NBN Strategy for data sharing with GBIF

October 2015

By Rachel Stroud
NBN Data Liaison Officer
E: rachelstroud.nbn@gmail.com

The following paper outlines the proposal to resume data sharing between the NBN and GBIF.

Executive Summary

- Previous data sharing with GBIF and the communications strategy surrounding this has been inadequate.
- This strategy sets out to reaffirm the NBN commitment to data sharing globally as per the NBN Strategy 2015-2020 and NBN Action Plan.
- This strategy has been circulated to NBN members and Data Partners for wider consultation.

Historic relationship with GBIF

Through GBIF, billions of records are served in downloads yearly - around 234B in 2014, and 174B so far in 2015. The UK currently publishes about 50 million records to GBIF, with roughly 35 million records published via the NBN (about 16% of the total GBIF holdings). GBIF was given a copy of only publicly downloadable data (excluding sensitive species) from NBN Gateway (Version 4) prior to July 2013, before the change to the new version of the NBN Gateway (version 5) in October 2013. At this time, GBIF Terms & Conditions were nearly identical to those of the NBN Gateway. No further data has been supplied to GBIF since this time.

During the NBN Gateway access controls consultation in 2012-13 the NBN Secretariat agreed that all downloads would be logged in return for removing the 'view only' control. GBIF allows download of data however, the detail on data download is not as comprehensive as on the NBN Gateway. Some UK data partners have previously identified this as an issue when sharing data with GBIF.

Recent changes within GBIF

There is now a need for greater clarity both for data publishers and users on how data may be used when shared via GBIF. At the 21st meeting of the GBIF Governing Board, a new set of principles and next steps relating to the licensing and endorsement of data published through GBIF were agreed based on extensive consultations on both issues during 2014.

GBIF have started work to ensure that all species occurrence datasets within the network are associated with digital licenses equivalent to one of the following three choices supplied by Creative Commons:

- **CC0,** under which data are made available for any use without restriction or particular requirements on the part of users
- **CC-BY,** under which data are made available for any use provided that attribution is appropriately given for the sources of data used
- **CC-BY-NC,** under which data are made available for any use provided that attribution is appropriately given and provided the use is not for commercial purposes

Where does the NBN want to be?

The NBN Strategic Action Plan Objective '2E: Share all biological data internationally and collaborate with partners in Europe and GIBF' sets out the NBN's commitment to sharing data with GBIF. This includes working with NBN members to maximise visibility of UK biological data on international platforms.

Work began to align the NBN data sharing principles more readily with GBIF at a NBN Gateway Terms and Conditions workshop in November 2014. The Network is supportive of assigning data licenses to NBN Gateway datasets and this functionality has now been developed within the NBN Gateway and is currently in beta testing phase. Since July 2015, NBN data partners have being consulted with to ask if they will allow their data to be shared with GBIF and what their preferred license would be from the three GBIF license options (CC0, CC-BY and CC-BY-NC).

To date the majority of NBN data partners are supportive of sharing their data with GBIF as long as they have permission from their recorders and other third parties involved. The favoured license is CC-BY-NC (n=34, 72%) with support for CC-BY (n=8, 17%) and CC0 (n=5, 11%).

Recent changes within GBIF on reporting and Digital Object Identifiers (DOIs) are believed to alleviate outstanding concerns within the Network surrounding tracking data download, and removal of data from GBIF. See Appendix 1 for more information on tracking downloads and Digital Object Identifiers (DOIs).

Options

There are two issues that need to be resolved before data sharing with GBIF can recommence:

- Are the existing facilities on GBIF close enough to the requirements of the UK data providers?
- How will data flow from the NBN to the GBIF portal?

Are the existing facilities on GBIF close enough to the requirements of the UK data providers?

The key differences are around license models and tracking downloads. The former is to a large extent resolved – the existing NBN Gateway now allows a user to select the license that they wish to associate with data downloads – these match those implemented by GBIF. The Gateway still documents more detail on who is downloading the data; GBIF records this but does not currently expose it to data providers.

Option 1A: Remove all data from GBIF and do not reload
The Secretariat permanently cease to offer NBN Data Partners the service of uploading their data to GBIF and instead ask NBN data partners to provide these data through their own routes. This would be the most cost effective mechanism but may result in little UK data being shared with GBIF due to the cost burden falling on individual Data Partners.

Option 1B: Remove all data from GBIF. Upload data, with licenses following consultation with NBN data partners
The Secretariat provides a means for NBN data partners to identify which datasets they wish to be shared with GBIF and under which license. This may require some further development to provide this mechanism and Data Partners will need to be educated about the implications of uploading data to GBIF.

Option 1C: Remove all data from GBIF and only reload government data on GIBF
The Secretariat removes all data from GBIF and only reloads only government data. NBN Data Partners will be required to upload their own datasets to GBIF.

Option 1D: *GBIF propose a license to which they would default the data set if there is no response from the data publisher that it should be handled differently.*
GBIF provide an inventory to the NBN outlining which datasets would be assigned to each of the three categories for circulation to the relevant data publishers. Data can be shifted to another license or removed if necessary. Thereafter, new NBN datasets would be given a licence or not as they are added to the NBN. **NB*:*** This is GBIF proposed position for all their data partners

Option 1E: *Switch all NBN sourced data to CC-BY-NC.*
This license condition is effectively the same as the NBN Terms and Conditions. The data are quite out of date but there would be no visible drop in GBIF holdings. A strategy to gradually switch off the datasets as they are refreshed through whatever route is agreed would need to be developed.

The preferred option moving forward is Option 1B: *Remove all data from GBIF. Upload data, with licenses following consultation with NBN data partners.* The data are already quite out of date and there is a need to establish a new mechanism for data upload. While this is being planned, the current data would be removed and the NBN would have a fresh start. The main issue would be the drop in the total GBIF

data holdings (i.e. a 16% drop) that would be bad for the UK's profile. GBIF's proposed position (Option 4) is not feasible given the NBN do not own the data on GBIF - whatever happens we would take the data off during the GBIF change period.

What next - How will data flow from the NBN to the GBIF portal?
Currently the NBN Secretariat is promoting the establishment of an alternative UK biodiversity data infrastructure based on the Atlas of Living Australia software stack. This effectively abandons the access controls and instead promotes full public access to the data. The implementation of this is currently being piloted in Scotland. If successful the aim would be to deprecate the existing Gateway.

The data flow to GBIF could either be through the current NBN Gateway (as has been the case to date) or through the implementation of the new Atlas. The latter has the advantage that the mechanism already exists though it is unclear what sort of volume of data will be available. Additionally, if the current Gateway is to be the publication route, some additional development would be required to update the publication route.

There are two options here:

Option 2A – Update publication through existing Gateway. Additional functionality has been added to the Gateway to allow a dataset administrator to select the license type although and the Terms and conditions updated to implement the change. The mechanism to publish to GBIF could be updated to allow refresh of data on GBIF. The cost of this is approximately XX which is unbudgeted.

Option 2B – Begin to publish to GBIF via the Atlas of Living Scotland. The Atlas of Living Scotland (currently under development) could be used to publish data to GBIF. This would require the licensing to fit with one of the Creative Commons licenses used by GBIF. An additional decision would be needed on whether data from beyond Scotland could be held in the underlying database even if not displayed on the Atlas itself.

The preferred option is Option 2B for the Atlas of Living Scotland infrastructure be used as the mechanism for data transfer to GBIF.

Data Partners will be able to use the Atlas platform to select which datasets they wish to transfer, and under which license. These data can then automatically be pushed to GBIF. There lies a risk that fewer datasets be transferred to GBIF through an 'opt in' approach, however there lies a greater risk of data partners distrusting the NBN if an 'opt out' clause automatically transfers their data to GBIF outside of their immediate control.

Until an Atlas of Living UK is developed it is currently unclear if the Atlas of Living Scotland could be used for data outside of Scotland but in a year's time this would be the route.

It is proposed that all data for the UK are removed from GBIF and are reloaded with appropriate data licenses chosen by the data provider. This may have to be a manual process until a more dynamic one through the atlas is developed.

There is a serious risk that this creates a gap in GBIF data holdings. The NBN will work closely with GBIF throughout this time to minimise this gap and will develop a GBIF data transfer plan. The Secretariat of the NBN will consult with data partners (already underway), identify which datasets are to be uploaded to GBIF before any data are removed/transferred back to GBIF.

Data Transfer Frequency

This will be determined in consultation with data partners. If an automated process were achievable Data Partners could then set the frequency. If the process is a manual process, the frequency needs to be discussed further.

- Finish consulting with data partners over their attitude to sharing data with GBIF
- Clarify if the Atlas of Living Scotland can hide data from outside Scotland
- Circulate proposed approach in October eNews
- Discuss with data partners which datasets they wish to be visible on GBIF and under which license
- Discuss with GBIF to clarify the process of removing data and reloading data
- Discuss communication plan with GBIF to prevent data seemingly to 'disappear' from GBIF
- If the Atlas of Living Scotland can publish data to GBIF from outside Scotland transfer data from NBN Gateway with data partner permission for upload to GBIF
- If the Atlas of Scotland cannot publish data to GBIF from outside Scotland, transfer data manually to GBIF for those data partners who wish their data to be visible

Appendix 1

Tracking dataset downloads

Data accessed through GBIF.org is tracked by dataset and each dataset page contains an 'Activity' tab where all downloads are visible. For example, for the "UK Cetacean Strandings Investigation Programme" dataset, all download access can be viewed here: http://www.gbif.org/dataset/5a87c36c-7785-41ac-bc6f-353516c8a33f/activity

This is also available through the API should data partners wish to build their own custom reporting aggregations:
http://api.gbif.org/v1/occurrence/download/dataset/5a87c36c-7785-41ac-bc6f-353516c8a33f

Digital Object Identifiers (DOIs)

Recently GBIF have started issuing Digital Object Identifiers (DOIs) to downloads themselves. For example, http://doi.org/10.15468/dl.y14xzk is a download that served some data from the example dataset above. GBIF's citation guidelines now instruct people to use those download DOIs in any published work. GBIF's intention here is to ultimately connect published work to the originators of the datasets i.e. provide data publishers an easy way to find research making use of their data. GBIF have set the foundations for this through the DOI implementations and will work now work with publishers to try and identify those linkages as people start to cite using DOIs.

If a dataset is removed from GBIF the DOI still resolves (http://doi.org/10.15468/fv2e59) but the user is simply presented with a page showing the metadata of the dataset at the time of deletion. See the following example of a deleted dataset: http://www.gbif.org/dataset/091913aa-dad3-47e7-8dab-a930151f2676.