Generic Research Data Infrastructure

Forschungsdatenkolloquium am 7. Juli 2017

Wilhelm Hasselbring
The GeRDI Team

Prof. Dr. Klaus Tocthermann
Prof. Dr. Wolfgang E. Nagel
Prof. Dr. Hans-Joachim Bungartz
Dr. Christian Grimm
Prof. Dr. Wilhelm Hasselbring
Prof. Dr. Arndt Bode
Prof. Dieter Kranzlmüller
GeRDI Vision and Mission

• Vision:
  • Multidisciplinary and FAIR research data management principles are widely accepted.
  • The common application of these principles results in great benefits for research, industry, society and our environment.

• Mission:
  • Contribution to the vision (not yet explicitly defined) with initial focus on the present evaluation communities.
Envisioned GeRDI Architecture
(From the Proposal)
Evaluation Communities
GeRDI Evaluation Communities

Economics  
- SOEP

Life science, Humanities  
- CBG

Marine science  
- SEA around us
- FAO

Environmental science  
- Virtual Alpine Observatory

Christian-Albrechts-Universität zu Kiel

Technische Universität Dresden

Funded by
Generic Research Data Infrastructure · www.gerdi-project.eu
Munich communities (LRZ)

Alpine Environment Data Analysis Center (AlpEnDac)
- Environmental Science and Medicine
- Geophysical health data from Climate, Glaciology, Radiology research

Hydrology and River Basin Management (HIOS)
- Environmental Science
- Geophysical data from Hydrology: Water Resource and Decentralized Flood Management

UN International Strategy for Disaster Reduction (UNISDR)
- Socio-economic and geophysical data from research in: Disaster Risk Reduction and Sociology of Disaster Modelling
Dresden communities (TUD)

Microscopy and Bioinformatics (CBG)
- Cell Biology and Genetics
- Images and sequence data from Gentic

Digital Humanities (Prof. Crane)
- Automatic text analysis and philology
- Textual data from image, text and speech analysis

National Center for Tumor Diseases (NCT)
- Medical research in tumors
- Clinical and epidemiological studies and collection of biomaterial
Kiel communities (ZBW, CAU)

- Socio-Economic Panel (SOEP):
  Prof. Schupp, DIW
  - Socioeconomics
  - From wide ranging representative studies of private households: Qualitative & quantitative data on composition, occupation, earnings, ...

- Environmental, Resource and Ecological Economics (EREE):
  Prof. Quaas, Future Ocean

- Paleoceanography:
  Prof. Dullo, GEOMAR
Requirements Engineering for Community Involvement

- Identification of ...
  - research workflows
  - relevant data repositories

- Elicitation of community-specific use cases

- Extraction of generic workflows
EREE
Scientific Workflow, as is.
Intended enhanced WWF workflow
Use cases

<table>
<thead>
<tr>
<th>Use Case</th>
<th>Filter the results of the search</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>🔄 SAI-121 - Filter the search results</td>
</tr>
<tr>
<td>Community</td>
<td>All communities</td>
</tr>
<tr>
<td>Precondition</td>
<td>There are relevant search results that can be filtered</td>
</tr>
<tr>
<td>Trigger</td>
<td>The researcher wants to use the filter function</td>
</tr>
<tr>
<td>Goal</td>
<td>The researcher gets the filtered results of the search</td>
</tr>
</tbody>
</table>
| Main Scenario            | 1. The researcher clicks on filter he/she wants to apply  
                           | 2. The researcher applies the filter, he/she has chosen  
                           | 3. The results of the filtering are displayed |
| Extension                |                                   |
| Successful               | The researcher was able to limit the search results and so get the most relevant items |
| Postcondition            |                                   |
| Failure                  | The results of the filtering are not relevant or are empty |
| Postcondition            |                                   |
From requirements to implementation and automatic regression tests

• Use Case 2.0: Combination of use cases with agile software development

• Use Case Slices: Various flows through main and extended scenarios

• Described using acceptance tests formulated using BDD (test-driven development)

Feature: Main Workflow

As a researcher
I want to get filtered results of the search
So I can find the relevant data easier

Scenario: The researcher clicks on filter he wants to apply

Given I am on the search page
And there are search results
When I click on filter button
Then the filter options are displayed

Scenario: Applying filter option [...]
Modeling polyp activity of *Paragorgia arborea* using supervised learning

Arne N. Johanson\textsuperscript{a,}\textsuperscript{*}, Sascha Flögel\textsuperscript{b}, Wolf-Christian Dullo\textsuperscript{b}, Peter Linke\textsuperscript{b}, Wilhelm Hasselbring\textsuperscript{a}

\textsuperscript{a} Software Engineering Group, Kiel University, Germany
\textsuperscript{b} GEOMAR Helmholtz Centre for Ocean Research, Kiel, Germany
Data available in OceanTEA
http://maui.se.informatik.uni-kiel.de:9090/
Design and Implementation
GeRDI services and workflow cycles

Research data lifecycle

- creating data
- processing data
- analyzing data
- preserving data
- re-using data
- giving access to data

Research workflow

- define research question
- re-test (done by others)
- gather information and resources
- form hypothesis
- perform experiment
- analyze data
- interpret data and draw conclusions
- publish results
- Preprocessing
- Analysis
- Publication
- Collect / Staging / Save / Store
- Filter
- Search
- Harvest
- Repository

UK data archive

Crawford S, Stuck L (1990), “Peer review and the changing research record”
GeRDI’s Architectural Style

API Gateway, Page Assembly Proxy, ...

Repository
  Repo API
  Repo DB

Harvest
  Harvest API
  Harvest DB

Search
  Query/Index API
  Elastic-search DB

Filter
  Query API
  Query DB

Collect / Staging / Save / Store
  Store API
  Local Storage

Preprocess
  Jupyter API
  Local Storage

Analysis
  Jupyter API
  Local Storage

Publication
  JIRA API
  JIRA Storage

REST, Messaging, Prospector AAI, Scalability/Elasticity, Monitoring, Control Center, ...

Funded by
Generic Research Data Infrastructure · www.gerdi-project.eu
Design Rationale

• Entry / Exit Options for users at various points in the workflow

• Self-contained Systems (SCS)
  • Also known as microservices (http://scs-architecture.org)

  • Each SCS is an autonomous (web) application: For its domain all data, the logic to process that data and all code to render the web interface is contained within the SCS.

  • Communication with each other or 3rd party should be asynchronous: This decouples the systems, reduces the effects of failure, and thus supports autonomy.

  • SCS should not share business code to avoid tight coupling.
GeRDI services and community workflows (current snapshot at CAU)

WWF Paper (EREE, Prof. Quaas)

- Repository
  - • SEA AROUND US (SAU)
  - • FAO Stat
  - • FAO FishStatJ
  - • SSP Scenarios
  - • GIS data

- Harvest
  - • Prototype: Uses generic library and adapter for FAO Stat and SAU

- Search
  - • Prototype: Elastic Search

- Filter
  - • LMEs
  - • Catch data
  - • Prices
  - • Trade data
  - • GIS for LMEs & Countries

- Collect / Staging / Save / Store

- Preprocessing
  - • Union of GIS data
  - • LME catches and global prices
  - • ...

- Analysis
  - • Data analysis
  - • Combining models
  - • Prediction

- Publication

OceanTEA (Bio & Environment, Prof. Dullo)

- Repository
  - • Manage Time Series
  - • Time Series Exploration

- Harvest

- Search

- Filter

- Collect / Staging / Save / Store

- Preprocessing
  - • Spatial Analysis

- Analysis
  - • Temporal Pattern Discovery
  - • P. Aborea Activity

- Publication
Mediator Architecture and Metadata

GeRDI as a software product line

- “A software product line (SPL) is a set of software-intensive systems that share a common, managed set of features satisfying the specific needs of a particular market segment or mission and that are developed from a common set of core assets in a prescribed way.”
  [Software Engineering Institute, CMU]

- Software product lines or application families are applications with generic functionality that can be adapted and configured for use in a specific context.
Demo

Live demo our search prototype
Without Gerdi

(1 / 5)

Making the international trade in plants and seeds a safer venture

International plant health body adopts new global standards
Without Gerdi

(2 / 5)

International Conference on Agricultural Statistics VII
Modernization of Agricultural Statistics in Support of the Sustainable Development Agenda

Statistics at FAO

FAO develops methods and standards for food and agriculture statistics, provides technical assistance services and disseminates data for global monitoring. Statistical activities at FAO include the development and implementation of methodologies and standards for data collection, validation, processing and analysis. FAO also plays a vital part in the global compilation, processing and dissemination of food and agriculture statistics, and provides essential statistical capacity development to member countries.

FAO has a decentralized statistical system and statistical activities cover the areas of agriculture, forestry and fisheries, land and water resources and use, climate, environment, population, gender, nutrition, poverty, rural development, education and health as well as many others.

FAO Statistical Programme of Work

The FAO Statistical Programme of Work is a collaborative effort that is overseen by the Chief Statistician and supported by the inter-Departmental Working Group on Statistics. These two mechanisms ensure strengthened coordination and cooperation on statistical matters and guarantee the high quality of FAO data.

The Statistical Programme of Work provides a summary of all the principal statistical activities at FAO, and a detailed description of all the individual statistical activities carried out by FAO Divisions active in the field of statistics. It presents the organization's operational activities according to different statistical categories and domains.

Related links

- FAO statistics website
- Food security statistics
- Forestry statistics
- Fisheries and aquaculture
- Trade and markets
- Water: information and statistics
- Coordinating Working Party on Fishery Statistics

Data products
Without Gerdi
(4 / 5)
Integrated search in multiple repositories with GeRDI
Work Packages

- Requirement & Communities
  - Cases Studies
  - Analysis and Specifications
  - Community Management
  - Data Curation
  - Evaluation and Proof of Concept
- System Architecture & Integration
  - Software Management
  - Software und System Architecture
  - IT Security
  - Setup and Integration of Pilot Center
- Data & Metadata
  - Data Management
  - Metadata Management
- Sustainability
  - Training Concept
  - Future Operation Model
  - Roll-Out Preparation
- Project Management

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

Funded by DFG
Summary & Outlook

• GeRDI as Contribution to the European Open Science Cloud
  https://www.youtube.com/watch?v=SC4-O8Bml4I
• Funding: 3 M€ for three years (first phase)
• KOLab
  • Kiel Open Data and Software Lab
• Digital Ocean 2017
  • 20. September 2017, 12-18 h at ZMB
    https://www.kms.uni-kiel.de/de/veranstaltungen/digital_ocean2017
• More to come at
  • http://www.gerdi-project.de/
References


