# Towards Institutional Data Tracking: A Case Study from the Leibniz Centre for Tropical Marine Research (ZMT)

**Birte Hemmelskamp-Pfeiffer[*], Helen Clara George[*,**], Finn Opätz[*]**

[*] Leibniz Centre for Tropical Marine Research (ZMT)
[**] University of Bremen

*This work was supported by the German Research Foundation DFG under the grant agreement number 442032008 (NFDI4Biodiversity). The project is part of NFDI, the National Research Data Infrastructure Programme in Germany.*
*The ZMT use case #27 in the NFDI4Biodiversity focuses on institutional research data management. The aim is to identify challenges in dealing with research data, develop corresponding concepts and guidelines and make these available to the community.*

## 1 Overview

As open science practices become more prevalent, research institutions face an increasing need to systematically track and promote research datasets generated by their researchers. This need is especially acute when datasets are disseminated through a variety of general-purpose and disciplinary repositories, often without centralised institutional oversight or aggregation mechanisms. At the Leibniz Centre for Tropical Marine Research (ZMT) researchers have long been publishing datasets. These include data supplements associated with journal articles, dedicated data papers, and data deposits in general repositories such as Zenodo or disciplinary repositories like PANGAEA. At the same time, many datasets were not published – either due to legal, ethical, or technical constraints (e.g. the need for anonymisation), or due to a lack of awareness or guidance on how to publish research data appropriately. In the absence of an institutional repository or a Current Research Information System (CRIS), it is challenging to identify and monitor – especially historical – research data publications. This impedes institutional visibility and hinders efforts to leverage data assets for knowledge exchange and reporting.

This paper presents a case study on institutional research data tracking accomplished by the Research Data Service at ZMT. It first discusses the strategic relevance of systematic monitoring and introduces a working definition of institutional research data. By demonstrating how metadata from published datasets can be integrated into the ZMT DataPortal, the study highlights how visibility, discoverability,

and institutional oversight of research outputs can be strengthened. Practical approaches for harvesting metadata from major repositories are outlined, with a focus on scalable and sustainable solutions. In this context, the paper examines the use of Application Programming Interfaces (APIs), web-based search interfaces, Python libraries, and open-source tools such as Zotero.

# 2 Background: The Need for Systematic Research Data Monitoring

## 2.1 Why Monitor Research Data?

Research data monitoring, also referred to as research data tracking, is the process of systematically identifying and recording datasets produced by an institution's researchers in order to build or enhance a data catalogue or institutional registry. Such monitoring serves multiple strategic functions, including visibility of institutional output, improved data governance, and facilitation of reuse and reporting through transparency and trust. Additionally, it can foster consistency, simplify project planning, and over time result in a natural integration of research data management into the research process.

Tracking can be conducted through various methods – ranging from manual curation to automated in-house workflows or third-party solutions (Elsevier, 2023). The presence of a Current Research Information System (CRIS) considerably facilitates this process by enabling structured recording, linking, and management of datasets. Conversely, institutions lacking a CRIS or institutional repository have considerable difficulties, especially when datasets are distributed via several disciplinary repositories (e.g. PANGAEA), general-purpose platforms (e.g. Zenodo), or individual project websites. In this context, the role of comprehensive and standardised metadata becomes critical. In particular, the use of persistent identifiers – such as the Research Organization Registry ID (ROR) for institutional affiliation and the Open Researcher and Contributor ID (ORCID) for author identification – is essential for reliably attributing research data to specific institutions and individuals (Research Intelligence, Elsevier, 2023).

## 2.2 Institutional Affiliation and Research Data Publications

This section proposes a framework for attributing research data publications based on institutional affiliation. In general, the assignment of a publication to an institution is related to the role of the institution in the research process (e.g. legal, financial) and affiliation of the author(s) to that institution. In this paper, we do not dive deeper into the broader and complex question of legal ownership, which remains the subject of ongoing scholarly and legal debate (see Grünberger *et al.*, 2024; Hübner, 2024).

Historically, ZMT researchers published research data independently, without institutional mediation, which often resulted in incomplete or incorrect attribution of institutional affiliation in the metadata. The establishment of dedicated data management positions at ZMT has significantly improved this process. Consequently, current datasets are more reliably attributed to ZMT than historical ones. For the systematic documentation of historical research data at ZMT, it is necessary to establish a precise

definition of what is to be considered institutional data.

In the case of first authorship, institutional attribution is generally well substantiated. The first author is typically responsible for the conception and design of the study, the collection and analysis of data, and the preparation of the manuscript. As these core research activities are most often conducted within the facilities and under the auspices of the author's affiliated institution, the connection between the publication and the institution is both plausible and justified. Furthermore, datasets resulting from projects either directly funded by the institution or from third-party funding with a principal investigator employed at the institution can likewise be considered institutional research data.

The complexity of institutional attribution increases when different publication policies are considered, i.e. co-authored or co-developed publications. In the case of co-publication, the main research may have been carried out at another institution. However, while it may not be institutional research data in the narrower sense, such publications can be shown as part of the research conducted at an institute (co-authored publication). For example, the University of Bremen requires all academic members to state the university as their institutional affiliation in all scholarly outputs, irrespective of the publication policy. This policy does not confer legal ownership to the university, but reflects an affiliation-based model for attributing scholarly contributions.

Another challenge is related to a change in institutional affiliation during the research process. The University of Bremen states that "former members of the University of Bremen whose publications are based on research generated at the University of Bremen should name the University of Bremen in addition to the affiliation of their new employer. (…)" Similarly, the publication guideline of the Johannes Gutenberg University Mainz (n.d.) states that in the case of a change in institutional affiliation during the research or publication process, the institution at which the research was ultimately conducted is named. Both institutions have to be named if relevant parts of the research were conducted in both institutions – while the institution with the main research is named first. We follow these arguments and institutionally assign ZMT research data as noted below:

- If a dataset was published by a researcher as first author who was employed at ZMT at the time of publication, the dataset is assigned to the institute.
- If data has been collected and published by a ZMT researcher as the first author within third-party funded projects that have been acquired by ZMT researchers, it can be assigned as a ZMT data publication, too (also see MacColl *et al.*, 2011).
- If a data publication is only accomplished after the employment relationship at ZMT has ended, but the main work on data has been done at ZMT, the first affiliation should always be the institute where the research has been carried out (ZMT data). If the data work is

equally distributed on both institutions, these publications are considered as multi-institutional data.

- If a ZMT researcher is only a co-author of the data published, but research and funding are attributed to another institution, the data is considered as a co-authored publication.

In all cases where the funding and extent of the research contribution of a ZMT researcher acting as a co-author are unclear, the data publication is considered a co-authored dataset.

## 3 Research Data Oversight and Visibility at ZMT

Metadata, or data about data, could be considered the eyes and mouths of datasets, giving valuable insights about them. In the process of identifying institutional datasets, metadata plays a crucial role, providing essential information such as title, abstract, and authors of a dataset. It also provides information about technical identifiers like DOIs and URLs, which could make the datasets more visible as well as accessible. Metadata should also include information about the institution through author affiliations, funding agency and project information.

In the past, the absence of an institutional data repository at ZMT, combined with no DOI assignment role, resulted in a lack of oversight of research data either remaining unpublished or being distributed across various disciplinary repositories. Furthermore, the lack of a Current Research Information System (CRIS) also makes it difficult to gain an overview of data publications.

However, providing an overview of and easy access to ZMT data and metadata was desired, so a single point of access was envisaged: the ZMT DataPortal. It was established with IndiScale GmbH and is based on the open-source software LinkAhead. The portal provides a searchable database of ZMT datasets irrespective of their storage location, i.e. data can be stored both on internal servers or public data repositories. It currently mainly provides metadata of and links to research data by ZMT researchers published in PANGAEA. These are identified through a mixture of manual tracking and automated retrieval solutions provided by IndiScale. Metadata from research data published in other data repositories, as well as internal data, will soon be included in the ZMT DataPortal.

## 4 General Methods for Metadata Harvesting

Metadata harvesting encompasses a range of strategies designed to collect descriptive information from heterogeneous sources and subsequently normalise it into an institutional schema.

One of the most established approaches is protocol-based metadata harvesting. Standardised machine-readable communication protocols like OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting), specifically designed for structured metadata exchange, are a reliable method. This protocol enables repositories to expose metadata records in a uniform manner, formulating

requests through HTTP and returning responses as well-formed XML documents.

Beyond protocol-based solutions, an increasing number of repositories and data platforms provide programmatic interfaces such as RESTful APIs, and in some cases SOAP or GraphQL APIs. These programmatic access points allow users to query, filter, and retrieve metadata in widely supported formats, including JSON, XML, or CSV (ESAP, 2023). APIs further facilitate scheduled and scalable harvesting while affording sophisticated filtering mechanisms such as by institutional name, unique identifiers, or variant spellings. Prominent platforms like OpenAlex and DataCite facilitate access to metadata with these APIs in addition to offering powerful search and filtering capabilities – such as narrowing results by institution (see also Lemonidou, 2025). There are tools like Zotero that make use of both protocols like OAI-PMH and RESTful APIs (Nag and Guhathakurta, 2024).

When reliable protocols and APIs are not available, web scraping and web crawling could also be utilised to harvest metadata (Tamasevicius, 2025). With the help of crawlers or scrapers web pages can be visited and relevant information extracted and parsed to retrieve the metadata. This can later be structured into institutional format and made visible.

To further streamline integration into research workflows, Python wrappers and libraries are often available. For instance, the *pangaeapy* library simplifies access to metadata from the PANGAEA repository within custom Python scripts. Additionally, browser-based tools and extensions, such as browser inspection utilities and Zotero connectors, offer alternative methods for extracting structured metadata directly from web content.

The following sections present the use of these techniques and platforms for retrieving metadata from datasets published by ZMT researchers across various repositories over time.

## 5 ZMT Datasets in Disciplinary and General-Purpose Repositories

The landscape of digital research data repositories is highly complex and interconnected. Certain repositories specialise in specific disciplines, e.g. life sciences or earth system research. Other repositories accept a wide range of datasets, including those from multiple disciplines, and are categorised as general-purpose repositories. They provide broader support across diverse data types and disciplines. For research that is multidisciplinary or does not fit into a specific disciplinary framework, general-purpose repositories serve as valuable platforms for dataset publication.

### 5.1 Monitoring Data Publications in PANGAEA

Many researchers at ZMT have published, and continue to publish, their research data in the disciplinary repository PANGAEA, the Data Publisher for Earth and Environmental Sciences. Its focus lies on archiving, publishing, and distributing georeferenced observational and experimental data from earth system research. The majority of datasets is freely accessible and all entries conform to a strict

catalogue of metadata annotation and a high degree of structural data harmonisation, which is ensured through a careful editorial process (Felden *et al.*, 2023).

Datasets in which the keyword "ZMT" was included as institutional affiliation, or which are linked to the respective ZMT project, were easily harvested by a crawler developed in cooperation with IndiScale GmbH and subsequently made visible in the ZMT DataPortal. However, when attempting to identify historical research data publications in PANGAEA, we realised that not all datasets in PANGAEA produced by ZMT researchers had been assigned to ZMT. One reason is that, in earlier decades, institutional affiliation with ZMT was not always consistently recorded[1]. Consequently, the presence of a considerable number of datasets in PANGAEA authored by ZMT researchers but lacking institutional attribution has created the need to retrospectively identify these datasets and assign the ZMT keyword.

In order to identify and label datasets from ZMT researchers in PANGAEA, the Python library *pangaeapy* has proven to be an efficient tool. The library was used as the primary resource to automate the retrieval of historical datasets from ZMT without explicit ZMT affiliation. To identify these datasets, the following strategies were applied, adhering to the affiliation-based model described in the previous chapter:

> [1] Identification via Journal Publications: ZMT datasets in PANGAEA were identified through journal articles authored by ZMT researchers, as listed in the institutional publications database. Using article DOIs as queries in *pangaeapy*, corresponding dataset DOIs were retrieved, including both directly linked ("Supplement to") and referenced ("Related to") entries. These datasets were then assessed for institutional attribution based on author affiliations, project associations, and related metadata.
>
> [2] Identification via Researcher Names: To locate ZMT datasets in PANGAEA without institutional keywords, selected researcher names were used as queries in *pangaeapy*. The resulting dataset DOIs were then reviewed to verify institutional attribution.

Datasets identified via *pangaeapy* that already contained the keyword "ZMT" were excluded. The remaining entries were verified based on funding source, first authorship, or co-authorship by ZMT researchers. Final attribution required manual inspection. Verified datasets were assigned a technical keyword (not visible in metadata) to mark ZMT authorship; co-authored datasets are currently being tagged with an additional keyword. These classifications will be integrated into the ZMT DataPortal. In sum, this verification process enables the systematic identification of historical datasets linked to ZMT

[1] Nowadays, a data manager for the natural sciences at ZMT additionally acts as a PANGAEA data editor, curating data submissions from ZMT researchers and ensuring their proper designation as institutional data publications.

in PANGAEA.

## 5.2 Monitoring Data Publications in Biological & Life Sciences Repositories

Public repositories that serve as archives for molecular biology data employ highly specific metadata schemas tailored to the characteristics of biological datasets. These schemas enable interoperability and standardisation of metadata across various biological data sources. Nucleotide sequencing datasets are hosted under the International Nucleotide Sequence Database Collaboration (INSDC). The ENA repository, provided by the European Bioinformatics Institute (EMBL-EBI) of the European Molecular Biology Laboratory, is part of INSDC. GenBank, maintained by the NCBI is also part of INSDC (Cummins *et al.*, 2022). These repositories are not merely data storage platforms; they are comprehensive databases containing all publicly available nucleotide sequences. They provide annotated DNA and RNA samples, nucleotide sequence data, sequencing assembly details, protein translations and related information. Researchers from ZMT have contributed and submitted datasets to these repositories. The metadata associated with these datasets can be accessed using the methods discussed earlier. The datasets in GenBank are visible and available through the efficient interface of NCBI. ENA offers REST API, through which users can access the metadata. A straightforward method for retrieving metadata is to query the corresponding web-portals with accession numbers or keywords. Search by the institutional name and its variants revealed datasets from ZMT research in these repositories (for details, see Table 2 in the Appendix).

## 5.3 Monitoring Data and Metadata Retrieval from Zenodo

Researchers at ZMT have frequently deposited research data in Zenodo[2], the open-access repository developed within the OpenAIRE initiative for open data dissemination. Several strategies were employed to systematically identify historical datasets associated with ZMT, beginning with direct searches on the Zenodo platform and the extraction of metadata via the Zotero reference management tool (see below). Beyond manual searches within Zenodo, the OpenAlex platform provides additional functionality, offering both an extensive API and elastic search capabilities to locate datasets hosted on Zenodo.

Zenodo itself supports dataset discovery through its publicly accessible REST API, which enables programmatic searches. In practice, however, the primary approach has involved querying both the Zenodo and OpenAlex user interfaces. A persistent challenge in this process is ensuring accurate attribution of datasets to the institution, as the affiliation of dataset creators often provides the initial

[2] The ZMT Community in Zenodo was recently introduced.

indicator of institutional association. This is facilitated by Zenodo's query functionality, which allows filtering by the field "creators.affiliation". Additionally, OpenAlex enhances this process by supporting refined filtering options such as institutional name, resource type (dataset), and year of publication. These features significantly streamline the identification and retrieval of ZMT-related datasets, provided that institutional information is explicitly linked to the dataset metadata.

## 5.4 Zotero for Metadata Harvesting in other General-Purpose Data Repositories

The reference management tool Zotero provides an efficient way to harvest and organize metadata from multiple repositories. For historical ZMT datasets, the Zotero Browser Connector extracts metadata from repository and publisher webpages into Zotero libraries, which are later verified and integrated into the ZMT DataPortal. The Zotero Web API allows programmatic access to saved metadata in formats such as JSON, CSL, BibTeX, or RIS, supporting both interactive and small-scale automated harvesting (Zotero, 2025a). Zotero employs multiple strategies to obtain metadata from repositories. It uses hundreds of *translators* – small programs that parse the structure of website-specific metadata formats (Zotero, 2025b). These translators scan webpages for embedded metadata tags and extract the information directly. Zotero can also extract metadata from PDFs when available (Zotero, 2025c) and use DOI or ISBN lookups via APIs, querying services such as CrossRef or DataCite to retrieve structured metadata. Additionally, it can import citation files (e.g., BibTeX, RIS) directly from websites and programmatically access bibliographic data through API calls (Zotero, 2025d). Zotero's capabilities are employed to systematically locate and harvest metadata of datasets generated by ZMT researchers, including supplementary data linked to journal publications. Using Zotero and its Browser Connector, datasets from interdisciplinary projects are collected from multiple repositories, with metadata automatically imported into Zotero libraries for later integration into the ZMT DataPortal (for details, see Table 1 in the Appendix).

## 6 Conclusion

With the growing importance of research data publications as both academic output and a marker of institutional visibility, this paper has underscored the need for systematic monitoring of research data. We highlighted the role of institutional affiliation in data publications and presented the ZMT approach for increasing transparency and recognition of its research data outputs. Particular attention was given to general strategies of metadata harvesting and their application to monitoring datasets in PANGAEA, Zenodo, other discipline-specific and general-purpose repositories.

Our analysis shows that APIs, Python libraries, and metadata platforms such as OpenAlex enable institutions to effectively harvest, consolidate, and present datasets. Tools like Zotero complement

these approaches for smaller-scale or manual collection tasks. When combined within a unified metadata aggregator, these resources strengthen the visibility, discoverability, and recognition of institutional research across disciplinary boundaries.

Despite systematic efforts to identify historical ZMT datasets across repositories and to make them accessible through the ZMT DataPortal, additional datasets remain to be located and integrated. Early work focused on the harmonisation of institutional name variants to establish consistent repository matching and to define institutional datasets. Metadata were harvested from multiple sources using APIs and wrappers, followed by normalisation into a unified institutional schema to ensure standardised access through persistent links. At present, clearly identifiable ZMT datasets in PANGAEA are available through the ZMT DataPortal. Looking ahead, the implementation of automated workflows promises to extend this visibility to historically published datasets from ZMT researchers in other repositories, reducing reliance on manual querying and retrieval.

## References


Cummins C., Ahamed A., Aslam R., Burgin J., Devraj R., Edbali O., Gupta D., Harrison P.W., Haseeb M., Holt S., Ibrahim T., Ivanov E., Jayathilaka S., Kadhirvelu V., Kay S., Kumar M., Lathi A., Leinonen R., Madeira F., Madhusoodanan N., Mansurova M., O'Cathail C., Pearce M., Pesant S., Rahman N., Rajan J., Rinck G., Selvakumar S., Sokolov A., Suman S., Thorne R., Totoo P., Vijayaraja S., Waheed Z., Zyoud A., Lopez R., Burdett T., Cochrane G. (2022). The European Nucleotide Archive in 2021. Nucleic Acids Res.;50:D106–D110. doi: 10.1093/nar/gkab1051.

Felden, J.; Möller, L.; Schindler, U.; Huber, R.; Schumacher, S.; Koppe, R.; Diepenbroek, M.; Glöckner, F.-O. (2023): PANGAEA – Data Publisher for Earth & Environmental Science. Scientific Data, 10(1), 347, https://doi.org/10.1038/s41597-023-02269-x

Grünberger, T., Ebert, B., Brand, O., Tschink, D. (2024) NFDI4Biodiversity-Podcast zu rechtlichen Aspekten im Umgang mit Biodiversitäts- und Umweltdaten. Episode 4 - Institutionelle Perspektive. https://t1p.de/4biodiv-rechtspodcast4

Hübner, A. (2024). Who "owns" research data?. Zenodo. https://doi.org/10.5281/zenodo.11059315

Johannes Gutenberg University Mainz (n.d.) *Publication Guidelines of Johannes Gutenberg University Mainz*. https://www.ub.uni-mainz.de/en/publication-guidelines-johannes-gutenberg-university-mainz#:~:text=If%20researchers%20change%20their%20institutional%20affiliation%20during,research%20was%20carried%20out%20is%20named%20first, last accessed: 29.09.2025.

Lemonidou, K. (2025). How can we monitor where Leiden researchers deposit their data? A comparative analysis between open sources. *Digital Scholarship@Leiden.* https://www.digitalscholarshipleiden.nl/articles/data-monitoring-comparative-study, last accessed: 29.09.2025.

MacColl, J. and Jubb, M. (2011). *Supporting Research: Environments, Administration and Libraries* (Dublin, Ohio: OCLC Research) http://www.oclc.org/research/publications/library/2011/2011-10.pdf, last accessed: 29.09.2025.

Nag, R., Guhathakurta, R. (2024). Metadata Harvesting: Applications and Influence in Digital Publishing. *Open Access Cases*, *1*(4). https://oacases.com/index.php/cases/article/view/15

Tamasevicius, J. (2025). Web Crawling vs Web Scraping: What's the Difference? *Decodo.* https://decodo.com/blog/crawling-vs-scraping, last accessed: 29.09.2025.

University of Bremen, https://www.suub.uni-bremen.de/uploads/cms/files/Richtlinien_Forschungspublikationen.pdf, last accessed: 29.09.2025

ZMT, Leibniz Centre for Tropical Marine Research (2025). ZMT Research Data Policy (1.0). Zenodo. https://doi.org/10.5281/zenodo.14672213

Zotero (2025a). *Zotero Web API Documentation.* https://www.zotero.org/support/dev/web_api/v3/basics, last accessed: 29.09.2025.

Zotero (2025b). *Zotero Translators.* https://www.zotero.org/support/translators, last accessed: 29.09.2025.

Zotero (2025c). *Retrieve PDF Metadata.* https://www.zotero.org/support/retrieve_pdf_metadata, last accessed: 29.09.2025.

Zotero (2025d). *Exposing Your Metadata.* https://www.zotero.org/support/dev/exposing_metadata, last accessed: 29.09.2025.

*Use of AI to polish written content.*

# Appendix

## Overview: General-Purpose and Disciplinary Repositories

**Table 1: General-Purpose Repositories Harvested Using Zotero.** These include Dryad, Figshare, the Center for Open Science (OSF), and Zenodo. Table 1 provides a brief overview of repositories queried by Zotero in which ZMT researchers have published datasets over time. It also indicates the availability of API support, Python wrappers or packages, and institutional name filtering options to date.

| Repository Name | API | Python libraries | Institution filtering | Note |
|---|---|---|---|---|
| **Dryad (Dryad API)** | RESTful API is available | No official Python wrapper, but dryad2dataverse package can be used for metadata transfer | Institutional filtering is possible through author affiliations | Metadata of datasets in Dryad can also be obtained from OpenAIRE \| Find and Share research |
| **Figshare (API Figshare)** | API and OAI-PMH support available | Ldcoolp-figshare facilitate Python-based access and curation tasks | Institutional data can be found through author affiliation | Figshare is also part of OpenAir and the datasets are findable there, too |
| **OSF (Open Science Framework) (OSF Client)** | Public REST API is available | Community wrapper osfclient | Indirect (author-based or keyword search) | Institution filtering is possible for member institutions (see here) |
| **Zenodo (Zenodo.Developers)** | Public REST API and OAI-PMH support | Frictionless Framework provides the ZenodoControl class (Zenodo_Portal) | Possible via communities, author affiliations or contributor fields | Communities provide direct link to institutions |

**Table 2: Discipline-Specific Repositories Harvested Using Zotero.** These include ENA and GenBank. Table 2 provides a brief overview of repositories queried by Zotero in which ZMT researchers have published datasets over time. It also indicates the availability of API support, Python wrappers or packages, and institutional name filtering options to date.

| Repository Name | API | Python libraries | Institution filtering | Note |
|---|---|---|---|---|
| **European Nucleotide Archive (ENA)** | ENA Portal API and ENA Browser API support are available | The python package ffq supports metadata retrieval | Submitter ID and centre name in metadata provide institutional information | Search with institutional name returns data when submitter ID is available with metadata |
| **GenBank / NCBI (Biopython)** | Entrez (eUtils) | BioPython (Bio.Entrez) | Search via author names and parse affiliations | The metadata extraction from NCBI is easy using the Zotero tool |

## Useful Tools and Links

OpenAlex Explorer: https://explore.openalex.org

DataCite Search: https://commons.datacite.org

Zotero: https://www.zotero.orgpy

Dataverse: https://github.com/AUSSDA/py

DataverseBiopython: https://biopython.org

OSFClient: https://github.com/osfclient/osfclient