GeRDI
Generic Research Data Infrastructure

Manifesto of the GeRDI GO FAIR Implementation Network

Introduction:

The project deals with the development of a Generic Research Data Infrastructure (GeRDI) in Germany. The aim is to enable scientists in Germany, especially those who hold only small amounts of data, to store, share and re-use research data across disciplines.

In a first phase, three pilot data centres supporting the management of research data will be linked up with each other so that research data can be used across disciplinary boundaries, enabling new opportunities for multi-disciplinary research.

In a second phase, the developed solution can be rolled-out in Germany and – if appropriate funding will be available – serve as a model for future Research Data Infrastructures in Germany and beyond. GeRDI will be able to support universities and research institutes in providing research data, in linking up their existing data stores and in establishing new research data stores.

Purpose of the Implementation Network:

Available data and efficient infrastructures are important success factors for the scientific knowledge process. This requires a sustainable system of integrated research data infrastructures, offering both reliable working structures for researchers as well as enabling further development.

Various initiatives are currently being funded to develop an interconnected research infrastructure, ensuring access and providing services for successful research data management throughout the data lifecycle towards the Internet of fair data and services. However, standards for metadata, service catalogues or persistent identifiers, to mention just a few, do not exist yet. Against this background, the GeRDI project will develop a pilot infrastructure for Germany that links existing and future research data repositories via a federated system, so that research data can be **found**, **accessed** (if the license allows it) and **reused** (if the data is documented accordingly and provided in a format supporting reuse). With GeRDI the researchers do not need to know in which repositories the required research data is stored or know about repositories that could potentially be considered for a query regarding their current research question.

GeRDI will enable universities and non-university research institutions to actively accept the task of providing and working with research data. The developments aim at researchers with moderate data volumes and at associated research communities, who do not yet have a pronounced degree of organization.

Various research communities are involved in GeRDI to identify the very heterogeneous needs of the most diverse disciplines. Currently, social sciences, economics, marine sciences, geosciences, environmental sciences, medical research as well as linguistics and ethnology contribute to the development of GeRDI. The exact paths that research data take during a research process are being investigated. The derived case studies get assigned to the different phases of the data lifecycle and result in the overall architecture and software concept.

According to the project's mission and in line with the FAIR principles, GeRDI will provide generic, sustainable and open software connecting research data repositories to enable multidisciplinary and FAIR research data management. These services and this software will be based on common standards and be developed in close collaboration with various research communities to ensure the best match to the requirements of different disciplines. GeRDI will promote a wide usage of its software and thus contribute to establishing an active GeRDI community which will continue to flourish beyond the lifespan of the project. All project results, particularly software, training support and business model, will form a German contribution to the European Open Science Cloud.

Overarching Principle of Operation

We commit to comply with the Rules of Engagement[1] of GO FAIR Implementation Networks.

Targeted Objectives for the Internet of FAIR Data & Services (IFDS):

1. Central Index

We offer knowledge and technology to build up a central index containing metadata from connected repositories.

In GeRDI, research data repositories will be harvested to build a central index containing metadata from the connected repositories. In this way we enable research communities to make their data findable.

Therefore, an extensive analysis of the research data landscape and metadata usage was undertaken. Very different formats of metadata (e.g. Dublin Core, DataCite, ORE, Mets) and interfaces for access (e.g. OAI-PMH, OpenData) have been identified in various existing research data repositories. Even within standards there is a very strong diversification in both the structure of metadata and the implementations of software solutions.

The research data repositories of our research communities were first connected via case-specific software solutions for collecting metadata ('harvesting'). In a second step - with the aim to set up a standard connection of these research data repositories to GeRDI - a generic software package for collecting the federated metadata was implemented prototypically. The implementations are based on the standard protocol OAI-PMH and a harvesting library to encapsulate micro service-compliant functionalities.

[1] https://www.go-fair.org/implementation-networks/rules-of-engagement/

At this stage, the metadata of three research data repositories is being transferred into a common search index:

- ENA (www.ebi.ac.uk/ena)
- FAOSTAT (www.fao.org)
- Pangaea (www.pangaea.de)

Currently, the "Elasticsearch" technology best meets the search requirements of the project.

To cover usability, adaptability and reusability, an architecture for the organization of the harvesters has been designed. This architecture is integrated into the overall architecture and is responsible for collecting and managing metadata as well as for controlling the required software components. The architectural component used to build the search index enables third parties to access the index via an API.

2. Architectural Design - Micro services

We offer knowledge in architectural design and developing microservice architecture. Also, we offer existing technologies for reuse.

GeRDI is being designed with a micro service architecture. The developments cover both the requirements raised by the research communities as well as fundamental non-functional requirements. The individual working steps of researchers are represented as abstract services. Currently, the architecture consists of eight services:

Harvest – Search – Bookmark - Store - Pre-process – Analyse - Publish - Archive

The services communicate via open interfaces, which allow it to integrate external services and other research infrastructures (e.g. high-performance data centres).

Harvest, Search and Bookmark offer generic functions (e.g. searching or collecting metadata) while the other services focus on discipline-specific functions (e.g. the analysis of research data or the publication of new data).

3. Metadata Schema across Disciplines

We are willing to share our expertise in developing an interdisciplinary metadata scheme for research data.

Due to the generic nature of GeRDI, an appropriate metadata standard is needed to underpin the cross-disciplinary efforts in describing research data and building services on top of it.

In GeRDI both generic (e.g. Dublin Core, DataCite, CERIF, DCAT, etc.) and discipline-specific (e.g. DDI, SDMX, etc.) metadata standards of our research communities have been evaluated. The DataCite schema proved to be the best starting point for the GeRDI context.

Furthermore, multidisciplinary RDI requirements dictated a metadata extension of extra elements (e.g. elements for identification in GeRDI, specification of the source repository, research data details and research data discipline. Together with the DataCite schema, these extensions form the core GeRDI schema.

Metadata, such as discipline-specific elements that cannot be represented via the core part (e.g. including metadata about provenance), is handled by a discipline-specific schema part. This is maintained during metadata entry to ensure the integrity of metadata and avoid name conflicts between similar metadata from different subject communities.

A clear flow of metadata between the services in GeRDI clarifies the systems architecture, as well as guides 3rd party implementers regarding service usage and adoption.

4. Community Requirements

We provide the methodology to conduct user requirements studies and the results of the requirements studies conducted in GeRDI.

The research communities involved in the GeRDI project can be characterized as very heterogeneous and several rounds of interviews are being conducted to identify their requirements on a multidisciplinary RDI. The results provide insights into concrete research questions and the workflows required to answer these research questions.

The use cases have been elaborated in detail along the data lifecycle of the UK Data Archive and form the basis for the GeRDI requirements specification. Up to now, four generic and 13 discipline-specific use cases were combined into case studies.

The case studies have been implemented as mock-ups (click prototype). They visualize the research processes and can be used for concrete discussions with research communities on RDM.

In future there is the potential that discipline-specific use cases develop into generic use cases that apply across the disciplines.

Membership list:

We consider this Manifesto to be one way by which the undersigned stakeholders **can speak with one voice** on several critical issues that are of generic importance to the objectives of FAIR, and on which we feel we have reached consensus.

Project Partners:

- Professor Dr. Klaus Tochtermann
  ZBW – Leibniz Information Centre for Economics

- Professor Dr. Dieter Kranzlmüller
  Leibniz Supercomputing Center of the Bavarian Academy of Sciences and Humanities

- Professor Dr. Wolfgang E. Nagel
  Technische Universität Dresden (TUD),
  Centre for Information Services and High-Performance Computing (ZIH)

- Professor Dr. Wilhelm Hasselbring
  Christian-Albrechts-University of Kiel (CAU), Software Engineering Group

- Prof. Dr. Hans-Joachim Bungartz
  Verein zur Förderung eines Deutschen Forschungsnetzes e. V.

Community Partners:

1. Environmental, Resource and Ecological Economics, Kiel University
2. GEOMAR, Helmholtz Centre for Ocean Research Kiel
3. German Socio-Economic Panel (SOEP),
   at German Institute for Economic Research (DIW Berlin)
4. Hydrology and River Basin Management, TUM Department of Civil, Geo and Environmental Engineering, Technical University of Munich
5. Alpine Environmental Data Analysis Center
6. VerbaAlpina, Ludwig-Maximilians-Universität München
7. Digital Humanities, University of Leipzig
8. Max Planck Institute of Molecular Cell Biology and Genetics
9. National Center for Tumor Diseases (NCT)

Timeline:

The GeRDI project runs from November 1, 2016 – October 31, 2019 with an option for extension for another three years. The design, development and implementation processes on the targeted objectives happen continuously:

- Open research data repositories are harvested to build a prototype for a central index.
- A concept of micro service architecture is being designed and will be implemented along the data life cycle.
- An interdisciplinary metadata scheme has been developed.
- Interviews along the data life cycle to identify community requirements and to set up use cases are conducted throughout the whole project period.

Within the GeRDI project the open science principles are applied. The software is made available in open source and publications are published in open access. Pertinent expertise and code can be continuously shared and extended.

Date:

June 2018

Signed by: