Article for Geoinformatics - December 2006 issue

Data Compression and OGC Standards

Author: Sam Bacharach
Executive Director, Outreach and Community Adoption
The Open Geospatial Consortium, Inc. (OGC)

Geospatial data is often voluminous – it takes up large amounts of disk space, requires a long time to process, and occupies a long time to transfer across networks. This is a longstanding problem which is aggravated in some network-based applications because network bandwidth is usually less than the bandwidth inside a computer (bus bandwidth).

However, the service interface and encoding standards developed in the Open Geospatial Consortium (OGC) have changed the usual paradigm from bulk transfer of files to service-based access to data residing on a Web server. A Web service query typically results in a response consisting of a relatively small amount of data, and Web services applications can provide near real-time interactivity, so overall work efficiency is greatly improved.

This ability to "get only the data I want" is one way around the problem of large geospatial data files. But sometimes the amount of data requested is still too large, and data compression must be part of the solution.

Data compression is widely used to deal with the problem of bulky data files. Data compression uses efficient methods of representing repeating patterns in data to reduce the amount of memory required to store the data. For example, an image that has a resolution of 200x200 at 8 bits per pixel requires 40 Kbytes of space. If this image can be compressed at a compression ration of 20:1, the amount of space required is only 2 Kbytes. Naturally, a compressed file can be transferred more quickly over a network.

Compressing data sometimes also reduces processing time, but the main benefits are in storage and transmission. It should be noted that compression and decompression require time to process, which may be significant or insignificant in an application.

Below we look at work that has been done in the OGC to deal with data compression for imagery and for vector-based data.

GML in JPEG 2000

Though there are many image compression formats, JPEG 2000, a product of the Joint Photographic Experts Group standardization committee (not the OGC) is rapidly becoming the standard for high-quality image compression in the geospatial domain. Like most image formats, JPEG 2000 is used in other industry domains, such as medical imaging, pre-press and even full motion video, but it includes a number of enhancements important for geospatial applications. One enhancement is the ability to embed eXtensible Markup Language (XML) data in a text-based "box" that accompanies the compressed image data.

XML is a key Web standard for encoding data in a format that enables both humans and computers to read it. In industries of all kinds, experts have worked together to define XML schemas suitable for encoding their industries' data. XML enables one insurance company's accident report, for example, to be converted "on the fly" into the accident report format used by another company.

Independently of the JPEG 2000 effort, geospatial experts in the OGC developed the Geography Markup Language (GML), an XML grammar for encoding geographic information. (GML is on track to join JPEG 2000 as an ISO standard.) After both JPEG 2000 and GML had been developed, a joint effort between the JPEG working group and OGC led to a new OGC standard called "OpenGIS(R) GML in JPEG 2000 for Geographic Imagery (GMLJP2) Implementation Specification," released in February, 2006. This standard defines the means by which the OGC Geography Markup Language (GML) is to be used within JPEG 2000 images for geographic imagery.

JPEG 2000 doesn't specify mechanisms for georeferencing an image, describing the sensor model used to collect the image, specifying styling, defining feature attributes, or correlating features within the imagery to other GIS datasets. These and other things are possible with "GML in JPEG 2000".

It's important to distinguish GML from metadata formats. GML is a language used to construct definitions for features, geometries, etc. By itself, it doesn't define specific features. GML is used to construct "application schemas" for use within a given application or system. A GML application schema is an XML schema specific to a particular geospatial application. It describes the object types whose data the application must expose. For example, an application for hydrology would define object types such as streams, lakes, and oceans in its application schema. Those object types in turn reference the primitive object types defined in the GML standard. GML, a very large specification, is made manageable by GML application schemas that pare away unnecessary XML tags, tailoring the GML data for efficient use in a particular application. The application schema accompanying a JPEG 2000 image can be accessed via standard XML linking and referencing mechanisms.

Many simple uses of JPEG 2000 don't need GML, but as both standards become more widely used, there will be great synergy in being able to use them together. As user communities begin to use the Web for accessing geospatial services and data, and as they become familiar with application schemas that simplify the use of GML, GML is becoming the geospatial language of the Web. It is used to describe regions and extents, define and label features, and express queries. GML makes it easier to use catalogs to publish and discover large archives of distributed geospatial data, and to access this data based on query parameters. Both vector based and raster based services can be accessed, such as mosaicking and layering operations, feature classification, description and extraction, and styling for presentation.

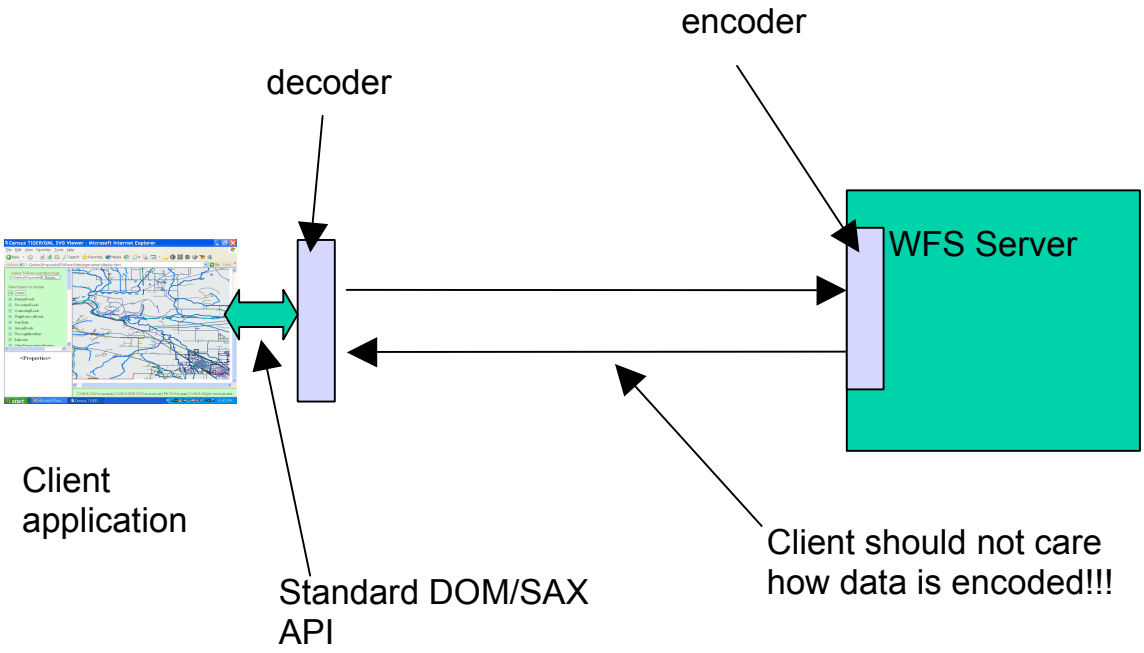
Much of the imagery functionality available in the world of open distributed geoprocessing will depend on being able to use GML for characterizing the imagery and its associated features.

BXML for Compressing GML

Sometimes, even with a well-designed application schema, GML files can be unacceptably large for an application. XML is ASCII text, and GML is XML, so, compared to binary encoded geospatial data, GML is both bulky and slow to process.

To help GML realize its potential, members of the OGC looked for ways to compress XML. No standard, open source tool was available, so OGC member CubeWerx (Canada) developed the "Binary eXtensible Markup Language" (BXML) and submitted it into the OGC's consensus process. BXML is a patent-unencumbered binary-encoding format for XML data which is easy to implement and for which an open-source C language reference implementation is freely available as an OGC Best Practices Paper (http://www.opengis.org/techno/discussions/03-002r8.pdf).

Processing and exchanging GML data in a more efficient way is certainly the most immediate motivation for encoding GML data in BXML. One of BXML's important features is that it allows lists of floating-point geospatial coordinate values to be directly represented as raw-binary arrays. This provides both compactness and improved processing throughput since converting coordinate values to and from a text representation is time-consuming.
BXML provides "markup-for-markup-compatible" binary encoding of XML data. That is, the encoding is "lossless". An XML file that has been encoded and decoded is the same as the original unencoded file. BXML is remarkably efficient: compression ratios of 5:1 to 10:1 are common, and some files compress at a ratio of 25:1. BXML works with GML and with any other XML data.



###