Andrea D'Andrea – Achille Felicetti – Matteo Lorenzini – Cinzia Perlingieri

Spatial and Non-Spatial Archaeological Data Integration using MAD

Abstract: GIS is an efficient tool for the management of complex geo-spatial datasets, but geographic information is stored in heterogeneous environments which makes sharing very difficult. To overcome this lack of interoperability, the Open Geospatial Consortium (OGC) has created an XML-based Geographic Markup Language (GML) to provide an XML-based encoding of geo-spatial data and make them portable and flexible enough to be used in different contexts. Data encoded in GML can be integrated with non-spatial data using MAD (Managing Archaeological Data), an application designed to manage structured and unstructured archaeological excavation datasets in order to create complete XML-based systems. This paper will present the GIS extension of MAD enabling the integration with non-spatial excavation information to preserve the native web compliancy of data and the possibility of managing unstructured documents (excavation diaries and reports) in a spatial context.

GIS Integration Problems

We could broadly define GIS as an integrated system ideal for managing the cartography and related information of a particular territory. The evolution of GIS, like that of all types of informatics systems, is undoubtedly directed towards the transmission and sharing of data through the Internet. This objective however requires the standardization of geographical data, currently underway thanks to institutions such as the ISO with its ISO 19100 standards (very important, even if still scarcely used today by archaeologists) and the OGC (Open Geospatial Consortium, http://www.opengeospatial.org/) with their OpenGIS project and the development of the GML standard.

Data storage in a GIS environment is still conceived as the juxtaposition of two different types of information. At the outset, the basic GIS cartography is set up with a reduced semantic content, often supplied in proprietary formats such as CAD (DWG) or ESRI (SHP). The information necessary for georeferencing is managed by files with a .twf extension, as distinct from the .shp, .dxf, or .dwg files and which are combined during the elaboration phase. Then we find data regarding the area depicted in the map (elevation, archaeological sites, etc.) contained in a table inside the database and "attached" to the map in a .dbf . This table cannot contain graphical data, which must instead be placed in an external file with a different format.

The structure is therefore very rigid, unsuitable for potential web-based publications or data development, and elaborated in closed formats unable to interact with the code and compatible only with the platforms originally used to create them. There is no standardization or interoperability between the different software platforms.

The Open Geospatial Consortium and GML

One of the possible solutions to this problem is the use of the GML format, carried out by the Open Geospatial Consortium, both at the "local" level (that of the private user who uses a web-GIS on his/her computer) and in a "public" environment (when a public administration or corporation publishes an online web-GIS).

The GML standard defines a language aimed at objects, in which every entity contains both geographic and also other kinds of information. This means:

- every "entity" is a separate class; and
- the objects are differentiated according to the class they belong to and not according to their code.

The GML structure is very flexible: the objects are grouped together not because they belong to the same information table, but according to logical criteria. Moreover, an object may contain other objects: spatial and dimensional data may be added to cartographic data and structured according to the ISO 19107 and ISO 19111 standards, instead of being recalled by external files.

The syntax followed by GML is based on the XML grammar devised by the World Wide Web Consortium (http://www.w3.org/XML/ [29 Nov 2007]), a further element of standardization, as XML is compatible with any kind of database.

At present, all major software, whether proprietary or open source, is able to handle the GML 2.0 format. However, the GML 3.0 standard has been available since 2003, allowing the handling of more information, including:

- complex, non-linear, three-dimensional geometries,
- topology for bi-dimensional elements,
- the ability to visualize elements with temporal and/or dynamic components,
- the use of referencing and/or unit of measurement systems, and
- compliance with other prescriptive standards.

It also possible to define the most appropriate data structure for every single application.

An XML grammar follows, expressed in XSD and used to convey, model and utilize geographical information. The basic concepts used by GML for modelling are taken from the OGC's Abstract (available at http://www.opengis.org/techno/abstract.htm). GML provides numerous types of objects to describe geography, amongst which are entities, systems of coordinates, geometry, topology, time and units of measurement. The GML 3.0 specification document is also available online (https://portal.opengeospatial.org/files/?artifact_id=7174).

In order for it to be represented, the GML data must be "styled" through the use of a rendering tool that interprets GML. The GML elements must simply be re-encoded in a format that can be interpreted by a graphical display in a web browser (map styling).

This requires the interpretation of GML contents by means of graphic symbols, line styles, area and volume filling. It often also requires the transformation of the geometry of the data into its visual presentation. Extensible Stylesheet Language Transformation may be used to execute this "styling" operation. It uses Extensible Stylesheet Language (XSL), a language for the transformation of one XML document (for example, GML) into another type of XML document (for example, SVG) according to specifically defined transformation rules.

Generally this graphical rendering procedure transforms GML data into an XML graphical format. Fundamentally, there are three open-source solutions: the first entails the use of OpenJump GIS viewer, which saves any project directly as a .gml file, and is also able to view any .gml file directly without modifying its structure. The second is GRASS which, since version 6.0, allows the user to save projects with a .gml extension by means of OGR libraries that manage vector files. The third solution is PostgreSQL and PostGIS paired together, which as they completely support the OGC's "Simple Features Specifications for SQL" allow vector data stored using PostgreSQL to be viewed directly with PostGIS. This also makes it possible to export files in SVG and GML format. The files are then interpreted by map servers aided by GDAL 2.0 libraries and published to the web.

The MAD Application

GML, as explained above, is also the ideal format for data integration due to its XML nature. To put to work real frameworks of integrated data using XML, we need powerful and flexible XML-oriented tools able to store and manage large and complicated information and make them available for any kind of application: i.e. a native XML data management application is required to preserve its native format for fully-featured GML-native geospatial operations alongside the traditional non-spatial ones.

The environment used to achieve this is MAD (Managing Archaeological Data), a tool developed as part of the EPOCH project through a fruitful collaboration among PIN (University of Florence-Prato, Italy) and CISA ("L'Orientale" University of Naples, Italy). MAD is the result of two years' research on open source technology, XML and international standards for cultural heritage data management and was originally designed as a web-based application to natively store and query XML-based archaeological datasets while preserving their original structure in order to be easily shared and reused in the new Semantic Web scenarios.

MAD archaeological records are encoded using the CIDOC-CRM, a standard ontology created to describe concepts and relationships used in cultural heritage documentation (Crofts et al. 2005). The application provides a complete set of advanced web interfaces which make it possible to perform structural and semantic queries using XQuery, SPARQL and RQL query languages: thus MAD works on XML data in the same way relational databases deal with tables and records.

One of the main advantages of an XML-native approach is its ability to manage unstructured archaeological excavation datasets, such as excavation diaries, as well as complex XML structures like GML geospatial information.

MAD is XPath/XQuery-aware (http://www.w3.org/TR/xpath; http://www.w3.org/TR/xquery/ [29 Nov 2007]) and features dynamic XSLT transformation (http://www.w3.org/TR/xslt [29 Nov 2007])and presentation of documents and query results. XML and RDF documents are indexed and stored in MAD in a UNIX-like set of folders and subfolders, a structure which is easier to manage than the traditional RDBMS.

GML Management in MAD

MAD can completely replace any traditional RDBMS, since GML data can be stored and queried directly in its native XML database. Furthermore, GML fragments can be generated on the fly according to the user's request using extended XQuery functions to implement geographic queries. XQuery, a language specifically created to extract relevant information from complex XML documents, is flexible enough to query a broad spectrum of XML information sources, including GML documents. The XQuery processors can also be easily extended by calling external function libraries without modifying or recompiling their source codes. Extended XQuery processors can deal with geospatial data, query and extract information from GML and other non-spatial data documentation, as well as preparing query results for presentation on the web via a map server web engine. A short description of how the application works follows:

- Users create their own semantic and geographic queries using the web interfaces provided by MAD, selecting the kind of presentation to be returned;
- both spatial and non-spatial documents within the MAD XML archives are then searched for the queried information;
- the query results are transformed and combined to return relevant archaeological data and to build archaeological maps based on GML;
- spatial information encoded in GML is then sent to the presentation framework to be shown in the browser.

To return the results of the geographic queries and to create the required maps and layers we used the MapServer environment, an open source framework for building spatially-enabled web applications *(Fig. 1)*. MapServer builds upon other popular open source or freeware systems (Shapelib, FreeType, Proj.4, GDAL/OGR) and supports several OGS web specifications (including GML) and popular scripting and development environments (PHP, Python, Perl, Ruby, Java, and C#). It excels at rendering spatial data (maps, images, and vector data) for the web.

Thanks to the extended XQuery functions, MAD can generate on the fly all the spatial data needed by MapServer to handle the user's request (MAP files, templates, GML documents, raster data), querying either existing GML documents or extracting spatial information from the data recorded during the archaeological excavation work (such as coordinates of excavation layers). In the latter case, brand new GML documents or fragments can be dynamically generated from non-spatial archaeological records and assembled by the application in order to be sent to the map server framework for the geographical visualization. This demonstrates the complete integration of the data, since the same information can be taken from any document, either geographic or non-geographic.

The system is also fully portable thanks to the ability of MAD to transform the GML documents or fragments via XSLT stylesheets directly into KML, to be exchanged and used in a Google Maps context, or into SVG, avoiding the implementation of a map server and using the MAD XML transformation features instead *(Fig. 2)*. Graphical objects can be grouped, styled, transformed and composited into previously rendered objects. The feature set includes nested transformations, clipping paths, alpha masks, filter effects, template objects and extensibility. SVG drawings can be dynamic and interactive, as the Document Object Model (DOM) for SVG allows straightforward and efficient vector graphics animation via JavaScript and a rich set of event handlers such as *onmouseover* and *onclick* can be assigned to any SVG graphical object directly from MAD.

Conclusion

The integration of both geographic and non-geographic data within a single database, managed using the same interfaces and the native web compliancy provided by MAD, makes archaeological information ready to be queried, updated and ex-

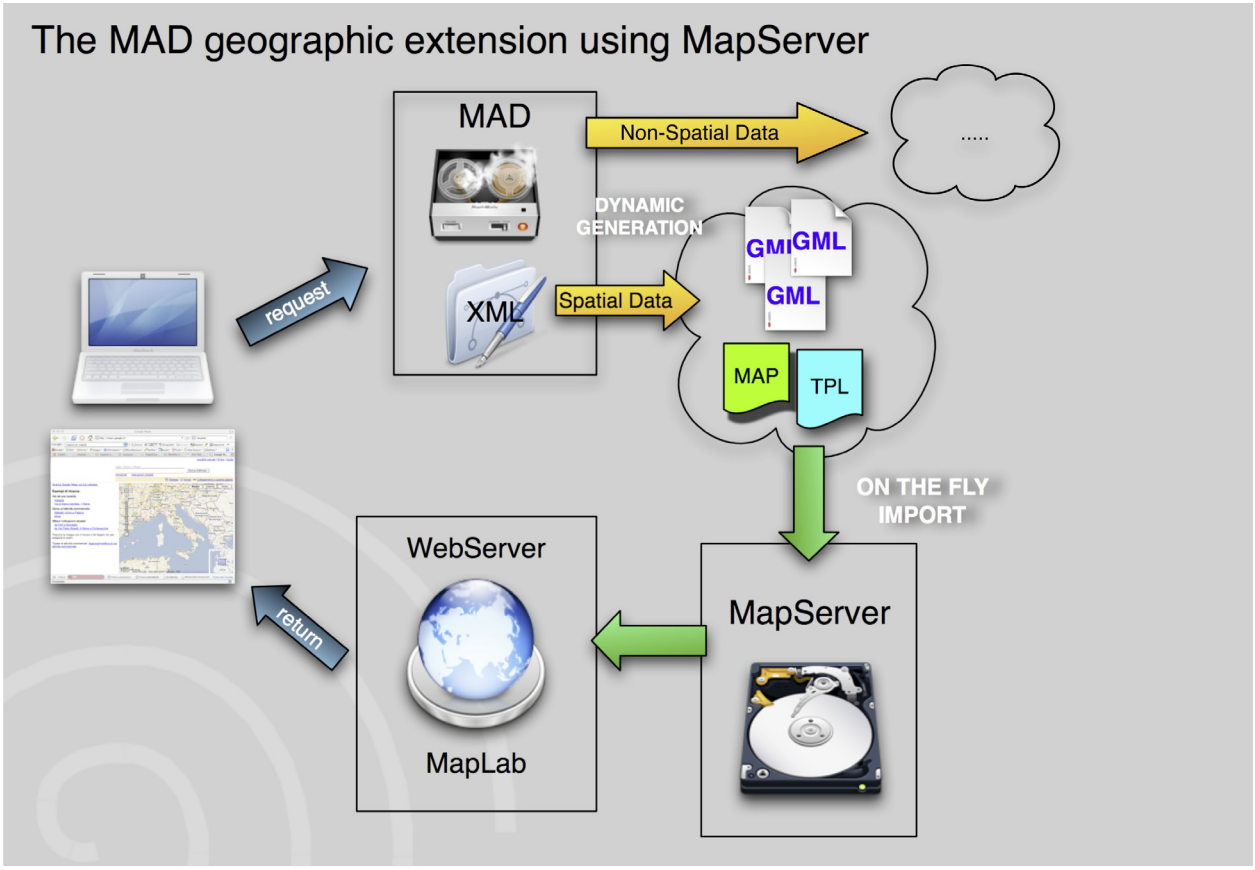

Fig. 1. MAD, GML, and MapServer in action.

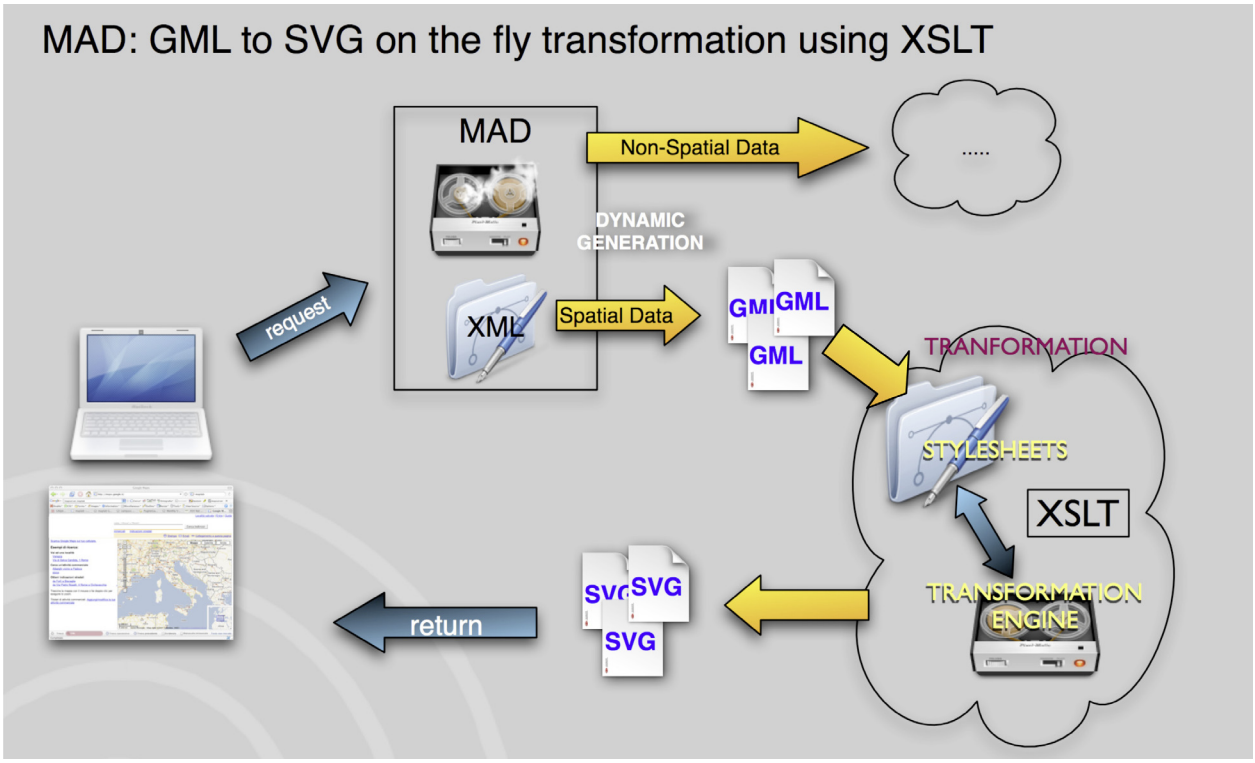

Fig. 2. GML to SVG serialization using MAD.

changed over the web, promoting the semantic evolution of geospatial web services.

The development of MAD for the management of both spatial and non-spatial archaeological data is the first step towards the full implementation of the geospatial semantic web, a future where the World Wide Web will be machine-readable and fully integrated, allowing the returning of both spatial and non-spatial resources to semantic queries.

The Semantic Web will offer something not previously available on a grand scale: interoperability that joins not only tightly structured spatial data such as that in online spatial databases, but also the unstructured, informal geographic information distributed throughout many web archives but not originally intended by its authors for geospatial processing.

References

Abiteboul / Buneman / Suciu 1999
S. Abiteboul / P. Buneman / D. Suciu, Data on the Web: From Relations to Semistructured Data and XML (San Francisco 1999).

D'Andrea et al. 2006
A. D`Andrea / A. Felicetti / S. Hermon / F. Niccolucci / T. Zoppi, I linguaggi del W3C e gli strumenti Open Source per la gestione dei dati archeologici. In: R. Bagnara / G. Macchi Janica (eds.), Atti del I Workshop su Open Source, Free Software e Open Format nei processi di ricerca archeologici, Grosseto, Italy, May 8, 2006 (Grosseto 2006).

Berners-Lee 1998
T. Berners-Lee, Semantic Web Roadmap: an attempt to give a high-level plan of the architecture of the Semantic WWW. http://www.w3.org/DesignIssues/Semantic.html [29 Nov 2007].

Chaundri / Rashid / Zicari 2003
A. B. Chaundri / A. Rashid / R. Zicari, XML Data Management: Native XML and XML-Enabled Database Systems (Boston 2003).

EPOCH
The European Research Network of Excellence in Open Cultural Heritage (EPOCH), Homepage. http://www.epoch.eu [29 Nov 2007].

Felicetti 2006
A. Felicetti, MAD: Managing Archaeological Data. In: M. Ioannides / D. Arnold / F. Niccolucci / K. Mania (eds.), The e-volution of Information Communication Technology in Cultural Heritage. Where Hi-Tech Touches the Past: Risks and Challenges for the 21st Century (Budapest 2006) 124–131.

Hermon / Niccolucci 2000
S. Hermon / F. Niccolucci, The Impact of Web-shared Knowledge on Archaeological Scientific Research. In: Proceedings of the Fifth Conference on Current Research Information Systems, Helsinki, Finland, May 25–27, 2000 (Helsinki 2000).

Crofts et al. 2005
N. Crofts / M. Doerr / T. Gill / S. Stead / M. Stiff, Definition of the CIDOC Conceptual Reference Model 4.2. Heraklion , June 2005. http://cidoc.ics.forth.gr/docs/cidoc_crm_version_4.2.pdf [29 Nov 2007].

OGC
Open Geospatial Consortium, Geography Markup Language (GML) Implementation Specification, 2003. https://portal.opengeospatial.org/files/?artifact_id=7174 [29 Nov 2007].

Salminen / Tompa 2001
A. Salminen / F. W. Tompa, Requirements for XML Document Database Systems. In: ACM (ed.), Proceedings of the 2001 ACM Symposium on Document engineering, Atlanta, USA, November 9–10, 2001 (New York 2001) 85–94.

Zlatanova / Prosperi 2005
S. Zlatanova / D. Prosperi, Large-scale 3D Data Integration: Challenges and Opportunities (Boca Raton 2005).

Andrea D'Andrea
Cinzia Perlingieri

CISA
Università di Napoli L'Orientale
Palazzo Corigliano
Piazza S. Domenico Maggiore, 12
80134 Naples
Italy
dandrea@unior.it
cinzia.perlingieri@fastwebnet.it

Achille Felicetti
Matteo Lorenzini

PIN, University of Florence
VAST-LAB
Piazza Ciardi, 25
509100 Prato
Italy
achille.felicetti@pin.unifi.it
matteo.lorenzini@pin.unifi.it