0

Introduction creating spatial information infrastructures

- towards the spatial semantic web

Paul Scarponcini, Styli Camateros and Oscar Custers (Bentley Systems)

Sisi Zlatanova and Peter van Oosterom (TU Delft)

Agreeing on high level concepts of spatial data and the development of systems handling these is the first step towards the Spatial Information Infrastructures (SII). OGC and ISO/TC211 have developed a rich set of standards in this area (independent of specific themes or domains). Parallel to this development has been the growth of the Internet and all its protocols that have created the foundation of the SII. This does not mean we do understand each other's information as for this we also have to agree on the domain (or thematic) models. In the context of these models the data get more meaning, and it is fair to state that data become information. Today these models are often expressed as UML class diagrams, often limited to just the data side. The next step is applying knowledge and inference engineering technology. This chapter gives an overview of the different aspects of the SII. Next attention is paid the question what is intended with the spatial semantic web. Finally an overview of the spatial semantics evolution is described. Throughout this introduction chapter references are made to the other chapters in this book, which are going into further detail.

0.1 Aspects of the SII

Besides agreeing on the information content there are a number of other aspects, which are needed to realize the SII. These could be subdivided into technical and non-technical aspects. Among the technical aspects the SII needs metadata, described in catalogues and made available via registry services. The web-based services (including data delivery services) form another technical aspect of the SII. Among the non-technical aspects of the SII, the legal and organizational issues are important: copyright, pricing policy, access rights, etc. The BE 2007 research seminar focused on the technical aspects, and specifically those of the core element of the SII: the spatial information itself. Figure 0.1 gives an impression of how the information within the SII can be combined and used in various application contexts.

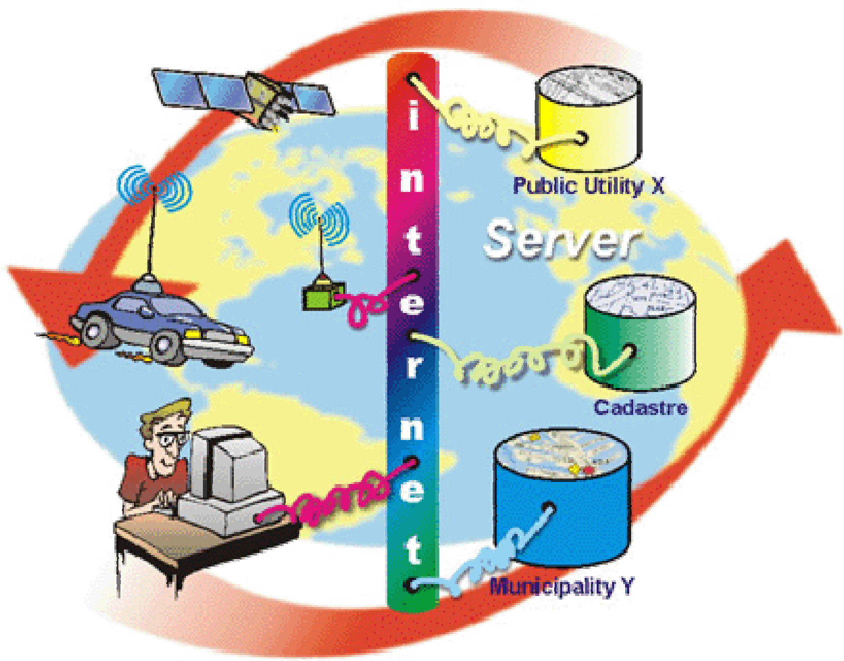

Figure 0.1 Using the Spatial Information Infrastructure

Standardization of themes, a number of examples

After the theme independent aspects of standardizing spatial information, there is now a wave of attempts to agree on complete themes, covering both the spatial and non-spatial aspects. A number of examples of standardizing themes/domains are now briefly presented. Within the (road) navigation sector the Geographic Data Files (GDF, ISO/TR 14825:1996) has been developed by ISO/TC204 Intelligent Transportation Systems (ITS). More recently, 'Recording and exchange of soil-related data' was submitted to ISO/TC190 Soil Quality. A third example is the submission by the FIG of the Core Cadastral Domain Model (CCDM) to ISO/TC211 (Lemmen and van Oosterom 2006). Note that the CCDM has recently been renamed to the Land Administration Domain Model (LADM), see Chapter 9. Also within the private sector, there are numerous development efforts for 'domain' models; though the acceptance or integration of these into the de jure standards may be problematic. Two important advantages of agreeing on domain models are: 1. it becomes easier to understand the information of others within the domain, 2. system developments may be shared as many partners base their system on the same model. The benefit of domain models (and ontologies) for facilitating information discovery and building knowledge-based systems is clear. The drawback of different ISO/TCs (or other organized sectors) for different geo-information themes, is that there is no/difficult harmonization between themes (perhaps confusing overlap and also double work). Anyhow, it will not stimulate interoperability between these themes as needed for a wide spatial semantic web. The development of thematic (semantic meaningful) models is the future of geo-information standardization.

Unpreceeded programs: INSPIRE and US DHS Geospatial Data Model

However, recently there are a number of large initiatives started to develop harmonized (interoperable) model specifications covering many themes. For example, within INSPIRE, 34 different themes are to be covered; see Chapter 1 and http://inspire.jrc.it/. It will be an incredible challenge for the 27 countries of the European Union to realize this: first agree on the harmonized models and next deliver information according to these models. But the situation is not unique for Europe; see the US Department of Homeland Security (DHS) Geospatial Data Model, which also covers quite a broad number of themes; see Chapter 4 and http://www.fgdc.gov/fgdc-news/geo-data-model/. Parts of DHS are based upon the FGDC Framework Data Content Standard, which depends upon the ISO TC211 feature model upper level ontology, described in more detail in Chapter 4. INSPIRE will also adopt various ISO TC211 standards. These large programs will create the infrastructure of which many applications and users will benefit, both within government, private sector and individuals.

Model Driven Approach

Creating the harmonized models and specifying them as UML class diagrams (and documenting them further with the help of feature and attribute catalogues) is in essence capturing (and agreeing on) human knowledge within a certain domain (or closed world). Note that UML is less suitable as basis for creating bridges between domains (semantic mapping, translation and transformations); other tools are needed for this; e.g. OWL and reasoners. These UML class diagrams can be used for the implementation of information systems according to the model driven architecture approach. The same model can be the basis of a database schema (SQL Data Definition Language), an exchange format (XML schema), or most of the user interface and associated behaviour in an edit environment (e.g. automatically generating specific types of forms to enter valid attribute values of a specific feature). Within Bentley, the XFM technology is an important indication of this development.

The scope of SII

Note that the SII does not only cover traditional geo-information, but also (geo-referenced) designs/models, subsurface information (geo-technical, geological, etc.). Clearly, we have the issue of the 3D aspect in many of the relevant themes. Further, as things do change over time, the temporal element is also very important. How does this all fit into a usable interoperable infrastructure? The semantic aspect of information (what does it mean) is not only important for human beings to understand each other, but semantics is also essential if we do want that machines do useful things with that information. Therefore, the semantics will have to be formalized beyond what is currently expressed in UML, especially when trying to harmonize across domains. This is the essence of the semantic web: ontologies, perhaps expressed in OWL.

0.2 The Spatial Semantic Web

So what exactly is the 'Spatial Semantic Web' and how is this different from the proposed Semantic Web and even the current World Wide Web (WWW)? First, it is important to understand the limitations of the WWW.

The World Wide Web

Try to Google the first author's name, Scarponcini, in order to see all of the papers he has written. You will get close to 100,000 hits. Today's search engines merely look for keywords in documents, and, if found, return the entire document. It turns out that 'Scarponcini' is Italian for small (low cut) boot. So most of the documents returned are in Italian and are about small boots. Searching instead for Paul Scarponcini will return a more manageable set of just under 1,000 documents, one of which is a story about *Paul* McCartney buying a pair of *small boots* in Italy. Yes, you could eliminate this one by putting quotes around the entire name. But what if a document contains only P. Scarponcini instead of Paul, or Scarponcini, Paul? The computer, search engine that is, has no idea that a successful find is a document which is a paper written by Paul Scarponcini. It also finds documents which are papers, but which are authored by someone else; Paul Scarponcini appears in the bibliography. Some documents are meeting minutes where Paul Scarponcini is in attendance. The other limitation of the current WWW is that it can only return (whole) documents. It is up to the user to read and interpret the result to see if there is anything of value in the document. The only information contained in the document (other than the document content itself) is the information the computer needs to be able to properly present the document on the screen. In its tremendous success in being able to make information readily available, the WWW may now be suffering because it returns just too much information for humans to consume.

The Semantic Web

The next logical step then is to augment the information contained in documents with additional information which would allow the computer to understand the document content. This can include information about the document (metadata) as well as (ontological) information about the information contained in the document. A search engine could then make better decisions about what documents to return. Metadata and ontologies are discussed in greater detail in subsequent chapters, particularly metadata in Chapter 10 and ontologies in Chapter 3. The web should also be able to provide more than just information in a document. Services will also be provided which can access, manipulate, integrate, and present information as the user wants to see it, instead of just how it appears in a single document. Searching for the right service poses an even greater challenge than searching for information, especially if multiple services need to be chained together to achieve the desired result (Chapter 7). Additional information is available about the Semantic Web from various sources, so it does not have to be repeated in detail here. However, to summarize where we have been and where we may be headed, Chapter 3 provides an assessment of significant achievements on semantics technologies to date as well as important challenges that remain to be solved. What is germane to this book are the spatial aspects of the Semantic Web.

The Spatial Part

Almost every piece of information has a location associated with it. This might be the home address of a person, the business address of a restaurant, the current position of my car as I am driving down the road, or where the speed limit changes. One would therefore expect location information to play an important role in the Semantic Web. In fact, many have proposed using location as a means of integrating other information – see Chapter 6. Augmenting the information that a restaurant only accepts cash with the locations of the restaurant and the nearest ATM might be helpful. But it is not quite that simple. There are numerous ways of expressing locations, from simple street addresses to more precise latitude and longitude coordinates. They can be absolute or relative, as two blocks down on the right (from your current location). So having the street address of the restaurant and the latitude/longitude of the ATM will not help unless you also have a way of correlating these two locations. What is special about spatial information that warrants its recognition in the Semantic Web? First, there are spatial types (point, line, polygon, polyhedron) but these can be described similar to non-spatial data using ontologies. Then there are fundamental spatial (intersects, within, touches) and topological (connected, adjacent) operators which require agreement on their precise meaning. These can then be augmented by more robust combinations using inferencing rules. For example, near might be defined using both distance and connectivity, such as determining if the ATM is near (within driving distance of) the restaurant. Spatial information is tied to a spatial referencing system which, in the case of geo-information, deals with the fact that distances between locations on the earth's curved surface may not be the same as their distance apart on a flat map. Many indexing techniques have been developed for non-spatial or one-dimensional information but when two- and three-dimensional spatial representations need to be indexed, special methods are required. Additionally, much of the nonspatial data in the Semantic Web relates to discreet objects, which are easier to classify in an ontology. Many spatial objects, such as areal coverages and linear events require special treatment both at the (continuous) object level as well as the location dependent property value level (e.g. elevation map), adding to the ontological complexity. And for spatial data sets, relationships are more frequently computed rather than stored. This presents problems for reasoners, which assume that all relationships are explicit.

0.3 Spatial Semantics Evolution

Associating semantics with spatial is not a new concept. As early as the 1970's, ontologies were part of CAD, though not recognized as such. The Building Design System, BDS, software from Applied Research of Cambridge, as well as the follow up General Drafting System, GDS, required the user to identify an object before drawing any linework (McDonnell Douglas 1988). The object name was comprised of up to six facets, effectively creating a type hierarchy. When purchasing such a system, the user was faced with the daunting task of creating a relevant taxonomy of objects, such as BUILT:TRANS:ROAD:HIGHWAY:RT66:EAST for the eastbound carriageway of the Route 66 highway in the transportation domain of the built (vs. natural) environment. The benefit was the ability to then obtain a drawing of just those object classes desired, instead of the more rigorous layered approach common at that time. So you could get all the TRANS objects, no BLDGs, and highways could be rendered green whereas local roads could be brown and dashed. Properties could be attached to objects at either the object or instance level. Eventually, this nonspatial data could be stored in a Relational Database using SQL*CAD (Scarponcini 1989). This was taken to an extreme when GDS/SQL*CAD/Oracle were customized for the design of the $6Billion Wastewater Treatment Plant in Boston (Scarponcini 1990). Object facets were used to identify both the facility being constructed as well as the system to be maintained and operated. All menus were data driven by queries to the database so that only the appropriate objects could be used for a particular facility. The resultant ontology for wastewater facilities was intended as a transition from design/construction into maintenance/operation. And this was in 1988. PCN data structures (for polygons, chains, and nodes) were added to GDS and with the SQL*CAD connection to Oracle, GDS was a predecessor to feature-based GIS.

Integrated data management

Related to the data and operation model driven approach indicated above is the fact that the integrated management of data, both spatial and non-spatial, is the preferred approach. The GEO++ system (Vijlbrief and van Oosterom 1992) based on the Postgres DBMS (Stonebraker et al 1990, de Hoop and van Oosterom 1992) is an early example of this approach. The spatial functionality within the DBMS is provided by Abstract Data Types (ADT) for point, line and area geometry. Note that the spatial functionality was before standardized within SQL (ISO/IEC 1999, ISO 2002a). The interface of GEO++ is completely model driven based on the DBMS catalogues describing the defined tables and the available operators for each data type. This is true for both the 2D and the 3D version of GEO++; see Figure 0.2. The Cadastral query tool is an example of a system based on the GEO++ model driven approach, though the research DBMS Postgres has been replaced by the production DBMS OpenIngres (van Oosterom et al 2002).

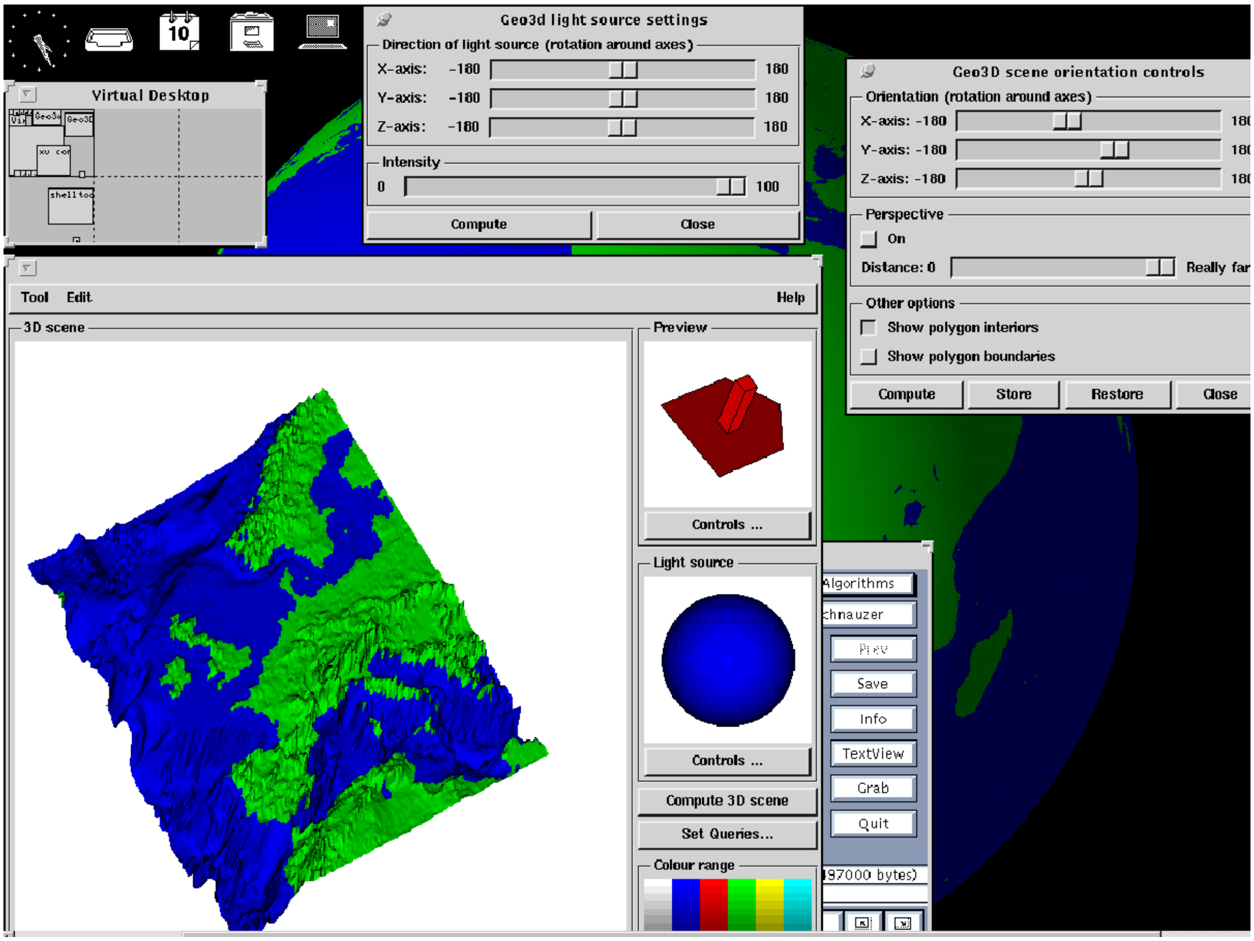

Figure 0.2 Impression of the 3D version of GEO++, a model driven GIS (van Oosterom et al., 1994)

Automated Spatial Reasoning

To extend the semantics of GDS, the *Dafne* prototype was developed in 1991 (Scarponcini, 1995). It captured spatial objects from GDS and their associated properties from Oracle, via SQL*CAD, and instantiated objects in Neuron Object, an AI software package. Inferencing rules were added for reasoning about the spatial data. Sample CAD drawings and GIS maps were created in GDS and *Dafne* was able to infer information from the data that was persisted. The near predicate mentioned earlier was implemented as a rule based on geometry and topology of the roads on the GIS map or rooms on the CAD architectural plan. What made *Dafne* a challenge was that it pre-dated any standards, like OGC and ISO TC211. There were no standard spatial data types, no fundamental spatial operators, and as yet no consensus on the validity of a feature-based approach. These were all proposed as part of the *Dafne* prototype, perhaps a bit ahead of its time.

Spatial Standardization Begins

This all changed in the mid 1990's with the adoption of OGC Simple Features (OGC 1998), ISO/IEC SQL/MM Part 3:Spatial (ISO/IEC 1999), and ISO TC211 (ISO 2002a). The 'simple' describes the geometries, not the features themselves, being limited to points, linestrings and polygons with linear interpolation only and collections of these. Perhaps their greatest initial contribution was a (simple) geometry type hierarchy (ontology), complete with properties and fundamental spatial operators. Perhaps most unknown at the time, the feature model that was specified in TC211 (ISO 2003) and supported by OGC and SQL/MM, changed the fundamental basis of GIS. This is explained in Chapter 4. A feature can have properties, including attributes, operations, constraints, and roles, as well as associations to other features. Of significance is the fact that spatial representation is considered to be an attribute of a feature, rather than being the central organizing type typical in early GIS – see Chapter 2. The feature model allows for multiple geometric representations for the same abstracted real world entity but offers no solution on how to 'harmonize them'. Chapter 8 elaborates on the need of semantic standardisation.

Geometry versus Object (Feature) first modelling approaches

When we look at spatial modelling in the past, in principle we can distinguish between three different approaches: 1) geometry- (topology) first, 2) object- (feature) first, and 3) hybrid approach. Because the related models have quite a different starting point, there is sometimes confusion between modellers.

In the geometry-first approach, the models start from the geometry (topology). Attributes are added to these geometries in order to classify the objects. The result is typically a set of tables in the database such as point/symbol table, text/label table, line table and area table. Within a table all objects (records) have the same set of attributes. For example in the area table there may be houses and roads, all having the same attributes. In this approach, it is also possible to explicitly model the topological structure (e.g. linear network, or partition of space) with well-known advantages (explicitly connectively, avoiding redundancy, better guarantees for quality under updates). The Dutch cadastral map in LKI is a typical example of this geometry (topology) first approach (Lemmen et al., 1998). In this solution objects may share, via topology, their geometry with other objects. It could be argued that map representations (on paper or screen) themselves, i.e. the visualization of the spatial data, is also a geometry-first type of model as all objects are considered together in a geometry model.

The second approach, the object-first approach, models the object classes first with added geometry. Every object class can have its own set of thematic attributes, which may vary for the different object classes. Also every object has its own geometric description independent of any other object. The TOP10NL model is an example of this approach. Typically the result is a set of tables (or relational objects) in the database such as houses, roads, waterways, which have among others their own simple object geometry type attribute. Sometimes additional rules (constraints) are added in order to avoid un-wanted situations (often topology based); e.g. a house polygon should not overlap with a road polygon at same level/layer. The drawback is that all these constraints have to be explicitly stated (and checked when updates are performed) and are not embedded in the main structure of the object-first type of model. Also the model does not explicitly contain the topological relationships, which may support various types of analysis (e.g. quality control of updates). It must be noted that topological relationships are very important for map generalisation; e.g. what are the neighbours of this object (candidates for aggregation), is the network connectivity damaged when this road segment is removed, etc. The ISO 19109 standard makes properties (including geometries and topologies) composites of features, thereby precluding their sharability.

The third approach is the hybrid approach, which treats the geometry and object class equally. It combines the strengths of both approaches: the (thematic) attributes are specifically designed for every object class, but the model also enables shared geometry and use of embedded structure. The spatial domain is partitioned and the result is described using tables for nodes, edges, and faces (and solids in 3D). The objects are modelled in the same way as in the object-first approach with the exception that the objects do not have their own independent geometry-attributes, but refer to primitives in the geometry/topology part of the model (node, edge, face,…). This is the approach as described in the 'formal data structure' (FDS) theory of Molenaar (1989) and quite recently implemented in products such as 1Spatial (LaserScan) Radius Topology and Oracle spatial topology (first introduced in version 10g). It is also supported by the new topology part of the SQL/MM standard and implemented in the ISO TC204 GDF standard. Unfortunately the security system enforced by the SQL language (i.e., GRANTs) did not allow the topological primitives to be implemented as relational objects like the geometry types, so tables were used instead in order to enable sharability of topology. The OGC Simple Feature standard only supports geometry, and sharing it between features is not precluded, as is the case with SQL/MM geometry.

De Hoop et al (1993) discuss the different modelling approaches and the consequences for realization and use. It cannot be claimed that one model is 'better' than another model. This depends on the application context and use. If one specifies a number of important characteristic of the application domain and typical use, then it is possible to state which approach is preferred. Considerations could be:

1. allow exceptional overlapping of objects in certain cases (e.g. bridge over water in 2D),

2. allow modelling of systematically overlapping sets of object classes (e.g. topographic objects at one hand and administrative units at the other hand),
3. enable multiple geometry representations of single objects (e.g. road area polygon and road centre line, or building footprint polygon, building rooftop polygon, and building centroid),
4. support consistent updating/maintenance,
5. support efficient querying, analysis and viewing of data,
6. avoid storage space consuming representations (might also be expensive for data transfer),
7. support easy delivery for customers (simple objects might be easier to receive in another system than topology structure), etc.

Standards Progress

After Simple Features, OGC was successful in launching the WMS, Web Map Service (OGC 2004). Chapter 11 presents the comprehensive Danish strategy for SII based upon WMS and augmented by various registries. WMS, though service based web enabled, suffers from the same limitations mentioned earlier for the WWW – the map, like a document, is retrieved in its entirety, with no knowledge of its content. So now the standards seem to be in a bit of free fall. OGC has proposed WFS, Web Feature Service (OGC 2005), a more semantically oriented version of WMS, and a WCS, Web Coverage Service (OGC 2006), and resultant confusion over whether a coverage is a map or a feature. This has led to the creation of an Architecture Board to help sort out a conceptual model for a family of Web Service specifications but even more fundamentally a strategy for writing specifications and even more fundamentally, identifying the mission of OGC. TC211 appears to be equally astray. The initial 20 domain-neutral standards in the ISO 191nn family presented a concise, consistent abstract model of all aspects of geographic information. The value of their contribution of the feature model as an upper level ontology suitable for the spatial semantic web cannot be underestimated. But the subsequent 30 standards appear to be all over the place, including domain specific topics such as transportation tracking and navigation. Perhaps it is appropriate that standard number 19150 is focused on semantics (ISO 2007). Some of the challenges beyond simple features and the feature model are now being wrestled with, both in OGC and in TC211. This includes observations and measurement, temporal, and services. Chapter 5 focuses on earth sciences. Here, it is often inadequate to model the data resulting from an observation without details of the instrument or process used to generate the data. OGC is working on reconciling an abstract model of observation and measurements with their current GML implementation. TC211 has managed a temporal schema, ISO 19108 (ISO 2002b), defining fundamental temporal types and operators. Moving features, ISO 19141 (ISO 2005), considers changes in location over time. The more difficult problem of features morphing over time has yet to be standardized. Chapter 7 transitions us into the world of services, identified by the operations they support. Each operation can be defined by its input an output parameters which are conveniently based upon the aforementioned feature model. The whole semantics of individual services and service chaining is certainly in the research phase, though solutions to non-spatial service semantics may provide hints to solve spatially enhanced services.

References

de Hoop, S. and van Oosterom, P. 1992. Storage and Manipulation of Topology in Postgres. In *Proceedings EGIS'92*, Third European Conference on Geographical Information Systems, Munich, Germany pages 1324-1336.

ISO 2002a. *ISO IS 19101:2002, Geographic information—Reference model,* International Organization for Standardization, Geneva.

ISO 2002b. *ISO IS 19108:2002, Geographic information—Temporal Schema,* International

Organization for Standardization, Geneva.

ISO 2003. *ISO IS 19109:2003, Geographic information—Rules for application schema,* International Organization for Standardization, Geneva.

ISO 2005. *ISO CD 19141, Geographic information—Schema for moving features,* International Organization for Standardization, Geneva.

ISO 2007. *ISO New Work Item Proposal, Geographic information—Ontology,* ISO document 211n2163, International Organization for Standardization, Geneva.

ISO/IEC 1999. *ISO/IEC IS 13249-3:1999(E) Information technology — Database languages — SQL Multimedia and application packages — Part 3: Spatial,* International Organization for Standardization,

Geneva. Lemmen, C.H.J., E.P. Oosterbroek, and P.J.M. van Oosterom, 1998. Spatial data management in the Netherlands Cadastre. In Proceedings of the *FIG XXI International Congress*, Commission 3, Land Information Systems, pages 398–409, Brighton, United Kingdom, July 1998.

Lemmen, C. and P. van Oosterom 2006. Version 1.0 of the FIG Core Cadastral Domain Model, *XXIII International FIG congress*, Munich, October 2006, 18 p.

McDonnell Douglas 1988. *GDS Reference Manual,* Release 4.10, McDonnell Douglas Corporation, St. Louis, MO.

Molenaar, 1989, Single valued vector maps: a concept in geographic information systems, in: *International Journal of GIS,* Vol.2, No.1, pp.18-27.

OGC 1998. *OpenGIS® Simple Features Specification For SQL Revision 1.0*, Open GIS Consortium, Inc., Wayland, MA.

OGC 2004. *OGC Web Map Service Interface,* Version 1.3, OGC 03-109r1, Open GIS Consortium, Inc., Wayland, MA.

OGC 2005. *OGC Web Feature Service Implementation Specification,* Version 1.1, OGC 04-094, Open GIS Consortium, Inc., Wayland, MA.

OGC 2006. *OGC Web Coverage Service (WCS) Implementation Specification*, Version 1.1, OGC 06-083r8, Open GIS Consortium, Inc., Wayland, MA.

van Oosterom, Peter, Wilco Vertegaal, Marcel van Hekken, and Tom Vijlbrief 1994. Integrated 3D Modelling within a GIS. In Proceedings of the *International GIS workshop AGDM'94* (Advanced Geographic Data Modelling), Delft, The Netherlands, 12--14 September, pages 80-95.

van Oosterom, P.J.M., Maessen, B. and Quak, C.W. 2002. Generic query tool for spatiotemporal data. In*: International Journal of Geographical Information Science*, Volume 16, 8, pp. 713-748

Scarponcini, Paul 1989. 'CADD/DB: Fourth Generation Application Platform for Software Integration,' *Computing in Civil Engineering: Computers in Engineering Practice, Proceedings of the Sixth Conference*, American Society of Civil Engineers, New York NY.

Scarponcini, Paul 1990. 'CADD/DB: The Boston Harbor Cleanup Project,' *1990 International Oracle User Week Proceedings*, Anaheim CA, September.

Scarponcini, Paul 1995. 'An Inferencing Language for Automated Spatial Reasoning about Graphic Entities,' *Advances in Spatial Databases, Lecture Notes in Computer Science #951,*

Egenhofer, M.J. and J.R. Herring (eds.), Springer, N.Y., pp 259-278.Fourth International Symposium on Large Spatial Databases - SSD'95.

Stonebraker, M., Rowe, L.A. and Hirohama, M. 1990. The Implementation of Postgres. *IEEE Transactions on Knowledge and Data Engineering*, 2(1), 125-142.

Vijlbrief, T. and van Oosterom, P. 1992. GEO++: An extensible GIS. In proceedings *5th International Symposium on Spatial Data Handling*, Charleston, South Carolina, pages 40-50.