DRAFT

Dealing with Sensitive Primary Species Occurrence Data

Report

Arthur D. Chapman

1 August 2006

Prepared for the Global Biodiversity Information Facility

Version 0.2

CONTENTS

1. The Issues

The **GBIF Secretariat** has concerns about the unprotected distribution of Sensitive Primary Species Occurrence Data (for example the exact localities of rare, endangered or commercially valuable taxa). This as an important issue and needs to be addressed in relation to data to be shared through the GBIF network and made visible through the **GBIF Data Portal**.

A review of current approaches for obscuring or generalizing such data was initiated in February 2006 and an on-line survey conducted through Survey Monkey[1]. A separate report on the results was made available via the GBIF Web site[2] in early June 2006 (Chapman 2006). It is important to also understand the possible impact that such approaches may have on biodiversity science, and while restricting the availability or resolution of certain data, not overly restricting the uses to which the data may be put. The second stage in the process has been the development of a report that will eventually lead to a best practice recommendations and hence this document.

Using the on-line survey, The GBIF Secretariat wished to examine:

- which data are regarded as 'sensitive'
- which approaches are currently used by GBIF data providers to protect sensitive data (and the associated advantages and disadvantages of these approaches)
- the extent to which each approach may be reversed through co-relational analysis
- the extent that generalization may restrict various analyses
- the level of generalization that may be appropriate for different types of data
- the best ways of documenting generalization of data and the methods used
- whether a standard approach can be promoted for all sensitive data provided through the GBIF network
- whether changes should be made to the TDWG ABCD and Darwin Core schemas (used by GBIF for exchange of Primary Species Occurrence Data) to facilitate sharing generalised data

A. What are Sensitive Taxa?

The survey identified several categories of data that institutions regard as being sensitive. These can be split into two groups, viz. those that are 'permanently' sensitive (such as endangered species, etc.) and those that are 'temporarily' sensitive (such as data subject to ongoing research or awaiting publication) (figure 1). The majority of respondents (around 75%) identified that less than 10% of their collections were sensitive taxa. Those with more than10% were generally agencies dealing specifically with endangered species, fossils, etc., had a large proportion of their collection as 'subject to ongoing research', or a large proportion of their collection subject to third-party agreements.

[1] Survey Monkey http://www.surveymonkey.com
[2] http://www.gbif.org/prog/digit/sensitive_data/Summary_of_Responses_-_03.pdf

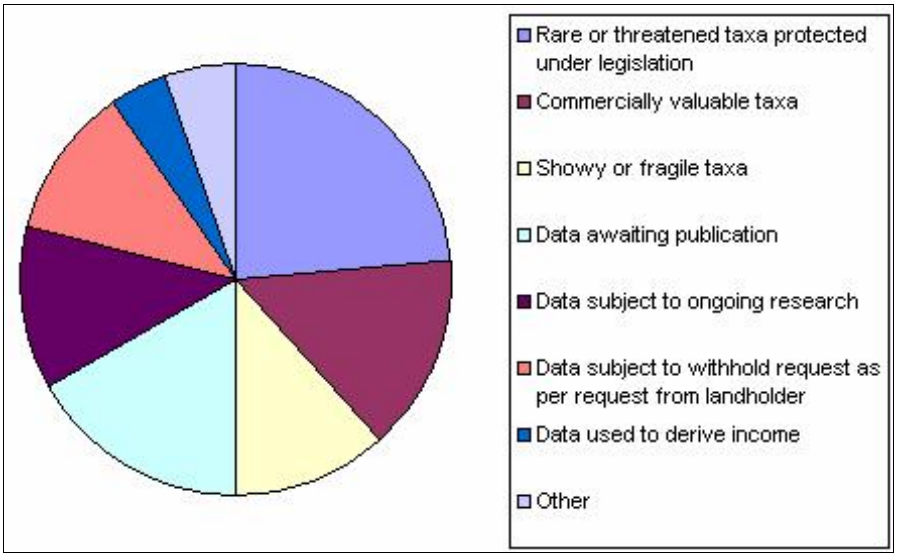

Fig. 1. Summary of 93 responses to GBIF on-line Survey on Dealing with Sensitive Primary Species Occurrence Data.

1. Permanently Sensitive Data

Categories of permanently sensitive data include

- Rare or threatened taxa protected under legislation (86% of respondents)
- Commercially valuable taxa (54% of respondents)
- Showy or fragile taxa (43% of respondents)
    - Orchids, cacti, cycads and some animal nesting/roosting sites were particularly identified
- Data subject to withholding request from landholder (41% of respondents)
- Data used to derive income (16% of respondents)
- Other (19% of respondents), including:
    - Traditional and cultural knowledge
    - Locations in protected areas, EEZ, hotspots
    - Data under agreements with third parties
    - Quarantine interceptions and quarantine records
    - Privacy and IP rights of others
    - Material from private breeding companies
    - Data that may be used to identify localities of related collections (sequential collections of a collector; dates of collection, etc.)
    - Possible misuse of the data

2. Temporarily Sensitive Data

Categories of temporarily sensitive data include

- Data awaiting publication (61% or respondents)
- Data subject to ongoing research (44% of respondents)
- Incomplete or unchecked data (5%)

B. National/Regional versus Global

An issue that arose with the identification of sensitive taxa was in knowing what was sensitive in different areas – for example an institution holding data from another jurisdiction (country, state, etc.) may not know what is protected or sensitive within that jurisdiction. Suggestions were to have a global list of sensitive taxa (or use the

IUCN Red List[3]) however; some taxa may only be sensitive in one region and not in another. Not all sensitive data are now included on any one global list and the difficulties in developing such a list (both taxonomically and geographically), and exchanging and maintaining that information is not a simple issue.

C. Privacy Concerns

A number of respondents raised the issue of personal privacy. Many countries are introducing privacy legislation in their jurisdictions and these may restrict the ability of institutions to make information on individuals (names of collectors, names of those who carried out the identifications, etc.) available. Some also raised the issue in the light of protecting individuals who may have inadvertently collected in areas where they may not have valid permits, etc. This has huge implications for data quality. Subsequent to the survey, a discussion on this issue occurred on the Taxacom listserver[4] and that discussion was used in the development of the recommendations in this report (see below).

D. Associate Information

Several respondents raised the issue of protecting associated information to restrict possibilities for co-relational analysis and data mining. These include collector, collector number, date of collection, etc. which may be used to identify sequences in collecting and thus be used for deductions on the localities of missing numbers. Other data may include habitat and community information.

E. Intellectual Property Rights

A GBIF Experts Meeting on biodiversity data held in 2004 (GBIF 2004a) identified a number of Intellectual Property Rights (IPR) issues with respect to sensitive data. A consultant at that meeting identified, among others, that IPRs "*may become relevant under one or more situations*".

At that meeting it was identified that data providers needed to understand the sensitivity of certain taxa and to restrict the release of data if necessary to protect vulnerable biodiversity or to respect confidentiality as may be required through contractual obligations.

It was agreed that users of the GBIF Network needed to respect "*data providers' restrictions of access to sensitive data*" and this has now been added to the GIF Data Use Agreement (GBIF 2004b), viz:

> *2. Users shall respect restrictions of access to sensitive data.*

F. Need to make data useable

Most users of the data appeared to understand the need for data providers to restrict certain information on sensitive taxa, however stressed the need for good documentation so that users knew what taxa were restricted and how, allowing them to make decisions on the value and/or usefulness of the data for their particular uses and analyses. At present it would appear that a lot of data are being generalized, but there is no associated documentation informing users that this has been done, and

[3] IUCN Red List of Threatened Species http://www.redlist.org/
[4] Taxacom Archive – June 2006 http://mailman.nhm.ku.edu/pipermail/taxacom/2006-June/thread.html

how. The data may thus be used in inappropriate ways and produce false results without the user being aware that the data have been modified.

2. Background

Prior to the introduction of electronic networks, curators could control (at least to some extent) who had access to the data from specimen collections in their institutions. However, there was still a culture of scientific openness and detailed location information was usually published in association with the publication of new species, and in floras and faunas, etc.

The internet introduced the possibility of exchanging large amounts of data and as a consequence data on sensitive taxa soon became available in seconds to those looking for them. This was a great boon to scientific research, but also opened up the possibility for unscrupulous users to use the information for nefarious purposes as there was no control or even identification of who was using the data or for what purpose. Many institutions began to hide all information on sensitive taxa (usually rare and threatened) while others (usually larger institutions with more sophisticated computing resources) developed systems for generalizing data in a number of ways in order to 'fuzzy-up' the detailed location information. One problem that has arisen is that data are already distributed around the globe through duplicate specimens, etc. and although data may be restricted from some institutions, others holding duplicates may be releasing the same information. This may be through ignorance of what may be regarded as sensitive in the home ranges of the taxon concerned, as no universal list of what is regarded as 'sensitive' is available. Difficulties are compounded by the fact that a taxon may be sensitive in one area, but not in another (and indeed may even be a weed or pest species in the second location).

Until now, no attempt has been made to standardize methods for generalizing data or for providing guidance to institutions on how they may safely make data on sensitive taxa available to those who need it in a way that makes the data useable, but at the same time restricts possibilities for nefarious use.

The Global Biodiversity Information Facility (GBIF) has a vested interest in making data available via its portals, but at the same time respecting the wishes of data providers to restrict information on sensitive taxa. As a result, GBIF has decided to conduct a survey on what institutions are currently doing to protect data on sensitive taxa, and to explore ideas for developing guidelines and standards as well as recommending methodologies that institutions may use in developing their data management and data release policies.

This document is one stage in the process and will eventually lead to a document on Best Practices for Dealing with Sensitive Primary Species-Occurrence Data. The first draft of this document will be used as a basis for a workshop to be held in the second half of 2006 from which it is hoped detailed recommendations on methodologies may be developed. Any feedback on this draft is welcomed by GBIF and the author.

A. Survey

In March 2006, GBIF conducted an on-line survey on 'Dealing with Sensitive Primary Species-Occurrence Data'. The results of that survey have been made available separately on the GBIF Web site[5]. The results of that survey have been used

[5] Questionnaire on Dealing with Sensitive Primary Species Occurrence Data − Summary of responses. http://www.gbif.org/prog/digit/sensitive_data/Summary_of_Responses_-_03.pdf).

extensively in the preparation of this document. Individuals representing more than 100 institutions completed the survey with another 48 supplying some information and requesting copies of the results. Respondents came from 28 countries and two International organizations and represented fairly evenly botanical and zoological collections, specimens and observations, small and large collections. Some living and paleontology collections were also represented. There was a strong skewing toward English-speaking countries and from the Northern Hemisphere with North American and European collections being by far the largest groups to respond. There were very few respondents from Africa, Asia, Central or South America (with the exception of Argentina) (figure 2).

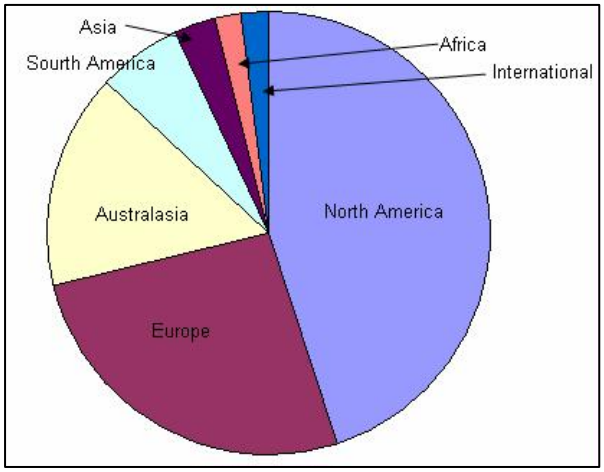

Fig 2. Summary of 93 responses to GBIF on-line Survey on Dealing with Sensitive Primary Species Occurrence Data showing countries of respondents.

It was evident from the survey, and from follow-up correspondence and discussions that dealing with sensitive data is seen as a critical issue in the distribution of primary species occurrence data. Responses ranged from the view that all data should be made freely available through to those that said they would not make any data available on sensitive taxa under any circumstances. Most though were looking for an acceptable way of generalizing information so that the exact localities of sensitive data were hidden from the general public while still making the information available to bona-fide researchers and users such as governmental conservation agencies. Many suggested that they would make data available if a suitable method could be found (possibly by GBIF) of registering and/or identifying bona-fide users, and if suitable security measures could then be introduced that would allow only those users to access the data. Others would prefer users wanting the data to contact curators individually and provide reasons for requiring the data.

B. Other Information Sources

Follow–up discussions were held with numerous people, either by email, through listserver discussions, at meetings of collection managers in Australia and the USA, through on-line internet and library searches and through discussions with programmers and others dealing with computer security.

3. The Present Situation

Currently, many institutions are not making any data on sensitive taxa available on-line or through the GBIF Data Portal at all. For data that are being made available, some fields are removed (including locality, fields including names or dates and, occasionally, taxonomic fields), georeferencing information is generalized or randomized in a variety of different ways, and very little, if any, documentation is supplied on what is being done. The lack of documentation is perhaps the most disturbing, as it means the data may not be suitable for the uses to which people are putting them, but the information is not available for the user to know that. This contrasts with another concern that many data custodians have about making sensitive data available – the problem of the possible misuse of the data.

It would appear that herbaria are more inclined to restrict their data than mammal or insect collections. Perhaps this is because plants don't move and the exact location of a collection is likely to lead one to an actual plant on the ground, whereas mammals and insects tend to move around. One entomologist commented that professional collectors and amateur groups often know more than the scientists about the location of rare species. However there are categories of animals where the exact locations were thought to be sensitive and included bat roosting and maternity sites, nesting sites of falcons, and the location of various lizards, tortoise and butterfly species. With plants, there is also a strong leaning towards not making information available for plants likely to be collected (pirated) such as cacti in Arizona (noted by several institutions), orchids and cycads. The protection of sensitive fossil sites was also identified.

On the other hand, some institutions have found benefit in working with the general public to gather information and to protect rare taxa, using the public and special interest groups to survey existing locations and to help locate new locations. There are good examples with birds, lizards, frogs, butterflies and various plant species (including orchids) in a number of countries. Several people have raised the issue of the balance between protecting taxa through knowledge of where they occur as opposed to protection through restricting knowledge of their occurrence at a location. This is very taxon (and maybe region) specific and certain taxa may be in greater danger due to inadvertent destruction through lack of knowledge than through deliberate collection and destruction through knowledge of locations. For this reason, a list of sensitive taxa may be quite different to a list of rare or threatened taxa, although there is likely be considerable overlap between the two.

A. Reasons for releasing or restricting data on sensitive taxa

As noted in a recent article in *Science* (Stuart *et al.* 2006), three newly discovered amphibian and reptile species rapidly appeared in commercial trade shortly after their descriptions in the scientific literature. This is an issue of concern to biologists and especially to taxonomists (Guteman 2006) – how much information should they release in publication when describing a new taxon.

The GBIF survey identified many reasons for restricting access to sensitive taxa. These can be categorized as follows with the number of respondents identifying them in brackets:

- Protect threatened species, economically important species and reduce the impact on wild populations of sensitive species and sensitive communities (37).
- Preclude deliberate sabotage or collection by unscrupulous and commercial collectors, poaching, hunting, disturbance, over exploitation, etc.; and to control bio-prospecting (35).
- Protect third party data held by the institution, abide by confidentiality, commercial-in-confidence and data agreements, protect the sources of the data and rights of data providers, and protection of IP rights, including need for proper attribution and citation (16).
- Allow for publication of research results and to maintain competitive advantage (14).
- Protect the rights and gain the cooperation and trust of landholders (10).
- Protect people's names and privacy (8).
- Fear of the user making inappropriate use of the data; not knowing purpose to which data will be put; fear of misinterpretation; can't guarantee data are 'fit-for-purpose' (5).
- Biosecurity, quarantine and trade issues (3).
- Wouldn't release under any circumstances (2).
- Benefit-sharing and need to maintain good relations with countries of origin, etc. (1).

The survey also identified the reasons institutions may grant access to sensitive data. This may not necessarily be through on-line access but through individual requests by bona-fide users, etc. The main reasons identified were:

- For scientific research and analysis; scientific advancement, collaborative projects (33).
- For species and conservation planning and management, and conservation assessment (21).
- Management of the environment, biological resources and land; need for continued conservation actions to maintain species and populations; environmental impact studies; biosecurity management (12).
- Inquiries from Government agencies and professional organizations, e.g. for policy making and environmental management (8).
- Species distribution studies, species modeling; vegetation survey and mapping; global scale analysis; monitoring and resurvey (6).
- Entire database should be available (free data policy) (6).
- Should be available to bona-fide individuals where there is reasonable assurance that data will be put to a non-commercial, serious scientific/scholarly use (3).
- Protection of species – where lack of disclosure could endanger species (2).
- For data contributors, benefit sharing, and data repatriation to countries of origin (2).
- For law enforcement and protection (1).
- Freedom of Information Act (1).
- Difficulty in restricting some and not all records (1).

Individual comments included:

- The data in most collections-based datasets are often fragmentary or incompletely representative of the population distributions or relative protection on conservation estate. Occasionally it is incompletely identified, mis-identified or not provided with the most recent applicable name. Without adequate knowledge and discussion of the project for which the data are sought, we cannot guarantee the data are 'fit-for-purpose'.
- Making land managers and agency biologists aware of rare species is essential to improve their chances of protection.
- An important concept is that making biodiversity data available should reduce the risk of damage to the environment.
- Free access to the data increases the utility, usefulness and value of the data which increases the value of the institute itself.
- We believe that biodiversity data needs to be freely available to anyone, anywhere, anytime.
- Collectors and poachers are usually ahead of scientists, not behind them.
- Environmentally sensitive information often relates to those species and habitats that are particularly vulnerable to land management activities. It is important that such information is made available to those that control land management activities at a level of detail that is useful.
- We do not make any sensitive data available - people can come here if they want it. They can read the specimens and locate the information. May restrict some science; better than lose some species.
- Access is granted only to data contributors who are vetted for professional qualifications.

Most institutions (over 80%) said that they would be prepared to make all categories of data available to Government Agencies, Universities and Research Organizations; around 60% to non Government Organizations and 25-50% to Commercial Consultants and the Public with the use of suitable protection methods such as password access, or single downloads. Most indicated that they would require some form of data agreement before release, or at least some way of identifying bona-fide users. Only 5% responded to the effect that they would supply no data of this nature under these circumstances.

Lawrence Way[6] of the Joint Nature Conservation Committee, suggests that there is a need for greater efforts in deciding what should truly be sensitive, including the use of evidence-based approaches and in developing education and leadership roles for data providers. There is also evidence (for example from the National Biodiversity Network in the UK and elsewhere) that collaboration with amateur groups rather than a confrontational approach can be beneficial in conserving sensitive taxa and in reducing pressure on wild populations through joint efforts at breeding and cultivation. A good example can be seen with the Wollemi Pine (*Wollemia nobilis*) in Australia[7].

Most institutions wished to retain control over the release of their data on sensitive taxa, and suggested that different levels of generalization of the data may be made available to the different categories of users. Many were happy to provide access electronically as long as a secure method of doing so was available, such as using

[6] *Pers. comm*. Lawrence Way, Joint Nature Conservation Committee, UK. July 2006.
[7] The Wollemi pine – a very rare discovery – Royal Botanic Gardens, NSW. http://www.rbgsyd.nsw.gov.au/information_about_plants/wollemi_pine

'username+password+ipaddress' login, and if there was some way of registering and/or verifying bona-fide users. Some saw an advantage in a system of verification being conducted through the GBIF Data Portal, whereas others were not prepared to "*hand over the ability to do any vetting of a request*".

B. Generalization

Two-thirds of the respondents to the question said that they currently generalized at least one field when making data on sensitive taxa available. Of these 64% deleted or altered the locality and/or the georeferencing information and 24% restricted information on collector's or observer's names. Other fields restricted included determiner's names, dates, taxonomic information, habitat information, sex of individuals, hosts, traditional uses and some others. Four percent did not show any information at all for sensitive taxa whereas another 7% restricted everything except the name and accession id.

The reasons given for restricting collector's and determiner's names included

- to protect the privacy of living people;
- restrict possibility of tracking itineraries and thus collections before and after a sensitive species;
- privacy legislation;
- to shield people from possible reprisals by animal-rights activists;
- observational data are sometimes interpreted to include (possibly illegal) collecting of material, whereas it usually consists of only photographic or observational records;
- to protect collectors of birds and mammals, etc.

Others note that they NEVER suppress this information.

About half of all respondents used individual data sharing agreements or data licenses for making data available to bona-fide users. Most are developed on a case-by-case basis, although some are general agreements signed across a number of programs (usually where data are made available through National Heritage programs or Data Centers, etc. or across collaborative programs that involve a number of agencies). In the majority of these, sample agreements are available to GBIF on request.

In some cases institutions saw a conflict between governmental requirements for data to be freely available, and the institution's desire to restrict certain information for what they saw as valid reasons.

1. Generalization of Locality Descriptions

The majority of institutions that generalize the locality descriptions do so by either not making the field available at all (60%) or by altering the wording (23%), for example to something like:

> *This specimen represents an endangered or threatened species. The specific locality has been removed from the on-line record to protect this species from over-collection. These data may be supplied to researchers on request.*

Many other institutions believed that they should do something similar, but currently were not exchanging the locality field at all.

2. Generalization of Georeference Information

There were 46 responses to the survey question on methods of generalizing georeferencing information (Table 1). Percentages do not add up to 100 as many respondents reported more than one category.

Response	No.	Response	No.
Report by a geographic region or bioregion	54%	Remove altogether	37%
Report by standard grid or map sheet	47%	Move to nearest named place	6.5%
Round down (to 1 minute, 10 minutes, 30 minutes, degree, etc.)	37%	Some other method (see comments)	17%

Table 1. **Reponses to the question on generalization of georeferencing information from the GBIF survey on 'Dealing with Sensitive Primary Species Occurrence Data'.**

In general institutions tend to generalize rather than randomize (see glossary for definitions), although a number of cases of randomization were reported. By far the majority who reported using a grid, reported using a 10 by 10 km grid or smaller (some as small as 100 or 200 meters), or a 1 minute grid (created by dropping off the seconds). There were a small number at smaller scales such as 0.1 degree (6 minutes) or 10 minutes with some rare cases where half degree or degree grids were used.

A large proportion or respondents said that they provided data on sensitive taxa by political region (often a county, parish or district) or by a biogeographic region or watershed.

3. Restricting information on Determiners

A small (but significant) number (ca. 8%) of respondents to the survey said that they did not make the names of living people (including the name of people who determined the specimen) available for privacy reasons. These came from a number of countries including Argentina, Canada, Spain, Sweden and Switzerland. Subsequent to the survey an extensive discussion was held on this topic on the Taxacom Listserver[8]. The listserver discussion put forward many convincing arguments for the need for information on the name of the determiner and the date to be exchanged, and considerable concern was expressed about any restriction of this information.

Privacy legislation has been introduced into a number of countries and it is not clear how such legislation may affect the distribution of information on those who have identified or confirmed an identification of a specimen or observation. This includes the name of such an individual. In many countries scientists have so far just ignored the legislation in the belief that it does not apply to scientific license and no one has yet been able to report a case where a scientist has been prosecuted under privacy legislation for releasing such information. A related case in Sweden is often cited as an example of what may occur under such legislation (see article discussing this issue in the *American Reporter* of November 24, 1998[9]).

[8] Taxacom Listserver Archive Discussion on 'Privacy Laws and Science' <http://mailman.nhm.ku.edu/pipermail/taxacom/2006-June/thread.html>

[9] *American Reporter* November 24, 1998. <http://www.praxagora.com/andyo/ar/privacy_sweden.html>

Apparently much of the privacy legislation referred to has arisen (at least in European countries) as a result of a European Union Directive 95/46 of 1995[10]. There would appear to be some disagreement as to the extent such a directive relates to scientific information, and to what degree it may apply to the names of collectors and determiners of biodiversity specimen and observational information. This would appear to need clarification.

C. How are others handling this issue?

A number of third-party agencies around the world have examined the issue of sensitive data and treat the data in various ways in an attempt to provide a balance between making information available while at the same time protecting the locations of sensitive taxa. For example, the Calflora project out of California:

> "*We want college students to have the information at their fingertips in a time frame that allows them to do term projects on the ecology of rare species. On the other hand we want to make sure that our actions do not unnecessarily contribute to vandalism or destruction of vulnerable species*" (Malpas 2004).

The Califlora has established an Advisory Board to address the topic to help decide what information needed to be restricted and to ensure that whatever is done "*is critically needed to prevent irreparable harm, and that benefits of suppression substantially outweigh the benefits of having this information available*" (Malpas 2004).

> "*The committee will review and decide on the merits of proposals to suppress location information for particular taxa. Such proposals must be supported by identification of specific threats to that taxon or its habitat, and must be supported by justification for the position that a change in CalFlora's display will materially reduce those threats*". (Appendix VIII. From Malpas 2004[11]).

The NBN Gateway[12] in the UK – an organization that links 66 public, private and voluntary bodies sharing 160 datasets containing 20 million records has developed a framework of online administrative tools that providers can use to control the availability of their own data. These controls are most developed for species records. The controls can be used to set different access levels for the public, specific registered individuals and specific registered organizations. Data providers are able to[13]

a) "*Limit the resolution of locality for records within a dataset. Resolution can be set at 10km square, 2km square, 1km square and full (actual) resolution.*
b) *Set whether or not a copy of their records can be downloaded from the NBN Gateway website.*
c) *Set whether or not attributes (additional fields of information within records) can be viewed (standard access gives the user the species taxa, the date recorded and the geographic location at the set resolution).*

[10] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data <http://aspe.os.dhhs.gov/datacncl/eudirect.htm>
[11] http://www.calflora.org/goalsAndAchievements.html
[12] http://www.nbn.org.uk/downloads/files/NBN%20Standard%20Exchange%20Format%201.pdf
[13] From NBN response to GBIF Survey on *Dealing with Sensitive Species-Occurrence Data* (Mar 2006).

d) *Whether or not records flagged as confidential or sensitive within the resource can be seen.*

e) *Whether or not the name of original recorders and determiners (if included) can be seen.*"

To identify sensitive taxa they have a field for "Sensitive Taxa (True or False)" which unless set is assumed to be False.

4. The Case for a Standard Method of Generalizing

From the GBIF survey, eighteen percent (18%) of respondents stated that they were required for legal reasons to limit access to information on sensitive taxa. Several reported that these were due to Acts of Parliament that restrict release of information on endangered species, etc., and some were due to privacy acts on the release of personal information. The majority or reasons, however, appeared not to be due to legislative instruments *per se*, but legal instruments such as data agreements with data suppliers, landholders, or traditional owners. Eighty two (82) percent said they weren't so required, but many of these did restrict information in the interests of protecting the species concerned. Several responses mentioned that there was a requirement in the United States to make data available in the public domain.

While many saw a need to restrict information on the detailed localities of sensitive taxa, they were keen to make information available to users in a generalized form and especially to users who may carry out studies or analyses that would benefit the long-term conservation of those taxa. Responses included:

- "*offers a good balance between protection and still making the data available for some purposes such as occurrence within a region, or even coarse distribution studies*";
- "*For the very sensitive species, 10km is a good compromise between planner needs for detailed data (or at least a flag that sensitive data are there) and the need to protect exact locations*."
- "*generalization creates/retains "true" data, whereas randomization creates deliberately false data. Generalization can be implemented (or not) on a case-by-case basis, depending on the intended use(s) of the data*".
- "*Generalization is easily implemented by simply omitting data fields containing more precise data, and supplying data in tabular instead of georeferenced format*".
- "[Generalisation is] *Simple, and provides information that is still useful at medium scales, without giving away the exact location of populations*".
- "[Generalization will lead to] *improved credibility of studies based upon GBIF data*".

An issue that was constantly raised by users of the information was that any generalization should be documented so that the users knew what reliability they could place in the data for their uses. A number of respondents also suggested that having one (or several) recommended methods for generalizing would lead to consistency between collections, would provide collections with guidelines on how best to do it (many suggested that they were just doing something that was simple without having "*investigated and dedicated much time to really focus on the issue and implement a more elaborate policy*"), and provide users of the data with a degree of certainty in the data.

Two-thirds of those who responded to the question (42 respondents) and who were not now generalizing said that they would likely generalize data on sensitive taxa if a standard and reliable method of doing so was recommended. The majority of respondents currently does not release any data at all on these taxa, or remove locality (and other information) completely from on-line distribution.

Quite a number of respondents would like to continue to restrict locational information, or generalize it quite coarsely, but would like to be able to make detailed data available to bona-fide users who could be vetted in some way. They would like to see recommendations on secure methods of allowing access to bona-fide users and for vetting the bona-fides of such users, either by a third body such as GBIF, or by the institution itself.

Some of those in agreement with a standard method of generalization being recommended, suggested that there was need to ensure that the moderate restriction that would result be worth the investment.

It appears to the author that many collections would like to make data available, but were concerned that if it was not done securely or in a manner whereby the information could not be deduced through co-relational analyses, then taxa could be threatened through the release of information. Many collections will continue to restrict locational data from being available over the internet, others will make data available to bona-fide users if a secure method of doing so can be recommended, and others are willing to make data available in a generalized way if this document can recommend a suitable method of doing so.

5. The Case against a Standard Method of Generalizing

Responses to the Survey identified several areas where generalization of data may not be appropriate. For example, when the region encompasses political units of very small area (e.g. some island nations), or areas with small remnant vegetation patches in an otherwise cleared area.

Some suggested that having to generalize data for internet distribution could be time consuming and may involve them in having to do a lot of extra work. It was also suggested that only those familiar with the species can judge what information is too revealing and that it may have to be considered on a case by case basis.

Responses were overwhelmingly of the view that data custodians must maintain control over what data may be generalized, and what should continue to be hidden from view, and many were of the opinion that any recommendations on generalization be just that – recommendations and guidance only – and that if possible, several methods and scales of generalization should be recommended.

Responses included:

- "*A standard method would be very helpful, especially for smaller institutions without the time to evaluate/develop their own standards. But institutions should also be allowed to deviate from the standards, especially to allow greater protection of sensitive species data when desired*".
- "*Information being withheld by one institution would be withheld by all, but I can see a lot of time and effort being expended at developing the standards. Be sure the moderate restriction that would result is worth the investment*".
- "[Any standard should be] *explicit, accurate and implementable*".
- "*Two levels of dealing with sensitive data would be required …. Quarantine records must be completely hidden. Rare or threatened taxa protected under legislation and commercially valuable taxa would need to have localities generalized*."
- "*If the standard becomes known, then commercial dealers will more easily deduce the exact location we shall try to protect*."
- "*Whether to generalize should be up to individual institutions, and so should the level of generalization. For example, 0.1 degree precision might get you within the home range of an individual large mammal, which is dangerous, but nowhere close to the only tiny patch inhabited by a really rare plant, which is fine*."
- "*Only those familiar with the species can judge what information is too revealing*."
- "*Globalisation in database structures is more and more becoming absurd; the costs far outweigh the benefits*".

6. Technical Issues

Depending on the method (or methods) recommended the technical issues may be easy or difficult. Generalization by geographic grids such as 10 minutes, 1 minute, 30 seconds, are very simple to implement – metric grids such as 100m, 500m, 1 km, 10 km, etc., are a little harder, whereas developing a system of vetting of bona fide users and providing systems with secure log-ins etc., is the most difficult. But all are possible. Generalizing to a biogeographic or political region is simple to implement, and is generally only done in text thus restricting possibilities for spatial searching using bounding rectangles or some other method. If georeferences are given for data that are generalized to a biogeographic or political region, the result can be quite misleading – a coastal species, for example, may end up with a georeference that is hundreds of kilometers inland, reducing usefulness for analysis or data cleaning. Making such data available without suitable documentation can lead to quite disastrous results for users. It is probably better in these cases to not supply a georeference. Note that reporting by a geographic grid (using latitude or longitude) without randomization, moves the point in basically one direction, i.e. toward the south west as geographic grids are reported using the bottom-left hand corner of the grid (figure 3A) (Chapman *et al.* 2005). A metric grid on the other hand is often referenced from the center of the grid, however this may vary (figure 3B).

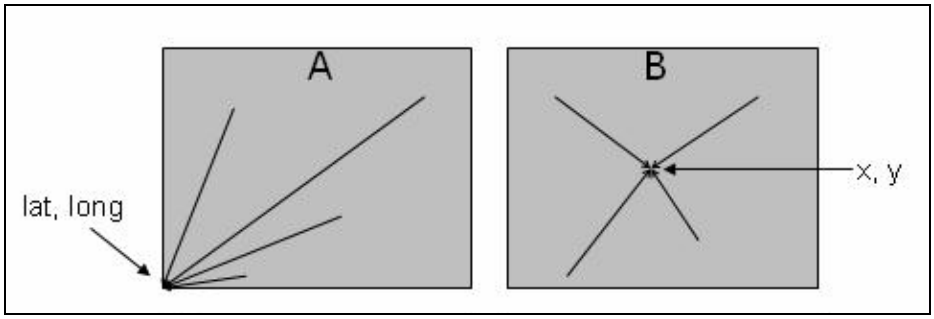

Fig. 3. Two generalization methods. A. a geographic grid where all records are referenced to the bottom right-hand corner. B. a metric grid where all records are referenced to the centroid.

Perhaps the greatest technical issue to overcome is having the data digitized in the first place. Any form of geographic generalization of the data depends on having the geographic coordinates available and it is currently estimated that only about 1% of the estimated 2.5 billion biodiversity collection records carry any geographic coordinates at all (Guralnick *et al.* submitted). On the other hand, this provides a unique opportunity to begin a process now that will be useable for the majority of data still to be georeferenced.

7. Social and Political Issues

The social and political issues with respect to the generalization of data are probably the most difficult to confront. It is largely for this reason, more than any, that there needs to be a series of choices on the scale, and perhaps method of generalization to suit all groups. What is most important, however, is the need to document what is being done in each case.

A. One case doesn't fit all

It is obvious from the survey that 'one case does not fit all'. There are good reasons why some agencies wish to hide data on certain taxa completely (quarantine is one case that was cited), or provide different levels of generalization for the different categories of threatened species, etc.

In some cases, national legislation may dictate what can be done and what data may not be legally released. This may be from the geographic locations of endangered species through to information on living people through Privacy Acts. It is not possible or desirable to dictate to institutions on how they should deal with their data; however guidelines on how best to restrict certain information while at the same time making the data useable appear to be wanted and needed.

On the other hand, some institutions have found benefits in working with the general public to gather information and to assist in the protection of rare taxa, using them to survey existing locations and to help locate new locations. There are good examples with birds, lizards, frogs, butterflies, corals and various plant species in a number of countries. Several people have raised the issue of the balance between protecting taxa through knowledge of where they occur as opposed to protection through restricting knowledge of their occurrence at a location. This becomes very taxon specific and certain taxa may be in greater danger due to inadvertent destruction through lack of knowledge than through deliberate collection and destruction through knowledge of the locations.

B. Competing Politics

There are many examples of cases where species have been endangered through knowledge of where they occur. For this reason, the locations of many of these species have been kept secret. A good example is the Wollemi Pine (*Wollemi nobilis*)[14] whose location and distribution have been kept secret while cultivating large numbers for the nursery trade[15]. In this way, it was hoped that the likelihood of piracy and pressure on the native population would be reduced.

On the other hand, many species have been endangered through lack of knowledge of their occurrence at a particular place. This often occurs through incidental destruction during road maintenance, farming and grazing, urbanization, etc. One reason often cited for species loss is from amateur collectors and biodiversity pirates, but often amateur collectors and interest groups know more about the locations of many sensitive taxa than do the professionals. But which groups are responsible groups and

[14] The Wollemi pine – a very rare discovery – Royal Botanic Gardens, NSW. http://www.rbgsyd.nsw.gov.au/information_about_plants/wollemi_pine

[15] Growing the Wollemi Pinehttp://www.rbgsyd.nsw.gov.au/information_about_plants/wollemi_pine/growing_it

which are not is a difficult question and is probably parallel to knowing who the criminals are in our society and who are the genuine citizens. Wherever possible there would seem to be advantages in organizations working closely with amateur organizations in the protection of sensitive taxa. There are many examples of successful partnerships that are aiding the protection of threatened and other species. In the UK, many amateur insect and plant groups aid in the recording and documenting of threatened species[16]. Across the world, amateur bird-groups assist in the recording of the locations of birds[17]. In Australia, the USA, Canada, etc, amateurs have been used for the recording of frogs[18]. In Australia, *The Banksia Atlas* (Taylor and Hopper 1988), involved over 400 amateurs and resulted in two new species and several new varieties.

There are a number of examples in coral-reef fishes where a new species has appeared in the commercial trade soon after, or in some cases even before, the species is scientifically described (*pers comm.* Richard Pyle[19] - see box).

A few examples with which I have had direct experience include:

Centropyge boylei
http://www2.bishopmuseum.org/natscidb/?pt=i&iID=-1386875360

Centropyge narcosis
http://www2.bishopmuseum.org/natscidb/?pt=i&iID=-2078333864

Belonoperca pylei
http://www2.bishopmuseum.org/natscidb/?pt=i&iID=-2140711607

...among a number of others.

Often in these and other cases, the existence of the new species is brought to the attention of the scientific community *by* the commercial (aquarium) trade; rather than the other way around. Thus, it is usually not considered so much of a "problem", but rather a sort of "symbiotic" relationship between the commercial trade and the taxonomists. Moreover, in most such cases in reef fishes, the species has eluded prior discovery not so much because it is rare or has an extremely restricted distribution, but because it simply lives somewhere that scientists have not yet been able to survey. Hence, there are usually few, if any, conservation implications in this context.

C. Duplicates

In plants, especially, (but also with other taxa such as insects) many collections are carried out in bulk and 'duplicates' (or parts of sets) are sent to many institutions around the world. This is usually in the order of 4-6, but examples of up to about 80 have been cited[20]. It has been recorded that 66% of collections in US Institutions that were collected outside of the USA are duplicates from another institution[20]. The

[16] Wikipedia – The Banksia Atlas. http://en.wikipedia.org/wiki/The_Banksia_Atlas
[17] Atlas of Australian Birds http://www.birdsaustralia.com.au/atlas/index.html;
North American Bird Breeding Atlas Explorer http://www.pwrc.usgs.gov/bba/.
[18] Frog Watch (Northern Australia) http://www.frogwatch.org.au/ ;
Frogwatch (USA) http://www.nwf.org/frogwatchUSA/ ;
Frog watch (Ontario) http://www.naturewatch.ca/english/frogwatch/on/ .
[19] *Pers. comm*. Richard Pyle, Bishop Museum, Hawaii (June 2006).
[20] *Pers. comm.*. P.J.Morris, Academy of Natural Sciences, Philadelphia, PA, USA.

problem that arises is that the originating institution loses control of what may happen to the information (including locality information) that may be distributed with those collections from those secondary institutions. In most cases this is not a problem, but with sensitive taxa, it often is. The secondary institution may not know what are regarded as 'sensitive taxa' in the jurisdiction of the originating institution, or may not have flagged that information. Sensitivity is not always information that can be distributed along with the collections, as it may not be known till much later that the species is endangered, etc. and thus sensitive. This is a difficult issue as just labeling a taxon as sensitive may not be the answer as a taxon that may be endangered in its native area (and thus sensitive), may be a weed or pest in other areas and locality information may be important for its control.

Perhaps the only real way of handling this is via the use of Globally Unique Identifiers (GUIDs) – see discussion on TDWG Web site[21], and possibly using filtered push technologies (Macklin, *et al.* 2006). Thus the originating institution could (automatically) notify collections holding duplicates of any change in status of the taxon, allowing for flagging in those institutions. Alternatively, as the originating institution makes data available via the GBIF Data Portal, the GUID could be used to identify duplicates and thus automatically generalize the duplicate records as well. This method may not be satisfactory if the originating institutions are not making their data available via GBIF. This issue needs further discussion.

[21] Globally Unique Identifiers (GUID) http://www.tdwg.org/TDWG_GUID.htm

8. Options

> Responsibility for information about the accuracy and reliability of the data, and restrictions of access to *sensitive data,* resides with the *data provider*. (GBIF 2004)

There are many ways of dealing with sensitive primary species occurrence data. Several are suggested below, and the workshop that will follow the production of this draft document is sure to identify more.

A. Standards versus Guidelines

When this project began, it was suggested that a standard may be developed (possibly through the Taxonomic Databases Working Group (TDWG)), however, the further the project has gone, the more apparent it has become that such a formalized process may not be the best solution.

Many of the respondents to the GBIF survey suggested that a series of recommendations and guidelines on methods for dealing with sensitive data would be more acceptable to data providers, and may encompass a range of methodologies, including

- generalization of data (both spatial and non-spatial),
- providing secure access to bona-fide users,
- dealing with temporarily sensitive data (such as awaiting publication),
- dealing with privacy issues for living persons,
- using data sharing licenses and agreements,
- developing and maintaining lists of both globally and regionally sensitive taxa, and
- providing record level metadata and documentation of what is being done.

Recommendation:

1. *A guide to best practices for dealing with sensitive primary species occurrence data be developed and made available via the GBIF Web site.*

B. Identifying what are Sensitive Taxa

As discussed above, a method for identifying what taxa are sensitive is needed. Suggestions have included:

- Developing a global list of sensitive taxa (somewhat akin to the CITES Appendices[22])
- Developing a global list of sensitive taxa as above with additional fields for geographic extent of sensitivity.
- Tagging of ECat[23] records with a sensitivity code
- Using the IUCN Red List of Threatened Species[24]

[22] CITES Appendices http://www.cites.org/eng/app/index.shtml
[23] GBIF Electronic Catalogue http://www.gbif.org/prog/ecat
[24] IUCN Red List of Threatened Species http://www.redlist.org/

- Leaving it up to individual data providers and use GUID and filtered Push Technologies to identify duplicates of sensitive records

All of these methods have their strengths and weaknesses.

1. Geographically sensitive data

In some cases taxa may only be sensitive over part of their range – in one country and not another, for example. A species may be common in the United States, but have a restricted distribution just across the border in Mexico and may be regarded as endangered within Mexico. For this reason, the Mexican Government may want to restrict information on localities in Mexico, whereas the American Government may have no such restrictions in the United States. For this reason, it is sometimes not practical to restrict or generalize information based just on a taxonomic name. For this reason, a list of sensitive taxa may require to be annotated with a geographic region of sensitivity.

2. Global lists of sensitive taxa

The production and maintenance of a global list of sensitive taxa is a resource intensive task. It is unlikely (nor is it desirable) that such a list have legislative backing, so its use would be purely voluntary – the final decision on whether data are made available or not must be up to the data providers and custodians. Drawbacks of such as list include its resource intensive nature, political and social issues in developing agreement on what should and what should not be included, and the regional nature of sensitivity for some taxa. I believe that if such a list is to be used, it should include provision for the addition of regional restrictions.

An advantage in using a list like the CITES Appendices is that it is hierarchical and thus all species of a genus, or all species of a family can be included with just one word and then look-up tables to ECat etc. used for any filtering. Ideally such a list would be able to cater for "include" and "exclude" capabilities so that all of one genus may be included except for one species, for example. This would make it easier to develop such a list in the first place and for it to be maintained.

3. Tagging of ECat Records

Similar drawbacks apply to this method as to the Global lists – viz, the difficulty in developing and maintaining such a list. Such a list could be used either as a guide to data providers prior to their making data available, or as a filter on the GBIF Data Portal. The use of a filter may be seen as over-riding the wishes of data providers to make data available, and is unlikely to work on a spatial or geographic basis (where taxa are only sensitive over part of their range). I believe that this is less desirable than the development of a separate list. [The separate lists could, of course be linked to the ECat].

4. Use of IUCN Red List of Threatened Species

Not all sensitive taxa are threatened species, and thus the Red List won't cover all taxa involved – although it may make a good starting point for development of such a list. It does not cover the other categories of sensitive taxa identified in the on-line survey for example. Also, the Red List does not cover all species regarded as threatened (and sensitive) at a regional or national level. In addition, as discussed previously, not all threatened species are endangered through knowledge of their

locations. For these reasons, I don't believe that the Red List itself is suitable as a list of sensitive taxa.

5. GUID and Push Technologies

There are very few examples of these technologies being used for the type of task that is being suggested here. GBIF and TDWG are about to introduce GUIDs for linking information on species, and have discussed the possibility of using them to identify duplicates, although how this might be done has still to be determined. An example of using filtered push technologies (see glossary) with duplicate records was presented at the SPNHC-NSCA meeting in Albuquerque, New Mexico in May 2006 (Morris *et al.* 2006). A combination of the two methodologies (although as yet untried) may be well worth exploring. These methods, of course will only work with duplicates of a record identified by the originating (or maybe secondary) institutions – it would be applied at the record level rather than the taxon level.

6. Combination of a Global List, GUID and Push Technologies

The ideal method, from the author's view, would be a combination of a global list of sensitive taxa (with the optional inclusion of geographic attribution), along with GUID and push technologies at the record level. The practicality of this needs to be further discussed at the forthcoming workshop, and perhaps through further listserver discussions.

Recommendation:

2. *That the development of a global list of sensitive taxa similar to the CITES Appendices (with optional geographic attribution) be explored.*
3. *That the use of GUID and Push Technologies for the identification of duplicate/related records and the (automatic) exchange of information (including sensitivity) be explored.*

C. Randomization versus Generalization

Few of the respondents to the on-line survey recorded that they randomize data as opposed to generalizing it. Reasons for not randomizing included the extra work and computation involved, the increased chance of mistakes being made, and the less reliability that users may be able to place in the data. One respondent suggested that:

> "*generalization creates/retains 'true' data, whereas randomization creates deliberately 'false' data. Generalization can be implemented (or not) on a case-by-case basis, depending on the intended use(s) of the data. Generalization is easily implemented by simply omitting data fields containing more precise data, and supplying data in tabular instead of georeferenced format*".

Others[25] pointed out that they were comfortable with displaying presence/absence of sensitive data within large polygons or grids squares, etc., because it still reflected the real data, but were aghast at the idea of deliberately 'faking' point coordinates such that locations appear as precise representations, but are randomly offset from the real data – i.e., they represent the deliberate introduction of error.

[25] For example, *pers. comm.*, James Morefield, Nevada Natural History Program, (Jun 2006).

At the University of Colorado, a "jitter" algorithm is used that randomly offsets both the x and y coordinates for the point with the option to specify both minimum and maximum distance variables for the offset[26]. There has been some suggestion that such an algorithm could be applied at the time data are requested from a provider (for example via GBIF), so that x, y values might change each time the data are requested. The value of this would need to be weighed against the drawbacks and difficulties of implementation. Would a different 'x, y' each time make a difference to an individual data requester? How often would the one person request the same data and thus get a different value? And what are the advantages of such a method over randomizing each record just once. One drawback may be that if a user carried out enough samples, and averaged, it could lead to a degree of precision close to the original for the record. These are issues that need exploring further.

A similar set of fuzzy algorithms are used by the Georgia Natural Heritage Program in the United States, using GIS Algorithms (mainly scripts using ESRI's ArcView® 3.x) (Krakow 2003). A drawback of such methods being universally recommended is that many natural history collection institutions have a low level of GIS implementation and knowledge, and thus may find the use of such technologies difficult or impractical.

Generalization (at least in a spatial sense) is usually of one of two types, viz.

- Generalization to a grid (metric or geographic)
- Generalization to a polygon (socio-political region, country, biogeographic region)

Many respondents to the survey argued for the simplicity of generalization to a grid, the simplicity of being able to vary the scale for different categories of sensitivity, the ease of maintenance and training, and the simplicity of creating suitable documentation. Some also suggested that while protecting the exact locations of sensitive taxa, it provided data in a format that was still useable for a majority of users, especially where a standard grid was used.

Where data are generalized to a geographic or biogeographic region (a polygon), the data have less usability for many analyses, but was seen by many as a more secure way of 'hiding' sensitive data locations. There are some parallels with this method with the reporting of census results in many countries where summaries are reported using Statistical Local Areas to restrict possible identification of individuals. A difference is that results are summarized over many individuals within a region, whereas with biological data we want to hide the location of a single entity within an area. It does *de facto* produce a summary, but this is not the primary intent. One problem with this method is that there is no guarantee that political (or even biogeographic) boundaries will remain constant over time and this further reduces the value of the data for many purposes. This has been found to be a problem when comparing some census data over time.

Another parallel is with geographic mapping:

> *Generalization is closely related to map scale. As we move from larger scale to smaller scale maps, we cannot show all of the detail that could be represented on the larger scale map. In order to maintain map legibility, it becomes necessary to generalize features on the map. As a result, maps at*

[26] *Pers comm.* David Neufeld, University of Colorado (April 2006).

different scales are useful for different purposes and the map designer must carefully balance the choice of map scale and consequent generalization requirements with the needs of the intended map user[27].

This relationship between generalization and use is an important one to keep in mind.

Recommendations:

4. *That generalization be the method of choice for protecting the exact localities of sensitive taxa in cases where data are made available via the internet.*
5. *That methods of randomization be explored for possible recommendation for those that wish to use randomization techniques.*

1. Methods of Generalizing Data

There are several main methods for generalizing data. The first three are grid based (i.e. have regular, or rectangular boundaries), and the last two are polygon-based.

- Geographic grid
- Metric or similar grid
- Map sheet
- Political or sociological region
- Biogeographic region or watershed

Generalizing to a geographic grid is the easiest to implement, especially if the data are stored as geographic coordinates such as degrees, minutes and seconds or decimal degrees. Generalizing can be done by simply removing the seconds, or last decimal place of the minutes, or by rounding decimal degrees to one or two decimal places. See Table 2 for the approximate area covered by different sized metric and geographic grids. Because the size of a geographic grid varies with latitude (largest at the equator), the area has been estimated at 30 degrees of Latitude.

Generalizing to a political, sociological or biogeographic region is usually done through use of text, with no georeferencing information supplied (although in rare cases, a bounding box or polygon may be supplied as part of the data). It is very easy to implement and provides a high level of security to the data, however, it also greatly restricts the uses to which the data may be put – especially as the regions vary greatly in size and shape. As mentioned previously, such regions can vary over time, leading to misleading information (if the data are stored that way), and makes comparison over time more difficult. The use of biogeographic regions would appear to make more sense from a biological point of view; however such regions are not universally accepted except for some areas of the world. If such methods are used, it is better that they be generated on export from the database to cater for any changes that may occur over time, however, this may require levels of technology beyond many institutions.

These are issues that need further discussion at the workshop before a recommendation can be made.

[27] *Cartographic Abstraction* in Dudycha (2003). http://www.fes.uwaterloo.ca/crs/geog165/cartabs.htm

Grid Size	Approximate area (at 30 degrees Latitude)
0.1 second	16.8 sq m
1 second	1,681 sq m
0.01 minute	600 sq m
0.1 minute	60,000 sq m
1 minute	6.0 sq km
10 minutes	600 sq km
30 minutes	5395 sq km
1 degree	21,580 sq km
0.1 degree	215.8 sq km
0.01 degree	2.16 sq km
0.001 degree	21,600 sq m
0.0001 degree	216 sq m
0.00001 degree	2 sq m
100 X 100 m	10,000 sq m
200 X 200 m	40,000 sq m
1 X 1 km	1 sq km
10 X 10 km	100 sq km

Table 2. Approximate area covered by geographic and metric grids of varying sizes at 30 degrees of Latitude.

2. Handling People's names and Determinavit Data

There is great resistance throughout the biodiversity community to the idea of hiding the names of determiners of specimens[28]. There would appear to be less resistance to hiding data on the names of collectors for a number of reasons. It seems that the biological community may have been inadvertently caught up in aspects of privacy legislation in a number of countries. The implications of these laws, and their applicability to our science, need to be explored. That is beyond the scope of this report; however, the author has begun a process of developing a simple standard for reporting on taxonomic verification with the aim of developing a TDWG standard in the near future. This is still in the early stage of discussion and thus there is a long way to go.

The current suggestions are that:

1. Where possible, the name and the date of the determiner be cited (but see discussion above on privacy considerations)
2. The basis on which a determination was made be cited, for example[29]
    a. Holotype or part of the type collection
    b. Compared with the holotype, isotype, etc.
    c. Compared with material from herbarium/museum xyz
    d. Run through so-and-so's key
    e. Identified using xyz Flora
    f. Compared with a figure in such-and-such field guide
    g. So and so told me it was this species, etc.

[28] See discussion on Taxacom Listserver Archive , June 2006, Discussion on '*Privacy Laws and Science*' <http://mailman.nhm.ku.edu/pipermail/taxacom/2006-June/thread.html>.
[29] Derived from *Pers comm.*. Dan Janzen, University of Philadelphia.

3. The level of expertise and certainty in the determination be recorded, for example

A-1	identified by **World expert** in the taxa	with **high certainty**
A-2	identified by **World expert** in the taxa	with **reasonable certainty**
A-3	identified by **World expert** in the taxa	with **some doubts**
B-1	identified by **regional expert**	with **high certainty**
B-2	identified by **regional expert**	with **reasonable certainty**
B-3	identified by **regional expert**	with **some doubts**
C-1	identified by **non-expert**	with **high certainty**
C-2	identified by **non-expert**	with **reasonable certainty**
C-3	identified by **non-expert**	with **some doubts**
D-1	identified by **the collector**	with **high certainty**
D-2	identified by **the collector**	with **reasonable certainty**
D-3	identified by **the collector**	with **some doubts**
U	**unknown**	

4. The reason why a determination may not be of high certainty, for example, "the specimen is damaged, poorly preserved, sterile, or is an undeveloped juvenile", etc.

The ideal would be some combination of these and suggestions as to how this may be done are welcome.

Recommendations:

6. *That GBIF begin a process to explore the implications of Privacy Legislation in different countries to the distribution of the names of collectors and determiners of biodiversity data.*
7. *That TDWG explore the worth of a standard on Taxonomic Verification along the lines outlined above.*

3. Generalizing Locality Data

Generalizing the georeferencing data, while leaving detailed locality descriptions, would appear to defeat the purpose of generalizing in the first place. If the georeferencing data are being generalized (or are not present for some reason), the locality data may also be generalised by:

1. Removing the locality information altogether and replacing with something like:

> *This specimen represents an endangered or threatened species. The specific locality has been removed from the on-line record to protect this species from over-collection. These data may be supplied to researchers on request.*

2. If the georeferencing has been generalized, then something like:

> *This specimen represents an endangered or threatened species. The specific locality has been generalized to presence within a grid of 1 minute resolution. Detailed data may be supplied to researchers on request.*

Alternatively, the detailed locality information may be removed and just the county or State information left, for example:

Orange County, California. [Data generalized]

I believe it is important to always include the information that the record has been generalized and there is more information available (See documentation below).

Recommendations:

8. *That when locality information is removed or generalized that this always be documented.*
9. *That standard forms of wording be developed and recommended for use where locality data are removed for modified prior to distribution.*

4. Dealing with Temporarily Sensitive Data

As previously noted, one large group of sensitive data are temporarily sensitive – e.g., data awaiting publication or results of research. It would appear, that in most cases at least, that these data not be released until the sensitivity no longer applies, or the sensitive portion of the data (be it the locality information, the name, etc.) be restricted or generalized as for other data. It is important that all temporarily sensitive data be time stamped such as "*for release after 1 Jan 2008*", etc. This provides users with some certainty and stops data being tied up for years and years from other researchers.

Recommendations:

10. *That data regarded as temporarily sensitive be time stamped for release at some definitive time in the future.*

5. Using Data Sharing Agreements and Data Licensing

Data Sharing Agreements and Data Licensing are common ways of managing access to sensitive data. Generally these are done on a one-to-one basis with data licenses drawn up individually in each case. This can be quite time-consuming, but provides the greatest control over who uses the data and how.

There are examples, however, of more general data licensing, usually by agencies making data available through third-party arrangements. These can be very broad (such as GBIF's current data license agreements) to more restrictive licenses such as are used by the National Biodiversity Network in the UK. This is similar to the way many software companies operate and provides little real control over who uses the data and how.

A third method is automatically generated on-line agreements such as used by the Australian Department of the Environment and Heritage for some non-biodiversity data (Freeman *et al.* 2000). This method, used in conjunction with secure logon by bona-fide users (see below), provides the most cost-effective method, while providing some control over who users the data and how.

The method developed by Freeman *et al.* (2000), requires the user to search the metadata directory for a dataset (see Discover Information Geographically), and then when they choose to download a dataset, they are asked to fill in a form on who they are, their address and email, and information on how they intended using the data. This information, along with parts of the metadata and standard licensing information are combined to automatically generate an individually crafted license agreement. Once the user accepts this, the agreement, the metadata and the data are packaged and

emailed to the user, with a copy of the agreement stored in the database. Advantages of the system are

- The user is required to enter into a legal license agreement with the Commonwealth;
- The agreement includes access and data use considerations and legal constraints as documented in the metadata within the license agreement;
- The data license is generated on-the-fly by the data dissemination facility at the time of data download;
- Overcomes paperwork and individual data packaging;
- Minimizes demands on the resources of both the client and the data supplier;
- Promotes ease and simplicity of data transfer between parties;
- Provides documentary evidence of who is using the data and for what purposes (although there is no guarantee that the information provided on proposed use is always truthful).

Agreements and Licenses

There are many different types of data license, data use agreements or data transfer agreements in operation. Llinás (2005) in a report to the Humboldt Institute in Colombia on Copyright, noted that there are two types of contract used for copyright – transfer contracts (similar to buying or selling) and license contracts (similar to leasing). However, he also notes, that many contracts now also impose restrictions on the way the information may be used by the recipient of the data which would appear to be opposed to the way strict copyright laws are written. The most common type of agreement used for the transfer of biodiversity data would appear to be a data license agreement in which the data are not transferred, but are provided for use by a second party. The difference is between an object being transferred to a second party, with the originator retaining no, or few, rights in it, and an object being transferred for use, but where there is no diminishment of the rights of the original owner to continue using it.

Creative Commons – a non-profit organization that promotes more flexible licenses for creative and scientific works, supports a form of transfer agreement that is somewhat less restrictive than would be implied by Copyright – for example, their slogan is "some rights reserved", as opposed to copyright that reads "all rights reserved" (Llinás (2005). Such a compromise would appear to be highly applicable to the type of data we are wishing to transfer here.

A contract is an indispensable tool to regulate relations between parties (Llinás 2005)

When creating agreements, it is important that users be clear regarding what the agreement intends to protect, what type of information can be made available to the public, what restrictions there are on how the data may be used, what the obligations are to cite the source and how such a citation should be made, if there are copyrights to respect and any authorizations that must be requested in the case of a usage different to the one for which permission has been granted, etc. (Llinás (2005).

6. Identifying (bona-fide) Users

One of most difficult issues to confront is the identification of who are 'bona fide' users and who not. As raised by several people, even some taxonomists cannot be trusted with the data, whereas many amateurs can be. It has been suggested that some form of 'certification' would be of value, such that trusted users could get special

access rights when logged into the GBIF Data Portal. Some respondents, however, said that they would not be prepared to allow someone else to decide who could get access, and wanted to retain the right to decide themselves.

Perhaps there is need for a two level certification – one at a higher level that may be managed through the GBIF Data Portal, and then a finer layer that is managed by the individual data nodes or providers. This would then allow three levels of data provision – open access, access to mid level data by GBIF 'certified' users; and then restricted and one-off access through the data provider's own certification.

One approach to certification[30] that may be worth examining is that used by organizations such as EBay® and PayPal®, to manage both sellers and purchasers through their Web sites. GBIF (or some other trusted party) could allow people to register themselves and create a profile. They could then connect to the GBIF network and be authenticated as that user (as is done by many on-line services at the moment). Other sites may then make use of this authentication infrastructure to authenticate a user and to retrieve the user profile information. They could then choose to open up additional data fields (either through a UI or a web service) only to users that they approve. This agreement may take place offline between the provider and the candidate user and would be entirely managed by the data provider. The provider site should probably have the ability to write their permissions back into the users profile rather than having to manage them locally. It may be possible (as is done by EBay®) for a user's profile to be added to with trust statements from various provider institutions and networks and for sites to review this information to determine whether to allow a user access. The whole process would need refinement and careful consideration of privacy issues, but ought to be fairly simple and to carry out in a very flexible fashion. To overcome privacy issues, mischievousness writings in trust statements, and possible libellous statements, there may need to be some form of moderation of trust statements before these are made public and a method for profile owners to seek redress and correction of any errors or misconceptions.

Doug Yanega[31] has suggested the establishment of a non-profit professional society for taxonomists where one of the conditions of membership would include signing an agreement that their use of data would be 'honorable'. This could be one input into certification, but certification would need to be much broader as membership of such a society would restrict 'bona fide' users to a limited group of users.

7. Methods of Restricting Access and/or Providing Secure Access

The SYNTHESYS[32] (Synthesis of Systematic Resources) project of the European Union produced two reports in 2004 and 2005 on developing authentication services for system access (Tolksdorf, *et al.* 2004, Tolksdorf & Suhrbier 2005).

Walter Berendsohn[33] suggests the possible use of the SYNTHESYS methods within the BioCASe access environment in three contexts:

[30] Suggested by Donald Hobern, GBIF Secretariat, *pers. comm.*. 31 July 2006.
[31] Doug Yanega, Entomology Research Museum, University of California, Riverside , TDWG Listserve discussion <http://mailman.nhm.ku.edu/pipermail/taxacom/2006-June/thread.html>
[32] SYNTHESYS: Synthesis of Systematic Resources <http://www.synthesys.info/>
[33] *Pers. comm*.., Walter Berendsohn, Botanischer Garten und Botanisches Museum, Berlin-Dahlem (Mar. 2006).

i) for GBIF-style open external access,
ii) within institutions to integrate various data sources, and
iii) within secured networks.

These reports looked largely at access rights within the BioCASe[34] scenario and thus have large implications for this document. Due to the technical nature of those documents, however, it is probably counter-productive to elaborate on them here, but suggest that they may be examined in more detail as a possible solution for restricting access, possibly at the workshop or during a subsequent process.

A second group of technologies that may be worth pursuing are filtered push technologies as mentioned previously (Macklin *et al.* 2006, Morris, *et al.* 2006). It is recommended that one of these authors be invited to the proposed Workshop to discuss these technologies and their possible role in restricting access to sensitive data.

The NBN Gateway website includes a framework of access controls developed to encourage data holders to have a go at sharing information over the Internet through the Gateway. The NBN Gateway access controls provide a secure environment through which data can be communicated from a data provider to different users at varying levels of detail. The data exchange principles[35] embody the important concept that you should always begin from a position of open access and then work back from that where truly necessary. This approach has recently been embodied within the UK Environmental Information Regulations 2004. The NBN Gateway access controls allow the sharing of detailed sensitive information over the Internet. Rather than the provider restricting the detail of the data, they can submit the full detail to the NBN Gateway and then use the access controls to block or limit public access to the full detail. They can give the public a summary level of access and share the full detail with registered individuals or organisations that they trust (e.g., have an exchange relationship with) or, in the case of environmentally sensitive information, that need to know (*pers. comm.* Oliver Grafton[36]). The Data Access Constraints and other documents used by the NBN Gateway (many of which are available from the NBN website) are worth examining as a possible parallel solution for GBIF, and I suggest that a representative be invited to the proposed workshop.

Recommendations:

11. *That the SYNTHESYS documents on authentication services for system access be examined for their applicability to establishing a system for restricting access to data on sensitive taxa through the GBIF Data Portal.*
12. *That representatives of the three systems discussed here, SYNTHESYS, Push Technologies and the NBN Gateway, be invited to the proposed workshop to discuss access constraints and authentication services.*

[34] BioCASe: Biological Collection Access Services <http://www.biocase.org/>
[35] NBN Data Exchange Principles <http://www.nbn.org.uk/downloads/files/DataExchange%20principles%202002.pdf>
[36] *Pers. comm..,* Oliver Grafton, National Biodiversity Network, in response to GBIF Survey.

D. Documentation

Documentation is one of the most essential, but most neglected aspects of dealing with generalizing information. It is essential that users know what has been done to the data in way of generalization, etc. to be able to determine if the data are fit for the use to which they want to put them. Without such documentation, the data are unreliable and thus reduces the benefit of making the data available in the first place.

The use of metadata to describe data and data sets is now common practice. The biodiversity community was a little slow to adopt it, but is now using it extensively to describe their datasets, along with access constraints, conditions of use, etc., but there is still a long way to go before all our datasets are consistently documented. One problem is that there is no universal standard for metadata in this area, although some standards have been developed in some regions (for example the NBII Metadata standards[37] in the USA). Perhaps TDWG could look at developing a universal metadata standard for biological collection and observation data, possibly using the NBII standards as a starting point.

Record-level metadata, however, is still not used extensively, and thus much of the data being distributed via the GBIF Data Portal is without detailed information as to what has been done to the data in the way of quality control, accuracy, data validation, or generalization, etc.

As stated by Llinás (2005),

> "*Metadata fulfils an essential function regarding communication to third parties, of access constraints and use conditions that the data generators intend to give to their data. It can be considered as an 'aid' in protecting data and information, since it will allow system users to visualize the conditions established by the data generator for access and use of the information. Additionally, in case the data are not accessible, the metadata allows knowledge of the conditions of access through other media (digital or not) as well as a summary of the content*".

Even though the metadata itself is not a mechanism of protection, it facilitates it.

1. Record Level Metadata

Record-level metadata is not an extensively used concept, but has been in use in some areas for distributed biodiversity data since the early 1990s if not earlier. Basically, instead of recording just information for the database or dataset as a whole, information that may be specific to each record is recorded with that record. Where this is used, it is often not regarded as metadata (such as accuracy or uncertainty in the georeferencing information), but it is metadata as it is added later about the data, and is not actually part of the data itself. Information that can be added in this way is extensive and should include such information as the accuracy of the identification, georeferencing, etc. Information that is consistent across the whole dataset, would, of course, continue to be recorded in the dataset level metadata. Only those aspects that differ from the dataset level metadata would be recorded at the record level.

With respect to sensitive taxa, additional information that should be included at the record level could include:

[37] National Biodiversity Information Infrastructure (NBII) Metadata Standards for Biological Data <http://www.nbii.gov/datainfo/metadata/standards/>

- any information on access constraints that may apply to individual records (or data fields) where these may differ from those of the database as a whole (such as for sensitive data)
- information on any modification to the record that leads to the provision of data that varies from the original data, such as[38]:
    - Information Withheld – where attribute information that exists in the source database is withheld from the public
    - Information Modified – where attribute information that exists in the source database has been modified in some way and which causes loss or alteration to the data that are made available to the public
    - Spatial Fit – for where georeferencing information has been altered or modified (for example, through generalization), and provides an indication of the goodness of fit of the resultant georeference compared to the original (see next section).
- details that expand on the last set, such as how the information is modified, what information is being withheld, etc. (see discussion above under '***Generalizing Locality Data'***) or conditions under which withheld information may be accessed, etc.

Recommendations:

13. *That TDWG consider developing a metadata standard for biodiversity collection data at the data-set level (with perhaps a record-level extension).*
14. *That the final 'Guidelines on Dealing with Sensitive Primary Species Occurrence Data' include recommendations on recording record-level metadata.*

2. Spatial Fit

Spatial Fit is a concept that has arisen out of the BioGeomancer project and provides a measure of how well a geometric representation matches the original spatial representation. In the case of sensitive data, where a georeference is generalized to a grid or biogeographic region, it is a mechanism to provide users with an indication of how well the information made available to the public matches the georeferencing information held in the original database. Spatial fit is a value of either zero, one or greater than 1, where 1 represents an exact match (i.e. the data has not been generalized).

Details on how Spatial Fit may be calculated can be found in Chapman and Wieczorek (2006), where the summary below can also be found:

> *A spatial fit with a value of 1 is an exact match or 100% overlap. If the geometry given does not completely encompass the original spatial representation, then the spatial fit is zero (i.e., some of the original is outside the transformed version, which we interpret as not being a fit). If the transformed shape does completely encompass the original spatial representation, then the value of the spatial fit is the ratio of the area of the transformed geometry to the area of the original spatial representation.*

[38] *Pers. comm.*, John Wieczorek, University of California, Berkeley (June 2006).

Special case: If the original spatial representation is a point and the geometry presented in not a point, then the spatial fit is undefined.

An example of its applicability is where a georeference with an uncertainty radius of 1 km (using a point radius method) is made available using a 10 km grid (which completely covers the uncertainty). In this case the Spatial Fit would be greater than 1 as it represents an area greater than the real uncertainty. Actually, in this case (as shown in figure 4),

Spatial Fit = 31.8 – i.e. (10^2/(PI*r^2))

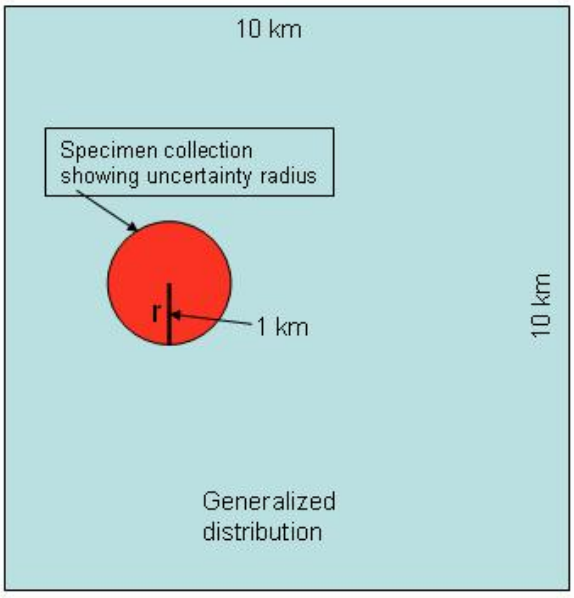

Fig 4. Example of calculating Spatial Fit for a collection with an uncertainty radius of 1 km (red circle), and which is distributed using a 10 km square grid (blue).

The smaller the grid size, the closer the Spatial Fit will be to '1'.

Note, that a record that has its georeference randomised or generalized such that a portion of the uncertainty radius falls outside the grid square would have a Spatial Fit equal to zero.

Recommendations:

15. That the final 'Guidelines on Dealing with Sensitive Primary Species Occurrence Data' include recommendations on using Spatial Fit to document how well generalized match fit the original data.

E. Interchange Standards

One of the questions asked in the on-line survey was what changes need to be made to the ABCD and Darwin Core standards to cater for the exchange of data on sensitive taxa.

Responses to the survey indicated that there needed to be extra fields, and there were a number of suggestions as to how this may be done. One suggestion (from John Wieczorek[39]) suggested fields to cater for

[39] *Pers. comm.*, John Wieczorek, University of California, Berkeley (June 2006).

- Information Withheld – where attribute information that exists in the source database is withheld from the public
- Information Modified – where attribute information that exists in the source database has been modified in some way and which causes loss or alteration to the data that are made available to the public, and
- Spatial Fit – for where georeferencing information has been altered or modified (see previous section)

Other suggestions were for just a yes/no field such as:

- GeoreferenceIntroducedError – Yes/No

And yet others provided more complicated solutions.

All suggestions were passed to the relevant TDWG Subgroup Convenors for further consideration.

Recommendations:

16. That TDWG consider modifying the ABCD and Darwin Core standards to cater for information on generalization, etc. of data on sensitive taxa.

9. Recommendations

It is recommended that:

1. *a Guideline to best practices for dealing with sensitive primary species occurrence data be developed and made available via the GBIF Web site.*
2. *the development of a global list of sensitive taxa similar to the CITES Appendices (with optional geographic attribution) be explored.*
3. *the use of GUID and Push Technologies for the identification of duplicate/related records and the (automatic) exchange of information (including sensitivity) be explored.*
4. *generalization be the method of choice for protecting the exact localities of sensitive taxa in cases where data are made available via the internet.*
5. *methods of randomization be explored for possible recommendation for those that wish to use randomization techniques.*
6. *GBIF begin a process to explore the implications of Privacy Legislation in different countries to the distribution of the names of collectors and determiners of biodiversity data.*
7. *TDWG explore the worth of a standard on Taxonomic Verification along the lines outlined above.*
8. *when locality information is removed or generalized that this always be documented.*
9. *standard forms of wording be developed and recommended for use where locality data are removed for modified prior to distribution.*
10. *data regarded as temporarily sensitive be time stamped for release at some definitive time in the future*
11. *That the SYNTHESYS documents on authentication services for system access be examined for their applicability to establishing a system for restricting access to data on sensitive taxa through the GBIF Data Portal.*
12. *representatives of the three systems discussed here, SYNTHESYS, Push technologies and the NBN Gateway, be invited to the proposed workshop to discuss access constraints and authentication services.*
13. *TDWG consider developing a metadata standard for biodiversity collection data at the data-set level (with perhaps a record-level extension).*
14. *the final 'Guidelines on Dealing with Sensitive Primary Species Occurrence Data' include recommendations on recording record-level metadata.*
15. *the final 'Guidelines on Dealing with Sensitive Primary Species Occurrence Data' include recommendations on using Spatial Fit to document how well generalized match fit the original data.*
16. *TDWG consider modifying the ABCD and Darwin Core standards to cater for information on generalization, etc. of data on sensitive taxa.*

10. Glossary

Generalization: — refers here to any modifications carried out to source data to conceal sensitive content, typically by reducing the precision of the data (such as reporting at the level of a mapsheet, grid or county, citing just the nearest named place, or by deleting some parts of the data). In geographic terms it refers to the conversion of a geographic representation to one with less resolution and less information content; traditionally associated with a change in scale. Also referred to as: *fuzzying*, *dummying-up*, etc.

Georeference: — to translate a locality description into a mappable representation of a ***feature*** (*q.v.*) (verb); or the product of such a translation (noun).

Globally Unique Identifier (GUID): — a pseudo-random 128-bit number often used in software applications and which is being examined by TDWG and GBIF as a possible way of uniquely identifying individual collections and biodiversity objects[40].

Randomization: — refers to a deliberate haphazard arrangement of observations so as to obscure their true location. Randomization leads to a falsification of the data. Also referred to as *falsifying*

Push Technologies: — a data distribution technology in which selected data are automatically delivered into the user's computer at prescribed intervals, based on some event that occurs (such as a change in determination). Proposed uses in biodiversity include distribution of information on duplicate collections, and may be usable in notifying collections of changes in status in sensitivity, etc. With **filtered push technologies**, users can define their interests by registering a filter that is applied to all notifications.

Sensitive data: — any data, that because of their nature, a data provider does not want to make available in their raw state, e.g. precise localities of endangered taxa.

[40] Globally Unique Identifiers: TDWG/GBIF. <http://www.tdwg.org/TDWG_GUID.htm>

11. Acknowledgements

The 150 people and institutions that responded to the on-survey, and the many people that have emailed the author personally, or have taken part in the Listserver discussions and meetings in Australia and the United States have made this study possible.

In particular, I would like to thank all those who took the time to fill out the on-line survey. It produced many unexpected results, all of which will aid in the development of robust recommendations and hopefully improved and consistent methods for dealing with sensitive species-occurrence data.

Secondly, the participants that were involved in discussions at the SPNHC-NSCA Meeting in Albuquerque, New Mexico in May 2006, and the NCRIS meeting in Sydney, Australia in March 2006, and especially Chris Frazier (University of New Mexico,), Paul Morris (Academy of Natural Sciences, Philadelphia), Hank Bart (Tulane University) and Jorge Sóberon (University of Kansas).

I would like to particularly also thank Oliver Grafton (National Biodiversity Network, UK) who supplied me with the many documents used by that organization, Lee Belbin (TDWG, Hobart, Australia), Dave Britton (Australian Museum), Alex Chapman (Western Australian Herbarium), Anita Cholewa (Bell Museum of Natural History, University of Minnesota), Jerry Cooper (Land Care Research, New Zealand), Dan Janzen (University of Pennsylvania, USA), Paul Kirk (CABI International, UK), James Morefield (Nevada Natural History Program, USA), Dave Neufeld (University of Colorado, USA), Richard Pyle (Bishop Museum, Hawaii), Lawrence Way (Joint Nature Conservation Committee, UK), John Wieczorek (University of California, Berkeley, USA), and Doug Yanega (University of California, Riverside, USA) as well as GBIF staff members, Else Østergaard Andersen, Per de Place Bjorn, Donald Hobern, and Hannu Saarenmaa, for informative email discussions and advice.

Lastly, I would like to thank Larry Speers (GBIF Secretariat, Denmark) and Bruce Stein and Lynn Kutner (Natureserve, USA), for endless assistance in planning the on-line survey and for on-going support and assistance on this project.

12 References

Chapman, A.D. 2006. *Questionnaire on Dealing with Sensitive Primary Species Occurrence Data − Summary of responses*. 61 pp. Copenhagen: GBIF. http://www.gbif.org/prog/digit/sensitive_data/Summary_of_Responses_-_03.pdf. [Accessed 26 Jul. 2006].

Chapman, A.D., Muñoz, M.E. de S. and Koch, I. 2005. Environmental Information: Placing Biodiversity Phenomena in an Ecological and Environmental Context, *Biodiversity Informatics* **2:** 24-41. http://jbi.nhm.ku.edu/viewarticle.php?id=9&layout=abstract [Accessed 27 Jun. 2006]

Chapman, A.D. and J. Wieczorek (eds). 2006. *Guide to Best Practices for Georeferencing*. Copenhagen: Global Biodiversity Information Facility.92 pp.

David P. 2003. The Economic Logic of "Open Science" and the Balance Between Private Property Rights and the Public Domain in Scientific Data and Information pp. 19-34 in Esanu, J.M. and Uhlir, P.F. (eds). *The Role of Scientific and Technical Data and Information in the Public Domain: Proceedings of a Symposium*. Washington D.C.: National Research Council of the National Academies, 238pp. http://darwin.nap.edu/books/030908850X/html/19.html [Accessed 26 Jul 2006].

Dudycha, D.J. 2003. Geography 165. Introduction to Cartography and Remote Sensing, University of Waterloo http://www.fes.uwaterloo.ca/crs/geog165/cartabs.htm [Accessed 10 Jul 2006].

Freeman, N., Boston, T. and Chapman, A.D. 2000 Integrating Internal, Intranet and Internet Access to Spatial Datasets via ERIN's Environmental Data Directory **in** *Proceedings of the 26th Annual Conference of AURISA, Perth, Western Australia, 23-27 November 1998*. AURISA.

GBIF. 2004a. *GBIF Experts' Meeting on biodiversity data, databases and property rights issues*. *Royal Botanic Garden, Madrid, Spain, 1-2 March 2004*. Meeting Report. Copenhagen: GBIF

GBIF.2004b. *Global Biodiversity Information Facility (GBIF). Data Use Agreement*: Copenhagen: GBIF. http://www.gbif.org/DataProviders/Agreements/DUA [Accessed 27 Jun 2006].

Guterman, L. 2006. Endangered by Research: Poachers mine the scientific literature for the locations of newly discovered animals. *The Chronicle of Higher Education* 52(46): A12. http://chronicle.com/free/v52/i46/46a01201.htm [Accessed on-line 19 Jul 2006].

Krakow, G. 2003. GIS Algorithms Useful for Producing "Fuzzy" Rare Species Locations on *Presentations from the 8th Annual Meeting of the Organization of Fish and Wildlife Information Managers, Rapid City, South Dakota 25-29 September, 2003*. PowerPoint Presetation at http://www.ofwim.org/docs/2003/PPT/Krakow_OFWIM_2003.ppt [Accessed 25 July 2006].

Llinás, J.V. 2005. Data and Information on Biodiversity and its Protection in the Digital Realm Ver. 1. Bogotá, Colombia: Biological Resources Research Institute Alexandre von Humboldt. 43pp.

Macklin, J.A., Rabeler, R. And Morris, P.J. 2006. Developing a framework for exchange of botanical specimen data to reduce duplicate effort and improve quality using a 'filtered push'. *Botany 2006. California State University – Chico. July28-August 2, 2006*. Abstract. http://www.2006.botanyconference.org/engine/search/index.php?func=detail&aid=587 [Accessed 17 July 2006].

Malpas, J. 2004. *The Calflora Project: Goals and Achievements.* http://www.calflora.org/goalsAndAchievements.html [Accessed 27 Jun 2006].

Morris, P.J., Macklin, J.A. and Rabeler, R.K. 2006. Filtered push: Exploiting technical methods for efficient use of community knowledge to improve the quality of collections data. Abstract. P. 54 in *The Road to Productive Partnerships. SPNHC-NSCA 2006. Program & Abstracts.Albuquerque, New Mexico 23-27 May 2006.*

Stuart, B.L., Rhodin, G.J., Grismer, L.L. and Hansel, T. 2006. Scientific Description Can Imperil Species. *Science* 312(5777): 1137.

Taylor, A. and Hopper, S.D. 1988. *The Banksia Atlas (Australian Flora and Fauna Series Number 8)*. Canberra: Australian Government Publishing Service.

Tolksdorf, R. and Suhrbier, L. 2005. *SYNTHESYS D 1.2 Develop authentication services for system access Milestone Report M22: A documented operational component integrated in the system*. Berlin: Freie Universität. 47 pp.

Tolksdorf, R., Suhrbier, L. and Langer, E. 2004. *SYNTHESYS D 1.2 Develop authentication services for system access Milestone Report: An analysis of the requirements of the users including national specifics and a concept for the rights management and control component.* Berlin: Freie Universität. 65 pp.

Index

L

M

N

O

P

Q

R

S

T

V

W

Appendix: Best Practice Guidelines for Generalizing Sensitive Primary Species Occurrence Data

To be written following Workshop