# GFBio pipelines for collection data

## a joint approach of 7 taxonomic facilities

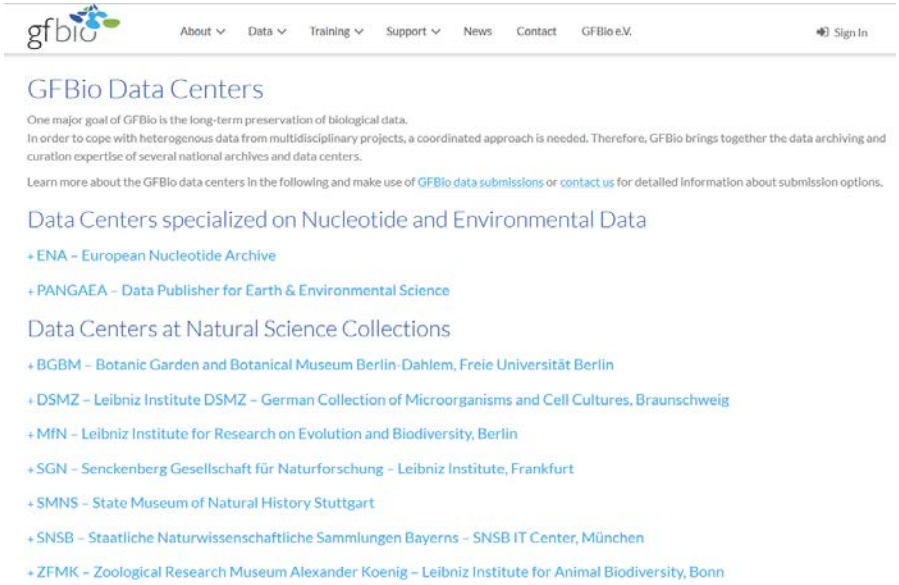
Dagmar Triebel, Peter Grobe, Anton Güntsch, Jana Hoffmann,
Joachim Holstein, Carola Söhngen, Claus Weiland

Funded by Deutsche Forschungsgemeinschaft

# GFBio Data Centers: research and collection data

**GFBio „German Federation for Biological Data“: Start in 2013**

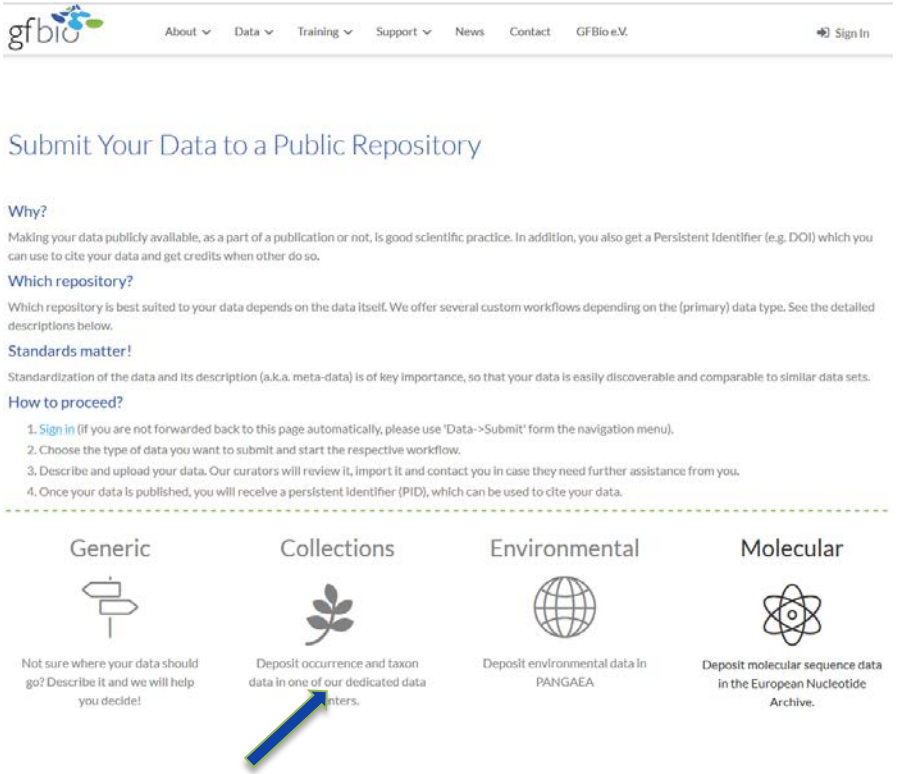
- Offers archiving and publication services for biological research and collection data
- Working on a long-term business model for data archiving and publication
- Help-Desk, training material, analysis tools, download of datasets
- Seven data centers at Taxonomic Facilities are partners for data handling, archiving and publication



# GFBio submission

**Data pipelines for collection data start with submission**

- Submission “Collections”
- Metadata EML Leaflet in GFBio portal database with web service APIs for ingest and access
- Datasets with units (collection data and monitoring data = type 1 data) offered as files for download and integration in the GFBio data centers´ systems

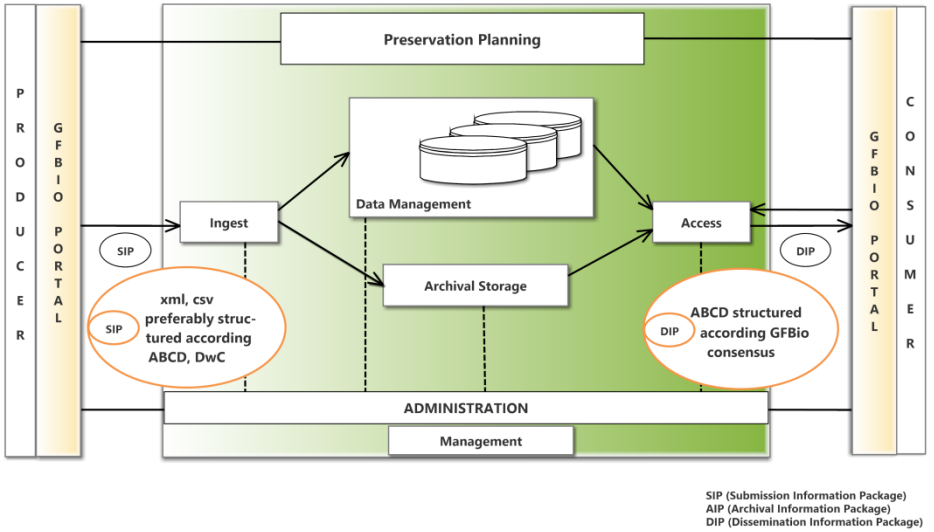


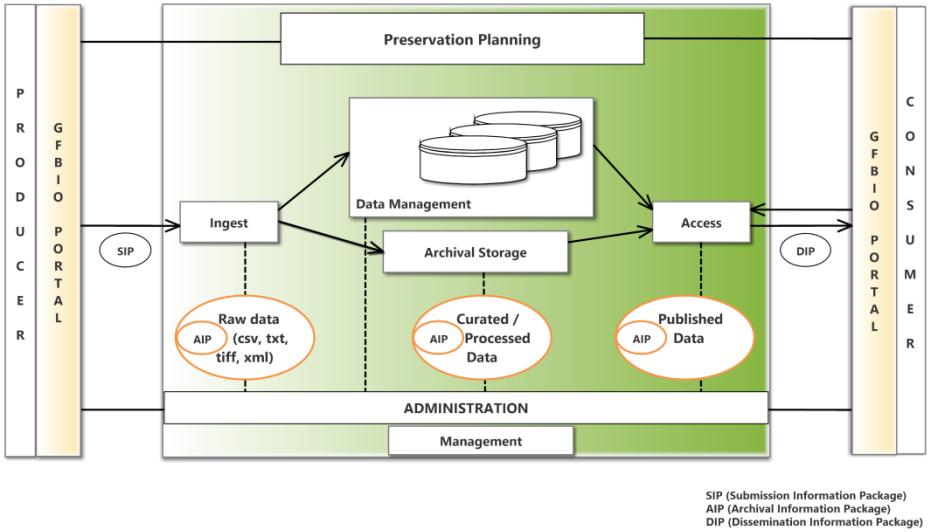
# Data centers/ archives organised according OAIS

**OAIS: Open Archival Information System**

**Information packages**

- Submission Information Package: **SIP**
- Dissemination Information Package: **DIP**
- Archiving Information Packages: **AIP + AIP + AIP**

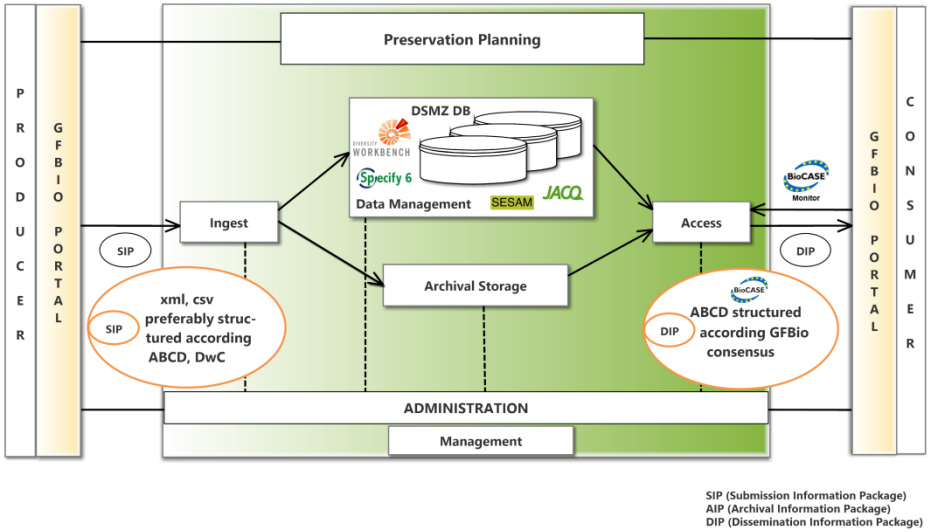


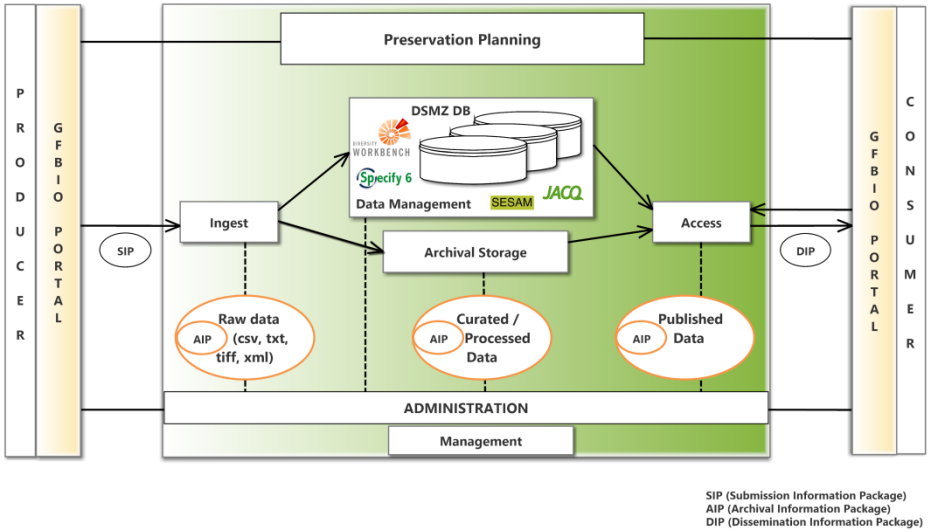



# Data archives organised according OAIS

**Data management of collection data**

- Data curation, quality control in various data management systems
- Import Wizards, import of various technical formats (xml, csv)

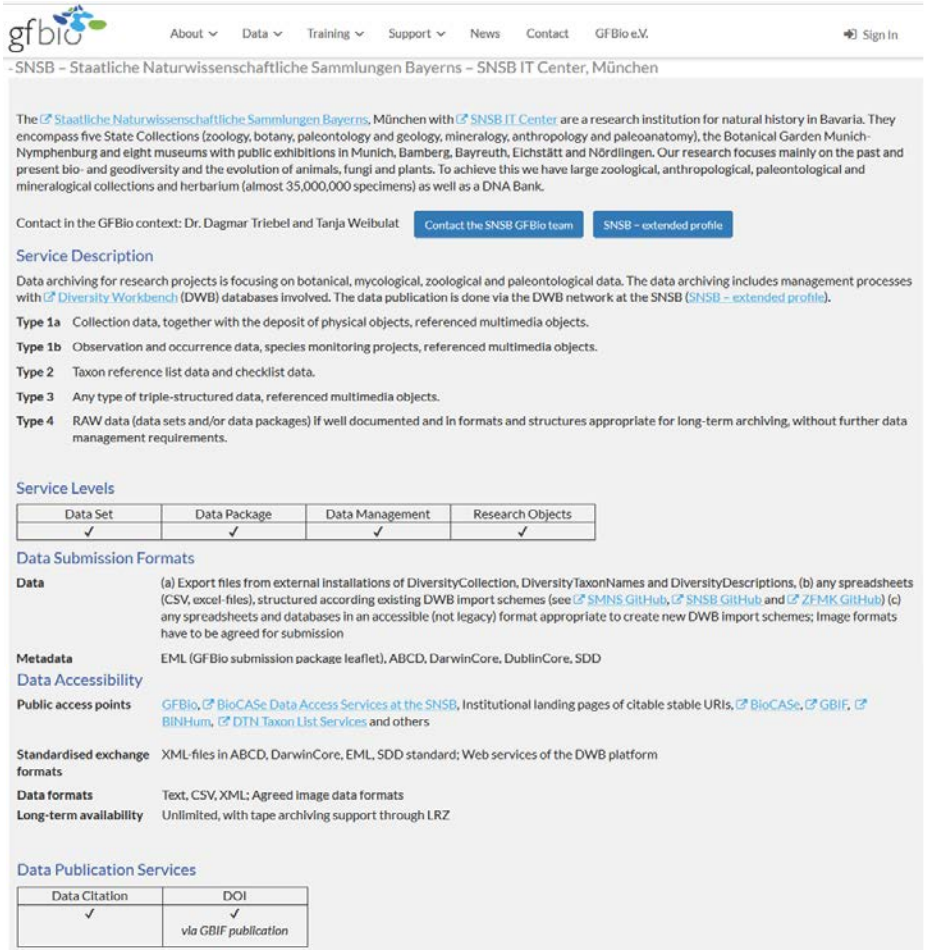


# Portfolios of the GFBio Data Centers

**Portfolio (example SNSB)**

- Service description, e. g., type 1-4
- Ingest formats
- Access and exchange formats
- Public access points



gfbio About Data Training Support News Contact GFBio e.V. Sign In

SNSB – Staatliche Naturwissenschaftliche Sammlungen Bayerns – SNSB IT Center, München

The Staatliche Naturwissenschaftliche Sammlungen Bayerns, München with SNSB IT Center are a research institution for natural history in Bavaria. They encompass five State Collections (zoology, botany, paleontology and geology, mineralogy, anthropology and paleoanatomy), the Botanical Garden Munich-Nymphenburg and eight museums with public exhibitions in Munich, Bamberg, Bayreuth, Eichstätt and Nördlingen. Our research focuses mainly on the past and present bio- and geodiversity and the evolution of animals, fungi and plants. To achieve this we have large zoological, anthropological, paleontological and mineralogical collections and herbarium (almost 35,000,000 specimens) as well as a DNA Bank.

Contact in the GFBio context: Dr. Dagmar Triebel and Tanja Weibulat [Contact the SNSB GFBio team] [SNSB – extended profile]

## Service Description

Data archiving for research projects is focusing on botanical, mycological, zoological and paleontological data. The data archiving includes management processes with Diversity Workbench (DWB) databases involved. The data publication is done via the DWB network at the SNSB (SNSB – extended profile).

**Type 1a** Collection data, together with the deposit of physical objects, referenced multimedia objects.

**Type 1b** Observation and occurrence data, species monitoring projects, referenced multimedia objects.

**Type 2** Taxon reference list data and checklist data.

**Type 3** Any type of triple-structured data, referenced multimedia objects.

**Type 4** RAW data (data sets and/or data packages) if well documented and in formats and structures appropriate for long-term archiving, without further data management requirements.

## Service Levels

| Data Set | Data Package | Data Management | Research Objects |
|---|---|---|---|
| ✓ | ✓ | ✓ | ✓ |

## Data Submission Formats

**Data** (a) Export files from external installations of DiversityCollection, DiversityTaxonNames and DiversityDescriptions, (b) any spreadsheets (CSV, excel-files), structured according existing DWB import schemes (see SMNS GitHub, SNSB GitHub and ZFMK GitHub) (c) any spreadsheets and databases in an accessible (not legacy) format appropriate to create new DWB import schemes; Image formats have to be agreed for submission

**Metadata** EML (GFBio submission package leaflet), ABCD, DarwinCore, DublinCore, SDD

## Data Accessibility

**Public access points** GFBio, BioCASe Data Access Services at the SNSB, Institutional landing pages of citable stable URIs, BioCASe, GBIF, BINHum, DTN Taxon List Services and others

**Standardised exchange formats** XML-files in ABCD, DarwinCore, EML, SDD standard; Web services of the DWB platform

**Data formats** Text, CSV, XML; Agreed image data formats

**Long-term availability** Unlimited, with tape archiving support through LRZ

## Data Publication Services

| Data Citation | DOI |
|---|---|
| ✓ | ✓ *via GBIF publication* |

# Data centers working with Diversity Workbench

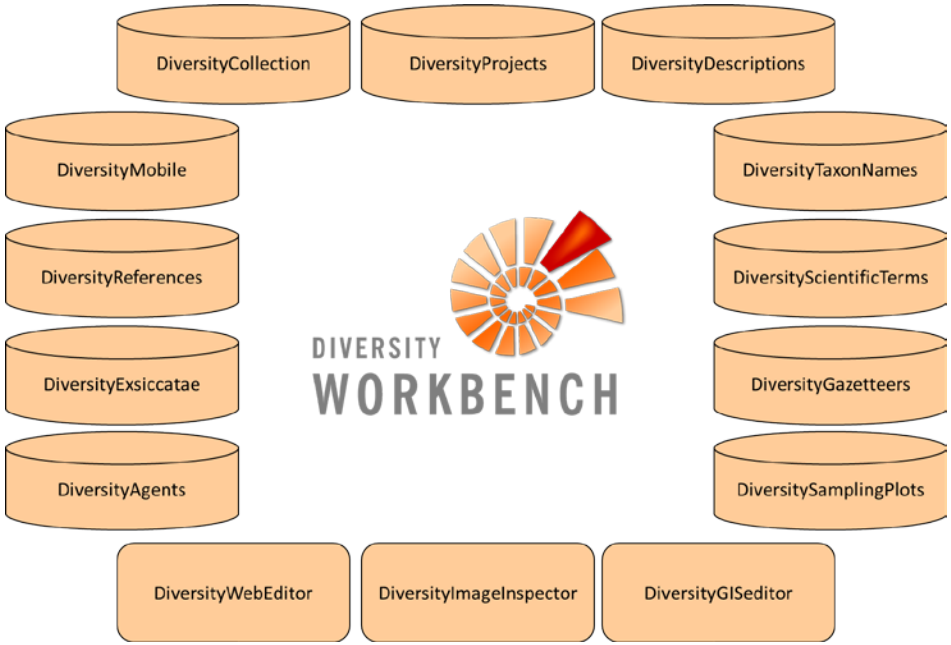
**Portfolios of SMNS, SNSB and ZFMK with similar data pipelines**

- Use of Diversity Workbench software for data ingest and management
- Use the same import schemes
- Implementation of agreed backup and archiving pipelines services
- Implementation of agreed publication pipelines

NATURKUNDE MUSEUM STUTTGART

FORSCHUNGS museum KOENIG

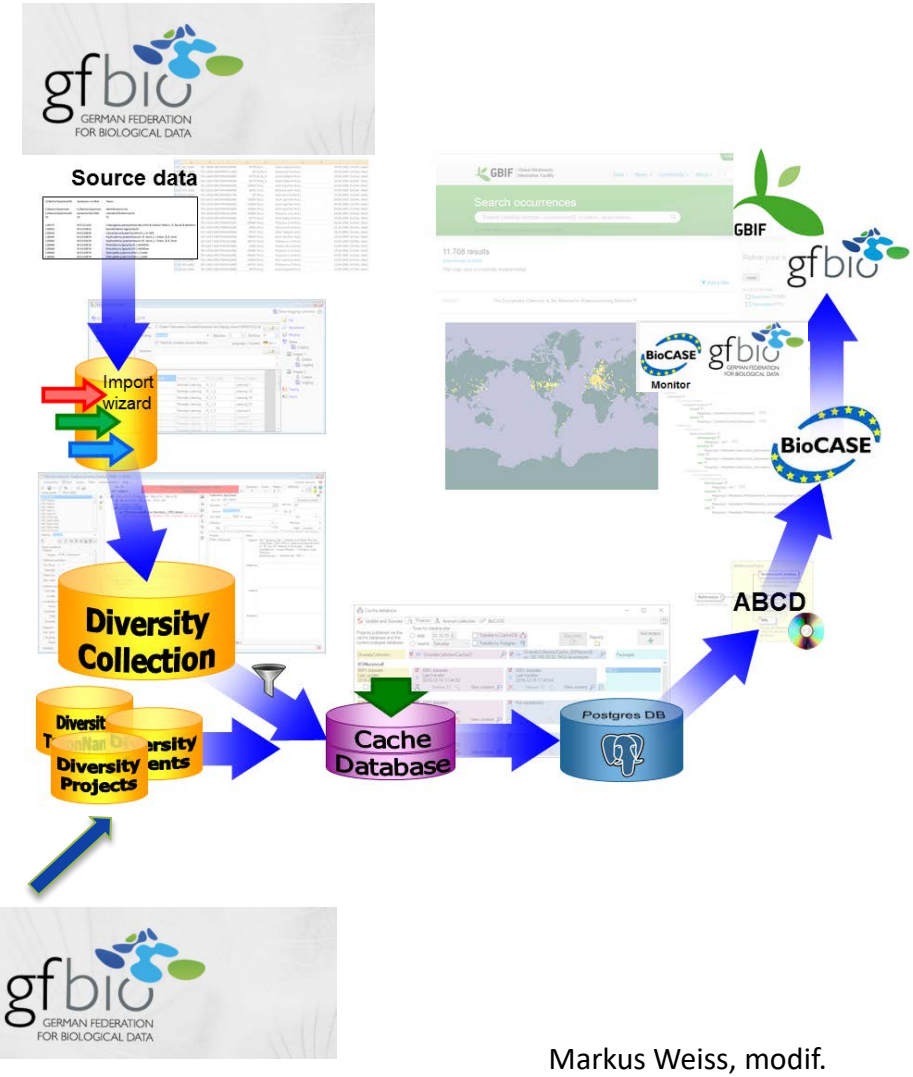
staatliche naturwissenschaftliche sammlungen bayerns



# Data centers working with Diversity Workbench

**Data pipeline for ingest and publication of „Type 1“ data**

- Metadata ingest from GFBio portal API in DiversityProjects
- Import of collection data (raw data) from xml files/ csv-files in DiversityCollection

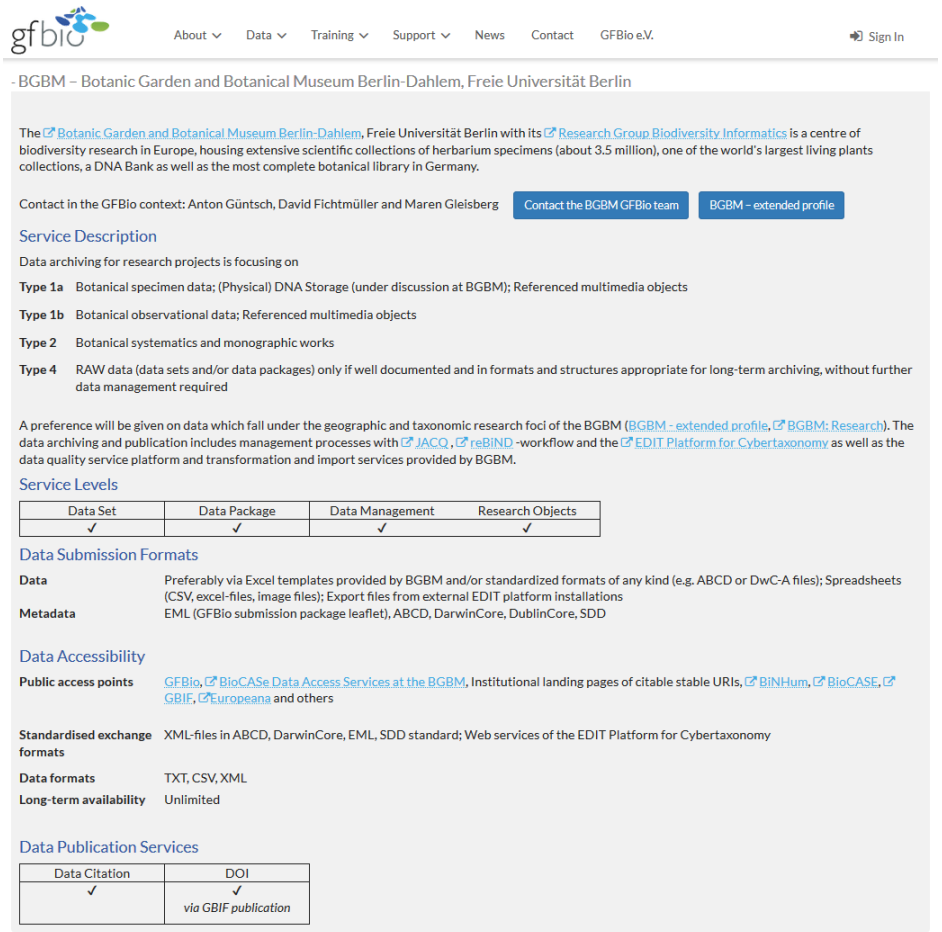



Markus Weiss, modif.

# Portfolio: GFBio Data Center BGBM

**Portfolio (example BGBM)**

- Service description, e. g., type 1-4
- Ingest formats
- Access and exchange formats
- Public access points



gfbio About Data Training Support News Contact GFBio e.V. Sign In

BGBM – Botanic Garden and Botanical Museum Berlin-Dahlem, Freie Universität Berlin

The Botanic Garden and Botanical Museum Berlin-Dahlem, Freie Universität Berlin with its Research Group Biodiversity Informatics is a centre of biodiversity research in Europe, housing extensive scientific collections of herbarium specimens (about 3.5 million), one of the world's largest living plants collections, a DNA Bank as well as the most complete botanical library in Germany.

Contact in the GFBio context: Anton Güntsch, David Fichtmüller and Maren Gleisberg | Contact the BGBM GFBio team | BGBM – extended profile

## Service Description

Data archiving for research projects is focusing on

**Type 1a** Botanical specimen data; (Physical) DNA Storage (under discussion at BGBM); Referenced multimedia objects

**Type 1b** Botanical observational data; Referenced multimedia objects

**Type 2** Botanical systematics and monographic works

**Type 4** RAW data (data sets and/or data packages) only if well documented and in formats and structures appropriate for long-term archiving, without further data management required

A preference will be given on data which fall under the geographic and taxonomic research foci of the BGBM (BGBM - extended profile, BGBM: Research). The data archiving and publication includes management processes with JACQ , reBIND -workflow and the EDIT Platform for Cybertaxonomy as well as the data quality service platform and transformation and import services provided by BGBM.

## Service Levels

| Data Set | Data Package | Data Management | Research Objects |
|---|---|---|---|
| ✓ | ✓ | ✓ | ✓ |

## Data Submission Formats

**Data** Preferably via Excel templates provided by BGBM and/or standardized formats of any kind (e.g. ABCD or DwC-A files); Spreadsheets (CSV, excel-files, image files); Export files from external EDIT platform installations

**Metadata** EML (GFBio submission package leaflet), ABCD, DarwinCore, DublinCore, SDD

## Data Accessibility

**Public access points** GFBio, BioCASe Data Access Services at the BGBM, Institutional landing pages of citable stable URIs, BiNHum, BioCASE, GBIF, Europeana and others

**Standardised exchange formats** XML-files in ABCD, DarwinCore, EML, SDD standard; Web services of the EDIT Platform for Cybertaxonomy

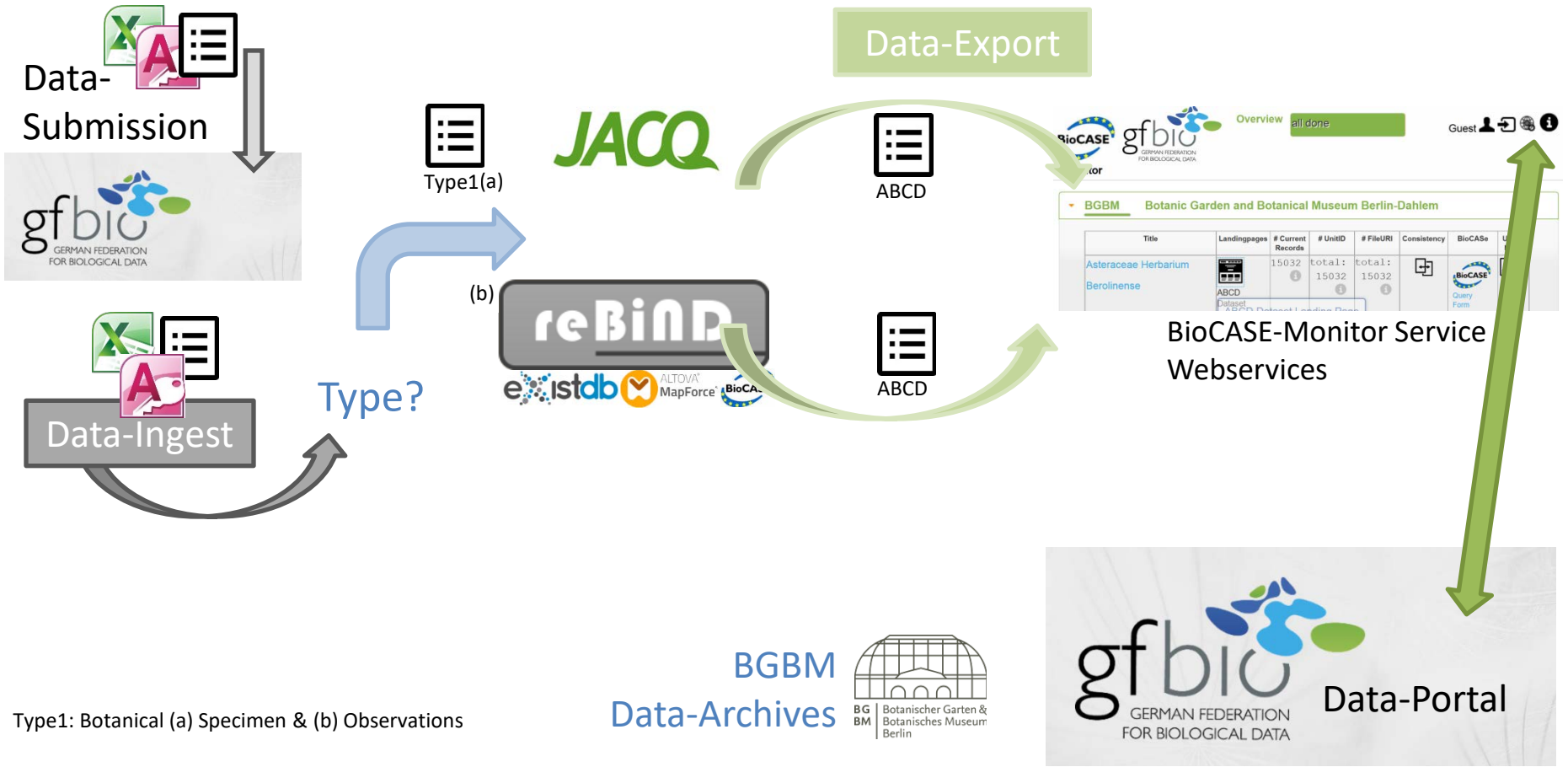
**Data formats** TXT, CSV, XML

**Long-term availability** Unlimited

## Data Publication Services

| Data Citation | DOI |
|---|---|
| ✓ | ✓ *via GBIF publication* |

# GFBio Data Center BGBM: Service offer for Type 1 data



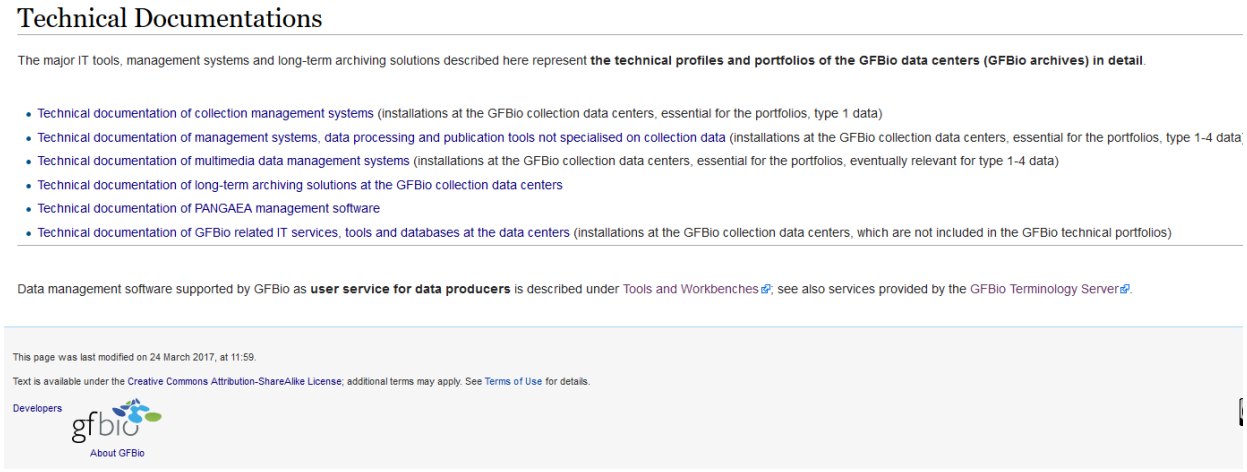

Type1: Botanical (a) Specimen & (b) Observations

Maren Gleisberg, modif.

# GFBio data centers – Technical profiles

**Documentation in Public Wiki**

- Information on technical profiles and portfolios is **important for data producers** (collection data, research data) for focused communication, e.g., on data quality and exchange
- GFBio Help Desk refers to this documentation
- Information on GFBio technical portfolios is **important for all GFBio partners** for focused communication, e.g., on data quality and exchange



Technical Documentations

The major IT tools, management systems and long-term archiving solutions described here represent **the technical profiles and portfolios of the GFBio data centers (GFBio archives) in detail**.

- Technical documentation of collection management systems (installations at the GFBio collection data centers, essential for the portfolios, type 1 data)
- Technical documentation of management systems, data processing and publication tools not specialised on collection data (installations at the GFBio collection data centers, essential for the portfolios, type 1-4 data)
- Technical documentation of multimedia data management systems (installations at the GFBio collection data centers, essential for the portfolios, eventually relevant for type 1-4 data)
- Technical documentation of long-term archiving solutions at the GFBio collection data centers
- Technical documentation of PANGAEA management software
- Technical documentation of GFBio related IT services, tools and databases at the data centers (installations at the GFBio collection data centers, which are not included in the GFBio technical portfolios)

Data management software supported by GFBio as **user service for data producers** is described under Tools and Workbenches; see also services provided by the GFBio Terminology Server.

This page was last modified on 24 March 2017, at 11:59.

Text is available under the Creative Commons Attribution-ShareAlike License; additional terms may apply. See Terms of Use for details.

Developers

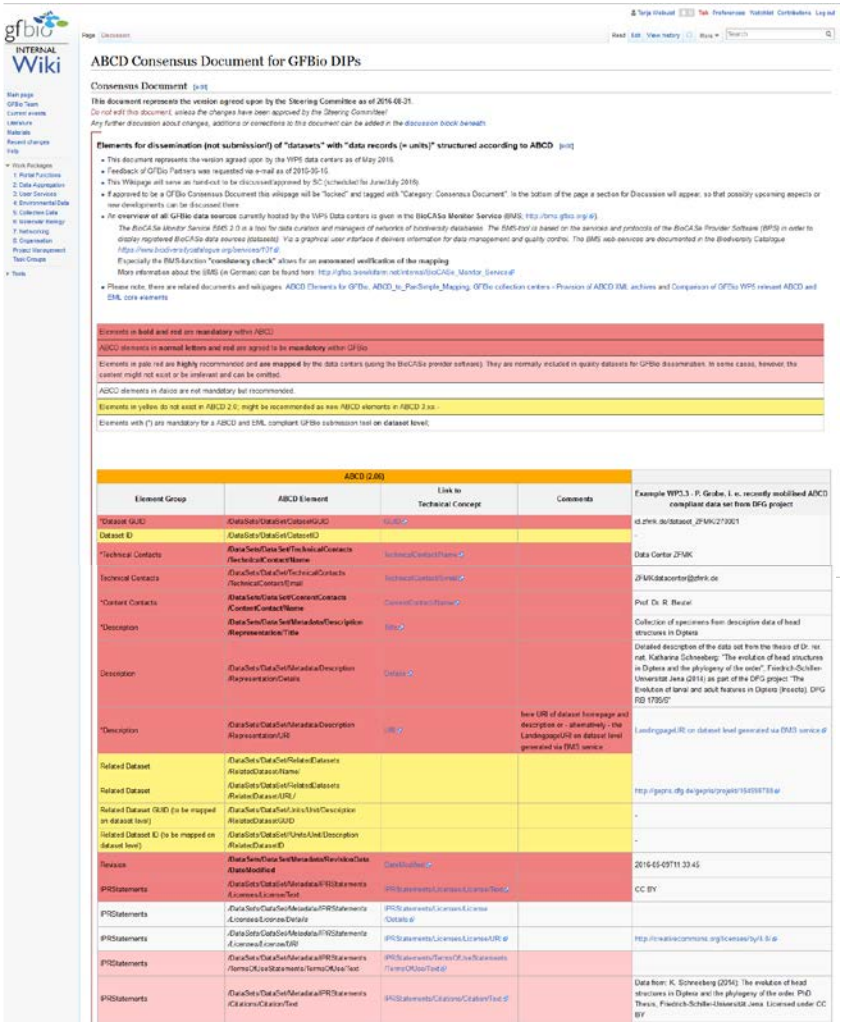
About GFBio

# ABCD Consensus Document for GFBio DIPs

**Consensus on Agreed Access for data „Type 1“**

GFBio DIPs for type 1 data “must”

- BioCASe Provider Software 3.6 installations
- ABCD 2.06
- Consensus on a number of mandatory ABCD elements
- The content of some ABCD elements is agreed, e g. citation scheme, basis of records, kingdom, landing page



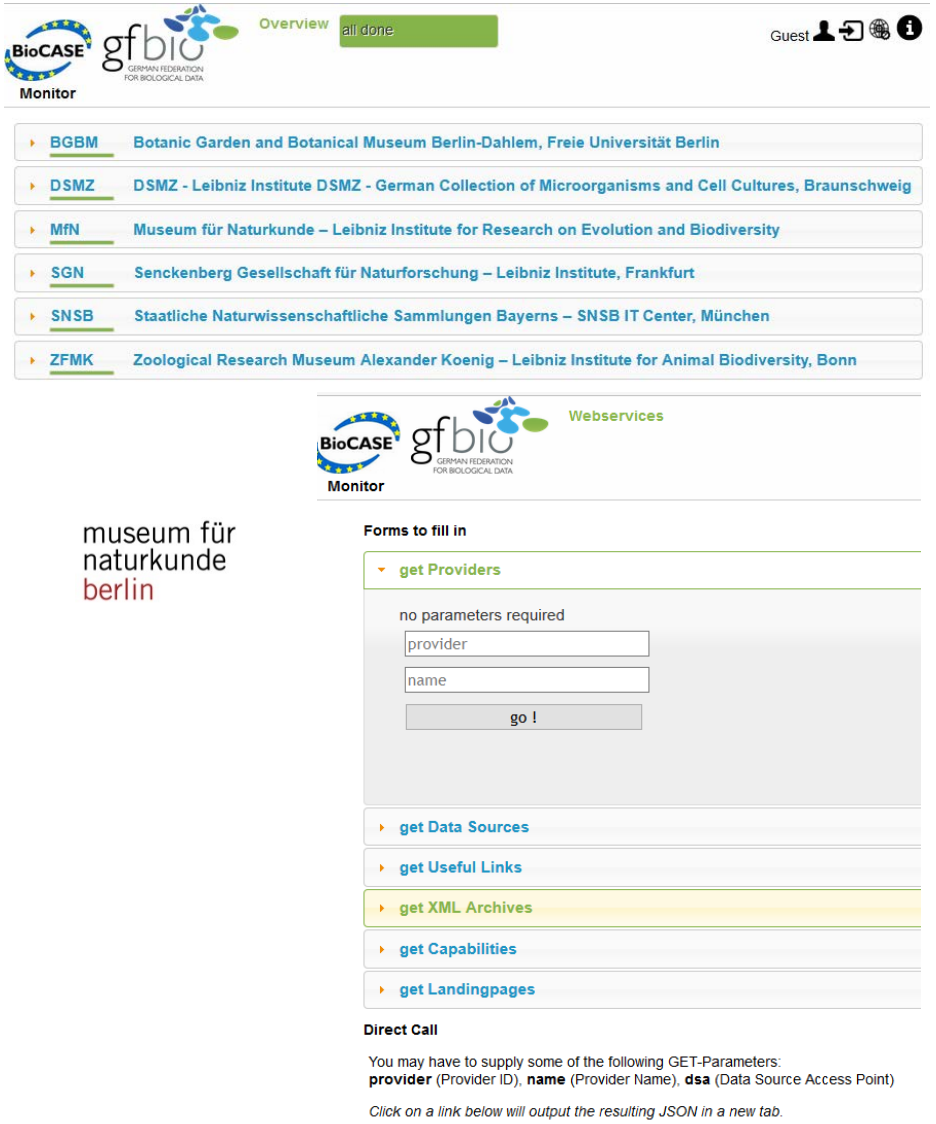



Dagmar Triebel

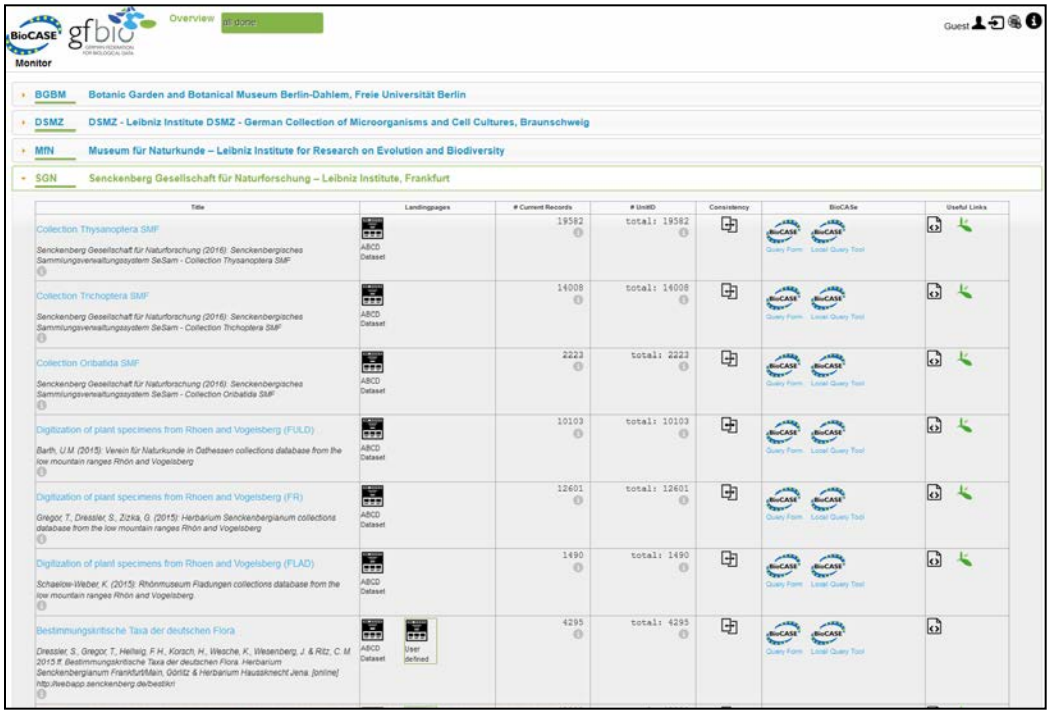
# GFBio BioCASE Monitor Service

**BMS Tool set up by MfN for control of DIPs**

- Tool for quality control and consistency check of DIPs, useful for data managers at GFBio data centers
- Access point to BioCASe provider software installations at the seven GFBio data centers
- Web services for GFBio portal partners
- Access point to ABCD xml archives for GFBio services
- ABCD → PanSimple → PANGAEA services for GFBio



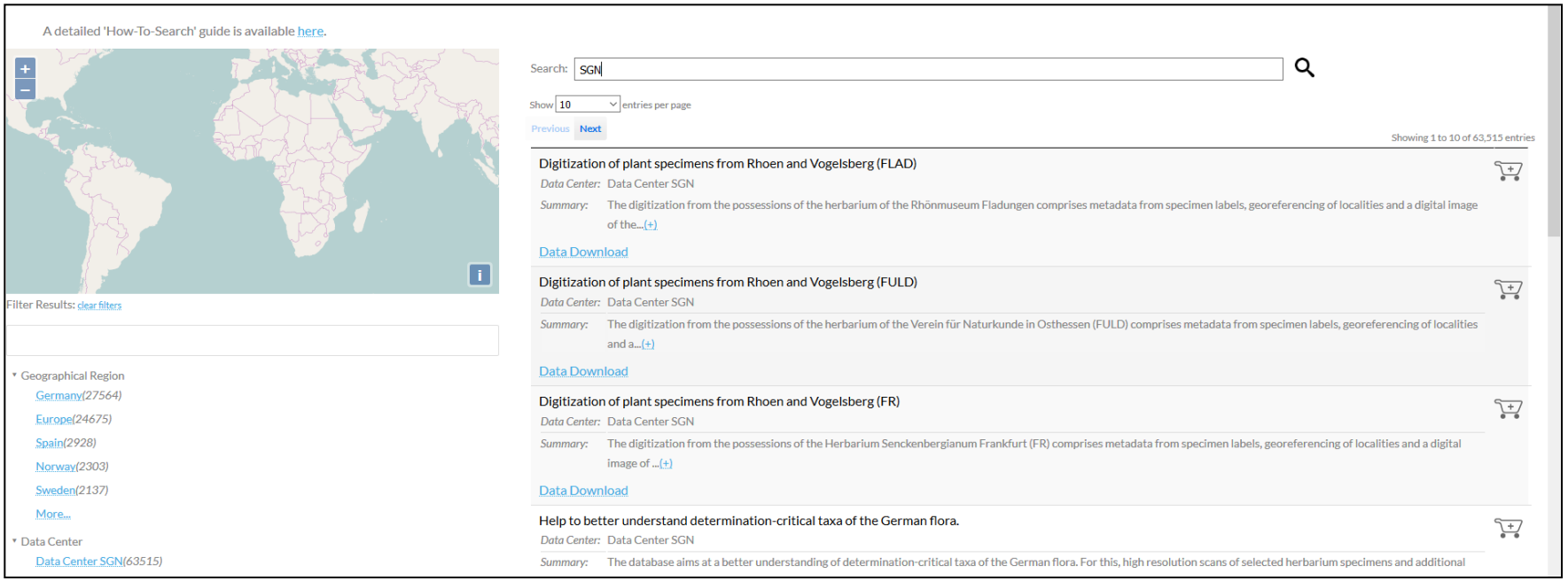
# GFBio BioCASE Monitor Service – Example SGN




museum für
naturkunde
berlin

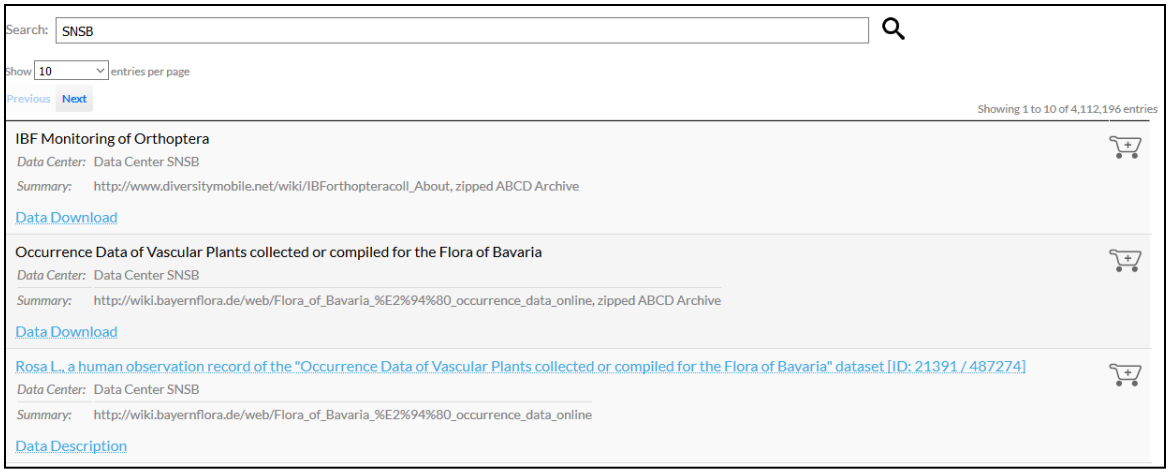
SENCKENBERG
world of biodiversity

# GFBio Data SearchPortal




SENCKENBERG
world of biodiversity

staatliche
naturwissenschaftliche
sammlungen bayerns



# The GFBio VAT (Visualization, Aggregation & Transformation) system

Addressing biodiversity research questions often depends upon aggregation, visualization and analysis of different data types at varying spatial and temporal resolutions with diverse formats. GFBio developed a Visualization, Analysis & Transformation System (VAT) that enables the synthesis of heterogeneous spatio-temporal data sets, and provides added value services via a GIS-like web-browser interface for biodiversity researchers.

# GFBio: www.gfbio.org

**Thank you!**


Team:
Christian Ebeling, David Fichtmüller, Eva-Maria Gerstner, Maren Gleisberg, Falko Glöckler, Birgit Klasen, Juan Carlos Monje, Thomas Pfuhl, Tanja Weibulat, Markus Weiss