Report of existing GIS standards and software – Deliverable 3.6.1

Synthesys NA-D 3.6
Javier de la Torre
Museo Nacional de Ciencias Naturales
CSIC

---

1 Introduction:

The goal of the NA-D 3.6 Synthesys task is to facilitate the geographical analysis of the networked specimen information.
This information is generally consumed through on-line applications, that is, through web portals were the user filters the available data in the network and downloads what he needs. If the user wants to do some geographical analysis of this data, normally, he will have to transform the data into some format that his favorite GIS application can read and then start working on it. This process of finding the information and adapting it to their working formats can be very long and tedious.

We aim at facilitating this process by providing tools and services that allow some basic analysis already on the web, at the portal level, and finally provide the data in a way they can directly consume in their desktop applications.
The movement to allow integration of different geographical information resources through Internet is often called *geospatial web* and has gained big momentum in the recent years. Therefore the software available, commercial and non-commercial, has increased a lot, making it difficult to choose between the different available solutions.

In the first phase of the project we reviewed existing standards for GIS data and available software for them. Because of the short duration of the project and the plethora of available software on the domain, it was decided to avoid the development of new products and preferably look for available software implementing those standards. The first part of this report is the result of this review.

In the second part we take a look at the different projects related to GIS in the GBIF community. It is our intend to avoid duplication of efforts in the community and therefore we would like to cooperate with other existing projects dealing with the same goals.

2 GIS standards:

The amount of different formats and standards available in the GIS community is very big, so we only have concentrated on those related to geospatial content and services on the web.

We will make special emphasis on describing the standards created by the **Open Geospatial Consortium** (OGC) because of the level of acceptance in the community and software implementing them.

We will also take a look at the other services, from companies like Yahoo or Google, that are providing geospatial content using their own formats and interfaces. These proprietary solutions have won a lot of importance due to the massive use of their services. One reason for that is the simple approach they have, in comparison with OGC standards, allowing developers to easily make use of them.
Therefore we have divided this chapter into OGC standards and Non-OGC standards.

2.1 OGC Standards

*"The **Open Geospatial Consortium**, Inc. (OGC) [1] is a non-profit, international, voluntary consensus standards organization that is leading the development of standards for geospatial and location based services*".
The OGC emerged to define standards to allow geoprocessing systems to communicate on the Internet through a set of open interfaces. Since 1994, the OGC has grown from 20 member organizations to over 250 from all over the world in commercial, academic, nonprofit, and government sectors.
All standards that the OGC adopts are freely available through their web site.

Among the different standards the OGC is developing, the following ones are considered interesting for our purpose:

Web Map Service (WMS) [2]: The way it generally works is that a client sends an HTTP GET request to a WMS server including a number of standard parameters. The server then returns an image based on those parameters. The parameters include such things as the geographic extent of the requested image, the image format you want, the projection, size and others.
There are two required request types that any WMS server must support: *GetMap* and *GetCapabilities*. *GetMap* returns a map image and *GetCapabilities* returns an XML document that contains information about the server and the data that is available.

Web Feature Service (WFS) [3]: With WFS the response you get from a request is an XML document in which each feature is represented with its geographic co-ordinates and its attributes. For example a request for a layer containing specimen locations with properties for them, like the scientific name, the specimen id, the collector and any other information that is associated with it. Line and polygon features are made up of multiple coordinate pairs but would otherwise appear similar in the XML document. These XML documents are, more specifically, made in **Geographical Markup Language (GML) [4]**.

Catalog Services for the Web (CSW) [5]: The OGC Catalog Service 2.0 specification defines a common interface that enables diverse but conformant applications to perform discovery, browse and query operations against distributed and potentially heterogeneous catalog servers. Most commonly a catalog service is used to provide access to metadata about geospatial data and/or geospatial services by exposing an XML/HTTP interface.

Web Coordinate Transformation Service (WCTS) [6]: Allows web-based transformation of geographic coordinates from one coordinate reference system into another. Useful in our community due to the differences among data providers.

2.2 Non OGC Standards

Despite the success of the OGC in bringing together the GIS community, there are some other standards (formats or services) following different approaches. Because of the extended use of them it is important to take them into consideration too.

ESRI Shapefile [7]: A shapefile stores a nontopological geometry and attribute information for the spatial features in a data set. The Geometry for a feature is stored as a shape comprising a set of vector coordinates.
The ESRI shapefile is a proprietary format from a commercial vendor, ESRI. The format specifications are publicly open so other software can also generate and read them.
This format is important because of the wide spread of GIS tools from ESRI, especially ArcGIS. Lots of users use this format to work with their geospatial data.

Google Maps/Google Earth KML [8,9]: More than a standard, Google Maps is a service for displaying maps. It is not using OGC standards at the moment, but the ease of use, the quality of the service and the prize, it is free, made it so popular that now it is being treated as a kind of *standard service*.
It has a lot of popularity among web programmers. They can incorporate the map service into their own web pages and integrate it with their own data. Thus creating rich geospatial information systems becomes pretty easy and fast to do.
Google maps can be used in several ways, but the most common one is through a Javascript API where data is fed in using simple XML.
Because the map is provided by Google with their enormous bandwidth, the quality of the service is very high, making the final user experience very satisfying.
On the other hand Google Earth is a Windows (being ported to Mac and Linux at the moment) program that shows the same imaginary data as Google maps, but that adds new navigation functionalities. Data can be mapped into Google Maps using a proprietary format called Keyhole Markup Language (KML), that can be compared to GML but is much simpler.

3 GIS software

Due to the lack of money for buying licenses and the possibility to distribute our results to other projects, one of the biggest considerations is the use of Open Source software. Therefore we have only analysed software packages with a certified OSI Open Source License [10].

The software is divided into different categories according to what it is used for. At the end of every category there is a conclusion paragraph discussing what is considered the best option for this specific task.

3.1 OGC Spatial servers

There is a set of available software that helps to create OGC services. This software is normally installed in a web server and is configured to publish your data trough OGC standards. More specifically for our purpose we consider the following standards necessary: **WMS, WFS, CSW, WCTS**.

- **deegree [11]:** free software initiative founded by the GIS and Remote Sensing unit of the Department of Geography, University of Bonn, and lat/lon. It is the reference implementation for the WMS standard. With more than one single software package, deegree is a set of building blocks to construct OGC

services. The software is right now under a big rebuild and a new version 2.0 is expected to be released before the end of the year.
    - **OGC Standards supported:** WMS, WFS, WCS, WCAS, WFS-G, WTS, WCTS, CSW.
    - **Read & Write interfaces:** ORACLE Spatial, PostGres/PostGIS, MySQL, other JDBC-enabled databases, ESRI Shapefiles, several raster data formats (JPEG, GIF, PNG, (Geo)TIFF, PNM und BMP).
    - **Architecture:** Java-servlet

- **Geoserver [12]:** again a free software initiative. It has a good community support behind and some projects funding its further development. It is the reference implementation for the WFS standard. The installation is easy and the documentation complete. Right now it only supports one feature per table and the mapping of complex schemas is not complete. There are people working on solving those limitations though.
    - **OGC Standards supported:** WMS, WFS.
    - **Read & Write interfaces:** PostGIS, ESRI Shapefile, ArcSDE and Oracle, VPF, MySQL, MapInfo, KML…
    - **Architecture:** Java-servlet

- **Mapserver [13]:** originally developed at the University of Minnesota (UMN) through the NASA-sponsored ForNet project, a cooperative effort with the Minnesota Department of Natural Resources. The software has grown and is maintained by an increasing number of developers (nearing 20) from around the world and is supported by a diverse group of organisations funding enhancements.
  The project started before the creation of these OGC standards and was later adapted to support them up to a certain degree. Right now the WFS service is read only and transactions are not supported. A bigger limitation is the lack of support of filters in feature attributes and POST operations.
    - **OGC Standards supported:** WMS, non-transactional WFS, WCS.
    - **Read & Write interfaces:** ESRI shapefiles, PostGIS, ESRI ArcSDE, TIFF/GeoTIFF, EPPL7
    - **Architecture:** CGI implemented in C, scripts in different scripting languages.

Conclusion and proposed software:

Considering the needs of the project it seems clear that MapServer will no be suitable. Choosing between GeoServer and deegree can be difficult. Both look stable and have been used in a lot of different projects. Deegree looks more complete than GeoServer according to the standards supported and the complex mappings that are possible in its configuration. On the other hand, GeoServer seems easier to deploy because of the very good documentation available for it.
The communities from both projects are active and look healthy. We have been posting emails in both mailing lists and there have always been fast and good responses.

We visited the lat/lon company in Bonn and talked about deegree. Our perception is that the deegree software is in good health and that there are new versions coming in

the next months. Especially there will be a major release, 2.0, before the end of the year, and they are working very hard on it to have it ready.

The *Geographisches Institut und Technologiezentrum GIS der Universitaet Bonn* is participating in a project with the German GBIF node related to Georeferencing of specimen data. They will use the deegree gazetteer service.

Considering the possibility of cooperation with projects related to GBIF in Germany we consider deegree a better choice. Using the same technology can make interoperability and integration of their services with ours easier.

In any case we do not discard the use of Geoserver if we encounter problems using deegree or if it performs better. It will not be difficult to switch from one to the other.

3.2 Spatial databases

Specimen data from data providers will have to be stored in an Spatial database where spatial queries can be performed. Typical spatial queries could be: give me all specimens of a certain taxon that had been gathered in this area.

- **PostGIS** [14]: adds support for geographic objects to the PostgreSQL object-relational database. In effect, PostGIS "spatially enables" the PostgreSQL server, allowing it to be used as a backend spatial database for geographic information systems (GIS). PostGIS has been developed by Refractions Research as a research project in open source spatial database technology.

- **MySQL** [15]: MySQL 4.1 introduced limited spatial functionality in MySQL which now is also available in version 5. The documentation available for these extensions can be found at the MySQL website.

Conclusion and proposed software:

MySQL spatial functionality is still very limited. It only works in MyISAM table type, that means no support for transactions. Also MySQL spatial indexes are not null-safe.

In general the MySQL spatial capabilities look as if they are still in an early development phase.
On the other hand PostGIS had been used for several years now and the software looks more complete. Also the support from tools like Geoserver and deegree are better for PostGIS.

It is still not clear how we will populate this spatial database. Possibilities are that we will have to use MySQL dumps from GBIF, therefore if we want to use PostGIS we will have to find or create tools to transform one into another. Because we will have to transform the data from simple coordinates fields in MySQL to geometry fields in PostGIS, it is clear that there is no direct advantage of using MySQL over PostGIS. Therefore we suggest to use PostGIS in our projects because of the better support in general which is available for it.

4 Related Projects

Since the beginning of the project we noticed that there are several other projects related to GBIF that share the goal of this project. Therefore we tried to establish contacts with them in order to avoid duplication of efforts and when possible get to common solutions. These contacts in most cases were done trough email and in the case of GBIF Germany we organized a meeting with the *Geographisches Institut und Technologiezentrum GIS der Universitaet Bonn* at the lat/lon company offices.

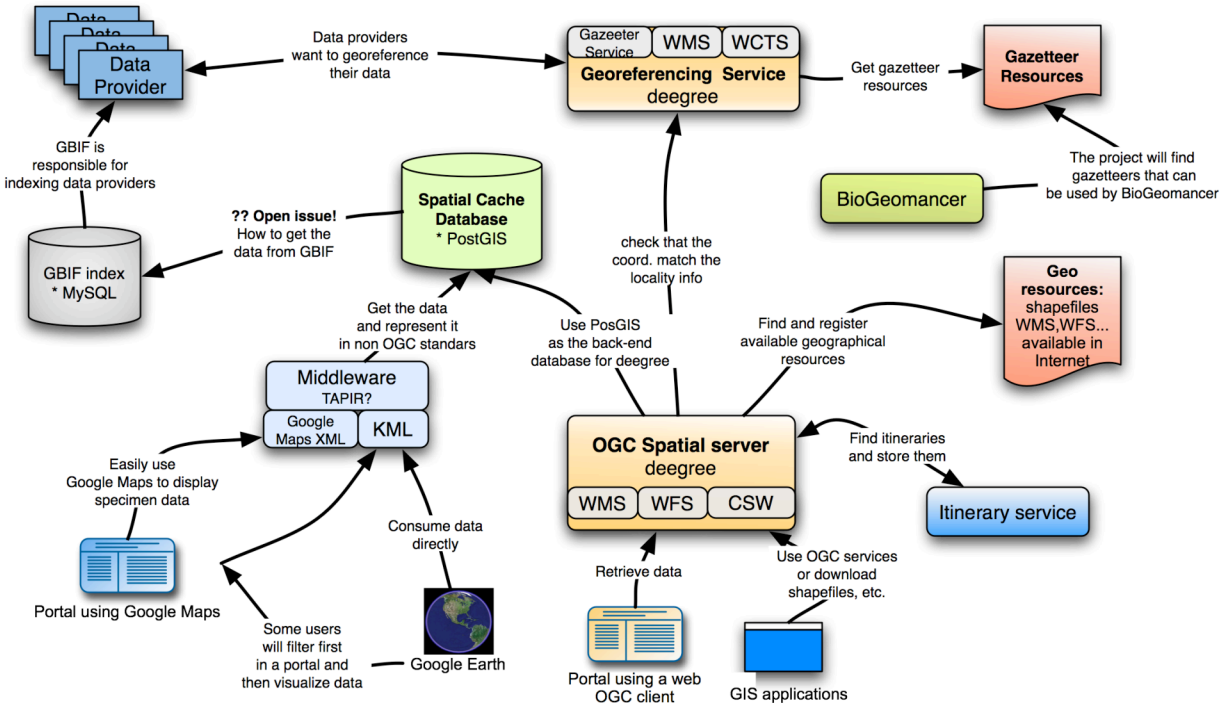
GBIF Germany [16]: They are running a project for "*Development of Web-based services for georeferenciation and map visualization of collection data*." To do that they will implement a gazetteer service and an interface where data providers can work to georeference their collection data. To do so they will use the tools available from deegree.

We had a meeting with them in October at the lat/lon offices in Bonn. It was very constructive because of their expertise on the deegree software and in OGC standards in general.
We would like to collaborate with them in setting up the needed GIS infrastructure and also use the gazetteer services that they will set up.

BioGeomancer [17]: From their web page, "*The BioGeomancer (BG) Project is a worldwide collaboration of natural history and geospatial data experts. The primary goal of the project is to maximize the quality and quantity of biodiversity data that can be mapped in support of scientific research, planning, conservation, and management. The project promotes discussion, manages geospatial data and data standards, and develops software tools in support of this mission*". With them we would like to cooperate in using the same GML application schemas to describe specimen data and, where possible, to make our services interoperate.

5 Conclusion

After this three months we have already a clear idea of what we need to implement in order to facilitate the geographical analysis of the networked specimen information. We have created a document with an architecture strategy that can found at http://www.biogeografia.com/synthesys/downloads/files/GeoSpatialProjectsForGBIF.ppt. We include here the diagram with the different software pieces we will set up in the next project phases.



We also set up a web page for discussion and where all this information is available http://www.biogeografia.com/synthesys/

As it is written in the diagram that we will use the **deegree** software to create our OGC web service infrastructure and PostGIS to run a cache database with the specimen data collected from the data providers.

Regarding the visualization of the data we foresee two different end users for the infrastructure that we are creating:

- **GIS users that wants to do analysis with the data:** These users have GIS knowledge and have their own tools to work with geographical data. For these users we will build the OGC services that they can consume trough web clients or directly trough their own GIS applications. Additionally we will configure the clients to be able to export data as shapefiles to make it easier for them to use the data.

- **Non GIS users:** Most of the users accessing GBIF data do not want to perform GIS analysis with it, but rather see "dots in maps" of specimen occurrence. We consider that this users will probably have a more satisfying experience using services like Google Maps [18] or Google Earth [19]. For them we want to make the data available through this non-OGC services.

We hope with this double strategy, OGC and non-OGC standards, to better meet the needs of the end users of specimen data.

We have already started playing with the available software described on this document and the results are promising. The GIS open source community is mature

and the software already stable. Together with services like Google maps, OGC standards will allow us to publish our georeferenced specimen data in a more practical way and meet the needs for data consumers.

6 Glossary of terms:

Portal: In this document we use the term portal to indicate a web site where specimen data can be retrieved. There are several portals with different focus like Geographical or taxonomic: a portal for Spanish data providers or a portal for Botanist in Europe.

GIS: Geographic Information System. This is normally a piece of software that represent maps and can do interactive queries, analyzed the spatial information, etc.

Feature and properties: In geographic information systems, a feature comprises an entity with a geographic location, typically describable by, for example, a set of coordinates. Properties of this feature is information attached to it, for example, the name of the feature. In our community a feature can represent an specimen that was gathered somewhere. The properties of this feature could be the scientific name, the collector, etc.

Layer: In GIS systems information is normally organized in different layers that when overlapped become a map. All the features in a layer have a common theme. When analysing data users work by overlapping and intersecting layers to get maps, or other layers.

XML: eXtensible Markup Language . Is a general-purpose markup language for creating special-purpose markup languages. It is primary prupose is to facilitate sharing of data across different systems.

GML: A special-purpose markup language based on XML to express geographical features and its relations. It is an standard from the OGC and is being used in different software to implement, mainly, the WFS standard.

KML: Keyhole Markup language. Another language to describe geographical features. It was created by a company called Keyhole, now bought by Google, and it is their representation format for their Google Earth software. It is comparable to GML, but simpler.

API: Application Programming Interface. A set of definitions of the ways one piece of software communicates with another. It can help to reduce the complexity of using a systems by defining simpler interfaces to interact with. Can be a method of abstraction.

Spatial index: When creating a database with spatial columns on it, it becomes necessary to create spatial indexes to speed up queries on it. They are comparable to normal database indexes.

Catalog server: a catalog server in the OGC argot is a server that helps users to discover geographical resources. Can be considered as registry where datasources are registered and where users can do searches and ultimately get to the data.

7 References:

1. **Open Geospatial Consortium**, Inc. (OGC): http://www.opengeospatial.org/
2. **Web Map Service (WMS):** http://www.opengis.org/docs/01-068r2.pdf
3. **Web Feature Service (WFS):** http://www.opengis.org/docs/02-058.pdf
4. **Geography Markup Language (GML):** http://www.opengis.org/docs/02-023r4.pdf
5. **OGC Catalog Service 2.0 (CSW):** ddd
6. **Web Coordinate Transformation Service (WCTS): dfdd**
7. **ESRI Shapefile:** http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf
8. **Google Maps:** http://maps.google.com
9. **Google Earth:** http://earth.google.com
10. **Licenses approved by OSI:** http://www.opensource.org/licenses/index.php
11. **degree:** http://deegree.sourceforge.net/
12. **GeoServer**: http://docs.codehaus.org/display/GEOS/Home
13. **MapServer:** http://mapserver.gis.umn.edu/
14. **PostGIS:** http://postgis.refractions.net/
15. **MySQL:** http://dev.mysql.com/tech-resources/articles/4.1/gis-with-mysql.html
16. **GBIF Germany:** http://www.gbif.de/weitere_projekte
17. **BioGeomancer:** http://www.biogeomancer.org
18. **Google Maps:** http://maps.google.com
19. **Google Earth:** http://earth.google.com