GBIF
www.gbif.org

GBIF Position Paper on Future Directions and Recommendations for Enhancing Fitness-for-Use Across the GBIF Network

August 2010

Suggested citation:
GBIF. 2010. GBIF Position Paper on Future Directions and Recommendations for Enhancing Fitness-for-Use Across the GBIF Network, version 1.0. authored by Hill, A. W., Otegui, J., Ariño, A. H., and R. P. Guralnick. 2010. Copenhagen: Global Biodiversity Information Facility, 25 pp. ISBN: 87-92020-11-9. Accessible on-line at http://www.gbif.org.

Layout: Ciprian Marius Vizitiu

ISBN: 87-92020-11-9

Contents

1. Introduction

In the midst of a biodiversity crisis of yet unknown magnitude (Pimm et al., 1995; Jenkins 2003), the community is working hard to coordinate the sharing and using of biological datasets from the diversity of natural sciences. In those efforts, geospatial data are a key component that can help us join biodiversity information with data from other sources to study where species exist and how they are responding to a changing environment (see, Soberón, 1999; Guralnick and Van Cleve, 2005; Green et al., 2005; van Zonneveld et al., 2009). Some results of this work are taxon or biome specific biodiversity networks (e.g. VertNet or OBIS) that feed to global resources such as the Global Biodiversity Information Facility (GBIF). Each of these data sharing networks arose to help efficiently publish, share, and discover data and information about biodiversity. Assuring that biodiversity data from these networks is as accurate as reported is essential given the myriad uses of such data in biological research, conservation assessment and education. Fortunately, the community has actively developed standardized approaches and methods for sharing biodiversity records. However, despite the best efforts of all involved, undocumented problems with geospatial data still persist. Each user therefore must vet records carefully to determine their fitness-for-use: often, a time consuming task. Although user vetting will always happen, the key discussion point in this white paper is what can be done prior to user access of data to enhance and better report the data's fitness-for-use.

Fitness-for-use refers to a scale of data quality that changes with the varying data accuracy, precision and intended use. For some applications, data quality can be relatively low and still fit for use. In the context of geospatial data, we can split fitness-for-use into two broad categories:

1. Are the geospatial data correct?
2. Are the geospatial data usable at the geographic scale of the question?

For example, coarse scale geospatial data may only be usable for continental or global analyses but certainly not for local analyses. In this system, incorrect data not flagged as such are a particularly vexing issue because these poor data may appear fit for use, feed into analyses, and cause errors in interpretation. The community knows errors without annotation exist in the GBIF network and this erodes the community confidence in all the data. While multiple methods of documenting fitness-for-use have been employed by both the primary institutions that curate the data as well as the organizations that assist in sharing that data, much more can and should be done. Three areas in particular require attention: improvement of revision and republication methods for data publishers, new and improved methods for documenting different areas of geospatial fitness-for-use, and adoption of new technology to increase the speed at which fitness-for-use enhancement can be performed on the entire available biodiversity information dataset. While much of the groundwork for discussing these concepts was developed in Chapman (2005), we attempt to build on that work to highlight several future directions for enhancing geospatial fitness-for-use in biodiversity data.

No matter the effectiveness of existing tools to detect geospatial errors, resolving those errors directly relies on better methods for data publishers to review external annotations, revise their records, and republish datasets. These technologies are becoming available but still need community attention. We focus on the Integrated Publishing Toolkit and the existing annotation schema as necessary components in this process. While the IPT may not be the solution, we are particularly focused on its ability to handle two way exchange of information at data publishing nodes in the biodiversity network, something that all nodes need to handle in the near future. To determine if geospatial data are correct, we explore technologies such as the GBIF filter and existing georeferencing tools (e.g. BioGeomancer and GEOLocate; Guralnick et al., 2006; Rios and Bart 2009) as a primary means to avoid geospatial errors and to generate georeferences according to best practices. Georeferencing technologies and solutions are far from complete, with the ongoing need to expand the number of languages represented and temporal coverage of the existing georeferencing tools. However, we do not delve too deeply into the best practices for georeferencing but refer readers to the Chapman and Wieczorek (2006) "Guide to Best Practices for Georeferencing" document. That document explains the procedures and necessity for, at minimum, a latitude-longitude-geographic uncertainty triplet. We explore several yet undeveloped tools, representing both algorithmic and workflow approaches for annotating errors and improving the quality of biodiversity data. Finally, we discuss the implications of cloud computing technologies to make these solutions scalable and widely accessible into the future.

There is still great potential for the community to generate many improvements to the quality of the geospatial components of biodiversity data. Scientists, policy makers, and educators rely on our community to

ensure that we document known errors and the quality of the data we publish. We feel that several key changes in the network could drastically reduce the cost of doing so while leading to major fitness-for-use improvements. If done effectively, these improvements could greatly increase the community's confidence in the data and will help

Recommendations and necessary action		
1.	Recommendation	**Continue support of georeferencing initiatives, geospatial data training services and outreach.**
	Action	Supporting georeferencing technologies and solutions is nothing new to GBIF. We do not spend much time developing ideas for the further development of georeferencing because it has been explored previously. However, it is clear that retrospective georeferencing will continue to be a primary method to increase fitness-for-use across the GBIF network. Similarly, we feel that GBIF and nodes have played an invaluable role in supporting community training and workshops to help data publishers improve data errors before they enter the network. These activities are still needed.
2.	Recommendation	**Improve existing data filter through the addition of formalized, core validation steps that can be expanded with future capabilities**
	Action	GBIF should consider reconfiguring the geospatial issues filter based on several recommendations detailed below. Primarily, deploy a series of tests on spurious records to determine if errors are resolvable prior to record exclusion and annotate exactly where records fail in the filter.
3.	Recommendation	**Promote and further develop methods of external record annotation across the GBIF network and methods for reporting those errors back to data publishers**
	Action	Explore the best method for exchanging annotated and corrected data with data publishers. A currently draft-form annotation schema can be more broadly applied so that many of the larger data publishers that do not use IPT can still use external annotations to enhance the quality of their data. Finalizing the annotation schema and providing outreach and education to data publishers on how to use annotations to enhance their data will greatly increase the speed at which we can improve fitness-for-use in the network. Also, developing methods to publish annotations at GBIF should be explored. In cases where single resources contain numerous but non-diverse errors, technical outreach may be required. Helping resource managers isolate and fix their problematic records would be a significant step forward.
4.	Recommendation	**Enable users to access data using cloud-based infrastructures through simple, reliable, and fast solutions**
	Action	Publish a portion of GBIF data in a publicly available cloud environment. Amazon Public Datasets or Google Public Datasets might be good candidates. To be successful we must make data more accessible. GBIF should solicit its current publishing partners to find the parties willing to allow data to be published as part of a unified, cloud hosted dataset. GBIF should then make that dataset available.
5.	Recommendation	**Support the development of methods for incorporating new information about species into our measurement of fitness-for-use in species occurrence data**
	Action	In addition to cloud based datasets and enhanced quality reporting tools, habitat preference maps and species distribution based methods may make it possible to rapidly scan biodiversity records for likely errors or highly valuable records for our knowledge of biodiversity. The community would be well served to fund the development of new methods for combining the various sources of data into meaningful fitness-for-use assessments and quality checks.
6.	Recommendation	**Draft and publish a Memorandum of Understanding for the biodiversity data publishing community.**
	Action	As the biodiversity publishing network coordinator, GBIF is the right choice for starting a larger community discussion on openness and availability of data. GBIF should support a meeting of several community leaders in data publishing and consumption to discuss the needs for a unified biodiversity dataset that is open, free, and easily usable by anyone. The MOU should consider both short-term understanding of concerns over data provenance and the long-term need to bring biodiversity data together in order to save it in the face of global change.

revolutionize the way in which biodiversity data is used to generate information and knowledge.

2. Nature and history of geospatial data quality issues

There is a long history behind geospatial data quality and the effect it has on geospatial fitness-for-use. A famous case is Darwin's finches. Darwin failed to fully document the geographic origin of his Galápagos Finch samples (Sulloway 1982) limiting the value of assessing multiple climatic and dietary changes that might have led to changes in the size, shape and functioning of the beaks of the finches over the last 150 years. However, since localities appear to have been reconstructed for some of the samples after being vouchered in the British Museum, the case of Darwin's Finches may also one of the first attempts of retrospective georeferencing for museum specimens.

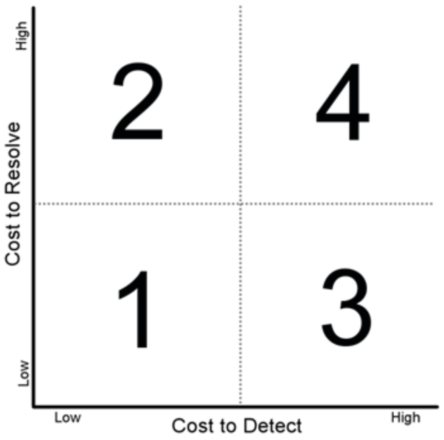
The errors that affect geospatial fitness-for-use are diverse, including the lack of or bad use of standardized methods, errors in geospatial data conversion (i.e. locality string to decimal degrees, UTM to decimal degrees, other datum conversions), errors introduced during data entry, and others. The majority of errors can be classified according to the degree of automatic detectability and resolvability. Automatically detectable and resolvable errors are those whose nature are clear and can be automatically detected and corrected with a predefined set of procedures. There are other errors which are easily detectable (i.e. land vertebrate record located in the ocean) but the method for resolving the error is unclear (i.e. no usable locality string and not a common data entry error). While still other errors are difficult to detect and need special tools or tests, and once detected can be resolvable or not. Here we will begin to divide geospatial fitness-for-use issues based on their origins and the difficulty associated with detecting and resolving them.

2.1. Classification criteria for geospatial data issues

In order to assess the geospatial fitness-for-use of biodiversity data, it is important to know the general level of error in the whole dataset and to determine whether existing errors can be detected and corrected. Each of these goals are achievable but would require deployment of error checking routines at different parts of the data publication process. Although just detecting errors is highly valuable, it is not enough. In addition, providing mechanisms to correct errors in an efficient manner is absolutely critical. Simply detecting errors and flagging them for data consumers has its uses, but it puts the burden squarely on the consumers to decide whether to spend time fixing problematic records or omit them from their record sets used in downstream applications. For many data consumers not conversant in georeferencing approaches, this can be a daunting task. Equally as daunting, however, will be designing and building approaches for network wide error detection. Resolving errors across the network will prove more difficult than detection because it requires the network facilitators and data publishers to be able to coordinate in ways that are not easily accomplished at the moment.

Two essential steps to improve fitness-for-use for geospatial data are to detect (including annotation) and then resolve mistakes. We refer to these processes as detectability and resolvability (see Appendix II, glossary of terms for formal definitions). The interplay of detectability and resolvability will determine the overall cost-benefit of attempting to address the given error type. We generalize this interplay in Figure 1 and the four internal sections. Records that cost little both to detect and to resolve fall in Section 1, the low hanging fruit of error types. Error types in Section 2 may be relatively easy to detect, but may limit us to simply flagging the records or removing them from datasets, as they are difficult to resolve. This will require both filters to detect such errors during publishing

Figure 1. The breakdown of an error's difficulty to address. An error type's difficulty can be considered interplay between the difficulty to detect the error, on the x-axis and the difficulty to resolve the error, on the y-axis. Those error types that are easy to detect and easy to resolve, Section 1, should be considered the first priority. Similarly, easy to detect, but difficult to resolve, Section 2, should be the next priority, as annotating these records quickly will be an immediate service to the community. Followed by Section 3, where value can be gained from the difficult to detect errors and finally Section 4 where a large time investment or technological advancement will be needed.



steps and more widely employed methods for sharing annotations of records to make all information available to the consumers. Error types that fall in Section 3 are beneficial to address, because although they are difficult to detect, once detected we can provide the correct data to make the record once-again useful for a greater number of research questions. Records in Section 4 may never be detected or resolved as we will discuss below, but once the majority of detectable errors have been addressed, either data consumers can more appropriately account for these remaining errors as uncertainty in their results, or errors might eventually stand out and move to Section 2 if noise from other, corrected errors is effectively removed in a bulk analysis.

2.1.1. Detectability

Detectability is a function of the way that an error causes the record to stand out in a comparison with a standard result or a dataset of otherwise correct data. An error's detectability is based on several factors including the degree of the error (i.e. wrong hemisphere or wrong trail-head), the scope of the dataset to which it is being compared (i.e. all data from the same publisher or all data from the same species), and the ratio of correct to incorrect records with which the error is to be compared. The degree of detectability can be thought of as a gradient following the x-axis of Figure 1. Certain values which are out of strict boundaries (such as 90 degrees for latitude) or wrong content type (string instead of number) are easy to detect as errors and fall in Section 1 or 2. Coordinates without decimal figures can be a hard to detect type of error if it refers to a single record, but easy to detect if a large collection of records with similar characteristics is given. An error in the numeric value of a single record, especially without other associated information, can be a hard or even impossible to detect (Figure 1, Section 4).

2.1.2. Resolvability

Resolvability can be defined as the difficulty in trying to assign the correct values to a record with a known error. A main goal of fitness-for-use improvement initiatives should be the development of tools to increase the rate of resolving georeferencing errors, and thus providing the highest quality, vetted records to the community. This includes the greater adoption of documentation and exchange protocols and the development of technologies for the data publishers to enact changes at the source. Resolving some errors will be relatively trivial. For example, the substitution of "S" and "W" characters in a record with "N", "S", "E" or "W" values by a negative modifier or the reconciliation of swapped coordinates should be easy to solve tasks, both in manual or automated batch processes (falling to the lower end of the y-axis of Figure 1). Coordinates given in UTM, military or grid systems will require more calculations, so they are harder to resolve. Incomplete information or "default" values in some systems are errors which are impossible to solve without going back to the original source or data. Only detected errors can be resolved, so resolvability always relies on detectability.

2.2. Common geospatial data issues

For the purposes of this discussion, we have generalized the life stages of a primary biodiversity datum from the nine stages presented in Chapman (2005) into four primary stages: Collection, Digitization and Documentation, Mobilization, and Utilization. Using knowledge of the stage of an error type's origin, we can make a first assessment of the errors and our ability to address them to better document fitness-for-use of the geospatial data.

2.2.1. Issues at time of collection - detectable or not, not resolvable

Collection is the first moment of a datum's existence, and errors can easily occur here. Some of the more serious errors introduced at the time of collection include log mistakes, poorly documented uncertainty, errors derived from malfunction of electronic devices, etc. Errors introduced at the time of collection can lead to serious misinterpretations and while sometimes detectable, they generally are not resolvable. Detectability is improved when multiple sources of geographic information are compiled at the time of collection. For example, it is of great value to take both GPS readings and textual locality descriptions. Although these two sources might not fully resolve to the same exact location, large discrepancies should be relatively easy to detect (Hill et al. 2009) and can potentially be resolved (e.g. values from a faulty GPS could be demoted and coordinates based on the textual description promoted as the main source of information for a particular record). In addition, records that lack coordinate uncertainty and precision information may remain useless for a large proportion of future research, while multiple sources of geographic information may allow us to measure these data retrospectively.

2.2.2. Issues at time of digitization and documentation - detectable or not, resolvable referring to the original

The second stage when an error can appear is at the digitization and further documentation of data. An error when transcribing the record or an uncaught OCR error can be added to a specimen record. In other cases, a record may have no georeferencing information but if the data entry system requires such information, the program may insert a default, perhaps absurd value (e.g. 9999) but which is not null. While these programs have mechanisms to ignore such values and it has no further importance when working locally, this can lead to trouble when data are made accessible and are part of a distributed network (see further in 2.3.3.). Documentation includes all information added to or abstracted from a record once it joins a collection. This can include newly added information, as is the case with retrospectively georeferenced records. Errors of these types may be detectable or not and if detected can be corrected if the original source of the data can be consulted.

2.2.2.1. Issues at time of local or collaborative georeferencing - detectable or not, often resolvable

Recent advances have changed the way we assess and treat geospatial data quality. The introduction of both Global Positioning Systems (GPS) into field biology collection practices and Geographic Information Systems (GIS) technology for mapping have been two such advances that have revolutionized our ability to examine biodiversity patterns. Primarily, these technologies increased the geographic resolution at which biodiversity data could be standardized, shared, and studied.

As a result, these technological developments have dramatically affected the workflows and processes of natural history collections and other organizations that collect and distribute point occurrence biodiversity data.

Mappable coordinate data associated with specimen records is essential. However, out of the estimate of 1.23 to 2.81 billion natural history records in museums around the world (Ariño, 2010), an analysis of data already indexed by GBIF show that 47% might not have computer-mappable coordinates (e.g. latitude and longitude). Instead, collectors provide text descriptions of their localities. A major endeavour has been to develop the means to convert these descriptions into coordinates in a consistent format similar to output from a GPS, the process known as retrospective georeferencing. Retrospective georeferencing, the generation or regeneration of latitude, longitude and uncertainty from textual locality information (see Chapman and Wieczorek, 2006), has been one of the most widely applied data enhancement methods for biodiversity data.

Although retrospective georeferencing best practices and tools (e.g. GEOLocate and BioGeomancer) have been developed, the often uncoordinated processing of effectively new geospatial data from textual descriptions has inadvertently introduced numerous types of errors across the GBIF network. The errors that can be introduced to a biodiversity record at this point involve many of the issues presented above. It is critical that georeferencing capture precision and accuracy in replicable, standardized ways (Beaman et al., 2004). These two issues, georeferences lacking uncertainty and non-standardized georeferencing methods, still remain problematic and continue to impact the quality of data in global biodiversity repositories.

Network-wide collaborative georeferencing can help address both of these issues (see below). Georeferencing has become rapid (Wieczorek et al., 2004; Stein and Wieczorek, 2004). If broadly applied as an integrated tool for IPT users or other data publishers that support methods for annotating and reporting corrections to their records, standardized georeferencing will remain one of the absolute best ways to increase geospatial fitness-for-use. Retrospective georeferencing increases the number of error types that are both detectable and resolvable. If documented properly (including method and uncertainty), any problems that arise at this point can be quickly addressed as gazetteers become more complete in the future. The need to continue the support and development of these tools is hopefully apparent within the community.

2.2.3. Issues in the distributed network - in general, detectable and automatically resolvable

Some errors do not appear when the data are used locally, but only appear when the dataset becomes part of a distributed data network. In these cases, errors come from poor or lacking use of data interchange standards. Different ways of coding the hemisphere value, or the use of different units to store information provide means to easily detect outliers. In many cases, these errors are easy to detect and are able to be corrected. However, this group also encompasses some of the harder and more widespread geospatial issues that exist in GBIF, including ineffective UTM conversions. The widespread nature

of this issue has been documented elsewhere (see http://biodivertido.blogspot.com/2009/02/grid-data-shared-as-point-data-errors.html).

2.2.4. Issues at the time of use - detectable and resolvable

Although many errors can come about at the time of use, our primary concern is with the use of existing errors. Some examples are: records used irrespective of undocumented uncertainty; scale of study not matching scale of record uncertainty; or documented geospatial issues not properly addressed. Although these errors are detectable and avoidable, resolving them once the results of the work have been made public, published or otherwise, becomes very difficult. By introducing more comprehensive documentation of known geospatial fitness-for-use issues we hope that this final group can be more easily avoided. In addition to increased documentation, promotion of the existing documentation, further publication, and increased training should all be considered to deal with a unified community approach to fitness-for-use issues prior to use in research.

3. The GBIF filter - challenges and solutions

GBIF was established in 2001 with this general objective: to make biodiversity data freely available. As of December, 2009, GBIF's indices facilitate the sharing of information for nearly 190 million primary biodiversity records all over the world, making it the world's largest initiative of this kind (Lane, 2003; Yesson & Brewer, 2007). Thus, it is reasonable to propose using this source as a sample of the world's available primary biodiversity data (Ariño & Otegui, 2008), and as a proxy for the quality of the available global geospatial information. Geospatial components are some of the main parts of primary biodiversity data. Here, we attempt to estimate the general fitness-for-use of the world's available biodiversity geospatial data by assessing completeness and accuracy of the data contained in GBIF's indexes.

3.1. GBIF's filter and validation tools, operations and benefits

The GBIF index makes available a number of fields related to the geospatial components of the biodiversity records, such as:

- latitude and longitude of the record in decimal degree format,
- the degree of certainty of those coordinates,
- the altitude and/or depth of the sample,
- the country of provenance of the record,
- other administrative units, textual descriptions, and metadata components of the occurrence.

As Chapman (2003) says, errors are known and indeed expected. With this fact in mind, the infrastructure program of GBIF developed a filter to distinguish what information appears to be correct, and what is not. Depending on the fulfilment of several conditions and the degree of 'incorrectness', the GBIF filter determines whether the geospatial

Issue	# records	% of issued	% of total
{No issue}	182276324	--	96.20%
Latitude probably neglected	102702	1.43%	0.05%
Longitude probably neglected	249780	3.47%	0.13%
Latitude and longitude probably transposed	582850	8.10%	0.31%
Coordinates supplied as 0.0 , 0.0	2421605	33.66%	1.28%
Supplied coordinates out of range	206559	2.87%	0.11%
Coordinates fall outside specified country	3915635	54.42%	2.07%
Supplied altitude out of range	277768	3.86%	0.15%
Altitude value suspect	3314	0.05%	<0.01%
Minimum and maximum altitude reversed	13871	0.19%	0.01%
Supplied depth out of range	69	<0.01%	<0.01%
Minimum and maximum depth reversed	26297	0.37%	0.01%
Total issued records	7194999	--	--
Total records	189471323	--	--

Table 1. Overview of error types and quantities issued by the GBIF filter as of December, 2009. The most common combinations are as follows: 1. Latitude and longitude probably transposed + latitude negated = 48964 records. 2. Coordinates fall outside specified country + Latitude and longitude probably transposed = 482831 records. Note: since a record may be affected by more than one geospatial issue, the sum of all issued records is lower than the sum of the differently issued records.

information of a record is included or omitted when a user access the GBIF Data Portal. While some mistakes only elicit an annotation in the database, others prompt the filter to block the disclosure of geospatial information. If the information is clearly wrong in a record (e.g. UTM coordinates where decimal degrees were expected), the filter omits the geospatial information. If the record contains a possible but uncertain mistake (e.g. records with potentially inverted coordinates), the filter makes an annotation available in the "geospatial issues" field. GBIF does not alter the data shared by publishers, even if an easy-to-resolve issue is detected, mainly due to the fact that what GBIF intends to do is to encourage data providers to actively manage their data (A. Hahn, pers. comm.).
By the time of analysis, approximately 7.2 million occurrence records (3.80% of the total amount) presented at least one kind of geospatial issue. Within those, records were annotated due to 11 different reasons. Table 1 lists those issues, accompanied by the absolute and relative amount of records flagged.

Although in many ways useful, the filter is the source of two competing concerns. On the one hand, some errors, mainly those concerning possible but uncertain mistakes, do pass through the filter. In these cases, the presence of the filter may provide false comfort in the quality of the data. On the other hand, there are some erroneous records whose data are blocked by the filter but could be corrected with a few simple calculations. In these cases, the filter has removed potentially useful data while not ensuring that the errors will be fixed in the future. In the next section, we outline some of the most common mistakes - both those which are blocked by the filters, and those which are not.

3.2. Challenges associated with enhancing fitness-for-use

3.2.1. Types of errors filtered

By querying the raw and filtered databases, we have observed that 2.26% of the records indexed by GBIF (approximately 4.27 million out of 189 million at the time of the analysis) contain geospatial information that is blocked by the filters. The distribution, type, and source of these errors is highly uneven. Two data resources produce 61% of the filtered records (Figure 2) and in the vast majority of the cases a data resource contains only one type of error (e.g. all blocked records from a single provider will have swapped latitude and longitude fields). By targeting these high error producing resources, limited cleaning efforts may result in significant improvement of the overall data quality.

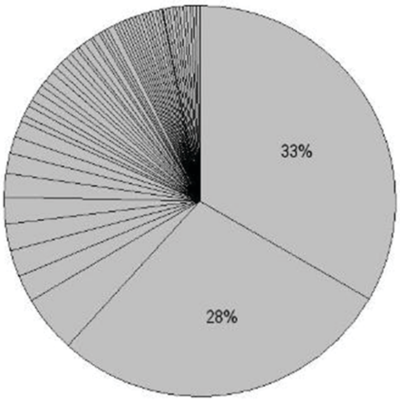


Figure 2: The total filtered-out records split by the number of records filtered from each particular resource in the network. The total amount of filtered records is 4.27 million. 61% of the blocked records belong to two data resources.

Most of the errors detected and blocked by the filter are both easily detectable and easily resolvable, except for those records where geospatial information has been inserted "by default" (as discussed in Section 2.2.3). Here are six common errors found and discussion about how each can be dealt with using automated services when possible:

Incomplete coordinates

Many records have only one of the two coordinate fields completed, while the other is empty. This makes it impossible to figure out the actual value. The only way to solve this issue is going back to the original source of the data. There are two ways to approach this sort of error. (1) The presence of an error should deem the entire record suspect, and thus should not be provided through the GBIF Data Portal. (2) A clear annotation of the error should be included with the record, while allowing the user to judge the usefulness of the record. In some cases, latitude alone or longitude alone may still be useful for scientific studies.

String in a number field

If the filters detect a string character, a record is blocked. To ensure predictable numerical coordinates for external services, this step is necessary. However, the detection of a string in the numeric coordinates does not always imply a significant issue. Consider the case of the cardinal direction strings 'N', 'S', 'E' or 'W'. The detection of these strings is simple, only requiring that we look for those characters in all coordinate fields. Once detected, the 'S' and 'W' can be converted into negative multipliers and the 'N' and 'E' values can be eliminated. This approach must also be

combined with a method for detecting a swapped latitude and longitude field. In fact, the detection of a ‘N’ or ‘S’ in the longitude field or the ‘W’ or ‘E’ in the latitude field can represent a very simple method for detection of swapped latitude-longitude fields. The combination of detection and resolution could be automatic in both of the above cases (e.g. a controlled dictionary of detected strings and swapped latitude-longitude combined with the presence of cardinal direction string).

Wrong coordinate system

Often, the issue of data providers using the wrong coordinate system is compounded by the above issue of string detection. It is relatively common to find coordinates in the DMS (Degree Minute Second) system, with values separated by either two dots, white spaces, colons, or the degree, minute, and second letter symbols. Sometimes, large sets have 6 figures in the latitude and 7 in the longitude. These can also be DMS coordinates that lack any separation between the elements. These errors could be easily detected using the small predictable set of patterns that signify DMS coordinates and resolved by recalculating the coordinates using decimal degree system. Other issues can arise from the reporting of UTM coordinates where conversions can often be difficult using available automated services.

Coordinate swap

It is also common to find swapped coordinates (i.e. latitude in the longitude field and longitude in the latitude field). These errors contain a mixture of easily detectable and difficult or impossible to detect cases. When the absolute value of the latitude is higher than 90 and the absolute value of the longitude is lower than 90, it is highly likely that a swap has occurred. However, in this case another, less likely, possibility is of a typing error in the latitude field. If both values remain lower than 90, a swap can only be revealed by ancillary checks, such as consulting a gazetteer to see if the coordinates point to the same country that the “country” field indicates. A first pass assessment of records (as already done by the GBIF filter) to formalize a hypothesis of a swap occurring could be further substantiated using ERM methods (as discussed in Section 6) to make a final, high quality assessment.

Numerical sign confusion

Another common mistake is the omission of a negative symbol (‘-’) in either or both of the coordinates (as noted also in Chapman, 2005). An exhaustive test to detect and resolve this problem may consist of checking all possibilities (latitude/longitude, -latitude/longitude, latitude/-longitude, -latitude/-longitude) against a gazetteer and see which one fits the “country” or lower geographic description field. In the same way as discussed in the Coordinate swap section, a first pass could be made using the ‘country’ field to formulate a hypothesis as to the source of the error (i.e. sign confusion on the latitude field). Following this first pass, the hypothesis could be tested using expert opinion range (ERM) methods discussed more fully below to determine if by correcting the suspected error the occurrence would then also be found within the accepted ERM of the species.

Out-of-map records

Some records have absurd coordinate values (e.g. any coordinate higher than 180). Here, there is a large set of possible sources of the error, although they can be reduced into three general categories: the batch entry of default values, a decimal symbol misplaced or deleted, or the coordinates in DMS with no separation among the component fields (discussed above). As we have pointed out, detecting records from the batch entry of default values is easy, while resolving them may be very difficult because original information gets masked by the default value. Fortunately, this is the least abundant of the three categories. Many more records exist in the second category, containing values of latitude and longitude above 180 but below 1000. This can be taken to indicate that the decimal symbol has moved one position to the right. This hypothesis could first by tested by dividing the coordinates by 10 and comparing to the result with information available from the country field. Again, ERM methods could further support the hypothesis.

3.2.2. Types of errors not filtered

The remaining 97.74% of records pass through the filter, indexed directly from the data publishers without any change. Nonetheless, this does not mean that those records are free from errors. Unfiltered records may have minor errors (generally in the categories discussed above) or ambiguous situations in which data may or may not be true. The filter labels 16.54% of records with an annotation in the “geospatial_issue” field that indicates the possible presence of an error. This field contains different values for different geospatial issues the records may have. These issues can be taken as suggestions to prompt action. A detailed list of the different issues follows:

- Latitude probably negated
- Longitude probably negated
- Latitude and longitude probably transposed
- Coordinates fall outside specified country
- Supplied altitude out of range
- Minimum and maximum altitude reversed
- Supplied depth out of range
- Minimum and maximum depth reversed

Still other errors are not detected by the filter largely because the filter applies to individual records while some issues only appear when dealing with large amounts of data. A global assessment of the geospatial components of biodiversity data can help us begin to detect these errors automatically. One of the best known issues is the uneven distribution of available biodiversity information among the countries, and the biases in the global patterns of biodiversity. Biodiversity data are disproportionately made available by institutions in developed countries (Ariño, pers. comm.) and leads to an appearance that biodiversity is affected by political boundaries and economic patterns, as shown Figure 3. This type of problem can be addressed by developing species and geographic based assessment of completeness available to GBIF users (Hill et al., in prep). We will discuss this further in Section 7.

Another not so well known issue is the omission of decimal figures in the coordinates of some datasets. When taken alone, these records may seem normal, but when there are large amounts of them, the spatial representations looks like a uniform matrix of equidistant high-density dots (Otegui et al., 2009)(Figure 4)

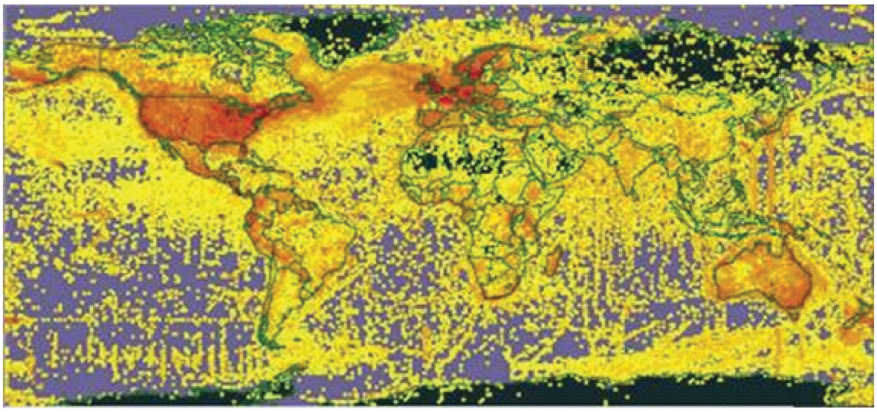


Figure 3: Spatial representation of all the available biodiversity data in GBIF (taken from the GBIF data portal, http://data.gbif.org). The data represented may lead to the incorrect conclusion that biodiversity is concentrated in North America and Europe.

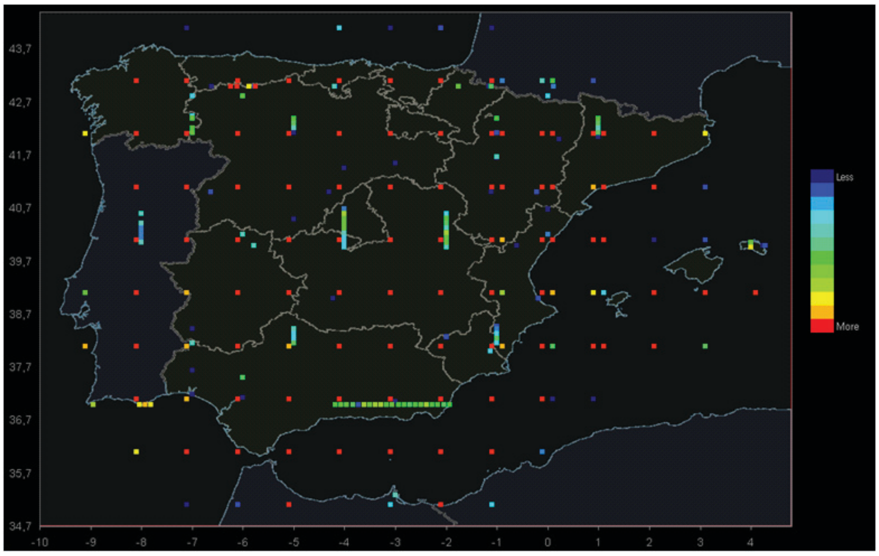


Figure 4: The spatial representation of a single resource of a Spanish data publisher. Most of the data forms a uniform high-density dot matrix (taken from Ariño & Otegui, in prep.) that does not resemble accurate spatial information.

3.3. Scope of Methods

According to the classification above, almost all described kinds of error arise at the time of digitization or in the distributed network. This means that they are mostly resolvable. It is quite easy, for example, to un-swap some coordinates, or to make an "S" become a "-" in latitude fields. Nevertheless, even if the records have passed one or more filters, the information they contain may still be erroneous. Since these tools only deal with minor transformations in data, if the original record is incorrect, the parsed value will still be incorrect. They have nothing to do with issues at the time of collection (there is no transformation which may turn right the wrong value).

It is not the aim of these methods to provide certainty of the correction and usability of the records. Instead, their aim is to reduce the incidence of some common mistakes that may happen during data manipulation, thus improving the potential usability of the records. The parsing of a record may lead to a potentially correct value in its geospatial information, but with these methods it is not possible to undoubtedly state that the processed record is correct.

In order to know if a piece of information is actually correct, there is no choice but to ask the information owner. The data owners are the ones who hold the master record and, thus, know if a record is right or wrong. This is in-line with the current implementation of the GBIF filter: data publishers should be aware of their data quality and involved in the improvement processes. We encourage data publishers to care about their data and keep them updated. For those reasons it remains ever important that no method overwrites original data, but instead offers alternate versions for consumers and ultimately data publishers to review.

4. Darwin Core and the Integrated Publishing Toolkit

The Darwin Core (DwC) body of standards provide the community a common format for sharing biodiversity data and information (see http://rs.tdwg.org/dwc/index.htm). It is not a transmission mechanism. However, DwC provides a standardized framework which can be used to address fitness for use enhancement. Here we provide a brief overview of geospatial information and the DwC standards and then discuss the annotation schema developed with the Integrated Publishing Toolkit (IPT; http://gbif-annotation-processor.googlecode.com/svn/trunk/api/annotation.xsd)

Darwin Core provides a set of fields that are used to share information about the sampling, identification and history of a published record (including date, location, and methods fields). For example, the locality elements within the DwC record (http://rs.tdwg.org/dwc/terms/index.htm#dcterms:Location) represent valuable information for assessing geospatial fitness for use (e.g. through recent retrospective georeferencing projects such as BioGeomancer and GEOLocate. As discussed above, the Country field alone will be important for the validation of suspected errors and their proposed resolution. Event information (http://rs.tdwg.org/dwc/terms/index.htm#Event) can provide temporal information that should preclude the use of any temporally insensitive gazetteers. Some elements of DwC record have yet to be fully realized in regards to their effect on fitness-for-use. For example, records from systematic surveys may have different measures of fitness-for-use than those recorded using ad-hoc methods. Although many records contain this information, GBIF only provides a small subset of DwC terms through the data portal.

Errors in digital records often precede the digitization of the record itself (discussed above), either arising from an error in recorded locality or imprecision of locality (Graham et al., 2004). As Graham and colleagues also correctly point out, these errors can often be quickly detected as outliers in a dataset, and moreover, can be corrected by referencing the details of the specimen and/or accompanying notes. This process is currently limited by the rate at which one can collect data from the GBIF data portal and the ability to collect complete records. While GBIF makes links to the original record available, a researcher is often faced with, 'We are sorry, but the requested data is not available at this moment due to connectivity problems with the provider'. Either a more complete availability of DwC terms directly through the GBIF data portal, better solutions for providers to mirror their data on external servers (see Cloud Computing below), or both can improve this situation in the near future.

While better methods for detected and resolving errors by processing DwC records will lead to better and more complete scientific analyses, the need for a means of two way communication of these corrected or annotated data still remains. A service tasked with annotating or improving fitness-for-use has very few options for reporting results to the data publishers. One possibility would be to create a new version of the complete DwC record containing needed changes and find methods for delivering these data back to the publishers. On large scale projects these solutions could represent a sizable undertaking and management issue. While we need to enable the annotation and correction of errors in DwC records, it is critical that original data should never be replaced, but only explicitly modified versions that allow roll-back and future comparison of changes

Currently, IPT has its own annotation schema (http://gbif-annotation-processor.googlecode.com/svn/trunk/api/annotation.xsd) that represents an important contribution to the fitness-for-use enhancement design. Like DwC and DwC-extensions, the annotation schema can currently be used as XML or easily shared as Fielded Text. This ability makes annotations easily

imported into tables and quickly, or as quick as the network allows, resolvable to the original records. The annotation schema could support all components of fitness-for-use enhancement discussed up to this point. Error detection and resolution services can operate on the original DwC records and provide annotated information back to the publishers in the standardized format. Next, data publishers can quickly skim records for the types of errors, any proposed resolutions and associated probabilities, and decide the priority of action. In the cases of providers whose data have primarily one error type that afflicts many records (see Section 3), results for records containing that single error type could quickly be parsed out and all resolutions enacted at once. For further discussion of the development and promotion of the annotation schema see Section 7.

The rate at which we can address geospatial fitness-for-use remains a primary concern. In their stand-alone state, most approaches to assessing fitness-for-use still require time investments beyond what is available to most data publishers. Even batch processing by data publishers can still be repetitive. As well, automated assessments run by external sources (e.g. the georeferencing workflow project BioGeobif, see below) require time consuming data harvesting as well as delivery to and reintegration of results at the source by the data publisher. Fully utilizing the elements of DwC in addition to the added IPT technologies may help decrease the time and repetitiveness of many data cleaning and fitness-for-use annotation methods. While adoption still is, and may remain, low in the community, the solutions it brings forth can be extended. For example, services for exchanging the annotation schema or other data correction solutions at any node within the network could be conceived such that the IPT was not a requirement. Such solutions will help us standardize our methods of communicating known errors while maintaining the integrity of the original data.

In addition to adopting new technologies, there exists a need for extended training. The data publishers, whether curators of the original data or organizations providing a middle publication layer, may not always be aware of the magnitude of the problem and the urgency of better standardized data. As standards and parties interested in providing data improvement services already exist, the education of community members in charge of the data will be essential. We believe this training mission is critical for the future success of biodiversity networks (Guralnick and Constable, 2010). Such training needs to not only include best practices for georeferencing, which GBIF has supported in the past. Training must also include building technical expertise to use workflow tools to incorporate data back into source databases.

5. Workflows and automated pipelines

To handle the scale of fitness-for-use annotations that is still required, novel services, automated pipelines, and better mechanisms for joining services and publishers with one-another will be necessary. In the past, we have argued that the building of automated pipelines is necessary for both assessing global biodiversity data (Hill and Guralnick, 2008; Guralnick et al., 2007) and for the improvement of particular fields of data within our biodiversity databases (Hill et al., 2009; Guralnick and Hill, 2009). By combining automated error detection services and error resolution services we can make robust data quality improvement tools linked to the global biodiversity publishing network. Technologies like the annotation schema will help us get the fitness-for-use information back to the owners of the data and decrease the number of times a record field will be vetted and the amount of time before errors are corrected at their source.

One such project that attempted to achieve some of this was entitled BioGeobif: a proof-of-concept that utilized a “harvesting” model in order to create georeferencing pipelines and lift some of the burden off of data publishers. Harvesting can either target the GBIF API or the data publisher’s protocols (DiGIR and TAPIR) directly. BioGeobif was designed to link together all parts of a workflow from data harvesting to data georeferencing and reporting to data publishers in formats that encouraged rapid incorporation of improved data (Hill et al., 2009). In cases where very high quality information could be easily gained from georeferencing (i.e. one highly likely coordinate pair and associated coordinate uncertainty) or where publisher reported coordinates were significantly different than those that could be found using georeferencing, BioGeobif was a promising automated solution for annotating and enhancing fitness-for-use.

The success of the above model was impeded by two solvable problems. First, delivering records to the data publishers can be complicated. Currently, records enhancements and annotations need to be stored on a site where data publishers can choose to download and ultimately reincorporate them into locally curated databases. This is an inefficient process, where publishers may remain unaware of

available fitness-for-use improvement services or may remain unable to commit the time needed to download and incorporate the externally generated results. In either case this leads to results going unused. Second, there are only a limited number of methods for automated services to access biodiversity data. A service can either access the records through the GBIF API or directly harvest records from a publisher's server. In both cases, record transmission is not as fast as it could be and the process can be complicated by different protocols. These problems can in part be solved by adopting modified storage and access methods for sharing data across the distributed network.

GBIF has already taken some of the necessary steps to improve result delivery to data publishers. With the increased adoption of the IPT or methods of exchanging the annotation schema independent of the IPT, the ability to report annotations to data providers may become a much easier process (see Section 4). Provided that the IPT lowers the cost of delivering results to data providers, one could also consider this a valuable method for providing annotations to data consumers that have not been fully incorporated into the source by the data provider. This could be done by including annotations from trusted services for download by data consumers, as XML or as fielded text, through the GBIF data portal and through a page in the IPT service. In fact, wide adoption of the annotations schema already being developed may enable advancements like queuing record improvement suggestions requiring action at the source. We will discuss some of these issues in Section 7.

The second problem we mentioned above is the inefficiency of aggregating complete records from across the network in order to be able to analyze and increase geospatial fitness-for-use. In short, services need to be able to perform a variety of analyses based on combinations of one to many data resources, species, or geographic regions while accessing many different components of the original DwC record. Furthermore, these processes would need to be repeated by any service that wishes to remain current with the existing data. Such a system might require a service to harvest many different data publishers very rapidly. These issues bring forward two primary needs: more complete representation of the original DwC fields and easier access to large user defined subsets of the available data. We understand the value of a fully distributed system. However, consolidation of the full set of published data in the cloud might be the easiest way to simplify this process while maintaining publisher control of their local databases.

The recent VertNet design prototype contains several exciting components that can provide design specifications for a provider in the cloud model (see Constable et al. 2010 for a full discussion). VertNet is taking a radical step away from the current data publishing models employed elsewhere; instead of publishers making digital records available at home institutions via local server hardware and relying on the central portals to re-access those records periodically, VertNet has proposed a data publishing model that adopts a cloud computing approach. In this model, contributors publish their dataset to the cloud utilizing existing common data standards. Publishers sync their local databases to the cloud store through mediator software so that any updates locally can be then propagated to the centralized cloud data store. Second, VertNet is more actively designing an API to enhance services that improve or document the quality of records published by VertNet participants. This gives services rapid access to records (via cloud-based connections) and methods for suggesting record annotations. These methods provide services with a higher likelihood of effecting rapid change. Although this solution may not be an exact fit for GBIF, the IPT may represent the key step for GBIF to move in a similar direction to VertNet as discussed in Section 7.

6. Effects and implications of non-point based geospatial information

To date, there are very few ways to assess geospatial fitness-for-use based on the reasonable extents of where an occurrence is expected to have originated. GBIF's development of the geospatial_issues field addresses fitness-for-use in non-biotic contexts (e.g. does a record's coordinates exist within the country it is reported to exist?). In the same way, BioGeomancer and other retrospective georeferencing tools can be used to assess a record's coordinates at greater precision by assigning latitude, longitude and uncertainty to more detailed locality descriptions. Importantly, neither of these methods take into account information that can be derived from the taxonomic components of the record, such as, where a species is known to exist or capable of existing. The technology and information necessary to perform such a study is becoming available. Here we discuss two further methods that hold high potential for the rapid evaluation of geospatial fitness-for-use: Expert

Opinion Range Maps (ERMs) and Species Distribution Models (SDMs).

6.1. Expert opinion range maps

Expert opinion range maps (ERMs) represent coarse spatio-temporal representations of where species occur (Hulbert and Jetz, 2007). Unlike occurrence points, it is accepted that ERMs contain errors of presence and over-generalize the occurrence of species thus rendering only broad-scale representations of species ranges. Until recently, there have been very few studies that have tried to evaluate the resolution(s) at which individual ERMs are appropriate for use (Lawes and Piper, 1998; Hulbert and Jetz, 2007). Additionally, the availability of ERMs has been generally limited to a small proportion of the entire world's species. Ongoing projects now seek to automate the processes of ERM generation and evaluation and also provide results to the broader community (see Map of Life Appendix II). Given that ERMs provide broad representations of species ranges, they could be used to quickly discover occurrence points that are highly likely to contain geospatial errors. This technology will not be ready immediately, but several groups are working to enhance the quality and reliability of existing ERM data. Unlike methods derived from purely geospatial information (e.g. comparing occurrence points to country of origin), a method that compares species occurrence data to ERMs allows us to start evaluating the likelihood of a record's accuracy utilizing biological information.

Applying ERM information will allow the community to incorporate a new set of processes for automated error detection based on taxonomically defined datasets. In the past, researchers have compared occurrence points for a given species to the ERM for the species and defined all outliers as errors (sensu Yesson et al., 2007). Such methods, although powerful (as discussed below) are hindered by the temporal scope of ERM information (i.e. a single ERM can only be used to evaluate records over a limited temporal window) and by the largely undocumented quality of ERM data. Despite these limitations, ERMs are likely the best estimate of distribution at coarse scales and over the last 25-50 years (the same time-frame when many of the records from GBIF were collected - see Figure 5) for many species.

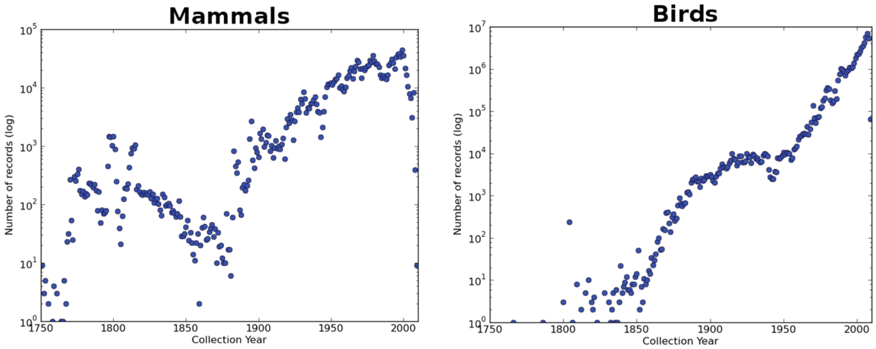


Figure 5. The distribution of record date-of-collection (log count) as reported to GBIF for Mammals (a) and Birds (b). In both cases, the overall number of records reported in the past 20 years is very high: 30% in Mammals and 88% in Birds.

Some of the authors and other researchers are actively developing methods that in the future will be transformed into services that GBIF and member institutions could utilize for geospatial record validation. Geospatial fitness-for-use may be greatly benefited by ERM methods through, (1) classifications of datasets, i.e. highest quality, potentially useful, potentially problematic, and containing known errors; (2) prioritize different datasets based on general estimates of the number of resolvable records versus the time investment needed to assess those records, and; (3) further validate automated data cleaning methods discussed in Section 3. Here we begin to outline one possible assessment method based on current GBIF occurrence points and ERMs published through the IUCN.

6.1.1. Fitness-for-use and scale based fitness-of-use

Overlaying an expert opinion range map and species occurrences for a taxon will lead to records either falling wholly or partially within the expert range map or wholly outside that map. We can use this information to begin classifying data and its geospatial fitness for use. The first grouping would be those records that fall entirely (including the extents of the record's uncertainty in meters) within the boundaries of the ERM polygon (Inset 1). Although not guaranteed to be without error, these represent a subset of records that could be generally trusted by consumers. For those records that do fall outside range boundaries the next step is to classify the scale of the uncertainty.

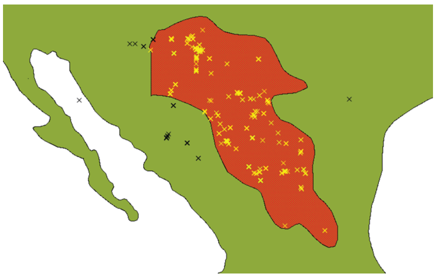
Here we present an example of how this could be done using a nested grid system developed previously.

The grids are made up three different cell sizes: 110km X 110Km grid cells (Inset 1a,b), 220km x 220km grid cells (Inset 1c) and 440km x 440km grid cells (Inset 1d). Now, excluding records within the range polygon, records that are found in a grid cell from the highest resolution (110km x 110km) and overlapping the ERM are classified as reliable to the 110km resolution. These records may contain geospatial errors, but before either they are corrected or the ERM is modified, they could be useful in biological studies where the absolute geographic position of the occurrence may not be necessary. As well, these records may be further vetted to determine if they are possibly accurate but represent a new report outside existing expert boundaries. Further examination of the georeferencing process may provide assessment of geospatial record quality. As well, an assessment of the habitat quality at the collection location could provide ancillary information about such records. Those records in suitable habitat but outside range boundaries might be more likely to be considered "range extension" records.

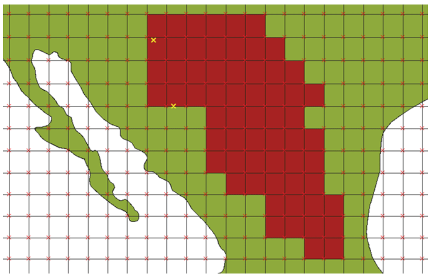
The previous step is repeated at ever larger scales (Inset 1c,d) giving the users of the records additional information about the minimum expected geographic error that is contained in a particular record. Such a method would allow researchers to still take advantage of the most information available as long as the data matches the scale of the study. We present the complete hypothetical workflow in Inset 1 below.

Inset 1

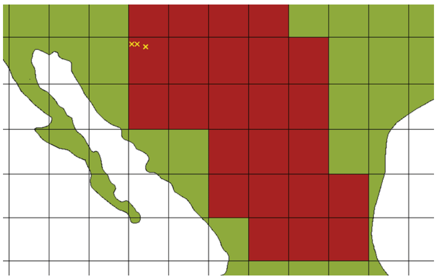
A classification system could be used to (1) provide geospatial fitness-for-use information for a majority of records available and (2) prioritize non-automated resources needed to address the possible errors. In Figure 1 we presented the interplay of detectability and resolvability when discussing errors in geospatial information, here we propose using this knowledge



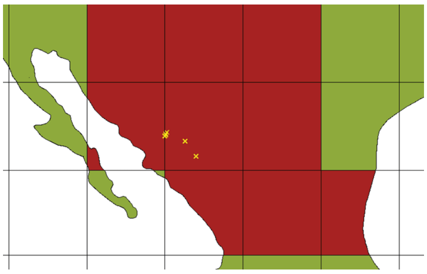
First, occurrence points for *Chaetodipus eremicus* available from GBIF data publishers are projected along with the currently accepted ERM (IUCN; Patterson et al., 2007). All records within the ERM boundaries are highlighted in yellow (right). Those records fall into Class I: the high quality datasets are considered to generally contain highly reasonable coordinates not requiring immediate evaluation beyond automated methods.



Next, the evaluation is repeated at a lower resolution. Here we use an approximately 110km x 110km grid. Grid cells that overlap the boundaries of the ERM create a lower resolution buffer around the ERM. Occurrence points highlighted in yellow are now those records that fall within this buffer. These records represent Class II: occurrence points that may be true but will likely need human evaluation to adjust the occurrence point or ERM properly. These records may or may not be useful in scientific study, but should be at least closely vetted by the user. In addition, these records may be highly useful for improving the ERMs, as they may represent unaccounted for species range limits. Ideally, an equidistant buffer could be drawn around the ERM, but here we present the computationally much simpler approach.



Those steps are then repeated at a second, even coarser resolution. Here we chose an approximately 220km x 220km grid. Again, occurrence points that fall with the grid cells forming the buffer are highlighted in yellow. These records represent Class III: occurrence points significantly distant from the accepted ERM boundaries. This scale could be chosen to generally detect records that should be left out of scientific studies that need higher resolution information. But again, these records could prove highly valuable for updating the accepted ERM and known species range extents.



Lastly, if only two stages of three stages of resolution are used to evaluate the records, the remaining records highlighted in yellow fall into Class IV: occurrence points that should be left out of current scientific studies unless at very low spatial resolution (e.g. global analyses) and only after further vetting of the record. Unless these records prove to be undocumented extensions to the ERM, these records need further vetting to determine if errors in the records can be resolved.

to prioritize the records we address manually.

6.1.2. Discussion

The steps provided above are a simplified conceptual implementation of how automated geospatial fitness-for-use annotation could be performed using ERM information. Several important factors need to be accounted for prior to undertaking such a method:

- The temporal nature of both the occurrence points and the ERM need to be well documented. Comparing occurrence points from outside the temporal coverage of an ERM could lead to poor inferences about GBIF geospatial data quality. ERM layers, generally describing species ranges as they are today, provide a promising starting set of information. Of mammal records shared through GBIF, roughly 30 percent have been collected within the past 20 years (Figure 5a) and in avian records the figure is closer to 90 percent (Figure 5b).
- Comparing occurrence records to ERMs requires that geospatial uncertainties from both types of data are available. Another approach would be to join the circle-radius of uncertainty with the ERM providing three different outcomes (circle fully within range polygon, circle intersects range polygon partially, circle does not intersect range polygon at all). ERMs have spatial uncertainty as well, depending on scale of the range, digitization of the range, etc. The method presented above, generating multiple staggered mesh grid cells, is one way of addressing ERM uncertainty. An alternative method would be to expand the ERM polygon equally on all sides and documenting the scale at which such "coarsening" leads to occurrences falling within the range polygon. We decided to not present such a method in favour of the computationally simpler approach and feel that as GBIF moves towards billions of records, computation will be a necessary consideration. Additionally, once a calculation of which grid cell contains a particular a record is performed, that information can be stored regardless of future changes to the shape of the ERM minimizing spatial calculations. This is already done in the cell_id and centi_cell_id columns provided by GBIF. Although a longer discussion of this method is warranted elsewhere, we feel these are the primary considerations to develop at this point.
- The availability of ERMs is currently limited and may remain taxonomically limited for at least some time to come. Focusing on validation tools where ERMs are available should be the priority and not the development of ERMs themselves.
- Evaluations based on ERMs (and many other methods) will only be as useful as the metadata available with the original records. For example, comparing exotic or invasive species with native ERM layers will be a risk that can only be solved by combining these methods with filters discussed above and extended metadata available with each record.
- To compare ERMs to any species, advanced knowledge regarding seasonality or potentially outdated occurrence data will be needed. Methods of partitioning the data will need to be explored during the development of these comparison methods. Other solutions (e.g. cloud based datasets) may help us find ways of more automatically detecting temporal limits to species data or partitioning data and ranges by seasonal constraints (i.e. wintering, migrating, and mating ranges).

Due to the points listed above and likely other points, the use of ERMs needs to be cautiously explored as an error detection tool. While obviously possible, the solution may not be immediate.

6.2. Species distribution models

A species distribution model makes predictions about habitat suitability in unsampled geographic areas using available information from presence or presence/known-absence data and associated environmental data. The development of species distribution modelling approaches has been rapid and led to a profusion of methods and applications (Guisan and Zimmermann, 2000; Soberón and Peterson, 2004). Even with these advancements, species distribution models based on biased sampling can be misleading (Phillips et al., 2009), thus limiting their ability to provide validation for incoming data records. The key benefit of species distribution modelling in the context of geospatial fitness-for-use is matching suitability values generated from training datasets with new "test" species occurrences. New occurrences in habitats that appear highly unsuitable based on models could be flagged as requiring further examination and potentially not-fit-for-use. In some cases, records in areas modelled as unsuitable may end up being verified as high quality leading to review of modelling results The development of a system where

automated niche model generation (see LifeMapper Appendix I) can be fed back into fitness-for-use annotation remains nascent. We would advocate a system where niche modelling workflows would only access and use those species occurrences that are annotated as fit-for-use at the scale of model construction. This might limit number of models that can be produced since a minimum number of records are needed for robust results. Once niche model outputs are permanently stored, they can be used for further fitness for use assessment of new data published to the network. Species distribution modelling could be an integrated step following evaluation using expert range map information. Records that fall within a set buffer around the ERM (see Section 6.1.) but not within the accepted distribution, can be checked against known environmental and habitat conditions in their location.
In addition, GBIF could allow users to perform more advanced queries, such as randomly sampling a species' occurrence points from within its distribution model, helping to mitigate biases (e.g. Northern or Western biases) in the data.

7. Recommendations to GBIF

Our current methods for assessing and reporting geospatial fitness-for-use are built around the need to improve information on the accuracy and precision of species-occurrence data. Up to this point there have been primarily two branches of addressing geospatial fitness for use: (1) Data publishers utilize existing services and methods to address their own records (e.g. using the BioGeomancer Workbench) and are then tasked with updating their databases with this information. (2) Fitness-for-use is addressed after the record has been shared through the network with little recourse for delivering the assessments back to the data publisher (e.g. independent research projects). Here we outline a path forward for GBIF to address many of the easily detectable and resolvable issues in three different ways: (1) An extension of the methods, coverage, and scope of their own filter services. (2) Expand upon the standards and methods for external services to access data and report results to data providers. (3) Promote new methods for assessing geospatial fitness-for-use and expanding the number of issues that we can detect and resolve.

7.1. Rethinking the GBIF Data Filter

The GBIF filter is an invaluable tool for documenting a record's fitness-for-use. However, if the filter is irregular in the way it addresses the errors it becomes a source of confusion over data quality rather than a firewall against erroneous scientific conclusions. We propose that the geospatial information filter be improved through the addition of a formalized set of tests having three validation phases. First, a modification of the values of coordinates (see below) is performed on records with suspected errors based (1) on patterns seen in the record (i.e. latitude greater than 90 and longitude less than 90) and (2) on the overall probability that the error is one type over another based on the known abundance of the error type in the shared dataset (i.e. coordinate swap is checked before decimal point modification). Second, the new coordinates are tested against a gazetteer to determine if they are found in the country of the record's origin. A subset of this test has been performed by GBIF in the past, but if computing resources become available the test could be expanded to higher resolution (i.e. using more specific locality information such as state or county information contained in the record). In the future, following the second test a third test could be performed in cases where ERM information is available for the species for the same time period of the record. Using a version of the methods described in Section 6.1, GBIF could determine if the new coordinates exist within the accepted range of the species.
The first step can be further broken down into four discrete tests,

1. Swap latitude and longitude
2. Change the signs of coordinates (all three possibilities)
3. Calculate the coordinates from DMS to decimal degrees, once the degrees, minutes and seconds are parsed, and
4. Move the decimal symbol.

GBIF can treat error detection and resulting suggested improvement as a hypothesis that will either be supported through the validation method or rejected by further testing. In this way, GBIF can encourage more validation methods beyond the ones touched upon here (i.e. the comparison of Locality string information to the final coordinate) as further tests of the hypothesis. Information regarding the history of a record's improvement can already be stored in DwC, although the onus will rest on the data publishers to ensure that this data are stored at the source. Currently, GBIF handles some reporting through methods such as the resource logs (see http://data.gbif.org/datasets/resource/1023/logs/) but this shares some of the same issues discussed in Section 4.

7.2. Standardising

error reporting to data publishers

Inability to efficiently provide record annotations and error resolutions is a major concern that will limit the rate at which biodiversity records will be improved. Automated methods have been explored elsewhere (see Hill et al., 2009) but one of the major difficulties has been efficiently delivering improvements to data publishers. We see both the IPT and the annotation schema as very promising moves toward a solution. However, in respect to data cleaning practices, those technologies have been under-explored. It is imperative that our community solve the problem of returning annotations to the source of the data.

The annotation schema is one solution to some parts of the problems discussed previously. Wider promotion and education about the annotation schema is one necessary step. The next will be to more explicitly detail regarding how error detection, error resolution, or both detection and resolution can each be documented by external services using the annotation schema. The annotation schema is a move towards a system of open peer review of published datasets (see recommendation in Chapman, 2005, a recent taxonomic example in Penev et al., 2009, and arguments for the benefits in Chavan and Ingwersen, 2009) by allowing external services to assess fitness-for-use (given a score in the IPT) at the dataset level. We strongly encourage GBIF to further develop and promote the schema. As services and providers become more familiar with how to develop and draw upon annotations, we feel the annotation schema will be a major method to increase the rate fitness-for-use assessment.

7.3. Into the cloud

The biodiversity data publishing community (including GBIF and other portals) must consider hosting a cloud based, unified biodiversity dataset. Given community concerns over replication of publisher data, doing so may need to be an opt-in service for only interested publishers. However, parts of the community have already seen the benefits of cloud based data, and we believe data publishers will quickly recognize the benefit of such a change. GBIF might consider formalizing a memorandum of understanding for data publishers and other large informatics initiatives to sign. This might allow the unified biodiversity dataset to grow beyond observation and occurrence data, to range maps, niche models and other common sources of biodiversity data. A MOU would also provide a common language to discuss the community commitment to free, open, and widely accessible data.

As discussed previously (see Section 3, 5 and 6), some methods of addressing fitness-for-use will require divisions of the biodiversity record beyond Provider/Resource. For example, building pattern recognition methods to detect transposed decimal points or UTM conversions would be benefited from the largest available record set and the most information about how data are partitioned. In another case, ERM based methods will likely seek to address issues one species at a time. Although this work could be done by harvesting data from the GBIF API, this remains a time-consuming task.

While there are some drawbacks to moving biodiversity datasets to the cloud, the benefits of lowering the bar to gaining access to biodiversity data far outweigh the drawbacks. One drawback for example, would be the added maintenance need for keeping the cloud dataset up-to-date. However, solutions can be easily conceived where the process of maintaining the cloud-based dataset could simply mirror the processes which keep the GBIF cache up-to-date. Increasing the number of people applying new methods to the data will be one of the greatest benefits. Biodiversity informatics methods are often limited to those conceived or at least implemented within our community. There are many methods of error detection through machine learning (see Duda et al., 2001) that could still be applied to biodiversity records that have yet to be attempted. A wealth of knowledge about these methods exists outside of our community.

We would suggest that GBIF explore a means for making biodiversity data available using widely used cloud computing services that would lower the bar to access and analysis. The Amazon's EC2 cloud, Google AppEngine (see VertNet; Constable et al., 2010), and other widely accessible cloud services (Google's GS or the planned NSF-Microsoft Azure cloud); replicating across multiple services may become an essential future direction for making biodiversity data freely available to all. The bioinformatics community has been quicker to recognize the benefits of such a move (see Schatz 2009; Bateman and Wood, 2009). The benefits of this move within the biodiversity informatics community would be far beyond fitness-for-use purposes, but we will focus on those:

- Time-stamped snapshots of the complete biodiversity record available (e.g. GenBank's FTP

services) to reduce the time from publication to error detection
- Scalable access to computing resources allowing many automated data cleaning and annotation methods based on old or new (e.g. pattern matching and machine learning) techniques run in massively parallel operations.
- Lowering the cost of computing infrastructure and the necessary bandwidth to run large analyses.

For some community members, the computing infrastructure or even access to affordable hardware, software, and bandwidth can remain a limitation to addressing large scale datasets. Cloud computing solutions reduce these research overhead costs.

There are some concerns that exist within the community (i.e. concerns that clouds are proprietary or that ownership of the data is lost) that can be addressed with a minimal amount of active outreach and education about the technology. For many of the same reasons elaborated by Chavan and Ingwersen (2009), the path may not be direct and the need to ensure incentives and metrics for data originators will be a necessary concern. There are remedies for these concerns, including those explored in the proposed VertNet model (Constable et al., 2010). We feel that a move to a versioned release of a GBIF snapshot on publically available clouds will be another means to rapidly reduce the time required to detect errors and propose improvements to GBIF data and that cloud-based biodiversity data will eventually revolutionize the methods of biodiversity informatics.

7.4. Furthering external methods

GBIF has a history of actively promoting external solutions to fitness-for-use assessment. Progress toward OGC compliance may improve our abilities to integrate data published through GBIF into the growing network of geospatial tools and services. Although not a primary recommendation for improving geospatial fitness-for-use, OGC compliance will be of growing importance as GBIF begins to integrate more diverse data types. Some methods discussed previously, such as expert opinion range maps, habitat suitability information and distribution models, and large dataset analysis may become possible in the near future. Up to this point, many of these methods have been employed by independent research endeavours, while little work has been done to either make them available through services or standardized reporting of the results back to data providers. In addition, these methods each fill gaps in the current diversity of fitness-for-use improvement methods. Each area still needs directed research and funding avenues to ensure that they will work efficiently and effectively for the biodiversity data network. In addition to simply serving as data vetting tools, these services may also prove valuable for very quickly assembling knowledge about where species exist. GBIF's own endeavours, the Species Distribution Repository (SDR), may indicate that the value has already been recognized.

7.5. Conclusion

Although technologies for georeferencing and other fitness-for-use enhancements already exist, a major problem for our community has been the ability of data consumers to make annotations to accompany the original data at the source of the error: the original database. Darwin Core, its extensions, and the annotation schema contain the needed fields to document errors, and track external operations performed and data modifications. However, the responsibility to ensure data are stored at the source remains with the data publishers. Improvements made at the source, by data collectors and data publishers will remain invaluable and irreplaceable. Still, methods for offering data improvements back to data publishers are needed. For many reasons, IPT could be the right solution. IPT is already aware of the source schema and is operated by the publishers of the data. However, the IPT will not be the only solution. Further exploration and finalization the annotation schema or some variation of annotated DwC will help us solve the problem of communicating data improvements back and forth through the biodiversity network.

We foresee new analytical methods and workflow approaches to fill in the geospatial data quality gaps that exist. Some of these methods, especially those that can detect and resolve obvious errors (transpositions, coordinate system errors) may be implementable as part of the existing GBIF filter. For errors that are more difficult to detect and resolve, development of external services is a fundamental next step. New methods are being developed that could greatly facilitate assessing and reporting a record's fitness-for-use. We discussed the use of expert opinion range maps and habitat suitability models as semi-independent means for validating existing and new records. Another technology that could speed up data cleaning operations is the adoption of cloud computing solutions. Cloud computing consolidates all publisher resources, simplifying development of APIs and services that require complete (including many DwC terms)

or most up-to-date versions of the records. As GBIF already performs backup routines to store data, it may prove a minimal additional effort to send versioned subsets of the index to a publicly accessible cloud environment.

The increased publication of data from novel sources, such as field studies, will generate a need for technologies we have not yet considered. However, the proposed improvements to our current methods will help simplify the process of dealing with future data sources. We attempted to highlight areas where changes and improvements to our current practices could lead to rapid increases in data quality and where the least investment will result in the greatest rewards for our community. Above all else, it will be important that our community efficiently address each of the related but diverse topics discussed here.

As a community, we will need to prioritize and coordinate our methods to address each of these areas. Through community wide cooperation we can succeed in undertaking some of the larger elements of these proposals. From this work we hope that GBIF and the data publishing network will be able to both improve the current record of biodiversity and rapidly respond to future needs.

8. Literature cited

Ariño AH, Otegui J (2008) Sampling biodiversity sampling. Proceedings of TDWG online.

Bateman A, Wood M (2009) Cloud computing. Bioinformatics 25(12):1475.

Beaman R, Wieczorek J, Blum S (2004) Determining Space from Place for Natural History Collections. D-Lib Magazine 10(5) Available: http://www.dlib.org/dlib/may04/beaman/05beaman.html

Butler D, Gee H, Macilwain C (1998) Museum research comes off list of endangered species Nature 394:115-117.

Chapman AD (2003) Environmental Data Quality - b. Data Cleaning Tools. CRIA Report No. 6, Appendix I.

Chapman AD (2005) Principles and methods of data cleaning - primary species and species-occurrence data. Report for the Global Biodiversity Information Facility. GBIF, Copenhagen.

Chapman AD, Wieczorek J (2006) Guide to best practices for georeferencing (eds. Arthur D. Chapman & John Wieczorek). Contributors: Wieczorek J, Guralnick R, Chapman A, Frazier C, Rios N, Beaman R, Guo Q. Global Biodiversity Information Facility. Copenhagen, Denmark. Pp. 1-80.

Chavan VS, Ingwersen P (2009) Towards a data publishing framework for primary biodiversity data: Challenges and potentials for the biodiversity informatics community. BMC Bioinformatics 10(Suppl. 14):S2.

Chavan V, Krishnan S (2003) Natural history collections: A call for national information infrastructure. Current Science 84(1):34-42.

Constable H, Guralnick R, Wieczorek J, Spencer C, Peterson AT, and the VertNet Steering Committee (2010) VertNet: A new model for biodiversity data sharing. PLoS Biology 8(2): e1000309.

Duda RO, Hart PE, Stork DG (2001) Pattern Classification (2nd Ed). Wiley Interscience. New York, NY.

Green JL, Hastings A, Arzberger P, Ayala FJ, Cottingham KL, Cuddington K, Davis F, Dunne JA, Fortin MJ, Gerber L, Neubert M (2005) Complexity in ecology and conservation: mathematical, statistical, and computational challenges. BioScience 55(6):501-510.

Guisan A, Zimmermann NE (2000) Predictive habitat distribution models in ecology. Ecological Modelling 135: 147-186.

Guralnick RP, Constable H (2010) VertNet: Creating a data-sharing community. BioScience 60(4):258-259.

Guralnick RP, Hill AW (2009) Biodiversity informatics: Automated approaches for documenting global biodiversity patterns and processes. Bioinformatics 25(4):421-428.

Guralnick RP, Hill AW, Lane M (2007) Towards a collaborative, global infrastructure for biodiversity assessment. Ecology Letters 10:663-672.

Guralnick RP, Van Cleve J (2005) Strengths and weaknesses of museum and national survey data sets for predicting regional species richness: comparative and combined approaches. Diversity and Distributions 11(4): 349-359.

Guralnick R, Wieczorek J, Beaman R, Hijmans RJ and the BioGeomancer Working Group (2006) BioGeomancer: Automated georeferencing to map the world's biodiversity data. PloS Biology 4(11): e381. doi:10.1371/journal.pbio.0040381

Graham CH, Ferrier S, Huettman F, Moritz C, Peterson AT (2004) New developments in museum-

based informatics and applications in biodiversity analysis. Trends in Ecology & Evolution 19(9):497-503.

Hill AW, Guralnick RP. (2008) Distributed systems and automated biodiversity informatics: Genomic analysis and geographic visualization of disease evolution. In A. Gray, K. Jeffery, and J. Shao (Eds.) British National Conference on Databases. Springer-Verlag Lecture Notes in Computer Science series 5071, pp. 270-279. Berlin & Heidelberg.

Hill AW, Guralnick RP, Flemons P, Beaman R, Wieczorek J, Ranipeta A, Chavan V, Remsen D (2009) Location, Location, Location: Utilizing pipelines and services to more effectively georeference the world's biodiversity data. BMC Bioinformatics. 10(Suppl. 14):S3.

Hulbert AH, Jetz W (2007) Species richness, hotspots, and the scale dependence of range maps in ecology and conservation. Proceedings of the National Academy of Science, 104(33):13384-13389.

Jenkins M (2003) Prospects for biodiversity. Science 302:1175-1177.

Lane MA (2003) The Global Biodiversity Information Facility. Bulletin of the American Society for Information Science and Technology 30(1):22-25.

Lawes MJ, Piper SE (1998) There is less to binary maps than meets the eye: The use of species distribution data in the southern African sub-region. South African Journal of Science 94(5):207-210.

Otegui J, Robles E, Ariño AH (2009) Noise in Biodiversity Data. Poster presented at: e-Biosphere International Conference on Biodiversity Informatics. Jun 1-3, London, UK.

Patterson BD, Ceballos G, Sechrest W, Tognelli MF, Brooks T, Luna L, Ortega P, Salazar I, Young BE (2007) Digital Distribution Maps of the Mammals of the Western Hemisphere, version 3.0. NatureServe, Arlington, Virginia, USA.

Penev L, Erwin T, Miller J, Chavan V, Moritz T, Griswold C (2009) Publication and dissemination of datasets in taxonomy: ZooKeys working example. ZooKeys 11: doi: 10.3897/zookeys.11.210

Phillips SJ, Dudik M, Elith J, Graham CH, Lehmann A, Leathwick J, Ferrier S (2009) Sample select bias and presence-only distribution models: implications for background and pseudo-absence data. Ecological Applications 19(1): 191-197.

Pimm SL, Russell GL, Gittleman JL, Brooks TM (1995) The future of biodiversity. Science 269:347-350.

Rios NE, Bart HL Jr. GEOLocate. Georeferencing Software. User's Manual. Belle Chase, LA, USA: Tulane Museum of Natural History. Accessed December 2009.

Schatz MC (2009) Cloudburst: highly sensitive read mapping with MapReduce. Bioinformatics. 25(11): 1363-1369.

Soberón J (1999) Linking biodiversity information sources. Trends in Ecology & Evolution 14(7): 291.

Soberón J, Peterson AT (2004) Biodiversity informatics: managing and applying primary biodiversity data. Philosophical transactions of the Royal Society of London 359(1444):689-698.

Stein BR, Wieczorek J (2004) Mammals of the World: MaNIS as an example of data integration in a distributed network environment. Biodiversity Informatics 1:14-22

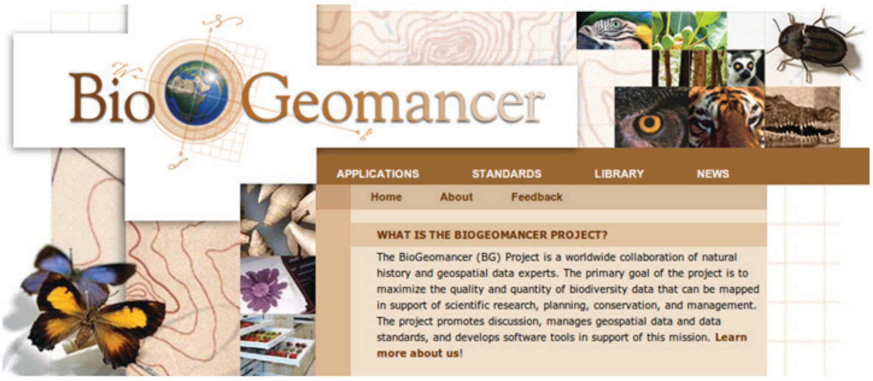
Sulloway FJ (1982) Darwin and his finches: the evolution of a legend. Journal of the History of Biology 15:1-53.

Wieczorek J, Guo Q, Hijmans RJ (2004) The point-radius method for georeferencing locality descriptions and calculating associated uncertainty. International Journal of Geographical Information Science 18(8): 745-767.

van Zonneveld M, Jarvis A, Dvorak W, Lema G, Leibing C (2009) Climate change impact predictions on Pinus patula and Pinus tecunumanii populations in Mexico and Central America. Forest Ecology and Management 257(7):1566-1576.

9. Appendix I: Catalogue of tools

i. BioGeomancer http://www.biogeomancer.org

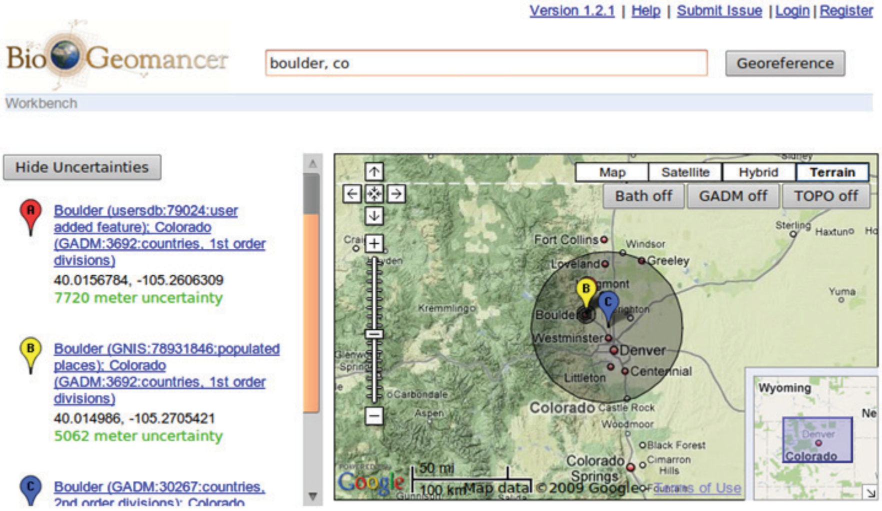


WHAT IS THE BIOGEOMANCER PROJECT?

The BioGeomancer (BG) Project is a worldwide collaboration of natural history and geospatial data experts. The primary goal of the project is to maximize the quality and quantity of biodiversity data that can be mapped in support of scientific research, planning, conservation, and management. The project promotes discussion, manages geospatial data and data standards, and develops software tools in support of this mission. **Learn more about us!**

The original BioGeomancer was established to convert textual descriptions of species occurrences (i.e. 3 miles NE of Boulder, CO) into decimal degree latitude and longitude and a statistically calculated coordinate uncertainty. BioGeomancer relies on a set of gazetteers to perform these conversions.

ii. BioGeomancer Workbench http://bg.berkeley.edu/latest/



boulder, co

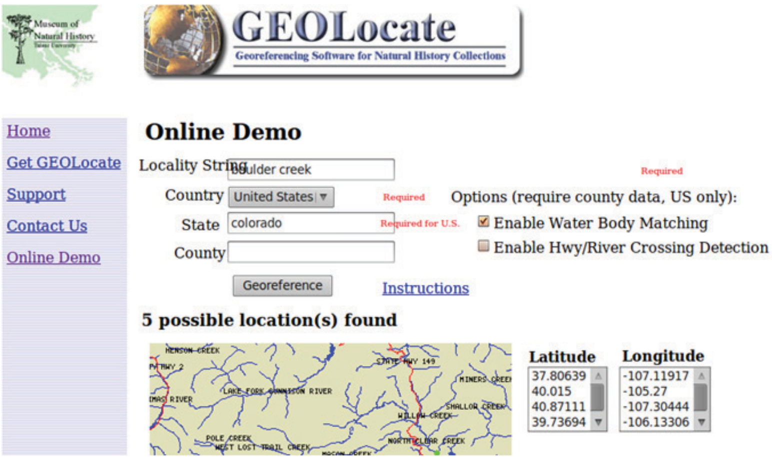
A Boulder (usersdb:79024:user added feature); Colorado (GADM:3692:countries, 1st order divisions)
40.0156784, -105.2606309
7720 meter uncertainty

B Boulder (GNIS:78931846:populated places); Colorado (GADM:3692:countries, 1st order divisions)
40.014986, -105.2705421
5062 meter uncertainty

C Boulder (GADM:30267:countries, 2nd order divisions); Colorado

The BioGeomancer workbench expands upon the functionality of BioGeomancer, by allowing users to modify coordinates and uncertainty based on known elements of the locality description. It has also expanded into batch georeferencing and the incorporation of user generated gazetteer information (Guralnick et al., 2006).

iii. GEOLocate http://www.museum.tulane.edu/geolocate/



Online Demo

Locality String Boulder creek

Country United States

State colorado

County

Options (require county data, US only):

☑ Enable Water Body Matching

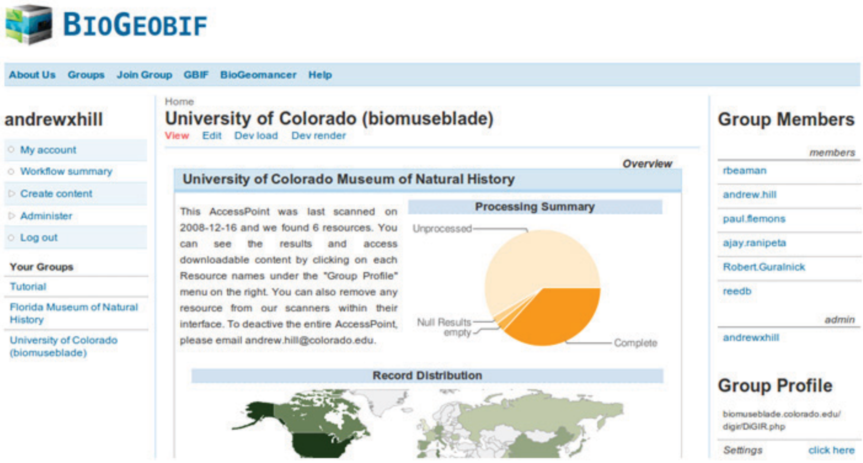
□ Enable Hwy/River Crossing Detection

5 possible location(s) found

Latitude	Longitude
37.80639	-107.11917
40.015	-105.27
40.87111	-107.30444
39.73694	-106.13306

GEOLocate provides an interface for georeferencing museum collections data. The program includes global gazetteer data and historically has had a strong focus on collections data sampled along water bodies and river-roadway crossings.

iv. BioGeobif http://biodiversity.colorado.edu/bgb/



andrewxhill

Your Groups

Tutorial

Florida Museum of Natural History

University of Colorado (biomuseblade)

University of Colorado (biomuseblade)

University of Colorado Museum of Natural History

This AccessPoint was last scanned on 2008-12-16 and we found 6 resources. You can see the results and access downloadable content by clicking on each Resource names under the "Group Profile" menu on the right. You can also remove any resource from our scanners within their interface. To deactive the entire AccessPoint, please email andrew.hill@colorado.edu.

Group Members

rbeaman

andrew.hill

paul.flemons

ajay.ranipeta

Robert.Guralnick

reedb

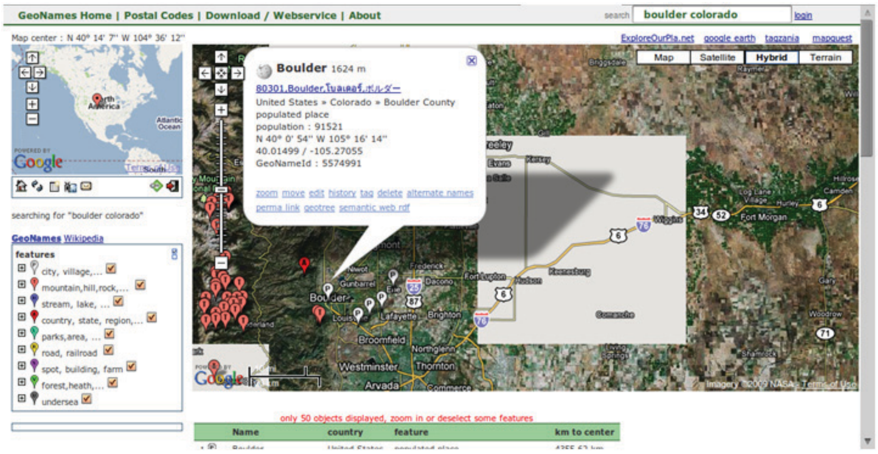
andrewxhill

Group Profile

biomuseblade.colorado.edu/digir/DiGIR.php

BioGeobif is an attempt to better link biodiversity collections data to the tools for georeferencing (BioGeomancer GEOLocate, etc.). The goal of the project has been to explore how to automate the methods of georeferencing, from collection to reporting back to publishers, with the hopes of easing many of the repetitive tasks associated with the process.

v. GeoNames http://www.geonames.org



Map center : N 40° 14' 7" W 104° 36' 12"

Boulder 1624 m
80301,Boulder,โบลเดอร์,ボルダー
United States » Colorado » Boulder County
populated place
population : 91521
N 40° 0' 54" W 105° 16' 14"
40.01499 / -105.27055
GeoNameId : 5574991

zoom move edit history tag delete alternate names
perma link geotree semantic web rdf

searching for "boulder colorado"

GeoNames Wikipedia

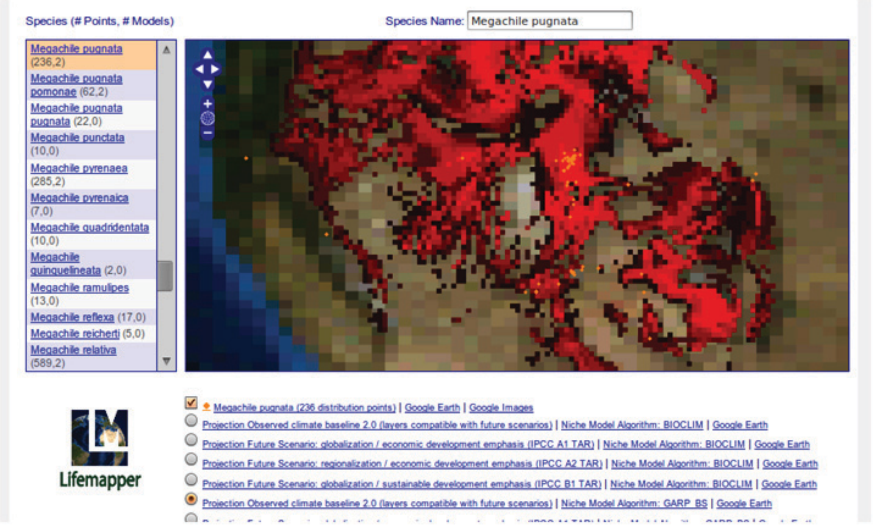
features
⊞ city, village,... ☑
⊞ mountain,hill,rock,... ☑
⊞ stream, lake, ... ☑
⊞ country, state, region,... ☑
⊞ parks,area, ... ☑
⊞ road, railroad ☑
⊞ spot, building, farm ☑
⊞ forest,heath,... ☑
⊞ undersea ☑

only 50 objects displayed, zoom in or deselect some features

Name	country	feature	km to center

GeoNames is another georeferencing tool. Unlike BioGeomancer or GEOLocate, GeoNames presents an example of a tool not developed specifically for biological collections.

vi. LifeMapper http://lifemapper.org/



Species (# Points, # Models)

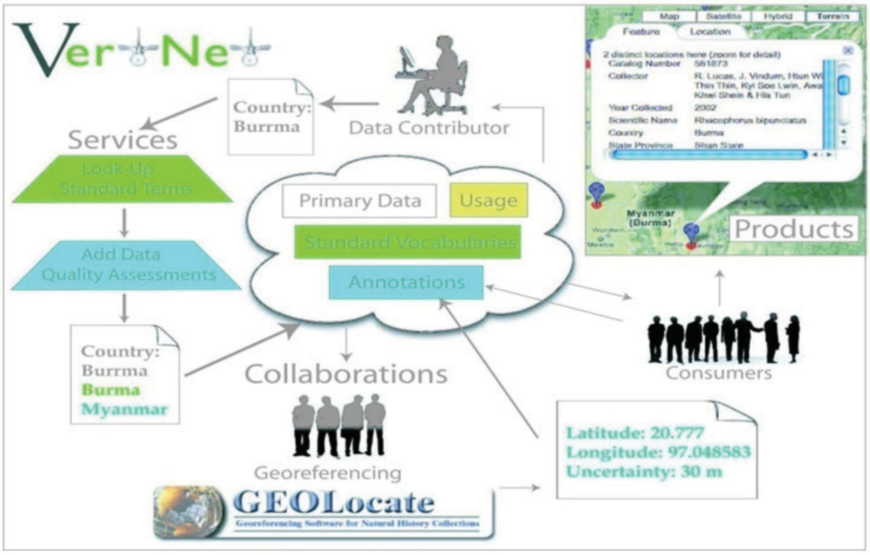
Species Name: Megachile pugnata

Megachile pugnata (236,2)
Megachile pugnata pomonae (62,2)
Megachile pugnata pugnata (22,0)
Megachile punctata (10,0)
Megachile pyrenaea (285,2)
Megachile pyrenaica (7,0)
Megachile quadridentata (10,0)
Megachile quinquelineata (2,0)
Megachile ramulipes (13,0)
Megachile reflexa (17,0)
Megachile reicherti (5,0)
Megachile relativa (589,2)

☑ Megachile pugnata (236 distribution points) | Google Earth | Google Images
○ Projection Observed climate baseline 2.0 (layers compatible with future scenarios) | Niche Model Algorithm: BIOCLIM | Google Earth
○ Projection Future Scenario: globalization / economic development emphasis (IPCC A1 TAR) | Niche Model Algorithm: BIOCLIM | Google Earth
○ Projection Future Scenario: regionalization / economic development emphasis (IPCC A2 TAR) | Niche Model Algorithm: BIOCLIM | Google Earth
○ Projection Future Scenario: globalization / sustainable development emphasis (IPCC B1 TAR) | Niche Model Algorithm: BIOCLIM | Google Earth
● Projection Observed climate baseline 2.0 (layers compatible with future scenarios) | Niche Model Algorithm: GARP_BS | Google Earth

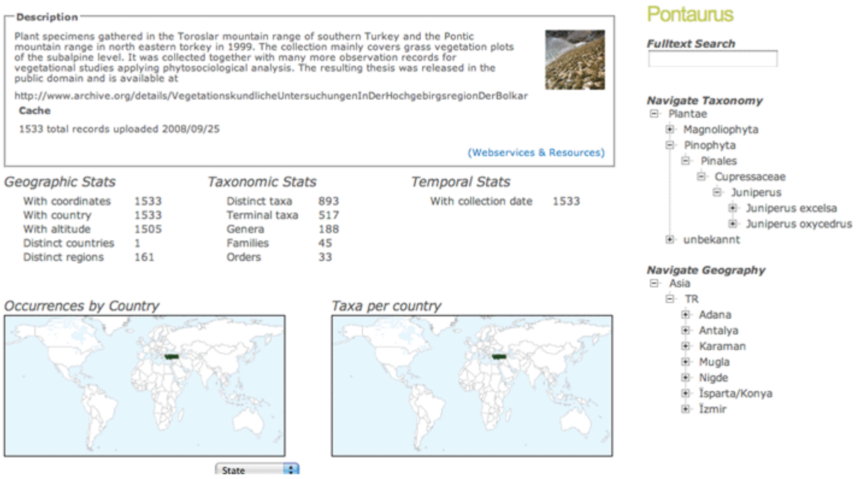
LifeMapper is a project focused on building species distribution models for all the species of the world. It uses species point data to predict a species niche with several algorithmic approaches (GARP, Bioclimatic envelopes, etc.) and many environmental layers. The project directly relies on high-quality geospatial data and will be benefited from better fitness-for-use documentation of point data.

vii. VertNet http://vertnet.org/



VertNet is a network bringing together four distinct consortiums of data publishers: MaNIS, HerpNET, ORNIS, and FishNET that collectively link data from 72 institutions. A newly proposed VertNet architecture (Constable et al., 2010) puts forth a new method of data publishing, where all data will be stored in a unified cloud resource.

iix. Integrated Publishing Toolkit http://ipt.gbif.org/



Description

Plant specimens gathered in the Toroslar mountain range of southern Turkey and the Pontic mountain range in north eastern torkey in 1999. The collection mainly covers grass vegetation plots of the subalpine level. It was collected together with many more observation records for vegetational studies applying phytosociological analysis. The resulting thesis was released in the public domain and is available at

http://www.archive.org/details/VegetationskundlicheUntersuchungenInDerHochgebirgsregionDerBolkar

Cache

1533 total records uploaded 2008/09/25

(Webservices & Resources)

Geographic Stats

With coordinates	1533
With country	1533
With altitude	1505
Distinct countries	1
Distinct regions	161

Taxonomic Stats

Distinct taxa	893
Terminal taxa	517
Genera	188
Families	45
Orders	33

Temporal Stats

With collection date	1533

Occurrences by Country

Taxa per country

Pontaurus

Fulltext Search

Navigate Taxonomy

```
⊟ Plantae
    ⊞ Magnoliophyta
    ⊟ Pinophyta
        ⊟ Pinales
            ⊟ Cupressaceae
                ⊟ Juniperus
                    ⊞ Juniperus excelsa
                    ⊞ Juniperus oxycedrus
    ⊞ unbekannt
```

Navigate Geography

```
⊟ Asia
    ⊟ TR
        ⊞ Adana
        ⊞ Antalya
        ⊞ Karaman
        ⊞ Mugla
        ⊞ Nigde
        ⊞ İsparta/Konya
        ⊞ İzmir
```

The Integrated Publishing Toolkit (IPT) represents a new way of publishing data to the GBIF distributed network. It gives publishers a web-browser based means of managing their resources and a simplified way for GBIF to monitor dataset updates and changes.

ix. Species Distribution Repository http://sdr.gbif.org/

The Species Distribution Repository (SDR) was an exploratory project run by GBIF to begin providing a web-interface to species range maps. Data included external sources (IUCN) as well as converting point data shared through GBIF into country level inventory data.

x. GBIF Training Manuals http://www.gbif.org/participation/training/resources/gbif-training-manuals/
GBIF has developed internally and commissioned a number of training manuals and best practice guidelines for the biodiversity publishing community. These are an important resource meant to lower the bar for new data publishers to enter into the network.

10. Appendix II: Glossary of terms

Annotation: A stored remark about any individual record or dataset. This can include a measure of fitness-for-use or a suggested correction.

API: (stands for Application Programming Interface) a set of tools, methods and functions of a program made available to enable other programs to interact with it.

Cloud computing: Cloud computing is a term that can be used to refer to one of many distributed computing solutions. In this work, it is primarily used to refer to the use of large-scale computing resources where data and software can be designed, stored, and employed by purchasing time or space from established computing and data centres (i.e. Google App Engine, Amazon Web Services, Force.com, etc.). We recommend readers review the VertNet publication (Constable et al., 2010).

Coherence error: A kind of error which may appear when gathering multiple sources of data, if those sources do not share a common data structure and/or definition. An example could be when a source uses degree-minute-second while other source uses decimal degree for latitude and longitude.

Darwin Core: A body of standards for sharing primary biodiversity data. Ratified by TDWG in 2009. Abbreviated DwC. (http://rs.tdwg.org/dwc/)

Data publisher: The organization making data digitally available but not always the data owner in cases where permission has been granted to an external organization.

Data resource: A subset of the data made available by a publisher. For example, each museum collection provided by the same data publisher may be considered different resources.

Datum: Here used as singular of data, not to be confused with the geodectic datum, the reference from which measurements are made in cartography.

Detectability: A relative measure of how much effort (human or computational) would be needed to detect an error of a given nature.

Error: A measure of the deviation of a piece of information from its true value.

Expert opinion range map: A hypothesis about the extents of a species range derived by numerous sources and vetted by one or many experts.

Gazetteer: A geographical dictionary that links the coordinates of places and their textual place names.

Georeference: The conversion of textual place names into the computer readable geospatial information; latitude, longitude.

Geospatial Extension: The non-core geospatial information stored in a separate file of the star schema for inclusion with Darwin Core records (http://wiki.tdwg.org/twiki/bin/view/DarwinCore/GeospatialExtension).

Harvest: an operation performed by many services where data is pulled from its source.

Precision: A measure of the granularity of given information, the degree of detail it provides. In numeric values, it would represent the number of significant digits an observation is recorded in.

Resolvability: A relative measure of how much effort (human or computational) would be needed to correct an error of a given nature.

Scale: A qualitative measure of the necessary granularity demanded in the information for a given work, research or study.

GLOBAL
BIODIVERSITY
INFORMATION
FACILITY

Secretariat
Universitetsparken 15
DK-2100 Copenhagen Ø
Denmark
Tel: +45 35 32 14 70
Fax: +45 35 32 14 80
Email: info@gbif.org
Web: www.gbif.org

ISBN: 87-92020-11-9