HARMONIZING POLAR BIODIVERSITY DATA FOR WIDER ACCESS AND INTEGRATION:

A COLLABORATION BETWEEN THE SPANISH POLAR DATA CENTER AND GBIF-SPAIN

O. Bermúdez[1], V. González-Álvaro[2], F. Pando[2], A. Barragán[1], K. Cezón[2], C. Lujano[2], S. Martínez de la Riva[2], C. Villaverde[2], P. Ríos[3]*

[1] National Polar Data Centre (NPDC), Spanish Geological Survey, Ríos Rosas 23, Madrid, 28003, Spain*

[2] GBIF Spain, Coordination Unit, Royal Botanical Garden-CSIC, Plaza Murillo 2, Madrid, 28014, Spain

[3] Oceanographic Center of Gijon, Spanish Institute of Oceanography (IEO), Avda Príncipe de Asturias 70bis, Gijón, 33212, Spain

Email: o.bermudez@igme.es

The Global Biodiversity Information Facility (GBIF) is an intergovernmental organization that promotes and facilitates the mobilization, access, discovery and use of information about organisms over time and across the planet, accomplishing the original purpose of its founding by governments in 2001, i.e. to encourage free and open access to primary biodiversity data over the Internet.

SCAR (Scientific Committee on Antarctic Research) is an Associated Participant of GBIF from 2008. Spain is a full member, and the associated Spanish Polar Data Centre (NPDT) is responsible of the management and upload of data (and metadata) from Spanish projects developed in both the Arctic and Antarctic regions in a, at the same time, national and international database (Antarctic Master Directory, Global Change Master Directory). As a result, the CNDP allows an excellent information flow to boost the polar research.

Here we present the collaborative work between GBIF Spain and NPDT. In the one hand, making polar biodiversity data available through GBIF seemed a logical and effective way to establish proved methods to manage biodiversity polar data. On the other hand, polar biodiversity data published by Spanish Institutions through GBIF can be identified and incorporated into the NPDT.

For those purposes, three datasets were selected in order to establish a methodology that could be later applied to other datasets.

ABOUT GBIF

The Global Biodiversity Information Facility (GBIF) is an international open data infrastructure, funded by governments. It allows anyone, anywhere to access data about all types of life on Earth, shared across national boundaries via the Internet. By encouraging and helping institutions to publish data according to common standards, GBIF enables research not possible before, and informs better decisions to conserve and sustainably use the biological resources of the planet.

http://www.gbif.org/

ABOUT GBIF.ES

GBIF Spain is a network of Centers, Institutions and Projects currently formed by 63 Spanish entities that make available more than 8.7 million of data records online.

Taking part in GBIF, the entities contribute with their data to the management of biodiversity data at different levels, being provided with technical support, visibility and acknowledgment. With that mission, the GBIF.ES Coordination Unit has developed several components.

http://www.gbif.es

ABOUT NPDC

The National Polar Data Centre (NPDC), has, among its commitments, the management of data that come from the Spanish scientific research, not only in the Antarctic field but also in the Artic field.

NPDC, as a member of the SCAR Standing Committe on Antarctic Data Managament (SCDAM), is responsible for the management and loading in the national and international database (Global Change Master Directory, Antarctic Master Directory).

Spain has been a member of the JCADM (SCADM now) since 1999, and has established its NPDC within the Spanish Geological Survey. In order to comply with the Antarctic Treaty Article III.1.c, the NPDC has developed a protocol during the last two years, for improving the management of data and sensor calibrations, and for generation of metadata. This data policy will bring about better efficiency in the management and safekeeping of data, by virtue of better knowledge conservation and availability to the scientific community. The NPDC has designed and developed a metadata management system for use by each principal investigator, to enable them to generate and manage their metadata by means of on-line tools. This is vital for achieving the data policy and allows the scientific community to generate their own metadata during the development of their campaigns.

http://hielo.igme.es/

METADATA

What is EML

Ecological Metadata Language (EML) is a metadata specification developed for the ecology discipline. EML is implemented as a series of XML document types that can be used in a modular and extensible manner to document ecological data. Thus, all EML modules together describe the metadata that should be included with any ecological dataset.

In order to publish biodiversity data through GBIF, metadata must be described accomplishing the EML format –some tools can be used for that purpose, e.g. Darwin Test or IPT (see below). This way, GBIF promotes and participates in the sharing of dataset-level metadata published in commonly used standards.

The metadata standard EML used by GBIF and implemented in IPT allows users to define taxonomic, geospatial and temporal scope, along with rights and citation, contact information and keywords. GBIF has defined its own profile of EML known as the "GBIF EML Profile".

http://code.google.com/p/gbif-metadata/

Format Conversion: DIF > EML

However, the CNDP must comply with another metadata format, which is the Directory Interchange Format (DIF), an ISO 19115 compliant format. DIF is a metadata initiative from the Earth sciences community, intended for the description of scientific data sets. It includes elements focusing on instruments that capture data, temporal and spatial characteristics of the data, and projects with which the dataset is associated.

As well as the EML format, DIF is defined according to an XDS specification and is expressed as an XML file. A conversion was feasible, mapping directly using PHP and XML, and was performed for the following two dataset metadata:

Database name	Origin	No°. of records	No. of species	Taxonomic groups
Antarctic lower plants	Spanish Polar Research Program	20	11	Lichens
Antarctic non-marine aquatic ecosystems	Spanish Polar Research Program	45	-	Various taxonomic groups

Once the EML files were generated, metadata were ready to be uploaded in the GBIF IPT in order to make the databases accessible through their publication in GBIF.

Format Conversion: EML > DIF

As per the objective of placing the available through GBIF Spanish polar dataset that was selected here (see the following table) in the CNDP, the EML metadata file was converted into a DIF file by direct mapping as well.

Database name	Origin	No. of records	No. of species	Taxonomic groups
Antarctic Porifera database from the Spanish benthic expeditions	Spanish Antarctic expeditions	766	74	Antarctic Porifera

We noticed that several fields could not be mapped directly though.

DATA

A dataset from GBIF incorporated to the CNDP

A dataset, containing polar biodiversity data, was selected in order to have it incorporated in the National Polar Data Centre: Antarctic Porifera database from the Spanish benthic expeditions (Oceanographic Center of Gijon, Spanish Institute of Oceanography).

For the adaptation of the polar dataset already published in GBIF into the SCAR databank, only metadata was a concern, since there is no standardized profile for biodiversity data within the Spanish Polar Data Center.

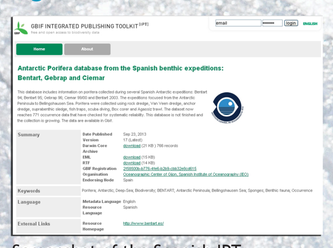
This datsets processed through the GBIF dataflows, will be incorporated into the Spanish Polar Data Center soon.

What is IPT

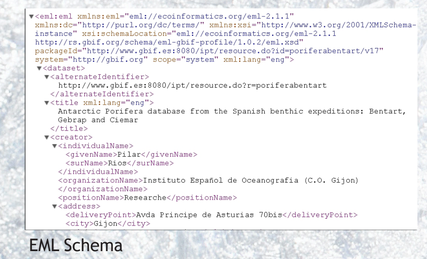
The Integrated Publishing Toolkit (IPT) was developed as a software platform to facilitate the efficient publishing of biodiversity data on the Internet, using the GBIF network. Following the Darwin Core standard, this free and open source web application can be used to manage and publish primary occurrence data, taxonomic checklists, and metadata.

Any kind of database can be imported as Darwin Core Archive, or as a delimited text file, e.g. csv, given the IPT allows field mapping to accomplish the Darwin Core Archive format.

Once the Darwin Core Archive (http://rs.tdwg.org/dwc/terms/guides/text/index.htm) and metadata have been made public, information becomes public, being not only accessible but downloadable, as Darwin Core Archive, EML and RTF (metadata structured as easily readable text) files. Also, the dataset becomes accessible through the GBIF data portal (http://data.gbif.org).



Screenshot of the Spanish IPT



EML Schema

Two datasets from CNDP published through GBIF

IPT service allows the user to upload, standardize, publish and register data easily and quickly.

Once the DIF metadata files were converted into EML files, and the datasets were mapped over the IPT, the two selected datasets and their metadata were published through the GBIF Spain IPT service (http://www.gbif.es:8080/ipt/).

Three files became accessible for each dataset, i.e. the Darwin Core Archive (a zip file that contains the occurrence txt file, the EML, and the META XML files), the EML, and the RTF file –metadata in a readable format.

Each dataset can be accessed following the addresses:

- http://www.gbif.es:8080/ipt/resource.do?r=cndp-limno
- http://www.gbif.es:8080/ipt/resource.do?r=cndp-lichen

CONCLUSSIONS

A procedure for connecting two large data avenues such as SCAR and GBIF has been defined in a way that can be easily expanded and replicated.

The work carried out also demonstrate to projects dealing with polar biodiversity data how the impact of their research can be multiplied.

GOBIERNO DE ESPAÑA · MINISTERIO DE ECONOMÍA Y COMPETITIVIDAD · CSIC Consejo Superior de Investigaciones Científicas · Instituto Geológico y Minero de España · GBIF · gbif.es