A Geographic Annotation Service for Biodiversity Systems

Fabiana B. Gil[1,2]**, Nádia P. Kozievitch**[2]**, Ricardo da S. Torres**[2]

[1] CPqD Foundation
Rod. Campinas - Mogi-Mirim km 118.5 Campinas, SP, Brazil

[2]Institute of Computing – University of Campinas
Av. Albert Einstein, 1251 Campinas, SP, Brazil

`fbellette@gmail.com, nadiapk@ic.unicamp.br, rtorres@ic.unicamp.br`

Abstract. *Biodiversity studies are often based on the use of data associated with field observations. These data are usually associated with a geographic location. Most of existing biodiversity information systems provides support for storing and querying geographic data. Annotation services, in general, are not supported. This paper presents an annotation Web service to correlate biodiversity data and geographic information. We use superimposed information concepts for constructing a Web service for annotating vector geographic data. The Web service specification includes the definition of a generic API for handling annotations and the definition of a data model for storing them. The solution was validated through the implementation of a prototype for the biodiversity area considering a potential usage scenario.*

1. Introduction

The term biodiversity – or biological diversity – describes the richness and variety of biological organisms in a given habitat. Some known issues in biodiversity are irreversible loss of species, loss of environmental services (production of oxygen by plants, the hydrological balance, soil fertility, and climate balance), and biopiracy.

Computer Science can be a great allied of Biologists, providing them with tools to analyze and report findings on species and their behaviors. Some challenges for the Biodiversity Information Systems (BISs) are [Torres et al. 2006]: (i) handle large volumes of information, (ii) integrate information from different sources and formats (heterogeneity), (iii) manipulate data and images, (iv) manipulate geospatial information reference.

In this paper we address the forth challenge. Biodiversity studies are often based on the use of data associated with field observations, later matched with geographic locations. Most of existing biodiversity information systems provides support for storing and querying geographic data. Annotation services, in general, are not supported.

Annotation has been recognized as one of the most important services in digital library systems to foster the cooperation among users and the integration of heterogeneous information resources [Agosti and Ferro 2008]. In this paper, we describe a new geographic data annotation Web service that can be easily integrated with other applications. The specification and implementation of the proposed Web service relies on two main contributions: (a) the proposal of a data model based on superimposed information [Maier and Delcambre 1999a] to manage geographic annotations; and (b) the definition of a generic API to manipulate annotations.

The rest of this paper is organized as follows: section 2 briefly presents the related work; section 3 describes the proposed geographic data annotation service; section 4 describes the use of the proposed annotation service in a Biodiversity Information System; and finally section 5 presents conclusions and summarizes future work.

2. Related Work

2.1. SuperImposed Information

Superimposed information (SI) [Maier and Delcambre 1999a] is the new information (or new interpretation) associated with the existing information or base (as for example, an annotation, term or keyword, a comment about a base document). In general, SI is not structured, does not have a pre-defined schema, and allows different types of SI under the same base.

Superimposed information has been used in several applications, named *Superimposed applications (SAs)* [Maier and Delcambre 1999b]. They allow the manipulation and creation of superimposed information, typically to highlight, annotate, draw, select, organize, connect, or reuse the information. Examples include SLIMPad [Delcambre et al. 2001], and RIDPad, Schematics Browser, SIMPEL, and Mash-o-matic from [SI]. None of these SAs, however, support interactions with geographic data.

The research described in this paper differs from related research in the sense that it takes advantage of superimposed information concepts to provide mechanisms for geographic data annotation in biodiversity applications. Furthermore, the use of a Web service implementation ensures appropriate encapsulation of data and procedures, which allows reuse of the developed component in other Digital Library (DL) initiatives involving geographic data annotation.

There are some other DL initiatives for the biodiversity domain. One example concerns floristic digital libraries (FDLs) [Sanchez et al. 1999]. FDLs use an agent-based infrastructure to manage botanical information about taxonomic keys, distribution maps, illustrations, and treatments (morphological descriptions). Another example is the Taiwanese digital museum of butterflies [Zhu et al. 2000]. This digital library offers a set of XML-based modules to manage data on butterflies. Geographic data annotation services, however, are not supported in these systems.

More recent DL-oriented initiatives include SuperIDR [SI], Ecopod [Yu et al. 2006] and BDiG-PELD [Barros et al. 2008]. The *Superimposed Image Description and Retrieval Tool* (SuperIDR) [SI] is designed to work with Tablet PC. With SuperIDR, users can navigate through the collection by species or taxonomic organization. Users can perform queries using part of an image or a textual description and recover the original content. Both systems work with annotation, but they still do not support geographic annotation.

Ecopod [Yu et al. 2006] aimed to replace the paper field guides with a mobile computing platform for identifying plants and animals in the field. In this tool, geographic information is used as an evidence in the species identification process. For example, some species are observed more frequently in specific regions. Annotation facilities, however, are not supported. Barros et al. [Barros et al. 2008] have proposed a georefer-

enced digital library, named BDiG-PELD that integrates network of ecological sites using OAI-PMH protocol and the ODL (Open Digital Library) framework. Even though geo-referencing facilities are used to determine species location, annotation services are not provided.

There are also examples of applications which work with geographic data annotation[Macário et al. 2009]. Besides having several standards for services and implementations, these applications still have lack of support for many features, like annotation of objects that are result of a query.

2.2. Geographic Data

The term spatial data denotes any type of data that describe phenomena associated with some spatial dimension. Geo-referenced data belong to a particular class of spatial data, describing events, objects and phenomena on the globe associated with their location on the surface, at a certain moment or period of time [Câmara et al. 1996].

2.2.1. Data Models

Within the context of applications which deals with geographic information, the real world is often modeled under two complementary views: the field-based model and object-based model. The field-based model (usually implemented according to the structure called *raster*) deals with a continuous surface on which the geographic phenomena varies according to different distributions (like atmospheric pressure). Each layer corresponds to a different theme (like vegetation, soil) [Gomes Jr 2007].

The object-based model (usually implemented by using vector representation) represents the world as an area occupied by identifiable objects that have their own geometry and characteristics. In the object-based model, a geographic object is typically represented in the vector format, that is its geometry is described using points, lines, and polygons. Lines are formed by sequences of points and polygons (open or closed), by a sequence of lines [Gomes Jr 2007].

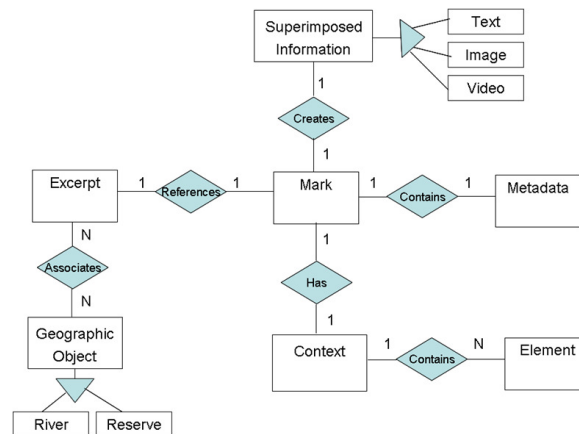The current version of the proposed annotation service is focused on vector data.

2.3. Spatial Operators

Spatial operators are used in our Web service to define which geographic objects (or parts of them) are being annotated. Several spatial operators have been proposed to process geographic objects based on their relationships. These relationships can be divided into three categories: metric, topological, and directional [Güting 1994]. Examples of topological relationships include *equals*, *disjoint*, *intersects*, *touches*, *crosses*, *overlaps*, and *contains*. Directional relationships include, for example, *above* and *north_of*. Metric relationships include *distance* and *area*.

Spatial operators are used in our annotation service to define regions which will be later associated with annotations.

3. Web Service for Annotating Geographic Data

This section describes the proposed Web service. First the data model is described. It includes annotation and geographic concepts, followed by the operations of the proposed

Figure 1. Proposed entity-relationship diagram.

annotation Web service.

3.1. Data Model

The data model is based on SI concepts, and the entity-relationship diagram diagram is presented in Figure 1. The entities are represented by rectangles; relationships are represented by diamonds; lines represent the connections between the entities and relationships; and the cardinality of the relationship is showed at the side of the diamonds.

The entities *Excerpt*, *Mark*, *SuperimposedInformation*, *Context*, and *Element* are inspired by the SPARCE model [SI].

When an annotation is inserted, a mark is created. A mark is used to link the superimposed information with the base information [Archer et al. 2008]. Different types of marks can be created when dealing with maps. Figure 2 illustrates four examples of marks associated with geographic data. The mark showed in Figure 2(a) refers to a rectangular region that includes several geographic objects represented as lines and points. In Figure 2(b), the mark includes parts of Brazilian states (i.e., parts of polygons). Figure 2(c) illustrates the definition of a mark associated with the result of applying a spatial operator (*intersection*) on two regions defined by the coverage of two antennas (i.e., intersection of two polygons). Finally, Figure 2(d) illustrates how to define a new mark associated with the intersection of the regions defined by two other marks.

An *Excerpt* represents the content of a marked region. In the case of vector data, the content may be a point, a line (or part of a line), a polygon (or part of a polygon), or sets of objects with different geometries. Figures 3 and 4 illustrate how excerpts are created, given defined marks.

Context entity defines a set of *elements* which provides contextual information associated with a mark. In the case of textual base document, context data may include, for example, information about font type, font size, or page number. For geographic data, context may involve information related to geographic coordinate system or visual properties (e.g., color, size, length) used for visualization purposes.

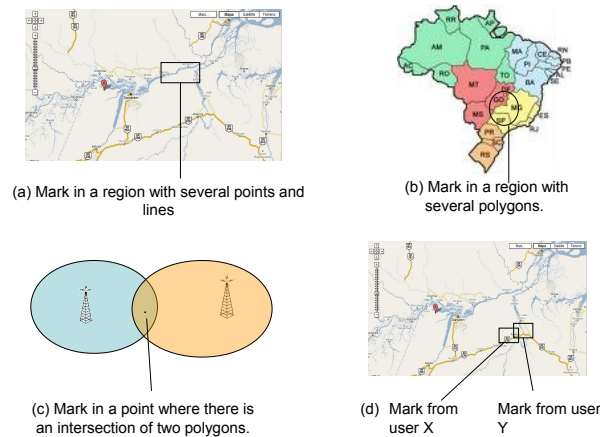*Metadata* attributes relies on *Dublin Core Metadata Initiative* elements (at

(a) Mark in a region with several points and lines

(b) Mark in a region with several polygons.

(c) Mark in a point where there is an intersection of two polygons.

(d) Mark from user X   Mark from user Y

Figure 2. Examples of marks.

http://dublincore.org). The metadata includes information about the annotation creator, language, annotation date, and rights held in and over the geographic data being annotated.

Superimposed Information refers to a new information associated with the information base (vector geographic data). In the current version of our annotation service, SI refers to textual annotations.

The spatial components of the data model (not shown in Figure 1 for clarity reasons) follow the Open Geospatial Consortium (OGC) model (available at http://www.opengeospatial.org/standards.). OGC defines the *Geometry* as the most generic entity. The geometry is associated with a Spatial Reference System, which specifies the coordinate system.

The basic representations of geometry are *point*, *curve*, *surface*, and *collection*. A *GeometryCollection* can be used to represent a collection of distinct geometries. *MultiPoint*, *MultiLineString*, *MultiPolygon*, *MultiCurve*, and *MultiSurface* are homogeneous geometry collections of *Point*, *LineString*, *Polygon*, *Curve*, and *Surface*, respectively. A *LineString* represents a sequence of connected points. A *Point* represents a simple location in the coordinate space. A *Curve* is a generalization of *LineString*. A *Line* is a specialization of *LineString* which has only two points. A *LinearRing* represents a *LineString* which is closed and complete. A *Polygon* is a close plane figure, defined by one exterior boundary and zero or more interior boundaries. A *Surface* is a generalization of a *Polygon*.

3.2. Web Service API

With the data model and annotation and geographic concepts presented in the previous section, we now present the interfaces for the insert, update, delete, and query operations.

A generic interface was first defined to manage annotations:

$< return >$ *operation* (**Text** Annotation, **Query** Query, **Metadata** Metadata, **Context** Context)

where *operation* can assume the insert (*InsertAnnotation*), update (*UpdateAnnotation*), delete (*DeleteAnnotation*), or query (*QueryAnnotation*) operations. The *Annotation* parameter refers to the annotation content. The *Query* parameter defines a query whose results include the set of geographic objects to which the target operation will be related. The *Metadata* parameter refers to attributes associated with the mark metadata. The *Context* parameter includes elements of the mark context.

Note that since several parameters are managed, the interface was later specialized for geo-queries (regions, excerpts, etc.) and non-spatial queries (metadata, annotation, etc.) and their combination. One remarkable aspect of the proposed API relies on the support of operations which are defined over geographic objects that are result of queries.

3.2.1. Annotation Insertion

The interfaces for annotation insertion are specializations of the generic interface, considering regions, list of objects and results of geographic queries. Figure 3 illustrates the insertion of an annotation associated with the region defined by the rectangular region (mark). As it can be observed, this annotation is linked to part of Street A, part of River X, and part of Street B (*Excerpt*). Table *IdGeographicObject* contains records related to geographic objects (e.g., Street A, Street B, and River X). Table *IdGeographicObject_has_Excerpt* contains records related to the relationship of geographic objects and excerpts, after inserting a mark.

- **boolean** *InsertAnnotation* (**Text** Annotation, **Geometry** Region, **Metadata** Metadata, **Context** Context)
  This operation inserts an annotation associated with a set of geographic objects within a given region.
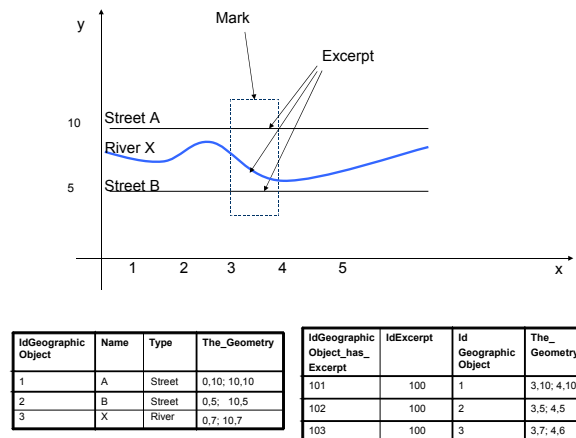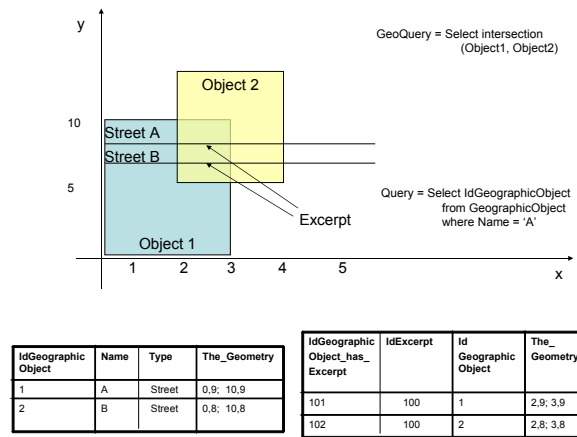

IdGeographic Object	Name	Type	The_Geometry
1	A	Street	0,10; 10,10
2	B	Street	0,5; 10,5
3	X	River	0,7; 10,7

IdGeographic Object_has_ Excerpt	IdExcerpt	Id Geographic Object	The_ Geometry
101	100	1	3,10; 4,10
102	100	2	3,5; 4,5
103	100	3	3,7; 4,6

Figure 3. Example of the execution of the operation *boolean InsertAnnotation (Text Annotation, Geometry Region, Metadata Metadata, Context Context).*

- **boolean** *InsertAnnotation* (**Text** Annotation, **Geometry** Region, **List** {ObjectType Type}, **Metadata** Metadata, **Context** Context)

Figure 4. Example of the execution of the operation *boolean InsertAnnotation (Text Annotation, GeoQuery GeoQuery, List {ObjectType Type}, Metadata Metadata, Context Context).*

This operation inserts an annotation associated with a set of geographic objects that can be found in a given region and that are of a given type.

- **boolean** *InsertAnnotation* (**Text** Annotation, **Geometry** Region, **Query** Query, **Metadata** Metadata, **Context** Context)
  This operation inserts an annotation associated with a set of geographic objects defined as the result of the query specified in the parameter *Query*.
- **boolean** *InsertAnnotation* (**Text** Annotation, **GeoQuery** GeoQuery, **Metadata** Metadata, **Context** Context)
  This operation inserts an annotation associated with geographic objects defined as the result of the geographic query specified in parameter *GeoQuery*.
- **boolean** *InsertAnnotation* (**Text** Annotation, **GeoQuery** GeoQuery, **List** {ObjectType Type}, **Metadata** Metadata, **Context** Context)
  This operation inserts an annotation associated with a list of geographic objects found in a given region and that satisfy the query defined in parameter *Query*.
  Figure 4 illustrates the insertion of an annotation associated with the region defined by the rectangular region (mark). The annotation is linked to parts of Streets A and B (*Excerpt*), which are the result of the query specified by parameter *GeoQuery*.

3.2.2. Query

This section describes the interfaces for querying annotations, considering a list of keywords, an object, a region, or results from a query.

- **List** {**Text, Metadata** } *QueryAnnotation* (**List** {**MetadataField** Field, **Operator** Operator, **MetadataValue** Value, **Conector** Conector})
  This operation returns the annotations (and associated metadata), which satisfy the given parameters.

39

- **List** {**Text, Metadata** } *QueryAnnotation* (**List** {**String** Keywords})
  This operation returns the annotations which contain a given list of keywords.
- **List** {**Text, Metadata** } *QueryAnnotation* (**String** ObjectName)
  This operation returns the annotations which are associated with a given geographic object.
- **List** {**Text, Metadata** } *QueryAnnotation* (**Geometry** Region)
  This operation returns the annotations found in a given region.
- **List** {**Text, Metadata** } *QueryAnnotation* (**Query** Query)
  This operation returns the annotations associated with objects defined by the results of the query defined in *Query*.
- **List** {**Text, Metadata** } *QueryAnnotation* (**GeoQuery** GeoQuery)
  This operation returns the annotations associated with geographic objects defined by the results of the geographic query *GeoQuery*.
- **MarkGeometry** *QueryMark* (**Text** Annotation)
  This operation returns the geometry of a mark whose annotation includes terms found in parameter *Annotation*. This operation can be used, for example, to obtain the geometries of the marks defined by users X and Y, as illustrated in Figure 2(d).

4. Case Study: Biodiversity Information System

This section presents a case study concerning the use of our annotation Web service to handle biological data. In our case study, species field observations are associated with geographic information.

4.1. Data Source

The data source contains observations of butterflies collected during field trips to the Municipal Reserve of Santa Genebra, a remnant forest fragment in Campinas, Brazil.

The biodiversity database model proposed in [Malaverri et al. 2009] was used for storing field observations. This data model considers the following elements: species taxonomic information, description of collection methodology and data on how, where, when, and by whom butterflies were collected.
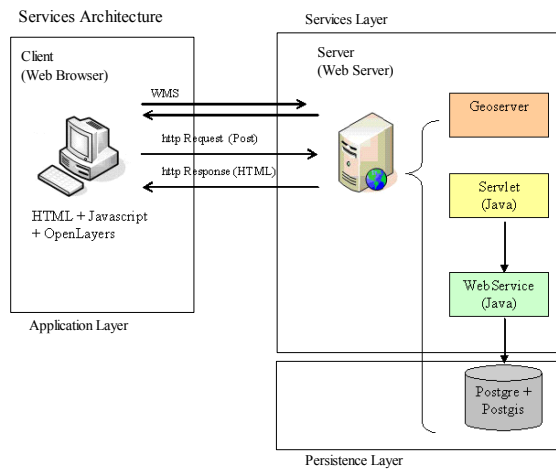
In our prototype, we consider the data provided by the study described in [Filho 2003]. The main objective of that study is to describe quantitatively the community of frugivorous butterflies with an emphasis on populations of *Anaea ryphea (Cramer) (Nymphalidae: Charanidae)*. Samples were held for 26 months. The data acquisition procedure considered the distribution of 12 traps in three areas with different levels of disturbance: 4 at the reserve boundaries, 4 in the forest interior and 4 near a central path (abandoned road), which crosses the whole reserve. The objective of this study is to evaluate butterfly features, given different luminosity, temperature, and wind conditions.

In this prototype, only text annotations were considered. Initially, the database was loaded with data obtained from digital maps of Brazil in Shapefile format (available at http://www.gismaps.com.br/). The following natural and environment resources were chosen for evaluation: Geography, Preservation Areas, and National and State Parks.

4.2. Architecture of the Biodiversity Information System

The architecture of the implemented Biodiversity Information System is composed of three layers: Application, Services, and Persistence (Figure 5).

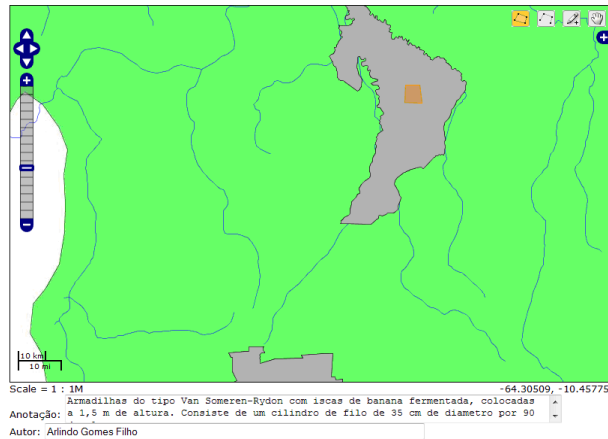**Figure 5. Architecture of the implemented Biodiversity Information System.**

The Application layer is responsible for the interaction with the user and the service layer. Usually, it can be a Web interface or a component of another system. In the case of our prototype, the client application layer was composed by HTML pages and Javascript running in a Web Browser. The rendering of the maps was performed on the server. The OpenLayers framework was used for map presentation. OpenLayers is an Open Source JavaScript library used to display spatial data in web pages. The OpenLayers API allows several tools to interact with the visual maps or performing actions on certain events. The HTML page invokes a servlet passing the desired operation and the required parameters.

The Service layer provides the APIs, which define how features can be invoked by other applications. This layer is composed by two components: a map server that is used to compose and render a map to be published on the web; and the proposed geographic annotation service that is used to insert annotations associated with geographic data. Geoserver (available at http://geoserver.org/display/geos/welcome) was used to implement the map server. GeoServer is compliant with the OGC open standards such as Web Map Service (WMS), Web Coverage Service (WCS) and Web Feature Service (WFS-T).
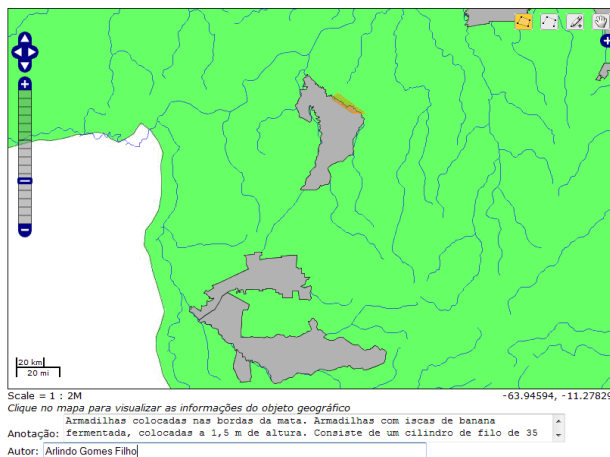
The Persistence layer includes database facilities to manage data about geographic objects, annotations, and marks. This layer uses PostgreSQL database system (available at http://www.postgresql.org) and PostGIS (available at http://postgis.refractions.net), whose provides mechanisms for processing spatial operators.

4.3. Usage Scenario

Figure 6 illustrates the creation of an annotation. Initially, a polygon is drawn within an ecological reserve (orange polygon). Later an annotation is inserted, indicating that the defined polygon contains special traps for capturing butterflies. In this case, the excerpt is a subregion of the ecological reserve.

41

Figure 6. Example of annotation of a region within an ecological reserve.

Figure 7. Example of annotation around the border of an ecological reserve.

A Biologist can insert annotation on traps located at the boundaries of the reserve (illustrated in Figure 7). In this example, the polygon was drawn around the border and is partly contained in the area of the ecological reserve. The excerpt is a subregion of the ecological reserve which *intersects* the defined mark region. The intersection operation takes advantage of the *InsertAnnotation* operation that uses a geographic query. A Biologist can also query annotations whose marks are contained within an ecological reserve (illustrated in Figure 8). In the example, the object was the geographical reserve REEXTE near the JACI-PARANA river.

A simple scenario is: suppose that a department has several biologists which worked in different Brazilian areas along the years, annotating reserves, rivers and regions. Using a simple map, it is easier to visualize which area was more explored, or visualize the intersection of regions which were explored by different researchers, for example.
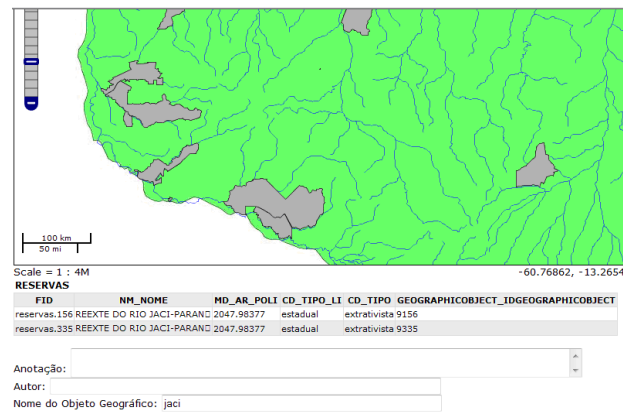
Figure 8. Example of an annotation query by textual term.

5. Conclusions

We have presented the specification and implementation of a Web service for managing geographic data in Biodiversity Information Systems. The service handles information in text format, referred as annotations.

The specification and implementation of the proposed Web service rely on two main contributions: the definition of a data model based on Superimposed Information and geographic concepts and the specification and implementation of a set of operations to manage them. One remarkable aspect of the proposed API relies on the support of operations (e.g., insertion, deletion, query, update, and deletion) which are defined over geographic objects that are result of queries. To validate our ideas, we developed a Biodiversity Information System that takes advantage of the proposed annotation service to correlate ecological data with geographic information.

Future work includes the extension of the proposed API to handle raster data, the extension for the SuperIDR tool [SI] and BioCore (available at http://www.lis.ic.unicamp.br/projects/biocore/). Another future work consists of applying the proposed service in other domains (e.g., Agriculture).

6. Acknowledgments

References

Agosti, M. and Ferro, N. (2008). Annotations: A Way to Interoperability in DL. In *ECDL '08*, pages 291–295.

Archer, D. W., Delcambre, L., Corubolo, F., Cassel, L., Price, S., Murthy, U., Maier, D., Fox, E. A., Murthy, S., Mccall, J., Kuchibhotla, K., and Suryavanshi, R. (2008). Superimposed Information Architecture for Digital Libraries. In *ECDL '08*, pages 88–99.

Barros, E. G., Laender, A. H. F., Gonçalves, M. A., and Barbosa, F. A. R. (2008). A digital library environment for integrating, disseminating and exploring ecological data. *Ecological Informatics*, 3(4-5):295–308.

Câmara, G., Casanova, M. A., Magalhães, G. C., and Medeiros, C. M. B. (1996). *Anatomia de Sistemas de Informação Geográfica*. 10a Escola de Computação, Instituto de Computação, Unicamp, Campinas.

Delcambre, L., Maier, D., Bowers, S., Weaver, M., Deng, L., Gorman, P., Ash, J., Lavelle, M., and Lyman, J. (2001). Bundles in Captivity: An Application of Superimposed Information. *17th ICDE*, pages 111–120.

Filho, A. G. (2003). A Comunidade de Borboletas Frugívoras da Reserva de Santa Genebra, Campinas, SP, Com Ênfase na Flutuação Populacional de Anaea Ryphea (Cramer) (Nymphalidae: Charanidae) e Sua Relação com as Plantas Hospedeiras. Master's thesis, Unicamp.

Gomes Jr, L. C. (2007). Uma arquitetura para consultas a repositórios de biodiversidade na web. Master's thesis, Unicamp.

Güting, R. H. (1994). An introduction to spatial database systems. *The VLDB Journal*, 3(4):357–399.

Macário, C. G. N., de Sousa, S. R., and Medeiros, C. B. (2009). Annotating geospatial data based on its semantics. In *GIS '09: Proceedings of the 17th ACM SIGSPATIAL*, pages 81–90.

Maier, D. and Delcambre, L. (1999a). Superimposed Information for the Internet. In *Proceedings of the ACM SIGMOD Workshop on the Web and Databases*, pages 1–9.

Maier, D. and Delcambre, L. M. L. (1999b). Superimposed information for the internet. In *WebDB (Informal Proceedings)*, pages 1–9.

Malaverri, J. G., Vilar, B. S. C. M., and Medeiros, C. B. (2009). A tool based on web services to query biodiversity information. In Filipe, J. and Cordeiro, J., editors, *WEBIST*, pages 305–310.

Sanchez, J. A., Flores, C. A., and Schnase, J. L. (1999). Mutant: Agents as Guides for Multiple Taxonomies in the Floristic Digital Library. In *Proceedings of the Fourth ACM JCDL*, pages 244–245, Berkeley, California, USA.

SI. Superimposed Information Projects at Virginia Tech, available at http://si.dlib.vt.edu/, last accessed on 01/10/2010.

Torres, R. S., Medeiros, C. M. B., calves, M. A. G., and Fox, E. A. (2006). A Digital Library Framework for Biodiversity Information System. *International Journal on Digital Libraries*, 6(1):3–17.

Yu, Y., Stamberger, J. A., Manoharan, A., and Paepcke, A. (2006). Ecopod: a mobile tool for community based biodiversity collection building. In *JCDL '06*, pages 244–253.

Zhu, B., Ramsey, M., and Chen, H. (2000). Creating a Large-Scale Content-Based Airphoto Image Digital Library. *IEEE Transactions on Image Processing*, 9(1):163–167.