Integrative Environmental Queries Using Geospatial Web Services

David Pullar, Jack Fan Zhang, Jane Hunter, Xiaofang Zhou

The University of Queensland
Brisbane, Australia 4072
{d.pullar@uq.edu.au}

Key words: geospatial web services, biological collections, distribution models

Abstract: The paper explores the use of geospatial web services for querying and analytical functions on distributed heterogeneous biological databases. A case study and web service implementation was tested on biological collections in Australia. Open geospatial specifications for web services were adequate for the test implementation, although some performance issues on XML-encoded data were identified. More significantly, we highlighted the need for domain specific formats to improve their precision and content to support queries that need to determine value-based identity. We also developed analytical functions well suited to geoprocessing web services that utilised statistical summaries and localised searches versus traditional algorithms that require access to the full dataset.

1. Introduction

Using geospatial information technology for large scale sharing of biological records is a worthy and challenging problem. Biological records are collected by many groups in society; including wildlife societies, museums, and environmental agencies. The scientific community would benefit from access to the huge worth of data collected by these diverse groups for research on evolutionary ecology and implications of global biodiversity. The geospatial IT community recognize that advances in scientific understanding hinge on scientists having access to information in a flexible way and being able to undertake complex analysis [12]. There is also an emerging trend with the democratization of information where web technology has enabled new opportunities for community involvement in data collection and maintenance. Agencies and museums are either establishing or expanding their capability for the public to enter biological records. This creates different challenges for authentication and validation of records as information is entered from a broader base. While improving access to biological collections we need to support better validation procedures.

These challenges form the backdrop for an e-Research project (support by an e-Science program in Australia) to investigate the next generation of geospatial technology for distributed access to integrated data repositories and enhancing capabilities to manage data. The project undertook a case study initially based upon an enterprise database called WildNet for the wildlife information in Queensland Australia, but this was expanded to include other museum and community information sources. Access to replica databases was implemented through web services using XML-encoded queries to discover and query data resources. The project undertook research and development of a prototype system to: i) assess new open GIS web technologies and protocols for integrated views of biological data and improving mechanisms for resource sharing, ii) integration of OGC and Grid technologies for geospatial data sharing [18], iii) use semantic web technologies to integrate biological records with physical environmental databases and taxonomic classifications for richer user queries [11]. This paper addresses the first task as a research problem. As the aim was to demonstrate the usefulness of new technology we focussed on two significant and highly relevant issues.

The first issue addresses support for queries to heterogeneous and distributed information servers. Data standards for representing biological information from the Global Biodiversity Information Facility (GBIF) [7] are used as the basis for a global schema to integrate queries. A major issue is that different organisations often have duplicate entries for the same real world sighting in different databases. Therefore this first issue involved addressing both technical and conceptual issues. The next section describes the nature of the data in this study and what GBIF is. Technical issues on integration are discussed based upon our experiences and an evaluation of existing data protocols like GBIF for query integration.

The second issue addresses support for maintaining biological databases where a significant problem is validating data as it is entered. Error-checking of biological data needs to go well beyond syntactical checks to deal with mistaken information. Error detection procedures may be used interpret the context for newly recorded species sightings to identify mistaken georeferences and detecting dubious records for further inspection and assessment [9]. We focused on analytical functions implemented as web services to check new entries against a wide set of integrated sources through web services. A geoprocessing service was implemented and utilised for error detection and for data interpretation. The service computes a local distribution map for a species that combines existing generalised predictions of species distributions with local sighting data. The novelty of the approach is that most species distribution models [10] are based upon algorithms that process the full set of data (species sightings and other related environmental data determinants) where we implemented a Bayesian updating classifier that combines a prior global statistical distribution with localised data. Using algorithms that analyse localised subsets of data is important in dealing with large heterogeneous data collections as it is impractical to access the full collection or even large subsets of it through web services.

The paper is organised around exploring these two issues. The next section describes the nature of biological information collections using an Australian example for herbarium and museum databases. We show how these single enterprise systems may mapped to an integrated resource for the wider scientific community. We also deal with the issue of record identification between databases and evaluate the GBIF data formats as a global schema for addressing this issue. A subsequent section describes a query geoprocessing service for providing species distribution maps. These may be utilised for visualising species distributions and for error detection in entering new occurrence data. The conclusion summarises our findings and discuss wider implications of our research.

2. Integrated access to biological collections with web services and the GBIF protocols

The project collaborated with a state Environmental Protection Agency (EPA) in Queensland Australia to obtain user requirements and data for biological collections. A prototype system was developed using Open GIS standards for web feature services and catalogue services [13] to assess integrated access to the EPA's collection and data sources. We describe these collections, the prototype implementation and issues that arose in the implementation.

2.1 Data Collections

The project identified three significant collections. Only limited subsets from these data collections were tested in the prototype, but it was sufficient to appreciate performance issues for queries. Table 1 lists the data description for datasets given from a data request (note that their internal data schemas may be different to the public exchange records). The collections are:

i) WildNet documents scientific information for the state's (Queensland , Australia) animals, including rare and threatened species. As of the writing of this paper it contains 3,700,000 records listed and managed by WildNet.

ii) BioMaps is part of the Australian Museum and aims to provide a broad sample of life and biological information for knowledge and biodiversity assessment. It houses collections from the Australian museum and many other collections from state museums. It is the most significant animal collection in the Southern Hemisphere with about five million biological specimens.

iii) Birdata is a web-based atlas documenting data and surveys on birds. It is operated by Birds Australia, a conservation society, and lists 7000+ community and scientist participants involved in contributing data records.

Table 1. Data descriptions for records in the three collections

WildNet species record description						
Common name	*Scientific name*	*Conservation status codes*	*Location (name, coordinates ϕ,λ, accuracy.*	*Time (start-end)*	*Record vetting status*	
BioMaps species record description						
	Scientific name	*Institution*	*Location ϕ,λ,*	*Date*		
Birdata species record description						
Common name	*Scientific name*	*Count (point and breeding)*	*Location (name, coordinates ϕ,λ, accuracy.*	*Time (start-end)*	*Search area extent*	*Observer ID*
GBIF species record description						
	Scientific name	*Institution*	*Location (name, coordinates ϕ,λ)*	*Date*		

2.2 Global Biodiversity Information Facility

The Global Biodiversity Information Facility (GBIF) is a web portal to facilitate the entry and dissemination of primary biodiversity data at a global scale. It integrates tens of millions of records of primary biodiversity data from hundreds of databases worldwide in museums, botanical gardens, and observation networks such as those of bird watchers [17]. The GBIF data portal disseminates sighting data in a simple format that lists the species, a location and date.

Although there are no formal standards for biological records, GBIF is a widely accepted format. The sighting data has a simple format, it basically records an identified species, a location and a date. As can be seen from Table 1; GBIF is a generalised subset of the three focus collections. We adopted the simple GBIF format as the exchange format for testing integrated access to our web service.

2.3 A prototype web service implementation using Open GIS Standards

The Open Geospatial Consortium (OGC) [13] defines open interface specifications for the implementation of web mapping services. The two main interface specifications of interest to this project are the Catalogue Discovery Service [14] for cataloguing access services, and the Web Feature Service (WFS) [15] for geospatial data access. We developed our own catalogue service that managed collections. It provided metadata on the contents of collections including taxa and geographical extent. The idea of using a catalogue is to allow listing of multiple data collections on the server and to support queries that have no domain knowledge of sources. Collection providers register that they have a web service. Queries on the server evaluate requests against the collections in the catalogue to find valid WFS's to query. The web server executes the WFS requests to the valid services, compiles results and sends back the requested information to the client. This process of services communicating with other services to complete requests is called service chaining [1]. Service chaining may incur performance drawbacks and difficult query processing issues, but independent access queries can be executed without the need for special tracking and error-reporting functions. A conceptual design for the web services is shown in Figure 1. None of the targeted organisations with biological collections support dissemination of data using WFS at the time of writing. Therefore we obtained sample collections from the respective organisations and created separate WFS servers.

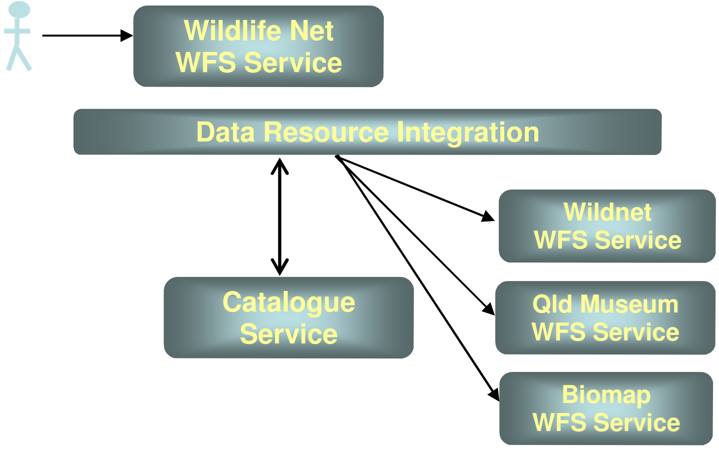

Figure 1: Conceptual diagram of test implementation components using web services

The prototype implementation was relatively simple. We implemented a Catalogue Discovery Service and a related WFS in Java\Eclipse IDE on a dedicated server. The chained services to collections were implemented with the open source server, Geoserver [8]. For convenience these services were on the same server machine, but could be more efficiently setup on other machines. All communication used XML-encoded strings and spatial data was sent as XML-based GML.

A number of performance issues were revealed with the test implementation. All XML data was transferred and handled as XML text so transfer of large data volumes was slow. Transfer rates could be improved with compression but parsing text data was still inefficient. Future support for a binary format for GML will remove some of these drawbacks. Other inefficiencies in query processing also arose. It was necessary to transform queries on the separate collections from their native schema's to the GBIF format, to pool these queries together as a response and send it to the client. Apart from schema mapping, very little data interpretation was required for queries. Being able to translate queries on the fly and stream them back to a request would have improved query performance greatly. The overall design for open geospatial web services is sound, but greater support for communication protocols based upon binary streaming is needed to give reasonable query response times.

2.4 Issues in handling heterogeneous collections

Beyond these technical issues the test implementation faced a semantic problem with the nature of biological collections, namely discerning the unique identity of records. The records represent an animal sighting and hence are very transient in nature. Even individual biological databases have difficulties with similar records on biological sighting. In some cases a sighting is recorded by different observers, or in other cases duplicate records find their way into databases [4]. In a heterogeneous environment these problems are compounded by a much greater likelihood of duplicate records in different databases. In some cases large subsets of the data are duplicated and if analysed without detection could cause serious misinterpretation for inventories and biogeographical studies [5]. To address this problem we evaluated GBIF format for establishing the uniqueness of records. Tests for equality may be based upon record content for the species name, location and data. But there is some ambiguity with each of these fields. Although the species name has a recognised scientific usage, it is still subject to changes. Over time as taxonomists study species they make more refined classifications and change the scientific name of the species according to different characteristics or functions of the species. If records are duplicated in different databases then it is possible one organisation as part of their quality control; updates the records to the new scientific name while the other database does not. There is greater ambiguity with regard to the date and location of sighting. All the target collections recorded the time and duration of a sighting observation. Several entries of separate occurrence data could be entered for the same day, generalising the time fields to a date loses vital data. The accuracy with which locations are recorded also varies. Even modern GPS georeferences are subject to operator error producing incorrect positions. Different organisations may have different protocols for entering the precision for GPS positions and the type of coordinates entered may be in different georeferencing systems. Again the situation arises that one organisation cleanses the data to conform to their own validation standards, and this exacerbates simple comparisons of equality between records in differing databases. Overlaid on these difficulties is the situation of mistakenly entered data or exchanged data.

Even with the simple GBIF format it is important to have codes of practice for entering and maintaining data. With greater acceptance of web based access to collections, the practice of copy records from other organisation should be reduced. GBIF is very active in providing guidelines and raising awareness of validation problems for sighting data [4][5]. The current simple format is too ambiguous and very little can be done to reliably assess record equality between databases. The amount of duplication was hard to determine and it does vary based upon regions, but we found in some areas where intensive surveys had been conducted that there was significant duplication in databases. For instance, surveys by Birds Australia were also entered into WildNet. It was possible to identify this duplication in the WildNet database (although ambiguities still existed), but it was impossible to do this using the simplified GBIF format. An additional field that is included in the enterprise Wildnet database, but not included in the exchange format is the observer’s identity. Birds Australia maintain a register for their ‘atlasser’ community and this observer ID is included with the exchanged record. If this approach was instituted at a National level then this would be extremely useful in assessing record equality.

In summary our implementation found that building online access to biological collections was technically feasible using open geospatial web services. There were some technical concerns with performance that could be overcome with binary streaming of geospatial data as part of the XML query encoding, but there were more significant semantic issues with data integration. A global data format like GBIF is vital, but it should be more detailed with actual times (as opposed to calendar dates) and GPS positions to a prescribed precision. Also including an observer ID is crucial.

3. Geoprocessing services for integrated collections

One of the organisations within this study, namely Birds Australia, actively used a wide community of observers to enter data. The trend is for greater utilisation of enthusiastic people within the community to get involved with providing incidental sighting data and doing surveys. However, the benefits of obtaining richer datasets should not be compromised by concerns for the quality of databases. Following the principles of database design, the most effective way to deal with data validation is at data entry. Data validation and quality control could be improved by keeping information about the observers and their reliability with regard to their biological expertise and observer accuracy. Another useful procedure would be to use an automated check on the consistency of a sighting with known species distributions. Species distributions models generate probabilities for occurrence of species at locations and this information could be used to validate new sightings. For instance you would expect a zero percent likelihood in finding a terrestrial animal in a marine environment. This is an obvious case, but species distributions models can generate a probability of occurrence as a means to authenticate new sightings. This section describes the development of a species distribution geoprocessing web services for this purpose.

3.1 Species Distribution Models

Understanding species distributions is at the heart of biological sciences. Whether it is for natural evolution or biogeographical studies it is important to explain why a species occurs in one location as opposed to another at different times. Species distribution models attempt to quantify the occurrence of species in line with ecological theories and the occurrence observations. Over the past twenty years the development of computer models for species distributions has been an active area of research [10]. The essence of models is to relate patterns of climate\environmental conditions and biological interactions with species occurrence. Distributions may be summarised by statistical probability densities; typically a binomial model is used for presence/absence occurrence or a Gaussian model if population is known [10]. To make sure they cover the full range of environmental conditions in computer models, it is desirable to include the full extent for sampled species occurrence. The models are good at distilling general ecological rules to explain distributions over large spaces, but at a localised scale their predictions should be seen a potential distribution versus an expected count of species. That is, one would expect to find that species somewhere within the predicted environmental niche but not necessarily at a particular location. Any expectation would be highly influenced by observed presences. Hence we have two separate pieces of information at particular location: the distribution potential and any local evidence of sightings. In the next section we describe how we combine this information for generating a refined distribution map.

3.2. Geoprocessing service for distribution map

Species distribution models require access to all sighting data to predict potential distribution. These models would normally be run by experts and output provided as a map coverage. A point query service can be developed to extract from a coverage a probability that a species utilizes that specific location in some way. BioMaps for instance has a demonstration project for a biodiversity analysis tool that maps species richness and taxonomic diversity via online queries. These models return broad coverage maps for all of Australia. A simple way to combine a generalised coverage with local sighting data is using a Naïve Bayesian classifier. The generalised model is treated as a prior and local sighting data treated as evidence in a classical Bayes's analysis. The advantage of building a query processing service based on this approach is that the prediction is made based upon all the available sighting data in the local vicinity of the query and it runs efficiently.

The geoprocessing service returns the potential of a species occurring at a location. The service may be run when a new sighting is submitted as a step, along with other validation checks, to validate the record before entering it in the collection. Acceptance levels may be set based upon probability to accept, reject or flag the entry for further inspection. The probability may be combined with other information, such as observer reliability, to define a quality assessment of the record.

The geoprocessing service was implemented as a web service. Given a XML-encoded GBIF record the service executed a server-side process that returned a probability value. In the Bayesian analysis we return a presence/absence probability by combining information from a global distribution coverage and a query of sighting data within the local search window. The Naïve Bayesian classifier computes the posterior probability of species *S* present at location *x* as:

$$\text{posterior probability of } x_S = \text{prior probability of } x_S \times \text{likelihood of } x_S$$

where the likelihood of x_S is estimated as the proportional area with the species *S* present in the vicinity of *x*.

If no records are found within a reasonable range then the prior probability, or coverage map query is used, otherwise if it finds other sighting data the likelihood is estimated and an updated probability is computed. Note that the binomial probability expresses a density with an assumed range of influence. The likelihood is computed as the proportion of the area of presence for the species in that vicinity to the total area of the vicinity [2]. Some interpretation of what is meant by a species vicinity is needed. This could be based upon the home range of an animal, but home range values may be unknown or vary significantly depending upon the region. We do not resolve this, but simply use a user-supplied geographical extent for the vicinity in the query. A client side viewer zooms in and out with the map window used as the geographical extent. In the absence of any better knowledge we leave it to the user to interactively explore the data and set this extent.

4. Conclusion

The project describes research problems and a technical issues for e-Research on query and analysis functions on biological collections. We used open geospatial standards for implementing web services. Although implementation of some of these standards was not widely available, i.e. the OGC Catalogue Discovery Service, we found the specified interface was adequate for developing our own web service catalogue and query capability. A test implementation was run to discover candidate web services with relevant collections in response to a spatial query, to request data from these services, synthesise the results and return a response. The individual collections were accessed by webs services implemented using GeoServer, otherwise we did our own web service implementation. As a proof-of-concept the test implementation fulfilled requirement for discovery and integrated queries on distributed collections. There were some performance issues that should be resolved with adoption of binary streams in exchange protocols. More significant issues related to the interpretation of data. We discuss how even establishing the equality of records is problematic, and could seriously distort interpretations and analysis of the data. A richer and more precise standard for encoding biological records is needed to improve matters. This would also open opportunities to intelligently interpret data by constructing ontology's for biological sightings [3]. For instance, the original sighting data often included free text descriptions of the site. Structured queries are unable to make use of this information, but the terminology is sufficiently formal and scientific to develop inference rules as part of an ontology. A separate project explored these possibilities for interpreting taxonomic descriptions [11]

The research also investigated the development of spatial analytical functions as geoprocessing web services. The novelty of these algorithms is that they do not access the full dataset as this is a prohibitively costly for large distributed data sources. We developed a species distribution model that combined prior information on statistical density with local query data to give an updated probability for the presence of a species at a location. This information may be used for querying a site or for validation of a new sighting record. The procedure uses a Naïve Bayesian classifier which despite being very simple, sometimes outperforms more sophisticated classification methods. But the main benefit is that the classifier is well suited to efficient spatial analysis by combining prior statistical information with a local vicinity-based query. We are currently testing the results from this classifier against experimental data for Koala distributions.

References

1. Alameh, A.: Chaining geographic information Web services, IEEE Internet Computing 7(5) (2003) 22- 29
2. Aspinall, R.: An inductive modeling procedure based on Bayes' theorem for analysis of pattern in spatial data. Int. J. Geogr. Inf. Syst. 6/2 (1992) 105–121.
3. Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D., Patel-Schneider, P. Stein, L.A.: . OWL Web Ontology Language Reference (2004)
4. Chapman, A.D. 2005a. Principles of Data Quality. Report for the Global Biodiversity Information Facility 2004. Copenhagen: GBIF. http://www.gbif.org/prog/digit/data_quality/data_quality (Accessed Feb. 2007)
5. Chapman, A.D. 2005b. Uses of Primary Species-Occurrence Data. Report for the Global Biodiversity Information Facility 2004. Copenhagen: GBIF. http://www.gbif.org/prog/digit/data_quality/uses_of_data (Accessed 1 Feb. 2007)
6. EPA: Environmental Protection Agency, WildNet Database, Environmental Protection Agency, Brisbane, Australia (2006)
7. GBIF: What is GBIF? http://www.gbif.org/GBIF_org/what_is_gbif (Accessed Feb. 2007) (2003)
8. GeoServer: http://docs.codehaus.org/display/GEOS/Home (Accessed 1 Feb. 2007)
9. Graham, C.H., Ferrier, S., Huettman, F., Craig Moritz, C., Townsend Peterson, A.: New developments in museum-based informatics and applications in biodiversity analysis, Trends in Ecology and Evolution 19/9 (2004) 497-503
10. Guisan, A., Zimmermann, N.: Predictive habitat distribution models in ecology, Ecological Modelling 155 (2000) 147-186
11. Henderson, M., Khan, I., Hunter, J.: Semantic WildNET- An Ontology based Biogeographical System. (In Preparation)
12. Muntz R R, Barclay T, Dozier J, Faloutsos C, MacEachren A M, Martin J L, Pancake C M, and Satyanarayanan M 2003 IT Roadmap to a Geospatial Future: Report of the Committee on Intersections Between Geospatial Information and Information Technology. Washington, DC, National Academy of Sciences Press .
13. OGC: Open Geospatial Consortium, http://www.opengeospatial.org/ (Accessed 1 Feb. 2007)
14. OGC: Catalog Interface Implementation Specification (Version 1.0), Document 99-051s (1999)
15. OGC: Web Feature Service Implementation Specification (Version 1.1) Document OGC 04-094 (2004)
16. Paul, M., Ghosh, S.: An approach for service oriented discovery and retrieval of spatial data. International Conference on Software Engineering, ACM Press (2006) 88-94
17. Saarenmaa, H.: Sharing and Accessing Biodiversity Data Globally through GBIF. ESRI User Conference, San Diego (2005)
18. Yanfeng Shu, Y., Fan Zhang, J., Zhou, X. (2006) A Grid-enabled Architecture for Geospatial Data Sharing, IEEE Asia-Pacific Conference on Services Computing (APSCC'06)