areas
- Water quality (Water Framework Directive). Similar diversity of involvement, but the spatial problems arise because of non-matching understandings and partitions of river networks.

There were two groups facilitated by Glen Hart (rapporteur Colin Bray) and Werner Kuhn (rapporteur Andreas Illert). The full outcomes are at (**Bray**) and (**Illert**) - key points can be summarised as follows:

- *Where do the semantics reside? How are they to be abstracted?*

  Application semantics reside with domain experts and documentation. Each discipline has its own segmentation which may not map on to that of the NMA's framework data, although the topographic view is the good common denominator as it is reasonably unspecialised. Data content is often available in documentation and schema. Structure and implicit information is harder to acquire. In the short term we have to derive semantics from specifications and metadata. It is really a modelling problem, but we need extensible models
- *Do we need Ontologies for Schema Translation?*

  Dependent on the data richness and intended uses. Ontologies may have other uses, e.g. for generalisation. Schema translation serves not only to combine views from different schema but also to serve up different views

for different users. Two approaches noted (IGN-F starting from conceptual model and formal specifications, OSGB starting from the domain ontology and working back to the data).

- *What are the guidelines for 'target schema' that are well-posed for translation purposes?*

Target schema has to meet the needs of target users - may be hard to pin down. We need to define what queries have to be asked of objects in the target schema. The design should not consider issues of compute v. pre-.store.

- *What is the nature of the multi-lingual aspects of the requirement?*

The issue is multi-cultural rather than multi-lingual - if we agree on concepts then 'no problem', although multi-lingual implementation specs are still needed. Synonyms can be a solution. The problem is communication with humans, not between machines

- *What other requirements exist?*

We should always be looking outside our own domains - not re-inventing wheels!

## *Session 3: Implementations and Expectations*

This session, chaired by Colin Bray, began with a second GiMoDig presentation by Tapani Sarjakosi of the Finnish Geodetic Institute on **'The GiMoDig Schema translations based on XSLT'**. This dealt with the service architecture including schemas and schema translations, queries and query translations and on-the-fly generalisation, taking the Global Schema as a given. The architecture uses the OGC WMS and WFS interface specifications with GML as the transport. A key problem to be solved is that of transforming on-the-fly from the local schemas to the Global Schema using XML technology, performing coordinate transformations on the way. This is accomplished using XSLT. XSLT is also used in the portal service to render GML into SVG. It is supplemented by Java functions (JUMP/JTS) for geometric processing. Query support is tricky and a two-phase approach might be advantageous.

Eddie Curtis of Snowflake Software, UK began his presentation on **'Schema Translation between Relational and GML Schema'** by describing an OGC test bed in the Critical Infrastructure Protection Initiative. This involved translation of transportation data from a set of local schema to a global schema, using a translating feature server and GML-3. Findings were that the required data must be present as inferring data is expensive and unreliable, and that creating the translations is currently labour-intensive particularly if metadata is poor or informal. He concluded that syntactic translation is feasible, but in need of industrial strength implementations. The next step is then to develop tools to derive the syntactic translation rules from semantic mappings. He presented an architecture capable of supporting this incremental development strategy, building semantic translation on top of syntactic tools (paper-Curtis).

David Remotti, of the Centro Interregionale, Italy, brought the session to a close with a presentation on **'Toward a National Spatial Data Infrastructure in Italy: on-going activities and research'**. The diversity of the situation in Italy provides a reminder of the practical issues arising from differential levels of development. A specialized UML model, called GeoUML, has been developed along with content definitions and derivation rules for other scales. Two organizational architectures have been tested, one centralised and the other distributed using the WWW. Maps are printed using selection filters and symbolization – work remains to be done on generalisation and on matching database objects at different scales (paper-Remotti).

## *Wrap-up Session*

The final session was facilitated by Peter Woodsford, with Saulius Urbanas as rapporteur. David Remotti advised the group that the upcoming FP6 call would include an action line on Semantics and invited organisations present with an interest in forming a consortium to bid for this to contact him. The full summary is at (**Urbanas**), the key points being;

- Ontologies are still at an early stage. Standards are only just emerging and not yet 'trusted'. The main technical advances will come from mainstream industries
- Users want multiple ways of representing data, together with easy means of publishing and finding data (automated metadata).
- There cannot be 'one semantics'. 'Global schema' will necessarily be per application domain.
- 'multi-cultural' differences are the key issue - progress has to be made on harmonising concepts

- In the near-term harmonising concepts of object identity and lifecycles will prepare the way for more systematic use of ontologies. Geospatial models must be extensible and explicitly represent the notion of identity and discrete information packages for data exchange within and outside the GI domain.
- Key areas to follow beyond syntactic tools such as UML and XML are RDF, RDF Schema, OWL, RuleML, reasoning engines
- A plausible goal is Semantic Reference Systems by the next decade.

## *Thanks*

Thanks are due to EuroGeographics for their splendid facities, to Isabelle Soquet for excellent local arrangements, to Paul Ayling for maintaining the website and preparing the CD, to the organising committee and to all presenters, rapporteurs and delegates for a very lively and stimulating workshop.

Peter Woodsford
Cambridge
27 May 2004
Reviewed by: Glen Hart and Saulius Urbanas
.