available at www.sciencedirect.com

ScienceDirect

www.elsevier.com/locate/ecolinf

A web-based GIS tool for exploring the world's biodiversity: The Global Biodiversity Information Facility Mapping and Analysis Portal Application (GBIF-MAPA)

Paul Flemons[a,*], *Robert Guralnick*[b,c], *Jonathan Krieger*[b], *Ajay Ranipeta*[a], *David Neufeld*[c]

[a] *Collection Informatics and Analysis Unit, Australian Museum, 6 College Street Sydney, NSW 2010 Australia*
[b] *Department of Ecology and Evolutionary Biology, 334 UCB, University of Colorado, Boulder, CO 80309 USA*
[c] *University of Colorado Museum, 265 UCB, University of Colorado, Boulder, CO 80309 USA*

ARTICLE INFO

Article history:
Received 13 December 2006
Received in revised form
7 March 2007
Accepted 11 March 2007

Keywords:
Biodiversity mapping
Online Geographic Information Systems
Species richness
Survey gap analysis
Global Biodiversity Information Facility

ABSTRACT

Legacy biodiversity data from natural history and survey collections are rapidly becoming available in a common format over the Internet. Over 110 million records are already being served from the Global Biodiversity Information Facility (GBIF). However, our ability to use this information effectively for ecological research, management and conservation lags behind. A solution is a web-based Geographic Information System for enabling visualization and analysis of this rapidly expanding data resource. In this paper we detail a case study system, GBIF Mapping and Analysis Portal Application (MAPA), developed for deployment at distributed database portals. Building such a system requires overcoming a series of technical and research challenges. These challenges include: assuring fast speed of access to the vast amounts of data available through these distributed biodiversity databases; developing open standards based access to suitable environmental data layers for analyzing biodiversity distribution; building suitably flexible and intuitive map interfaces for refining the scope and criteria of an analysis; and building appropriate web-services based analysis tools that are of primary importance to the ecological community and make manifest the value of online biodiversity GBIF data. After discussing how we overcome these challenges, we provide case studies showing two examples of the use of GBIF-MAPA analysis tools.

1. Introduction

> "Overall, we are locked into a race. We must hurry to acquire the knowledge on which a wise policy of conservation and development can be based for centuries to come." E.O. Wilson (1988)

In the above quote, Wilson (1988) urges the global biodiversity community to develop a wise set of policies for biological conservation. Wisdom, in a systems analysis perspective, is the top portion of a hierarchy that also includes data, information and knowledge (Cleveland, 1982) — the so called DIKW (data, information, knowledge and wisdom) hierarchy. In order to achieve wisdom, one must first have data — raw observations and measurements. Information is amassed when data are initially summarized or analyzed. Knowledge is created when summarized information is interpreted and used for decision making. Wisdom is continued utilization of knowledge to guide behaviors into the future. Ultimately, the wisdom of our biodiversity decision making will be based in part on the quality of the data we have, on how effectively we can mobilize that data and tools to create knowledge, and on how quickly we can

[*] *Corresponding author.*
E-mail address: paul.flemons@austmus.gov.au (P. Flemons).

1574-9541/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.ecoinf.2007.03.004

achieve this goal given the current biodiversity crisis (Wilson, 1988; Gaston, 2000; Heywood and Watson, 1995).

The raw data for much of biodiversity research are specimen occurrence records — when and where species are found (Gaston, 2000). Accumulating global biodiversity data is a key step towards a wise policy of biodiversity conservation. Much of these data exist in natural history and survey collections, but the cost to discover and acquire the data has often been very high. To meet the goal of providing low-cost access to global biodiversity data, the Global Biodiversity Information Facility (GBIF) has developed an infrastructure by which museums and herbaria can publish their databases to a global network of biodiversity data (as discussed by Graham et al., 2004; Guralnick and Neufeld, 2005). This in turn provides a vast, freely available, and queryable resource for analyzing the distribution of biodiversity. As of March 2007, the GBIF data portal provides access over the Internet to approximately 120 million species occurrence records from over 1000 separate collections (http://gbif.net).

Biodiversity scientists have also developed methods for quantifying biodiversity measures and testing hypotheses about causes and consequences of biodiversity changes. These methods represent best practices for converting data into information and knowledge. They too need to be made available to the community so that as many workers as possible are using the best approaches to generate information and knowledge about biodiversity. Computer programs like EstimateS (Colwell, 2005) which provide estimations of species richness, and DIVA-GIS (http://www.diva-gis.org/; Hijmans et al., 2001), which provide GIS capabilities and a suite of biodiversity analysis modules, are examples of toolkits that implement such methods.

The parallel development of biodiversity science and informatics, and the maturing of internet speed, accessibility and data synchronization have created the potential for a paradigm shift in the way in which researchers and managers acquire information and knowledge about biodiversity. We argue that one way this convergence will occur is through global online distributed biodiversity mapping and analysis applications that streamline the workflow of doing biodiversity science (Guralnick and Neufeld, 2005). Such streamlining will both hasten our ability to generate information and knowledge about biodiversity and avoid costly duplication of survey effort.

We have developed a first generation web-based application called GBIF-MAPA (Mapping and Analysis Portal Application; http://gbifmapa.austmus.gov.au/mapa/) to support survey planning and species richness assessment. The application is a web-based biodiversity workflow tool that provides users the means to semi-automate raw biodiversity data acquisition, geospatial visualization and deployment of core biodiversity analyses based on that data. Building such a global mapping and analysis tool required solving a series of technical challenges; in the rest of this paper we delimit and discuss possible solutions to these challenges and present the particular implementations we employed in developing GBIF-MAPA. We also more fully discuss how the GBIF-MAPA application can be used to acquire biodiversity knowledge in regions of the world known to have tremendous biodiversity but where sampling has been limited. In particular, we provide two case studies showing how GBIF-MAPA functions. One case study examines species richness of rodents in the Ethiopian Highlands, and the other case study shows the use of the survey gap analysis tool for Anura (frogs and toads) on Madagascar.

2. Technical issues and solutions in developing GBIF-MAPA

The GBIF-MAPA application was developed to provide an end to end workflow for doing biodiversity-based analyses. This workflow is shown in Fig. 1 and involves a number of separate operations. The first and second steps are to make initial selection of the region and taxa of interest. The third step is to accumulate georeferenced species occurrence results from the

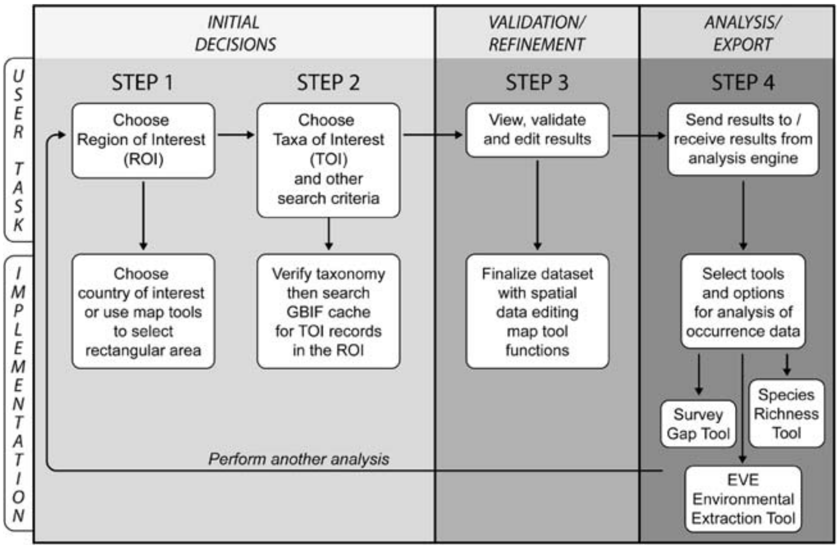

INITIAL DECISIONS
VALIDATION/ REFINEMENT
ANALYSIS/ EXPORT
USER TASK
IMPLEMENTATION
STEP 1
STEP 2
STEP 3
STEP 4
Choose Region of Interest (ROI)
Choose Taxa of Interest (TOI) and other search criteria
View, validate and edit results
Send results to / receive results from analysis engine
Choose country of interest or use map tools to select rectangular area
Verify taxonomy then search GBIF cache for TOI records in the ROI
Finalize dataset with spatial data editing map tool functions
Select tools and options for analysis of occurrence data
Survey Gap Tool
Species Richness Tool
EVE Environmental Extraction Tool
Perform another analysis

Fig. 1 – Diagram showing the user tasks and how they are implemented in the Global Biodiversity Information Facility Mapping and Analysis Portal. The tasks include: making initial decisions about the region of interest and taxon of interest; validating and refining the record set by mapping and checking records; performing one of three biodiversity analyses (species richness assessment, survey gap analysis and environmental extraction) and exporting results.

GBIF data cache, view results on a map and perform validation steps. The final step is to submit the species occurrence information, analysis parameters, and (where applicable) environmental information to one of three analysis tools which returns the results to the user. The first tool is a species richness analysis tool (SRA), the second a survey gap analysis tool (SGA) and the last a tool that extracts environmental conditions for selected species occurrences. At each step in the full process, there were critical technical decisions for developing a workable web application. These are discussed below.

We also had overarching application design strategies that guided our development process, including a focus on user interface and usability and a modular, fast back-end architecture. For user interface (UI) issues, we employed user interface experts when developing the information wireframe and workflow for the application, and when developing appropriate graphics that support user decision making. The other factor contributing to the user experience of a site is the architecture behind the interface. Designing an architecture that ensures suitable response times from database searches and application executables is key to the usability of an application. As important is a flexible architecture that could allow MAPA to grow beyond the life cycle of this initial project. Below we discuss the steps and technical challenges that were encountered during the development and deployment of the GBIF-MAPA application. We present our discussion in terms of a Challenge/Solution/Implementation breakdown of the issues. We begin with overarching issues, followed by those relating to implementing the GBIF-MAPA workflow (Fig. 1), and finally present case studies showing examples of how GBIF-MAPA can be used.

2.1. *Overarching issues for application development*

2.1.1. *Maximizing user certainty: Speed of application and user feedback*

2.1.1.1. Challenge. Web-based applications must provide the user with feedback within a reasonable timeframe. To fail to provide the user with adequate feedback and reward will discourage users and greatly reduce the application's acceptance and usability. Feedback can be in the form of results of a search or analysis, or simply be a symbol indicating that something is happening (preferably with an indication of when a result will be forthcoming).

2.1.1.2. Solution. Develop the back end architecture of the application to be as fast as possible and provide user interfaces that manage user expectation and as much as possible disallow long latencies. This includes developing user interfaces that are clearly reporting the application state (eg. waiting for next user step, collating and returning data to users).

2.1.1.3. GBIF-MAPA implementation. Major development decisions that applied across the application and helped increase speed of the application and effective UI reporting included:

- Locating the GBIF-MAPA application on the same network as the cache of raw biodiversity data and the taxonomic names database thus maximizing search speeds.
- Using Asynchronous JavaScript and XML (AJAX) for client-side development. AJAX increases speed of the application by avoiding unnecessary content refreshes. It also has built-in state reporting so that users know a process is "in action".
- Developing the user interface so that the user is unlikely to perform a search or analysis with long latency times (i.e. not longer than a minute or two at most). This required testing and adjusting search and analysis performance as the application was developed. For example the SGA tool was originally implemented with a 30 second cell size option, but once testing showed that the user may wait many minutes for a response, it was removed as an option.

2.1.2. *Maximizing user certainty: Intuitive and linear process flow*

2.1.2.1. Challenge. In an application like GBIF-MAPA that has a defined path to achieve its goal, it is essential that users know where to start, what steps are involved, what to expect at the conclusion of the steps, and where they are at any particular time in the process. The challenge is designing a user interface able to effectively lead users through the necessary steps to perform an analysis.

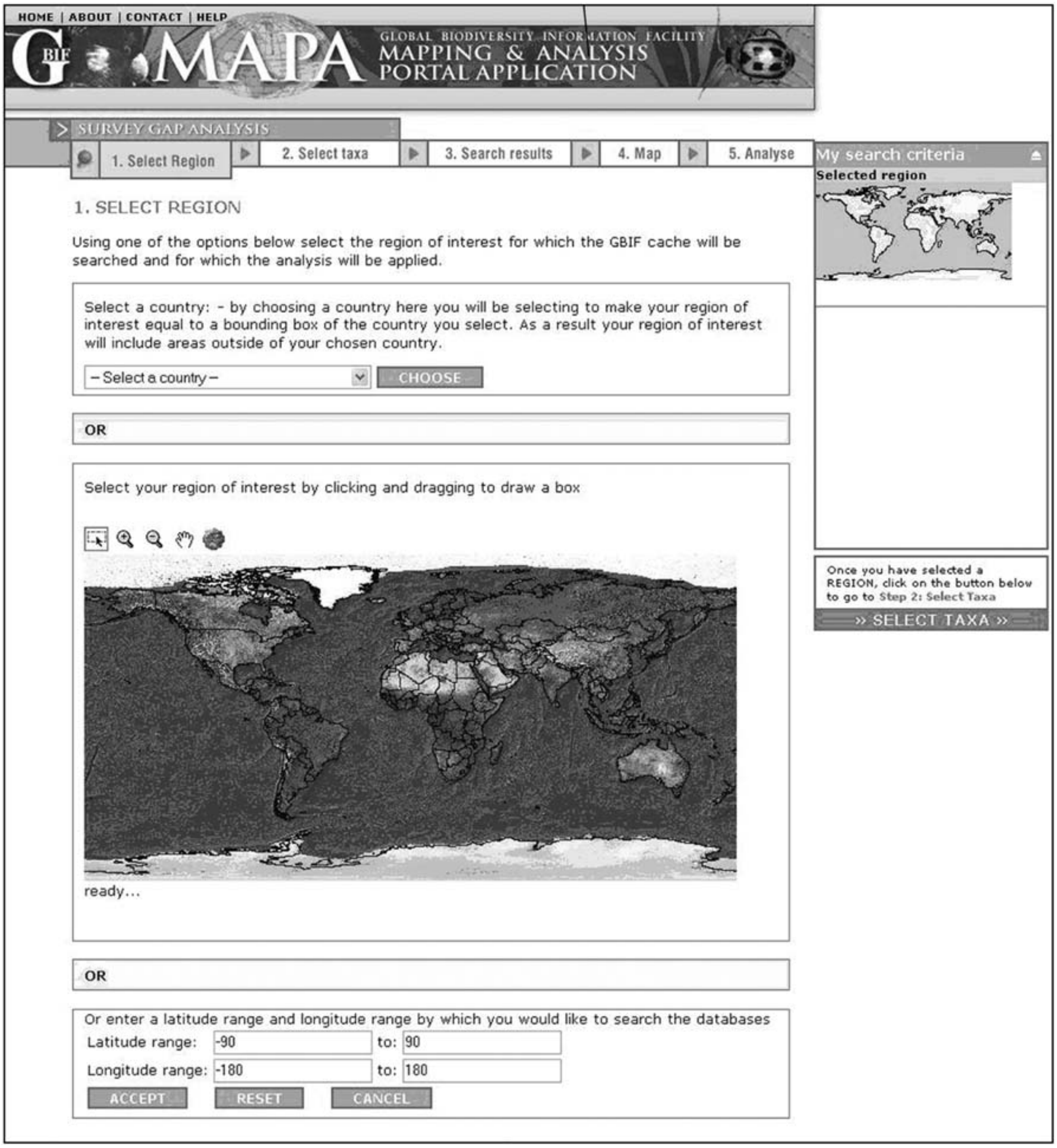
2.1.2.2. Solution. Use a linear process with clearly defined steps that provide the milestones in the process for key user input. Each webpage completes one step in the process and on each page the user interface has similar processes in the same place (eg. "move to next step" buttons) so that users can gain immediate comfort with the application.

2.1.2.3. GBIF-MAPA implementation. MAPA has a specific set of five steps that need to be followed when carrying out any of the three analyses. The first four are common to each analysis, with the fifth being unique. There were two options for leading the user through these five steps : 1.) The user chooses the analysis type at the beginning and then follows the five steps through to a result for the chosen analysis; 2.) The user completes the first four steps and then chooses the type of analysis they wish to use. We chose option 1 for MAPA because users' choices in the early steps are often driven by their analysis needs.

MAPA is designed so that the user can follow the progress through the steps in an analysis by following a "process flow bar" at the top of the page (top bar in Fig. 2). The navigation bar provides a simple way for users to track their progress in completing an analysis. Users can also click back to previous steps using the navigation bar providing a simple means to adjust a search or analysis parameter. In addition an updatable "shopping basket" type search criteria monitor box is always present on the far right of all pages so the user can always see what region and taxa they have chosen (Fig. 2 — My search criteria).

2.1.3. *Architecture for online GIS*

2.1.3.1. Challenges. Online GIS is a central component of GBIF-MAPA that cuts across many of the steps required to perform analyses. Users interact with the online GIS to select a region of interest, to validate their data, and again in some of the analysis tools (SGA, EVE). The challenge is to build an



1. SELECT REGION

Using one of the options below select the region of interest for which the GBIF cache will be searched and for which the analysis will be applied.

Select a country: - by choosing a country here you will be selecting to make your region of interest equal to a bounding box of the country you select. As a result your region of interest will include areas outside of your chosen country.

Select your region of interest by clicking and dragging to draw a box

ready...

Or enter a latitude range and longitude range by which you would like to search the databases

Latitude range: -90 to: 90

Longitude range: -180 to: 180

Once you have selected a REGION, click on the button below to go to Step 2: Select Taxa

Fig. 2 – GBIF-MAPA interface showing GUI design methods for maximizing user certainty. At the top of the screen under the GBIF-MAPA banner is the "process flow bar" and on the right of the page is the "search criteria summary box".

effective architecture for the online GIS that was consistent across the application but could also be flexible to perform new tasks (i.e. have new layers returned to it when an analysis is performed).

2.1.3.2. Solutions. Service oriented architectures (SOA) provide a framework for organizing a collection of services — in our case services which can respond to biodiversity analyses involving geospatial operations. While SOA technology is not new, the advent of web services and application frameworks like Spring have further simplified the ability of programmers to deploy SOA's quickly for a wide variety of purposes.

2.1.3.3. GBIF-MAPA implementation. The GBIF-MAPA implementation uses the Spring Application Framework (http://www.springframework.org/) to implement its SOA rather than a more traditional web services approach. This approach allows for a more simplified and compressed request/response mechanism between client and server. The user interface was designed using Java Server Pages (JSP) and requests are sent to the server using Asynchronous JavaScript and XML (AJAX). A simple controller class evaluates incoming requests and routes them to the appropriate service. Each service performs its operation or analysis and sends an XML response back to the client containing the results of analysis whether it be in the form of a map image, text attributes or a more complex rendering of a customized HTML page. Using this approach we have realized two significant benefits. The first benefit is a mechanism to more easily add additional services without the need to rewrite any of the existing services code as might occur in a more monolithic application. The second benefit is the ability to deploy additional mapping applications that have different user interfaces but require accessing the same core functionality.

2.2. Step 1: Make initial decision on region of interest

2.2.1. Enabling flexibility in selecting a region of interest (ROI)

2.2.1.1. Challenge. The selection of an appropriate region of interest (ROI) is crucial to any analysis tool that deals with spatial distribution of biodiversity. The ROI establishes the area in which an analysis technique will be applied. Allowing users the ability to select a meaningful ROI is an application development challenge, since the ROI can be relatively simple (eg. a rectangular bounding box) or more complex (eg. polygons representing boundaries of a national park).

2.2.1.2. Solutions. Selecting an ROI by which to search the GBIF cache has three possible approaches:

- Database field — the GBIF biodiversity cache has a field that defines the administrative units within which a collection record was recorded (eg. country, state, county). Stipulating an SQL query where the field “Country” is equal to “Italy” for example will return all specimen records matching the query. This method of selecting an ROI is potentially very fast and requires no spatial capability in the target database. The major downside to this approach is that the ROI is limited to predefined administrative areas rather than to user defined areas. The vast majority of biological questions do not follow politically defined units. Another downside to this approach is that it assumes that the administrative unit of interest is populated for the records in the database, which is not always the case.
- Bounding box — this approach defines the ROI by using a rectangle whose boundaries are defined by a lower left and an upper right coordinate. One advantage is that the bounding box is simple to implement in an online GIS. As well, the target database does not need spatial capabilities and searches on that database are typically fast. A downside is that large regions of interest that include water and land areas may require more sophisticated masking.
- Polygonal bounding area — users can define their ROI by uploading a shape file or interactively defining their ROI by drawing a polygon on a map interface. This option allows the user to most accurately and precisely define their ROI. However it is a major technical challenge to develop a UI for capturing a complex user drawn polygon. As equally problematic is the need for a spatially enabled target database to allow a polygon ROI query to be processed effectively.

2.2.1.3. GBIF-MAPA implementation. In MAPA, we use the bounding box approach because it did not require a spatially enabled target database, and was easier to implement given the time frame for development. Within MAPA however we supply three different means of deriving the bounding box (Fig. 2):

- The user selects a country by name, which is then converted into the best-fit rectangle around the country. Coordinates from this fit are used to search the GBIF cache. A limitation to this approach is that areas outside the country of interest but still within the approximating rectangle are also included in the analysis.
- Use GIS tools to draw a bounding box of interest using the map interface. The result is a rectangular area of interest. The great benefit of this approach is the availability of the visual interface and the associated GIS tools (zoom and pan) which provides users far greater flexibility in choosing an ROI than option 1.
- Manually type in the lower left and upper right latitude and longitude coordinates of the bounding box. This enables users to specify an exact ROI.

2.3. Step 2: Make initial decision on taxa of interest

2.3.1. Biodiversity data access

2.3.1.1. Challenge. Web-based biodiversity analysis applications typically require large amounts of data for analysis. Users may wish to access all records for a class, family or large genus which can result in large query returns — as many as hundreds of thousands of records. Currently the available standard protocols for distributed database queries return results in XML format. The nature of XML means that the XML package is considerably larger than the raw data. As more providers and records are available, the download times increase exponentially leading to long latency time. Remote access calls take a long time because the distributed database may be on the other side of the world, be hidden behind a firewall, or require a piece of software called a “wrapper” to interpret the XML and convert it into a language understood by the database. As a result, distributed biodiversity database mechanisms for accessing data for use in biodiversity analysis tools is particularly problematic. Web users demand fast response times.

2.3.1.2. Solutions. The current GBIF portal harvests key data from providers on a regular basis but is still centralized. The resultant cache provides an excellent data platform for building applications that use direct local JDBC (Java Database Connectivity) calls rather than remote access internet calls. JDBC calls to databases have considerable speed advantages.

2.3.1.3. GBIF-MAPA implementation. The GBIF-MAPA application is meant to be installed directly onto GBIF portals (which may be mirrored). Therefore when the application needs to request data, it can do so directly to the GBIF data cache using JDBC.

2.3.2. Validating names of taxa prior to analysis

2.3.2.1. Challenge. A common basis of biodiversity collection data is the scientific name and its associated taxonomic hierarchy. To improve search times and results it is important that the user validate the input taxon name against a list of current and accurate valid names. The major search issues include two types: obsolete names and incomplete and/or incorrectly spelled names. Obsolete names can be dealt with by the use of an accurate and up-to-date list of synonyms which link out-of-date names and incorrectly applied names with their correct current names. Incomplete and misspelled words require "contains-phrase" queries (or other fuzzy logic) which can generate many results.

2.3.2.2. Solutions. The ideal solution is to have a names web service that contains all of the known species names, and all of their associated synonyms. In reality the existing names lists are not comprehensive for either current names or the synonyms. The most suitable database is the Catalogue of Life (CoL) (Bisby et al., 2005), which covers approximately 50% of the named valid species (as of 2006) and will be nearly 100% by 2011. CoL is suitable for use in GBIF-MAPA given its global coverage. However, some areas of the world are better represented than others. In Australia, for example, CoL has limited coverage of faunal names in comparison to Europe or North America.

There are three ways of building names lists like CoL into a web-based search tool. The first way is to install the names database locally, allowing for fast search speeds but this approach has problems with maintaining currency. The second is to use a names web service. The web service approach may lead to slower result returns but is likely more up to date. A third alternative is a hybrid of these approaches. In this case the web-based names service is regularly harvested into a local database so that the names service is current and maintainable but still provides the fast search returns of a local names database. A further improvement would be implementing a "fuzzy" search engine to search for possible alternative spellings for commonly misspelled entries.

2.3.2.3. GBIF-MAPA implementation. MAPA uses a local version of the Catalogue of Life (CoL, http://www.catalogueoflife.org/search.php). Catalog of Life is a particularly good choice since it is the names database of choice for the GBIF portal therefore providing comparable search results in MAPA as for other GBIF tools.

2.4. Step 3: Map and validate species occurrences

After the user selects the region and taxa of interest, the application returns summary results for each taxon showing how many records are available in the ROI. The user is also provided with a drop-down list of map symbols for each taxon prior to mapping. After selecting the symbology, the user can then visualize the records on the map and do spatial data editing on the occurrence data (eg. moving the point where a species occurrence is found or deleting a point entirely). Once all these changes are made, the user is ready to send the data off to one of the analysis engines. The challenges here were to provide flexibility in mapping the records using a user-specified symbology and providing tools for spatial data editing. Mapping flexibility and spatial data editing operations are novel challenges in online GIS applications.

2.4.1. Providing user flexibility in mapping outcome

2.4.1.1. Challenges. The use of maps showing species occurrence records provides a powerful but simple way to visually validate location of species occurrences and examine relationships between those occurrences and environmental GIS layers (eg. rainfall, temperature, soils, vegetation, elevation). The challenge is to develop means to effectively view the occurrences of multiple taxa on multiple base layers and easily retrieve tabular data for those occurrences.

2.4.1.2. Solutions. There are many ways to optimize the effectiveness of viewing the distribution of taxa in an online GIS. For multiple taxa, one way to effectively show their distribution is to allow user selection of symbology for map search results. Another optimization is to allow the user to view or hide different basemap layers through a layer list. Finally, the online mapping toolkit can be developed to include a means for the user to select information on the map and have refreshed attribute data to view, or vice versa.

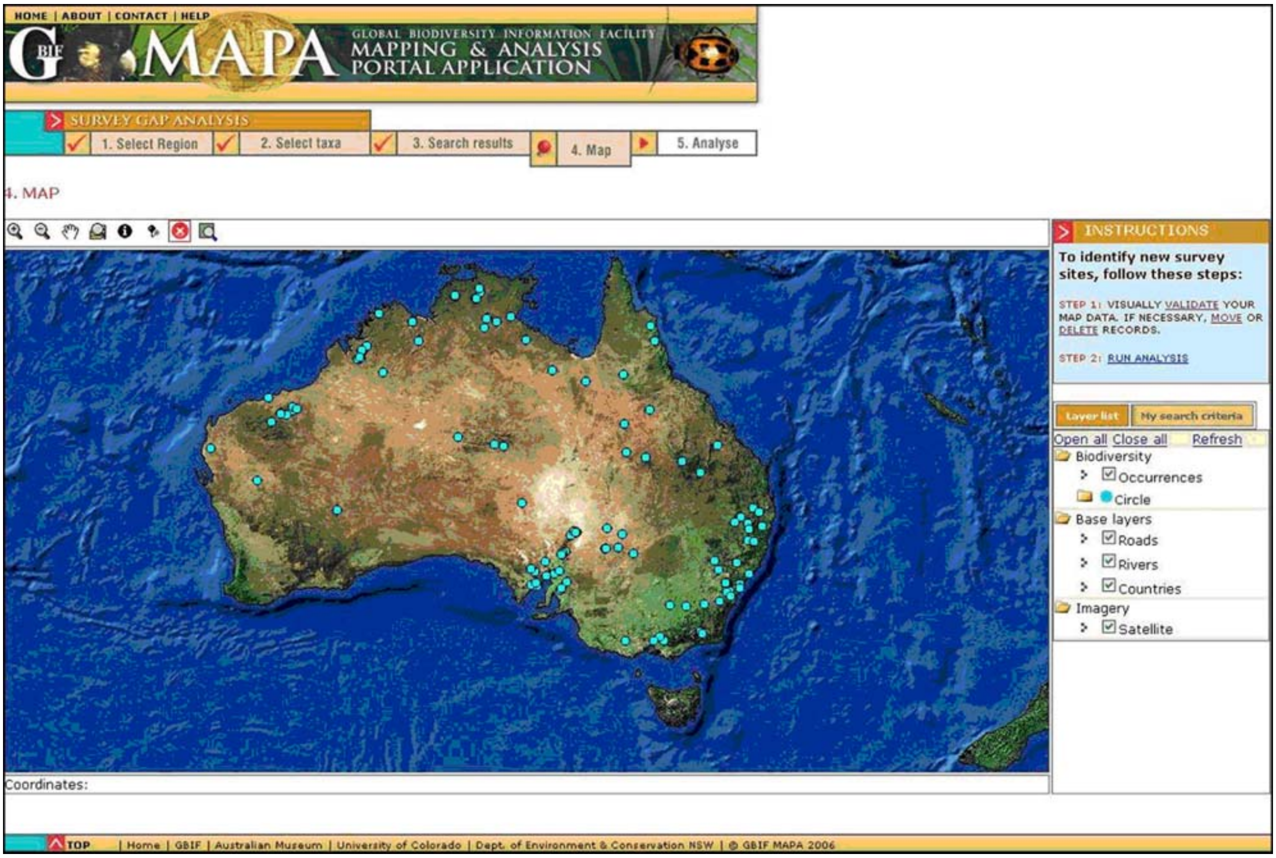
2.4.1.3. GBIF-MAPA implementation. For mapping species occurrence records, users are presented with a drop-down list of symbols for each taxon they have selected. However, only five symbols are available in that drop-down list. This prevents the map becoming overly cluttered, which would make interpretation more difficult. After the map is displayed with the user-selected symbol choices, the user may use the "select" tool to choose symbol(s), and retrieve the tabular data about species occurrences at the selected location(s).

MAPA also provides a limited set of base layers for users to toggle on or off, available as a layer list to the right of the map (Fig. 3), including global satellite imagery, roads, rivers, and country boundaries. The choice to include a relatively limited palette of base layers was intentional. Some tools do not require contextual layers (eg. environmental extraction). Other tools have need for only a few layers. For example, the essential contextual layer for the SGA tool is the road layer, providing necessary information about vehicle access to new survey sites.

2.4.2. Spatial data editing functionality in an online GIS

2.4.2.1. Challenges. Edit and delete capabilities are important for biodiversity analysis tools so that users may validate the quality of data records before using them in analyses. Tools that allow editing of spatial data have typically been relegated to desktop applications where edits are typically made to a flat file and multiple user conflicts are not an issue. A remaining challenge for web-based editing is in creating a secure and trusted environment for edits to be performed without sacrificing source data quality.

2.4.2.2. Solutions. Using a combination of open source GIS application software, web-based editing of geospatial data is no longer an expensive proposition. There are now viable, high quality freeware alternatives including PostGIS (http://postgis.refractions.net/) and MySQL (http://www.mysql.com/) that provide support for spatial data operations. UMN MapServer (http://mapserver.gis.umn.edu/) can be used for visualizing and selecting geospatial datasets stored in a PostGIS database. Edit requests sent from a web-based client application need only pass a unique identifier and the new geometry to PostGIS in order to update the geometry of the



To identify new survey sites, follow these steps:

STEP 1: VISUALLY VALIDATE YOUR MAP DATA. IF NECESSARY, MOVE OR DELETE RECORDS.

STEP 2: RUN ANALYSIS

Fig. 3 – GBIF-MAPA map interface showing tools and available layers.

feature. This operation can be performed by a spatial update service making use of underlying JDBC calls to the PostGIS database.

2.4.2.3. *GBIF-MAPA implementation.* In our case, edit and delete functions are being performed on a user-generated temporary data cache. After selecting an ROI and taxa, users are presented with a map of all species occurrences matching the criteria. From there, a user may click and drag a bounding box to display the attributes of the occurrences in a table below the map. Individual records may then be flagged using a check mark for editing or deleting. In the case of an edit, the user clicks on the map to identify the new location for the point, and an update request is sent to the spatial update service. The update service performs an SQL update of the point's geometry using its unique identifier in PostGIS and updates the web mapping client with an image showing the point's new location. Likewise, if a user wants to delete points, she simply submits a delete request which removes the record in the local cache and updates the map.

2.5. *Step 4: Data analysis and return results to the user*

After all the data has been validated, the user can then perform a set of analyses based on GBIF data including estimating species richness, determining new survey sites and determining environmental conditions based on species occurrence locations. For both the species richness analysis tool (SRA) and survey gap analysis tool (SGA), the user is required to specify some additional input parameters about how to perform and return the analysis results. The SGA and the environmental extraction (EVE) tool both require underlying environmental data in order to return results. The challenges in this phase include: how to decide the most effective tools to include in the application; how to choose development environments for the analysis tools; how to compile the environmental and contextual data layers needed for the analyses; how to deploy the analysis tools and; how to deal with latency issues with data returns to the user?

2.5.1. *Choosing analysis tools to meet biodiversity user needs*

2.5.1.1. *Issues.* Web-based biodiversity tools have focused primarily on simple visualization of data with some limited scope for summarizing that data. A next generation challenge is to provide a workflow for biodiversity research in a web-based environment. A first step in this process is to determine which web-based tools are relatively simple to implement but still meet a significant need for the biodiversity research community.

2.5.1.2. *Solution.* Analysis tools that already utilize species occurrence data available from GBIF and that are of the most use to the biodiversity research and management community are the obvious foci for deployment.

2.5.1.3. *GBIF-MAPA implementation.* We have focused on analysis tools that provide the user with two types of information — summary data on species richness in an area and a predictive modeling tool to assess the most likely locations where new biodiversity is to be found. Species richness estimation techniques not only give estimates of species numbers in an area, but also give insight into the probability of discovering new species (Colwell and Coddington, 1994). Thus a user may use the species richness tools for a taxon and region of interest and discover that different estimators do not converge on a single measurement of number of species present (see http://gbifmapa.austmus.gov.au/mapa/SRADetail.htm for the different estimators implemented in GBIF-MAPA). Such results suggest that more sampling is likely needed to determine the true species richness in that area.

The survey gap analysis (SGA) tool provides a powerful means for designing a biodiversity survey that will best complement the existing survey effort, by identifying those areas least well surveyed in terms of environmental conditions (Faith and Walker, 1996). Most surveys are carried out with the main deciding factors for site location being a combination of geographic coverage and available transport access. This approach is ineffective in providing a survey design that samples the breadth of potential environmental niches. SGA builds a multidimensional environmental space for the region of interest and then determines how well the existing site data (in this case GBIF specimen location data) spans that space. It then creates a GIS layer classified into complementarity values. The areas with the highest complementarity values are recommended as potential locations for survey sites that will maximize the environmental representativeness of the survey effort.

The SGA tool makes the assumption of equal collecting effort at each site across the region of interest, and this is often not the case given the ad-hoc nature of GBIF data. Therefore care must be taken in interpretation of the environmental complementarity surface (ECS). Funk et al. (2005) describe different strategies for dealing with patchy collecting effort. Since GBIF-MAPA does provide editing and deleting functions for records, a user may run the SGA analysis using different subsets of data to test the effect on outputs. For example, a user may choose to remove “sites” that have limited sampling.

Interpretations of the SGA tool outputs will vary depending on whether one is considering a single species or higher taxon. If one was to choose the records of a single species and the species true distribution is well represented by the data in the GBIF cache then the ECS will simply represent where one is least likely to find that species. On the other hand if the data in the cache is a poor reflection of the species true distribution then the ECS would identify those areas that may most contribute to an understanding of the true distribution of that species. If one was to choose all the records of a higher taxon, such as Order, as the “existing” sites then the resulting surface will represent those areas where it is most likely that new species, or new “range expanding” records of existing species, may be found.

2.5.2. *Deploying analysis tools for use in GBIF-MAPA*

2.5.2.1. *Challenges.* There are multiple ways to bundle analysis tools with the rest of the application. The choice of solution depends on a number of issues already touched upon here, including size of datasets that need to be transmitted over the web, and the importance of providing analysis tools in a way so that other developers could access them for their own applications.

2.5.2.2. *Solutions.* Given these issues, three reasonable implementation models are: single server installation where all components are installed on one computer; two server installation where web mapping server, cache and cache data query tool are installed on a local computer, and the analysis tools are installed on another internet-accessible computer and; a web services approach with each analysis tool and the cache query tool available as a service. The web services model is not bound within any infrastructure or software implementation and can be customised into any interface that utilizes them as necessary.

2.5.2.3. *GBIF-MAPA implementation.* For MAPA we have combined the use of web services for the analysis tools, with a two server implementation. The primary MAPA site runs in the Australian Museum network, with the SGA and EVE web services on a dedicated analysis server, and the database cache, web server and query application on a separate server. The final component of MAPA, the SRA tool, runs as a web service at the University of Colorado.

2.5.3. *Latency issues with analysis tools*

2.5.3.1. *Challenge.* Analysis tools deployed as web services can have multiple bottlenecks. There can be bottlenecks in: delivering the data from the map application to the tools over the network; time to process and analyze the data depending on development choices for the analysis engine, and; returning the summary results to the user. The best tools are ones that minimize those bottlenecks and provide the user the fastest possible return of analysis summaries.

2.5.3.2. *Solutions.* Developers need to make good programming language and coding choices to deliver analysis tools that are fast. As importantly, developers need to work with researchers to test how the tools perform with smaller and larger datasets to optimize the analysis services to the web-based environment. Limiting user interface choices to tried and tested parameter values is a crucial component in reducing latency issues.

2.5.3.3. *GBIF-MAPA implementation.* We developed the Species Richness Analysis (SRA) engine in Mathematica. Deploying SRA in Mathematica, which uses a higher-level language based primarily on mathematical functions, makes the software very fast and easy to modify in a testing environment. The species richness analyses performs resampling across grids of species data, the larger the number of the grids, the slower the application in returning results. Tests on multiple datasets showed that anything above 25 cell by 25 cell grids (625 cells) took a very long time to analyze, and the speed slowed exponentially with linear increases in samples per cell. We initially restricted the tool to two grid sizes, 100 and 625 cells. After further testing, even the 625 cell grid would sometimes lead to very long analysis times, and so we decided to only use the 100 cell grid. Although this decision limits flexibility, it also means that users typically get results quickly (in <60 s in most cases) and the tool is not processing multiple large datasets for long periods of time. We plan to continue experimenting with the trade-off between speed of analysis versus number of grid cells so we can provide finer grain analysis options.

Optimizations to code were also essential for the Survey Gap Analysis tool. The initial SGA executable was relatively slow in processing the large datasets involved. It was optimized by taking advantage of large available Random Access Memory (RAM) to implement a look up table that increased the speed of the analysis approximately 10 fold.

2.5.4. *Abiotic data access*

2.5.4.1. Issue. Providing access to environmental data layers is an essential task since tools like Survey Gap Analysis and Environment Extraction require these data for generating results. Many climate, topographic and landscape rasters are freely available but are often stored in a bewildering variety of formats and require expertise to properly understand and decode the values stored in the files.

2.5.4.2. Solution. An advantage of web-based applications is that environmental and contextual (eg. roads, rivers) layers may be managed centrally by experts instead of distributed along with software to each end-user. Desktop applications require that each user have some specialized expertise in formats and contents of raster environmental data. In web-based applications, experts can more quickly accumulate, prepare and present raster and vector environmental datasets to less knowledgeable users through the mapping interface. Another advantage is that multiple web applications can share a single set of compiled raster layers including near or real-time environmental layers like those available through the USDA Crop Explorer (http://www.pecad.fas.usda.gov/cropexplorer/). The disadvantage of pure web-based analysis is that user upload of new, unwarehoused environmental raster data to an online application is slow and potentially more difficult than local Desktop upload.

2.5.4.3. GBIF-MAPA implementation. A large set of freely available environmental data layers have been accumulated and are bundled with the GBIF-MAPA application. These layers include: Worldclim (Hijmans et al., 2001) monthly mean, minimum and maximum temperature, precipitation and altitude raster data available at square kilometer resolution; IPCC (2002) data layers for climate at .1 degree resolution (~11 km scale); and USGS (2002) hydrological and global land-cover data also at .1 degree scale. A global roads and rivers dataset, for accessing information in the SGA tool, is also included from the Digital Chart of the World (http://www.maproom.psu.edu/dcw/dcw_about.shtml#DCW).

3. Case studies

3.1. *Introduction*

The point of developing GBIF-MAPA was to provide a means for users to select any region in the world and ask core biodiversity research questions like: “Is there enough existing biodiversity data to determine accurate measures of species richness?” or “Where is the most likely spot to survey for more biodiversity given the current species occurrences and environmental conditions?” So far we have detailed how we developed GBIF-MAPA but we have not shown how it could be used to answer questions like the ones posed above. To make this more clear, we discuss two case studies – one utilizing the survey gap analysis tool and the other utilizing the species richness analysis tool – that exemplify how GBIF-MAPA functions for acquiring biodiversity knowledge. We chose these case studies based on two criteria: 1.) The regions and taxa were not ones with which we were intimately familiar; 2.) The regions are known to be hotspots for biodiversity.

3.2. *Where to look for new frog and toad species on Madagascar?*

Our first cast study focuses on Madagascar, one of the world’s 34 biodiversity hotspots (http://www.biodiversityhotspots.org/xp/Hotspots/). Madagascar developed a unique suite of fauna and flora after its divergence from the African mainland 160 million years ago, and much of that diversity is now known to be threatened or endangered (http://www.iucnredlist.org/). Not least of this diversity are the amphibians, which are almost entirely endemic to the island. The amphibian Order Anura (frogs and toads) was chosen for this case study because there are only a few Anura records from the GBIF data cache on mainland Madagascar and all but one of these is located in the bottom fifth of the island. This distribution suggests that there may be undiscovered anuran species on Madagascar.

We used the Survey Gap Analysis tool to determine best survey locations on Madagascar for anurans. First, Madagascar was chosen from the countries drop-down list as the region of interest (ROI). This sets a bounding box around Madagascar as the ROI. After choosing Anura from the Catalogue of Life, as the taxa of interest, the GBIF cache was searched returning 51 records. Upon mapping these records it was found that 12 of them were located in the ocean. Assuming these to be incorrect they were selected and deleted. This left 39 records, 38 of which were clustered in the bottom fifth of the island, suggesting poor coverage of the islands’ environmental niches. The final step is to choose demand points and a scale of analysis. Demand points are simply the number of cells chosen to create the environmental space into which the existing survey points are projected. We chose the default which is 1000. Analysis scale options are 2.5, 5 and 10 min grid cells. We chose 2.5 to give as detailed an analysis as possible.

The analysis returned a complementarity surface and a selected “highest complementarity” survey location (Fig. 4A). The surface grades from dark red where the complementarity is the highest to blue where it is the lowest. The results of iterative analyses, including the newly selected survey

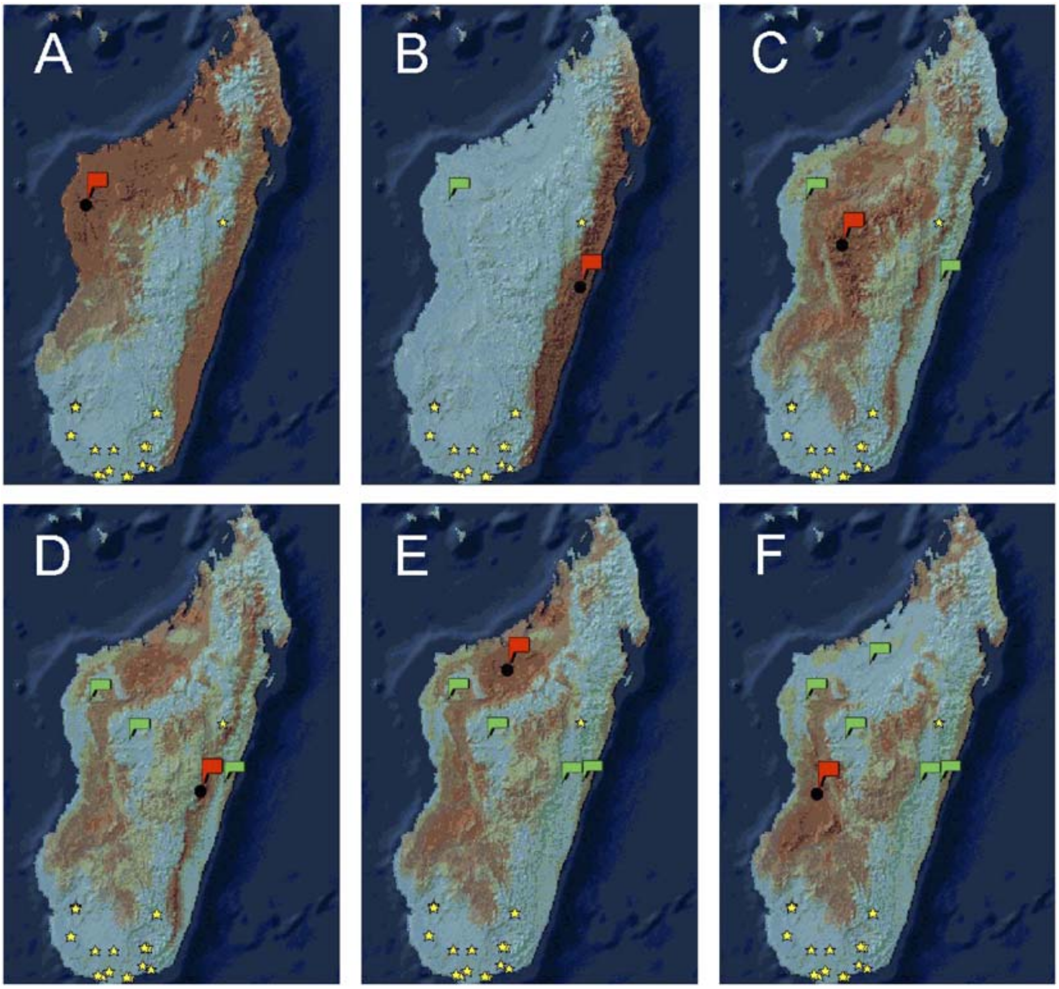


Fig. 4 – Panel of maps showing the results of the iterative selection of points by the Survey Gap Analysis algorithm. The colour gradation from blue to brown indicates areas of low complementarity to high complementarity. The red flags represent the optimal survey location in terms of most complementing existing survey effort. The green flags mark previously chosen new survey locations. A — Red flag marks the location of the site of highest complementarity value relative to the existing survey effort (yellow stars) B to F — addition of one new site per iteration, each iteration identifies the next highest site of complementarity. Each green flag represents a previously chosen new survey location. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

locations in each run, can be seen in Fig. 4. The maps (Fig. 4) show the point locations identified as optimal by the SGA algorithm. Each of the new survey sites could have been moved to accommodate access (defined by roads and rivers data layers), which would have changed the next iteration's selected survey site. SGA identified the central east west band of the island as most requiring survey effort. The SGA does not suggest the utility of sampling in the northernmost portion of the island, highlighting that more remote or distant geographical areas may be environmentally similar (in terms of the climate data used) to other areas already sampled and thus not likely to be where new biodiversity is discovered.

3.3. *Can Ethiopian rodent species richness be estimated using GBIF data?*

The case study for species richness assessment is rodent richness in the country Ethiopia. A relatively large portion of the country comprises the Ethiopian Highlands, which is one of thirty four global biodiversity hotspots (Williams et al., 2005) especially rich in vertebrate endemics (Brooks et al., 2001). It is also an area of global conservation concern given high species richness and number of endemics, and given increasing and dramatic human impact on the area (Williams et al., 2005). Of the 48 mammal species listed by Conservation International as threatened endemics in the Ethiopian Highlands, 17 are rodents.

For this case study, we selected the country “Ethiopia” as our region of interest and “Rodentia” as our taxon of interest. The search under these criteria returned 1053 specimen occurrence records. These records were then mapped and spot validation performed to make sure the records appeared to be valid species known to occur in the area. Examining all the records proved important because some of the records had “Rodentia” as Order but did not list a scientific name for the records. These were deleted. As well, some records were only identified to genus (eg. "*Mus* sp."). These records were also deleted so that they would not be mistaken as individual novel species in the analysis. This reduced the number of records from 1052 to 998. It also reduced the number of species present from 77 in the original sample to 63 in the validated one.

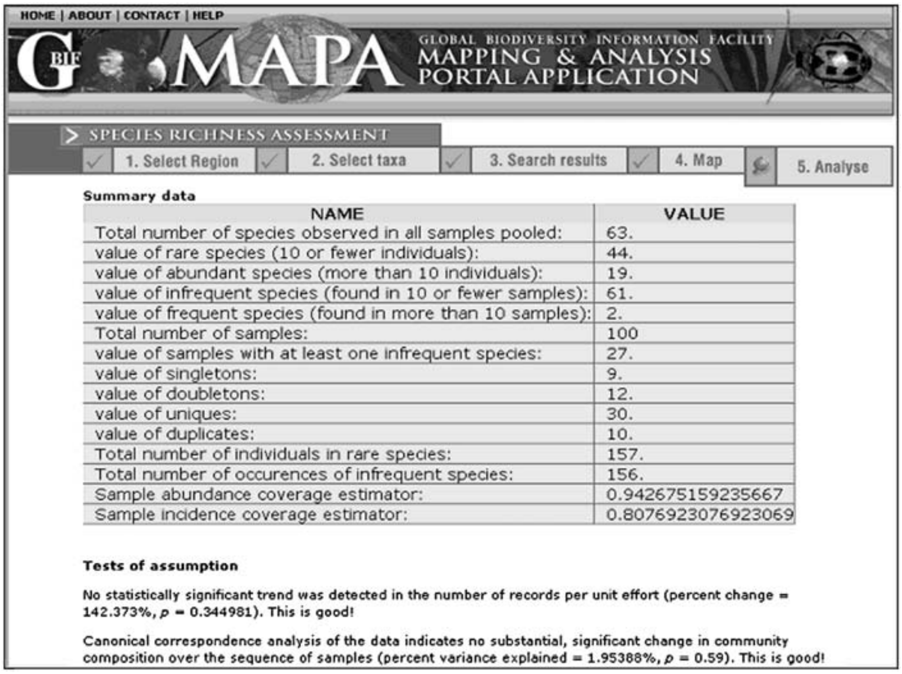
The next step was to perform the species richness analysis, using a one hundred cell grid (each grid cell was167 km^2). Some of the outputs from the analysis after specifying the grid size are shown in Fig. 5. The main reason for performing such an analysis is to gauge whether further sampling may lead to returns in terms of finding new species not already in the collected sample, and to estimate the potential number of species in the sample using non-parametric richness estimators. These questions can be examined by generating sampling accumulation curves and determining if the curves asymptote such that as new records are added, no new species are being found. An asymptoting curve is not meant to suggest that the dataset is representative of richness for the area. Instead, an accumulation curve that plateaus only implies that more sampling using the same approaches and methods will likely lead to negligible returns. If observed species richness does not appear to plateau, species richness estimators can provide an assessment of true species richness.

The non-parametric estimation results from this case study (randomized 100 times with and without replacement to account for sample order) suggest that inclusion of more records will have appreciable returns in terms of species richness. That is, most of the non-parametric species richness estimators show higher values of species richness than observed values. However, it is difficult to give a reasonable number for true species richness since species richness values differ among the estimators and none of the estimates gave a stable asymptote. Similar to the results from Guralnick and Van Cleve (2005), incidence estimators (eg. Incidence Coverage Estimator, Chao2) give much higher estimates of species richness than the abundance based estimators (eg. Abundance Coverage Estimator, Chao 1) when randomizing with replacement. This likely reflects that species occurrences are not randomly distributed across the environment (eg. there is patchiness); the number of unique occurrences (found in only one sample) is much higher than the number of singletons (only one record). Given this, it seems likely that the incidence-based measures may provide a better estimation of actual richness since these measures are generally less sensitive to patchiness (Magurran, 2003).

Datasets collected using different methods may lead to radical reassessment of the regional diversity. With data from GBIF, the collecting method is ad-hoc aggregations from multiple collecting events, some of which may be more systematically collected than others. Guralnick and Van Cleve (2005) noted that at least for birds these ad-hoc collections tend to show skewed abundance distributions due to a bias toward "rare representation". Collectors often focus more on collecting rare species than accumulating all species common or rare in systematic fashion. The results for Ethiopian rodents are consistent with those from Guralnick and Van Cleve (2005) suggesting that similar biases may exist for rodents as for birds.

4. Conclusions

1. We address two fundamental questions in ecological and biodiversity informatics: How can high quality information in a global biodiversity portal support research, conservation management, and education missions, and how can



Summary data

NAME	VALUE
Total number of species observed in all samples pooled:	63.
value of rare species (10 or fewer individuals):	44.
value of abundant species (more than 10 individuals):	19.
value of infrequent species (found in 10 or fewer samples):	61.
value of frequent species (found in more than 10 samples):	2.
Total number of samples:	100
value of samples with at least one infrequent species:	27.
value of singletons:	9.
value of doubletons:	12.
value of uniques:	30.
value of duplicates:	10.
Total number of individuals in rare species:	157.
Total number of occurences of infrequent species:	156.
Sample abundance coverage estimator:	0.942675159235667
Sample incidence coverage estimator:	0.8076923076923069

Tests of assumption

No statistically significant trend was detected in the number of records per unit effort (percent change = 142.373%, p = 0.344981). This is good!

Canonical correspondence analysis of the data indicates no substantial, significant change in community composition over the sequence of samples (percent variance explained = 1.95388%, p = 0.59). This is good!

Fig. 5 – Tabular summary output from the species richness analysis (SRA) of rodents in Ethiopia as generated from GBIF-MAPA. After validation, there were 998 usable species occurrence data points from the GBIF data cache. The summary data is part of a much larger set of output from SRA.

ecological informaticians build tools utilizing the GBIF data portal to support those missions? We argue that maturation of web application development technology, accessibility of biodiversity data, and enhanced data synchronization create the potential to build web-based workflows for acquiring information and knowledge about biodiversity.

2. GBIF-MAPA is a first generation tool for performing an end-to-end workflow for biodiversity analysis online. The tool allows users to accumulate biodiversity data from global portals for a region of interest and then perform three types of analyses: environmental extraction, survey gap analysis, and species richness assessment.
3. GBIF-MAPA was developed based on the following design goals: the back-end application needed to maximize speed of acquiring biodiversity data, and performing analyses; the user interface needed to minimize user uncertainty and provide flexibility in representing results; the analysis engines needed to be available to other developers to use for their own application. To maximize data access, we developed an online GIS that resides on the same server as the GBIF data cache. To maximize analysis speed, we developed analysis tools in programming languages explicitly built to handle math functions that form the core of the analysis engines. To minimize user uncertainty, we built a very simple step-by-step process for analysis that provides feedback to the user about the state of the application and that gives the user flexibility in designing outputs (eg. user selection of map symbols). Finally, the analysis engines were built as web services linked to the online map and GBIF portal so that other developers could access those services for their own applications.
4. The survey gap analysis tool shows the power of linking GBIF species occurrence data, online GIS, and analysis engines together. The tool uses species occurrence data and raster environmental data layers to create new map layers in the online GIS that show those areas least well surveyed in terms of environmental conditions. The user iteratively runs the SGA analysis, utilizing both the original species occurrence and new survey sites, to generate new maps and a final list of potential survey areas. The ability to iteratively run map-based biodiversity analyses in a web-based environment had not been achieved before.

Acknowledgements

We are indebted to the Global Biodiversity Information Facility for demonstration project funding to carry out this work. Donald Hobern and his developers from GBIF provided access to a static copy of the GBIF data cache to test use of the tools. Beatriz Torres, the GBIF Outreach and Capacity Building Programme Officer, was a constant supporter and excellent beta-tester of the tool. Thanks to Simon Ferrier (NSW Department of Environment and Conservation), and Glenn Manion for building the Survey Gap Analysis application which was used in GBIF-MAPA; to Ruby Blessing of Kinetic Media for making MAPA look fabulous; and Michael Elliott for his assistance in preparing the figures. The SRA tool was built on Gareth Russell's species richness tool (Estimating Species Richness v2.1, 2006; http://www.eco-tools.net).

REFERENCES

Bisby, F.A., Ruggiero, M.A., Wilson, K.L., Cachuela-Palacio, M., Kimani, S.W., Roskov, Y.R., Soulier-Perkins, A., van Hertum, J. (Eds.), 2005. Species 2000 and ITIS Catalogue of Life: 2005 Annual Checklist. CD-ROM; Species 2000: Reading, U.K.

Brooks, T., Balmford, A., Burgess, N., Fjeldsa, J., Hansen, L.A., Moore, J., Rahbek, C., Williams, P., 2001. Towards a blueprint for conservation in Africa. BioScience 51, 613–624.

Cleveland, H., 1982. Information as resource. Futurist 34–39 December 1982.

Colwell, R.K. 2005. EstimateS: Statistical estimation of species richness and shared species from samples. Version 7.5. User's Guide and application published at: http://purl.oclc.org/estimates.

Colwell, R.K., Coddington, J.A., 1994. Estimating terrestrial biodiversity through extrapolation. Philosophical Transactions of the Royal Society (Series B) 345, 101–118.

Faith, D.P., Walker, P.A., 1996. Environmental diversity: on the best-possible use of surrogate data for assessing the relative biodiversity of sets of areas. Biodiversity and Conservation 5, 399–415.

Funk, V.A, Richardson, K.S., Ferrier, S., 2005. Survey-gap analysis in expeditionary research: where do we go from here? Biological Journal of the Linnaen Society 85, 549–567.

Gaston, K.J., 2000. Global patterns in biodiversity. Nature 405, 220–227.

Graham, C.H., Ferrier, S., Huettman, F., Moritz, C., Peterson, A.T., 2004. New developments in museum-based informatics and applications in biodiversity analysis. Trends in Ecology and Evolution 19, 497–503.

Guralnick, R.P., Neufeld, D., 2005. Challenges building online GIS services to support global biodiversity mapping and analysis: lessons from the mountain and plains database and informatics project. Biodiversity Informatics 2, 56–69.

Guralnick, R.P., Van Cleve, J., 2005. Strengths and weaknesses of museum and national survey datasets for predicting regional species richness: comparative and combined approaches. Diversity and Distributions 11 (4), 349–359.

Heywood, V.H., Watson, R.T., 1995. Global Biodiversity Assessment. Cambridge University Press, Cambridge.

Hijmans, R.J., Guarino, L., Cruz, M., Rojas, E., 2001. Computer tools for spatial analysis of plant genetic resources data: 1. DIVA-GIS. Plant Genetic Resources Newsletter 127, 15–19.

Intergovernmental Panel on Climate Change (IPCC). 2002. Climate data archive. IPCC, Geneva, Switzerland. Available from http://www.ipcc.ch/ (accessed December 2006).

Magurran, A.E., 2003. Measuring Biological Diversity. Blackwell, Malden, Massachusetts. 256 pp.

U.S. Geological Survey (USGS), 2001. HYDRO1k Elevation Derivative Database. USGS, Washington, D.C.. Available from http://edcdaac.usgs.gov/gtopo30/hydro/ (accessed March 2006).

Williams, S.D., Vivero Pol, J.-L., Spawls, S., Shimelis, A. and Kelbessa, E. 2005. Ethiopian Highlands. In Hotspots Revisited (eds. Mittermeier, R., et al.). Conservation International. Cemex Press.

Wilson, E.O., 1988. The current state of biological diversity. In: Wilson, E.O (Ed.), Biodiversity. National Academy Press, Washington, D.C., pp. 3–18.