CURATION AT GFBIO DATA CENTERS

Tanja Weibulat[1,2], Maren Gleisberg[3], Birgit Klasen[4], Ivaylo Kostadinov[2], Jimena Linares[2], Anke Penzlin[5], Judith Weber[6]

[1]SNSB IT Center , [2]GFBio e.V., [3]BGBM, [4]ZFMK, [5]SGN, [6]Uni Bremen-MARUM/PANGAEA

GFBIO DATA CENTERS

The ten GFBio Data Centers (fig. 1) act as infrastructure partners in the GFBio broker network. Data submitted to GFBio are transmitted to and curated by data curators at the GFBio Data Centers (fig. 2). The submitted data are enriched with metadata, dependent on the data type, data format and data content. This is done in close collaboration with the data producer.

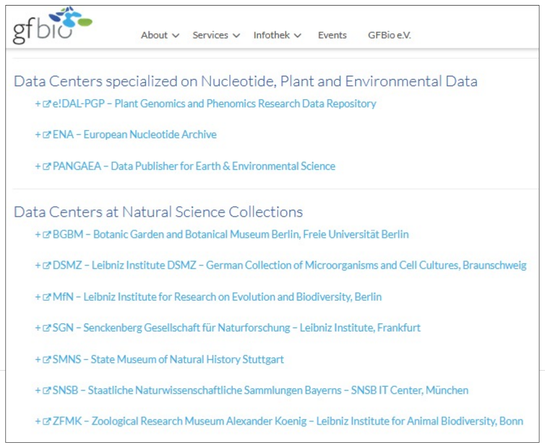

Figure 1: List of the GFBio Data Centers (https://www.gfbio.org/data-centers)

DATA CURATION DESCRIPTORS

Data curation is done in various ways and at all steps of the data life cycle. The efforts may improve data and metadata quality but are no direct indication for the quality of the original research data. Descriptors concerning metadata, research data and long-term data curation offered by GFBio Data Centers are listed in figure 2.

Descriptors of Curation Levels	Level 1	Level 2	Level 3
Exchange with data producer regarding metadata	■	■	■
Metadata are curated by data curator	■	■	■
Metadata are assigned to GFBio consensus elements of biodiversity community agreed standards for data exchange	■	■	■
Stable Identifiers (e.g. DOI, ENA-accession numbers) are assigned to published datasets	■	■	■
Exchange with data producer regarding research data content and quality		■	■
Research data are curated by data curator		■	■
Research data are assigned to GFBio consensus elements of biodiversity community agreed standards for data exchange		■	■
Research data are semantically enriched (e.g. by linking to ontologies or identifier services)		■	■
Long-term collaboration between data producer and data curator regarding dynamic datasets			■
Remote curation by the data producer			■
Continuous versioning of dynamic datasets			■

Figure 2: Curation descriptors of GFBio Data Centers arranged according to three curation levels (See also https://kb.gfbio.org/pages/viewpage.action?pageId=39256097)

COMMUNITY STANDARDS

As part of the GFBio project, approximately 40 community agreed and domain-specific standards for metadata and data exchange were identified and documented. The profiles of the ten GFBio Data Centers address 20 of them (fig. 3, see table in the GFBio Public Wiki). Community standards as well as domain-specific vocabularies are determining the results of the curation work and of the reusability of the delivered data packages.

Data exchange standards, protocols and formats relevant for the collection data domain within the GFBio network

The Natural History Collections and Culture Collection BGBM, DSMZ, MfN, SGN, SNMS, SNSB, ZFMK with their evolving GFBio Collection Data Centers' Data Archives are partners of several national and international initiatives and projects developing and using data exchange protocols and standards. During the first months of the GFBio project the partners collected and evaluated relevant technical documentations of existing domain-specific data exchange formats, interfaces and protocols with relevance for the interaction and harmonization between collection management systems and archive infrastructure as a whole (see table below). This table deals with collection standards, but it was decided to include **SDD**, **EML** and **GML** as well, as they are used and closely connected to the work within GFBio Collection Data Centers' Data Archives.

This documentation is part of the process to identify existing and to install and integrate new data exchange mechanisms and protocols appropriate for the GFBio network. It provides useful information for software developers and data scientists to set up and run GFBio agreed standard exchange software solutions at the GFBio Collection Data Centers.

Table: Data exchange standards, protocols and formats relevant for the collection data domain within the GFBio network [edit | edit source]

Note: Cells highlighted in grey indicate standards and protocols for which the GFBio Collection Data Centers' Archives have expertise or which they use directly or indirectly.

Standard/ Protocol - Acronym	Full name/ Version	Short description	Documentation/ Schema (URL)	Category	Data domain	Status	GFBio Collection Data Centers – Expertise	Notes/ References
ABCD 2.06	Access to Biological Collections Databases v2.06 (2007-06-13)	Standard for the access to and exchange of data about specimens and observations	XSD File, Schema; see also certain subversions of ABCD 2.06	Data Exchange Standard	Collection Archives	accepted (TDWG)	BGBM, DSMZ, MfN, SGN, SNMS, SNSB, ZFMK	[1] [2] [3]
ABCD 2.1	Access to Biological Collections Databases v2.1 (2014-05)	Enhanced version developed for GGBN and BiNHum, but it will not be used by GBIF	XSD File, Schema	Data Exchange Standard	Collection Archives	published	BGBM, MfN, SMNS, SNSB, ZFMK	
ABCD 3.0	Access to Biological Collections Databases v3 (2019-01-31)	New version now described as an XML Schema and an Ontology. This allows the access of the standard through semantic queries, encourages element reuse and serves as basis for future software and services in the area of semantic web.	XSD File, Schema, GitHub	Data Exchange Standard	Collection Archives	published	BGBM, MfN	[4]
ABCDDNA	DNA Extension for ABCD v2.06 (2009-05-27)	Standardised XML Schema extension for ABCD to facilitate storage and exchange of data related to DNA collection units. It offers a rudimentary set of DNA-specific data (Sequences).	XSD File, Schema	Data Exchange Standard	Collection Archives	draft (TDWG)	BGBM, DSMZ, SGN, SNSB (ZSM)	[5]
ABCDEFG	Access to Biological Collection Databases Extended for Geosciences	Standard developed for use with palaeontological, mineralogical and geological digitalized collection data	XSD File, Schema	Data Exchange Standard	Collection Archives	proposed (TDWG)	BGBM, DSMZ, MfN, SGN, SMNS, SNSB	[6] [7]
ABCDHISPID	Herbarium Information Standards and Protocols for Interchange of Data (HISPID). HISPID5 was presented during TDWG 2007	HISPID5 is a file format serving as extension for ABCD v2.06. It was developed by Australian herbaria to enable the interchange of plant specimen data.	Documentation ; google code ZIP	Data Exchange Standard	Collection Archives	published	(BGBM expertise; HISPID is used in Australia)	[8]
AC	Audubon Core Multimedia Resources Metadata Standard v1.0	A set of vocabularies designed to represent metadata for biodiversity multimedia resources and collections.	Documentation, AC on github	Metadata Standard	Collection Archives	accepted (TDWG)	BGBM, SMNS, SNSB, ZFMK (all BiNHum-Partners)	[9]
BioCASE UAP[10]	Unitlevel Access Protocol v1.31 (2012-11-14)	The protocol used in the BioCASE unit-level network for communication between the central software and the wrapper software sitting on top of the providers databases.	Documentation, XSD	Data Exchange Protocol	Collection Archives	published	BGBM, DSMZ, MfN, SGN, SMNS, SNSB, ZFMK	[10]
CDM-light	CDM-light v2.04 (2019-12-14)	A light version of the Common Data Model (CDM) developed in the context of the EDIT platform for cybertaxonomy	Documentation	Data Exchange Standard	Collection Archives	published	BGBM	[12]
DC	Dublin Core: Metadata Element Set[13] v1.1 (2013-06-13)	Dublin Core is a Metadata Standard that was originally developed for libraries but its elements have been reused in many other formats as well, e.g. DWC. Dublin Core	Documentation, XSD	Metadata Standard	Literature; General	ISO Standard	(BGBM, DSMZ, MfN, SGN, SMNS, SNSB, ZFMK	[14]

Figure 3: Data exchange standards, protocols and formats relevant for the biodiversity and collection data domain within the GFBio network (https://gfbio.biowikifarm.net).

GFBIO CURATION LEVELS

The curation levels set up by Task Group 2 – Curation Levels & Criteria in GFBio are intended to provide guidance for data producers. They give them a first impression of the kind and amount of curation efforts needed for handling different types of data and metadata structured according various standard schemas and formats (fig. 3).

Level 1 curation addresses metadata curation only. It is applied for example to data packages with metadata structured according community agreed standards for metadata exchange, e.g. EML (Ecological Metadata Language). The research data themselves are structured according proprietary schemes. After curation at the GFBio Data Centers data will be published once with PID assignment.

Level 2 curation comprises metadata curation as for level 1. In addition the curation efforts include aspects of research data curation, e.g. conceptual schema and ontology assignment. After curation at the GFBio Data Centers data will be published once with PID assignment.

Level 3 curation is meant in addition to level 1 and 2 for data packages from long-term research projects. The data are curated for a longer period and published dynamically with 'snapshots' (versions).

Not all GFBio associated Data Centers offer Level 3 data curation. See the Data Center profiles for more details.

THE FUTURE

The GFBio Data Centers are partners of the NFDI4Biodiversity consortium and will continue their work on data curation and publication according to their profiles and portfolios.
NFDI4Biodiversity is widening the scope and network of data repositories which adds a variety of curation approaches, metadata standards and data schemas. Ongoing documentation of work will be essential.

gfbio GERMAN FEDERATION FOR BIOLOGICAL DATA

Funded by DFG Deutsche Forschungsgemeinschaft German Research Foundation