Improved access to arachnological data for ecological research through the ARAMOB data repository, supported by Diversity Workbench and NFDI data pipelines

Alexander Bach, Martina Roß-Nickoll, Joachim Holstein, Richard Ottermanns, Florian Raub, Dagmar Triebel, Markus Weiss, Ingo Wendt & Hubert Höfer

doi: 10.30963/aramit6609

Abstract. The scientific community has been developing and refining digital data standards to ensure that biodiversity data can be easily exchanged between different databases, systems and institutions. However, scientists still face the challenge of effectively analysing this vast amount of data. Variations in the quality, documentation and availability of metadata often make it difficult for scientists to compile appropriate datasets for their research. One contribution towards this task is the research data repository ARAMOB of the Arachnologische Gesellschaft (AraGes), which focuses on systematically collected data on spider assemblages. Mandatory requirements have been developed to ensure the quality and utility of the data for ecological research during a given project. A next step towards enhancing the data basis for convincing analyses of spider assemblages in Central Europe is the offer to now publish data in the society's open access journal Arachnologische Mitteilungen/Arachnology Letters, which are integrated into the data repository and thus made effectively accessible and usable. These data papers will be presented as one printed page in the journal, accessible on the website of the AraGes and from the BioOne Digital Library, accompanied by a PDF-document containing metadata to effectively use the published data. The original dataset is published as spreadsheet tables, but also deposited in the ARAMOB data repository, which is managed with the modularized database software and virtual research environment Diversity Workbench. By this means, the published data packages are also accessible and analysable within a wider context through the ARAMOB web portal. On request, scientists can also exploit data with the free and well-documented Diversity Workbench software tools. The data pipelines involved are defined in the context of the National Research Data Infrastructure (NFDI).

Keywords: arachnology, data collection, data publication, ecology, research environment

Zusammenfassung. Verbesserter Zugang zu arachnologischen Daten für ökologische Forschung über das von Diversity Workbench und NFDI Datenpipelines unterstützte ARAMOB Datenrepositorium. Mit Beginn des digitalen Zeitalters haben Wissenschaftler Datenstandards entwickelt und verfeinert, um sicherzustellen, dass Biodiversitätsdaten zwischen verschiedenen Datenbanken, Systemen und Institutionen ausgetauscht werden können. Es stellt jedoch immer noch eine Herausforderung dar, umfangreiche Datenmengen aus verschiedenen Quellen aufzunehmen und für effektive Analysen in Datenspeicher zu übertragen. Unterschiede in der Qualität und Dokumentation (Verfügbarkeit von Metadaten) erschweren es Wissenschaftlern häufig, geeignete Datensätze für ihre Auswertungen zusammenzustellen. Einen Beitrag zu dieser Aufgabe leistet das Forschungsdatenrepositorium ARAMOB der Arachnologischen Gesellschaft (AraGes), das sich auf systematisch erhobene Daten zu Spinnenzönosen konzentriert. Verbindliche Anforderungen zur Sicherstellung von Qualität und Nutzbarkeit der Daten für ökologische Forschung wurden in einem vorausgegangenen Projekt formuliert. Ein nächster Schritt, um die Datenbasis für fundierte Analysen von Spinnenzönosen in Mitteleuropa zu verbessern, ist nun das Angebot der Publikation von Daten in der frei zugänglichen Vereinszeitschrift Arachnologische Mitteilungen/Arachnology Letters, die in das Datenrepositorium integriert und damit effektiv zugänglich und nutzbar gemacht werden. Die als data paper veröffentlichten Daten werden auf einer Druckseite der Zeitschrift präsentiert, die auf den AraGes Webseiten und in der BioOne Digital Library öffentlich zugänglich ist. Dieser Artikel wird von einem zweiten PDF-Dokument, das die Metadaten zur effektiven Nutzung der veröffentlichten Daten enthält, sowie den Originaldaten in Tabellenform ergänzt. Die Daten werden im ARAMOB Datenrepositorium hinterlegt, das mit der modularen Datenbanksoftware und virtuellen Forschungsumgebung Diversity Workbench gemanagt wird. Über das ARAMOB-Webportal sind die veröffentlichten Datenpakete somit auch in einem breiteren Kontext zugänglich und analysierbar. Mitglieder der AraGes können darüber hinaus Daten mit den frei verfügbaren und gut dokumentierten Diversity Workbench Software-Tools erschließen. Die beteiligten Datenpipelines sind im Kontext der Initiative der Nationalen Forschungsdateninfrastruktur (NFDI) definiert.

Biodiversity, the wealth of life on Earth, is crucial for the functioning of ecosystems and the provision of services that sustain human well-being. But as human activities continue to threaten biodiversity, it is more important than ever to understand and monitor this valuable resource. One key tool in this effort is the sharing and gathering of data to provide insights into the distribution, ecology and characteristics of different species and ecosystems. Since ecological patterns can only be identified through the aggregation of many individual observations (Underwood et al. 2000) sampled data should be stored systematically and well documented in easily accessible, centralized data repositories to gain a comprehensive understanding of our planet's biodiversity (von Wettberg & Khoury 2022). Therefore, Orr et al. (2022) identified the interoperability of biodiversity data as one of the bottlenecks of the Post-2020 Global Biodiversity Framework which is still under development as a tool of the Convention of Biological Diversity (CBD) to significantly lower biodiversity losses by 2050. To take this into account the DFG (German Research Foundation), one of the largest national research funding bodies in Europe, has now set out extensive requirements on the handling of research data when approving research proposals (DFG 2022).

In general, two different types of scientific database concepts can be distinguished (adapted from Porter 2000) for the storage of biodiversity data. "Wide" databases collecting as much data as possible from the relevant scientific field and

Alexander BACH, Richard OTTERMANNS, Martina ROSS-NICKOLL, Institute for Environmental Research, RWTH Aachen University, Worringerweg 1, 52074 Aachen, Germany; E-mails: alexander.bach@bio5.rwth-aachen.de, https://orcid.org/0009-0001-3585-3148, ottermanns@ifer.rwth-aachen.de, https://orcid.org/0000-0002-2168-9455, ross@bio5.rwth-aachen.de, https://orcid.org/0000-0002-6425-1728
Hubert HÖFER, Florian RAUB, Staatliches Museum für Naturkunde Karlsruhe, Erbprinzenstr. 13, 76133 Karlsruhe, Germany; E-mails and ORCIDs: hubert.hoefer@smnk.de, https://orcid.org/0000-0003-3962-151X, florian.raub@smnk.de, https://orcid.org/0009-0004-4226-2698
Joachim HOLSTEIN, Ingo WENDT, Staatliches Museum für Naturkunde Stuttgart, Rosenstein 1, 70191 Stuttgart, Germany; E-mails: joachim.holstein@smns-bw.de, https://orcid.org/0000-0002-1541-7170, ingo.wendt@smns-bw.de, https://orcid.org/0000-0002-8367-4048
Dagmar TRIEBEL, Markus WEISS, Staatliche Naturwissenschaftliche Sammlungen Bayerns, IT Center, Menzinger Straße 67, 80638 München; E-mails and ORCIDs: triebel@snsb.de, https://orcid.org/0000-0003-1980-3148, weiss@snsb.de

Academic editor: Tobias Bauer

submitted 10.11.2023, accepted 23.12.2023, online 29.12.2023

"deep" databases that specialize in certain types of data. An example of a "wide" database in the biodiversity context is the GBIF database infrastructure. From primitive unicellular organisms to slime moulds, invertebrates, and vertebrates to the fossil *Tyrannosaurus rex*, any occurrence record or checklist or sampling event dataset of any species that has ever lived on our planet can be stored there (Telenius 2011, see also data types in the GBIF Integrated Publishing Toolkit (IPT) User Manual). This concept of a mega-diverse data aggregator leads to an enormous heterogeneity of the spectrum of available datasets. Another limitation is the data quality, i.e. the heterogeneous use of the standard information elements, e.g. within Darwin Core and ABCD (Access to Biological Collection Data), within and across the single datasets, which is regularly criticized in studies (Beck et al. 2013, Ferro & Flick 2015, Maldonado et al. 2015, Yesson et al. 2007, Zizka et al. 2020). Therefore, a more complex analysis of the data is not easy and requires careful data selection and several data cleaning and data preparation steps, which can be time-consuming.

"Deep" databases focus on a specific type of data, in the biodiversity context, for example occurrence or sampling-event data restricted to specific taxonomic groups, single habitats or geographic regions or traits of specific taxa. Examples are Edaphobase (Burkhardt et al. 2014) focusing on soil zoological data, Critterbase (Teschke et al. 2022) for benthic taxa, GlobalAnts for ant trait data (Parr et al. 2017) or the Azorean Biodiversity Portal (Borges et al. 2010). Since the datasets here are limited to taxonomic groups, habitats or regions, the data naturally exhibit an increased homogeneity compared to data in wide databases. Also improved are data quality and data validation, as the total data flow is considerably smaller and therefore much more controllable.

The ARAMOB data repository

Here we present the database or data repository ARAMOB of the "deep" kind, restricted to spiders (Araneae) as a taxonomic group and focussing on Central Europe as the geographic region (although so far with a very strong predominance of data from Germany) and systematically sampled community data (spider assemblages). Spiders are an appropriate model group for this kind of database project, as they have several advantages. First, arachnologists are well-organized worldwide. There is the International Society of Arachnology (ISA) and many continent and country-related societies, e.g. the European Society of Arachnology (ESA) and the German language-based Arachnologische Gesellschaft. Second, there is a strong and commonly agreed taxonomic backbone – the online World Spider Catalog (2023). Taxonomy and faunistics of this group have been extensively studied in Central Europe, resulting in a plethora of online resources available to the arachnological community, e.g. araneae Spiders of Europe (Nentwig et al. 2023) and the Picture Gallery of Belgian and French spiders (Oger 2023) for identification, the Atlas of the European Arachnids (Arachnologische Gesellschaft 2023a) for occurrence data and the World Spider Trait database (Pekár et al. 2021). However, a systematic collection of existing sampling-event data on spiders was so far lacking, despite a large known quantity of dark data, which is mainly opportunistically collected data (Bowser 1986, Michener 2015) from small-scale studies, not available after publication of the results (Heidorn 2008, Reichman et al. 2011). Dark data are maintaining the situation of severe impediments in invertebrate conservation (Cardoso et al. 2011).

The ARAMOB data repository is one outcome of the DFG-funded project "Semantic enrichment and mobilization of data in distributed repositories for taxonomy and ecology of spiders" (2016 - 2019, grant number: 316372061). The main objective of this project was to create a curated data repository for spider assemblage (i.e. sampling-event based) data that can be used by researchers to analyse the ecology, habitat preferences and temporal changes in population sizes of these organisms. To this purpose, research groups from Staatliches Museum für Naturkunde Stuttgart (SMNS), the IT Center of Staatliche Naturwissenschaftliche Sammlungen Bayerns (SNSB) and the Institute for Environmental Research (IFER) of RWTH Aachen University, under the leadership of Staatliches Museum für Naturkunde Karlsruhe (SMNK), have joined forces. The SNSB and SMNS are established GFBio Data Centers (see Web links and abbreviations) with a focus on biodiversity data (Weibulat et al. 2023).

The volume of data that could be mobilized and made accessible to date with the help of workflows developed in the project (and reported elsewhere) shows the mass of biodiversity data that is still waiting to be tapped. The ARAMOB project, in collaboration with its partners, has likely amassed one of the largest and most easily accessible collections of systematically collected spider datasets in Central Europe, with community data available from more than 1170 sites with over 450.000 individuals from 660 species. This is particularly noteworthy given that only three institutions contributed their previously immobilized data to the ARAMOB data repository. Other notable collections of similar sizes are the dataset of Hänggi et al. (1995), which is not publicly accessible, and the Atlas of the European Arachnids (Arachnologische Gesellschaft 2023a) which incorporates mainly occurrence data for understanding spider distribution and is already frequently used by scientists (Narimanov et al. 2021, Purgat et al. 2021, Štokmane & Cera 2018, Wersebeckmann et al. 2021, Wiśniewski et al. 2018). Most of the atlas data were not systematically collected, posing challenges for their application in addressing ecological research questions requiring quantitatively comparable datasets. Datasets originating from non-systematic and non-standardized sampling approaches are often deposited in natural history museums, such as unsystematically collected collection specimens or observation data from citizen science portals like inaturalist.org (iNaturalist 2023). For niche modelling, biogeographic, conservation or taxonomical research, this kind of data still represents immeasurable value (Casas-Marce et al. 2012, Rowe 2005, Sillero et al. 2021). The SMNK and SMNS regularly deliver such data to the Atlas, or deliver the data in a machine-readable format to GBIF (see Web links and abbreviations).

ARAMOB was intended to be a supplement to the existing Atlas and GBIF by providing high-quality data packages from studies under a specific research hypothesis or question that can be used in meta-studies. Currently, data exchange between the two arachnological data collections (Atlas, ARAMOB) is done manually, but plans to automate this process in the future are being considered. ARAMOB is focused on *opportunistic* biodiversity data (Bowser 1986, Michener 2015). This is usually a self-contained dataset that has been

collected over a short or long period of time, to address a specific research hypothesis or question. Although these data have well-documented information on sampling sites and protocols, environmental variables, and other relevant details, like their collection and compilation are influenced by the scientific question being addressed. Opportunistic datasets are often generated in small-scale research projects that are completed within a defined timeframe, typically three to six years. Heidorn (2008) outlined the usual manpower behind this kind of projects as follows: "… with one lead researcher with part-time commitment to the project and perhaps two or three graduate students or part-time staff scientists". The collection, processing and evaluation of the data is also performed with the same care as in long-term studies. However, various reasons lead to the situation that the data from these kinds of projects are often not accessible at the end. One of the main reasons can be found in the reward system of the academic environment. Scientists get recognition and funding when they publish findings derived from their raw data in highly condensed scientific articles. Making the raw data publicly available does not initially benefit the responsible scientist. This is where data papers come into focus.

To our knowledge ecological data on spiders are collected regularly and in large quantities (not only in Germany), but there is a pressing need for action to ensure that these collected datasets are not left unused or stored locally after the conclusion of research projects and theses, published or not. Only biodiversity data published and stored in standardized formats will be usable for science in the future. And it is crucial to store such data in a structured manner that retains its complexity and critical metadata for cross-study analysis, while also making it easily accessible to researchers and decision makers. As a case study, the research data repository ARAMOB, which focuses on systematically collected spider assemblage data, can be utilized to evaluate its advantages, challenges, and potential applications. Members of the Arachnological Society, and all other arachnologically interested scientists, may carry out such work.

Which software is used by ARAMOB?

Various types of database management systems exist, with relational databases historically holding a predominant position as the most extensively utilized (Porter 2018). In Germany, the German Federation for Biological Data (GFBio e. V.) recommends two systems, Diversity Workbench (Triebel et al. 1999) and BEXIS2 (Chamanara et al. 2021), for the management of biodiversity data (Diepenbroek et al. 2014, Lotz et al. 2012, Karam et al. 2016). Diversity Workbench (DWB) is a modular, free (GPL 3 licence) virtual research environment aimed at managing and analyzing bio- and geodiversity data and was started in 1999 at the SNSB IT Center and has been continuously developed ever since (Triebel et al. 1999). It supports a variety of data types including Taxon data, Biodiversity and Occurrence data, Environmental Biological and Ecological Data, Non-Molecular Analysis Data and Molecular Sequence Data, and allows users to customize data entry forms and the database structure to suit their specific needs. To cope with the possible volume and large heterogeneity of biodiversity data, there are different modules within the DWB framework depending on the managed type of information (Triebel et al. 2007).

The utilization of Diversity Workbench has the advantage of facilitating standardization of data through the specification of terminologies, ontologies and taxonomies at an early stage of the data life cycle (Recknagel & Michener 2018), thus promoting the creation of FAIR data (Harjes et al. 2020, Karam et al. 2016, Schneider et al. 2019). This data can subsequently be published via a BioCASe (see Web links and abbreviations) data pipeline as an output of the DiversityCollection module, based on the ABCD (Access to Biological Collection Data) standard schema, and accessed via GBIF and NFDI for increased FAIRness (see Harjes et al. 2020, Novotný et al. 2022, Weiss et al. 2018). Modern data standards such as Darwin Core and ABCD have emerged to further standardize and improve the management and sharing of biodiversity data (Holetschek et al. 2012, Wieczorek et al. 2012). Both standards provide a common framework for describing and sharing primary biodiversity data, allowing for greater interoperability and accessibility of data across different platforms and systems. Given the significance of data security and the requirement of Microsoft licences, the operating environment of the Diversity Workbench server incorporates PostgreSQL cache databases. Before external release, data in the primary operating environment must be transferred to those cache databases.

Data quality in ARAMOB

To ensure the quality of the data in the ARAMOB database, only datasets sampled systematically and in a methodologically consistent manner can be included. In a first step, structures, templates, and workflows were developed to mobilize data from a wide variety of sources, e.g., publications, studies, project reports, or theses, and to transfer them into a consistent terminology. When possible, data was enriched with additional information, such as coordinates or habitat types, if not available beforehand. By minimizing the heterogeneity and inconsistency that can plague biodiversity databases, ARAMOB aims to provide a freely accessible 'clean' research environment that allows robust and reliable analyses with community datasets. In the further course of this work, specific methods, queries and algorithms were developed to automate the data extraction from the database and convert it into formats that are suitable for statistical analysis.

The concept of data quality may seem intuitive, but it encompasses numerous nuances that require closer examination. A dataset that comprises a species name, date, and coordinates, generated by a photo identification app, for instance, may have a high level of data quality (see RfII 2019), if all information is accurate and may be suitable for, e.g., modelling species distributions. Nonetheless, this type of dataset may not be usable for meta-analysis with systematically collected community data. Hence, data quality is primarily contingent upon the user's specific requirements for the analysis. According to Veiga et al. (2017), assessing biodiversity data quality involves using a set of quality metrics to determine its suitability for a specific use. However, even though data may have the same quality level, it may still have varying levels of suitability for different purposes. To determine the fitness of data for use, three crucial interrelated components must be considered: the intended use, the type of relevant data, and the criteria for determining 'fitness' for that data in the context of the use. In ARAMOB, the primary focus is on systematically

collected spider assemblage data obtained through standardized methods. Thus, when referring to data quality here, the emphasis is placed on not just the accuracy of the data, but also on determining the additional information necessary in community data to facilitate meta-analyses and enhance their validity.

How can research data become and remain sustainable in the long term?

One early initiative to support the needs of researchers and institutions in managing and sharing their biodiversity and environmental data was the German Federation for Biological Data (GFBio), which emerged from a DFG-funded project and later became an established association (GFBio e. V.) in 2016 (Diepenbroek 2015). The goal of GFBio e. V. is to provide a centralized infrastructure to help scientists managing and sharing biodiversity data along the data life cycle (Fig. 1).



Fig. 1: Idealized, schematic data life cycle. GFBio Training Materials: Data Life Cycle Fact-Sheet: Data Life Cycle: Plan. Retrieved 27 Jun 2022 from https://www.gfbio.org/training/material/data-life-cycle/plan/.

This included the development of a portal for accessing data, as well as tools for data management and analysis (Authmann et al. 2015, Diepenbroek et al. 2014). GFBio also serves as a facilitator for various data centers, providing a platform for researchers to submit their data for permanent storing and archiving. Later, in 2019 the National Research Data Infrastructure (NFDI) was founded with a broader scope and the goal of creating a sustainable infrastructure for research data management in Germany across all domains, not limited to biodiversity. It aims to create a network of research data infrastructures to support the storage, management, and preservation of research data in a variety of disciplines (Hartl et al. 2021). To date, the NFDI is split into 26 different consortia, each of which is focused on a specific research discipline. For biodiversity and environmental data, this is done by the NFDI4Biodiversity consortium, supported by 49 partner organizations from (citizen) science and government (Glöckner et al. 2020). This also includes GFBio e. V., whose services and tools will be integrated into the broader infrastructure being developed by NFDI4Biodiversity, providing researchers with a comprehensive set of data management tools and services (Weber et al. 2021).

NFDI4Biodiversity carries out a bottom-up approach, by involving different actors like research institutions and projects, natural science societies and government agencies in this process (Weber et al. 2021). One of the so-called Use Cases is no. 16 by the Arachnologische Gesellschaft e. V. (AraGes), SMNK and SNSB. In this way, the ARAMOB project continues to be integrated as a use case for data provisioning within the NFDI4Biodiversity consortium (Glöckner et al. 2020). This integration aims to enhance the interoperability of ARAMOB and Atlas data, enabling the seamless transfer of data between the NFDI4Biodiversity tools and services currently under development. In order to ensure data quality and long-term access of the ARAMOB database as a data provider at the SMNK, a partnership has been established between the SMNK and the AraGes with involvement of the SNSB and SMNS respectively as data centers. This agreement for distributed responsibility intends to ensure permanent data storage through the integration of the arachnological data stock into the technical infrastructure of the three museum institutions, realized as an autonomous Diversity Workbench (DWB) Data Repository ARAMOB, administrated and technically curated by two GFBio data centers: SMNS (data in DiversityCollection, DiversitySamplingPlots and DiversityProjects) and SNSB (freely available thesauri and name services, data in DiversityTaxonNames and DiversityScientificTerms as well as DiversityGazetters – via database server with open access). The scientific curation and the public data portal and website ARAMOB is under the responsibility of the SMNK. The set of explorative data analysis tools (see below) was developed and integrated by the working group at the RWTH Aachen university. The AraGes society provides expertise to aid in the collection, organization and quality assurance of ecologically relevant data for ARAMOB (Fig. 2).

Participation and benefits of ARAMOB for researchers

The main, and most significant, advantage for scientists lies in the provision of high-quality data packages dedicated to research purposes. By adhering to the aforementioned stringent quality criteria, these datasets become readily usable for meta- and cross-studies, enhancing the robustness and reliability of scientific investigations and results.

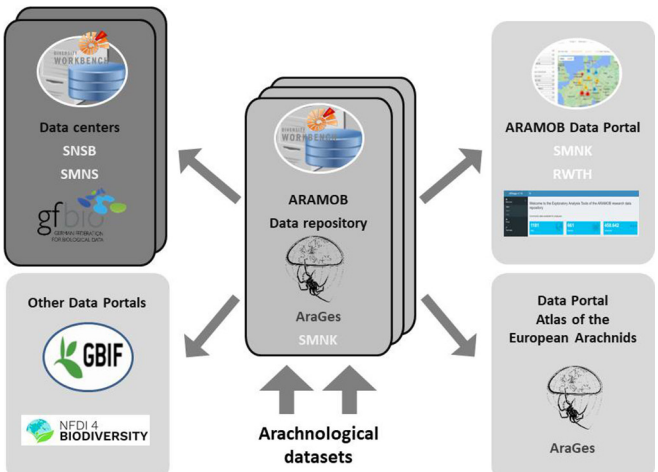


Fig. 2: Distributed responsibility for service, tool and data management, archival and publication of ARAMOB datasets.

Moreover, this initiative extends its benefits to (citizen) scientists, professionals from public authorities, and nature conservation organizations, empowering them to evaluate the dataset using a specially designed explorative tool package independently and interactively (Bach et al. in press). This sophisticated tool package enables comprehensive data analysis, encompassing various essential parameters. For instance, researchers can focus on study of phenology, ecological parameters (including the factors shading and humidity), habitat preferences, altitudinal distribution patterns, and encounter companion species of the diverse spider species documented within the database. By facilitating such thorough and customizable analyses, this comprehensive approach serves as a valuable asset in advancing our understanding of spider species diversity and ecological dynamics. Consequently, the implementation of this system fosters collaborative research efforts and empowers researchers and stakeholders alike in making informed decisions concerning biodiversity conservation.

There are currently two options for making your own data available through the ARAMOB data repository. Institutes may either become part of the ARAMOB repository through their own Diversity Workbench database or contribute their datasets through the SMNK. A new and attractive way to make data from systematic ecological studies (results published or not) available is the publication of a data paper in Arachnologische Mitteilungen/Arachnology Letters (Raub et al. 2023).

Access to the ARAMOB data repository

Data from the ARAMOB data collection can be accessed at https://aramob.de/en/data/data-exploitation. The data portal is bilingual (German, English) and includes information on the former project, background and networks of data management and – central – the part where the data can be accessed. Directions for use help users to effectively work with the filter and search functions and explore the resulting data collection. Search results are shown as a map of sampling localities, a study data list (including number and gender of the individuals of a species per sampling event, sampling dates, localities, habitat types, sampling method and the specimen-ID) or a species list. The lists can be exported as CSV files for individual data management and further analysis. A set of metadata of all project-based datasets included in the search results are offered to facilitate the selection of appropriate data for specific analytical approaches and interpretation of the results. As mentioned, the available data were quality-checked by us, however we cannot give a guarantee for correctness and accuracy. The responsibility for a data selection, the results and consequences of use have to be taken by the user. All the data presented on the ARAMOB website can be used under a Creative Commons 4.0 International Licence (CC BY 4.0). Downloaded/analysed data collections have to be cited as: "ARAMOB data accessed via https://aramob.de [Date of access]".

The complete dataset can also be analysed with a set of explorative data analysis tools (Bach et al. in press) accessible at https://aramob.de/en/data/statistic-tools/. Operating on a server, it removes the necessity for users to install or download extra software. Furthermore, members of the AraGes retain the ability to access the data within the Diversity Workbench virtual environment, which offers additional information such as methodological metadata and the full range of tools integrated in the complex software.

Concretely, for ambitious users, ARAMOB offers:

a) a read-only guest access to the DWB module DiversityTaxonNames, where the German spider species are listed, together with a set of Red Lists and traits and quick links to the selected species at the websites: World Spider Catalog (2023), araneae (Nentwig et al. 2023), Atlas and wiki of Arachnologische Gesellschaft (2023a, b);
b) a read-only guest access to the ARAMOB database in the DWB module DiversityCollection. Registered users can order a personal password by mail from the last author. Login is enabled via the freely available DiversityCollection client software and the full use of its features to search, explore, analyse the ARAMOB data.

Acknowledgements

We thank all people working in the background in all our teams, namely Anton Link, Dieter Neubacher, Stefan Seifert and Tanja Weibulat at the SNSB IT center, the spider specialists Tobias Bauer and Franziska Meyer, Thomas Stierhof and our volunteer Benjamin Riedel at SMNK, the active members of AraGes and supporters of our vision Michael Hohner (Atlas), Christoph Hörweg, Christoph Muster and the editors of Arachnologische Mitteilungen Tobias Bauer, Petr Dolejš and Konrad Wiśniewski. We are grateful to Aletta Bonn, Barbara Ebert and Thore Engel from the NFDI4Biodiversity consortium. We also thank Heiko Metzner, a dedicated arachnologist who always shared our vision and helped realizing the ARAMOB portal beyond his function as project manager of the media service provider psbrands. The ARAMOB research project was supported by the German Research Foundation (DFG: 316372061). The AraGes use case is supported within the project "Establishment of the National Research Data Infrastructure (NFDI)" in the consortium NFDI4Biodiversity (DFG: 442032008). Finally, we would like to express our sincere appreciation to Pedro Cardoso and Michael Hohner for their valuable contributions in reviewing and enhancing this article.

Web links and abbreviations

ABCD – Access to Biological Collection Data – https://abcd.tdwg.org (8. Nov. 2023)
AraGes – Arachnologische Gesellschaft/Arachnological Society – https://arages.de (8. Nov. 2023)
ARAMOB project and portal – https://aramob.de (8. Nov. 2023)
BioCASe – Provider software – https://www.biocase.org/products/provider_software/ (8. Nov. 2023)
BioOne Digital Library – https://complete.bioone.org/ (8. Nov. 2023)
Darwin Core – Darwin core standard – https://dwc.tdwg.org (8. Nov. 2023)
DFG 2022 Deutsche Forschungsgemeinschaft – http://www.dfg.de/foerderung/info_wissenschaft/2022/info_wissenschaft_22_25 (4. Nov.2022)
DWB – Diversity Workbench – https://diversityworkbench.net (10. Oct. 2023)
ESA – European Society of Arachnology – https://www.european-arachnology.org/esa/ (8. Nov. 2023)
GBIF – Global Biodiversity Information Facility – Datasets of SMNK and SMNS: https://www.gbif.org/dataset/8277b324-f762-11e1-a439-00145eb45e9a, https://www.gbif.org/dataset/7a681cd5-9781-489f-a43b-dfd117967550 (8. Nov. 2023)
GBIF Integrated Publishing Toolkit (IPT) User Manual – https://ipt.gbif.org/manual/en/ipt (8. Nov. 2023)
GFBIO e. V. – German Federation for Biological Data – https://www.gfbio.org/data-centers (8. Nov. 2023)
ISA – International Society of Arachnology – https://arachnology.org (8. Nov. 2023)

NFDI – Nationale Forschungsdaten Infrastruktur – https://www.nfdi.de/konsortien (8. Nov. 2023)
NFDI4Biodiversity – https://nfdi4biodiversity.org (8. Nov. 2023)

References

Arachnologische Gesellschaft 2023a Atlas der Spinnentiere Europas. – Arachnologische Gesellschaft – Internet: https://atlas.arages.de (09.11.2023)

Arachnologische Gesellschaft 2023b Wiki des Spinnen-Forums. – Internet: https://wiki.arages.de (09.11.2023)

Authmann C, Beilschmidt C, Drönner J, Mattig M & Seeger B 2015 VAT: A System for Visualizing, Analyzing and Transforming Spatial Data in Science. – Datenbank Spektrum 15: 175-184 – doi: 10.1007/s13222-015-0197-y

Bach, A, Raub F, Höfer H, Ottermanns R, Roß-Nickoll M. (in press) ARAapp: Filling gaps in the ecological knowledge of spiders using an automated and dynamic approach to analyze systematically collected community data – Database

Beck J, Ballesteros-Mejia L, Nagel P & Kitching IJ 2013 Online solutions and the „Wallacean shortfall": What does GBIF contribute to our knowledge of species' ranges? – Diversity and Distributions 19: 1043-1050 – doi: 10.1111/ddi.12083

Borges PAV, Gabriel R, Arroz AM, Costa A, Cunha RT, Silva L, Mendoņa E, Martins AMF, Reis F & Cardoso P 2010 The azorean biodiversity portal: An internet database for regional biodiversity outreach. – Systematics and Biodiversity 8: 423-434 – doi: 10.1080/14772000.2010.514306

Bowser CJ 1986 Historic Data Sets: Lessons from the Past, Lessonst for the Future. In: Michener WK (ed.) Research Data Management in the Ecological Sciences. University of South Carolina Press, Columbia, South Carolina. pp. 426

Burkhardt U, Russell DJ, Decker P, Döhler M, Höfer H, Lesch S, Rick S, Römbke J, Trog C, Vorwald J, Wurst E & Xylander WER 2014 The Edaphobase project of GBIF-Germany-A new online soil-zoological data warehouse. – Applied Soil Ecology 83: 3-12 – doi: 10.1016/j.apsoil.2014.03.021

Cardoso P, Erwin TL, Borges PAV & New TR 2011 The seven impediments in invertebrate conservation and how to overcome them. – Biological Conservation 144: 2647-2655 – doi: 10.1016/j.biocon.2011.07.024

Casas-Marce, M, Revilla, E, Fernandes, M, Rodríguez, A, Delibes, M & Godoy, JA 2012 The Value of Hidden Scientific Resources: Preserved Animal Specimens from Private Collections and Small Museums. – BioScience, 62(12): 1077–1082 – doi: 10.1525/bio.2012.62.12.9.

Chamanara J, Gaikwad J, Gerlach R, Algergawy A, Ostrowski A & König-Ries B 2021 BEXIS2: A FAIR-aligned data management system for biodiversity, ecology and environmental data. – Biodiversity Data Journal 9 (e72901) 1-32 – doi: 10.3897/BDJ.9.e72901

DFG 2022 Deutsche Forschungsgemeinschaft. – Internet: http://www.dfg.de/foerderung/info_wissenschaft/2022/info_wissenschaft_22_25 (04.11.2022)

Diepenbroek M 2015 Orientierung im Datenmeer – Infrastruktur für Umwelt- und Ökosystemdaten. – BIOspektrum 21: 467 – doi: 10.1007/s12268-015-0601-z

Diepenbroek M, Glöckner FO, Grobe P, Güntsch A, Huber R, König-Ries B, Kostadinov I, Nieschulze J, Seeger B, Tolksdorf R & Triebel D 2014 Towards an integrated biodiversity and ecological research data management and archiving platform: the German federation for the curation of biological data (GFBio). In: Plödereder E, Grunske L, Schneider E & Ull D (eds) Informatik 2014. Gesellschaft für Informatik e.V., Bonn. pp. 1711-1721

Ferro ML & Flick AJ 2015 „Collection Bias" and the importance of natural history collections in species habitat modeling: A case study using *Thoracophorus costalis* Erichson (Coleoptera: Staphylinidae: Osoriinae), with a critique of GBIF.org. – Coleopterists Bulletin 69: 415-425 – doi: 10.1649/0010-065X-69.3.415

Glöckner FO, Diepenbroek M, Felden J, Güntsch A, Stoye J, Overmann J, Wimmers K, Kostadinov I, Yahyapour R, Müller W, Scholz U, Triebel D, Frenzel M, Gemeinholzer B, Goesmann A, König-Ries B, Bonn A & Seeger B 2020 NFDI4BioDiversity – A Consortium for the National Research Data Infrastructure (NFDI). – doi: 10.5281/ZENODO.3943645

Hänggi A, Stöckli E & Nentwig W 1995 Lebensräume Mitteleuropäischer Spinnen. Centre Suisse de cartographie de la faune, Neuchatel. 460 pp.

Harjes J, Link A, Weibulat T, Triebel D & Rambold G 2020 FAIR digital objects in environmental and life sciences should comprise workflow operation design data and method information for repeatability of study setups and reproducibility of results. – Database 2020 (baaa059): 1-20 – doi: 10.1093/database/baaa059

Hartl N, Wössner E & Sure-Vetter Y 2021 Nationale Forschungsdateninfrastruktur (NFDI). – Informatik Spektrum 44: 370-373 – doi: 10.1007/s00287-021-01392-6

Heidorn PB 2008 Shedding Light on the Dark Data in the Long Tail of Science. – Library Trends 57: 280-299

Holetschek J, Dröge G, Güntsch A & Berendsohn WG 2012 The ABCD of primary biodiversity data access. – Plant Biosystems 146: 771-779 – doi: 10.1080/11263504.2012.740085

iNaturalist 2023 Inaturalist - Internet: https://www.inaturalist.org (23 Dec 2023)

Karam N, Müller-Birn C, Gleisberg M, Fichtmüller D, Tolksdorf R & Güntsch A 2016 A Terminology Service Supporting Semantic Annotation, Integration, Discovery and Analysis of Interdisciplinary Research Data. – Datenbank-Spektrum 16: 195-205 – doi: 10.1007/s13222-016-0231-8

Lotz T, Nieschulze J, Bendix J, Dobbermann M & König-Ries B 2012 Diverse or uniform? - Intercomparison of two major German project databases for interdisciplinary collaborative functional biodiversity research. – Ecological Informatics 8: 10-19 – doi: 10.1016/j.ecoinf.2011.11.004

Maldonado C, Molina CI, Zizka A, Persson C, Taylor CM, Albán J, Chilquillo E, Rønsted N & Antonelli A 2015 Estimating species diversity and distribution in the era of Big Data: To what extent can we trust public databases? – Global Ecology and Biogeography 24: 973-984 – doi: 10.1111/geb.12326

Michener WK 2015 Ecological data sharing. –Ecological Informatics 29: 33-44 – doi: 10.1016/j.ecoinf.2015.06.010

Narimanov N, Hatamli K & Entling MH 2021 Prey naïveté rather than enemy release dominates the relation of an invasive spider toward a native predator. – Ecology and Evolution 11: 11200-11206 – doi: 10.1002/ece3.7905

Nentwig W, Blick T, Gloor D, Hänggi A & Kropf C 2023 Spinnen Europas. – Internet: www.araneae.unibe.ch (06.02.2023)

Novotný P, Seifert S, Rohn M, Diewald W, Štech M & Triebel D 2022 Software infrastructure and data pipelines established for technical interoperability within a cross-border cooperation for the flora of the Bohemian Forest. - Biodiversity Data Journal 10: e87254 – doi: 10.3897/BDJ.10.e87254

Oger P 2023 Les araignées de Belgique et de France. – Internet: https://arachno.piwigo.com (10.11.2023)

Orr MC, Hughes AC, Costello MJ & Qiao H 2022 Biodiversity data synthesis is critical for realizing a functional post-2020 framework. – Biological Conservation 274 (109735): 1-7 – doi: 10.1016/j.biocon.2022.109735

Parr CL, Dunn RR, Sanders NJ, Weiser MD, Photakis M, Bishop TR, Fitzpatrick MC, Arnan X, Baccaro F, Brandão CRF, Chick L, Donoso DA, Fayle TM, Gómez C, Grossman B, Munyai TC, Pacheco R, Retana J, Robinson A, Sagata K, Silva RR, Tista M, Vasconcelos H, Yates M & Gibb H 2017 GlobalAnts: a new database on the geography of ant traits (Hymenoptera: Formicidae). – Insect Conservation and Diversity 10: 5-20 – doi: 10.1111/icad.12211

Pekár S, Wolff JO, Černecká L, Birkhofer K, Mammola S, Lowe EC, Fukushima CS, Herberstein ME, Kučera A, Buzatto BA, Djoudi EA, Domenech M, Enciso AV, Pinanez Espejo YMG, Febles S, Garcia LF, Goncalves-Souza T, Isaia M, Lafage D, Líznarová E, Macias-Hernández N, Magalhaes I, Malumbres-Olarte J,

Michálek O, Michalik P, Michalko R, Milano F, Munevar A, Nentwig W, Nicolosi G, Painting CJ, Petillon J, Piano E, Privet K, Ramirez MJ, Ramos C, Řezáč M, Ridel A, Růžička V, Santos I, Sentenská L, Walker L, Wierucka K, Zurita GA & Cardoso P 2021 The World Spider Trait database: a centralized global open repository for curated data on spider traits. – Database 2021: 1-10 – doi: 10.1093/database/baab064

Porter JH 2000 Scientific databases. In: Michener WK & Brunt JW (eds.) Ecological data: Design, management and processing. Blackwell Science, Oxford. pp. 48-69.

Porter, JH 2018 Scientifc Databases for Environmental Research. In: Recknagel, F. & Michener, W.K. (eds.) Ecological Informatics. Springer International Publishing AG. pp. 27-53.

Purgat P, Ondrejková N, Krumpálová Z, Gajdoš P & Hurajtová N 2021 *Tegenaria hasperi* Chyzer, 1897 and *Zoropsis spinimana* (Dufour, 1820), newly recorded synanthropic spiders from Slovakia (Araneae, Agelenidae, Zoropsidae). – Check List 17: 775-782 – doi: 10.15560/17.3.775

Raub F, Bach A, Bauer T & Höfer H 2023 Editorial. Data paper publication in Arachnologische Mitteilungen – Goals, review criteria, editorial procedures, format, data management and mobilization. – Arachnologische Mitteilungen 66: iii-iv

Recknagel F & Michener WK 2018 Ecological Informatics: An introduction. In: Recknagel, F. & Michener, W.K. (eds.) Ecological Informatics. Springer International Publishing AG pp. 3–10.

Reichman OJ, Jones MB & Schildhauer MP 2011 Challenges and Opportunities of Open Data in Ecology. – Science 331: 703–705 – doi: 10.1126/science.1197962

RfII Rat für Informationsinfrastrukturen 2019 Herausforderung Datenqualität – Empfehlungen zur Zukunftsfähigkeit von Forschung im digitalen Wandel. Göttingen. 172 pp.

Rowe, RJ 2005 Elevational gradient analyses and the use of historical museum specimens: a cautionary tale. – Journal of Biogeography, 32(11): 1883–1897 – doi: 10.1111/j.1365-2699.2005.01346.x

Schneider FD, Fichtmueller D, Gossner MM, Güntsch A, Jochum M, König-Ries B, Le Provost G, Manning P, Ostrowski A, Penone C & Simons NK 2019 Towards an ecological trait-data standard. – Methods in Ecology and Evolution 10: 2006–2019 – doi: 10.1111/2041-210X.13288

Sillero, N, Arenas-Castro, S, Enriquez-Urzelai, U, Vale, CG, Sousa-Guedes, D, Martínez-Freiría, F, Real, R & Barbosa, AM 2021 Want to model a species niche? A step-by-step guideline on correlative ecological niche modelling. – Ecological Modelling 456: 109671. doi: 10.1016/j.ecolmodel.2021.109671

Štokmane M & Cera I 2018 Revision of the calcareous fen arachnofauna: habitat affinities of the fen-inhabiting spiders. – ZooKeys 802: 67–108 – doi: 10.3897/zookeys.802.26449

Telenius A 2011 Biodiversity information goes public: GBIF at your service. – Nordic Journal of Botany 29: 378–381 – doi: 10.1111/j.1756-1051.2011.01167.x

Teschke K, Kraan C, Kloss P, Andresen H, Beermann J, Fiorentino D, Gusky M, Hansen MLS, Konijnenberg R, Koppe R, Pehlke H, Piepenburg D, Sabbagh T, Wrede A, Brey T & Dannheim J 2022 CRITTERBASE, a science-driven data warehouse for marine biota. – Sci Data 9: 1–7 – doi: 10.1038/s41597-022-01590-1

Triebel D, Hagedorn G & Rambolt G 1999 Diversity Workbench – A virtual research environment for building and accessing biodiversity and environmental data. – Internet: http://www.diversityworkbench.net (23.01.2023)

Triebel D, Peršoh D, Nash TI, Zedda L & Rambold G 2007 LIAS – an interactive database system for structured descriptive data of Ascomycete. In: Biodiversity databases. Techniques, politics, and applications. CNC Press, Boca Raton. pp. 99–110

Underwood AJ, Chapman MG & Connell SD 2000 Observations in ecology: you can't make progress on processes without understanding the patterns. – Journal of Experimental Marine Biology and Ecology 250: 97–115 – doi: 10.1016/S0022-0981(00)00181-7

Veiga AK, Saraiva AM, Chapman AD, Morris PJ, Gendreau C, Schigel D & Robertson TJ 2017 A conceptual framework for quality assessment and management of biodiversity data. – PLOS ONE 12 (e0178731) 1–20 – doi: 10.1371/journal.pone.0178731

Weber J, Ebert B, Diepenbroek M, Kostadinov I & Glöckner FO 2021 NFDI4BioDiversity - NFDI-Konsortium für Biodiversitäts-, Ökologische und Umweltdaten. – Bausteine Forschungsdatenmanagement 2: 98–109 – doi: 10.17192/bfdm.2021.2.8334

Weibulat T, Triebel D & König-Ries B 2023 Das Netzwerk aus Anbietern von Daten, Diensten und IT Werkzeugen in NFDI4Biodiversity. NFDI4Biodiversity-Jahreskonferenz 2022 (NFDI4Biodiversity AHC 2022), Berlin, Germany. – doi: 10.5281/zenodo.7644275

Weiss M, Weibulat T, Seifert S, Monje JC, Ruff M, Neubacher D, Reichert W & Triebel D 2018 A flexible Diversity Workbench tool to publish biodiversity data from SQL database networks through platforms like GFBio. – doi: 10.22032/dbt.37797

Wersebeckmann V, Kolb S, Entling MH & Leyer I 2021 Maintaining steep slope viticulture for spider diversity. – Global Ecology and Conservation 29 (e01727): 1–12 – doi: 10.1016/j.gecco.2021.e01727

von Wettberg E & Khoury CK 2022 Biodiversity data: The importance of access and the challenges regarding benefit sharing. – Plants, People, Planet 4: 2–4 – doi: 10.1002/ppp3.10241

Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T & Vieglais D 2012 Darwin core: An evolving community-developed biodiversity data standard. – PLoS ONE 7 (e29715): 1-8 – doi: 10.1371/journal.pone.0029715

Wiśniewski K, Rozwałka R & Wesołowska W 2018 Distribution, habitat affinities and phenology of the *Micrargus herbigradus*-species group (Araneae: Linyphiidae) in Poland. – Biologia 73: 151–164 – doi: 10.2478/s11756-018-0026-5

World Spider Catalog 2023 World Spider Catalog. Version 24 – Natural History Museum, Bern – Internet: https://wsc.nmbe.ch (23. Jan.2023)

Yesson C, Brewer PW, Sutton T, Caithness N, Pahwa JS, Burgess M, Gray WA, White RJ, Jones AC, Bisby FA & Culham A 2007 How global is the global biodiversity information facility? – PLoS ONE 2 – doi: 10.1371/journal.pone.0001124

Zizka A, Antunes Carvalho F, Calvente A, Rocio Baez-Lizarazo M, Cabral A, Coelho JFR, Colli-Silva M, Fantinati MR, Fernandes MF, Ferreira-Araújo T, Gondim Lambert Moreira F, Santos NMC, Santos TAB, dos Santos-Costa RC, Serrano FC, Alves da Silva AP, de Souza Soares A, Cavalcante de Souza PG, Calisto Tomaz E, Vale VF, Vieira TL, Antonelli A, Carvalho FA, Calvente A, Baez-Lizarazo MR, Cabral A, Ramos Coelho JF, Colli-Silva M, Fantinati MR, Fernandes MF, Ferreira-Araújo T, Lambert Moreira FG, da Cunha Santos NM, Borges Santos TA, dos Santos-Costa RC, Serrano FC, da Silva APA, de Souza Soares A, de Souza PGC, Tomaz EC, Vale VF, Vieira TL & Antonelli A 2020 No one-size-fits-all solution to clean GBIF. – PeerJ 8: e9916 – doi: 10.7717/peerj.9916