# Research Data Management at LRZ and beyond

Megi Sharikadze,

20.04.2018 | Plan-E Meeting, Paris

# LRZ - Leibniz Supercomputing Centre, Bavarian Academy of Sciences and Humanities

HPC

Network

Storage

Virtual Reality & Visualisation

Training

Know-How

Consulting

Support


The LRZ has been operating world-class supercomputers for decades. The current supercomputer, the SuperMUC (Phases 1 and 2), is one of the most powerful computers in the world. With a peak performance of 6.8 Petaflops (almost 7 quadrillion operations per second), 500 Terabyte main memory, 20 Petabyte external data storage, and a high speed interconnect, the SuperMUC provides first-class information technology for researchers in the fields of e.g. physics, chemistry, life sciences, geography, climate research, and engineering.

# HPC and Big Data Competence Centre

**SuperMUC Phase 1**
- 3.2 PFlops performance
- 147.456 Sandy Bridge CPU cores

**SuperMUC Phase 2**
- 3.6 Pflops performance
- 86.016 Haswell CPU cores

**Bavarian Big Data Competence Centre**
- Compute Cloud, DGX-1, GPU-Cloud

**SuperMUC-NG**
- 26.7 Petaflop cluster
- 6500 nodes, ThinkSystem SD650 servers
- 45 percent greater electricity savings cf.  air-cooled system

**CoolMUC-3**
- warm-water cooled racks, with inlet temperature of at least **40 °C**.
- With 4.96 GFlops/Watt (according to the strict Green500 level-3 measurement methodology) CooLMUC-3 is **one of the most efficient x86 systems worldwide.**

# Storage and Archival Systems

**Altogether 200 Petabyte Storage**

- HPC Cluster Data systems
- Data Science Storage
- Archive and Backup systems
- Bayern Cloud

# Research Data: Reproducibility and Re-use

Main problem:

Research data is hard to reproduce and re-use

Challenges:

- Missing documentation und description (Metadata)
- Scientific results are hard to cross-check
- Heteregeneity and missing standards
- Repetition of research studies?
- Costs arise in terms of time and Infrastructure

Technical challenges

  - Interfaces (Interoperability)
  - Formats and Metadata standards
- Workload und costs
- No clearly defined responsibilities and structures
- Saving of sensitive data
- Intelectual property rights

# Research Data Lifecycle: Creating Data

- Design Research
- Plan Data Management (formats, storage, etc.)
- Consent for data sharing
- Locate existing data
- Collect data (experiment, observe, measure, simulate)
- Capture and create metadata


UK Data archive

# Research Data Lifecycle: Processing Data

- Enter data, digitise, transcribe, translate
- Check, validate, clean data
- Anonymise data where necessary
- Describe data
- Manage and store data


UK Data archive

# Research Data Lifecycle: Analysing Data

- Interpret data
- Derive data
- Produce research outputs
- Author publications
- Prepare data for preservation


UK Data archive

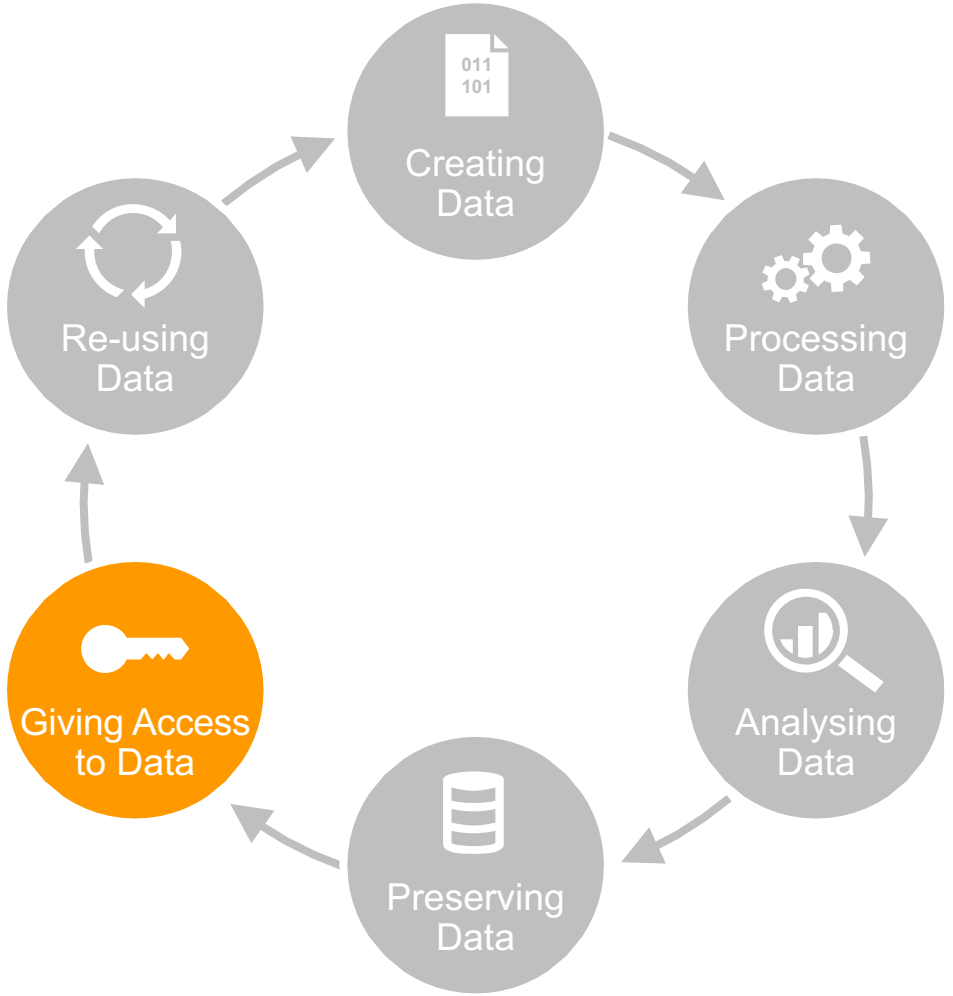
# Research Data Lifecycle: Preserving Data

- Migrate data to best format
- Migrate data to suitable medium
- Backup and store data
- Create metadata and documentation
- Archive data


UK Data archive

# Research Data Lifecycle: Accessing data

- Distribute data
- Share data
- Control access
- Establish copyright
- Promote data




UK Data archive

# Research Data Lifecycle: Re-using data

- Follow-up research
- New research
- Undertake research reviews
- Scrutinise findings
- Teach and learn

# Research Data Management (RDM) and Infrastructures

## Research data Management

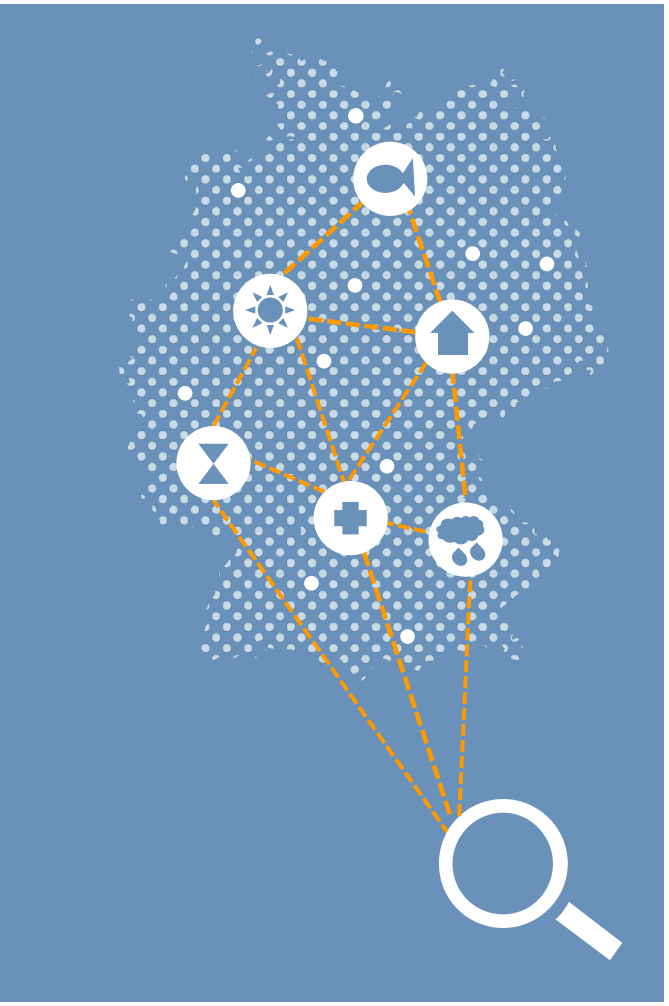
„all activities which deal with preparation, storage, achiving and publishing of Research Data“[1]

## Creating and connecting RD infrastructues

Personal, Services and Tools for support of RDM activities

**1** Simukovic, E., Kindling, M., & Schirmbacher, P. (2013); Umfrage zum Umgang mit digitalen Forschungsdaten an der Humboldt-Universität zu Berlin

# Generic Research Data Infrastructure – GeRDI (DFG funding)



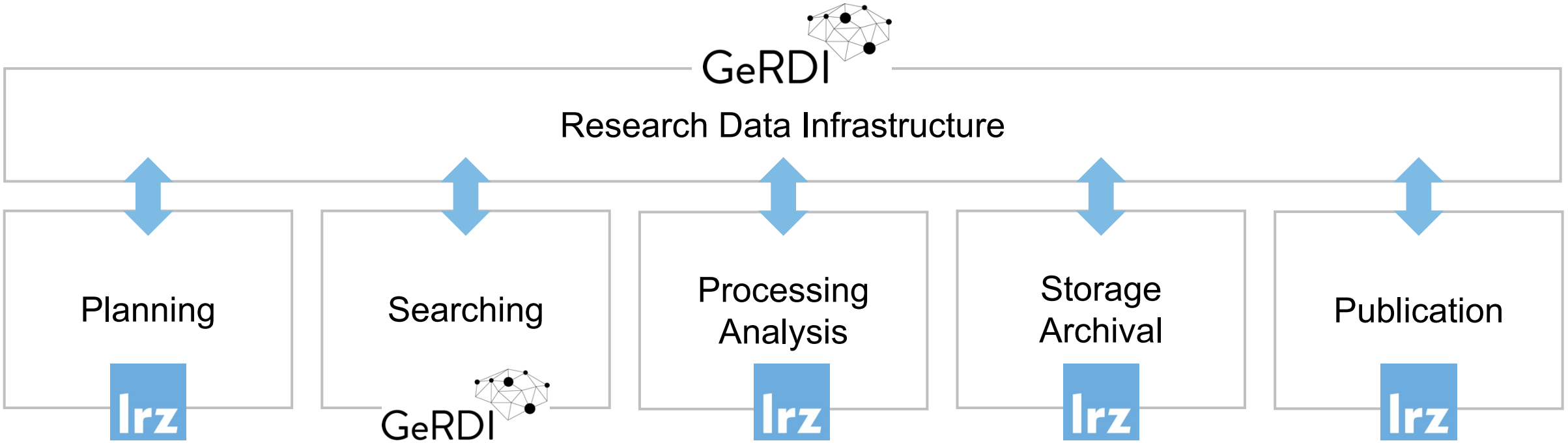
- Aim: to build RD infrascructure in Germany
- First Phase: 2016 - 2019
- Based on FAIR Principles
  **F**indable | **A**ccessible | **I**nteroperable | **R**e-usable
- Inclusion of new communities in development
- Model structure und achieve connectability with national and European RD infrastructures

Project coordinator: Prof. Dr. Klaus Tochtermann

The German National Library of Economics (ZBW – Leibniz Information Centre for Economics)

# RDM with LRZ and GeRDI

- Integration of RD repositories with LRZ services
- Uniting with already existing LRZ services
- Supporting RDM Workflows



# German Council for Scientific Information Infrastructures (RfII)

Rat für Informations Infrastrukturen

TO FACILITATE AND ACCELERATE DIGITAL TRANSFORMATION IN THE SCIENCES ON A NATIONAL LEVEL

**The Council monitors transitions in the German academic system at large and gives practical recommendations to academia and the government. Specifically it**

- provides foresight on the **development of digital science**;
- promotes **coordination** of existing activities;
- identifies **potential synergies** between the diverse actors and new fields of action;
- intends to stimulate **cooperation** within the academic system;
- monitors **international policy** developments.

UNION DER DEUTSCHEN AKADEMIEN DER WISSENSCHAFTEN

**Positionspapier**

**der Union der deutschen Akademien der Wissenschaften zur Schaffung einer Nationalen Forschungsdateninfrastruktur (NFDI)**

Die Wissenschaftsakademien in Deutschland sind sowohl Gelehrtengesellschaften als auch Forschungseinrichtungen. Die Forschung der Akademien ist überwiegend geisteswissenschaftlicher Art. Dabei kommt den Akademien gerade dadurch, dass sie Langzeitforschungsprojekte durchführen, eine besondere Stellung innerhalb der geisteswissenschaftlichen Forschung zu. Mit ihren Editionen, Wörterbüchern, Lexika und Erschließungsprojekten leisten die Akademien wichtige Beiträge zur geisteswissenschaftlichen Grundlagenforschung. Schon vor Jahren haben die Akademien hierzu spezielle IT-Fachabteilungen aufgebaut bzw. betreiben sogar große wissenschaftliche Rechenzentren wie das Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften, die unter anderem für die Bereitstellung, Vernetzung, Pflege und Sicherung von Forschungsdaten und elektronischen Publikationen Sorge tragen. Indem die Akademien im eigenen Hause (und teilweise auch darüber hinaus) Infrastrukturdienste anbieten und damit de facto auch als Datenzentren fungieren, sind sie sowohl mit wesentlichen Anforderungen aus der Nachfrageperspektive als auch mit den Herausforderungen aus der Angebotsperspektive vertraut.

# Summary

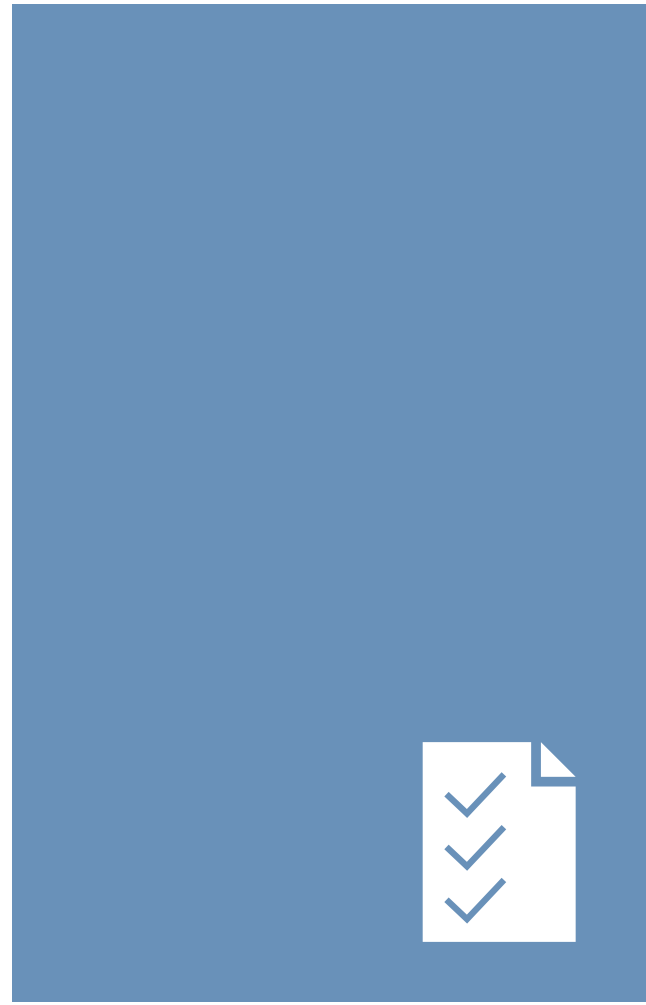
- RDM is an important aspect of sustainable research
- Data and computing centres are important partners in RDM
- RDM at LRZ is combination of LRZ und GeRDI services
- Available Interfaces guarantee federated RDM

## Future steps (LRZ)

- Proceed with GeRDI Phase 2 and expand in whole Germany
- Provide training and consulting on RDM related themes
- RD repository for FAIRe data storage

## National level

- National Research Data Infrastructure (Nationalen Forschungsdateninfrastruktur, NDFI)



EnviroInfo2$^5$ www.enviroinfo2018.eu

5-7 September 2018, Munich, LRZ


## Workshop 6

### Open data and models – dealing with uncertainties

Transparency is essential for repeatability and thus of vital importance to the practice of science. This is an argument for using open data and open models. In this special session, we will discuss the strengths and weaknesses of existing models, which questions can or cannot be answered through simulations using those models, what forms of uncertainty exist, and how these uncertainties might be addressed. How can they be localized? What formal techniques are there for considering uncertainty when performing experiments? How can they be characterized and quantified when considering results of a simulation? This special track offers an opportunity to present your approaches to these problems.

The main session topics are:

- Possible approaches for quality assurance of data
- How to evaluate uncertainties of input and output data
- Structural uncertainty and complexity of models
- Spatial and temporal aggregation and granularity vs. accuracy of results
- How to communicate, document and visualize uncertainties

# Thank you for your attention!


Contact: Dr. Megi Sharikadze

Research coordination and project management,

Leibniz Supercomputing Centre,
Bavarian Academy of Sciences and Humanities,

Boltzmannstr. 1, 85748 Garching near Munich, Germany
Phone: +49-89-35831-8873, Fax: +49-89-35831-8673

ms@lrz.de