# Aims

There have been a lot of publications over the last 20 years which a focus on what we should be doing in this space and the need for standardisation of practices. The purpose of this workshop was to try to move towards some community-agreed outcomes.

The aims of the workshop were:

* to generate discussion and possible endorsement of the metadata recommendations made by Arthur Chapman and his 2023 paper discussed and try to endorse approaches to managing restricted access data such as generalisation and the proposal for global sensitive species list.
* to produce a report based on these workshop discussions that request TDWG to consider adopting a Restricted Access Data Extension to Darwin Core and if adopted that can be endorsed by TDWG and organisations such as GBIF and the Atlas Living Australia, and tabled with international conservation organisations and agencies as a global approach.

# Introduction / Background

The growth of biodiversity data aggregators and citizen science projects over the past two decades has led to an exponential expansion in the application of big data to conservation assessment and biodiversity research. Subsequently, expectations that biodiversity data should be made available have risen, as has the recognition of the challenges of both releasing and not releasing sensitive species data.

As biodiversity data has been democratised, the motivations of individuals, researchers, and organisations in restricting access to certain types of data have remained constant. Data producers and custodians may have a more conservative view of access than data consumers. There's those in favour, those opposed. But the basic reality is a lot of institutions are restricting access to a certain proportion of their data and that is not likely to change the behaviour in the near future. Factors to be taken into consideration include type and level of threat, vulnerability, type of information, and public availability. Additionally, there is a growing recognition of the need to enable Indigenous peoples and local communities to assert data sovereignty over traditional knowledge and biodiversity data gathered by, about or within areas managed by them.

Two of the most recent attempts to deal with the overlapping issues in this space have been Arthur Chapman's 2020 [Current Best Practices for Generalizing Sensitive Species Occurrence Data](https://docs.gbif.org/sensitive-species-best-practices/master/en/#listing-sensitive-taxa), and in Australia the development of a standardised [national framework for the Sharing of Restricted access species data in Australia](rasd.org.au).

## The Australian Experience (Cam Slatyer, ALA)

Australia is symptomatic of the many issues and restricted access data. Australia has three tiers of government comprising hundreds of local government areas. There are eight state and territory governments as well as the national government. In addition, there are 43 universities, over 20 major collecting institutions, around nine major non-government groups managing land for conservation purposes, 50 Land Management agencies, indigenous Ranger groups and thousands of companies undertaking environmental consulting or Land Management. Most agencies have restricted access policies of some description, and eight of the 9 jurisdictions and at least four NGO's maintain their own lists of sensitive species. Then there are international organisations like GBIF, Bold and GenBank to consider. The Grey Falcon was used as an example to illustrate the issue – this species occurs in every state and territory in Australia except Tasmania. Different sensitivities rules are applied across the country including generalisation, withholding records and withholding breeding records. Some data custodians don’t generalise the data at all and raw records are available in international databases such as GBIF and Genbank.

The term restricted access species data (RASD) was chosen as opposed to the sensitive species data as we found that most organisations conflated sensitive with government sensitivity and were withholding records based on fears of legislative repercussions. The national framework, which was developed earlier this year sets out principles that aim to be consistent with government requirements and FAIR and CARE principles (including publishing of metadata for all data regardless of whether it is generalised or withheld). It provides a consistent classification of RASD.

In addition to FAIR and CARE, the framework aims to consistently classify and identify RASD nationally, make the data discoverable, to apply consistent transformations and to provide as complete a dataset as possible to requestors.

We defined 4 classes of RASD in the framework.

* personal identifiable information. The reason that's included is in Australia personal identifiable information is protected by legislation nationally in most States and territories except for WA. This doesn’t mean that datasets are withheld, rather that the fields including this information are withheld.
* Indigenous data – the framework recognises that there's a category of culturally important sensitive data that will need to be considered more comprehensively but requires extensive consultation with communities. Application of the CARE principles are encouraged.
* Usage- restricted categories. So that generally means categories where the data aggregator holds a data set that belongs to someone else. There's a legal agreement in place for access to that data, and they can't pass it on to a third party.
* Species-related categories – these include data about species where access to information about the exact location causes sensitivity (e.g. orchids at risk of being collected or nest trees that might be present in forestry operations), data that contains species whose identification has major ramifications (e.g. biosecurity species that aren't found in a country where an incursion can actually threaten export markets ), and data which contains additional information that causes additional sensitivity (e.g. location of pest control activities)

During the development of the framework, it became apparent that there was a need (at least nationally) for a single restricted access species list. In Australia there are about seven and a half thousand species that are identified as sensitive by one or more jurisdictions or organisations. Of those, only 162 occur in more than one jurisdiction. The single list will allow all organisations wishing to apply restricted access obfuscations are applying it consistently to the right species in the right jurisdiction. The reason why there's seven and a half thousand is purely because some jurisdictions include everything that is identified as a threatened species as sensitive. Other jurisdictions have very well-developed policies and expert panels for listing sensitive species. Not all threatened species are sensitive and not all sensitive species are listed as threatened. It depends on the reason for the sensitivity of the data.

## GBIF – International Approach (Andrew Rodrigues, GBIF)

After concerns being raised that GBIF may be making potentially sensitive data publicly available, a report was commissioned to undertake an overview of existing sensitive species frameworks at national and organizational levels, the utility of using the global IUCN Red List and CITES appendices as global trigger lists for identifying sensitive species and to assess the degree to which data publishers are applying Chapman’s guidelines for generalising sensitive data. This work was undertaken by Dr Francisca Astorga, University of Mayor, Chile and consultation on the recommendations of this report are due to be released by the end of October for public consultation.

The study identified 9 national or organisational frameworks and associated protocols for identifying and generalising sensitive data and included Australia, New Zealand, France, Sweden, United Kingdom, South Africa, Finland, Switzerland, eBird and iNaturalist. All 9 used different models, often involving national stakeholder processes and consultation processes. A number of these frameworks included user feedback mechanisms to allow users of the system to flag potentially sensitive species data.

We compared these 9 to Chapman's guidelines which include three main elements of identifying potentially sensitive taxa:

1. identifying whether there is a risk to that specific taxon
2. assessing the impact that that that risk will have on the taxon
3. assessing whether releasing the data on that taxon would result in more harm coming to that to that species.

We found that the majority of these systems were applying the risk of harm element, using the IUCN Red List or a national Red List to identify potentially sensitive taxa. And the other two elements – impact of harm and sensitivity of data were used less frequently.

We looked at the IUCN Red List and the CITES appendices to evaluate whether there might be some utility in using these for identifying sensitive species at a global level. This is particularly relevant given that we only found nine frameworks across the entire world and so there seems to be little information on sensitivity within data for most countries. To do this we chose a smaller subset of the IUCN Global Red List specifically focused on those species which are targeted for exploitation where we might assume that these species are more sensitive to the public release of data. We used the biological resource use filter on the IUCN Red List to identify 12,890 potentially sensitive species (41% of those were categorised as Threatened and the rest were Least Concern, Data Deficient or Near Threatened). We then identified nearly 41,000 potentially sensitive taxa listed on the CITES appendices and used these as global trigger lists to interrogate the GBIF data.

We checked to what extent these global trigger lists overlapped with national and organisational lists to see whether the same taxa were appearing on both. The compiled list of species on the national and organisational lists was 9233 taxa, many of which were from Australia. The majority of taxa on this compiled list i.e. 5715 had not been evaluated by IUCN so they wouldn't have been picked up if we were using the IUCN Global Red Lists and with only 220 were identified as potentially sensitive using our IUCN Red List filter. We also found significant biases in the taxonomy, with a much higher representation of plant taxa on national and organisational lists as when compared to potentially sensitive species identified using IUCN Red List filters, likely reflecting taxonomic biases within the global Red List. In addition, non-vertebrate species are generally less well represented on the global Red List.

Around 1200 species from the compiled national and organisational list were also listed on the CITES appendices so this is a better representation of globally sensitive species, however still not very representative. Of the national and organisation lists we found that there were no sensitive species lists in Africa, South America, Asia and the Global Red List would have more information on taxa in these regions.

We analysed to what extent data from sensitive or potentially sensitive species were being generalised in GBIF across geographies and different organisations and the extent to which Darwin core terms as recommended by Chapman were being used. We investigated a subset of 5728 taxa (38 million records) found on either national or organisational lists or the subset of IUCN trigger list taxa. We then identified which species had been generalised and looked at five generalisation fields – data generalisation, information withheld, coordinate uncertainty, coordinate precision and footprint WKT to indicate that some generalisation had been applied to the locality data.

We found that only 360 taxa had either their information withheld or their locality data generalised, and these species had generally a very small number of records. 40-58% of records were not generalised at all for even when taking coordinate uncertainty into account. Data for species on national lists that was being published outside of the country tended to be better generalised than data at a national level.

We then took a subset of the data (270,000 records) to analyse how the information withheld and data generalisation terms (the 2 fields that are recommended within Chapman for providing information on sensitivity) were being used. Around 187,000 of these were using information withheld however only 3% referred to any sensitivity within the data. Around 107,000 occurrences were using the data generalisation field and 75% were referring to sensitivity within the data.

We then analysed two specific examples - the rhinoceroses and orchids. For rhinoceros we found that for the five extant species of rhinoceros,89% of the occurrences were being generalised using the information withheld and data generalisation terms. However, there were still around 10% of records that were being provided with generalisations under 250 metres, which isn’t generalised sufficiently to protect that taxon.

We analysed 16,647 occurrences of threatened orchids and found that 22% of these records were using the information withheld term and another 52% were using the data generalisation term. The free text often referred to national frameworks and the reason why the data was generalised. 32% of occurrences were being provided with no generalisation whatsoever.

The key takeaway points from this review are:

* there are a number of national approaches out there with some maturity now in developing national frameworks and that we should be supporting these and building on these efforts.
* that global lists can provide some signal however; local national lists are providing the best information as it’s through these that you get a better understanding of the local context of what species or what sort of threatening activities are going on at a national level.
* The Darwin core terms are not being used consistently across different taxa and across different publishers.

## Operationalising RASD Management – an Australian Example (Piers Higgs, Gaia Resources)

A solution to part of the problem of handling sensitive species data is the [RASD Service](https://service.rasd.org.au/#/) that is now up and running via the Atlas of Living Australia. When we talk about the FAIR principles, this is addressing the Findable part. We’re starting to look at solving at least one part of the complex problem that's out there through this approach.

The RASD Service, as outlined on the web site, aims:

“..to provide a single place for users to discover and request Australian restricted access species data. The service enables data custodians and data requestors to track the progress of and manage their requests. Data custodians can also track the use of their data via a digital object identifier (DOI) through the service.”

The RASD Service essentially provides a single place to discover and request RASD. It is the first step in dealing with accessibility and enabling people to have access to data. It is targeted to data requestors as people who are seeking data and data custodians as those who hold that data to make it findable and accessible, noting that the data service doesn't hold any data. It's a request facilitation service. It does not actually pass data around, it's simply helping manage the requests and make that data findable.

It enables users to filter by key words / topics, search for locations, descriptions to see metadata for RASD datasets which have been nominated by the data custodians. The user selects a dataset(s) and then is taken to a form to request the data. This form gathers information about with the requestor wants the data and what they want to use it for and that gives the data custodians the opportunity to really evaluate the request. Another feature of the form is that if the request is not approved as a full resolution delivery, which asks if they requestor would you be happy to receive denatured data.

The main way that it works is through two sets of tools. If you're a data custodian, you're either managing the data sets that you enter metadata for, or you're managing requests that come through for your datasets. If you're a data requester, you can see the status of the requests that you’ve made and search for datasets. It is a fairly simple service but has quite complex tools to manage the request process. At any stage data custodians can see who has been requesting data sets from them, what's the status of that request, the purpose of it. It's very transparent.

The purpose of the RASD Service is about firstly making the RASD findable, and then facilitating making it accessible to a broader audience.

# Discussion / suggestions

Generalisation vs Randomisation

Generalisation – reducing the number of decimal places on latitude and longitude to reduce the precision of the record (lossless way of desensitising a records – where you can say it falls within a grid cell)

Randomisation – randomly assigning a point to a record (a lossy way of desensitising a record where you cannot get back to the precise location if you need to)

Cave Fauna

Mentioned as a special case – if you know stygofauna taxa you can easily locate caves e.g. on the Nullarbor plain. There is a need to approach various different groups e.g. cave groups / speleologists etc to ensure that the Australian National Framework covers off on RASD issues specific to their data.

Aggregated data or individual data requests to custodians?

In Australia we found that data custodians wanted to maintain control over their data so instead of aggregating RASD we provided a centralised request tracking system instead so that a data requestor could place one request that covered one or more datasets. Custodians and requestors can track progress of their requests

Linked data

Need to be cognisant when sharing RASD where it has been collected at the same time as records for other non-sensitive species by the same collector – in this case you may have to generalise more than just the locality – may have to generalise the date / time etc as well.

Regionally applied lists best approach

Sensitivity is a taxon plus locality issue e.g. a highly sensitive Australian species may be a weed in the USA. So it would not make sense from a management point of view to restrict access to occurrences in the US for this taxon. So using a global list would mean you would be restricting access to useful data in locations where the taxa are not sensitive.

Restricted / sensitive species list

Need to stress the importance of these checklists for conservation – e.g. to recommend DWC extension to TDWG, IUCN, groups involved with Convention on Biological Diversity etc.

Need something that is driven through local knowledge and local processes

Look at CITES listed species

Need for this list – why not just have each country have their own – e.g. Genbank – international records / databases don’t have a list to obfuscate Australian records – BUT most countries don’t have one – GBIF only found 9 countries which had one which were publicly available. So no info on which spp are sensitive in which country

Indigenous data

Further consultation required to further define indigenous reasons for sensitivity

In the interim, the dataSensitiveReason would be a good place to flag that a record is sensitive from an indigenous perspective

May not be latitude and longitude that needs obfuscating it may be other attributes e.g. medicinal use

Embargo date

Guidance on what period of time you’d be setting for different types of reasons for embargo

Was included in proposed extension due primarily to withhold records prior to publication – delisting might be another reason e.g. extinct taxa. (see RASD Framework) or commercial in confidence reasons for fisheries locations, or records subject to court cases

automatic release post embargo end date

Consideration of obfuscated pre-embargo release product prior to embargo end date instead of embargoing entire dataset / and all other records in the dataset that isn’t sensitive

Release of data prior to licensing permits are released tied in with commercial sensitivities

Clarity around the publication which the data is going to be used in so that others can make a call if the embargo period has ended once the paper is published.

Internal use of standard or for aggregators only?

Eg records subject to legal cases that can’t be released to aggregators at all until after the court case is resolved.

# Recommendations

DWC working group to work towards ratifying sensitive species extension

* Take Arthur’s extension and make amendments to it based on this workshop discussion
* Defined set of fields – pick lists for dataSensitiveReason (this is defined in Arthur’s report)
* Consider the creation and adoption of national checklists

Global list with a series of criteria – work towards this – need to establish a global interest group to do this perhaps in partnership with relevant international organisations such as IUCN and CITES amongst others

# Next Steps

Speak with TDWG Executive about establishing a TDWG working group