Generic Research Data Infrastructure

Forschungsdatenkolloquium am 7. Juli 2017

Wilhelm Hasselbring

Funded by DFG

The GeRDI Team

Deutsche Forschungsgemeinschaft DFG

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

Prof. Dr. Klaus Tochtermann

TECHNISCHE UNIVERSITÄT DRESDEN

Prof. Dr. Wolfgang E. Nagel

DFN Deutsches Forschungsnetz

Prof. Dr. Hans-Joachim Bungartz
Dr. Christian Grimm

CAU Christian-Albrechts-Universität zu Kiel

Prof. Dr. Wilhelm Hasselbring

lrz

Prof. Dr. Arndt Bode
Prof. Dieter Kranzlmüller

Funded by DFG

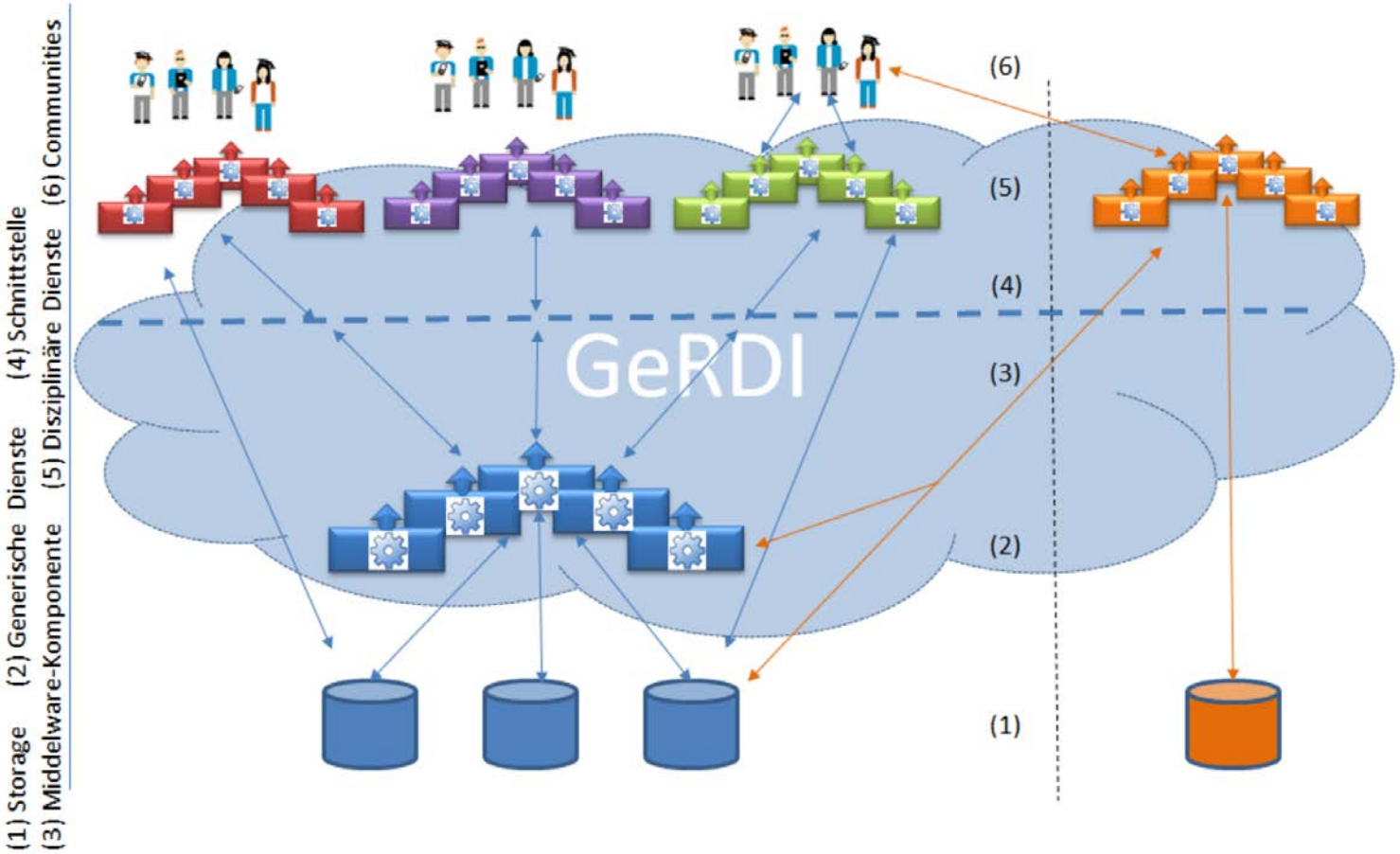
GeRDI Vision and Mission

- 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.

Funded by DFG

Envisioned GeRDI Architecture (From the Proposal)



Evaluation Communities

GeRDI Evaluation Communities

ZBW Leibniz-Informationszentrum Wirtschaft
Leibniz Information Centre for Economics

TECHNISCHE UNIVERSITÄT DRESDEN

Economics	Life science, Humanities
SOEP	CBG Max Planck Institute of Molecular Cell Biology and Genetics; dh
Marine science	**Environmental science**
SEA AROUND US; FAO FIAT PANIS	Virtual Alpine Observatory; 

CAU Christian-Albrechts-Universität zu Kiel

lrz

Funded by DFG

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 Gentics

Digital Humanities (Prof. Crane)

- Automatic text analysis and philology
- Textual data from image, text and speech analysis

National Center for Tumor Diseases (NCT)

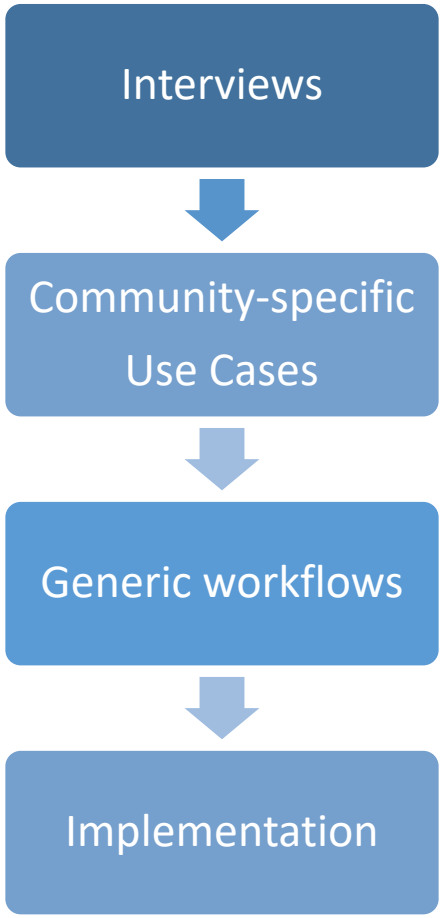
- Medial research in tumors
- Clinical and epidemiological studies and collection of biomaterial

Funded by DFG

Kiel communites (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, …
- Enviromental, Resource and Ecological Ecomomics (EREE):
  Prof. Quaas, Future Ocean
- Paleoceanography:
  Prof. Dullo, GEOMAR

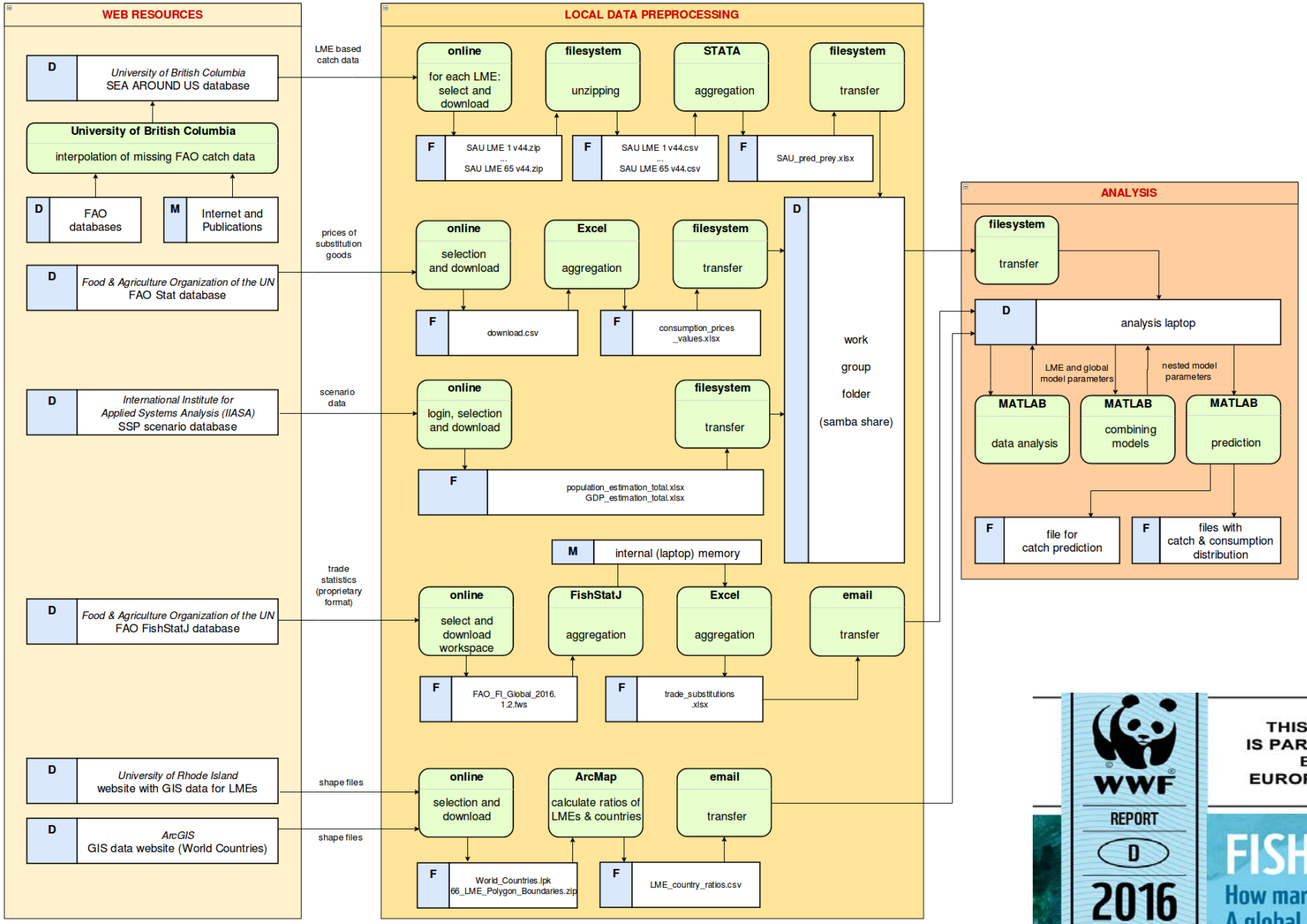
Requirements Engineering for Community Involvement



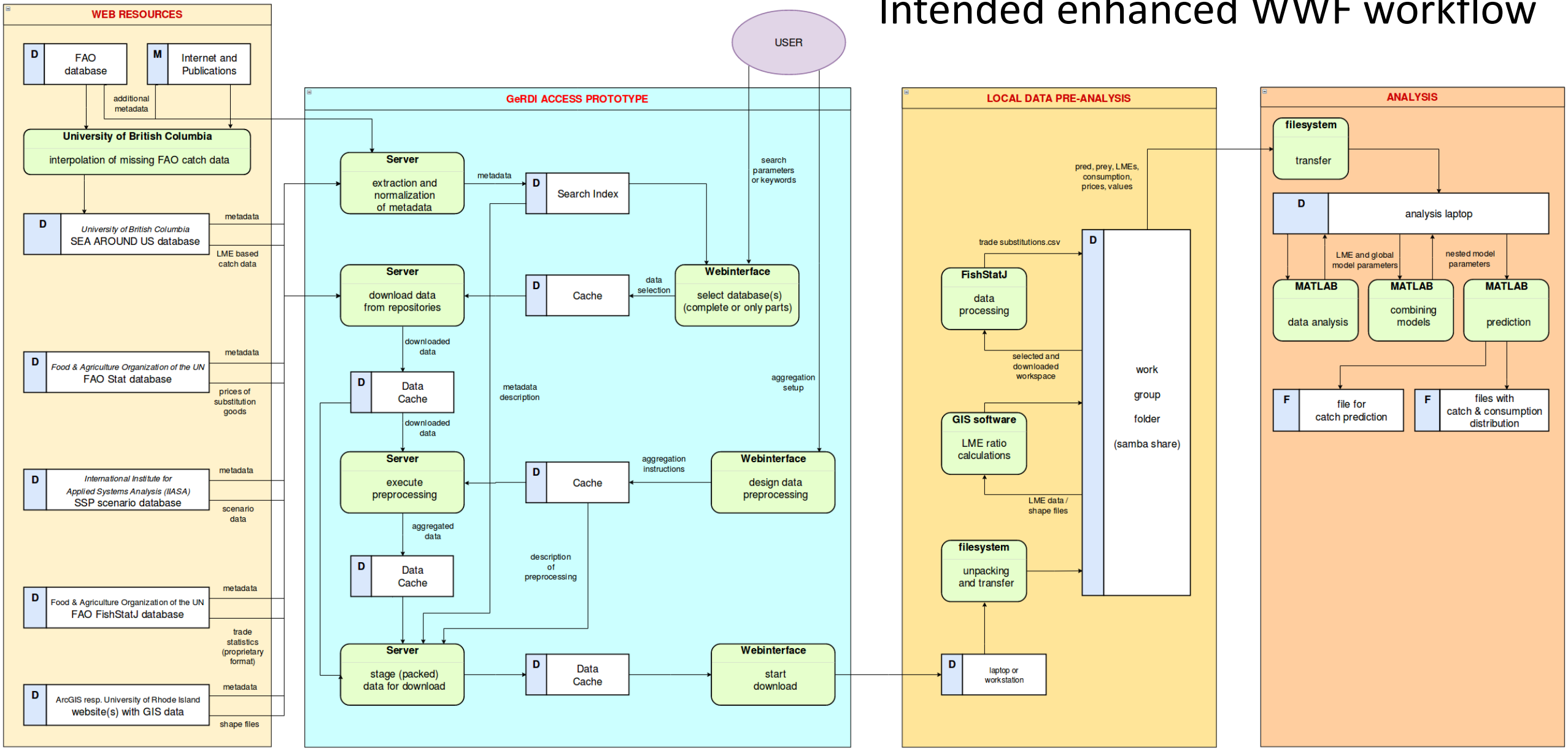
- Identification of …
  - research workflows
  - relevant data repositories
- Elicitation of community-specific use cases
- Extraction of generic workflows

Funded by DFG

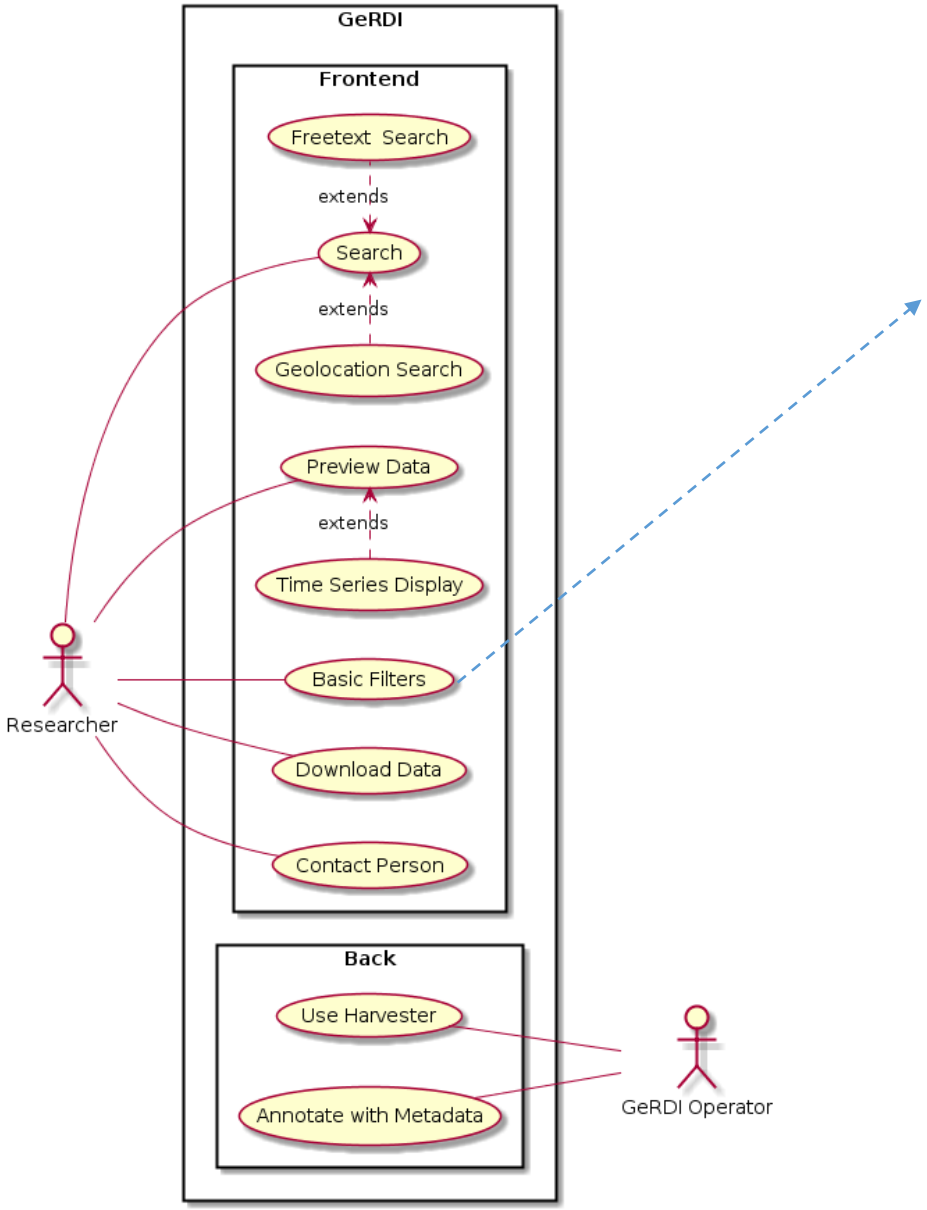
EREE Scientific Workflow, as is.



Intended enhanced WWF workflow



Use cases



Use Case	Filter the results of the search
ID	SAI-121 - Filter the search results BACKLOG
Community	All communities
Precondition	There are relevant search results that can be filtered
Trigger	The researcher wants to use the filter function
Goal	The researcher get the filtered results of the search
Main Scenario	1. The researcher click on filter he/she wants to apply 2. The researcher apply the filter, he/she has chosen 3. The results of the filtering are displayed
Extension	
Successful Postcondition	The researcher was able to limit the search results and so get the most relevant items
Failure Postcondition	The results of the filtering are not relevant or are empty

Funded by DFG

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 […]

Use Case 2.0
https://www.ivarjacobson.com/publications/white-papers/use-case-ebook

Funded by DFG

Contents lists available at ScienceDirect

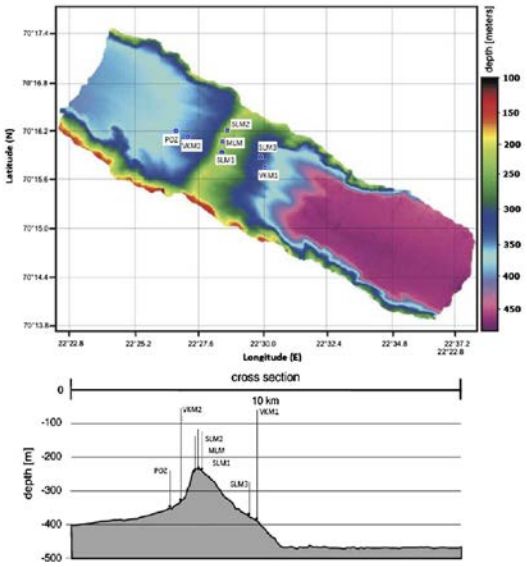
Ecological Informatics

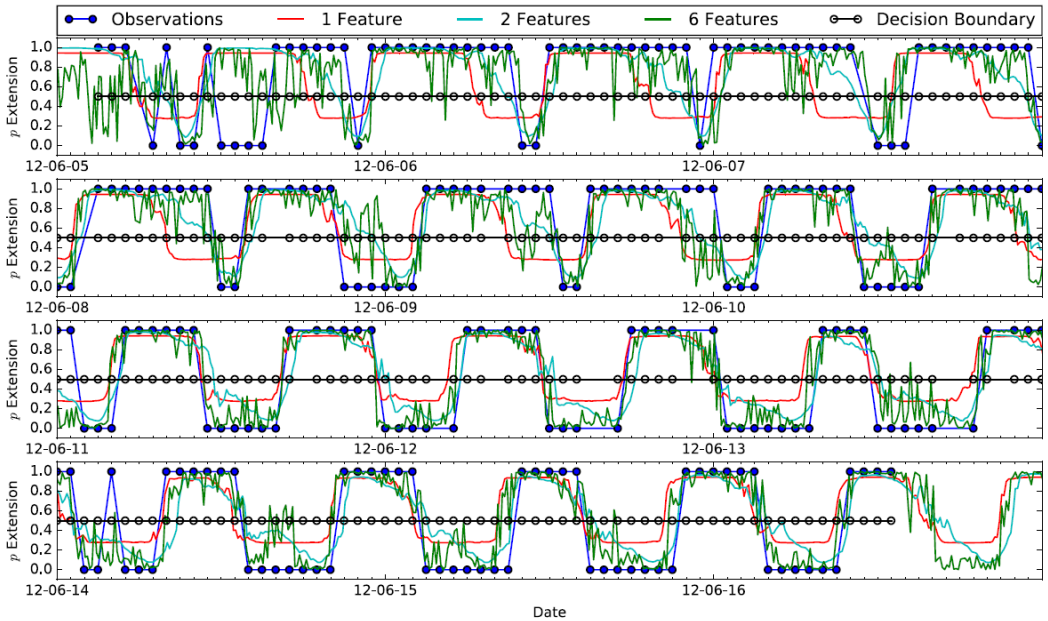
journal homepage: www.elsevier.com/locate/ecolinf

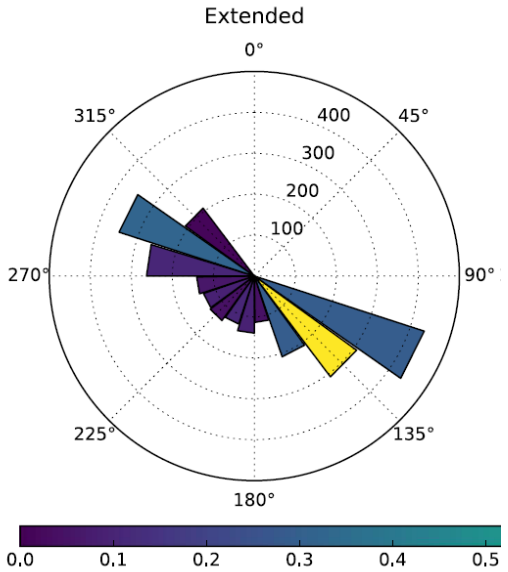
Modeling polyp activity of *Paragorgia arborea* using supervised learning

Arne N. Johanson[a,*], Sascha Flögel[b], Wolf-Christian Dullo[b], Peter Linke[b], Wilhelm Hasselbring[a]

[a] *Software Engineering Group, Kiel University, Germany*
[b] *GEOMAR Helmholtz Centre for Ocean Research, Kiel, Germany*

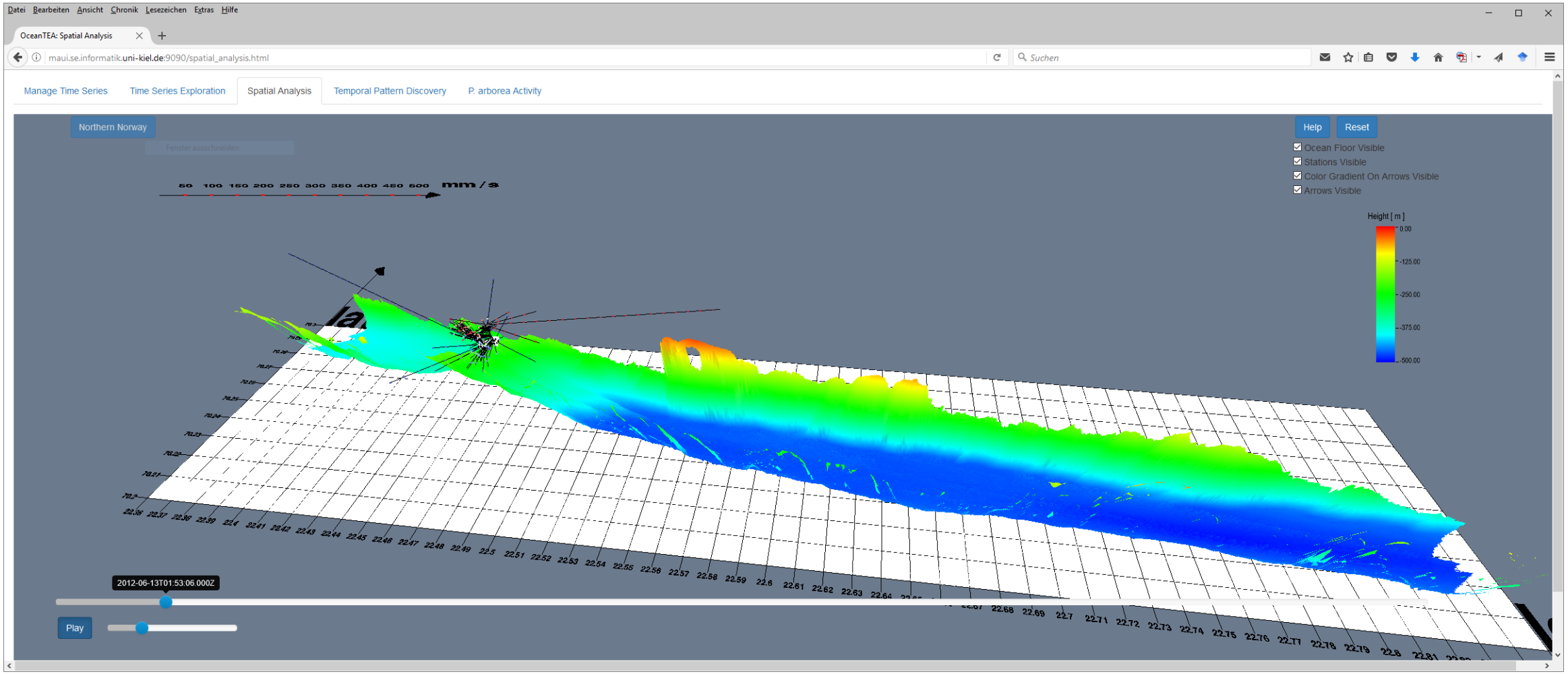




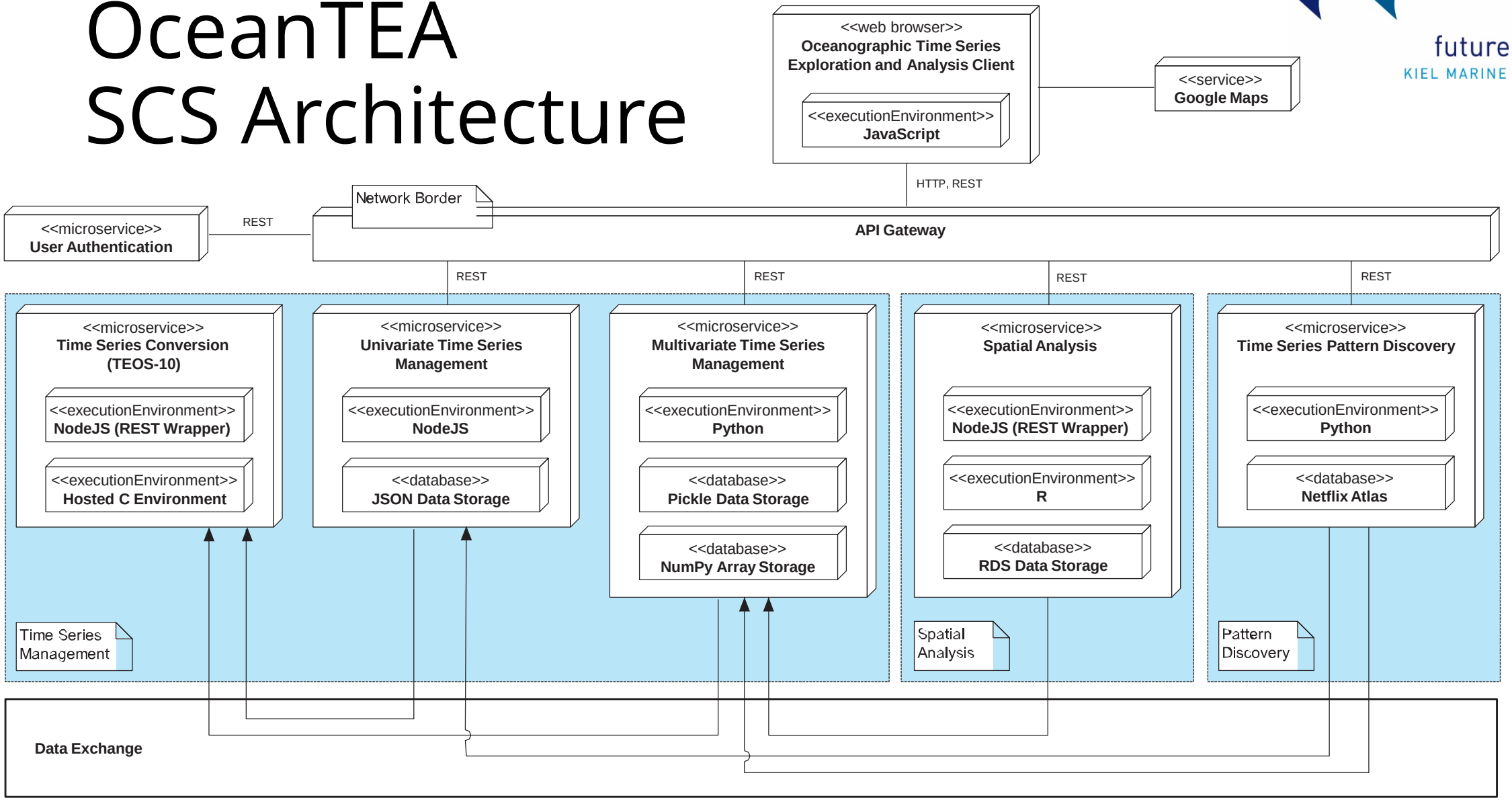


Data available in OceanTEA

http://maui.se.informatik.uni-kiel.de:9090/



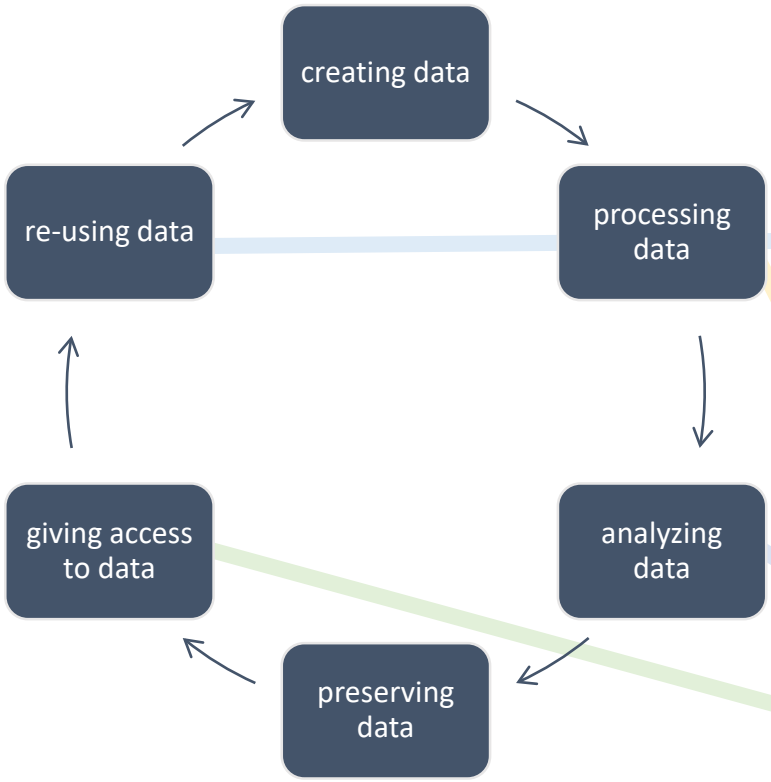
OceanTEA SCS Architecture



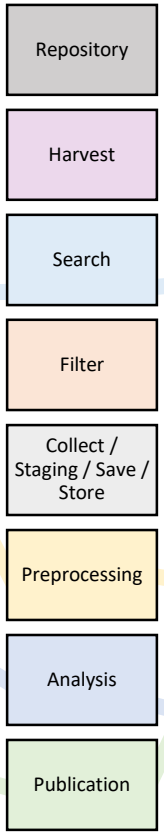
Design and Implementation

GeRDI services and workflow cycles

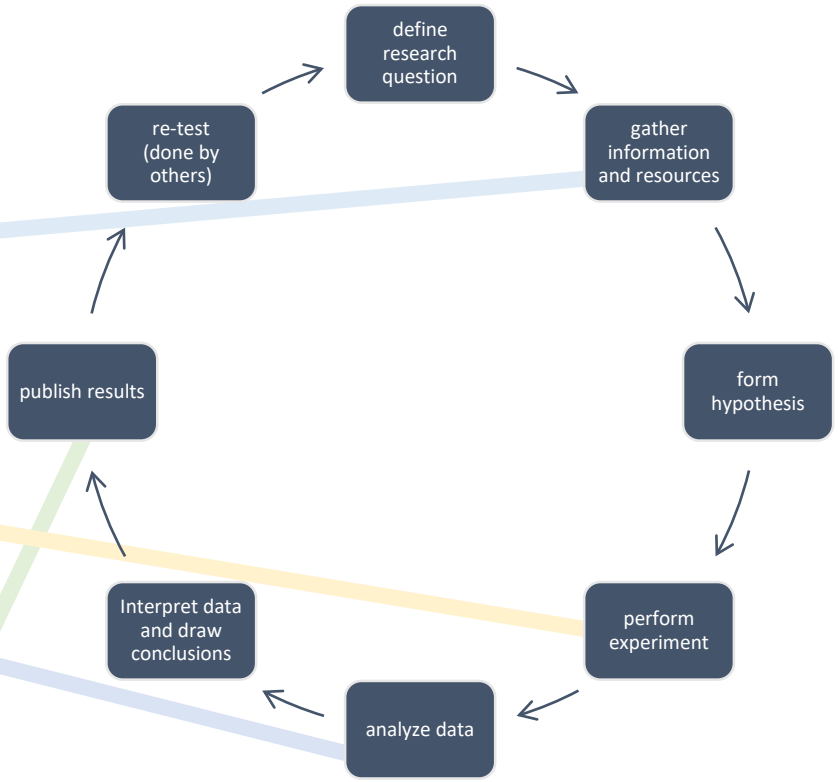
Research data lifecycle



Services



Research workflow

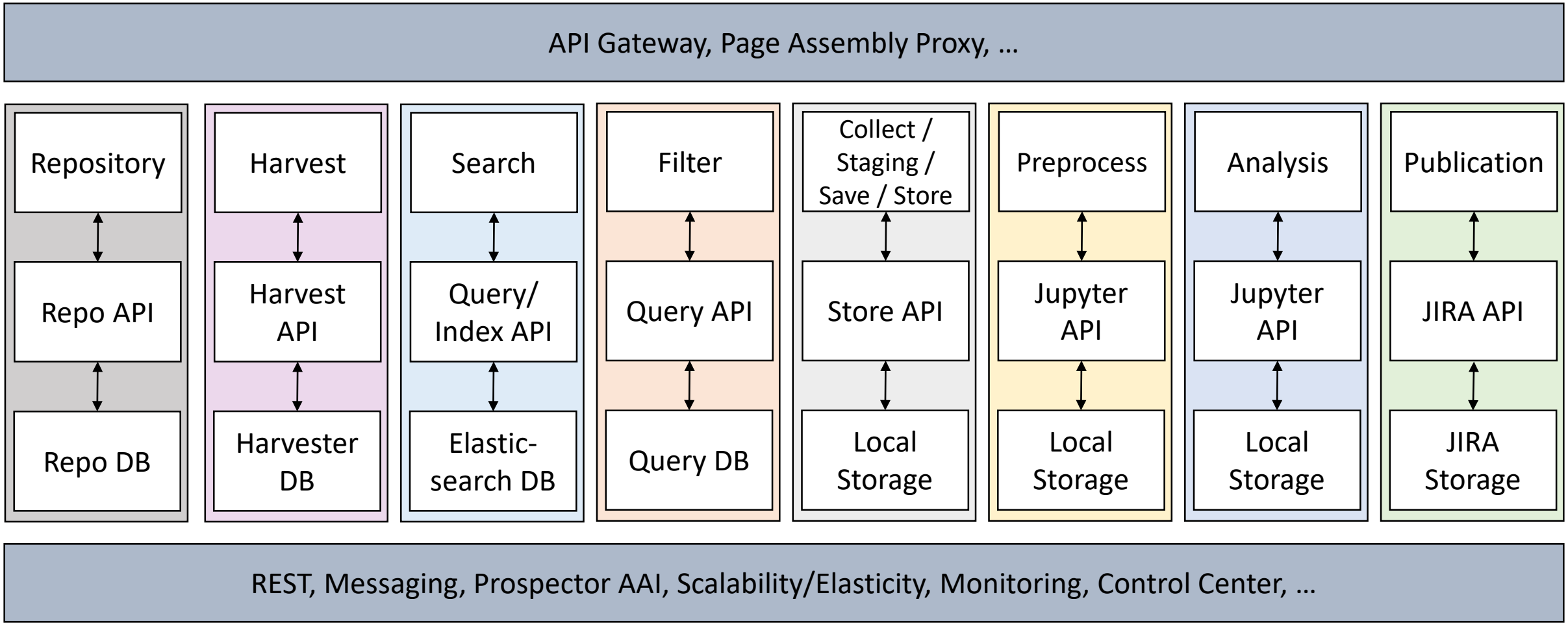


UK data archive

Crawford S, Stuck L (1990), “Peer review and the changing research record”

Funded by DFG

GeRDI's Architectural Style



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.

Funded by DFG

GeRDI services and community workflows (current snapshot at CAU)

WWF Paper (EREE, Prof. Quaas)

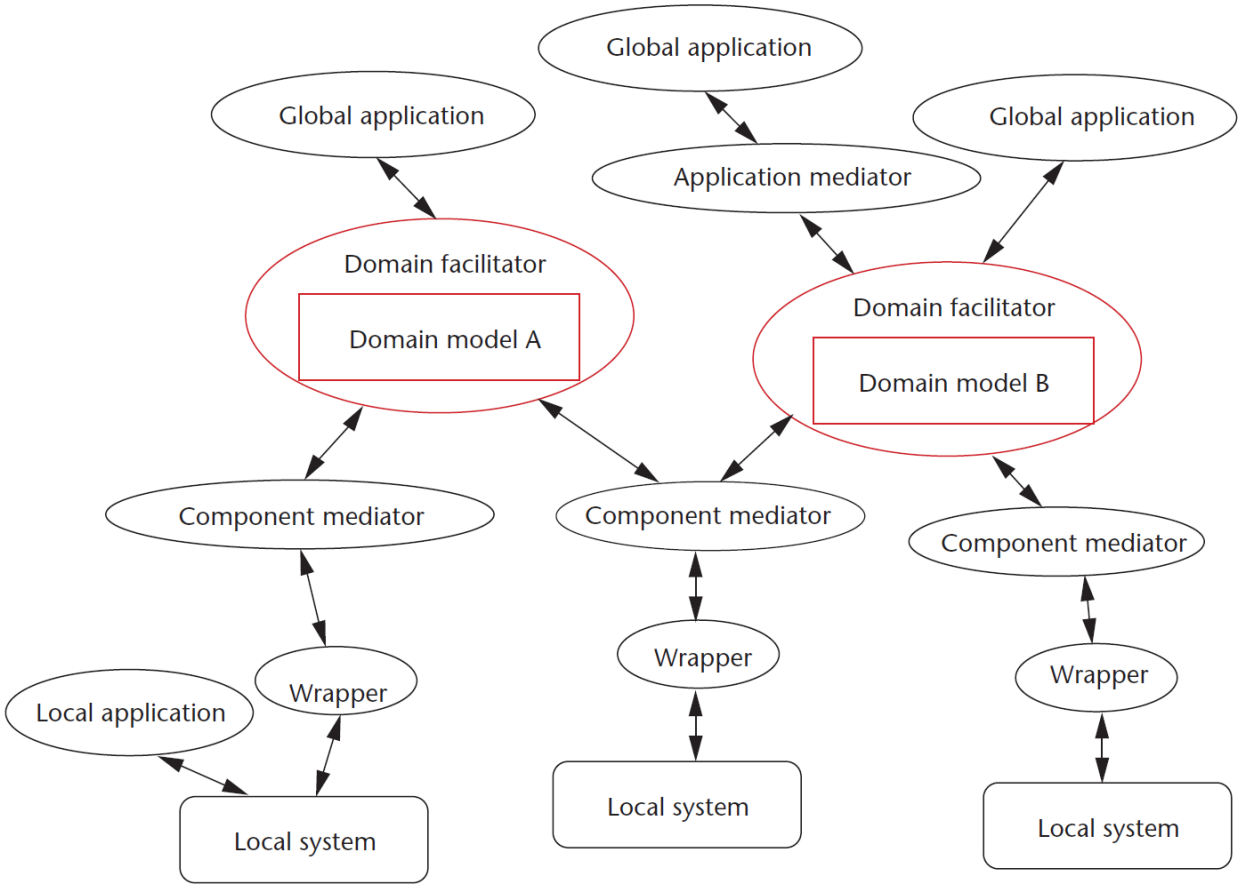
Repository	Harvest	Search	Filter	Collect / Staging / Save / Store	Preprocessing	Analysis	Publication
• SEA AROUND US (SAU) • FAO Stat • FAO FishStatJ • SSP Scenarios • GIS data	• **Protoype**: Uses generic library and adapter for FAO Stat and SAU	• **Protoype**: Elastic Search	• LMEs • Catch data • Prices • Trade data • GIS for LMEs & Countries		• Union of GIS data • LME catches and global prices • …	• Data analysis • Combining models • Prediction	

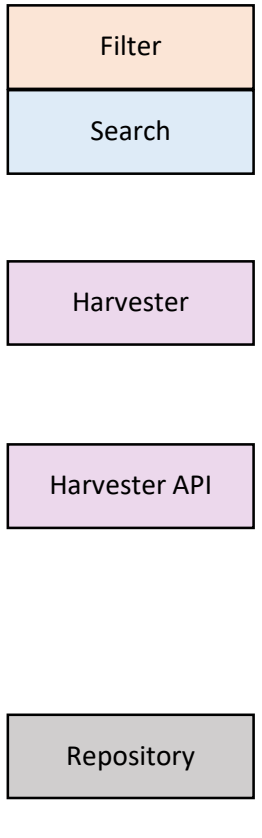
OceanTEA (Bio & Environment, Prof. Dullo)

Repository	Harvest	Search	Filter	Collect / Staging / Save / Store	Preprocessing	Analysis	Publication
• Manage Time Series • Time Series Exploration					• Spatial Analysis	• Temporal Pattern Discovery • P. Aborea Activity	

Funded by DFG

Mediator Architecture and Metadata





Source: Hasselbring, W. (2002) “Web Data Integration for E-Commerce Applications”, IEEE Multimedia, 9 (1). pp. 16-25. DOI 10.1109/93.978351.

GeRDI as a software product line

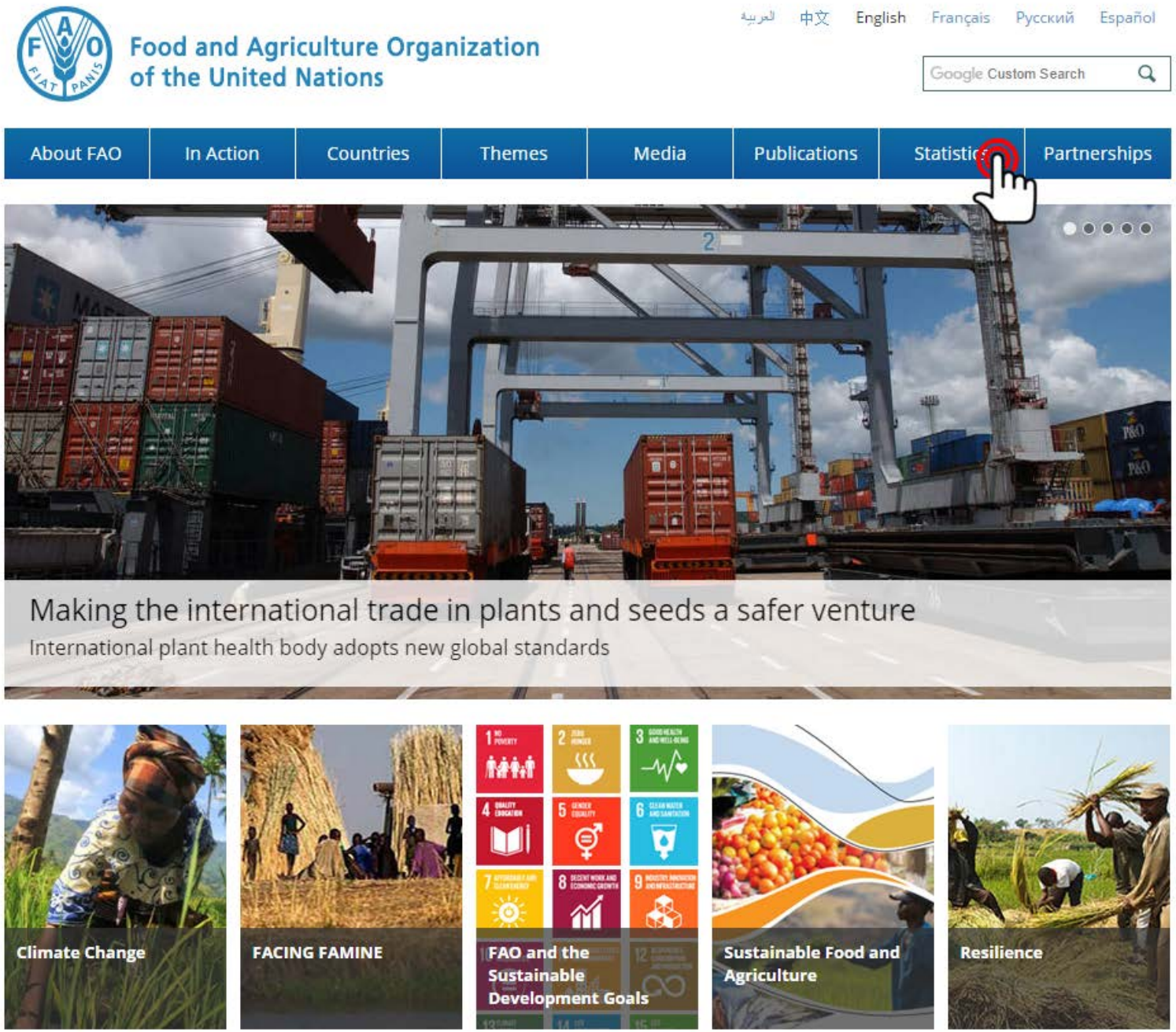
- “A software product line (SPL) is a set of software-intensive systems that share a common, managed set of **features** satisfying the specific needs of a particular market segment or mission and that are developed from a common set of core assets in a prescribed way.”
  [Software Engineering Institute, CMU]
- Software product lines or application families are applications with **generic** functionality that can be adapted and configured for use in a specific context.

Funded by DFG

Demo

Live demo our search prototype

Without Gerdi (1 / 5)



Without Gerdi (2 / 5)

Home > Statistics

Send Print

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 of 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.

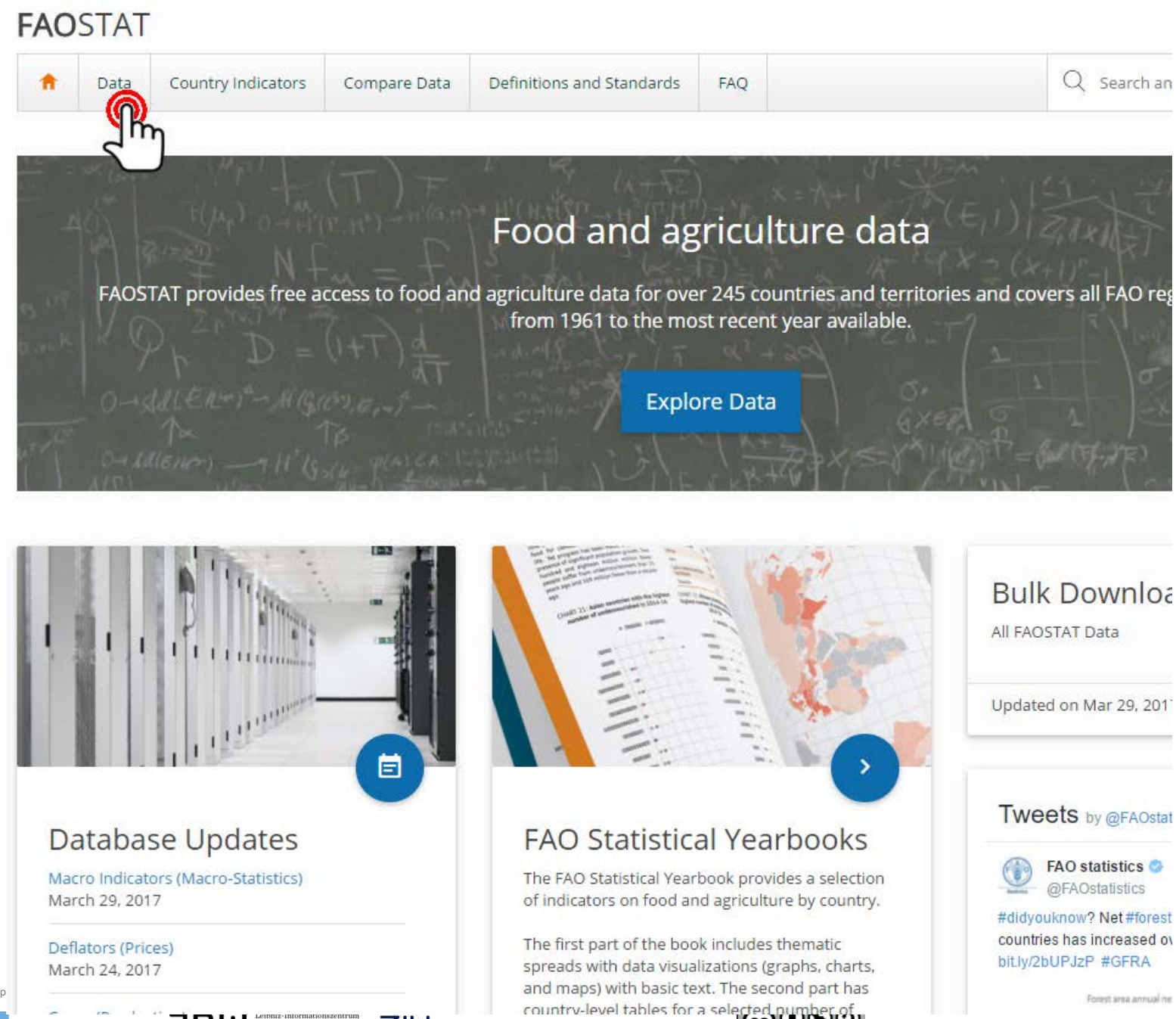
Get FAO statistics

Related links

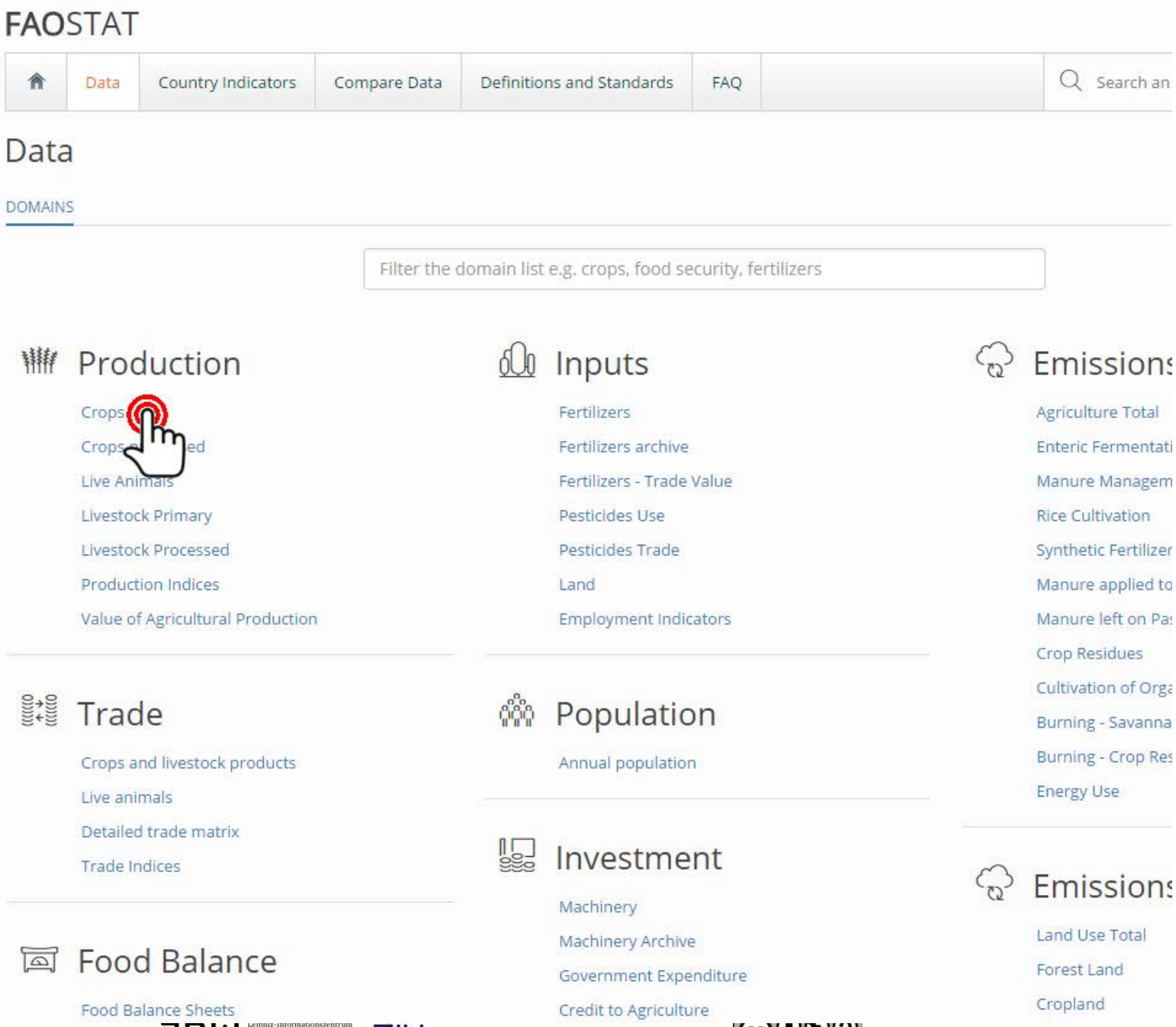
- FAO statistics website
- Food security statistics
- Forestry statistics
- Fisheries and aquaculture
- Trade and markets
- Water information and statistics
- Coordinating Working Party on Fishery Statistics

Data products

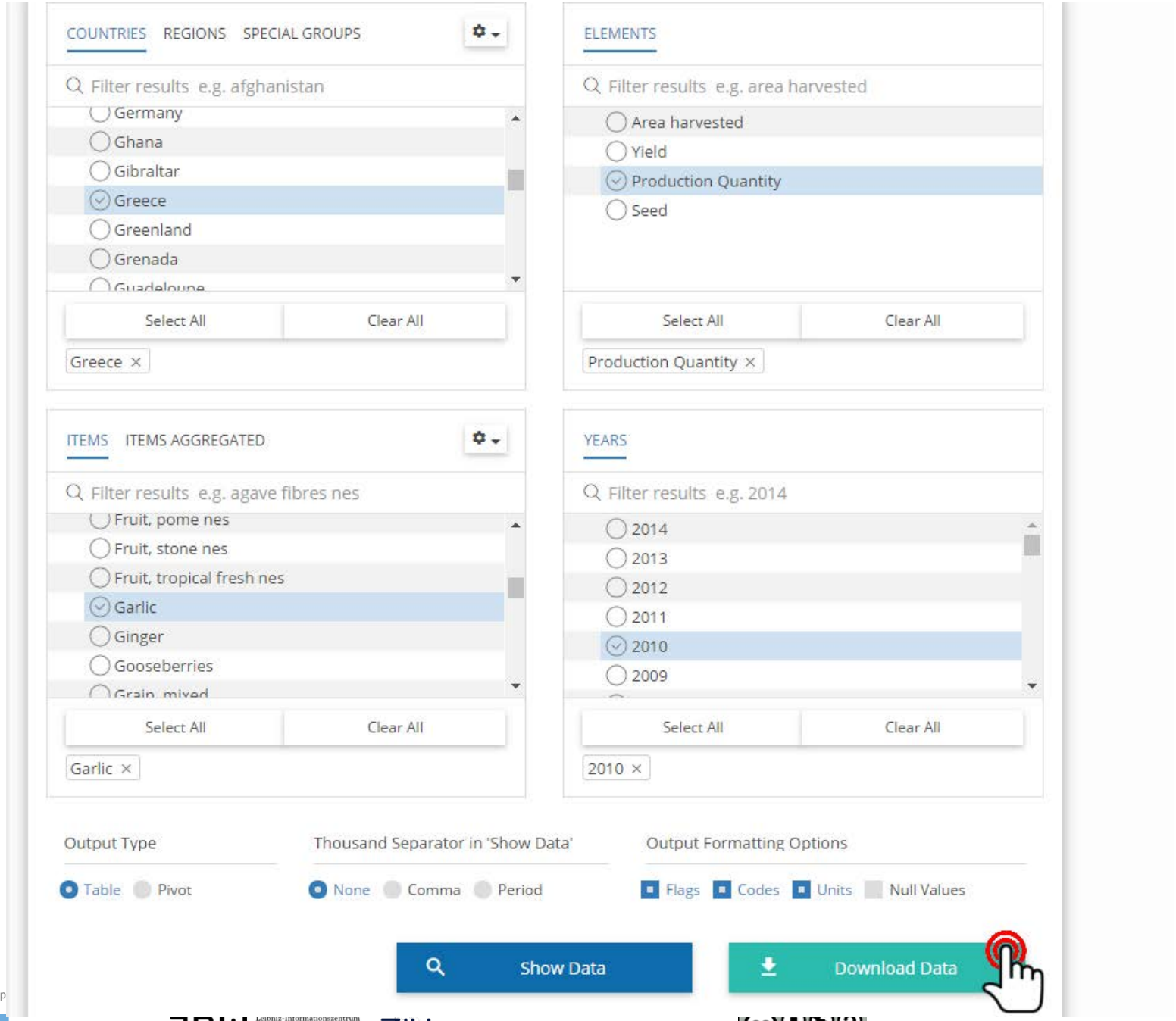
Without Gerdi (3 / 5)



Without Gerdi (4 / 5)



Without Gerdi (5 / 5)



Integrated search in multiple repositories with GeRDI

fish AND Norway

6 results found in 25 ms

Commodity Balances - Livestock and Fish Primary Equivalent

http://www.fao.org/faostat/en/#data/BL

Food supply data is some of the most important data in FAOSTAT. In fact, this data is for the basis for estimation of global and national undernourishment assessment, when it is combined with parameters and other data sets. This data has been the foundation of food balance sheets ever since they were first constructed. The data is accessed by both business and governments for economic analysis and policy setting, as well as being used by the academic community.

Food Supply - Livestock and Fish Primary Equivalent

http://www.fao.org/faostat/en/#data/CL

Food supply data is some of the most important data in FAOSTAT. In fact, this data is for the basis for estimation of global and national undernourishment assessment, when it is combined with parameters and other data sets. This data has been the foundation of food balance sheets ever since they were first constructed. The data is accessed by both business and governments for economic analysis and policy setting, as well as being used by the academic community.

Food Balance Sheets

http://www.fao.org/faostat/en/#data/FBS

Food Balance Sheet presents a comprehensive picture of the pattern of a country's food supply during a specified reference period. The food balance sheet shows for each food item - i.e. each primary commodity and a number of processed commodities potentially available for h
foodstuffs produced in a country added to the total quantity imported and adjusted to any change
the supply available during that period. On the utilization side a distinction is made between the c
and non-food uses, losses during storage and transportation, and food supplies available for hun

Live animals

http://www.fao.org/faostat/en/#data/TA

The food and agricultural trade dataset is collected, processed and disseminated by FAO accord
is mainly provided by UNSD, Eurostat, and other national authorities as needed. This source dat
missing cells, and data on food aid is added to take into account total cross-border trade flows. T
import quantity and import value. The trade database includes all food and agricultural products i

Trade Indices

http://www.fao.org/faostat/en/#data/TI

The food and agricultural trade dataset is collected, processed and disseminated by FAO according to the standard International Merchandise Trade Statistics Methodology. The data is mainly provided by UNSD, Eurostat, and other national authorities as needed. This source data is checked for outliers, trade partner data is used for non-reporting countries or missing cells, and data on food aid is added to take into account total cross-border trade flows. The trade database includes the following variables: export quantity, export value, import quantity and import value. The trade database includes all food and agricultural products imported/exported annually by all the countries in the world.

Crops and livestock products

http://www.fao.org/faostat/en/#data/TP

The food and agricultural trade dataset is collected, processed and disseminated by FAO according to the standard International Merchandise Trade Statistics Methodology. The data is mainly provided by UNSD, Eurostat, and other national authorities as needed. This source data is checked for outliers, trade partner data is used for non-reporting countries or missing cells, and data on food aid is added to take into account total cross-border trade flows. The trade database includes the following variables: export quantity, export value, import quantity, and import value. The trade database includes all food and agricultural products imported/exported annually by all the countries in the world.

Commodity Balances - Livestock and Fish Primary Equivalent

FAOSTAT

Go to | Download

DESCRIPTION

Food supply data is some of the most important data in FAOSTAT. In fact, this data is for the basis for estimation of global and national undernourishment assessment, when it is combined with parameters and other data sets. This data has been the foundation of food balance sheets ever since they were first constructed. The data is accessed by both business and governments for economic analysis and policy setting, as well as being used by the academic community.

ADDITIONAL METADATA

Funded by DFG

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**

Funded by DFG

Summary & Outlook

- GeRDI as Contribution to the European Open Science Cloud https://www.youtube.com/watch?v=SC4-O8BmI4I
- 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/

Funded by DFG

GeRDI

References

(1) Grunzke, R., Adolph, T., Biardzki, C., Bode, A., Borst, T., Bungartz, H. J., Busch, A., Frank, A., Grimm, C., Hasselbring, W., Kazakova, A., Latif, A., Limani, F., Neumann, M., de Sousa, N. T., Tendel, J., Thomsen, I., Tochtermann, K., Müller-Pfefferkorn, R. and Nagel, W. E. (2017): "Challenges in Creating a Sustainable Generic Research Data Infrastructure" In: Softwaretechnik-Trends, 37 (2).

(2) Hasselbring, W. and Steinacker, G. (2017): "Microservice Architectures for Scalability, Agility and Reliability in E-Commerce" In: IEEE International Conference on Software Architecture 2017, April 03-07, 2017, Gothenburg, Sweden.

(3) Hasselbring, W. (2002): "Web Data Integration for E-Commerce Applications", IEEE Multimedia, 9 (1). pp. 16-25. DOI 10.1109/93.978351.

(4) Hasselbring, W. (2016): "Microservices for Scalability" In: International Conference on Performance Engineering (ICPE 2016), March 2016, Delft, Netherlands.

(5) Johanson, A., Flögel, S., Dullo, W. C., Linke, P. and Hasselbring, W. (2017): "Modeling Polyp Activity of Paragorgia arborea using Supervised Learning" In: Ecological Informatics, 39 . pp. 109-118. DOI 10.1016/j.ecoinf.2017.02.007.

(6) Johanson, A., Flögel, S., Dullo, C. und Hasselbring, W. (2016): "OceanTEA: Exploring Ocean-Derived Climate Data Using Microservices" In: Sixth International Workshop on Climate Informatics (CI 2016), September 22-23. 2016, Boulder, Colorado.