GeRDI Generic Research Data Infrastructure

Research Data Infrastructures – How Generic, How Specific? Overview of the GeRDI Project

Hans-Joachim Bungartz

Department of Informatics, Technical University of Munich

Leibniz Supercomputing Center (LRZ)

German Research & Educational Network (DFN)

Personal Background

Research

- Education in Mathematics & Informatics (sorry – no physics, no materials … ☹)
- Scientific Computing, High-Performance Computing
- Computational Science and Engineering
- Data analytics via “numerical machine learning”

IT infrastructure

- Heading DFG’s Commission for IT Infrastructure 2006-2013
- Heading DFN since 2011 (network, services – EduRoam, DFN-AAI, …)
- Task Force of the Wissenschaftsrat that led to NHR / National Supercomputing
- Various advisory boards

Funded by DFG

Attention! Brief philosophical detour (concerning the 4 paradigms) ☺

3 Paradigms … (my list counts only 3 … ☺)

#1: Experiment … somehow data-based

(cf. statistics: conclude about reality, based on experiments)

#2: Theory … somehow model-based

(cf. stochastics: conclude about reality, based on models)

#3: Computation … bridging the gap

A Brief History of “Computational”

(Computational ←→ Computation-aided)

1st generation of “computational”: qualitative (forward) simulation

2nd generation of “computational”: optimization, inverse problems

3rd generation of “computational”: quantitative through analytics/ML

Ultimate goal of “computational”: predictive science

The current AI/ML hype is also due to the fact that we now have the computational power (HW, algorithms) that was absent in the early days of AI/NN/…

End of detour – back on Earth again… ☺

Topics

1. Motivation
2. Vision & Mission
3. GeRDI at a Glance
4. Important Aspects
5. Outlook

Funded by DFG

GeRDI Generic Research Data Infrastructure

1. Motivation

What was the motivation to initiate GeRDI?

CAU Christian-Albrechts-Universität zu Kiel

DFN Deutsches Forschungsnetz

lrz Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften

ZBW Leibniz-Informationszentrum Wirtschaft Leibniz Information Centre for Economics

ZIH Zentrum für Informationsdienste und Hochleistungsrechnen

European Developments: European Open Science Cloud

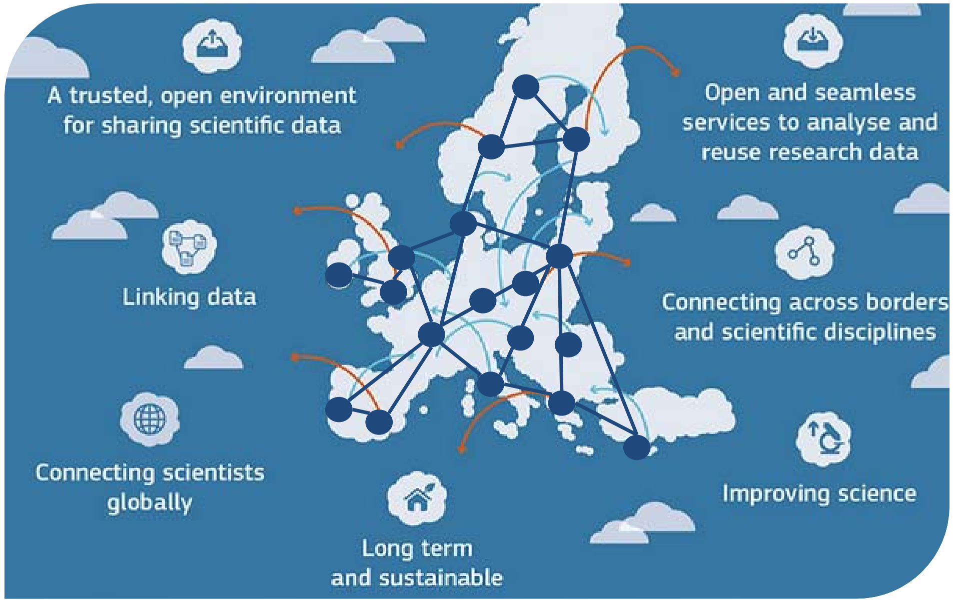
Idea: European Research Area via a Network of Research Data Centers

European Commission

Realising the European Open Science Cloud

First report and recommendations of the Commission High Level Expert Group on the European Open Science Cloud

Research and Innovation



A trusted, open environment for sharing scientific data
Open and seamless services to analyse and reuse research data
Linking data
Connecting across borders and scientific disciplines
Connecting scientists globally
Improving science
Long term and sustainable

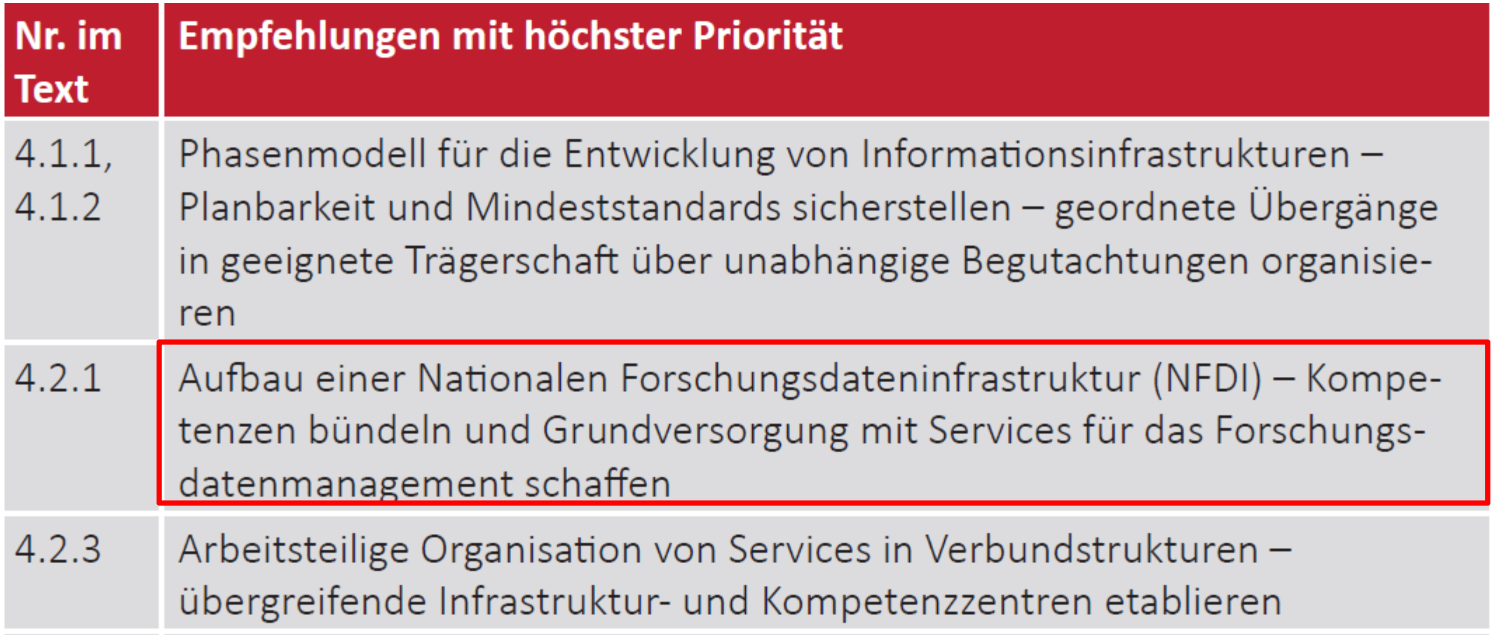
Funded by DFG

Nat‘l Developments: Recommendations of the German Council for Research Information Infrastructures (RfII)

rfii Rat für Informations Infrastrukturen

LEISTUNG AUS VIELFALT

Empfehlungen zu Strukturen, Prozessen und Finanzierung des Forschungsdatenmanagements in Deutschland



Nr. im Text	Empfehlungen mit höchster Priorität
4.1.1, 4.1.2	Phasenmodell für die Entwicklung von Informationsinfrastrukturen – Planbarkeit und Mindeststandards sicherstellen – geordnete Übergänge in geeignete Trägerschaft über unabhängige Begutachtungen organisieren
4.2.1	Aufbau einer Nationalen Forschungsdateninfrastruktur (NFDI) – Kompetenzen bündeln und Grundversorgung mit Services für das Forschungsdatenmanagement schaffen
4.2.3	Arbeitsteilige Organisation von Services in Verbundstrukturen – übergreifende Infrastruktur- und Kompetenzzentren etablieren

Research Data Management Initiatives in Germany (excerpt)

Hessen, Hamburg, Baden-Württemberg, Helmholtz, Leibniz, Fraunhofer, ...

Industrial Data Space
FhG

LeibnizData

Helmholtz Data
Federation

Hamburg Open Science

HeFDI – Hessische
Forschungsdateninfrastrukturen

eScience
Baden-Württemberg

Funded by DFG

On the one hand …

- A lot of political initiatives (open, data, FAIR, …)
- Increasing awareness of importance of research infrastructures
- Even more: RI as an essential part of science – not just "something that has to be there"
- Emerging funding frameworks – DFG-LIS/KfR, EOSC, NFDI, …
- Several institutional responses
  - Research Associations
  - States
  - Projects (CERN-LHC, Copernicus, SkA, …)
  - "well organized" communities
- Rather mature level for "Huge-Data communities" (Helmholtz, e.g.)

Funded by DFG

On the other hand …

- Various isolated (community-driven) attempts
- In Germany often via DFG funding
- Creates increasing concern at DFG:
  - in their bodies (AWBI/LIS, KfR/WGI, …)
  - among reviewers: “*How many of such research data-related database systems are we expected to fund*?”
  - Lack of sustainability – standard issue in scientific software
  - Lack of interoperability (the “I” in FAIR …)
  - Too many similar developments, missing economy of scale
- Not very well developed situation for the “long tail” (universities, “Small-Data communities”) ➔ primary target group for GeRDI

Funded by DFG

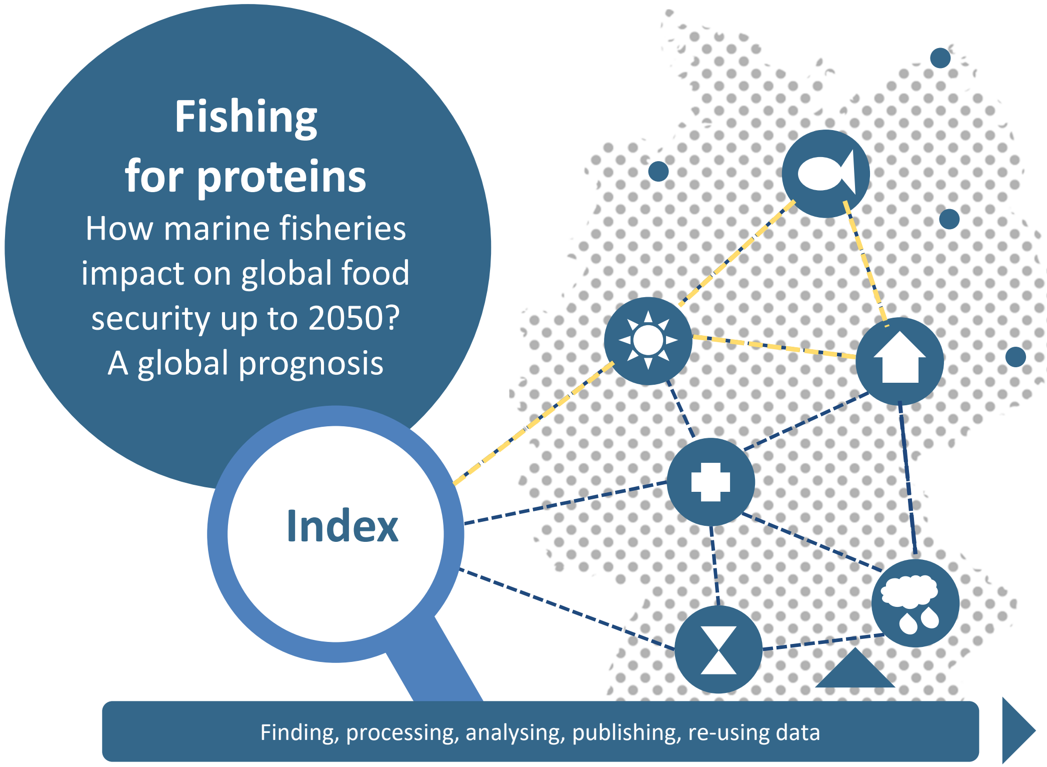
Generic vs. Specific

- The specialist's / domain scientist's view:
  - Starting point: "My science is special, my data are special, my requirements are special … hence my Research Data Infrastructure has to be special, too."
  - Consequence: "Let's get some help from experts and develop something for us."
- The generalist's / IT person's view:
  - Starting point: "Data are data, we can work with abstractions of concrete things; data need data base / data management systems. That's it."
  - Consequence: "Let's build a generic RDI useful for all fields."
- Truth: somewhere in-between
  - At the top: domain-dependent data ➔ specific top
  - At the bottom: bits & bytes ➔ generic bottom
  - Explore the frontier!

Funded by DFG

2. Vision & Mission

GeRDI – The basic idea



Fishing for proteins

How marine fisheries impact on global food security up to 2050? A global prognosis

Index

Finding, processing, analysing, publishing, re-using data

2016

ÜBERFISCHT UND UNTERVERSORGT

Wie viel Fisch wir in Zukunft fangen und wer [illegible]
Eine weltweite Prognose

Connect repositories across disciplines

- Collaboration with various research communities
- Focus on the data life cycle
- Contribution to the European Open Science Cloud and national initiatives

Funded by DFG

Vision

Significant increase of interoperability.

Principle “as generic as possible, as specific as necessary” allows for effective and efficient usage of research data cross disciplines.

Funded by DFG

Mission

GeRDI will provide **generic**, **sustainable** and **open** software connecting research data repositories to enable multidisciplinary and FAIR research data management.

This **software** will be based on common standards and be developed in close collaboration with several research communities to ensure a 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 life span of the project.

All project results – in particular software, training support and business model – will form a German contribution to the European Open Science Cloud.

CAU Kiel University Christian-Albrechts-Universität zu Kiel · DFN Deutsches Forschungsnetz · lrz Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften · ZBW Leibniz-Informationszentrum Wirtschaft Leibniz Information Centre for Economics · ZIH

Funded by DFG

GeRDI Generic Research Data Infrastructure

3. Project at a Glance

How GeRDI is organised

CAU Christian-Albrechts-Universität zu Kiel   DFN Deutsches Forschungsnetz   lrz Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften   ZBW Leibniz-Informationszentrum Wirtschaft Leibniz Information Centre for Economics   ZIH Zentrum für Informationsdienste und Hochleistungsrechnen

Facts and Figures

- Initiated by funding agencies (BMBF, later DFG) – “pre-NFDI” initiative
- Funded by DFG within its programme „Wissenschaftliche Literaturversorgungs- und Informationssysteme“ (DFG-LIS)
- Outline:
  - Phase 1 (2016-2019): concepts, pilot with selected communities
  - Phase 2 (2019-2022): production, roll-out
  - Budget: ca. 3 Mio Euro (Phase 1)
- 5 partners + associated community partners
- Phase 2 delayed – revision of proposal required

Funded by DFG

Project Consortium

Plus Advisory Board gathering other key players in Germany

Infrastructures

C A U
Christian-Albrechts-Universität zu Kiel

DFN
DEUTSCHES FORSCHUNGSNETZ

lrz Leibniz Supercomputing Centre
of the Bavarian Academy of Sciences and Humanities

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

ZIH
Zentrum für Informationsdienste und Hochleistungsrechnen

Communities

Virtual Alpine Observatory

envcomp

UNISDR
United Nations Office for Disaster Risk Reduction

GEOMAR

CBG
Max Planck Institute of Molecular Cell Biology and Genetics

ozean der zukunft
DIE KIELER MEERESWISSENSCHAFTEN

SOEP

TUM Lehrstuhl für Hydrologie und Flussgebietsmanagement
Ingenieurfakultät Bau Geo Umwelt
Technische Universität München

NCT
NATIONAL CENTER FOR TUMOR DISEASES
PARTNER SITE DRESDEN
UNIVERSITY CANCER CENTER UCC

dh Digital Humanities
UNIVERSITÄT LEIPZIG

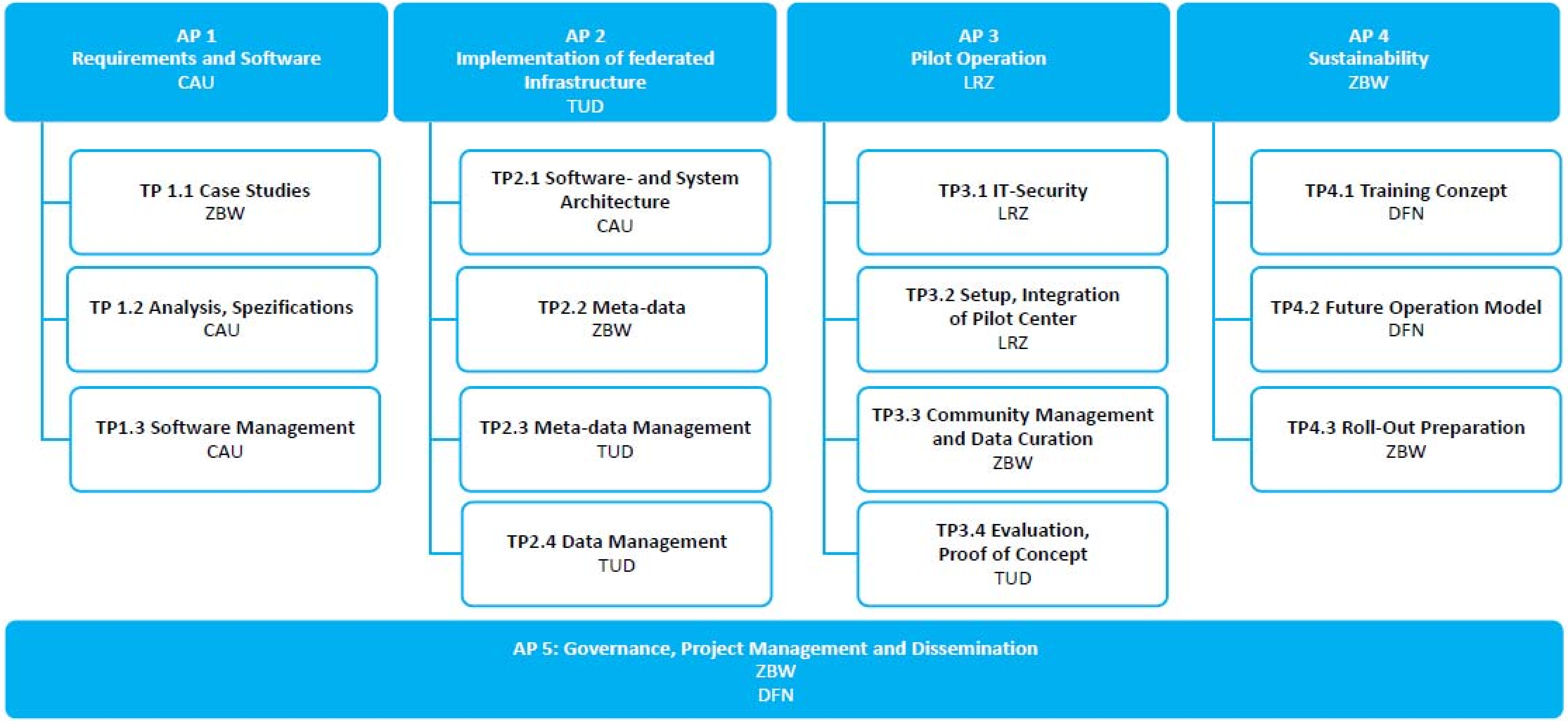
GeRDI Perspective & Embedding

	Phase I 2016-2019	Phase II 2019-2022	from 2022
GeRDI Project Concept and Evaluation of a Research Date Infrastructure	Services, Pilot, Operation Model, Training	Transfer of Phase I: Put GeRDI into Operation, Roll-Out, New Communities	
BMBF/DFG Setup, Operation, Organization of a National Research Data Infrastructure	Expert discussions about Research Data Management	Tendering of a National Research Data Infrastructure, Roll-Out, Sustainable Organization	Self-sustained and self-organized • Operation • Maintenance • Development

Image: www.digitalbevaring.dk

Funded by DFG

Work Packages



AP 1
Requirements and Software
CAU

- TP 1.1 Case Studies
  ZBW
- TP 1.2 Analysis, Spezifications
  CAU
- TP1.3 Software Management
  CAU

AP 2
Implementation of federated Infrastructure
TUD

- TP2.1 Software- and System Architecture
  CAU
- TP2.2 Meta-data
  ZBW
- TP2.3 Meta-data Management
  TUD
- TP2.4 Data Management
  TUD

AP 3
Pilot Operation
LRZ

- TP3.1 IT-Security
  LRZ
- TP3.2 Setup, Integration of Pilot Center
  LRZ
- TP3.3 Community Management and Data Curation
  ZBW
- TP3.4 Evaluation, Proof of Concept
  TUD

AP 4
Sustainability
ZBW

- TP4.1 Training Conzept
  DFN
- TP4.2 Future Operation Model
  DFN
- TP4.3 Roll-Out Preparation
  ZBW

AP 5: Governance, Project Management and Dissemination
ZBW
DFN

Funded by DFG

Aspects of Sustainability (cf. NFDI discussions)

- Software
  - Maintenance?
  - Training?
  - Modifications, extensions, variants, …? Open source?
- Hardware
  - Centralized at “hubs” (cf. HPC in Germany – PRACE, Gauß Center, Gauß Alliance, NHR)?
  - Semi-centralized at community hubs?
  - Decentralized with users / institutions / universities?
- Governance
  - Decision-making?
  - Integrated in NFDI governance vs. independent?

Funded by DFG

GeRDI Generic Research Data Infrastructure

4a. Communities & Scenarios

Who are the (initial) users of GeRDI's services?

Our Research Communities

Various disciplines are involved

- Alpine Environmental Data Analysis Centre
- Digital Humanities
- Environmental Resource and Ecological Economics
- Hydrology and River Basin Management
- Molecular Cell Biology and Genetics
- Paleoceanography
- Socio economics
- Tumor Diseases

Long tail – no existing solutions, but openness

Existing links to partners

Why these?

Long tail – no existing solutions

Needs & openness

Existing links to partners

Certain breadth of topics

Community Partner

Virtual Alpine Observatory

envcomp

UNISDR
United Nations Office for Disaster Risk Reduction

GEOMAR

CBG
Max Planck Institute of Molecular Cell Biology and Genetics

ozean der zukunft
DIE KIELER MEERESWISSENSCHAFTEN

SOEP

TUM
Lehrstuhl für Hydrologie und Flussgebietsmanagement
Ingenieurfakultät Bau Geo Umwelt
Technische Universität München

NCT
NATIONAL CENTER FOR TUMOR DISEASES
PARTNER SITE DRESDEN
UNIVERSITY CANCER CENTER UCC

dh Digital Humanities
UNIVERSITÄT LEIPZIG

CAU Kiel University Christian-Albrechts-Universität zu Kiel · DFN DEUTSCHES FORSCHUNGSNETZ · lrz Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften · ZBW Leibniz-Informationszentrum Wirtschaft Leibniz Information Centre for Economics · ZIH

Funded by DFG

Research Case 1: Fishery Management

Fishing for proteins: How marine fisheries impact on global food security up to 2050?

Research

Predict the amount of fish available for worldwide nutrition

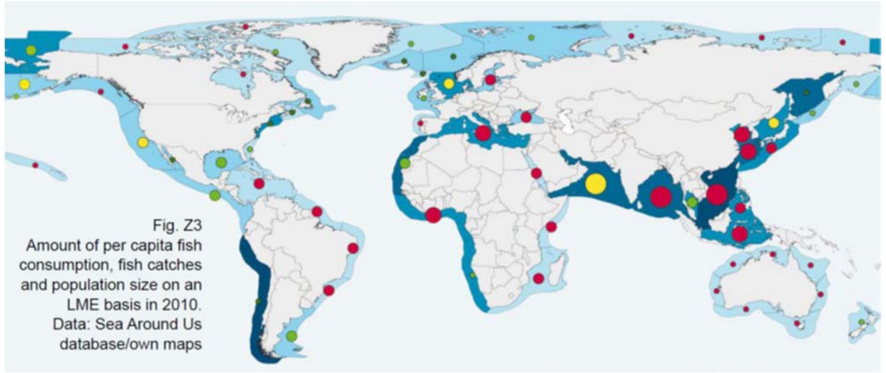


Fig. Z3
Amount of per capita fish consumption, fish catches and population size on an LME basis in 2010.
Data: Sea Around Us database/own maps

Activities

- Collect data from heterogenous sources (FAOStat, SeaAroundUs, …)
- Preprocess, analyze, and predict
- Publish
  - WWF-Report
  - Research paper, doi: 10.1111/gcb.13060
  - Research data, doi: 10.1594/PANGAEA.856741

Funded by DFG

Research Case 1: Fishery Management

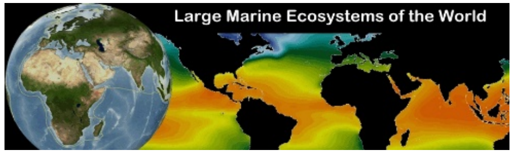
Fishing for proteins: How marine fisheries impact on global food security up to 2050?

SEA AROUND US
FISHERIES, ECOSYSTEMS & BIODIVERSITY

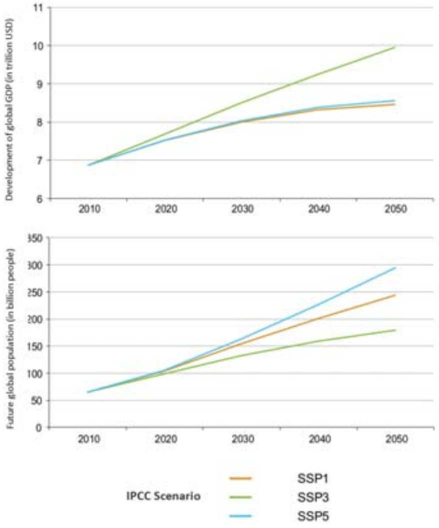
International Institute for Applied System Analysis

FAO STAT

FishStatJ



- large marine ecosystem based catch data
- socio-economic data of 5 future scenarios
- prices of substitution goods
- trade statistics
- large marine ecosystem shape files



WWF
THIS PROJECT IS PART-FINANCED BY THE EUROPEAN UNION
2016
FISHING FOR PROTEINS
How marine fisheries impact on global food security up to 2050. A global prognosis

PANGAEA

Funded by DFG

Research Case 2: Hydrology

Hydrology: How can we simulate & predict flash flood events?

Research

Modeling of flash flood events and assessment of impact factors

Photograph by Timothy Swinson, CC-BY 2.0 license

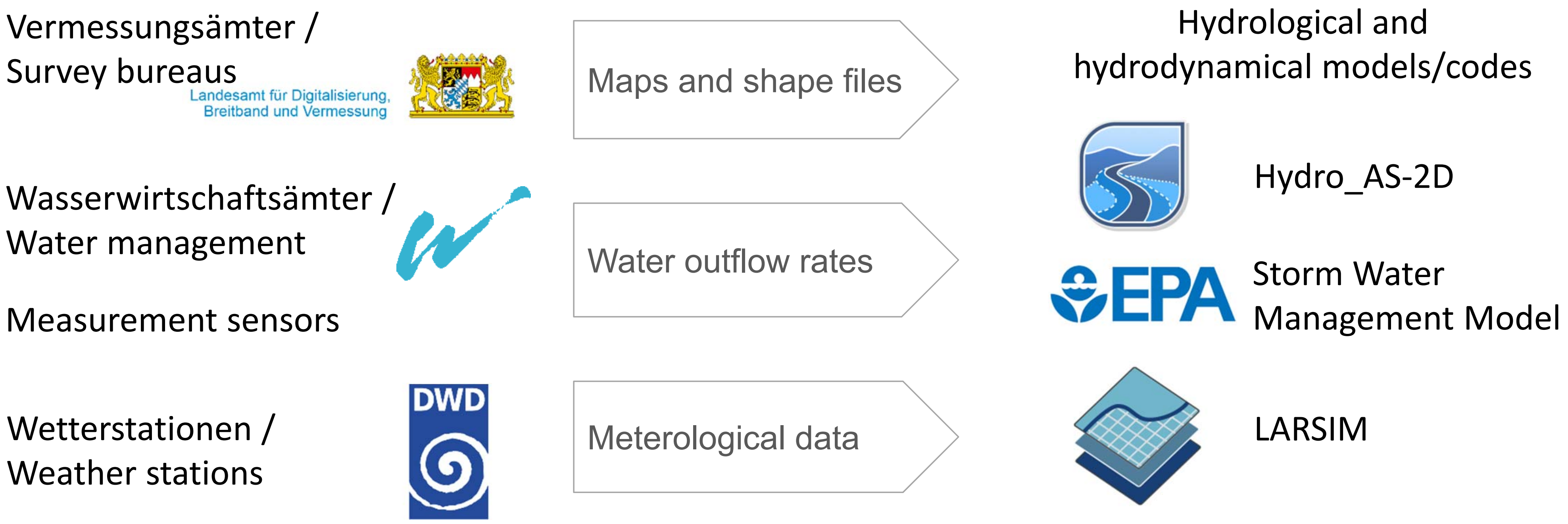
Tasks

- Aggregate data from various sources
- Preprocess and format data
- Provisioning of HPC systems
- Run hydrological and hydrodynamic simulations
- Analysis of the simulations
- Learning from federated data (simulations and measurements)
- Publish results on web portals

Funded by DFG

Research Case 2: Hydrology

Hydrology: How can we simulate flash flood events?



Vermessungsämter / Survey bureaus

Landesamt für Digitalisierung, Breitband und Vermessung

Wasserwirtschaftsämter / Water management

Measurement sensors

Wetterstationen / Weather stations

DWD

Maps and shape files

Water outflow rates

Meterological data

Hydrological and hydrodynamical models/codes

Hydro_AS-2D

EPA Storm Water Management Model

LARSIM

Funded by DFG

Research Case 3: Microscopy

Cell Imaging: How cells form tissue – Understand biological mechanisms

Research

- Development of novel microscopes
- Advanced imaging for biological samples (cells, tissues)

Tasks

- Navigate microscopy data using preview images
- Filter for image properties (magnification ratio, time resolution, size)
- Select images manually based on visual properties
- Preview of selected images as animated image sequence
- Interface to analysis service (e.g. cell tracking)

Funded by DFG

Research Case 3: Microscopy

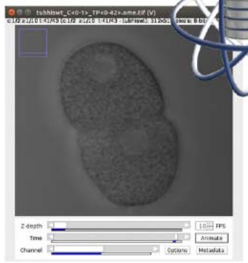
Cell Imaging: How cells form tissue – Understand biological mechanisms

Digital light sheet microscopy (DLSM)

- Produces image data or image sequences
- 3D images of molecules
- 10GB to 100GB per image sequence
- Metadata encapsulated in images (OME-TIFF)
- Ingest of microscopy data into KIT Data Manager
- Setup KIT Data Manager as community-specific repository solution

OME-TIFF
Open Microscopy Environment

KIT Data Manager


tubhiswt_C<0-1>_TP<0-42>.ome.tif (V)

C A U Kiel University Christian-Albrechts-Universität zu Kiel · DFN Deutsches Forschungsnetz · lrz Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften · ZBW Leibniz-Informationszentrum Wirtschaft Leibniz Information Centre for Economics · ZIH

Funded by DFG

FAIR principles	GeRDI support
Findable	Harvesting data sources, Search Index
Accessible	Data download support
Interoperable	Mapping of Metadata to GeRDI-Metadata Schema
Reusable	Publishing results to a GeRDI-indexed repository

Funded by DFG

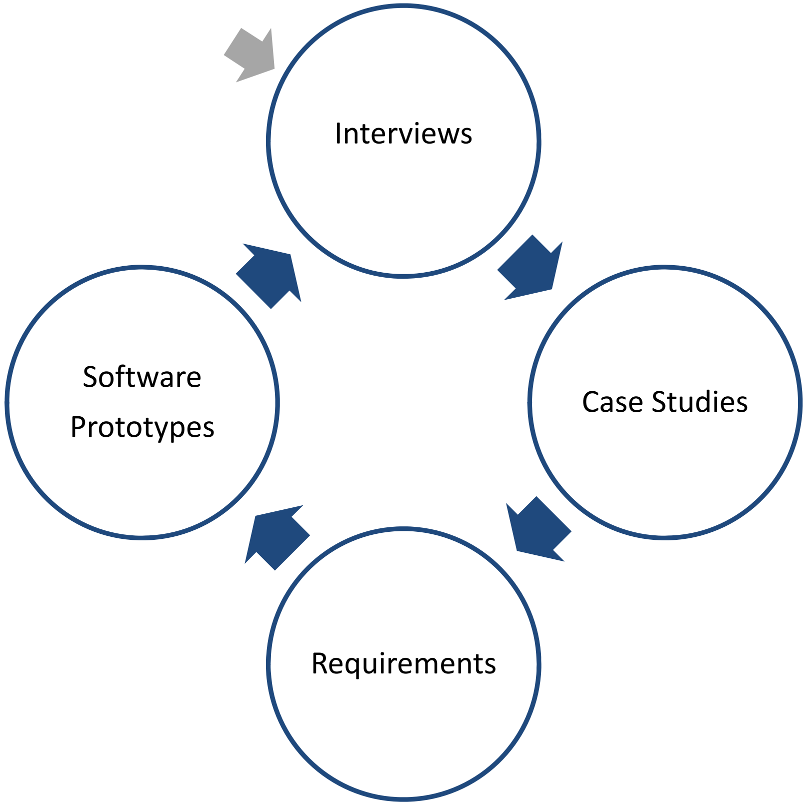
GeRDI Generic Research Data Infrastructure

4b. Requirements Engineering

How do we define the requirements on GeRDI?

CAU Christian-Albrechts-Universität zu Kiel DFN Deutsches Forschungsnetz lrz Leibniz-Rechenzentrum der Bayerischen Akademie der Wissenschaften ZBW Leibniz-Informationszentrum Wirtschaft Leibniz Information Centre for Economics ZIH Zentrum für Informationsdienste und Hochleistungsrechnen

Use (Research) Case Driven Approach



Interviews

Case Studies

Requirements

Software Prototypes

Christin is a researcher at a tumor center.

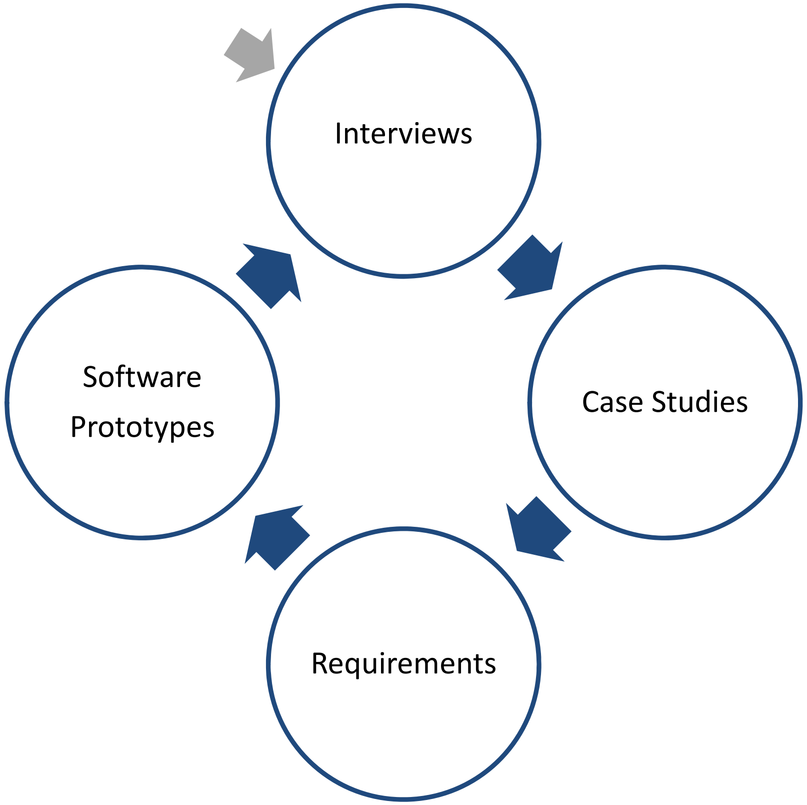
How does the environment influence the development of tumors?

Jannis is a researcher in the environment, resources and economics group.

How marine fisheries impact on global food security up to 2050?

Funded by DFG

Requirements Analysis



Stakeholders

- Researchers
- Data Providers (repositories and archives)
- Developers (GeRDI project team)

Continuously …

… present state of the art

… derive new usage scenarios

… determine scenarios

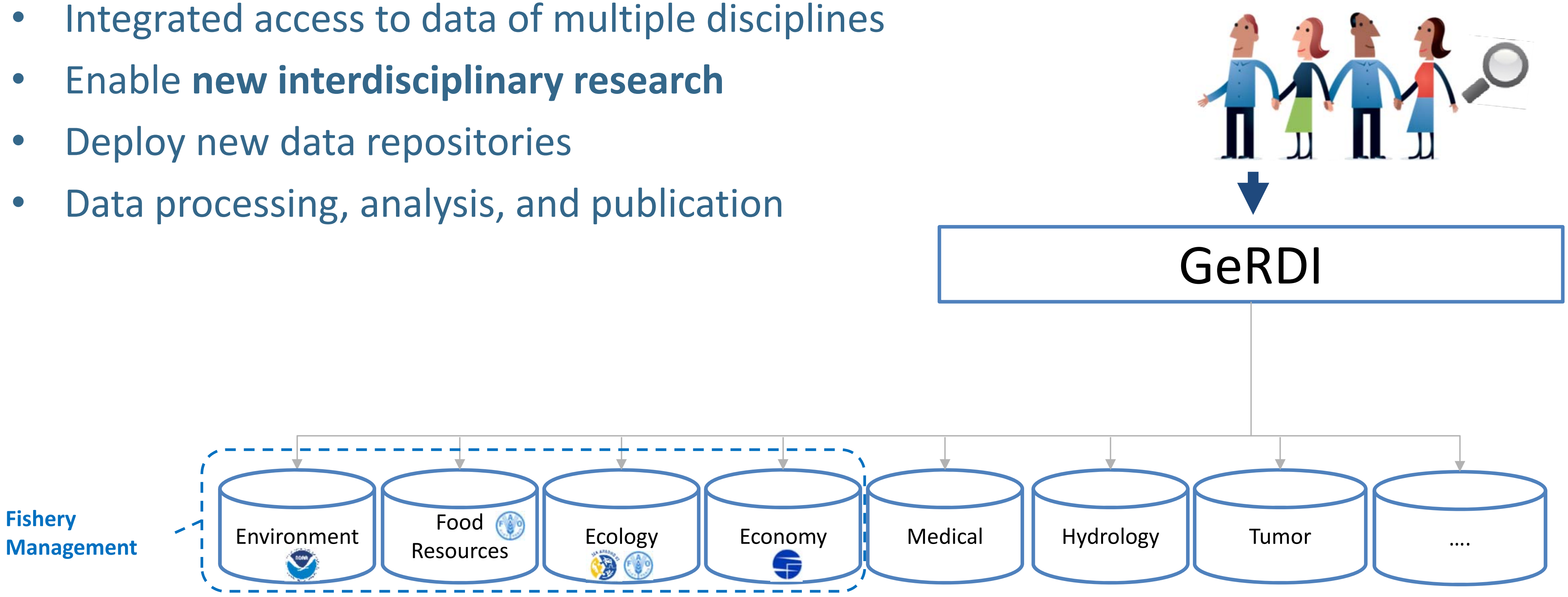
… add to Backlog

… release new software versions

Funded by DFG

Main Usage Scenarios

- Integrate and harvest existing repositories
- Integrated access to data of multiple disciplines
- Enable **new interdisciplinary research**
- Deploy new data repositories
- Data processing, analysis, and publication



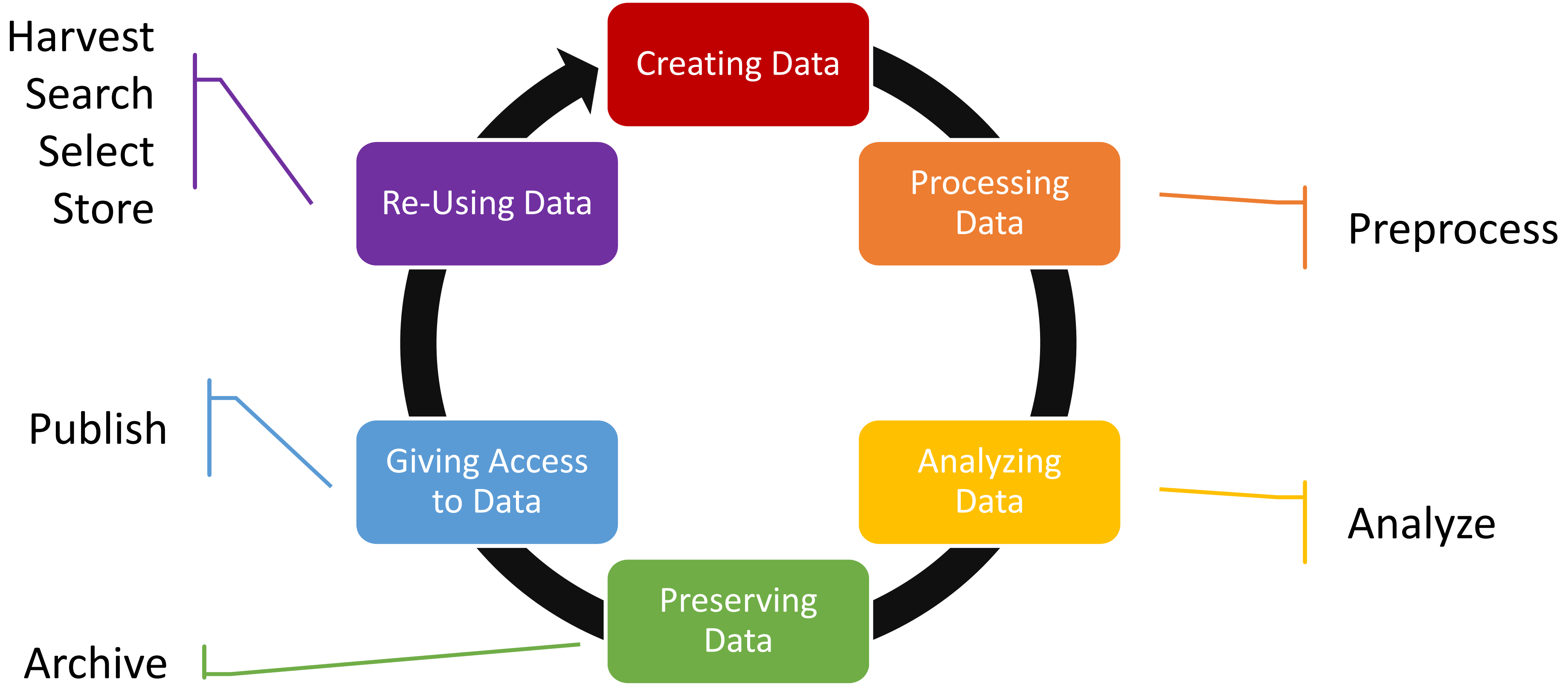
Funded by DFG

GeRDI Generic Research Data Infrastructure

4c. Architecture Design

Which services will GeRDI provide, and how are they interconnected?

GeRDI Services: Follow the Research Data Life Cycle



Creating Data

Processing Data

Analyzing Data

Preserving Data

Giving Access to Data

Re-Using Data

Harvest
Search
Select
Store

Preprocess

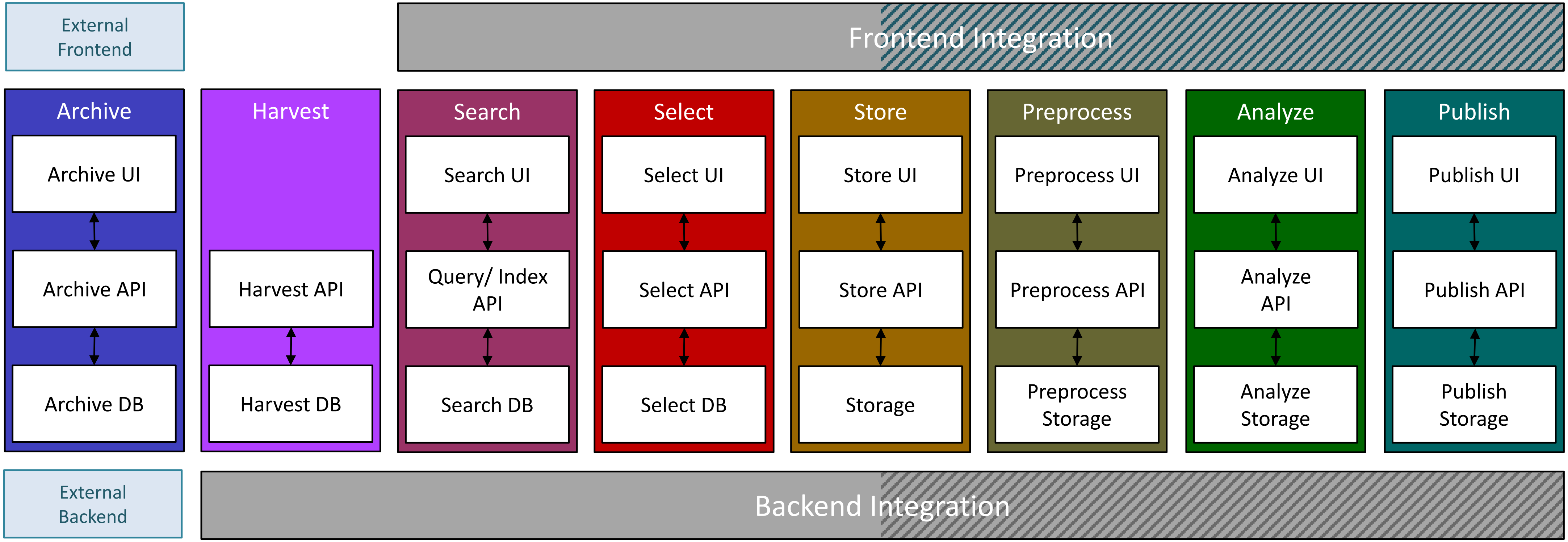
Analyze

Publish

Archive

Funded by DFG

GeRDI Services: Microservice Architecture Overview



External Frontend

Frontend Integration

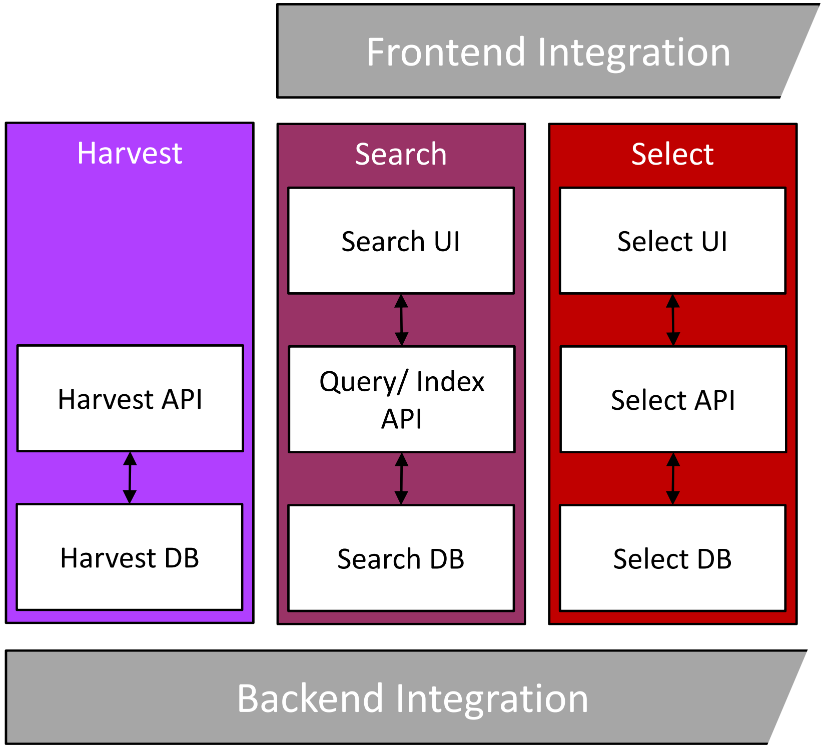
Archive	Harvest	Search	Select	Store	Preprocess	Analyze	Publish
Archive UI		Search UI	Select UI	Store UI	Preprocess UI	Analyze UI	Publish UI
Archive API	Harvest API	Query/ Index API	Select API	Store API	Preprocess API	Analyze API	Publish API
Archive DB	Harvest DB	Search DB	Select DB	Storage	Preprocess Storage	Analyze Storage	Publish Storage

External Backend

Backend Integration

Funded by DFG

GeRDI Core Services



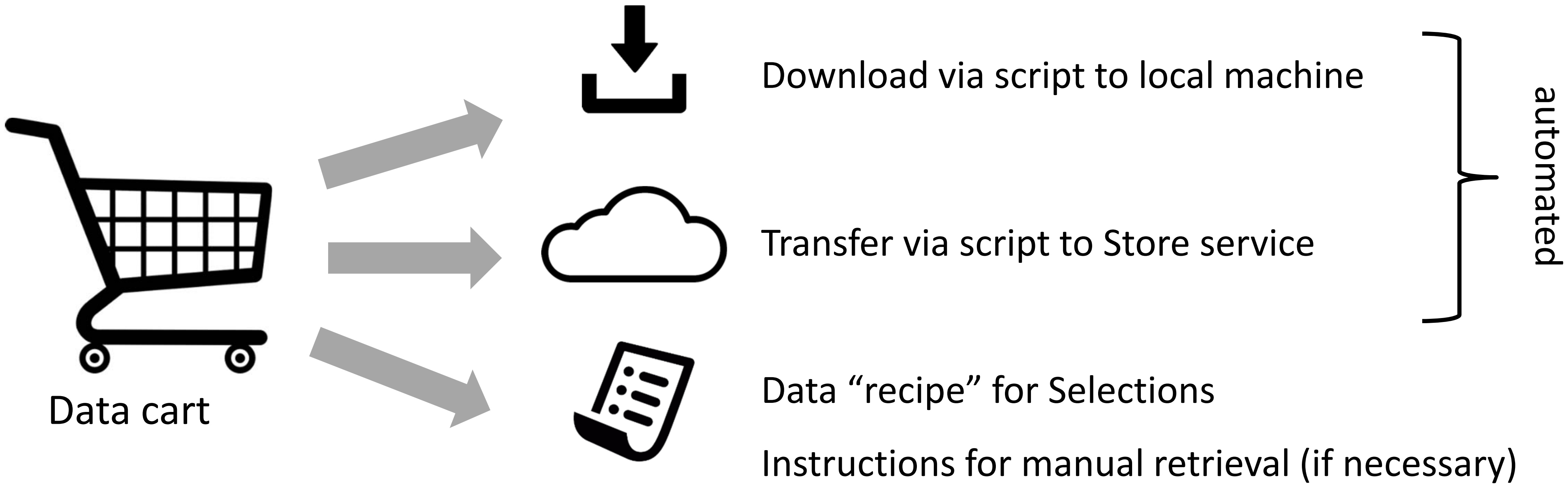
Components that are developed and operated by the GeRDI project team:

- Integration of data archives and repositories, based on standards such as OAI-PMH
- Meta data collection, integration, normalization, and enrichment, based on standards such as DataCite
- Search over multiple sources
- Selection of (meta) data into a data cart

Funded by DFG

The Select Service

General idea: select and save a named compound collection of data sets



Data cart

Download via script to local machine

Transfer via script to Store service

automated

Data “recipe” for Selections

Instructions for manual retrieval (if necessary)

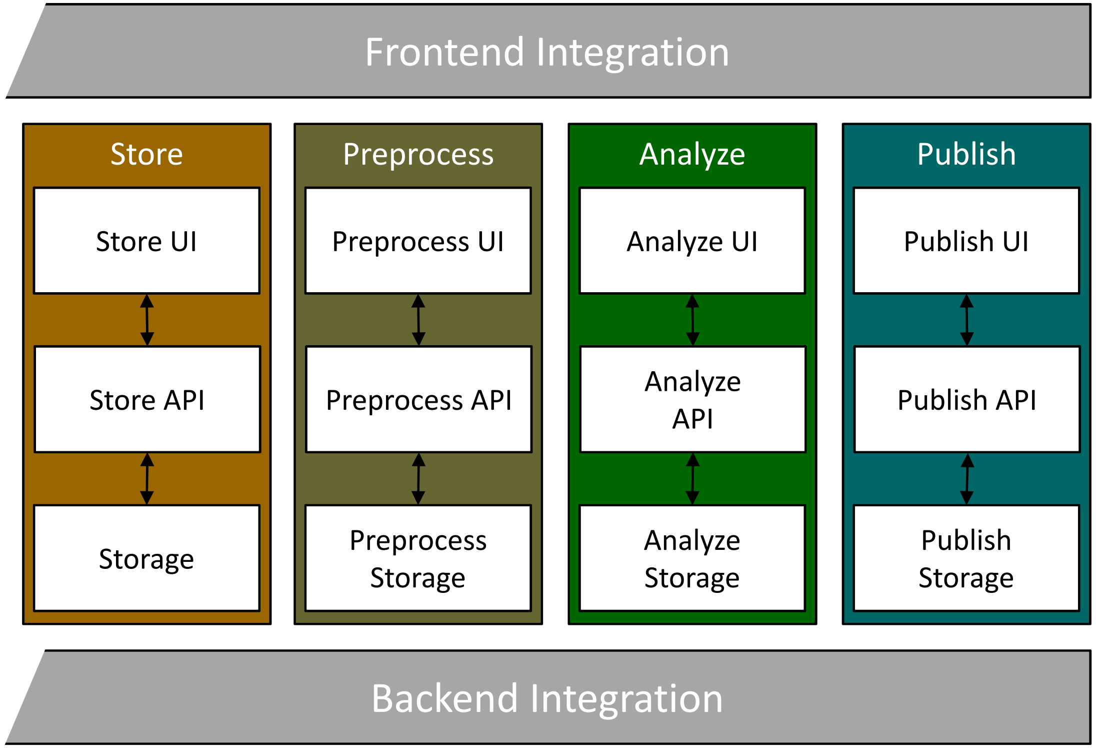
https://openclipart.org/, https://www.shareicon.net/

Funded by DFG

Extended Services

Components for which the GeRDI team provides reference implementations:

- Storage on different systems (e.g. DFN-Cloud or HPC Staging) based on data recipes
- Merging, truncating, normalizing data
- Analyzing data (Jupyter, R-Studio, Simulations, etc.)
- Publish the derived data product: Reusability of research data



Funded by DFG

Research Case Mappings

Fishery Management (Environmental Resource and Ecological Economics)

Archive	Harvest	Search	Select	Store	Preprocess	Analyze	Publish
• Sea Around Us • FAO Stat • FAO FishStatJ • SSP • GIS Data	• Adapter to the Repos	• LMEs • Catch Data • Prices • Trade Data • GIS LME & Countries	• Prices and Trade data for fish commodities	• Download onto Laptop	• Union GIS • LME Catches	• Model Combination • Prediction	• Back to Repos e.g. Pangaea

Hydrology – Flash flood modeling

Archive	Harvest	Search	Select	Store	Preprocess	Analyze	Publish
• DWD – German Weather Service • Sensor data	• Adapter to DWD repository	• Location • Date	• Time range • Area size	• Data Staging to LRZ Linux Cluster	• Prepare and format data as input for WaSiM and LARSIM	• Run the simulation software WaSiM and LARSIM • Benchmarks of the simulation	• Results are used to create a hazard map

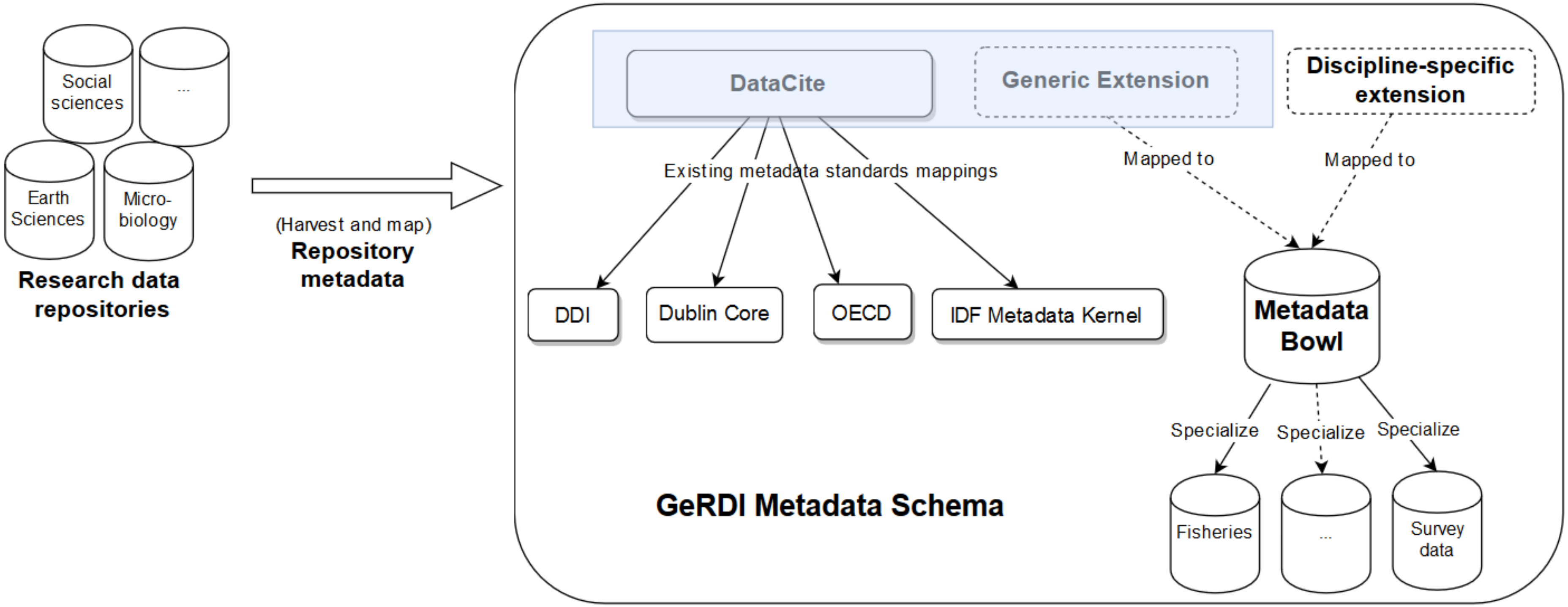
Funded by DFG

GeRDI Generic Research Data Infrastructure

4d. Implementation and Software Management

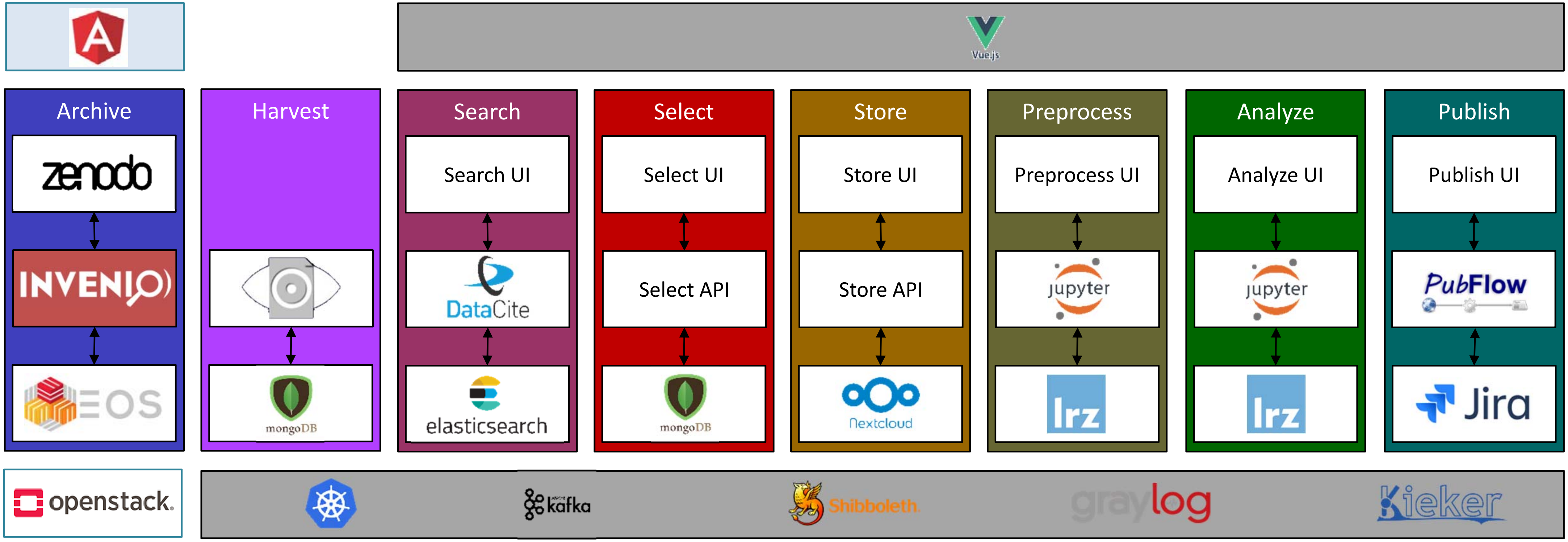
How do we implement the services?

Schema for Harvested Meta Data



Funded by DFG

Possible Implementation View

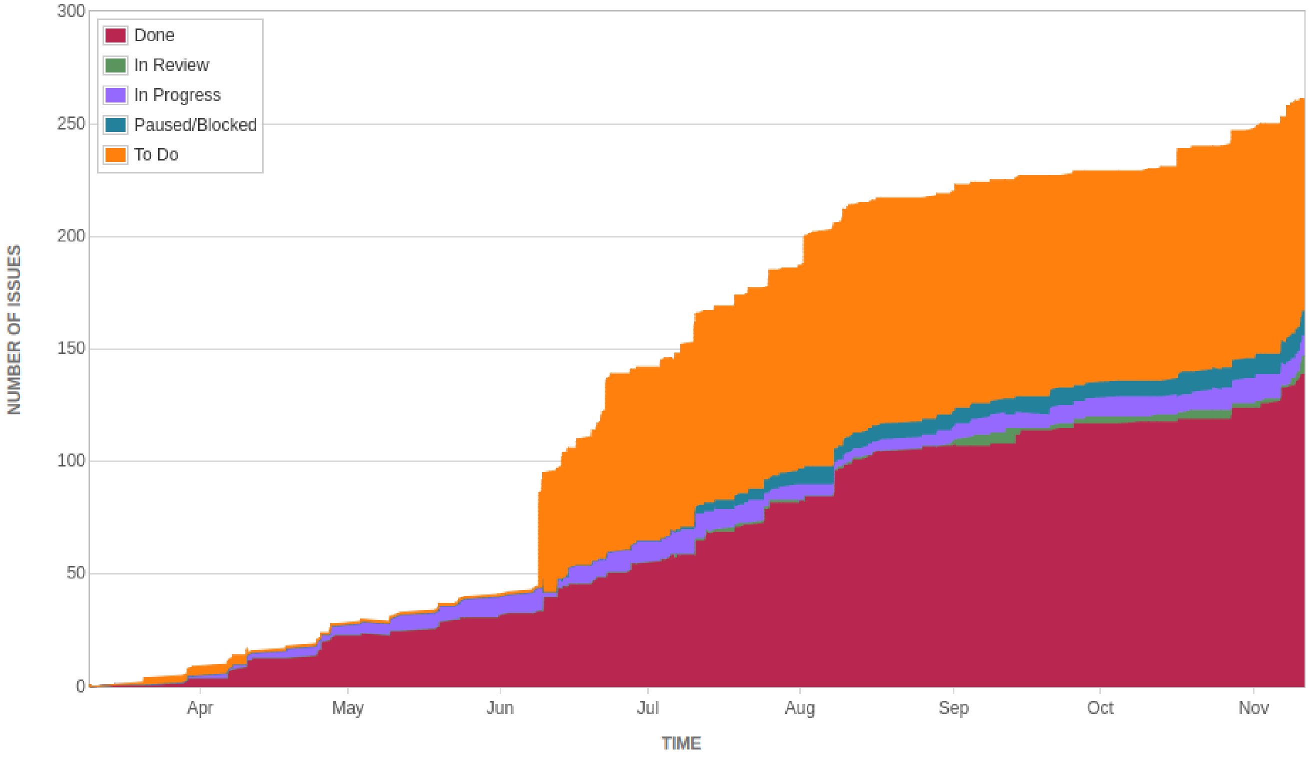


Funded by DFG

Software Management

Agile Workflow Sprints every other week

- Estimation of task
- Assignment of tasks
- Review of the sprint via Atlassian tools



Continuous Integration with nightly builds

- based on Bamboo and Kubernetes

Funded by DFG

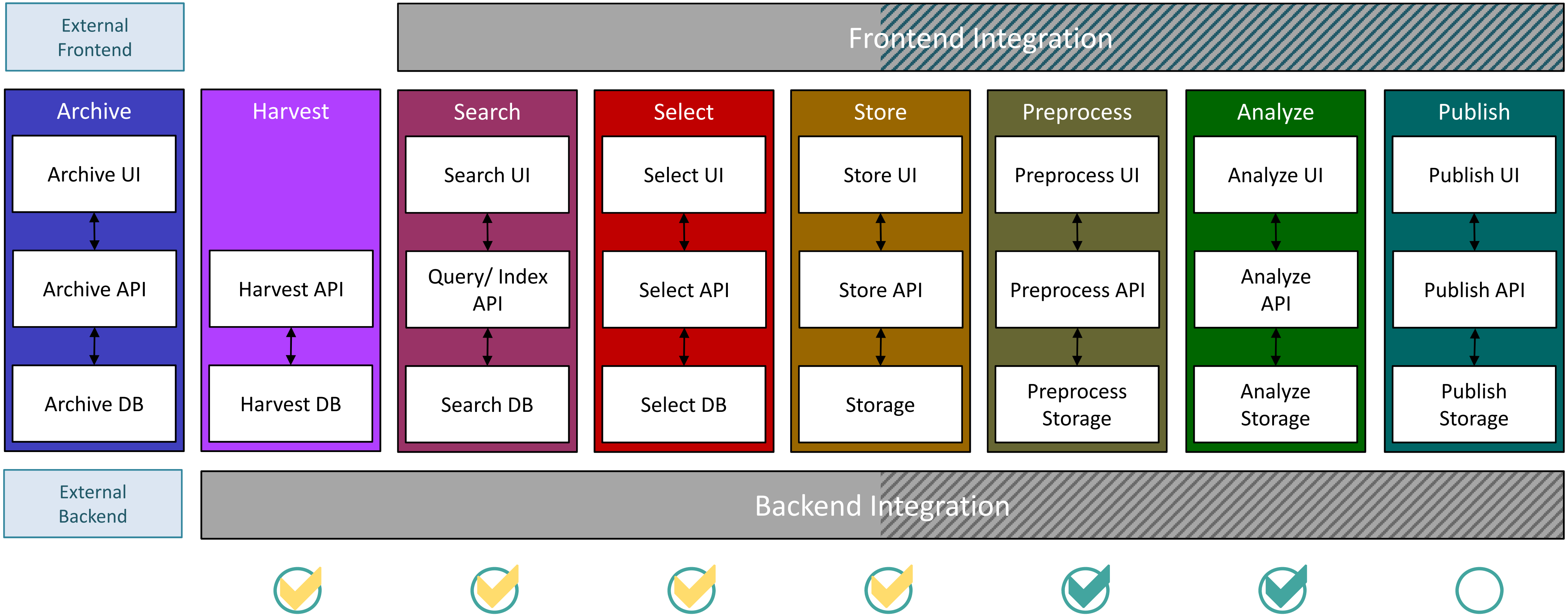
5. Status

Achievements of GeRDI

- Generic metadata schema based on Datacite
- Harvester technology, including metadata mapping capacity
- Microservice architecture
- Services for Search, Bookmark, and Store
- First version of AAI services
- Accepted as GO FAIR implementation network

Funded by DFG

GeRDI Services: Microservice Architecture Overview



External Frontend

Frontend Integration

Archive: Archive UI ↕ Archive API ↕ Archive DB

Harvest: Harvest API ↕ Harvest DB

Search: Search UI ↕ Query/ Index API ↕ Search DB

Select: Select UI ↕ Select API ↕ Select DB

Store: Store UI ↕ Store API ↕ Storage

Preprocess: Preprocess UI ↕ Preprocess API ↕ Preprocess Storage

Analyze: Analyze UI ↕ Analyze API ↕ Analyze Storage

Publish: Publish UI ↕ Publish API ↕ Publish Storage

External Backend

Backend Integration

Funded by DFG

Success criteria

	Target End of 2019	June 2019
Number of current or pending collaborative projects (incl. NFDI consortia) that would like to use GeRDI software	4	9
Number of interviews / feedback meetings with the community partners	90	79
Number of connected research data repositories (without Zenodo)	15	13 (16)
Number of harvested metadata records (without Zenodo) ◀	700.000	567.036
Number of presentations about GeRDI	45	39

Funded by DFG

Success criteria

Number of current or pending collaborative projects (incl. NFDI consortia) that would like to use GeRDI software

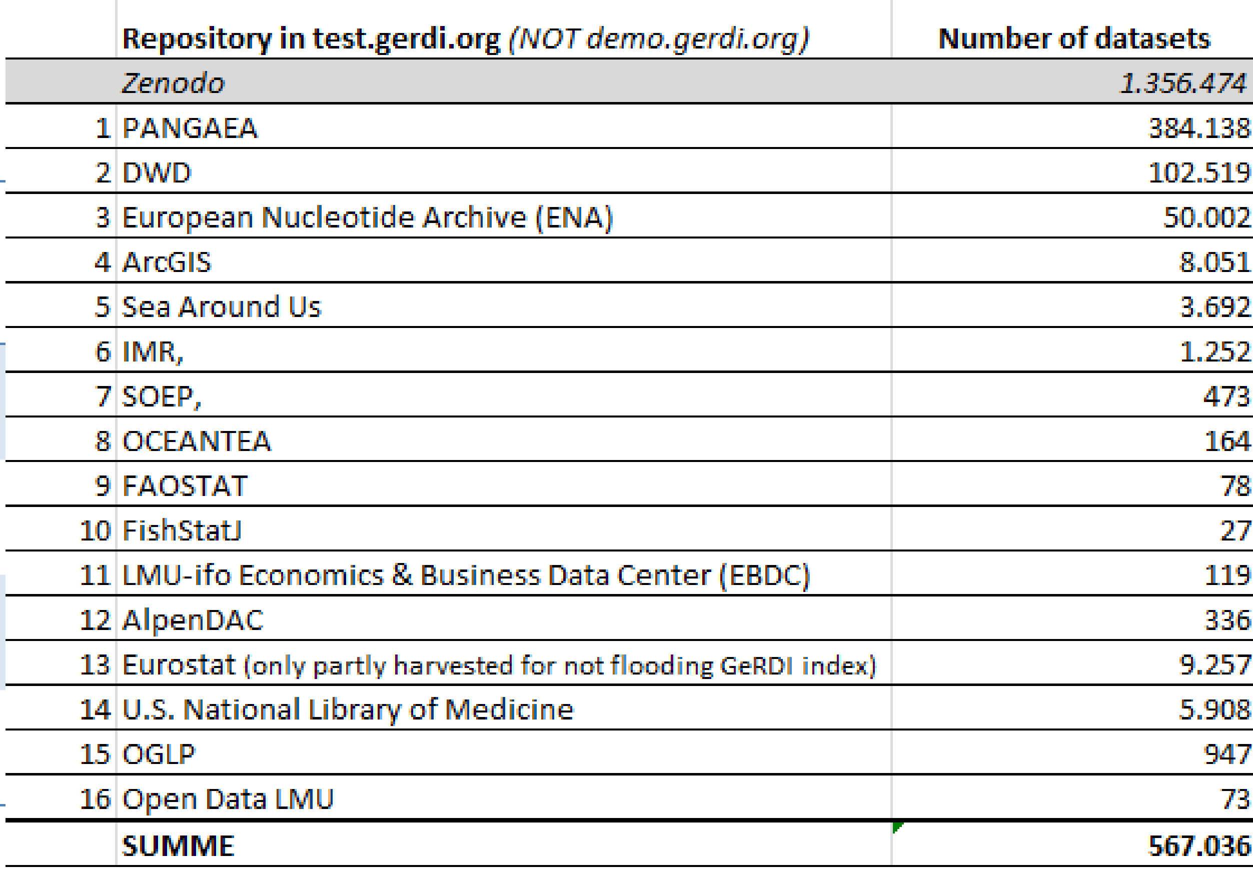
Number of interviews / feedback meetings with the community partners

Number of connected research data repositories (without Zenodo)

Number of harvested metadata records (without Zenodo)

Number of presentations about GeRDI

Target end of 2019 June 2019



	Repository in test.gerdi.org *(NOT demo.gerdi.org)*	Number of datasets
	Zenodo	*1.356.474*
1	PANGAEA	384.138
2	DWD	102.519
3	European Nucleotide Archive (ENA)	50.002
4	ArcGIS	8.051
5	Sea Around Us	3.692
6	IMR,	1.252
7	SOEP,	473
8	OCEANTEA	164
9	FAOSTAT	78
10	FishStatJ	27
11	LMU-ifo Economics & Business Data Center (EBDC)	119
12	AlpenDAC	336
13	Eurostat (only partly harvested for not flooding GeRDI index)	9.257
14	U.S. National Library of Medicine	5.908
15	OGLP	947
16	Open Data LMU	73
	SUMME	**567.036**

Thanks for your attention!

Cocktail & dinner time …