GUIDE TO BEST PRACTICES FOR GENERALISING SENSITIVE SPECIES OCCURENCE DATA

Arthur D. Chapman and
Oliver Grafton

GBIF
www.gbif.org

GUIDE TO BEST PRACTICES FOR GENERALISING PRIMARY SPECIES-OCCURRENCE DATA

Chapman, A. D. and O. Grafton. 2008. Guide to Best Practices for Generalising Primary Species-Occurrence Data, version 1.0. Copenhagen: Global Biodiversity Information Facility, 27 pp. ISBN: 87-92020-06-2.

Introduction

The unprotected distribution of Sensitive Primary Species Occurrence Data (for example the exact localities of rare, endangered or commercially valuable taxa) has been a concern of the GBIF Secretariat since its beginning. In early 2006, GBIF initiated a process to address this issue, especially 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 generalising 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). An experts' workshop was then held in early March 2007 that focussed on the various technical issues involved (Chapman 2007a).

A final report on Dealing with Sensitive Primary Species Occurrence Data was developed following these processes and discussions, and was presented to GBIF in April 2007 (Chapman 2007b). It is available via the GBIF Web site. This report made a number of recommendations, and many of these are included in this document.

The final step in this process has been to develop a Guide to Best Practices. This document should be seen as an overriding guideline for institutions, data providers and GBIF Nodes to use to develop their own in-house guidelines. Organisations and institutions should produce their own internal document that incorporates the practices outlined in this document and related documents such as the Guide to Best Practices in Georeferencing (Chapman and Wieczorek 2006) and incorporate them into their own working environment.

It is also important to understand the possible impact that approaches for restricting sensitive data 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. For that reason, a set of principles are elucidated below. Key among these is the need to make biodiversity information freely available wherever possible, in the interests of science, the environment and the biodiversity itself.

Two issues that this document has not covered, because they will need further discussion and agreement before robust recommendations can be made, are the issues of the privacy of living individuals and the development of Data Sharing and Data License Agreements. Both of these issues have legal implications and vary considerably from jurisdiction to jurisdiction. Recommendations were made in the Report on Dealing with Sensitive Species Occurrence Data (Chapman 2007) for GBIF to further explore these issues.

> *"The term best practice generally refers to the best possible way of doing something; it is commonly used in the fields of business management, software engineering, and medicine, and increasingly in government. [...] The [qualified] term, 'best current practice', often represents the meaning in a more accurate way, showing the possibility for future developments of 'better practice'."*
>
> (Wikipedia: Best Practice).

[1] Survey Monkey http://www.surveymonkey.com

[2] http://www.gbif.org/prog/digit/sensitive_data/Summary_of_Responses_-_03.pdf

Principles

Biodiversity information should be made freely available to be shared globally to enable their use for not-for-profit decision-making, education, research and other public benefit purposes. Making the full detail of biodiversity information available should reduce the risk of damage to the environment and help safeguard a sustainable future. Where release will have the opposite effect, access to the full detail may need to be controlled.

Below are a set of high level principles related to the sharing of data generally, and the sharing of sensitive data in particular.

1. *Wherever possible, environmental information should be freely available to all. Generally this benefits the environment by increasing awareness, enabling better decision-making and reducing risk of damage.*
2. *In a small number of cases, public access to information can result in environmental harm. It should be recognised that in such cases, availability of information may need to be controlled; although the presumption remains in favour of release and any restrictions should be interpreted rigorously.*
3. *All data regarded as being sensitive should include a date for review of their sensitivity status, along with documented reasons for the sensitivity status. The date for review may be short or long depending on the nature of the sensitivity. Whenever a data provider receives an application for enhanced access to restricted data they should avoid assuming continued sensitivity and use it as an opportunity to revisit the determination.*
4. *If the data are to be restricted for distribution, then this should only be done to a copy of the data at the time of their distribution. Data should never be altered, falsified or deleted from the stored record.*
5. *Documentation is essential for many reasons, and where data have been restricted or generalised it is important that that information is recorded as metadata that remains with the record.*
6. *Where data are restricted or generalized for distribution (such as the name of a collector, textual locality information, etc.) this should be documented by replacing with appropriate wording − the field should not be left blank or null.*
7. *There are extremely strong reasons <u>not</u> to restrict data on related collections (e.g. collector's numbers in sequence, collector's name, etc.) because of the restrictions this places on data quality/ data validation procedures and the limits it places on the effectiveness of filtered Push Technologies.*
8. *Users of sensitive data should respect any and all restrictions of access that the data provider has placed on the data. If granted enhanced access to restricted information users must not compromise or otherwise infringe the confidentiality of such information.*
9. *Data providers should respect the needs of data users to have access to data and documentation in order to determine the 'fitness for use' of the data, and to ensure that analyses are robust and not misleading.*

Determining Sensitivity

As a first step, information holders need to identify any data which are regarded as ‘sensitive’. Sensitive information is any which if released to the public, would result in an ‘adverse effect’ on the taxon or attribute in question or to a living individual. A number of factors need to be taken into account when determining sensitivity, including type and level of threat, vulnerability of the taxon or attribute, type of information, and whether it is already publicly available. Determining these factors leads us to a criteria-based approach.

Two examples of sensitivity criteria that provide a starting point for the development of criteria are those developed by the National Biodiversity Network (NBN) in the UK (National Biodiversity Network 2002, 2004), and the Department of Environment and Conservation in New South Wales, Australia (Department of Environment and Conservation 2007).

Below are a series of criteria for determining the sensitivity of taxa and data along with recommended metadata statements for documenting the reasons for the determination. The first two are for use by biodiversity data holders and those creating trigger lists of potentially sensitive taxa and refer largely to the taxa themselves. The last two are for use by biodiversity data holders and deal with an assessment of the data they hold and are considering making available – they are not suitable for the creation of trigger lists.

The criteria are used to determine:

1. **Risk of Harm**	An assessment of whether the taxon is subject to harmful human activity.
2. **Impact of Harm**	An assessment of the sensitivity of the taxon to the harmful human activity.
3. **Sensitivity of Data**	An assessment on whether the release of data will increase harm.
4. **Decision on release & Category of sensitivity**	A balanced decision regarding the release of the data and a determination of the category of sensitivity, and thus the level of generalisation, of the data for release.

A set of scenarios using Criteria 1 and 2 below for determining triggers for sensitivity of taxa is attached as an Appendix to this chapter.

Criteria for Determining Sensitivity

The first step in the process of determining sensitivity is to make an assessment on whether or not the taxon is subject to a harmful human activity and if the availability of related biodiversity data will increase the likelihood of the harmful activity occurring. If it is not then there would appear no reason to list it as a potential environmentally sensitive taxon. It is recommended that you use the documented wording supplied but with additional supporting rational documenting the specifics of the threat, for example:

> “*The taxon is at risk from harmful human activity –it is subject to attack by Phytophthora which is transported by human operated vehicles.*”

1. RISK OF HARM **Assess whether the taxon is subject to a harmful human activity.**	
1.1. Is the taxon subject to a harmful human	**Yes:** ***Document using statement 1a with***

activity?	***supporting rationale.*** **Go to 1.2**
	No: ***Document using statement 1b*** **[Taxon is not sensitive] Go to 3**
1.2. Is there established evidence of current or recent occurrences of the harmful human activity?	**Yes:** ***Document using statement 1c with supporting rationale.*** **Go to 1.3**
	No: ***Document using statement 1d with supporting rationale.*** **Go to 1.3**
1.3. Will availability of related biodiversity data increase the likelihood of the harmful human activity taking place?	**Yes:** ***Document using statement 1e with supporting rationale.*** **Go to 2**
	No: ***Document using statement 1f with supporting rationale.*** **Go to 2**

1a – The taxon is at risk from a harmful human activity.

1b – There is no significant risk of a harmful human activity.

1c – There is established evidence of actual or recent harm to the taxon

1d – There is currently no established evidence of actual harm to the taxon.

1e – Availability of biodiversity data will increase the likelihood of the harmful human activity taking place.

1f – Availability of biodiversity data will not increase the likelihood of the harmful human activity taking place.

The next step is to determine if the taxon is sensitive to that human harm or whether it is suitably robust so as not to be adversely affected.

2. IMPACT OF HARM

Assess the sensitivity of the taxa to the harmful human activity.

2.1. Does the taxon have characteristics that make it significantly vulnerable to the harmful human activity?	**Yes:** ***Document using statement 2a with supporting rationale.*** **Go to 2.2**
	No: ***Document using statement 2b and supporting rationale.*** **Go to 2.2**
2.2. Is the taxon vulnerable to harmful human activity over its total range, or are there areas (such as in conservation zones, or other parts of the world) where the taxon is not at the same level of risk?	**Yes:** ***Document using statement 2c with supporting rationale.*** **Go to 3**
	No: ***Document using statement 2d with supporting rationale.*** **Go to 3**

2a – The taxon has characteristics that make them significantly vulnerable to the harmful human activity.

2b – The taxon is not significantly vulnerable to the harmful human activity.

2c – The taxon is vulnerable to harmful human activity over its total range.

2d – The taxon is not vulnerable to harmful human activity over its total range **and/or** there are areas where the taxon occurs but is not at significant risk.

Once it has been decided that the taxon is subject to a significant risk and impact from harm or not, then a decision needs to be taken on whether the release of specific data on that taxon – or other related data – will increase the risk and impact of harm.

3. SENSITIVITY OF DATA

Assess whether the release of data will increase harm.

3.1. Is the content and detail of the biodiversity data such that their release would enable someone to carry out a harmful activity upon the taxon or attribute?	**Yes:**	***Document using statement 3a with supporting rationale.*** **Go to 3.2**
	No:	**[Data are not sensitive]** ***Document using statement 3b with supporting rationale*** **Go to 4**
3.2. Is information already in the public domain, or already known to those individuals or groups likely to undertake the harmful activity?	**Yes:**	***Document using statement 3d with supporting rationale.*** **Go to 3.3**
	No:	***Document using statement 3c with supporting rationale.*** **Go to 3.3**
3.3. Would disclosure damage a partnership or relationship (especially where the maintenance of which is essential to helping achieve a specific conservation objective)?	**Yes:**	***Document using statement 3e with supporting rationale.*** **Go to 3.4**
	No:	***Document using statement 3f with supporting rationale.*** **Go to 3.4**
3.4. Would disclosure allow the locations of sensitive features to be derived through combination with other publicly available information sources?	**Yes:**	***Document using statement 3g with supporting rationale.*** **Go to 4**
	No:	***Document using statement 3h with supporting rationale.*** **Go to 4**

3a – The content and detail of the data is such that their release would enable someone to carry out a harmful activity upon the taxon or attribute.

3b – The content and detail of the data if released would not enable someone to carry out a harmful activity upon the taxon or attribute.

3c – The information is not in the public domain, and is not already known to individuals or groups likely to undertake harmful activities.

3d – The information is already in the public domain, or is already known to the individuals or groups likely to undertake harmful activities.

3e – Disclosure of the data is likely to damage a partnership or relationship the maintenance of which is essential to helping achieve a specific conservation objective.

3f – Disclosure of the data will not damage any partnership or relationship essential to conservation.

3g – Disclosure would allow the locations of sensitive features to be derived through combination with other publicly available information sources

3h – Disclosure will not allow the locations of sensitive features to be derived through combination with other publicly available information sources

The final step is to make an overall assessment based on the three criteria above and to document the overall decision using the combined information documented in making each of the earlier decisions. Once it has been determined that the data should or should not be released, then it is important that a decision is made on the Category of Sensitivity, and the level of generalisation for the release of the data.

4. DECISION ON RELEASE & CATEGORY OF SENSITIVITY

Make a balanced decision regarding the release of data and determining the category and level of generalisation

4.1. On balance, considering criteria 1 to 3 above and any important wider context, will releasing the information increase the risk of environmental harm or harm to a living person?	**Yes:**	***Document using statement 4a.*** **Go to 4.2**
	No:	***Document using statement 4b.*** **Go to 4.5**
4.2. Is the taxon distinctive and of high biological significance, under high threat from exploitation/ disease or other identifiable threat where even general locality information may threaten the taxon? Or could the release of any part of the record cause irreparable harm to the environment or to an individual?	**Yes:**	***Document using statement 4c, collate all supporting rationale and document the decision to withhold the data.*** **Go to Category 1**
	No:	**Go to 4.3**
4.3. Is the taxon such that the provision of precise locations at finer than 0.1 degrees (~10 km) would subject the taxon to threats such as disturbance and exploitation? Or does the record include highly sensitive information, the release of which could cause extreme harm to an individual or the environment?	**Yes:**	***Document using statement 4d, collate all supporting rationale and document the decision to release the data.*** **Go to Category 2**
	No:	**Go to 4.4**
4.4. Is the taxon such that the provision of precise locations at finer than 0.01 degrees (~1 km) would subject the species to threats such as collection or deliberate damage? Or does the record include sensitive information, the release of which could cause harm to an individual or the environment?	**Yes:**	***Document using statement 4e, collate all supporting rationale and document the decision to release the data.*** **Go to Category 3**
	No:	**Go to 4.5**
4.5. Is the taxon subject to low to medium threat if precise locations (i.e. locations with a precision greater than 0.001 degrees or 100m) become publicly available and where there is some risk of collection or deliberate damage?	**Yes:**	***Document using statement 4f, collate all supporting rationale and document the decision to release the data.*** **Go to Category 4**
	No:	***Document using statement 4g, collate all supporting rationale and document the decision to release the data.*** **Data should be publicly released**

4a – On balance, release of the information will, or is likely to, increase the risk of environmental harm or harm to a living person.

4b – On balance, release of the data will not increase the risk of environmental harm or harm to a living person.

4c – The species is a distinctive species of high biological significance, is under high threat from exploitation/ disease or other identifiable threat and even general locality information may threaten the taxon, or the release of the information could cause irreparable harm to the environment, an individual, or some other feature. [**Category 1**]

4d – The species is classed as highly sensitive, and the provision of precise locations would subject the species to threats such as disturbance and exploitation, and/or the record includes highly sensitive information, the release of which could cause extreme harm to the environment or an individual. [**Category 2**]

4e – The species is classed as of medium to high sensitivity, and the provision of precise locations could subject the species to threats such as collection or deliberate damage, and/or the record includes sensitive information, the release of which could cause harm to the environment or to an individual. [**Category 3**]

4f – The species is classed as of low to medium sensitivity, and the provision of precise locations could subject the species to threats such as disturbance and exploitation. Detailed data may be made available to individuals under license. [**Category 4**]

4g – The species is classed as of low sensitivity, and the distribution of precise locations is unlikely to subject the species to significant threat, **and/or** the record includes information of low sensitivity, the release of which is unlikely to cause harm to the environment or to any individual. The data should be released to the public 'as-held'
[**Not Environmentally Sensitive**]

In the on-line survey, a number of respondents identified data awaiting publication, data subject to ongoing research, and incomplete or unchecked data as data that they would class as sensitive, and thus subject to restrictions on release. These are data whose sensitivity has a short time frame and it is important that a time for release or review be clearly documented. They would most likely fall under criterion 3.3 above and would be documented accordingly with the supporting rationale being "*awaiting publication*", etc.

NB. *All data regarded as being sensitive, should include a date for review of their sensitivity status, along with documented reasons for the sensitivity status. The date for review may be short or long depending on the nature of the sensitivity.*

The Categories of Sensitivity (below) are largely based on those from the New South Wales Department of Environment and Conservation (2007).

1. Categories of Sensitivity

Criterion	Reasoning
Category 1 –Species or records for which no records will be provided at all, or which are only released as present within a large region such as a county, watershed, etc.	The reason for non-disclosure is that: 1. a ***distinctive*** species of ***high biological significance*** is under ***high threat*** from exploitation/ disease or other identifiable threat where even general locality information may threaten the taxon. 2. the information in the record is of such a nature that its release could cause irreparable harm to the environment, to an individual or to some other feature. Data may only be supplied under strict License conditions or as presence in a large region such as a watershed, county, or biogeographic region.
Category 2 – Species or records for which	The reasons for restriction are that: 1. The species is classed as ***highly sensitive***, and the provision

Criterion	Reasoning
coordinates will be publicly available 'denatured' (to 0.1 degrees) and/or other information in the record is generalized. Finer scale data (Category 3 or 4 or detailed data) may be supplied to individuals under License.	of precise locations ***would*** subject the species to threats such as disturbance and exploitation. 2. The record includes ***highly*** sensitive information, the release of which could cause ***extreme*** harm to an individual or to the environment. Data are supplied to the public 1. with the georeference denatured to 0.1 degrees (~10 km) and/or 2. with sensitive fields generalized or removed and replaced with suitable replacement wording. Data may be supplied at finer scales on request under the conditions of a written data agreement, usually a Data Licence Agreement. When data are provided to clients, they will be advised which species or fields are sensitive and may have their coordinates denatured to that available under Categories 3 or 4. **NB.** In the case where the sensitivity is triggered by fields other than the georeference, it may be more appropriate to class the record as Category 3 or 4.
Category 3 – Species or records for which coordinates will be publicly available 'denatured' (to 0.01 degrees) and/or other information in the record is generalized. Finer scale data (Category 3 or 4 or detailed data) may be supplied to individuals under License.	The reasons for restriction are that: 1. The species is classed as of ***medium to high sensitivity***, and the provision of precise locations ***could*** subject the species to threats such as disturbance and exploitation. 2. The record includes ***sensitive*** information, the release of which could cause harm to an individual or to the environment. Data are supplied to the public 1. with the georeference denatured to 0.01 degrees (~ 1 km) and/or 2. with sensitive fields generalized or removed and replaced with suitable replacement wording. Data may be supplied at finer scales on request under the conditions of a written data agreement, usually a Data Licence Agreement. When data are provided to clients, they will be advised which species or fields are sensitive and may have their coordinates denatured to that available under Category 4. **NB.** In the case where the sensitivity is triggered by fields other than the georeference, it may be more appropriate to class the record as Category 4.
Category 4 – Species or records for which coordinates will be publicly available 'denatured' (to 0.001 degrees) and/or other information in the record	The reasons for restriction are that: 1. The species is classed as of ***low to medium sensitivity***, and the provision of precise locations could lead to risk of collection or deliberate damage. 2. The record includes ***sensitive*** information, the release of which could cause harm to an individual or to the

Criterion	Reasoning
is generalized. Detailed ‘as-held’ data may be supplied to individuals under License.	environment. Detailed data may be supplied under the conditions of a written data agreement, usually a Data Licence Agreement. When data are provided to clients, they will be advised which species or fields are sensitive.

Listing Sensitive Taxa

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 currently 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).

For these reasons it has been recommended that a trigger list of potential environmentally sensitive taxa should be created and linked through GBIF's Electronic Catalogue (ECat)[3]. This would have the advantages of alerting data providers in other jurisdictions that a species is potentially sensitive, and via ECat would provide links to synonyms. It is important to note that the list should be regarded as a trigger to flag the need for a decision on the actual sensitivity of sharing information using the criteria in the previous chapter, and not for generating blanket restrictions. Not all endangered species are threatened through knowledge of their locations and so should not be regarded as sensitive *per se* and thus the list of potential environmentally sensitive taxa should be much smaller than any existing list or rare and threatened species.

The list should be created using Criteria 1 and 2 (refer to the previous Chapter and scenarios in Annex 1), and should include additional information such as:

- Name of Taxon
- Criteria and supporting rationale for inclusion
- Name of person or organisation responsible for the taxon being included
- Geographic coverage of sensitivity (especially if only sensitive over part of its range or within one jurisdiction)
- Recommended Sensitivity Category
- Date for Review

Jurisdictions may also wish to maintain a similar list for their own purposes, and it is recommended that if they do so, they include the above information in all cases. The advantages of making the information more broadly available is that it will alert other data custodians that your jurisdiction regards the taxon as potentially sensitive, and alert users that they should take the sensitivity into account when publishing the results of their analyses, etc.

> **NB.** *Any list of potential environmentally sensitive taxa should be regarded as a trigger only, and any restrictions on availability of actual data should be made on a case by case basis taking into account the listed criteria.*

[3] GBIF Electronic Catalogue http://www.gbif.org/prog/ecat

Generalising Textual Information

In some cases, information in text fields might be regarded as sensitive under certain circumstances. This may include such information as:

- Names of living persons
- Locality information
- The date of collection
- The collector's number
- Habitat
- Landholder information
- Taxonomic names

Some of these may need to be restricted to stop co-relational analyses leading to deductions on the localities of records that are restricted or generalized – for example the collector's name, date, and collector's numbers in sequence. In other cases, it may be necessary to hide the name of a taxon in a list of collections in a biodiversity hot-spot or sensitive locality.

Such restrictions should not restrict the provision of the record as a whole. The data that needs to be hidden may be removed and replaced with suitable wording (see below), or generalized – for example, just giving the name of a higher level taxonomic rank where the species is to be restricted.

> **NB.** *Whenever data in a textual field are restricted or generalized for distribution (such as the name of a collector, textual locality information, etc.) it should be documented by replacing it with appropriate wording – the field should* **not** *be left blank or null.*

Examples of replacement wording include:

- "*name suppressed for reasons of privacy*";
- *"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";*
- *"This specimen represents an endangered or threatened species. The specific locality has been generalized to presence within a grid of 0.1 degree resolution. Detailed data may be supplied to researchers on request*".

> **NB.** *Where there is need to restrict a taxonomic name (for example, of sensitive taxa as part of a survey), it may be possible to replace it with a higher taxon name (genus/family, etc.), or to just report that there are 'x' sensitive taxa present without providing names.*

Occasionally, data providers may be tempted to restrict information in records related to a sensitive record (in addition to the sensitive record itself), such as the collector's name and numbers in a sequence of records collected at the same location and time as a sensitive record in order to reduce the possibility of the sensitive record being found through co-relational analysis. However, if the collector's name and number is removed from just the sensitive record and not the others, it is unlikely that these could be deduced unless the seeker of the information already has inside knowledge. For this reason, and others (see box below), it is recommended that the data on related records not be restricted.

NB. *There are extremely strong reasons not to restrict data on related collections (collector's numbers in sequence, collector's name, habitat, etc.) because of the restrictions this places on data quality/ data validation procedures and the limits it places on the effectiveness of filtered Push Technologies. Information in records related to a sensitive record (but not in the sensitive record itself) should not be restricted unless absolutely necessary.*

Generalising Spatial Information

One of the most common requirements for generalising biodiversity information is to generalise the spatial locality or georeference. Traditionally this has been done in many ways, and there has been little consistency in methodologies, and very little documentation as to what has been done in each case. This has considerably reduced the value of the data for analysis, and often users are unaware that the data has even been modified.

Good practice dictates that whatever you do to generalise the data that you document it so that users of the data know what reliance they can place in them.

Following considerable discussion among data providers and data users, it has been decided to recommend that data providers who are generalising their data do so using a standard methodology (see below), and to document this accordingly. As most biodiversity data are currently made available using decimal degrees, the recommended method means that protocols (such as Darwin Core) do not need modification, other than to allow for suitable metadata documentation.

The method recommended below allows for several levels of generalisation that conform to Categories 1-4 described in the earlier Chapter on *Determining Sensitivity*.

The recommended method for generalisation is:

Category	Sensitivity	Georeference
Category 1	Extreme	Georeference not released or data may be released by watershed/ bioregion/ county, etc. with no georeference coordinates.
Category 2	High	Georeference rounded to 0.1 degree
Category 3	Medium	Georeference rounded to 0.01 degree
Category 4	Low	Georeference rounded to 0.001 degree
Not sensitive	Not sensitive	Georeference unrestricted.

Documentation

It is important to document the method and level of generalisation so that users are aware of what has been done to the data, and what reliability they may be able to place in the data. Currently, neither Darwin Core nor the ABCD protocols provide fields for the recommended metadata. It has been recommended, however, that these protocols be modified to accept such metadata (see Chapter on *Documentation and Metadata*), but in the meantime, it is recommended that the information be recorded in Comments fields.

As far as the generalisation of georeferencing data is concerned it is important to record that the data has been generalized using a 'decimal geographic grid', and record both:

- Precision of the data provided (e.g. 0.1 degree; 0.001 degree, etc.)
- Precision of the data stored or held (e.g. 0.0001 degree, 0.1 minute, 1 second, etc.)

The recommendations for metadata for inclusion in the Geospatial Element Definitions Extension to Darwin Core (TDWG 2005) are set out in the next Chapter on *Documentation and Metadata*. Once they (or similar) have been adopted, then it is recommended that the appropriate fields be recorded and distributed with the data.

> **NB.** If generalizing to a large region such as a watershed, biogeographic region or a county, etc., then do not supply a georeference.

Documentation and Metadata

It is important that data be accurately documented so that users and others know exactly what the data represent, and the reliance that can be placed in them. For example, a user needs the information to determine if the data are suitable for the analysis they are about to run. Many data providers reported in the survey that one reason that they were reluctant to release some of their data was a fear that the data would be mis-used. If the data aren't adequately documented, then the likelihood of inadvertent mis-use is greatly increased as the user may use the data in an analysis mistakenly thinking they are getting accurate point records, when in reality, the data had been generalized to a 10 km grid square, and could be anywhere in a 100 square kilometre area. If running a climate modelling algorithm, for example, then this sort of error could result in a quite misleading result. For this reason alone, it is important to data providers, data users, and end users (such as environmental managers, policy makers, etc.) that the data are accurately described.

In particular, there should be a clear documentation of the 'Access Constraints' which could include, for example, an indication of which parts of the data are sensitive (if any), reasons for sensitivity and conditions under which release is possible.

Documenting Sensitivity

> "*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*". (Llinás, 2005).

Metadata has generally been used to refer to documentation of a whole dataset. Documentation at the record level has usually been referred to just in comments. I prefer, however, to term this '*record-level metadata'* (see glossary) and to formalise the process. In the previous chapter a recommendation was made that where data were generalized for distribution, to document the level of generalisation - for example, that the data had been generalized using a decimal geographic grid, and to record both the precision of the data provided and the precision of the data stored or held. Also, in the chapter on *Determining Sensitivity*, a series of documentation processes were recommended. Some of these may be more appropriate for documenting the reasons for regarding a taxon as a potential environmentally sensitivity taxon (Criteria 1 and 2), while the others (Criteria 3 and 4) are appropriate to the data themselves and belong as part of the broader record-level metadata. To fully document the reasons for restricting data, however, it may be necessary to inherit the documentation from Criteria 1 and 2 to the record level – for example, the reason that data are restricted may include that the taxon is subject to harmful human activity.

At the moment, neither the Darwin Core nor ABCD standards have fields for recording the type of record-level metadata that is recommended here. A number of recommendations have been made to the Taxonomic Databases Working Group (TDWG) for the inclusion of extra fields to the Geospatial Element Definitions Extension to Darwin Core (TDWG 2005) and/or to the Darwin Core itself. The recommendations included those shown in the table on the next page.

The '*DataSensitiveComments'* field here is perhaps equivalent to the '*Access Constraints'* field in most dataset level metadata. The sort of information at the dataset level may include something like:

> "*This dataset is only available to the public at a summary resolution for the following reason. Some of the information held within this dataset relates to species that are vulnerable to human disturbance or prejudice. Two species* (Adelanthus lindenbergianus, Athalamia hyaline) *are significantly vulnerable to collecting. The full detail of this sensitive information may be made available under licence to specific organisations and individuals that need to know to avoid harm to the environment. Please contact the provider for more information.*"

Until such time as these standards and protocols are modified, it is recommended that the data be documented in comment fields, and as far as possible to record the same type of information that would be included in the recommended fields above – i.e.

- That the data are sensitive;
- The primary reasons the data are regarded as sensitive (see Criteria 1-4 the Chapter on *Determining Sensitivity*) along with supporting rationale;
- The date that the sensitivity of the data should be reviewed;

- Precision of the data made available;
- Precision of the original data stored or retained.

Field	Comments
DataSensitiveIndicator	Y/N flag that the observation is sensitive.
DataSensitiveReason	The primary reason why the data are sensitive. Suggested format is either a picklist with values derived from Criteria 1-4 above (or a text field that combines the statements 1a-4g attached to those criteria).
DataSensitiveComments	Further information on the reason(s) or supporting rationale for determining relevance of the Criteria for this record as recommended above. [Free Text]
SensitiveDateForReview	A date field documenting when the sensitive nature of the date should be reviewed. Especially important where the sensitivity is just awaiting publication of results, etc.
PrecisionDataProvided	The scale or the precision of the data made available via the Darwin Core record – may be done as precision, e.g. • 0 = 1 degree • 1 = 0.1 degree • 2 = 0.01 degree • 3 = 0.001 degree • 4 = 0.0001 degree
PrecisionDataStored	The scale or the precision of the data made stored or retained by the data custodian – may be done as precision, e.g. • 0 = 1 degree • 1 = 0.1 degree • 2 = 0.01 degree • 3 = 0.001 degree • 4 = 0.0001 degree • Etc. or may be more free text, such as '1 minute', '0.1 minute', '1 second', etc. depending on how data are stored.

Authentication and Authorisation

As recommended by the experts' workshop, and identified by many in the on-line survey, responsibility for determining who may or may not have access to detailed data on sensitive data, possibly through use of secure log-on, or one-off data license agreements, must be with the data providers.

It was also agreed at the workshop that it is not the role of GBIF to manage the identification, verification or authorisation of users, nor to control authentication or log-on at the Data Portal, but it may have a role in providing guidance and a suitable authentication method to the Nodes.

It was reported at the experts' workshop that the technical issues relating to the authentication of a group or individual, and the use of roles, etc. is not a difficult task. There are several well established protocols and working systems for authentication in use and these could easily be adapted for use by data providers.

The main issue is in determining who the authorized users should be and how to determine who are *bona-fide* users and who are not. This is a difficult issue and one that will need to be explored over time. It is not something that can be recommended in this best practices document; however the earlier report (Chapman 2007b) did make a number of recommendations on how this issue may be further explored.

It has been recommended that GBIF explore the issue of authentication with the view to providing appropriate mechanisms that help data providers. It is therefore recommended that data providers wishing to develop secure authentication for their databases discuss the issue with GBIF, or with their GBIF Node.

The recommendation made to GBIF in the earlier report (Chapman 2007b) was that:

> *GBIF explore the issue of authentication with the view to providing appropriate mechanisms that help data providers identify users who can dig deeper and how. Although GBIF shouldn't have a role (at this stage at least) in vetting users, or in placing controls on the GBIF Portal, it does have a role in providing guidance and assisting Nodes in implementing a suitable and robust authentication method.*

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 8 Apr. 2007].

Chapman, A.D. 2007a. Workshop on Dealing with Sensitive Species Occurrence Data. Held at: NatureServe Offices, Arlington, Virgina, USA. 6-7 March 2007. Report. Copenhagen: GBIF. 30 pp. http://www.gbif.org./

Chapman, A.D. 2007b. *Dealing with Sensitive Primary Species Occurrence Data. Report*. Report to the Global Biodiversity Information Facility 60pp. http://www.gbif.org./. Copenhagen: GBIF.

Chapman, A.D. and Wieczorek, J. (eds). 2006. *Guide to Best Practices for Georeferencing*. BioGeomancer Consortium. Copenhagen: Global Biodiversity Information Facility. 90pp. ISBN: 87-92020-00-3. http://www.gbif.org/prog/digit/Georeferencing **See also Chapter 5 in this *Manual*.**

Department of Environment and Conservation – NSW. 2007. *Threatened Species Information Disclosure Policy* (Version 3 Amended March 2007). http://www.nationalparks.nsw.gov.au/npws.nsf/content/sensitive_species_policy [Accessed 15 Mar 2007].

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.

National Biodiversity Network Trust. 2002. *NBN Data Exchange Principles*. Version 3.2, April 2002. <http://www.nbn.org.uk/downloads/files/DataExchange%20principles%202002.pdf> [Accessed 27 Mar 2007].

National Biodiversity Network Trust. 2004. *The 'Environmental Exception' and access to information on sensitive features*. Version 1.3.2, Countryside Agencies' Open Information Network Environmental Information Regulations Guidance Note No. 1. Linked from www.nbn.org.uk/eir [Accessed 27 Mar 2007].

TDWG. 2005. *Geospatial Extension to Darwin Core*. Taxonomic Databases Working Group. http://wiki.tdwg.org/twiki/bin/view/DarwinCore/GeospatialExtension [Accessed 1 Apr 2007].

Appendix: Scenarios using Criteria 1 and 2 as Triggers

The following sets of scenarios show how the criteria statements given in the Chapter on *Determining Sensitivity* may be used to develop summary statements for documenting the reasons why a taxon may be regarded as sensitive. The summary statement (in the white boxes), should also include supporting rationale, such as specific types of harm, etc. For example in the second scenario (**B**) – the full statement may read something like:

> "***Taxa could be at risk from harm from diseases carried on the wheels of forestry machinery but occurrence is not affected by data availability***."
>
> This may apply to a species of plant in a forestry area susceptible to *Phytophthora* attack, the fungi being transferred on the wheels of forestry vehicles.

Criterion 1:

Scenario A

Criterion statement(s)	Summary statement
1a – There is no significant risk of a harmful human activity.	The taxon is not sensitive.

Scenario B

Criterion statement(s)	Summary statement
1a – The taxon is at risk from a harmful human activity.	The taxon could be at risk from harm but likelihood of harm is **not** affected by data availability.
1d – There is currently no established evidence of actual harm to the taxon.	
1f – Availability of biodiversity data will not increase the likelihood of the harmful human activity taking place.	

Scenario C

Criterion statement(s)	Summary statement
1a – The taxon is at risk from a harmful human activity.	The taxon could be at risk from harm and the likelihood of harm **is** affected by data availability.
1d – There is currently no established evidence of actual harm to the taxon.	
1e– Availability of biodiversity data will increase the likelihood of the harmful human activity taking place.	

Scenario D

Criterion statement(s)	Summary statement
1a – The taxon is at risk from a harmful human activity.	The taxon is at risk from harm and there is evidence to support this, but occurrence is **not** affected by data availability.
1c – There is established evidence of actual or recent harm to the taxon.	
1f – Availability of biodiversity data will not increase the likelihood of the harmful human activity taking place.	

Scenario E

Criterion statement(s)	Summary statement
1a – The taxon is at risk from a harmful human activity.	The taxon is at risk from harm, there is evidence to support this, and occurrence **is** affected by data availability.
1c – There is established evidence of actual or recent harm to the taxon.	
1e– Availability of biodiversity data will increase the likelihood of the harmful human activity taking place.	

Criterion 2:

Scenario F

Criterion statement(s)	Summary statement
2b – The taxon is not significantly vulnerable to the harmful human activity.	The taxon is not significantly vulnerable to the harmful activity, and is not vulnerable to that activity over its total range and there are areas where the taxon is not at significant risk from that activity.
2d – The taxon is not vulnerable to harmful human activity over its total range **and/or** there are areas where the taxon is not at significant risk.	

Scenario G

Criterion statement(s)	Summary statement
2a – The taxon has characteristics that make it significantly vulnerable to the harmful human activity.	The taxon is significantly vulnerable to the harmful activity, but is not vulnerable to that activity over its total range and there are areas where the taxon is not at significant risk from that activity.
2d – The taxon is not vulnerable to harmful human activity over its total range **and/or** there are areas where the taxon is not at significant risk.	

Scenario H

Criterion statement(s)	Summary statement
2a – The taxon has characteristics that make it significantly vulnerable to the harmful human activity.	The taxon is significantly vulnerable to the harmful activity, and is vulnerable to that activity over its total range.
2c – The taxon is vulnerable to harmful human activity over its total range.	

Glossary

Authentication: — refers to the determination of a user's identity, as well as determining what a user is authorized to access. The most common form of authentication is user name and password, although this also provides the lowest level of security.

Authorisation: — refers to the process of determining which individuals can be afforded different access rights for authentication and data access.

Generalisation: — 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 watershed, 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.

Record-level Metadata: — refers to documentation at the level or a record rather than for a complete dataset. In this document it largely refers to documentation of the sensitivity status of the record (or the species of which it is a part) along with access constraints pertaining to the record and details of any generalisation of the data.

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.

Index to Chapter 6