

CONFERENCE OF EUROPEAN STATISTICIANS

**Joint UNECE/EUROSTAT Work Session on Methodological Issues Involving the Integration of
Statistics and Geography**

(Tallinn, Estonia, 25-28 September 2001)

Topic (iv): Standards and metadata

**OGC: A FRAMEWORK FOR GEOSPATIAL AND STATISTICAL INFORMATION
INTEGRATION**

Submitted by Universitat Jaume I, Spain and Open GIS Consortium¹

Invited paper

**I. INTRODUCTION – THE ADVANCING TECHNICAL FOUNDATION FOR
INTEGRATING GEOSPATIAL AND STATISTICAL INFORMATION**

I.1. Benefits of integration

1. The premise of this paper is that statisticians and all who rely on their work will benefit from the provision of services for integrating geospatial and statistical information. Last year's conference summary report states that GIS was found to be a useful tool in many different areas of statistics, including population census, social and demographic statistics (health, justice, education, labor), economic statistics (business surveys, trade, transport, tourism, agriculture, etc.) and environment statistics. GIS is used in all the different phases of statistical production, and it is useful in cross-sectoral and inter-agency projects as well. The value of geospatial data in statistics is not surprising, because most data types (variables) studied by statisticians have a spatial component – indeed, everything and everyone is somewhere, and statisticians are uniquely aware of how boundary conditions can affect sampling and therefore their results. The emergence of inexpensive computing power, expanding network bandwidth and sophisticated component-based software can potentially offer statisticians extraordinary opportunities for collecting, analyzing, and presenting statistical data from a spatial perspective. This paper attempts to shed light on a standards-based process for designing and building software that best meets the needs of this, or any other, information community.

I.2. Historical limitations and Innovations

2. In the past, most statistical applications have not used geospatial processing², for the same reasons that geospatial processing has been slow to find its way into general purpose computing: lack of discoverable and available data; different processing approaches; different standards, formats and data dictionaries; different quality levels; lack of a common geospatial referencing framework; and lack of consistent metadata and data quality and heritage information. In Europe and elsewhere, diversity in data policy and its interpretation, in data specification, in pricing and access rules, and in private/public sector

¹ Prepared by prof. Michael Gould, Information Systems Department, Universitat Jaume I, Castellón Spain, and Louis Hecht, Jr., Vice President Business Development, Open GIS Consortium, Inc.

² SAS is a notable exception having developed some limited functionality that is present in their product offerings.

relationships have also inhibited wider use of spatial data, as was recently pointed out in the EC *Green Paper on Public Sector Information in the Information Society (COM(98)585final)*.

3. Several recent projects have addressed these problems. The CommonGIS project, funded by the European Commission under the Esprit Program, has yielded a treasure trove of operational and user-oriented insights about the technical foundations that could be made generally accessible as services for visualizing statistical information from a spatial perspective. This good work needs to be carried forward to enable the IT industry to provide services like these on more than a conceptual and demonstration basis. Supporting this and similar initiatives, the ETeMII project, also EC-funded, is creating and disseminating materials which facilitate the uptake of international standards supporting geographic metadata and reference data: two key ingredients for large-scale geospatial statistical analyses.

I.3. Interoperability removes limitations

4. Over the last seven years, the Open GIS Consortium (OGC), a global consortium of geoprocessing technology providers and users, has made important progress toward interoperability between geoprocessing systems, employing practical testbeds and a consensus specification development process to arrive at open specifications for standard interfaces and protocols that can be used by IT suppliers for particular information communities. Simultaneously, data coordination efforts worldwide have made progress toward semantic interoperability based on standard data dictionaries, metadata profiles and geospatial data modeling schemas (discussed below). All of this progress, when viewed together, benefits statisticians who seek to assimilate geospatial processing and geospatial data into their work.

5. There was common agreement at last year's conference that the main obstacles to data integration are not technical but managerial and organizational. However, it will be seen that advances in technology don't merely support managerial and organizational progress – they force such progress. In this paper we report on the tools and methodologies OGC and its members provide for the collection, evaluation and dissemination of geospatial data and its metadata. These tools and methodologies use Web-based distributed processing to “change the playing field” for the management and use of geospatial data and geospatial processing.

6. One limitation noted in last year's report concerns the fact that many commonly used statistical operations are not available in GIS software packages³. As we will see below, this is partly because of the lack of agreement in the definition of geospatial concepts -- hierarchies for example -- making it necessary at times to compare apples and oranges. The classic problem “Are two spatial patterns significantly different” is difficult to analyze when boundaries and/or spatial units are defined in incompatible manners. It is perhaps also because the statistics community has yet to fully communicate its needs regarding geospatial processing software.

II. TRANSLATING USER NEEDS INTO ADDRESSABLE TECHNOLOGY REQUIREMENTS – THE USE OF UML TO MODEL BEHAVIOR AND OTHER COMPONENT CHARACTERISTICS

7. The statistics community appears to be ready to work toward consensus on its requirements for common analytical methods that employ geospatial information and to work with other communities to reach consensus on data dictionaries and metadata schemas. Participants in these discussions will benefit from having a common-yet-open technology platform that encompasses certain software design approaches universally held by industry and employed by OGC and other consortia such as the Object Management Group (OMG), W3C and ISO. Modern object oriented software design tools address real-

³ In addition to SAS, offering mapping functionality as part of its statistical product offerings, ESRI, PCI, and ERDAS (three GIS and Image Processing Vendors) do offer limited statistical functionality as part of their product offerings.

world concerns and cultural complexities while also addressing the complexity of heterogeneous and evolving technology infrastructures. The OGC Technical Committee has followed the lead set by ISO Technical Committee 211 (Geomatics) in the use of the Unified Modeling Language (UML) as their primary method for providing a graphical and lexical description of software object classes and their relationships and interfaces. This helps to maintain overall consistency among the myriad specifications which are emerging from testbed programs (some 12 technical specifications this year alone). See http://www.omg.org/gettingstarted/what_is_uml.htm for more information on UML.

8. OGC's OpenGIS Specification is tiered into three conceptual levels. The Essential Model (a description of the real world) and the Abstract Model (a generic model of the software) are recorded in the OpenGIS Abstract Specification, a 16-volume compendium describing how geospatial processing components should behave and cooperate with one another. The Implementation Model (a model of the software objects in specific executing software environments, and how the software objects communicate in those software environments) is recorded in a growing collection of OpenGIS Implementation Specifications. Utilizing UML in addition to English-language descriptions (which are more prone to misunderstanding) has facilitated the convergence of geospatial models and interfaces between OGC and ISO TC/211, such that in many cases although text descriptions vary slightly, a software designer can rely on the UML class diagrams of either organization to build geoprocessing software which is compliant with both.

9. A similar approach could help the statistics community to record and communicate requirements: what you'd like geoprocessing software to be able to do for you. These requirements can be inserted into OGC's consensus process, resulting in new or modified OpenGIS Implementation Specifications defining open interfaces and protocols that are ready-made for the statistical community. This will prove to be a valuable methodology and process as statisticians begin to employ new Web-based work flows that involve geospatial data and online spatial processing resources. Hence, we are suggesting that through collaboration the statistical community can directly invoke the market to make its needs known and applications addressing those requirements will quickly follow!!

III. THE TECHNOLOGY CONTEXT OF THE METADATA PROBLEM

10. This section describes the approach OGC is taking to produce an open platform of interface specifications that will enable web-based integration of geospatial and statistical information to happen. OGC's task has always been to develop open software interfaces that constitute a spatial "lingua franca" for software systems. Distinct from OGC's task, data coordination – the development of standard geospatial feature naming conventions and metadata schemas ("semantic standards") -- has traditionally been the task of "data coordination" groups which are sometimes adhoc and other times recognized standards bodies such as the US Federal Geographic Data Committee and Eurogeographics. For some years, commercial software has been available that provides templates that support implementation of standard data dictionaries and standard metadata schemas. However, technology to support data coordination is now moving far beyond those simple tools. Recent web-based computing developments in OGC simultaneously support data coordination, increase the importance of data coordination, reduce the need for data coordination in some uses, and leverage the value of data coordination through automated processes that operate on metadata.

11. To begin, it is helpful to understand OGC's conception of an "Information Community." An information community is a group of people who share a common geospatial feature data dictionary (including definitions of feature relationships) and a common metadata schema. There are many information communities because different disciplines, professions, and sometimes regional groups necessarily will define roads, for example, in different manners. It is good for each discipline and profession to work toward becoming a single information community, because members of the community can then easily share data, in all its discipline-specific detail. And it is good for diverse information communities to seek some semantic commonality so that they can easily share data and reduce the burden of redundant data collection.

12. It is also important to understand the notion of “Framework Data”. Framework data is a limited set of data layers – transportation, hydrography, cadastral and administrative boundaries, elevation, digital imagery, and geodetic control -- which provide a base on which to collect, register, integrate and analyze statistical data. Framework layers are (or should be) publicly available, maintained for the common good, useful for many purposes, and each is likely to comprise at least a subset of that data layer for any particular Information Community. ISO TC/211’s metadata standard (ISO/CD 19115 Geographic Information -- Metadata, currently in a committee draft version) provides common schemas for describing these Framework layers, and ISO/CD 19107 Geographic information – Spatial Schema, provides standard definitions of the geometric and topologic characteristics of geospatial data, which can assist information communities in their quest to compare like entities: apples with apples.

III.1. Catalogs and Registries in the Spatial Web

13. Emerging technologies now make it possible to exploit the geospatial properties of data on the Web in order to facilitate retrieval and data fusion, creating a sort of Spatial Web. How will users (or the software components serving the user) of this Spatial Web find the most recent, or the “best,” or the most accurate, or a particular small subset of framework data for a particular region?

14. In the USA, before the Web became widely used, there was “The NSDI Clearinghouse,” Federal Geographic Data Committee’s online catalog to the US library of geospatial data sets in which data holders would register standard metadata describing their holdings. Data seekers could search the Clearinghouse, find a data set, and download the data using ftp, or perhaps they would order the data on tape or CD.

15. In Europe Megrin did something similar, creating the Geographic Data Discovery Directory (GDDD), and the Centre for Earth Observation (CEO) created the CIP protocol for geospatial information discovery. Diverse clearinghouse solutions had suddenly appeared in the geospatial information field, however there was little if any ability for one solution to utilize another: lack of interoperability.

16. Then OGC delivered the OpenGIS Catalog Services Specification for standard interfaces enabling open access to catalogs of geospatial data (such as the Clearinghouse). It then became clear that any user could then have access to not one Clearinghouse but many, that could be recursive (cascading servers), and that would be online and could operate like one big *virtual* Clearinghouse without the need to physically centralize the metadata. The specification provides for discovering data sets, but it also provides for discovering specific data about specific geospatial features, for example NUTS region boundaries or international highway segments. This decentralized approach allows for local data/metadata maintenance and global discovery.

17. Those responsible for the geoprocessing needs of a governmental, commercial, or industrial organization, will find it useful to know the difference between *catalogs* and *registries*; they will need to know about those catalogs that index their data and about those registries that define their metadata. A registry records types and a catalog records instances. For example, somewhere on the Web there will be a registry defining the standard types of geospatial features contained in a transportation layer. ISO TC/211’s metadata standard (ISO/CD 19115) will be in a registry, and it will contain the authoritative schema for the transportation data and metadata. But perhaps you work for a local highway department, and your schema, though based on the ISO standard, is particularized to accommodate 1) the schema used by the national highway department and 2) the schema used by the transportation department of the major city in your region. Both of those metadata schemas will be in online registries. Your region’s actual transportation data will be a collection of instances of such data, and it will be described in all its uniqueness in at least one catalog.

18. How will a query work? Imagine that an engineering firm is preparing a bid to repair a bridge in your region. The firm needs to know the details about the road supported by the bridge. Your region’s data is

online in a catalog, and the metadata describing your region's data is structured according to a schema in the aforementioned registries. The query will go to the catalog first, discover the schemas that must be accessed to make sense of the data, go to the registries to get those schemas, and then complete the query, returning to the engineer information about the road that crosses the bridge.

19. The transportation data returned to the engineer is Framework data. Its dictionary of feature types may be slightly non-standard by US DOT or similar national standards, but the differences will be documented in machine-readable text in the registries. The Web's XML standard will play an important role here: XML allows the unambiguous, self-documenting transfer of information between systems. This goes for the transfer of metadata as well as, in some case, of the geospatial data themselves. OGC's Geography Markup Language (GML) is essentially an XML dialect for geospatial information -- Spatial Web infrastructure layered on Web infrastructure. Key reference data agencies such as US Bureau of the Census and UK Ordnance Survey have already taken steps to provide geospatial data transfer in GML format.

III.2. Building a Foundation for Semantics and Ontologies in Web Services

20. Just as data will be found through catalogs and registries, servers holding geoprocessing applications services will be discoverable anywhere among the WWW, through service catalogs and registries. The OGC Web Services (OWS) Initiative is a group of projects, planned or already under way in OGC's Interoperability Program, designed to create methods to publish, retrieve, bind, and chain together geoprocessing services. General IT web service standards (including WSDL, UDDI, and SOAP) and the availability of broadband for Web-based distributed processing have matured to the point where such processing is beginning to be practical, and it makes sense now to apply these approaches to geospatial interoperability.

21. OWS sponsors and participants envision that geoprocessing applications can be dynamically composed of services discovered and marshaled "on demand" at runtime. A Web service is a self-contained, self-describing, modular application that can be published, located, and dynamically invoked across the Web. Web services perform functions ranging from simple requests to complicated business processes. Once a Web service is deployed, other applications and other Web services can discover and invoke the deployed service. Services will be described in service catalogs. Types of services will be described in service registries.

22. Data catalogs will point to particular online data sets. Data registries will provide the definitive data dictionaries and metadata schemas for particular information communities. Service catalogs will point to particular online services. Service registries will provide definitive descriptions of the types of services. All of this text data will need to be parsed automatically. The Web's XML standard and OGC's GML standard will play important roles here. GML version 3.0, due to be released by the end of 2001, will provide support for metadata. In essence, GML works by saying to a browser "I'm a collection of spatial features, as described in the following definitions (including a data dictionary and other metadata). Use me for analysis and display me as you are programmed to do." These web service-based technologies have been designed to offer just the right level of "glue" to permit service chaining through communication on the Web via standard interfaces, while at the same time allowing information communities to maintain their particular professional or regional semantics.

III.3. An Example: Language-neutral Modeling of Spatial Features using OGC and ISO specifications

23. The services mentioned above are all designed to be independent of the information represented: OGC does not dictate how information communities should think or represent their key concepts. OGC and ISO/TC211 Specifications (for example, spatial schema, GML, features and service architecture) enable the creation of consistent models representing common geospatial features (bridges, roads, rivers, etc.) in a manner which is language-neutral. This is accomplished using the IT-standard Unified Modeling Language (UML) to represent both exemplar "use cases" and more detailed "class diagrams". UML class

diagrams can be interpreted in a language-neutral manner and can represent indirect mappings of concepts from one information community to another, and can even be translated to application code (C++, Java, etc.). OGC is currently testing translations from XML code-based services to UML and vice versa, allowing developers even more liberty to create generic -yet-standard component software solutions meeting specific information community needs.

24. This language-neutral representation in UML can help solve one of the important problems encountered by statisticians in their use of geospatial data. For their studies to be valid, statisticians must compare *comparable* features (commonly spatial statistical units), and these models will facilitate this process. When using imagery (raster data) it is quite simple to rectify several layers and then vary (resample) the resolution of one data layer to match that of another. In raster data the basic geospatial unit is the pixel (cell), however when dealing with so-called vector data – features — this exercise becomes quite difficult, as the definition of geospatial units is a very diverse exercise: there will be many different representations of, for example, Germany, Northrhine Westphalia, Koln and the neighborhoods within Koln.

25. Beyond the modeling OGC does not expect a single homogeneous definition of any particular feature to be created, rather the schema modeling described above supports the mapping between features, which in most cases will not result in a 1:1 mapping. A simple example of this is the heterogeneous spatial coverage of a Eurostat NUTS region. Another example might be where one information community model represents the concept *forest* and another only supports *tree stand*, the geospatial model in UML would represent how *tree stand* could be aggregated to map to *forest*. In this case the second information community would not need to change its conceptual representations (its ontology), as the UML model allows to map concept to concept, facilitating the semantic interoperability necessary for the proper function of geospatial web services across information communities and crossing linguistic, cultural and even professional barriers.

26. When completed, these UML models will grease the wheels of many inter-information community projects. Statistical analysis crosses many of these same information community boundaries, and so, again, it is in the interest of the statistical community to make its needs known, and to join in on the development of new schema models.

IV. ENABLING STATISTICAL MARKETPLACES

27. OGC's commercial members participate in the consortium because they see commercial opportunity arising from a standards infrastructure. Interoperability enhances market activity in a number of ways:

- Vendors share the costs of developing the infrastructure-level elements of their software, so they have more to spend on product differentiation at the application level.
- Componentization of software based on open interfaces leads to less expensive solutions, best practices and best of breed service delivery. Because vendors offer, and customers choose services web and user customized to a particular task.
- Open interfaces provide the “standard hooks” that developers of non-spatial software can use to integrate components that provide their customers with particular spatial capabilities.
- Open interfaces remove the barriers between different kinds of geoprocessing software including GIS, earth imaging, automated mapping and facilities management, navigation, and location services. This expands the opportunities for everyone and increases the use of geoprocessing in the bigger “supermarket” of data and services. Participation in OGC initiatives allows “cooperation” among commercial entities which previously only competed. They now see tangible benefits in not reinventing basic infrastructure (common interfaces), in the same sense that recently several competing European telcos have agreed to share the costly 3G basic infrastructure instead of duplicating (or quadrupling in some cases) investment.

28. If statisticians make their requirements known in OGC and if vendor members see a significant market opportunity, all of the “market enablement” factors listed above can begin to operate in the area of geospatial/statistical integration.

29. Other special interest groups have sponsored testbeds in OGC’s Interoperability Program designed to yield candidate OpenGIS Specifications that specify, through open interfaces, all that is necessary to meet specific needs, stopping short of specifying the algorithmic approaches that might constitute a software vendor’s proprietary processing methods. Similarly, statistical offices could form an interest group to influence the software producers to implement GIS software functionality’s required for statistical purposes. As a special interest group, statisticians can address their needs and technology requirements.

30. It is worthy of statisticians’ notice that one such Interoperability Program sponsor launched a new domain of interoperable spatial processing – Geospatial Fusion – which enables spatial data to be collected from text documents and organized in useful ways. Place names in gazetteers, for example, are searched for in text documents to discover references to those places, and special folders are created that enable efficient integration of diverse data sources (including Web links and non-graphical data sources) whose data reference a particular place.

31. What will the next new domain be? If it is to be a statistical domain, then will your organization be one of the leaders?