# Towards an Integrated Biodiversity and Ecological Research Data Management and Archiving Platform: The German Federation for the Curation of Biological Data (GFBio)

Michael Diepenbroek, MARUM, Universität Bremen
Frank Oliver Glöckner, Jacobs University and MPI für Marine Mikrobiologie
Peter Grobe, Zoologisches Forschungsmuseum König, Bonn
Anton Güntsch, Botanischer Garten und Botanisches Museum Berlin-Dahlem
Robert Huber, MARUM, Universität Bremen
Birgitta König-Ries, Universität Jena and iDiv
Ivaylo Kostadinov, Jacobs Unversity
Jens Nieschulze, Universität Göttingen
Bernhard Seeger, Universität Marburg
Robert Tolksdorf, Freie Universität Berlin
Dagmar Triebel, Staatl. Naturwissenschaftl. Sammlungen Bayerns

**Abstract:** Biodiversity research brings together the many facets of biological environmental research. Its data management is characterized by integration and is particularly challenging due to the large volume and tremendous heterogeneity of the data. At the same time, it is particularly important: A lot of the data is not reproducible. Once it is gone, potential knowledge that could have been gained from it is irrevocably lost. In this paper, we describe challenges to biodiversity data management along the data life cycle and sketch the solution that is currently being developed within the GFBio project, a collaborative effort of nineteen German research institutions ranging from museums and archives to biodiversity researchers and computer scientists.

## 1 Introduction

Environmental and biological research is becoming central to major societal challenges related to the Earth's ecosystems and climate dynamics. To handle the scale and complexity of the scientific questions being addressed, there is a strong need to integrate knowledge. However, relevant data are currently scattered, difficult to share and often threatened to be lost. In this paper, we take a close look at these challenges and the state of the art. We then sketch a coherent infrastructure to improve scientific data integration and preservation that is currently being developed in the GFBio project. This project brings together national key players providing environmentally related biological data and services to develop the German Federation for the Curation of Biological Data (GFBio). The overall goal is to provide a sustainable, service oriented, national data infrastructure facilitating data sharing and stimulating data intensive science in the fields of biological and environmental

research. While GFBio is still in the early phases of implementation, it might well serve as blueprint for similar initiatives in other disciplines. The current setup and approach are the result of a long and thorough preparatory phase involving representatives of all important stakeholders in the field.

## 2 Importance of Biodiversity Data and its Management

With the rapid evolution of scientific methodologies and technologies during the last decades an enormous amount of data in bio-sciences has been and will continuously be generated. Further, complex scientific challenges require international, interdisciplinary approaches resulting in extremely heterogeneous data sources to be analyzed. This increase in experimental and observational data volumes and analytical complexity has leveraged a large shift in research practices from traditional hypothesis-driven approaches by domain experts towards data-driven approaches which involves computational analysis, networked sensors and data sources, data sharing, large numbers of researchers including crowd-sourcing, and interdisciplinary collaboration.

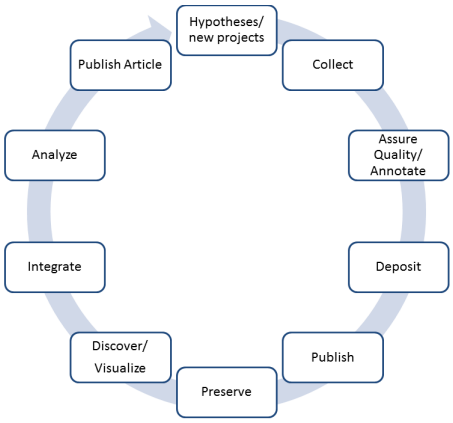


Figure 1: The Data Life Cycle

Ecosystem research, the study of organisms and their interactions with the environment, is one of the most fertile and receptive areas to such developments – this field is now key to meet global challenges such as climate change, particularly the loss of biodiversity and food resources and its impact on security. However, there is a significant lack of understanding of the complex interactions between organisms and the geo- and atmosphere. Environmental monitoring and observation systems show promising progress (e.g., GMES, EU-BON, GEO-BON, or ESONET/EMSO) but global data coverage is still patchy. Many species are still undiscovered or undescribed and even the extent of this knowledge gap is heavily disputed [MTA+11]. Our knowledge on the scales at which organisms and the environment interact is fragmentary. For any organism, a wide variety of information is of interest, describing for example their morphology, metabolism, genome, community structure and interactions, and ecological functions. An integrated analysis of this diverse knowledge requires a federated infrastructure that ensures general and long-term availability and usability of quality assured data and that supports the entire data life cycle as described by DataOne[1] (see Figure 1).

[1] DataOne: Tutorials on Data Management, `http://www.dataone.org/sites/all/documents/L01_DataManagement.pptx` [last access: 26.06.2014]

## 3 State of the Art and Challenges in Biodiversity Data Management

In this section, we will follow the data life cycle and will discuss for each phase or group of phases what it is about, how it is addressed today and what the major open challenges are that will be addressed in GFBio. We will focus on the German research landscape.

**Data Collection**

**What it is about:** Biodiversity data is very diverse, ranging from remote sensing and next-generation-sequencing data to field observations and collection object-associated data linked with categorical or numerical data. The data are from ecophysiological studies, describe functional traits or are molecular data collected for phylogenetic treatments. They are gained automatically by data loggers (e.g., sensor data from e.g., climate stations) or gathered manually by individual researchers. The result is a great heterogeneity between the single data packages or data sets which might comprise data of varying granularity and quality. Often, the single data packages/data sources are linked with other data packages, but the documentation of the linkage might be scarce. Traditionally, biodiversity research data and ecological data are often produced in spreadsheet and specialised local databases, which are not appropriate to manage complex data sets, e.g., with relations to other data sets or multimedia data, e.g., audio files.

Currently a few database systems and workbenches are on the way to support biodiversity data gathering already in the field, later in the lab or finally in the collection magazine. Such tools have to be flexible and usable in the field without internet access but also be able to integrate into a virtual research environment using network resources and services.

**How it is done today:** There exist first tools to support human data collection in the field fulfilling a number of requirements listed by [JKN+09]. Examples include DiversityMobile[2] or the recently released OSD App[3], which allows both scientists and citizens to easily contribute to a global, orchestrated scientific event – the Ocean Sampling Day – by collecting standardized metadata. Also, tools not specifically developed in the context of biodiversity to support automatic data collection from different types of sensors etc. are in use.

**Challenges addressed by GFBio:** Data collection is a particularly crucial step when looking at the data life cycle, i.e., problems occurring in this step like errors, missing data, or missing meta data, introduce severe, expensive to overcome problems later on in the life cycle. Thus, it is very important to guide researchers from the very beginning of the study through the process of collecting well structured data with appropriate technical tools. In the best case, the tools should support the data collection, improve the data quality, maybe also enable the collection of a higher data volume without restricting the scientific goal. Challenges are thus both technical (develop better tools) and educational or sociological (training scientists in data management and creating a culture that acknowledges the importance of data management). GFBio will work on both issues by extending existing tools and by providing a virtual helpdesk with guidance and training.

[2] Diversity Mobile, `http://www.diversitymobile.net` [last access 26.06.2014]
[3] Ocean Sampling Day, `http://www.oceansamplingday.org` [last access 26.06.2014]

**Quality Assurance**

**What it is about:** Research data are often managed by single researchers without support by data scientists or IT experts. Thus, the data are often not well structured and the management concepts and solutions used are not appropriate. Often researchers spent little time on data management, because they are under pressure to publishing their results as quickly as possible. Unfortunately, long-term handling of research data has very low priority. Once an article has been published, the underlying data are no longer in focus. On the other hand, large portions of the raw data and the accompanying results are not being published for various reasons. As a result, errors and gaps in primary research data are often detected late, if ever. This seriously hinders reuse of data.

**How it is done today:** Today, biodiversity and ecological research is often organised in larger projects which have a dedicated data manager. These experts try to ensure high data quality by (mostly manually) checking data produced by the scientists in their projects. Tools supporting data quality management start to appear but are as of yet poorly integrated into the data collection tools.

**Challenges addressed by GFBio:** It is essential to offer flexible and user-friendly data management systems to single scientists and to larger research groups already at the beginning of their research activities. This will allow them to manage and quality-control their research data. Already at that time, the researchers have to be convinced that a maximised high quality data pool is relevant for their own research, now and in the future. The possibility for an early publication of high quality data with adequate citation mechanism (see below) might encourage the researchers to provide their data to the scientific community after the end of their research studies. Our goal with GFBio is to provide appropriate mechanisms so data is made public as early as possible. Also, GFBio aims to promote data collection and management tools that support and integrate data quality control.

**Data Deposition**

**What it is about:** Once data has been collected and the first level of management and quality assurance is done by the single researcher, it needs to be managed in a larger project context. Typically, this first management level is done for the duration of a project.

**How it is done today:** Traditionally, data management within small and medium scale projects is achieved by providing tailored common data management systems or simply by file sharing among partners. During the last years, many collaborative projects have developed their own data deposition or data management systems, including [LNB+12, NRB+13]. In addition to these project-specific tools, several flexible tools have been developed, which are highly configurable, reusable and can be adapted to the specific needs of research projects. Diversity Workbench[4] and BExIS 2[5] are examples for such multipurpose data management working environments. Both tools are currently used by a large number of organizations and consortia and provide an important basis for mid-term and long-term data management.

**Challenges addressed by GFBio:** Whether data is stored in a project-specific platform

[4] Diversity Workbench, `http://diversityworkbench.net` [last access 26.06.2014]

[5] BExIS++, `http://fusion.cs.uni-jena.de/bexis` [last access 26.06.2014]

or in one of the more adaptable ones, the long-term fate of data managed locally or with these tools is often unclear and dependent on the capacities of the organization hosting them. Many large scale projects are using tools for their internal data workflow without a clear long term data hosting strategy beyond their project lifetime. Therefore, a large amount of data needs to be preserved and transferred to a suitable long term data holding infrastructure and possibly to a data management infrastructure maintained by a scientific data center or a long-time institutional data repository. Strong support for the development of standard conform and advanced software for long-time data management and deposition as well as the establishing of content standards, management plans and archiving concepts is therefore urgently needed. Within GFBio we will extend BExIS and the Diversity Workbench to meet these requirements and to serve as blueprints for other tools.

**Data Publication**

**What it is about:** There is a cultural change in the research community promoting the idea of data publishing as a clear incentive for scientists to share their data. Only in the past few years have scientists began calling for data "citation" and referring to data "publication" rather than data "sharing" and "availability" [CMG+13].

**How it is done today:** Data publication can be similar to the conventional publication of articles in journals that includes online submission, quality checks, peer-review, editorial decisions, and an equivalent of page proofs. In fact, data storage in central databases is getting increasingly important or is yet mandatory for the acceptance of peer reviewed publications in specific fields of biodiversity research as e.g. molecular sciences or ecology [Whi11]. PANGAEA has built up an extensive collaboration with science publishers and implemented services to link articles and data. The impact on citation rates could be shown in a recent bibliometric study on science articles having supplementary data [Sea11]. Eventually an index for science data would be needed similar to Data Usage Index (DUI) [IC11]. Data publishing is also strongly supported by GBIF and the ICSU World Data System (WDS).

**Challenges addressed by GFBio:** One major challenge is to make data publication and access to published data as easy as possible for scientists. This requires a strong integration of data publication in data management tools which should support a "one click" publication. In addition to fostering such tools, GFBio's helpdesk will provide a wide range of information on options for data publication.

**Data Preservation**

**What it is about:** Much of biodiversity data is hard or impossible to reproduce. For instance, it is impossible to repeat an observation of the species richness in a certain area for a point of time in the past. It is thus extremely important that data is preserved beyond the end of projects. Since only a small percentage of data end up being part of a scientific article or a data publication, other means to ensure long-term availability of data are needed.

**How it is done today:** Globally, the backbone of long-term archives for environmentally related biological data is mostly restricted to specific disciplines, data types, and/or

regional coverage. Prominent are the International Nucleotide Sequence Database Collaboration (INSDC) with their archives. Other networked resources or platforms include the Global Biodiversity Information Facility (GBIF), Dryad and some others devoted to certain data domains [THR12]. Prominent global infrastructures for environmental data include the WMO and ICSU World Data Centers. Besides there are various national data centers, e. g. the Australian National Data Service (ANDS) and the US DataOne initiative.

In Germany, a number of data centers and archives are available to researchers: the University of Bremen jointly with the AWI operates PANGAEA, a data archive specialized on environmental data. Founded in 1992 PANGAEA has demonstrated its long term perspective by a certification of ICSU World Data System and was accredited by the WMO as Data Collection and Processing Center (DCPC). Currently, PANGAEA hosts more than 450.000 data sets comprising of more than 6 billion individual measurements collected during approx. 160 EU, international and national projects. Currently three main databases for molecular biodiversity analysis exist in Germany. They are mainly used for the treatment of sequences from microorganisms, which make up the majority of currently available molecular data. The SILVA rRNA database[6] provides quality-controlled sequence data sets for molecular biodiversity analysis. The Megx.net portal[7] integrates microbial genomic and environmental data based on georeferencing. Both systems are hosted and maintained at the Max Planck Institute for Marine Microbiology and Jacobs University Bremen. The Bacterial Diversity Metadatabase "BacDive"[8] is a highly curated, comprehensive data resource for all cultured Bacteria and Archaea in Biologial Resource Collections (BRC), currently under development by the German Collection of Microorganisms and Cell Cultures (DSMZ) in Braunschweig. The German Natural History Collections, organised in the Consortium of the Deutsche Naturwissenschaftliche Forschungssammlungen (DNFS) guarantee sustainability in hosting more than 140 million of life science objects together with adjacent information, mainly occurrence data of taxa in space and time, as well as all kinds of additional ecological data. The DNFS institutions also manage enormous amounts of biodiversity- and organism-related information, e.g., concerning DNA samples and data, monitoring data, taxon- and object-related metadata and multimedia data which also include (meta-) data from past and ongoing third-party-funded research projects. Promising progress has been achieved regarding standardization and networking among these national history collections which led to harmonized access to these archives with more than 10 million data records within e.g. GBIF and BioCASE.

**Challenges addressed by GFBio:** For individual projects, resources and funding are limited. That is one reason why – despite a growing number of projects having produced large volumes of data during the last years – only part of this data is stored in data centers such as the ones mentioned above and generally available for reuse. A lot of data is lost. Funding agencies as well as research organizations put increasing pressure on the research community and demand long term storage and open access on research data. The majority of researchers are generally willing to archive and share their data. However, to get acceptance by the researchers, the data archives and scientific data centers have

[6] Silva: High Quality Ribosomal RNA Databases, `http://www.arb-silva.de` [last access 26.06.2014]

[7] Megx - Marine Ecological Genomics, `http://www.megx.net` [last access 26.06.2014]

[8] BacDive - The Bacterial Diversity Metadatabase, `http://www.bacdive.dsmz.de` [last access 26.06.2014]

to provide appropriate functionalities and prove their trustworthiness as well as a long term perspective of operations. It is important for them to comply with international standards, technical formats and guidelines for long term archives. In Germany, only few data archives/data centers operate in a sustainable and organized way, especially in the fields of environmental sciences and natural history collections and museums. Germany's situation is essentially the same as the global one, where existing infrastructures are insufficient to cope with heterogeneous data from multidisciplinary projects, leading to the existence of numerous isolated, technically proprietary and unsustainable databases. A coordinated approach is needed to improve the current situation and expand the infrastructure of the archives and data centers. Such an approach is established in GFBio.

**Discovery and Visualisation**

**What it is about:** For biological sciences a major drawback is the lack of interconnections between the classical biodiversity knowledge resources such as natural history, environmental and molecular data providers as well as literature. Data generated in integrative biodiversity and ecosystem research are extremely heterogeneous, multidisciplinary, and collected by many different scientific disciplines using a large range of methods, technologies, and data standards. On the other hand, biodiversity research critically depends on a wide range of scientific disciplines providing the fundament for the understanding of organisms and their interactions with their environment in time and space. The lack of coordination between standards still hinders true interoperability which makes it impossible to enable unified mechanisms for discovery and visualisation across all relevant resources.

**How it is done today:** Harmonization and mapping of domain specific metadata standards as well as the integration of community profiles has to occur before interdisciplinary archives can successfully be realized. Several approaches have been initiated in the past e.g. by TDWG groups to integrate geospatial and biodiversity data or GEO-BON to integrate data resources from different domains. Domain specific approaches include the definition of complex namespaces and schemas such as the Diversity Workbench (DWB) or EDIT Common Data Model (CDM) suitable to be applied as a (maximum) standard for mapping biodiversity and ecological data. On the European level the INSPIRE initiative fosters the ISO19139 schema implementation. Efforts have been made to e.g. map between ISO19139 and other schemas such as Dublin Core or the GBIF metadata profile and to map genome and geographical information. In recent years, a number of initiatives started out to implement semantic services providing stable terminologies and ontologies supporting specific user demands in biodiversity research such as the "TDWG-Ontology" and the GBIF multilingual vocabulary server[9].

**Challenges addressed by GFBio:** Despite the progress made over the last decade, there is a general lack of formally represented knowledge about the relation between terms referring to similar entities used in different scientific communities. As a consequence, computer-based biodiversity studies often fail to exploit the potential of existing and highly valuable data resources ranging from data stemming from individual scientific projects to large scale data intensive biodiversity and environmental assessments. GfBio addresses

[9]TDWG LSID Vocabularies, `http://rs.tdwg.org/ontology/voc/Collection`, GBIF Vocabularies, `http://vocabularies.gbif.org/` [last access 26.06.2014]

this gap by developing a terminology server and services offering unified interfaces to existing and presently in-accessible vocabularies used in different relevant scientific disciplines. The terminology server will also provide a platform for ontology development with a special focus on collaborative aspects and consensus finding as well as maintenance processes in continuously evolving scenarios. In a community-based approach, GfBio works closely with experts in the different sub-disciplines providing prototypic domain-vocabularies and use-cases for the future terminology services.

**Data Integration and Analysis**

**What it is about:** Data integration provides means to access various data sources in a uniform manner to extract information that is implicitly hidden within their union. For biodiversity research, it is expected that the value of this implicit knowledge is large.

**How it is done today:** Today, the individual disciplines of biodiversity research collect large data collections, let it be by measurements or by digitalization of existing material. While these collections are large and technically advanced, they remain isolated and means for integrated access and analysis lack. Thus, today, data integration in biodiversity remains a largely manual task requiring considerable effort. An example for such an integration effort is the TRY database [KDL+11]. This is the largest compilation of plant trait data integrating data from more than a hundred contributors. Despite the clear thematic focus, integrating this data requires tremendous manual effort. The situation is even worse when trying to integrate data across disciplines. This is a clear hindrance to answering key questions of biodiversity research.

**Challenges addressed by GFBio:** Making archives interoperable and enriching data with (ontology-based) metadata is of little use, if there is no common interface that makes it easy for users to access the data. This means that sophisticated search interfaces are needed that allow searching in data and metadata, support browsing and faceted search and hide syntactic differences between the different data. The continuously increasing volume and complexity of biodiversity data this becomes more and more challenging. On the one hand, efficient harvesting and indexing techniques that scale with growing data volume are needed, on the other hand, new interfaces for navigating across search results are needed, including visualizations that allow a quick overview over potentially huge and complex search results. Community involvement is a proven measure to improve the overall quality of data services. Crowdsourcing has been tested in a variety of web pages and e.g. categorization and prioritization of content (so-called folksonomy) has a good potential to improve the find-ability of research data. Further, personalization can help to recommend contents and enable users to return to and build on earlier search results. Highly heterogeneous data in differing formats needs innovative integrative approaches such as text mining and indexing including Semantic Web technologies and Linked Data concepts using RDF and SPARQL. Such innovative concepts need to be further elaborated and adapted to the specific needs of the scientific community. Many of these new approaches have previously been evaluated for biodiversity research, however the adaptation of these technologies in a highly interdisciplinary data environment still is a challenge.

Data-intensive science [TTH11] is building on the integration of data from various sources. The real challenge in interdisciplinary data management is finding a common approach to

exchange data not only metadata. Only little progress has been made to agree on interdisciplinary content standards and corresponding exchange protocols for data. Even within one scientific community multiple, incompatible formats are used for machine-machine interaction such as NetCDF or OGC OM format. In addition there are a large number of web services offering proprietary formats. Emerging technologies such as grid or cloud technologies have made promising progress with regards to data exchange, storage, scalability and performance. But no successful attempt has been made to apply these technologies to integrate the complex and interdisciplinary data scenario related to biodiversity. However such concepts are a prerequisite for scalable handling of large data volumes and necessary to provide innovative commonly usable tools and services to analyze such data. GFBio will therefore investigate new ways to embed these technologies within the given data landscape in order to provide such added value services for the integration and analysis of heterogeneous environmentally related biological data.

## 4 GFBio – A proposal for an integrated platform

The need to address the challenges outlined above was recognized several years ago by the senate commission on biodiversity research of DFG. As a result, it encouraged a preliminary project to further investigate the conditions, in particular the willingness to share data and the availability of an appropriate infrastructure in Germany. Results of this project were published in [ETB+12, BSE+12] and showed clear need for action. This resulted in an initiative to build a national infrastructure for data management in biological and ecological research which brought together stakeholders from all involved communities, namely archives, national history collections, biodiversity research and computer science. This group of 21 principal investigators obtained funding by DFG to set up a ”German Federation for the Curation of Biological Data (GFBio)”.

Figure 2 provides an high-level overview of the GFBio components. GFBio will build on international standards and proven data archiving infrastructures and workflows such as the PANGAEA long term archive for environmental data, the resources of networked natural history collection data repositories and genome data archives. Interoperability among these resources, the integration of domain specific data bases as well as external, international data providers and a cooperation with relevant publishers are a key focus of the project. GFBio will follow innovative approaches to stimulate data submission among the scientific community. In particular it will pursue the promising efforts of data publication and cross-linking of research data with the scientific literature. Project data management and submission will further be supported by integrating and further developing innovative tools and platforms built on existing efforts such as the Biodiversity Exploratories Information System (BExIS) and Diversity Workbench (DWB).

Via the portal, we aim to establish a central point of contact for researchers and the interested public with respect to environmentally related biological data. The portal will offer the possibility to search across the connected data sources and will be closely integrated with visualization and analytical tools which will enable efficient data aggregation and analysis, including links between formerly non-connected data sets from different sub-

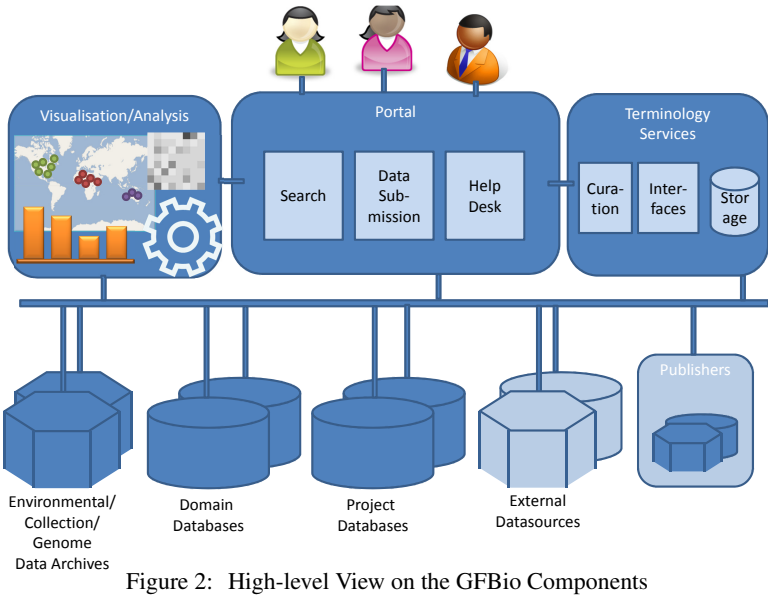


Figure 2: High-level View on the GFBio Components

disciplines. In addition, the portal will support data submission from project or domain databases to the underlying archives. Based on commonly defined workflows and procedures, the virtual helpdesk will address digital data curation issues as early as possible within the creation of research data and will provide researchers with guidance on how to do data management and what tools to use. The Terminology Services provide a semantic backbone supporting high quality data acquisition and data integration.

GFBio will thus provide an infrastructure that addresses not only the individual challenges identified above, but the data life cycle as a whole. We are convinced that only such an integrated solution with strong links to international initiatives has the potential to provide efficient, sustainable support for biodiversity (or any other) research. With the specific challenge to integrate data and metadata from several domains GFBio will also push relevant standards and innovative solutions. Special emphasis will be given to governance and organizational issues of the planned infrastructure. Taking into account the working conditions for every institution involved, a common organizational structure with common services and sustainable business models will be developed.

## 5 Summary and Conclusion

While research data management has gained a lot of interest over the last few years and initial building blocks for a solution exist, many challenges remain. In particular, up to now, integrated solutions that provide users with a ”one-stop” platform for all their needs

with respect to data management across all phases of the data life cycle are lacking. For Germany, GFBio aims at providing such a platform.

**Acknowledgements:** We thank all GFBio PIs (in addition to those co-authoring the paper these are: François Buscot, Johanna Eder, Stephan Frickenhaus, Christoph Häuser, Thomas Hickler, Jens Kattge, Heike Neuroth, Jörg Overmann, Jens Schumacher, Johann Wägele, Christian Wirth, Ramin Yahyapour) and everyone else involved in GFBio for their contributions. The work described here is funded by DFG.

## References

[BSE+12] Kerstin Bach, Daniel Schäfer, Neela Enke, Bernhard Seeger, Birgit Gemeinholzer, and Jörg Bendix. A comparative evaluation of technical solutions for long-term data repositories in integrative biodiversity research. *Ecological Informatics*, 11:16–24, 2012.

[CMG+13] Mark J. Costello, William K. Michener, Mark Gahegan, Zhi-Qiang Zhang, and Philip E. Bourne. Biodiversity data should be published, cited, and peer reviewed. *Trends in Ecology & Evolution*, 28(8):454–461, August 2013.

[ETB+12] Neela Enke, Anne Thessen, Kerstin Bach, Jörg Bendix, Bernhard Seeger, and Birgit Gemeinholzer. The user's view on biodiversity data sharing – Investigating facts of acceptance and requirements to realize a sustainable use of research data. *Ecological Informatics*, 11:25–33, 2012.

[IC11] Peter Ingwersen and Vishwas Chavan. Indicators for the Data Usage Index (DUI): an incentive for publishing primary biodiversity data through global information infrastructure. *BMC bioinformatics*, 12(Suppl 15):S3, 2011.

[JKN+09] Stefan Jablonski, Alexandra Kehl, Dieter Neubacher, Peter Poschlod, Gerhard Rambold, Tobias Schneider, Dagmar Triebel, Bernhard Volz, and Markus Weiss. DiversityMobile – Mobile Data Retrieval Platform for Biodiversity Research Projects. In *GI Jahrestagung*, pages 610–624, 2009.

[KDL+11] Jens Kattge, Sandra Diaz, Sandra Lavorel, IC Prentice, P Leadley, G Bönisch, Eric Garnier, Mark Westoby, Peter B Reich, IJ Wright, et al. TRY–a global database of plant traits. *Global change biology*, 17(9):2905–2935, 2011.

[LNB+12] Thomas Lotz, Jens Nieschulze, Jörg Bendix, Maik Dobbermann, and Birgitta König-Ries. Diverse or uniform? – Intercomparison of two major German project databases for interdisciplinary collaborative functional biodiversity research. *Ecological Informatics*, 8:10 – 19, 2012.

[MTA+11] Camilo Mora, Derek P Tittensor, Sina Adl, Alastair GB Simpson, and Boris Worm. How many species are there on Earth and in the ocean? *PLoS biology*, 9(8):e1001127, 2011.

[NRB+13] Karin Nadrowski, Sophia Ratcliffe, Gerhard Bönisch, Helge Bruelheide, Jens Kattge, Xiaojuan Liu, Lutz Maicher, Xiangcheng Mi, Michael Prilop, Daniel Seifarth, et al. Harmonizing, annotating and sharing data in biodiversity-ecosystem functioning research. *Methods in Ecology and Evolution*, 4(2):201–205, 2013.

[Sea11] JR Sears. Data sharing effect on article citation rate in paleoceanography. In *AGU Fall Meeting Abstracts*, volume 1, page 1628, 2011.

[THR12] Dagmar Triebel, Gregor Hagedorn, and Gerhard Rambold. An appraisal of megascience platforms for biodiversity information. *MycoKeys*, 5:45–63, 2012.

[TTH11] Kristin M Tolle, D Tansley, and Anthony JG Hey. The Fourth Paradigm: Data-Intensive Scientific Discovery [Point of View]. *Proceedings of the IEEE*, 99(8):1334–1337, 2011.

[Whi11] Michael C. Whitlock. Data archiving in ecology and evolution: best practices. *Trends in Ecology & Evolution*, 26(2):61 – 65, 2011.