

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



museum für
naturkunde
berlin

SENCKENBERG
world of biodiversity

NATURKUNDE
MUSEUM
STUTTGART



staatliche
naturwissenschaftliche
sammlungen bayerns

FORSCHUNGS
museum
KOENIG

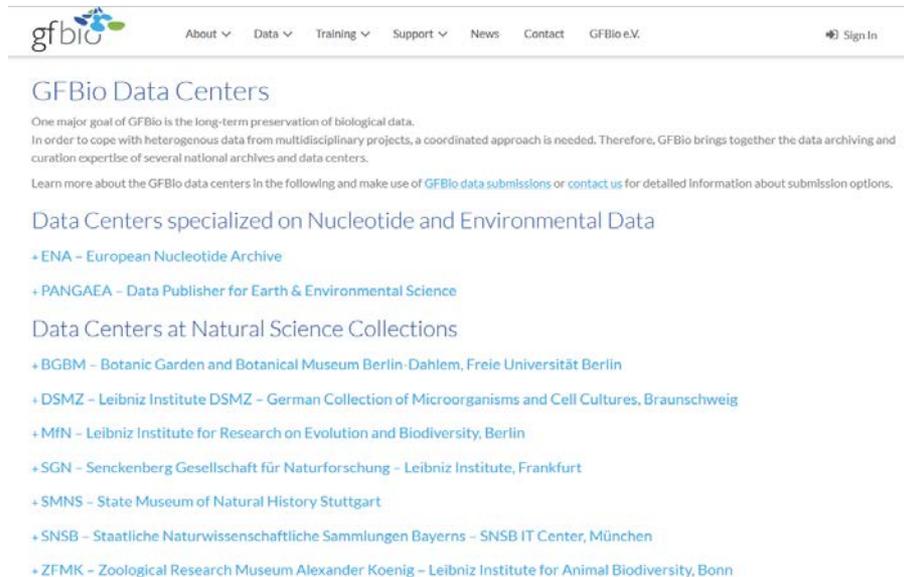
Funded by



Deutsche
Forschungsgemeinschaft

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



The screenshot shows the GFBio Data Centers website. At the top, there is a navigation bar with the GFBio logo and links for About, Data, Training, Support, News, Contact, and GFBio e.V. A Sign In button is also present. The main heading is "GFBio Data Centers". Below this, a paragraph states: "One major goal of GFBio is the long-term preservation of biological data. In order to cope with heterogenous data from multidisciplinary projects, a coordinated approach is needed. Therefore, GFBio brings together the data archiving and curation expertise of several national archives and data centers." This is followed by a link to learn more: "Learn more about the GFBio data centers in the following and make use of [GFBio data submissions](#) or [contact us](#) for detailed information about submission options." The page is divided into two sections: "Data Centers specialized on Nucleotide and Environmental Data" and "Data Centers at Natural Science Collections". The first section lists: ENA - European Nucleotide Archive and PANGAEA - Data Publisher for Earth & Environmental Science. The second section lists: BGBM - Botanic Garden and Botanical Museum Berlin - Dahlem, Freie Universität Berlin; DSMZ - Leibniz Institute DSMZ - German Collection of Microorganisms and Cell Cultures, Braunschweig; MfN - Leibniz Institute for Research on Evolution and Biodiversity, Berlin; SGN - Senckenberg Gesellschaft für Naturforschung - Leibniz Institute, Frankfurt; SMNS - State Museum of Natural History Stuttgart; SNSB - Staatliche Naturwissenschaftliche Sammlungen Bayerns - SNSB IT Center, München; and ZFMK - Zoological Research Museum Alexander Koenig - Leibniz Institute for Animal Biodiversity, Bonn.

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

Submit Your Data to a Public Repository

Why?

Making your data publicly available, as a part of a publication or not, is good scientific practice. In addition, you also get a Persistent Identifier (e.g. DOI) which you can use to cite your data and get credits when other do so.

Which repository?

Which repository is best suited to your data depends on the data itself. We offer several custom workflows depending on the (primary) data type. See the detailed descriptions below.

Standards matter!

Standardization of the data and its description (a.k.a. meta-data) is of key importance, so that your data is easily discoverable and comparable to similar data sets.

How to proceed?

1. [Sign in](#) (if you are not forwarded back to this page automatically, please use 'Data->Submit' form the navigation menu).
2. Choose the type of data you want to submit and start the respective workflow.
3. Describe and upload your data. Our curators will review it, import it and contact you in case they need further assistance from you.
4. Once your data is published, you will receive a persistent identifier (PID), which can be used to cite your data.

Generic



Not sure where your data should go? Describe it and we will help you decide!

Collections



Deposit occurrence and taxon data in one of our dedicated data centers.

Environmental



Deposit environmental data in PANGAEA.

Molecular



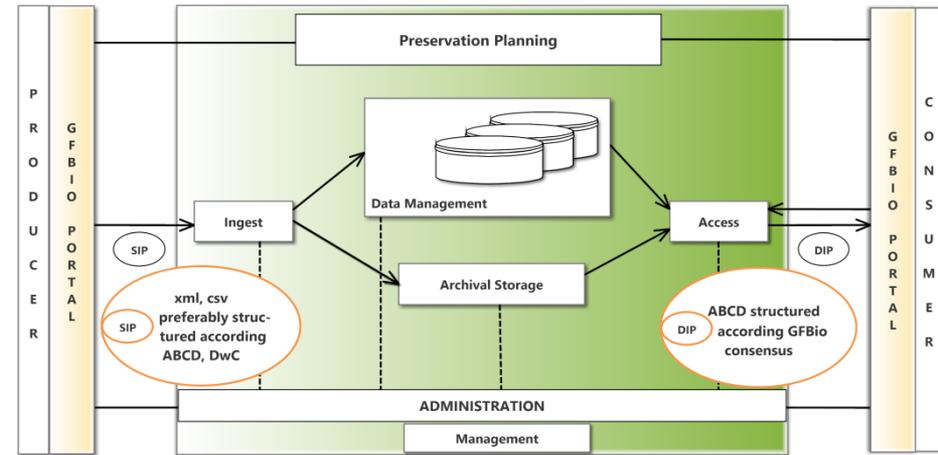
Deposit molecular sequence data in the European Nucleotide Archive.



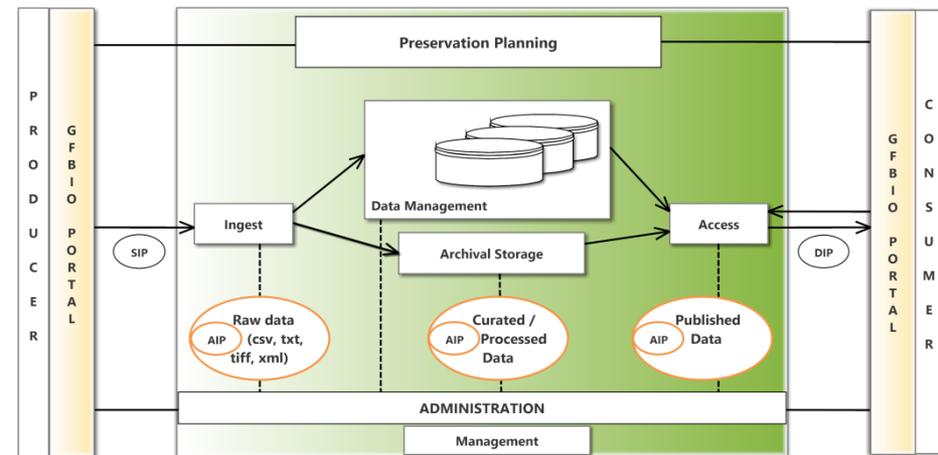
OAIS: Open Archival Information System

Information packages

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



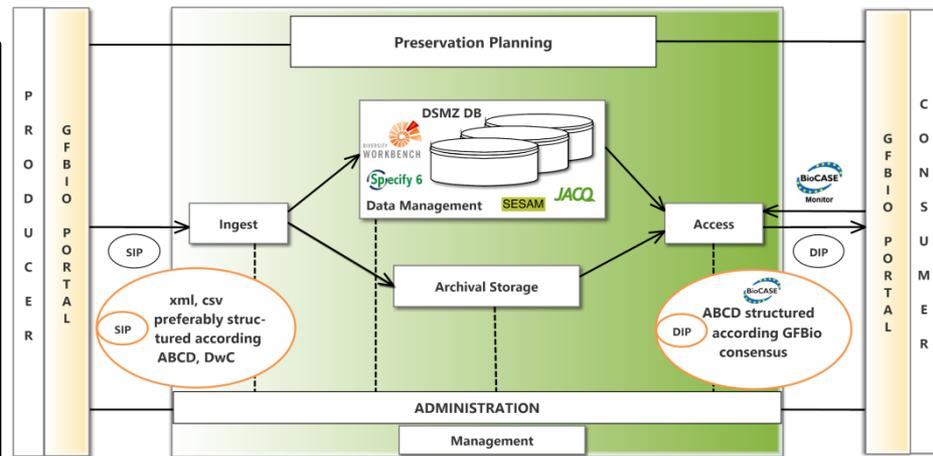
SIP (Submission Information Package)
AIP (Archival Information Package)
DIP (Dissemination Information Package)



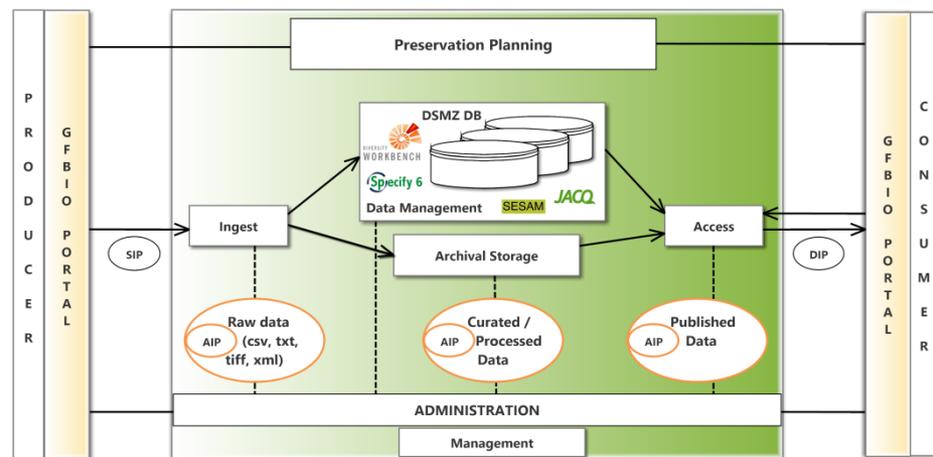
SIP (Submission Information Package)
AIP (Archival Information Package)
DIP (Dissemination Information Package)

Data management of collection data

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



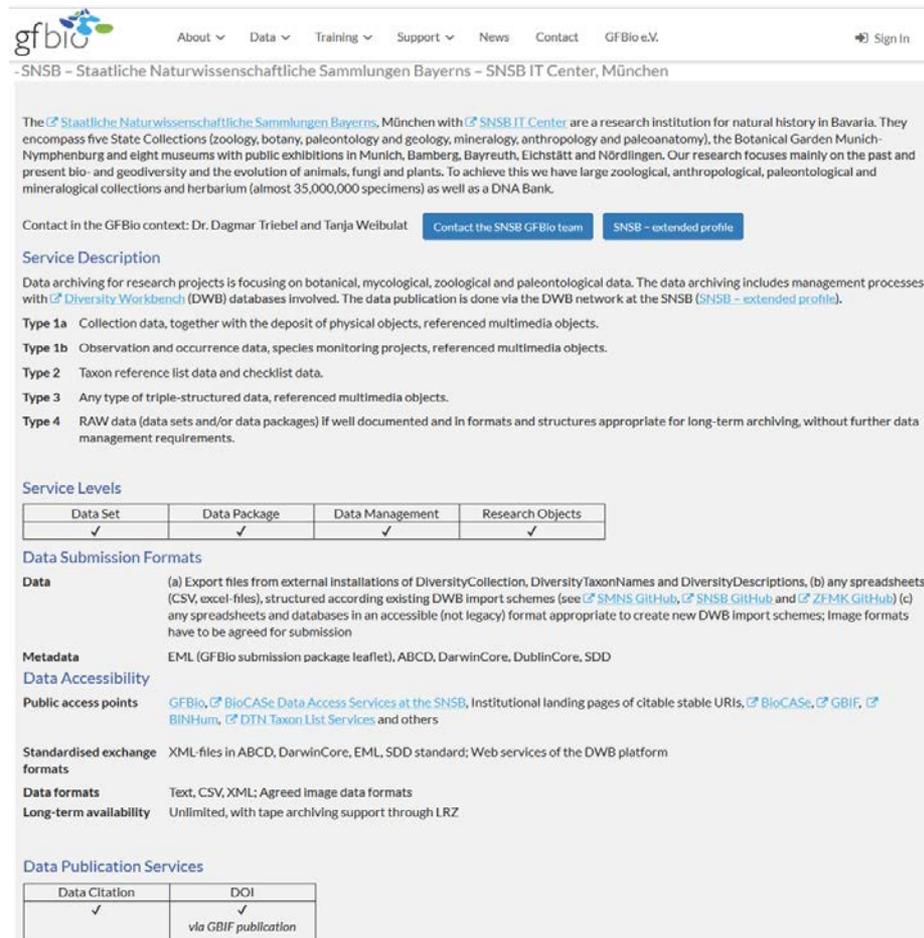
SIP (Submission Information Package)
AIP (Archival Information Package)
DIP (Dissemination Information Package)



SIP (Submission Information Package)
AIP (Archival Information Package)
DIP (Dissemination Information Package)

Portfolio (example SNSB)

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



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 |

Portfolios of SMNS, SNSB and ZFMK with similar data pipelines

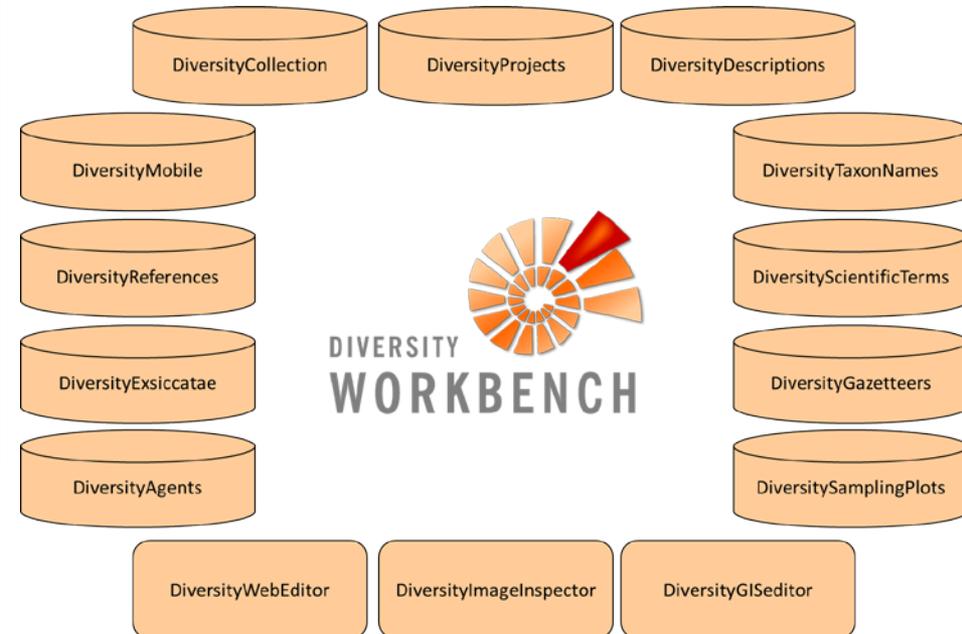
- Use of Diversity Workbench software for data ingest and management
- Use the same import schemes
- Implementation of agreed backup and archiving pipelines services
- Implementation of agreed publication pipelines

NATURKUNDE
MUSEUM
STUTT GART



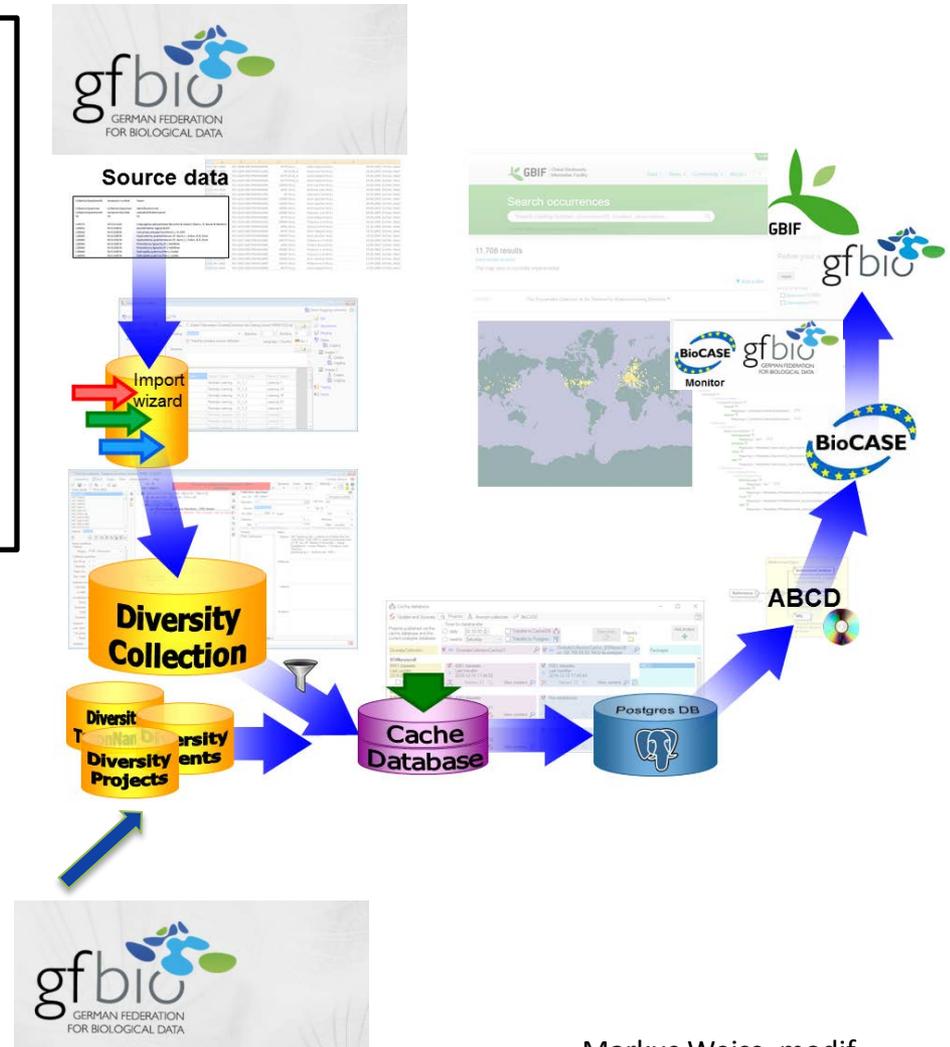
FORSCHUNGS
museum
KOENIG

staatliche
naturwissenschaftliche
sammlungen bayerns



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

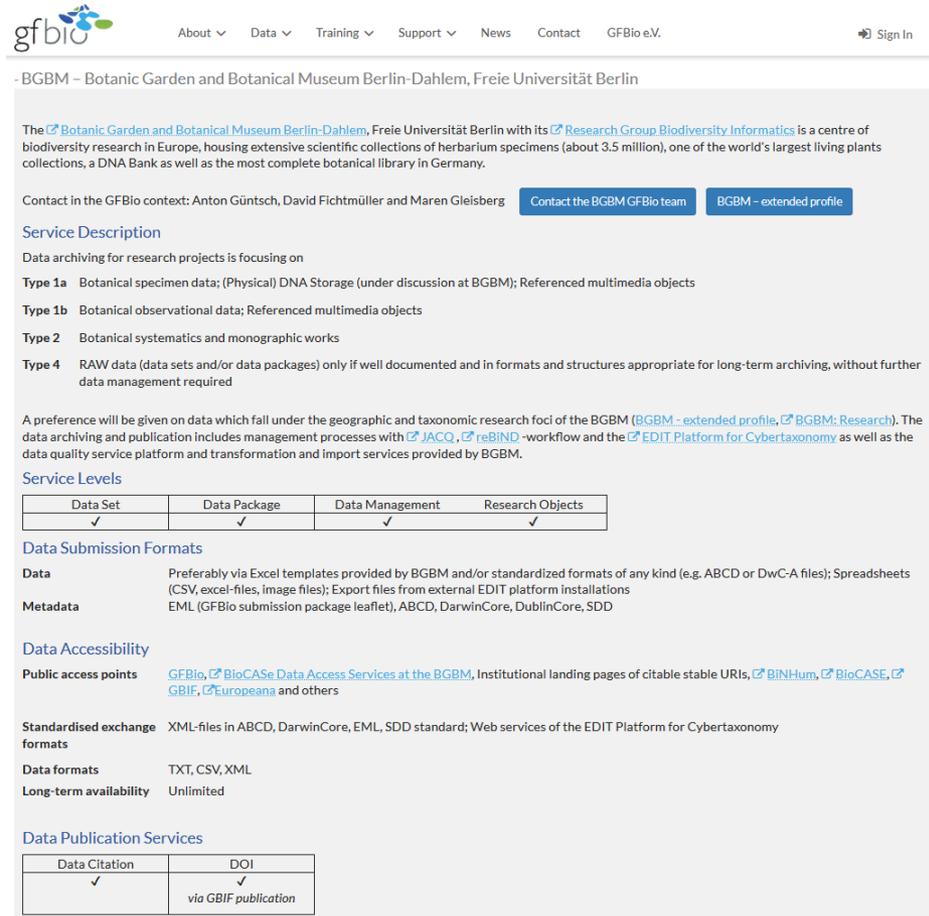
- Metadata ingest from GFBio portal API in DiversityProjects
- Import of collection data (raw data) from xml files/ csv-files in DiversityCollection



Markus Weiss, modif.

Portfolio (example BGBM)

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

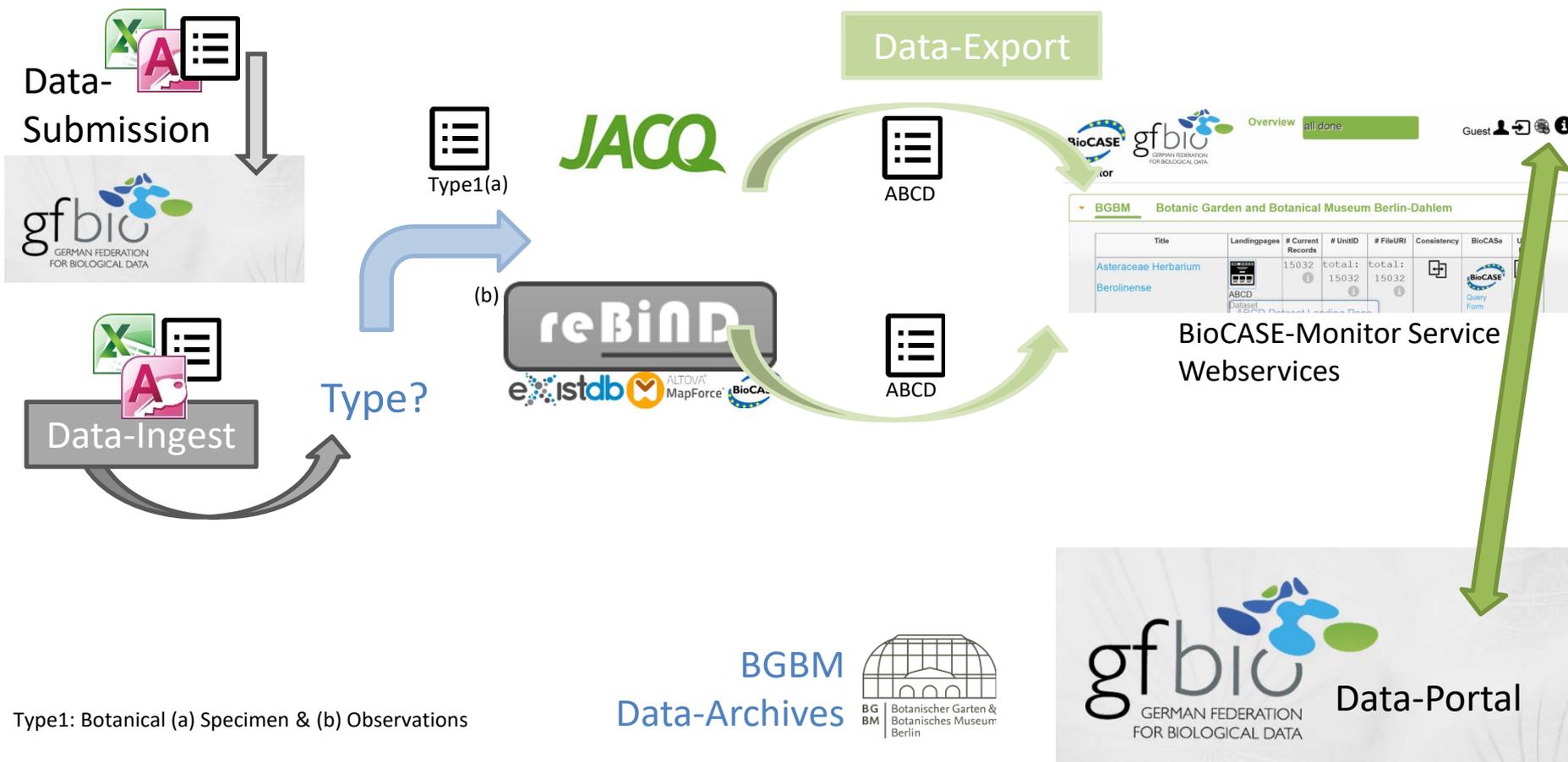


The screenshot shows the GFBio website for the Botanic Garden and Botanical Museum Berlin-Dahlem (BGBM). The page includes a navigation menu, a header with the GFBio logo, and a main content area with the following sections:

- About:** - BGBM – Botanic Garden and Botanical Museum Berlin-Dahlem, Freie Universität Berlin
- Description:** 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:** Anton Güntsch, David Fichtmüller and Maren Gleisberg. Buttons for "Contact the BGBM GFBio team" and "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
- Preference:** A preference will be given on data which fall under the geographic and taxonomic research foci of the BGBM (BGBM - extended profile, [GFBio: 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 (EML (GFBio submission package leaflet), ABCD, DarwinCore, DublinCore, SDD)
 - 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 URLs, [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 |



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

BGBM
Data-Archives
BG | Botanischer Garten & Botanisches Museum Berlin

Maren Gleisberg, modif.

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 [↗](#).

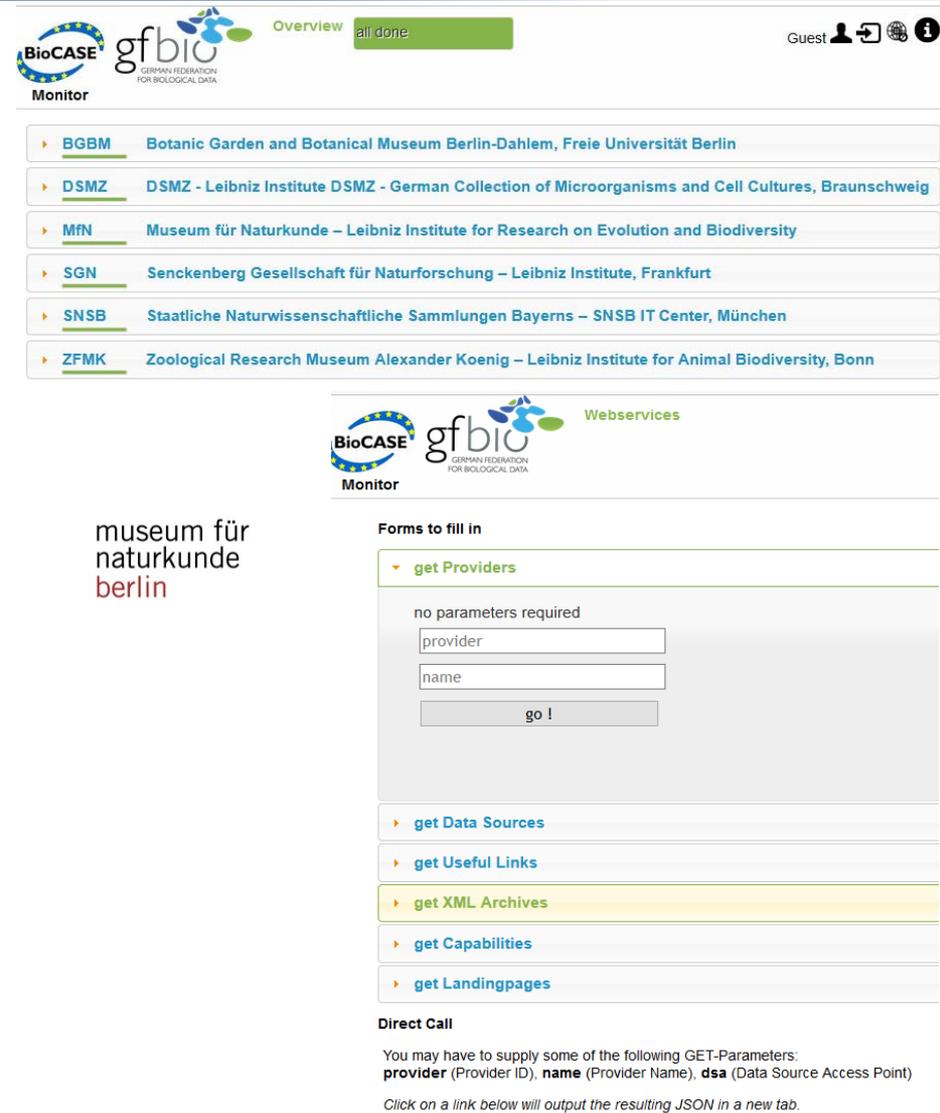
This page was last modified on 24 March 2017, at 11:59.

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



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



The screenshot shows the GFBio BioCASE Monitor Service interface. At the top, there is a navigation bar with the BioCASE and gfbio logos, the text "GERMAN FEDERATION FOR BIOLOGICAL DATA", and a green "all done" button. Below this is a "Monitor" section with a list of providers: BGBM (Botanic Garden and Botanical Museum Berlin-Dahlem, Freie Universität Berlin), DSMZ (DSMZ - Leibniz Institute DSMZ - German Collection of Microorganisms and Cell Cultures, Braunschweig), MfN (Museum für Naturkunde – Leibniz Institute for Research on Evolution and Biodiversity), SGN (Senckenberg Gesellschaft für Naturforschung – Leibniz Institute, Frankfurt), SNSB (Staatliche Naturwissenschaftliche Sammlungen Bayerns – SNSB IT Center, München), and ZFMK (Zoological Research Museum Alexander Koenig – Leibniz Institute for Animal Biodiversity, Bonn). Below the list is a "Webservices" section with the text "museum für naturkunde berlin" and a "Forms to fill in" section. The "Forms to fill in" section contains a form for "get Providers" with fields for "provider" and "name", and a "go !" button. Below the form are links for "get Data Sources", "get Useful Links", "get XML Archives", "get Capabilities", and "get Landingpages". At the bottom, there is a "Direct Call" section with a note: "You may have to supply some of the following GET-Parameters: provider (Provider ID), name (Provider Name), dsa (Data Source Access Point). Click on a link below will output the resulting JSON in a new tab."

GFBio BioCASE Monitor Service – Example SGN



gfbio BioCASE Monitor

Overview | [Help](#)

Guest | | |

- BOBM** Botanic Garden and Botanical Museum Berlin-Dahlem, Freie Universität Berlin
- DSMZ** DSMZ - Leibniz Institute DSMZ - German Collection of Microorganisms and Cell Cultures, Braunschweig
- MIN** Museum für Naturkunde – Leibniz Institute for Research on Evolution and Biodiversity
- SGN** Senckenberg Gesellschaft für Naturforschung – Leibniz Institute, Frankfurt

| Title | Landingpages | # Current Records | # InSD | Consistency | BioCASE | Useful Links |
|---|--------------|-------------------|--------------|-------------|---------|--------------|
| Collection Thysanoptera SMF Senckenberg Gesellschaft für Naturforschung (2016): Senckenbergisches Sammlungsvollständigungssystem SeSam - Collection Thysanoptera SMF | | 19582 | total: 19582 | | | |
| Collection Trichoptera SMF Senckenberg Gesellschaft für Naturforschung (2016): Senckenbergisches Sammlungsvollständigungssystem SeSam - Collection Trichoptera SMF | | 14008 | total: 14008 | | | |
| Collection Oribatida SMF Senckenberg Gesellschaft für Naturforschung (2016): Senckenbergisches Sammlungsvollständigungssystem SeSam - Collection Oribatida SMF | | 2223 | total: 2223 | | | |
| Digitization of plant specimens from Rhoe and Vogelsberg (FLUD) Berth, U M (2015): Verein für Naturkunde in Olfessen collections database from the low mountain ranges Rhoe and Vogelsberg | | 10103 | total: 10103 | | | |
| Digitization of plant specimens from Rhoe and Vogelsberg (FR) Gregor, T, Dressler, S, Zick, G (2015): Herbarium Senckenbergianum collections database from the low mountain ranges Rhoe and Vogelsberg | | 12601 | total: 12601 | | | |
| Digitization of plant specimens from Rhoe and Vogelsberg (FLAD) Schaefer-Weber, K. (2015): Rhönmuseum Fladungen collections database from the low mountain ranges Rhoe and Vogelsberg | | 1490 | total: 1490 | | | |
| Bestimmungskritische Taxa der deutschen Flora Dressler, S, Gregor, T, Hellwig, F H, Koroch, H, Wiesche, K, Wesenberg, J & Ritz, C M (2015): Bestimmungskritische Taxa der deutschen Flora. Herbarium Senckenbergianum Frankfurt/Main, Götting & Herbarium Hausrecht Jena. [online] http://webapp.senckenberg.de/beatol/ | | 4295 | total: 4295 | | | |

museum für naturkunde berlin

SENCKENBERG world of biodiversity

Summary

| #elements | checking... | errors | warnings | infos |
|-------------------|---|---|--|--|
| total: 47 | done! | 0 | 0 | 0 |
| searchable: 41 | | | | |
| not searchable: 6 | | | | |

| source element | searchable | datatype | properties/rules | counters | example values |
|---|------------|------------------|----------------------|--|-------------------|
| 1 /DataSets/DataSet/Content/Contacts/Content/Contact/Name | 1 | normalizedString | M notEmpty | count show | |
| 2 /DataSets/DataSet/Dataset/GUID | 1 | normalizedString | M notEmpty,unique | total: 4295 distinct: 1 #proposed: 0 | show |
| 3 /DataSets/DataSet/Metadata/Description/Representation/Details | 1 | normalizedString | M notEmpty | count show | |
| 4 /DataSets/DataSet/Metadata/Description/Representation/Title | 1 | normalizedString | M notEmpty,unique | total: 4295 distinct: 1 #proposed: 0 | show |
| 5 /DataSets/DataSet/Metadata/Description/Representation/URI | 1 | anyURI | M notEmpty | count show | |

A detailed 'How-To-Search' guide is available [here](#).



Filter Results: [clear filters](#)

- Geographical Region
 - [Germany\(27564\)](#)
 - [Europe\(24675\)](#)
 - [Spain\(2928\)](#)
 - [Norway\(2303\)](#)
 - [Sweden\(2137\)](#)
 - [More...](#)
- Data Center
 - [Data Center SGN\(63515\)](#)

Search: 

Show entries per page

[Previous](#) [Next](#) Showing 1 to 10 of 63,515 entries

Digitization of plant specimens from Rhoen and Vogelsberg (FLAD) 

Data Center: Data Center SGN

Summary: The digitization from the possessions of the herbarium of the Rhönmuseum Fladungen comprises metadata from specimen labels, georeferencing of localities and a digital image of the...[\(+\)](#)

[Data Download](#)

Digitization of plant specimens from Rhoen and Vogelsberg (FULD) 

Data Center: Data Center SGN

Summary: The digitization from the possessions of the herbarium of the Verein für Naturkunde in Osthessen (FULD) comprises metadata from specimen labels, georeferencing of localities and a...[\(+\)](#)

[Data Download](#)

Digitization of plant specimens from Rhoen and Vogelsberg (FR) 

Data Center: Data Center SGN

Summary: The digitization from the possessions of the Herbarium Senckenbergianum Frankfurt (FR) comprises metadata from specimen labels, georeferencing of localities and a digital image of ...[\(+\)](#)

[Data Download](#)

Help to better understand determination-critical taxa of the German flora. 

Data Center: Data Center SGN

Summary: The database aims at a better understanding of determination-critical taxa of the German flora. For this, high resolution scans of selected herbarium specimens and additional

SENCKENBERG
world of biodiversity



Search: 

Show entries per page

[Previous](#) [Next](#) Showing 1 to 10 of 4,112,196 entries

IBF Monitoring of Orthoptera 

Data Center: Data Center SNSB

Summary: http://www.diversitymobile.net/wiki/IBForthopteraColl_About, zipped ABCD Archive

[Data Download](#)

Occurrence Data of Vascular Plants collected or compiled for the Flora of Bavaria 

Data Center: Data Center SNSB

Summary: http://wiki.bayernflora.de/web/Flora_of_Bavaria_%E2%94%80_occurrence_data_online, zipped ABCD Archive

[Data Download](#)

Rosa L., a human observation record of the "Occurrence Data of Vascular Plants collected or compiled for the Flora of Bavaria" dataset [ID: 21391 / 487274] 

Data Center: Data Center SNSB

Summary: http://wiki.bayernflora.de/web/Flora_of_Bavaria_%E2%94%80_occurrence_data_online

[Data Description](#)

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.



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