Germany and its Open Science Cloud

PLAN-E Meeting, April 27th, 2017, Poznan Supercomputing and Networking Centre, Poznan

Arndt Bode  Member of the Board of Directors, Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities and Technical University Munich
Germany and the Open Science Cloud (1)

German Wording: „Informationsinfrastrukturen“

2011 „Kommission Zukunft der Informationsinfrastruktur“
Definition of a concept for an Information-Infrastructure in Germany
Submitted to GWK („Gemeinsame Wissenschaftskonferenz“)

2014 „Rat für Informationsinfrastrukturen“ (RfII)
RfII publishes position paper „Leistung aus Vielfalt“
Recommendation for Research Data Management (RDM): structures, processes, financials for RDM in Germany
Recommendation to establish „Nationale Forschungsdaten Infrastruktur“ (NFDI):
Governing Board to coordinate activities regarding research data. [ To be decided by GWK in April 2017 ]
Many activities regarding RDM by the „Laender“ (provinces)
- Baden Württemberg: eScience Initiative (coordinated project of BW universities)
- Hessen: HeFIS „Hessische Forschungsdaten Infra-Struktur“

Activities by Research Organizations
- Helmholtz: Helmholtz-Data-Federation
- Fraunhofer: Development of a private cloud FORDATIS
- Leibniz: FDM project: Leibniz Data

Cooperation NL-D: position paper „FAIR use of Research Data“

Participation in EOSC since 2015 „High level export group“: Klaus Tochtermann
Governing Board (to be established summer 2017) would correspond to NFDI

Preparation of the G7-Member States for a Recommendation: Action plan to be signed in September 2017
Open Science Cloud defined as a seamless scalable and federated infrastructure for HPC, HTC, data analytics, RDM is "work in progress" : GCS, GA, ........

GeRDI: a federated approach, mainly for RDM
Open Science needs Federated Infrastructure

Structure. Keypoints of the Talk

The Point of View of an Infrastructure Provider (LRZ)
The Open Data Challenge
Towards a Generic Research Data Infrastructure (GeRDI)
GeRDI Vision

Create an infrastructure to support open and interdisciplinary research.

Open Trusted Environment

Data
- findable
- accessible
- interoperable
- Reusable

Image: www.digitalbevaring.dk
The Leibniz Supercomputing Centre

What is the LRZ? IT-Service Provider / Research Institution / HPC Centre

Leibniz Supercomputing Centre of the Bavarian Academy of Sciences and Humanities

Computer Centre for all Munich Universities
- 160 employees and 43 extra staff
- More than 100,000 students
- 30,000 professors, scientific and other staff
Services also for big customers, eg. BSB/BVB (Bavarian libraries) or Hochschulstart

IT Services (amongst others)
- Mail & Web
- Internet Service Provider (MWN)
- Storage Cloud
- Compute Cloud
- gitlab
- Licences
- Backup and Archiving Centre
- HPC (SuperMUC)
  incl. application support
Munich Scientific Network (MWN)

Science needs Connectivity. Internet and Network for Scientists

Operation figures
- 16 Routers
- 1,500 Switches
- 2,900 Access points
- 76 leased dark fibres
- 40+ private dark fibres
- >150,000 systems connected
- 50 Locations with more than 540 groups of buildings

Connecting Munich’s universities to the internet
- with 10 Gbit/s
- 1,200/800 Terabyte per month incoming/outgoing
- 22 Petabyte per month via backbone
Code Hosting @ LRZ

Science needs Collaboration. Enable to share and improve scientific applications

The LRZ hosts a gitlab instance as an open platform for code sharing

• 4600+ users
• 6500+ code projects
• 420+ GB of source code
• 1800+ Merge Requests
• 1000+ Forks
• supports invitation of scientists outside of the LRZ Identity Management (gitinvited.lrz.de)

gitlab.lrz.de was introduced early 2016 and is now widely used in both research and teaching.
The LRZ: a Supercomputing Centre

Science needs Computing. Creating and Analyzing Big Data

SuperMUC
- Phase 1:
  - 3 PF, > 300 TB Memory, > 156,000 cores
- Phase 2:
  - 3 PF, 200 TB Memory, > 86,000 cores

Linux Clusters
- Heterogenous, Several flavours
- Continuously expanded
- 327 TF, 33 TB, 16,180 cores
- Housing for LCG and others
Backup and Archiving Systems at LRZ

Science needs a Digital Memory. Long Term Archival and Backup Solutions

**Occupancy**
- 48 Petabytes data
- 21 billions files
- 117 Terabytes data incoming per day
- 9,654 registered systems

**Configuration**
- 5 tape libraries
- 154 tape drives
- 46,000 tape cartridges
- approx. 3 Petabytes disk cache
Open Data Challenge 1

Diversity. Several research data repositories across Germany · Europe · World.
Open Data Challenge 2

Openness and Interdisciplinarity. How can data be made FAIR?

Data should be

- Findable
- Accessible
- Inter-operable
- Reproducible

Interdisciplinary data exchange adds even more complexity (e.g. what metadata standards should be used?)
Sceptical scientists. Why should I make My Data open?
How can these challenges be met?

The GeRDI-Approach. Open Science needs Federated Infrastructures

- A single service provider cannot hope to succeed
  → Federation of IT-service providers
- FAIRly shared data need standards
  → Interdisciplinarity and Openness reconciled in one infrastructure
- The right incentives for sharing data must be derived considering the whole data life cycle both from a specific and a global perspective
  → Community Management to foster open science
    (Listen to communities and make them talk to each other)
Project Consortium

5 Partners · CAU · DFN · LRZ · ZBW · ZIH

- Leibniz Supercomputing Center of the Bavarian Academy of Sciences and Humanities
- Verein zur Förderung eines Deutschen Forschungsnetzes e. V.
- Software Engineering Group, Christian-Albrechts-University of Kiel
- Centre for Information Services and High Performance Computing, Technische Universität Dresden
- German National Library for Economics – Leibniz Information Centre for Economics
Support the Data life cycle

What search abilities are needed by the communities?

SEARCH ENGINE

Index System

Node 1     Node 2

Public Search Mask

Environment, Resource and Ecological Economics

Chair of Hydrology and River Basin Management

Image: www.digitalbevaring.dk
Support the Data life cycle

What kind of (re)assembling procedures are needed?

SEARCH ENGINE
Index System
Node 1  Node 2

INTERIMSS TORAGE

Environmental, Resource and Ecological Economics
Chair of Hydrology and River Basin Management

Image: www.digitalbevaring.dk
Support the Data life cycle

How to interface (High Performance) Computing services?
Support the Data life cycle

Where to store derived data?

SEARCH ENGINE

INTERIMSS TORAGE

HPC

Process

Find

Analyze

Store

Environmental, Resource and Ecological Economics
Chair of Hydrology and River Basin Management
Involved Research Communities

We focus on their data and workflows to identify requirements.

- Environmental, Resource and Ecological Economics (EREE), CAU Kiel
- Paleoceanography, Ocean Circulation and Climate Dynamics, GEOMAR Helmholtz Centre for Ocean Research Kiel
- German Socio-Economic Panel (SOEP), German Institute for Economic Research (DIW)
- Hydrology and River Basin Management: ZEUS (Bavarian Flashflood Map)
- ...
Example Community: AlpEnDAC

Title:
*Virtuelles Alpen Observatorium II: Alpine Environmental Data Analysis Centre (AlpEnDAC)*

Partner:
- Prof. Michael Bittner
- Uni Augsburg / DLR
- UFS
- bifa

Duration:
2014 – 2017
GeRDI Project
Concept & Evaluation of a virtual RD Infrastructure

BMBF/DFG
Setup, Operation, Organization of a national RD-Infrastructure

Phase I
2016 – 2018
Storage Management, Services, Pilots, Operation Model

Phase II
2019 – 2021
Transfer of Phase I: Put GeRDI into Operation, Roll-Out, New Communities

Phase III
starting 2021
Self-sustained and self-organized
• Operation
• Maintenance
• Development

Outlook
Questions or Comments?
Importance for Research Data

Rat für Informationsinfrastrukturen (RFII)
Data management requires information infrastructures to support the digital transition in science. There are many aspects: e.g. management of increasing digital research data, development of tools and services.

High Level Expert Group „European Open Science Cloud“
There is increasing agreement among Member States and […] e-infrastructure stakeholders that an integrated approach to research data infrastructures is needed, going beyond layers (computing, data storage, use) and national and disciplinary silos.

What does bringing together mean?

GeRDI
A model for a future (German) Research Data Infrastructure
Generic Research Data Infrastructure

Link existing and future research data repositories with each other.
Across Disciplines

FAIR Data. Findable · Accessible · Interoperable · Re-usable

application of standards and development of new technologies to link existing and future research data centers
GeRDI Reference Implementation

Longterm Accessible · Sustainable · Interoperable Infrastructure

Connection of

- Existing Systems
- Software Components
- Standards
- Resources
- Technical Infrastructure
- Policies
- Protocols …

Image: www.digitalbevaring.dk
Infrastructure for interdisciplinary research

Support the professional management of research data.

Enable all scientists in Germany

• Search
• Process
• Analyze
• Store
• Share
• Manage
• Re-use
Project Idea

(6) Communities
(5) Disciplinary Services
(4) Middleware Components
(3) Interfaces
(2) Generic Services
(1) Storage
Challenges

A long-living expandable infrastructure.

- architectural decisions
- identification and evaluation of existing software systems for re-use in GeRDI
- community management and domain-specific requirements
- heterogeneous research data and corresponding metadata
- sustainability and operation model
Work Packages

AP 1: Requirements and Software
- CAU
  - TP 1.1 Case Studies
    - ZBW
  - TP 1.2 Analysis, Specifications
    - CAU
  - TP 1.3 Software Management
    - CAU

AP 2: Implementation of federated Infrastructure
- TUD
  - TP 2.1 Software- and System Architecture
    - CAU
  - TP 2.2 Meta-data
    - ZBW
  - TP 2.3 Meta-data Management
    - TUD
  - TP 2.4 Data Management
    - TUD

AP 3: Pilot Operation
- LRZ
  - TP 3.1 IT-Security
    - LRZ
  - TP 3.2 Setup, Integration of Pilot Center
    - LRZ
  - TP 3.3 Community Management and Data Curation
    - ZBW
  - TP 3.4 Evaluation, Proof of Concept
    - TUD

AP 4: Sustainability
- ZBW
  - TP 4.1 Training Concept
    - DFN
  - TP 4.2 Future Operation Model
    - DFN
  - TP 4.3 Roll-Out Preparation
    - ZBW

AP 5: Governance, Project Management and Dissemination
- ZBW
  - DFN

Generic Research Data Infrastructure · www.gerdi-project.eu

Funded by

DFG
Our aim is to setup a pilot center connecting different repositories.
Connect repositories

Allow interdisciplinary research
Create an infrastructure to support open and interdisciplinary research.
Situation in Germany

Open science needs a federated research data infrastructure.

Communication networks

Integrated national infrastructure for research data

High performance computing capacities