E-seminar: Data Exchange for Biodiversity Conservation in Freshwater Ecosystems: Introducing the BioFresh Platform and Data Portal

You mentioned that there is a lot of data out there. Could you give some examples, where this data exists and why it was compiled?

As far as I am aware, no one has a good overview of what data is out there. The BioFresh partners realised that there must be large datasets from environmental agencies, collected in the framework of monitoring programs like the European Water Framework Directive, the Habitats Directive, or for environmental impact assessments. Scientists know that the underlying data for scientific papers is often hard to access as it is stored on hard drives, pen drives, or CDs that eventually fail, get lost or are forgotten. Research institutes and organisation might maintain databases, of which the existence is only known to those involved or close to the organisation. All this data is a great potential for the BioFresh database.

How could data publishing from science be improved? Which framework conditions could be changed to engage scientists in data publishing?

Data from public institutes is often available upon request, but unfortunately an overview of existing datasets within the institute is not always available, and the process of getting the data you are interested in is likely to be rather tedious. In such a case, installing a data management policy which includes publishing data in public repository would be appropriate.

Researchers often refer to the need for publishing the data before making the data available. This is a genuine concern, but often the publication of data is unnecessarily delayed and scientists don't regard making it available as priority as they get little credit for it. Releasing the data through the publication of a data paper could partly diminish this concern.

On the other hand, clear guidelines from funding agencies and scientific journals could facilitate the taking up of data publishing as a standard scientific practice. Funding agencies could set clear recommendations in terms of embargo periods for publishing datasets, while scientific journals can play a role in formulating the request that data needs to be made available or archived at the time of publication.

The animation mentioned that the project also addresses policy makers. Could you explain how you see the contribution of scientists to policy making?

I do not really expect that water managers, let alone policy makers, would necessarily start from the raw data, which is the focus of the data portal. By making this data publicly available, we provide access to scientists (both within and outside the BioFresh network) who can use this data in their analyses and modelling work. The results of such analyses can be disseminated to a wider public through other components of the BioFresh platform such as the Global Freshwater Biodiversity Atlas and the blog.

Additionally, I suspect that knowledge on the distribution of species – as available through the data portal – will also be an important resource for identifying the needs for biodiversity conservation policy. For this reason, we have also teamed up with the IUCN freshwater group in BioFresh, to include the expert validated distribution ranges in the data portal for the species where this information is available. With respect to the occurrence data, I believe that a huge increase in the data that is made publicly available would be required to provide a better overview of the species' distribution. So, we are very much welcoming any contributions.

The e-seminar was a cooperation between WaterDiss (Grant agreement number 265167 under the EC 7th framework contract) and BioFresh FP7.

E-seminar: Data Exchange for Biodiversity Conservation in Freshwater Ecosystems: Introducing the BioFresh Platform and Data Portal

You addressed that BioFresh also prepared own data and has a lot more in the pipeline. Which data can we still expect to be published and which efforts have you taken in order to digitalise it?

Within BioFresh we initiated a large number of small data digitisation and acquisition projects. One of the major themes was completing the picture of the distribution of European fishes, which accounts for a quarter of a million occurrence records. Another major project focused on mobilising European caddisfly (Trichoptera) data, which will yield over half a million records. We also spent effort in obtaining data from the intercalibration exercise for the Water Framework Directive. While the rate of success in obtaining these data was quite variable, a significant amount of monitoring data was obtained in parallel to these efforts.

From these major datasets, the German monitoring data are ready to go, the fish data are undergoing a taxonomic quality control and the caddisfly data are under embargo until the end of the project (April 2014). But at this stage, we are setting up a system to integrate the data in the occurrence database independent of where the data is hosted. This should allow us to speed up the data processing towards the end of the project.

Although it is hard to give a number to estimate the data volume, with these datasets alone we are most probably reaching over a million occurrence records, which seems a considerable amount to me, given that we had 9 million records available through GBIF (Global Biodiversity Information Facility). I see this only as a start to set up an active and growing network, as we will have the infrastructure in place to publish and display these data.

The system includes the possibility to store data de-centrally and we are using this approach to harvest freshwater datasets that are already available through the GBIF network. However, despite the fact that we have this capability in place, most of the datasets that are currently in the pipeline will be hosted on a central BioFresh server. We are offering this possibility for data holders that are not eager to install a data publishing system on their servers , which is needed for the GBIF data platform (the Integrated Publishing Toolkit).

Carlos Rodriguez: As you are aware, we are now starting EU BON project and we are involved as testing sites in data mobilization (at least our own data). Do you think it is too challenging to have all data sets in a way that not only metadata but also data sets could be available for everybody?

Technically making the data publicly available is not very challenging. In the BioFresh project we are using GBIF's Integrated Publishing Toolkit to do so. Of course this tool is especially suitable for species collection and observation data and is currently not specifically build for exchanging monitoring or sampling data that also include environmental parameters. For this I understand that EU BON is referring to the DataOne model, with which I am less familiar. But again, I believe the biggest challenge is in convincing data holders to make their data available on-line.

Davide Speranza: I had no chance to play that much with the dataset. Does it come as species distribution? Is it possible to use a web-GIS-like tool?

For the BioFresh occurrence database there is indeed a mapping interface, which allows to visualise

The e-seminar was a cooperation between WaterDiss (Grant agreement number 265167 under the EC 7th framework contract) and BioFresh FP7.

occurrences. In the near future users will be able to combine occurrence data with expert validated distribution range maps available through IUCN. In addition, data can be downloaded for analyses and mapping on your own computer.

Davide Speranza: Well, just to be sure, the dataset relates to the EU situation only, right?

While most of the BioFresh partners with the exception of IUCN and WorldFish are European, the focus of the project has been global from the start. One of the aims to ensure the continuation of the BioFresh initiative and network after the end of the project will be to reach out to partners worldwide to establish a wider network

Carlos Rodriguez: Apart from the GBIF tool mentioned, is it any template/tool that help data bases to be harmonized in terms of field names (I guess that a high proportion of data sets would include a field "species", "place", "date", and so on). We need to translate all our data sets, so such tool would be useful

We don't have a specific tool to do so, but we have prepared Excel templates with a selection of fields from the Darwin Core exchange standard (http://data.freshwaterbiodiversity.eu/submitdata.html#submspreadsheet). In general, we try to assist data holders for mapping their datasets to the standards. Data holders of an established database would typically create an export query with a selection of the relevant fields. I also find that the GBIF Darwin Core Archive assistant (http://tools.gbif.org/dwca-assistant/) is quite useful to guide the choice of fields to be included.

Davide Speranza: For example, I am familiar with the IBATforBUSINESS online tool, which is another IUCN/UNEP-WCMC appraised tool (it integrates IBAs, KBAs, PAs, etc). Is it planned to intergate the Biofresh dataset with the IBAT ones in order to use just one global tool?

There are no immediate plans to do this, but BioFresh has contributed to extending some of the IUCN freshwater datasets and I believe that IBAT also links with GBIF data. As we plan to publish all occurrence data through the GBIF network, I believe most data would be accessible through this system. But, still it may be worthwhile exploring this idea. Thanks.

The e-seminar was a cooperation between WaterDiss (Grant agreement number 265167 under the EC 7th framework contract) and BioFresh FP7.