GFBIO COLLECTION DATA CENTERS: HARMONIZED DATA PIPELINES FOR OCCURRENCE DATA

Tanja Weibulat[1,2], Christian Ebeling[3], Maren Gleisberg[4], Falko Glöckler[5], Birgit Klasen[6], Juan Carlos Monje[7], Anke Penzlin[8], Dagmar Triebel[1]

FAIR ACCESS TO OCCURRENCE DATA

ABCD is a well recognized TDWG community standard and XML schema for collection and occurrence data. It is appropriate to exchange primary biodiversity data records that document digital specimens and species' occurrences in time and space. The GFBio Submission System (fig. 1) is supporting data producers to submit such data for FAIR access guided by the data centers at natural science collections.

ABCD OCCURRENCE DATA IN GFBIO

ABCD structured information packages with single FAIR records, as generated by the BioCASe Provider Software, are provided with Persistent Identifier (PID, e.g. DOI) assignment for download under CC BY license.

The access is organised via the GFBio Data Portal (fig. 2) supported by metadata as delivered by the data centers via the GFBio instance of the BioCASe Monitor Service.

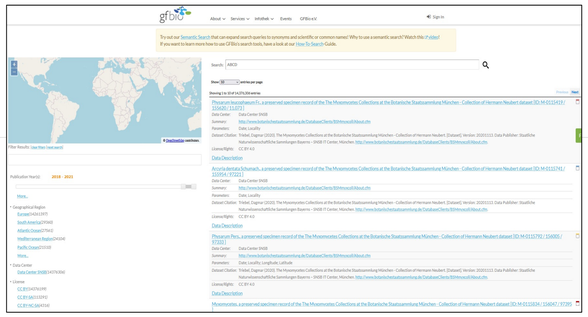

Figure 2: ABCD datasets for download via the GFBio data portal

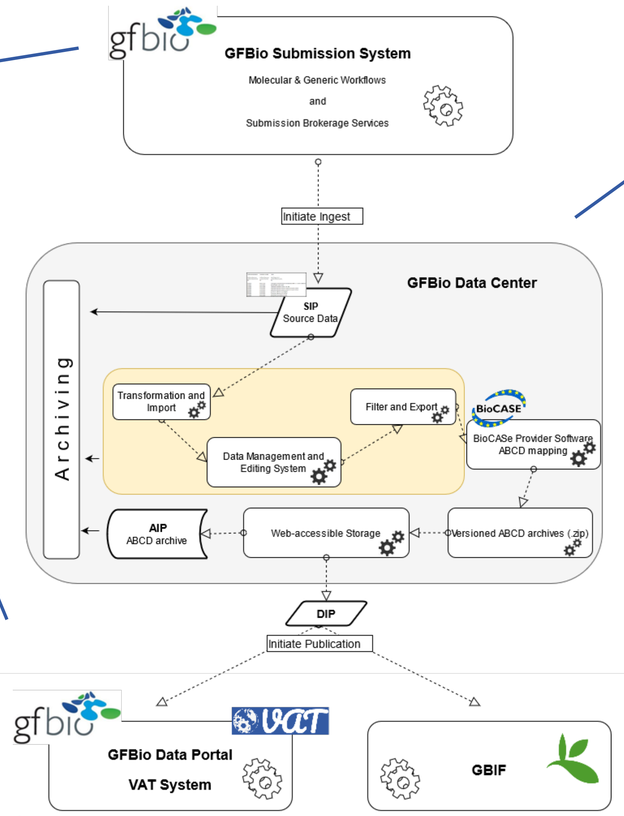

Figure 1: Harmonized data pipelines at the seven GFBio Collection Data Centers

SEVEN GFBIO COLLECTION DATA CENTERS

The seven GFBio collection data centers (see icons below) have established harmonised data pipelines for data ingestion, data management, archiving and publication. The dataflows follow the principles of the Open Archival Information System (OAIS). Figure 1 gives an overview, detailed diagrams are provided via the GFBio Wiki (https://gfbio.biowikifarm.net). The data centers are relying on their in-house solutions of (collection) data management and editing systems.

BIOCASE PROVIDER SOFTWARE

The installation of the BioCASe Provider Software (BPS, fig. 5) on a web server of each data center is mandatory. BPS is an XML data binding middleware and used as an abstraction layer in front of a database. It delivers ABCD structured XML information packages for archival and dissemination in zip archives. These can be updated by content and used by distributed networks. Data as well as linked multimedia data objects are stored on servers at the respective data centers.

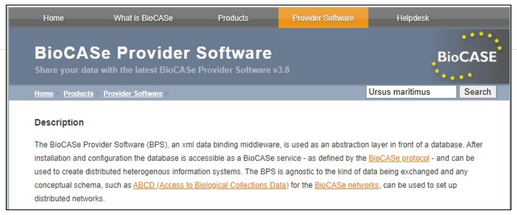

Figure 5: Website of the BioCASe Provider Software

VISUALISATION IN THE GFBIO VAT TOOL

The more than 15 million single ABCD collection and occurrence data records are valuable resources in the VAT Tool (fig. 3) and used as FAIR data backbone ("free and open data") for analysis purposes.

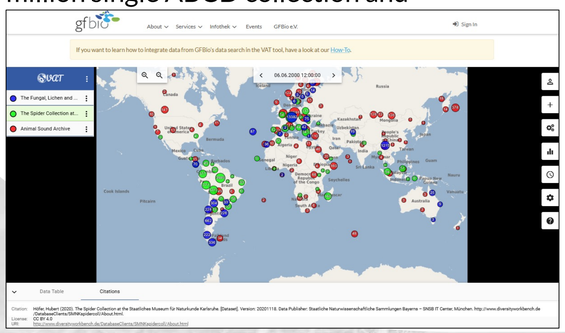

Figure 3: ABCD occurrence data in the GFBio VAT Tool

PERSPECTIVES

The harmonized GFBio data pipelines provide FAIR and dynamic access to occurrence data approved by natural science collections. Thus, they might be comfortable for planned data services of the NFDI4Biodiversity consortium, similar as those offered by the Living Atlases community (fig. 4).

Figure 4: Website of the Living Atlases Community portal

ABCD DATA IN GBIF

The long-term involvement of GFBio collection data centers within the GBIF network provides additional options for single researchers and research groups to publish identifiable occurrence data with PID assignment via the data portal of the Global Biodiversity Information Facility (GBIF, fig. 6).

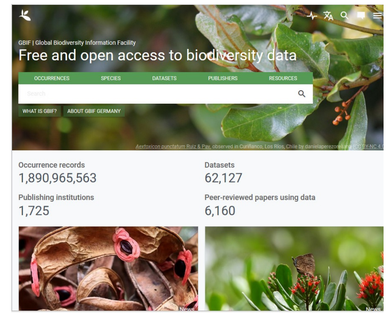

Figure 6: Open access via the GBIF portal

Funded by DFG Deutsche Forschungsgemeinschaft German Research Foundation