Latif, Atif; Limani, Fidan; Tochtermann, Klaus

Article — Published Version

A Generic Research Data Infrastructure for Long Tail Research Data Management

Data Science Journal

Suggested Citation: Latif, Atif; Limani, Fidan; Tochtermann, Klaus (2019) : A Generic Research Data Infrastructure for Long Tail Research Data Management, Data Science Journal, ISSN 1683-1470, Ubiquity Press, London, Vol. 18, Iss. 1 (article no. 17), pp. 1-11, https://doi.org/10.5334/dsj-2019-017

DATA SCIENCE JOURNAL

Latif, A, et al. 2019. A Generic Research Data Infrastructure for Long Tail Research Data Management. *Data Science Journal,* 18: 17, pp. 1–11. DOI: https://doi.org/10.5334/dsj-2019-017

RESEARCH PAPER

A Generic Research Data Infrastructure for Long Tail Research Data Management

Atif Latif, Fidan Limani and Klaus Tochtermann

ZBW – Leibniz Information Center for Economics, Kiel/Hamburg, DE

Corresponding author: Atif Latif (a.latif@zbw.eu)

The advent of data intensive science has fueled the generation of digital scientific data. Undoubtedly, digital research data plays a pivotal role in transparency and re-producibility of scientific results as well as in steering the innovation in a research process. However, the main challenges for science policy and infrastructure projects are to develop practices and solutions for research data management which in compliance with good scientific standards make the research data discoverable, citeble and accessible for society potential reuse. GeRDI – the Generic Research Data (RD) Infrastructure – is such a research data management initiative which targets long tail content that stems from research communities belonging to different domain and research practices. It provides a generic and open software which connects research data infrastructures of communities to enable the investigation of multidisciplinary research questions.

Keywords: Research Data; Research Data Management; Generic Research Data Infrastructure; Long Tail Content

1 The Rise of Research Data Infrastructures

From decades, the collection and organization of data has been an integral part in steering the research process. Typically, this data as the scientific findings is encapsulated in a scientific article and is further shared by means of conventional publication life cycle.[1] A scientific article usually describes the problem statement, methodologies and experimental tools used for solving the problem. Over completion, this research study goes through the peer reviewing for quality check and on acceptance is published and made available for other scientist to reference. Critically, the whole publication cycle promotes good scientific practices and help in dissemination of scientific findings but is weaker when it comes to the reproduction and transparency of the results. The main reasons are: the failure to mainstream the very idea of research data management and inability of making it an integral part of publication life cycle.

In recent years, the advent of data intensive research has successfully portrait digital data's great potential, if properly managed and shared among researchers (Smith Rumsey, 2010) (Buckland, 2011). Undeniably, the transparency and re-producibility of scientific results are now increasingly dependent on availability and management of research data. In general, data-enabled research as a 4th research paradigm (Hey, 2012) is driving the dissemination of research data (RD) as independent, publishable research artifacts. Many scientific disciplines are producing a lot of RD during or as an end goal of research projects; as a result, RD has now emerged as a 1st class research citizen, breaking away from the "confines" of research publications, with enough traction, added value and its own management ecosystem (metadata description, curation, licensing issues, etc.). The incentives that fuel this growth vary; the research visibility, RD reuse in validation efforts, or scenarios of it being used in novel ways, are just some of the typical drivers for this practice (Bobby Vocile, 2017).

RD has recently become a focal point in the EU Open Science Policy[2] which states the exchange of research data within scientific disciplines will create added value for the progress of science and innovation. This, however, requires adapting the research practice and prerequisites for the RD that data publishers need to

[1] http://www.lib.berkeley.edu/scholarly-communication/publishing/publishing-lifecycle.

[2] https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-policy-platform.

adhere to. In order to reap the benefits of RD, funding agencies and many (trans)national initiatives such as European Open Science Cloud (EOSC)[3] and GO FAIR Initiative[4] are already pushing for a set of criteria that RD needs to abide by.

In the midst of RD infrastructure initiatives, we see a need to target communities that lack established research management practices, including adopted metadata standards, research lifecycle and associated infrastructures. Long tail data communities consists of relatively small but heterogeneous research dataset which are curated individually and managed without adherence to particular RD standards. Relatively limited scale organization like universities, research institute and department, communities and libraries are the prime example of such cases. To an RDI that aims to provide seamless infrastructure-like support for long tail research data to be findable, accessible, interoperable and reusable (Wilkinson et al., 2016), the data heterogeneity present a biggest challenge. GeRDI[5] – the Generic RD Infrastructure – is a federated research infrastructure for long tail research data. It targets research communities from different domains, research practices, standards (metadata, research workflow, etc.), and research support services (Grunzke et al., 2017), exactly fitting the long tail research community profile.

The paper is structured as follows: we provide a state of the art on various research data management initiatives. Next, we provide a brief overview of GeRDI approach towards requirement gathering from its research communities. With reference to our pilot community requirements, We then discuss the GeRDI design, metadata standard options, metadata schema finalization and architecture. A brief review of implemented services with respect to research data life cycle is provided next. In the end, we discuss some of current and future challenges and conclude this paper.

2 State of the Art

Various initiatives already support research data sharing and reuse to help researchers in their daily research work. At the EU level, the EOSC,[6] an aspiring RDI undertaking, aims to federate existing (and future) RDIs – across disciplines – under a single umbrella, and provides (open) services for the European researcher community. Metadata standards, both generic and discipline-specific in scope, designed with RD in mind, create a "presence" among researchers and make RD findable for interested parties. DataCite is but one of the generic metadata standards that promotes research data citation (Starr et al., 2015). Furthermore, the availability of RD repositories, either institutional covering universities, research centers, libraries, etc., or repositories that reside "outside" the research environment are specifically-built to support services for RD and represents an additional enabler for RD sharing and reuse.

International repository initiatives like Figshare,[7] Zenodo[8] and Dyrad[9] are the notable examples for institutional RD archiving. Moreover, other significant national initiatives for RD are: RADAR – which is a DFG funded project (Brophy & Razum, 2017) and delivers a comprehensive archival and publication service of the long tail data to institutions and researchers to provide a cross disciplinary search capabilities. SowiDataNet is an other project (Linne & Zenk-Möltgen, 2017) operated by GESIS Leibniz Institute for Social Sciences, is a research data repository for social sciences and economics supporting discipline specific search.

PANGAEA[10] – Data Publisher for Earth Environmental Science is another disciplinary infrastructure hosted jointly by the Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research (AWI)[11] and the Center for Marine Environmental Science (MARUM) at the University of Bremen. PANGEA enables the long-term preservation, publication and dissemination of scientific data and provides scientific data management for projects and institutes. The Geo-referenced and quality-controlled data and metadata searching features are the stand out one for researchers. Finally, projects that work at the level of RD collections – repository registries – strive to further organize RD repositories from the same/different domains into a single (and rich) point of access and provide useful set of services to researchers. Registry re3data.org (Pampel et al., 2013) is one such example.

Currently, we can distinguish between two streams of RD infrastructures: one that targets Big Data and another that targets long tail content (e IRG, 2016). At present the greater focus has been on the accessibility

[3] https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud.
[4] https://www.go-fair.org/.
[5] http://www.gerdi-project.de/.
[6] https://www.eosc-portal.eu.
[7] https://knowledge.figshare.com.
[8] https://www.zenodo.org/.
[9] https://datadryad.org/.
[10] https://pangaea.de/.
[11] http://www.awi.de/en.html.

of Big Data which bids to target large scale organizations and enterprises which produce large but homogeneous RD. On the contrary, citing the reason of data diversity, less number of research data infrastructures (RDI) are targeting the provision of long tail research data (Horstmann, Nurnberger, Shearer, & Wolski, 2017). However, this is also a well established fact, that most of the important data has been produced by long tail projects which undeniably are the breeding ground for new ideas and never before attempted science.

3 GeRDI: Approach Overview

In this section, we give an overview of the activities in order to provide the necessary context, identify the challenges, and provide the corresponding solutions faced in the GeRDI project so far. We provide more details on each of these activities in the following sections.

In the beginning, we considered few **RD lifecycle** models that researchers typically follow. This helped us to better understand the context of RD management, set a common ground with researchers in terms of expectations, and identify areas for further examination via requirements gathering techniques. This also turned out to be an effective scope-definition mechanism in the project: it was both common/easy to understand and relate to by researchers, and easy for us to map it to multiple aspects in GeRDI.

The RD lifecycle provided us with enough context to support our next activity in the process – that of **requirements gathering**. The key driver for this activity are the research communities. The communities range across disciplines, from digital humanities, to bio-informatics, to social sciences and marine sciences. This closely reflects the setting for an RDI for long tail RD. Interview and prototype demonstrations were used as the main techniques for this part. This complemented the initial requirements identified from relevant projects and RD management-related practices (such as the RD lifecycle).

Metadata is another key RDI component that "glues" RD resources by means of a (structured) description, and supports services in it. This is one of the major challenges in the project, especially when considering that GeRDI focuses on long tail RD. With all the different pilot research communities that participate in the project (see **Table 1**), we propose an approach that maintains a reasonable balance between generic and discipline-specific metadata. In this way, we support key use cases across (and between, such as interdisciplinary ones) communities.

The RDI **architecture** brings all the different perspectives – from requirements identification, to metadata harvesting, to services deployment – in an operating environment. Self-contained services (SCS)[12] as an architectural approach nicely map to and encapsulate our RD lifecycle of choice. It enables services to operate independently from the whole suite in GeRDI, which means that research communities will have a fine-grained access to services in GeRDI. For the case where research communities decide to operate GeRDI locally, they can pick and choose services that apply to their RD lifecycle and do not have to deploy and operate RDI services as a whole, monolithic (software) solution.

We conclude this section with our approach over **operations model and sustainability**. Currently, the project is developing different scenarios and will evaluate various options for the financial sustainability. One scenario is to offer GeRDI with a service fee, to be covered by potential users. Another scenario is to include it as an integral part of other, existing initiatives such as the European Open Science Cloud, within the field of research data management. In this case GeRDI would benefit from the cost model of the respective initiatives. A third scenario considered is to exploit the open source strategy of GeRDI. In this case,

Table 1: GeRDI pilot communities.

Research area	Research community
Social sciences and Economics	- Socio-Economic Panel (SOEP)
Life sciences and Humanities	- Microscopy and Bio-informatics - Digital Humanities - National Center for Tumor Diseases
Marine sciences	- Environmental, Resource and Ecological Economics - Paleoceanography
Environmental sciences	- Alpine Environment Data Analysis Center - Hydrology and River Basin Management - UN International Strategy for Disaster Reduction

[12] http://scs-architecture.org/.

revenues could be generated by consultancy, customer-driven adaptations of the open source services, and the like. During the second phase of GeRDI, planned for 2019–2022 period, these scenarios will be extended and analyzed and a final solution will be provided.

4 Research Communities and Requirements Gathering

4.1 Research Data Lifecycle

One of the initial guides to understand RD practices was its life cycle model. This is the way to conceptualize the stages through which scientific data generally passes in a research process. After exploring multiple options, we adopted the UK Data Archive Data Life Cycle.[13] The model is quite straightforward and consists of 6 phases which we briefly describe here. *Creating Data* phase depicts the creation or collection of data and metadata of interest. *Processing Data* denotes the phase during which created RD is treated or filtered in order to make it usable for the research task at hand. Once completed, interpreting the outcomes and storing them for future reference are depicted in *Analyzing* and *Preserving Data* phases correspondingly. Dissemination related activities for distributing, sharing and ensuring access control to RD is part of phase *Giving access*. The last phase *Re-using data* ensures the secondary usage of RD as results scrutinizing which may precede by new follow up research activities.

Overall, the selected life cycle helped us to better understand the research practices of our pilot research communities and identify some of the core services that an RDI project like GeRDI could provide.

4.2 Pilot Research Communities in GeRDI

Diversity is one of the key descriptors of long tail RDIs. These projects deal with rich flavours of research disciplines, which in turn implies the presence of multiple metadata standards, RD management practices, RD repository solutions, etc. – that impact the requirements determination for the RDI.

Nine research communities across disciplines constitute the pilot communities in GeRDI. In order to present the research discipline diversity, we next provide a brief description of these communities (see **Table 1**). There we (loosely) group the research communities based on a broader research area, just to have it easier to present. We provide a brief description of the research goals for these "clusters" of communities:

- Social sciences and Economics: This community deals with (longitudinal) survey studies, including data about household composition, occupational biographies, employment, earnings, health and satisfaction indicators, etc.
- Life sciences and Humanities: includes datasets (images and image sequences) from genetics; images, text and speech analysis from linguistics; and medical research datasets of clinical studies, epidemiological studies and bio-material collections.
- Marine sciences: The scope includes biological and statistical datasets; as well datasets from meteorology, geo-sciences, and oceanography.
- Environmental sciences: This research "cluster" includes datasets from geophysical and health data, in the narrow fields of climate, glaciology, and radiology. Moreover, in this group we find socio-economic and geophysical datasets from the domain of environmental sciences.

As listed above, these GeRDI pilot research communities present a great case for a long tail RD, and potential for interdisciplinary use cases across disciplines and communities.

4.3 Requirements Gathering

Every pilot community in GeRDI has a **community manager** assigned to it. They are GeRDI project members and serve as a bridge between the project and the research community. From the community perspective, the activities include requirements gathering – features and services that they want GeRDI to support, whereas from GeRDI's perspective, it provides valuable insights and feedback for its functionality.

Once this relationship is established, the process is organized as follows:

- **Interviews:** Interviews are the technique of choice for requirements gathering process. Community managers conduct the process according to agile practices: iteratively – they interview users over a longer period (multiple times), and incrementally – they continuously refine existing and/or add

[13] http://www.data-archive.ac.uk/.

new requirements. Interviews have two roles, depending on the project phase: in the beginning, they elicit requirements from users; since GeRDI v.1 release, however, are used to gather feedback and further requirements via GeRDI demo sessions that community managers organize.

- **Requirements specification – "Refining" requirements:** Based on the interview session recordings and notes, community managers identify the initial use cases. At this point, user requirements start to form in more general features that users want to see in GeRDI. As an example, research communities require basic search functionality that users are accustomed to. From search services in general; a minimal selection functionality that would enable a research to choose resources of interest after a search; or a possibility to narrow down search results by certain parameters. During this activity, there could be other requirements that community managers identify, not specifically required from the users, such as user authentication.
- **Use case modeling:** To further complement the requirements from the previous step, community managers model the requirements via use cases. In addition to the main scenario, community managers are able to set (implementation) priorities for each use case, and decide to either propose it for implementation as a whole, or slice it in smaller functional units (use case slices) to be implemented as independent feature units. During this process we also developed "throw-away prototypes" to better showcase the different GeRDI features to the users.
- **Mapping use cases to GeRDI services:** In this final step, community managers coordination with their corresponding communities, map the use cases with services to be implemented in GeRDI.

With requirements identification completion, one of the main challenges that we face in the project is of specifying the metadata harvested from the research communities in a standard metadata schema.

5 Describing Resources in GeRDI: Metadata

Typically in RDI projects a schema contains (a minimal set of) common elements across RD communities, based on which services are then developed. This approach balances well between lowering the effort to describe the datasets and services offered for the user communities. When we deal with long tail RD, however, this approach leaves disciplinary metadata uncovered (metadata are not harvested, thus there are no services that rely on them). As a result, users benefit a minimal (very generic) set of use cases from the RDI, i.e., lack any disciplinary use cases. This limitation could seriously affect RDI user adoption, as mentioned in the (e IRG, 2016) report.

5.1 GeRDI Metadata Schema

From the metadata management perspective, we take a different approach for a metadata schema for GeRDI. GeRDI Schema consists of 3 parts that map to different requirements categories in GeRDI, meant to provide a better support for the research communities (from the metadata standpoint). It is worth mentioning that the schema is still under development as we collect disciplinary metadata from our communities. **Figure 1** presents the conceptual diagram of the schema, and following is the rationale behind each part:

- Generic Part: includes typical bibliographic metadata, suitable to support RDI services that are generic in nature (title, author, publication year, etc.). Generic metadata usually contain a

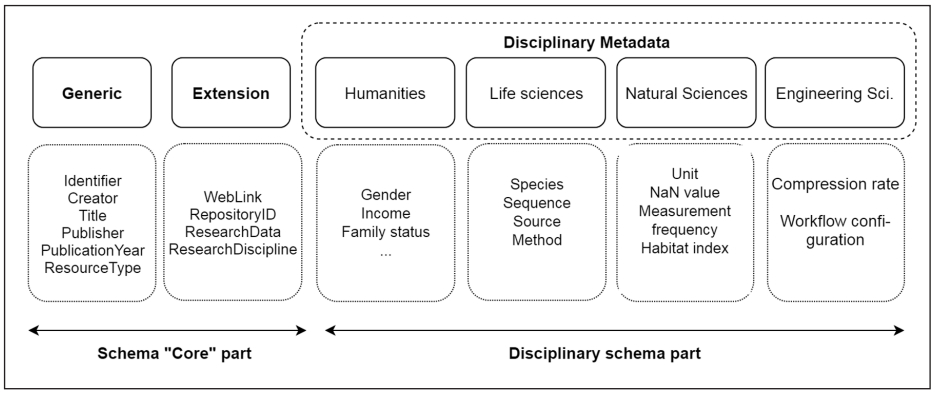

Figure 1: GeRDI Metadata Schema, a conceptual diagram.

smaller – but stable – number of elements. For this part we opted for schema reuse and DataCite[14] – a well-established and popular metadata standard that incentivizes RD exchange and citation.

- Operational Part: as an infrastructure, GeRDI requires certain metadata elements to support its operations, such as identify harvested resources, track resource origin, research discipline, different (URI) access links (download, view, etc.), and so on. Its role in metadata harvesting and maintenance and in general RDI services support is crucial. "Extension" box in **Figure 1** shows metadata elements for this part. In order to provide more details, **Table 2** lists the properties of this part that contains the property ID, name, obligation status (Mandatory – M; Recommended – R; and Optional – O), occurrence of the property, and a brief definition.
- Disciplinary Part: contains metadata that are specific for research disciplines. This is the most challenging part of the schema, one of the key factors that impacts both user requirements and supported services in GeRDI. "Disciplinary metadata" box in **Figure 1** shows metadata organized according to subject areas from the German Research Foundation.[15] The metadata elements provided for each area present examples of how this part of the schema could be structured. Requirements for this part are still under way.

As illustrated in **Figure 1**, the first two blocks form the GeRDI "core", whereas the rest forms the "disciplinary" part of the schema. This schema design provides capabilities for support of generic services – a more straightforward and "must have" component in an RDI. Morevoer, it accommodates the disciplinary metadata (to the extent possible) as well – thus provide support for disciplinary services which are of great importance to the (participating) research communities.

We expect this design to support a set of key services across disciplines (despite RD metadata diversity), as well as a set of more discipline- or community-specific ones (based on the second part of the schema). In this way we balance the breadth and depth of RD metadata in order to provide interesting prospect for researchers. For example, a discovery service could enable search functionality to users based on schema "core" elements, such as author, title, or publication year, to name a few; and a recommendation functionality based on relevant disciplinary metadata elements in the "disciplinary" elements to enrich the discovery process of users.

5.2 GeRDI Schema in Practice

As is the case throughout the requirements gathering process, research communities via their use cases weigh in on the disciplinary metadata to be considered in GeRDI Schema. This enables a GeRDI harvester implementer to specify the mapping between community metadata and GeRDI Schema. One would map as many metadata as possible, especially considering the limitations for disciplinary metadata.

Table 2: GeRDI generic extension metadata set.

ID	Property	Obl.	Occ.	Definition
1	WebLink	R	0–n	A string that identifies a resource in GeRDI, via a (name, URI) pair.
1.1	webLinkName	O	0–1	String value denoting the name of the link (sth that the user would see during browsing).
1.2	webLinkURI	M	1	The (access) URI of the resource.
1.3	webLinkType	R	0–1	The type fo the weblink.
2	RepositoryIdentifier	M	1	A unique human readable string that identifies the source repository.
3	ResearchData	M	1–n	A downloadable file from the source repository.
3.1	researchDataIdentifier	M	1	A universal unique identifier for the file.
3.2	researchDataURL	M	1	File download URL.
3.3	researchDataLabel	R	0–1	A human readable name of the file.
3.4	researchDataType	R	0–1	The file type of the research data.
4	ResearchDiscipline	R	1–n	The research discipline (s) of the data set.

[14] http://datacite.org.
[15] http://www.dfg.de/en/dfg_profile/statutory_bodies/review_boards/subject_areas/index.jsp.

Let's depict this process through one of our communities from the social sciences. Based on their feedback, we have selected 2 disciplinary metadata elements – variables and concepts that describe the variables – which we add to the disciplinary part of GeRDI Schema (a similar effort to identify and include metadata is currently being applied for the other pilot communities). When harvesting metadata for this community, in addition to the "core" schema part, one will be able to include these two disciplinary metadata elements.

Every metadata instance harvested is stored as a JSON document (in a MongoDB instance) that feeds the GeRDI search index (based on Elasticsearch). The structure of the search index maps the GeRDI Schema, including the disciplinary metadata part. The social sciences community in GeRDI can now use the disciplinary elements during GeRDI activities, such as search, bookmark, storage, etc. (see Section 6.2 for more details on GeRDI services).

GeRDI services (mainly) rely on available metadata to operate, although future work considers external resources to be used in addition. The schema support goes beyond just GeRDI services. The user interface is another part that directly depends on the underlying metadata schema. As an example, after an initial search, a user can decide to filter the results based on a certain (disciplinary) criteria. In order to enable this capability, we need to know what all the available metadata – including the disciplinary part of GeRDI Schema – are.

6 System Architecture Design

Architectural design decision are always critical for the success of an infrastructure project as they provide the necessary backbone where all components come together to provide user services. Specifically, we need an architecture design that closely maps services delivery to the research life cycle and accommodates the changing requirement of our research communities, while keeping the level of complexity as low as possible. Based on our requirements and design goals, the self-contained systems or micro service architecture seemed the most appropriate solution for GeRDI (de Sousa, Hasselbring, Weber, & Kranzlmüller, 2018). Importantly, this architectural approach is also compliant with the service marketplace of the European Open Science cloud (EOSC) which will provide us leverage for future services co-operations. A self-contained system (SCS)[16] emphasizes on the separation of major functionalities to many independent components to avoid the problem of complexity. Each self contained component is an autonomous web application with its own interface, business and logic layer. In nutshell, an SCS can fulfill its primary use cases on its own, without having to rely on other systems being available.

6.1 Mapping GeRDI Services to RD life Cycle

The workflow of identified research usecases during the requirement gathering phase provide us with enough feedback to identify potential services in GeRDI. As expected for our diverse RD workflows and collection, our communities expressed different preferences with regard to services that GeRDI should provide. We analyzed and clustered the research use cases based on required functionalities and mapped these services to our RD life cycle phases. The mapping of GeRDI services to RD life cycle is illustrated by **Figure 2**.

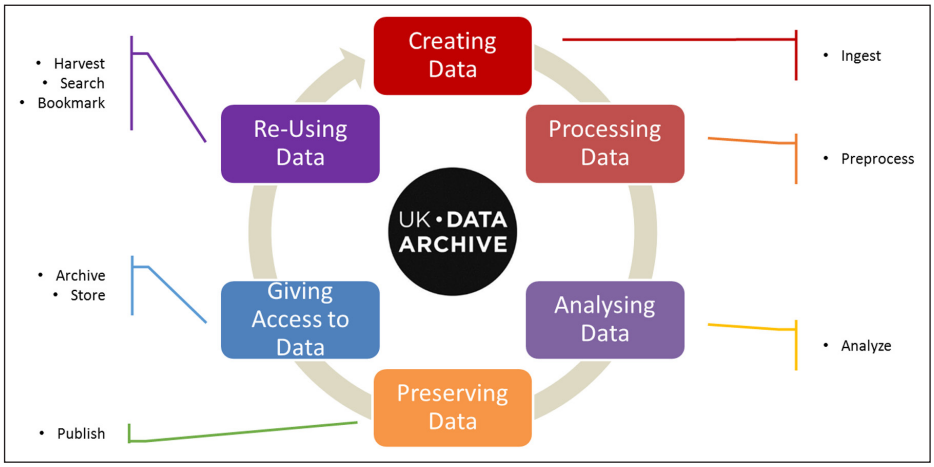

Figure 2: Mapping of GeRDI services to RD life Cycle.

[16] http://scs-architecture.org/.

6.2 GeRDI Services

The final set of services is shown in **Figure 3**. Each colored box depicts a (vertical) service domain that supports a part of the research data life cycle. Based on the workflow of researchers in our heterogeneous communities the identified services are further sub-categorized into core and extended services. Before we move on with the (vertical) services, it is important to mention authentication and authorization infrastructure (AAI) service that cross-cuts through and supports many of the SCS services. This service enables researchers to use their institutional credentials for different services in GeRDI and when inevitable support the communication between SCS services. Authentication and authorization infrastructure (AAI) service is deployed within a back-end integration layer. This rather horizontal service will introduce the concept for multi-protocol, single sign-on among the GeRDI services based on REST interfaces.

6.2.1 Core Services

The core services components are implemented and operated by the GeRDI team; these services set consist of:

- Harvest: this service provides an interface to access the metadata from community research data repositories and to ingest that in GeRDI infrastructure index. It also enriches the harvested metadata and forwards it to a search index. Currently, there are more than 377K datasets across communities harvested in GeRDI. To further enrich GeRDI index and to proof scalability (1 mio entries) from Zenodo respository are also harvested.
- Search: the search service is intended to facilitate researchers to find research data over searched queries. Special features like Geo based searching, filtering and recommendation will be part of this service as well.
- Bookmark: the bookmark service helps researcher to create a persistent selection of search results. In addition, it also holds a data recipe of which search terms were used to get to the selection and how was the data selected.

Figure 4 shows the running prototype of implemented GeRDI services, with **Search** and **Bookmark** already operational. Here you can see a typical search service, accompanied with filtering capabilities, such as collection repository, publication year, author, etc. Note that the harvesting service is also implemented, but that will not be available directly to users.

6.2.2 Extended Services

The extended Service components will be provided as a reference implementation to the interested project communities. this services set consist of:

- Store: this service helps researchers to download the bookmarked data either to a local machine or a remote storage system.
- Preprocess: helps researchers to normalize, filter and even preview the data before the analysis phase.
- Analyze: is a service to provide actual analysis on the pre-processed data to gain new scientific insights.
- Publish: provide publication and ingestion of newly generated research data into GeRDI accessible repositories.
- Archive: depicts the repository services for long-term research data archival.

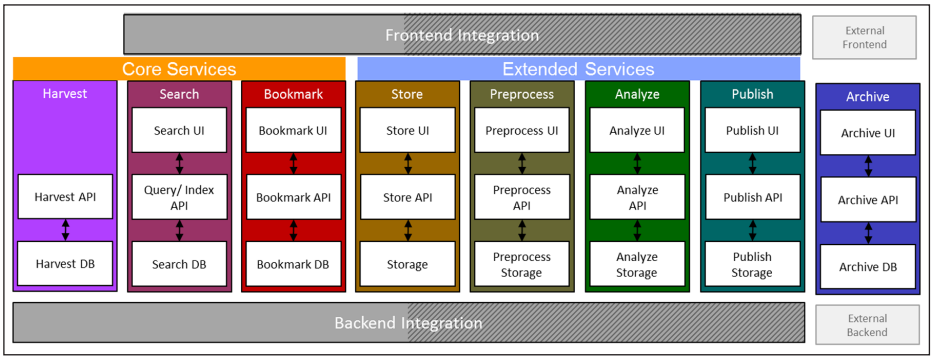

Figure 3: GeRDI services: A Self-Contained Systems Architecture view.

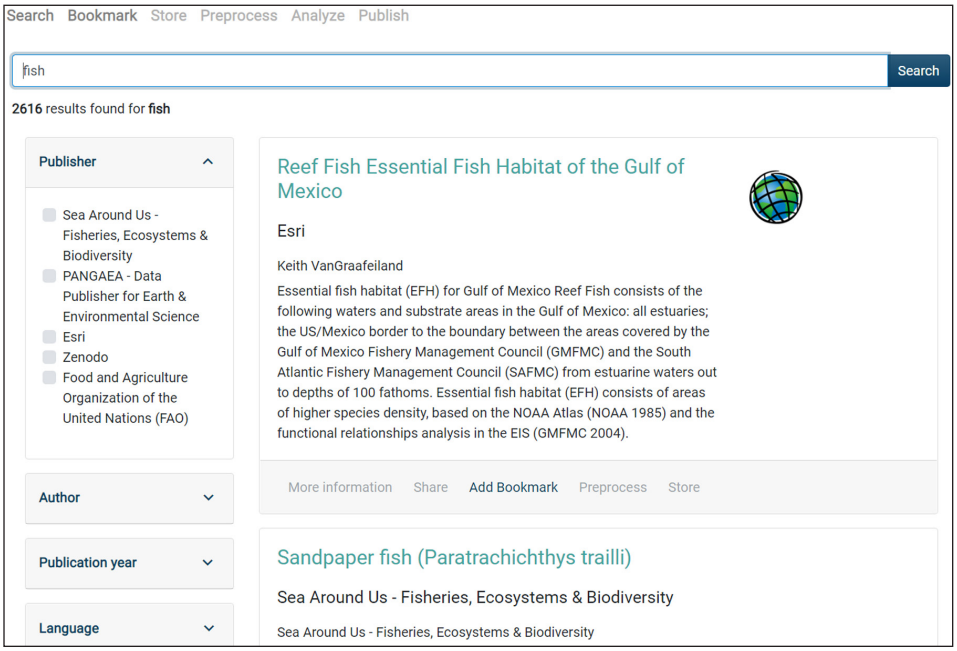

Figure 4: GeRDI services example: Search and Bookmark.

7 Challenges

From the **metadata standpoint**, there are few challenges that we need to address in the long run. Long tail RD does not have something like a central authority that can provide (final) requirements for a RDI to follow. As a result, pilot communities in GeRDI proved instrumental in driving this process both for metadata, and services aspects of the project. This, however, requires more effort and bigger engagement for the communities. On another note, maintaining a balance between metadata breadth and depth is a continuous theme in such a project. Namely, as new communities join GeRDI, or current ones propose new use cases, new (disciplinary) metadata will be proposed to be part of GeRDI Schema. This will require us to decide to what extent can we support this without breaking any of the GeRDI services, or limiting the corresponding UI for users. To this end, we are closely monitoring initiatives that deal with (meta)data type standardization, such as the RDA Data Type Registries Working Group,[17] and similar organizations. Finally, ontology application for both pilot communities (to improve search and discovery services, for example), and ones that could be applied at "GeRDI level" (generic in nature, supporting infrastructure operations) is another challenge/opportunity we are considering for GeRDI.

From the **architectural design standpoint**, at long run RDI may undergo a changing set of stakeholders with inclusion of new communities and therefore possibility of new researcher work flows with shifting requirements. Hence, the architectural design must be flexible and able to fulfill changing requirements of the heterogeneous communities. In addition, services architecture also needs to be reusable and shareable to other researchers and infrastructures to increase not only the impact of their research efforts, but also to make the research process more efficient, transparent, and reproducible. We have provided the solution of self contained system to tackle this challenge and with the implementation of all the service (end of GeRDI Phase 1) will be able to test the efficiency of our provided solution.

From the **competitiveness standpoint**, We are closely monitoring RDI initiatives – in different scopes, sizes, foci; commercial and non-commercial, etc. – to ascertain the unique value proposition for GeRDI. We target different aspects of these initiatives, such as: different delivery platforms (EUDAT,[18] RADAR, Google Dataset Search,[19] Mendeley Data,[20] etc.), standardization bodies (RDA Alliance,[21] etc.), as well as policy regulators (ESFRI,[22] etc.), that together shape the landscape of RDIs.

[17] https://rd-alliance.org/groups/data-type-registries-wg.html.
[18] https://eudat.eu/.
[19] https://toolbox.google.com/datasetsearch.
[20] https://data.mendeley.com/.
[21] https://www.rd-alliance.org/.
[22] https://www.esfri.eu/.

8 Conclusion

Research Data as a knowledge resource ensures trace-ability of existing scientific results and provides potential for more insights, if disseminated with and/or across research communities. As some of the impactful research happens at the cross roads of different disciplines, we consider infrastructure for long tail RD as an enabler for the modern (inter) disciplinary research initiative and practices. While infrastructure projects that target big or discipline specific RD are more prevalent, there is a considerate RD from the long tail of research – which are heterogeneous in nature, produced at "low scale" – that doesnot reside on RD repositories or infrastructure. It does not adhere to RD documentation and maintenance, or lacks the typical services for the different phases of a research lifecycle, etc. The GeRDI project attempts to fill this gap by providing a generic, federated research infrastructure for the long tail of RD. It deals with the RD from different domains, research practices and standards. Furthermore, it identifies and implements multidisciplinary use cases to ultimately provide services for researchers throughout RD lifecycle.

Acknowledgements

This work was supported by the DFG (German Research Foundation) with the GeRDI project (Grants No. BO818/16-1 and HA2038/6-1). All project partners i.e., ZBW – Leibniz Information Center for Economics, CAU – Kiel University, TUD – Technical University of Dresden, LRZ – Leibniz Supercomputing Centre Munich and DFN – German National Research and Education Network have contributed in the realization of GeRDI project. Specifically, the software architecture has been designed by the software engineering group of Prof. Wilhelm Hasselbring from Kiel University. TUD contributed in design of metadata harvester and results. ZBW contributes in requirement gathering and metadata schema on going development. LRZ supports the operation and infrastructure issues during the development and provide the pilot operation while DFN is working for the project sustainability with operational model.

Competing Interests

The authors have no competing interests to declare.

References

Bobby Vocile, W. 2017. Open science trends you need to know about. https://hub.wiley.com/community/exchanges/discover/blog/2017/04/19/open-science-trends-you-need-to-know-about.

Brophy, E and **Razum, M.** 2017. Radar: A research data management repository for long tail data. *Tage 2017*, 23.

Buckland, M. 2011, aug. Data management as bibliography. *Bulletin of the American Society for Information Science and Technology*, 37(6): 34–37. DOI: https://doi.org/10.1002/bult.2011.1720370611

de Sousa, NT, Hasselbring, W, Weber, T and **Kranzlmüller, D.** 2018. Designing a generic research data infrastructure architecture with continuous software engineering. In *3rd workshop on continuous software engineering (cse 2018)*.

e IRG. 2016. Long tail of data, e-irg task force report 2016. http://e-irg.eu/documents/10920/238968/LongTailOfData2016.pdf.

Grunzke, R, Adolph, T, Biardzki, C, Bode, A, Borst, T, Bungartz, HJ, et al. 2017. Challenges in creating a sustainable generic research data infrastructure. *Softwaretechnik-Trends*, 37(2): 74–77.

Hey, T. 2012. The fourth paradigm – data-intensive scientific discovery. In *Communications in computer and information science* (pp. 1–1). Berlin, Heidelberg: Springer. DOI: https://doi.org/10.1007/978-3-642-33299-9

Horstmann, W, Nurnberger, A, Shearer, K and **Wolski, M.** 2017. Addressing the gaps: Recommendations for supporting the long tail of research data. https://www.rd-alliance.org/groups/long-tail-research-data-ig.html. DOI: https://doi.org/10.15497/RDA00023

Linne, M and **Zenk-Möltgen, W.** 2017, mar. Strengthening institutional data management and promoting data sharing in the social and economic sciences. *LIBER QUARTERLY*, 27(1): xx–xx. DOI: https://doi.org/10.18352/lq.10195

Pampel, H, Vierkant, P, Scholze, F, Bertelmann, R, Kindling, M, Klump, J, Dierolf, U, et al. 2013, nov. Making research data repositories visible: The re3data.org registry. *PLoS ONE*, 8(11): e78080. DOI: https://doi.org/10.1371/journal.pone.0078080

Smith Rumsey, A. 2010. Sustainable economics for a digital planet: ensuring long-term access to digital information: final report of the blue ribbon task force on sustainable digital preservation and access.

Starr, J, Ammann, N, Ashton, J, Barton, A, Elliott, J, Jacquemot-Perbal, MC, Ziedorn, F, et al. 2015, 08. Datacite metadata schema for the publication and citation of research data. DOI: https://doi.org/10.5438/0010

Wilkinson, MD, Dumontier, M, Aalbersberg, IJ, Appleton, G, Axton, M, Baak, A, Mons, B, et al. 2016, mar. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. DOI: https://doi.org/10.1038/sdata.2016.18

How to cite this article: Latif, A, Limani, F and Tochtermann, K. 2019. A Generic Research Data Infrastructure for Long Tail Research Data Management. *Data Science Journal,* 18: 17, pp. 1–11. DOI: https://doi.org/10.5334/dsj-2019-017

Submitted: 29 June 2018 **Accepted:** 16 April 2019 **Published:** 08 May 2019

]u[*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS