Integrated search and analysis of multidisciplinary marine data with GeRDI

Ingo Thomsen, Software Engineering, Kiel University (Germany), int@informatik.uni-kiel.de
Whilhelm Hasselbring, Software Engineering, Kiel University (Germany), wha@informatik.uni-kiel.de
Jörn Schmidt, Economics, Kiel University (Germany), jschmidt@economics.uni-kiel.de
Martin Quaas, Economics, Kiel University (Germany), quaas@economics.uni-kiel.de

An exemplary research question: "How marine fisheries impact on global food security up to 2050"

Multidisciplinary research usually requires data from more than one data repository that has to be retrieved and analyzed. Figure 1 outlines the dataflow addressing such an exemplary research question: data from multiple discipline-specific repositories was aggregated and analyzed. The results were published as part of a WWF report [1].

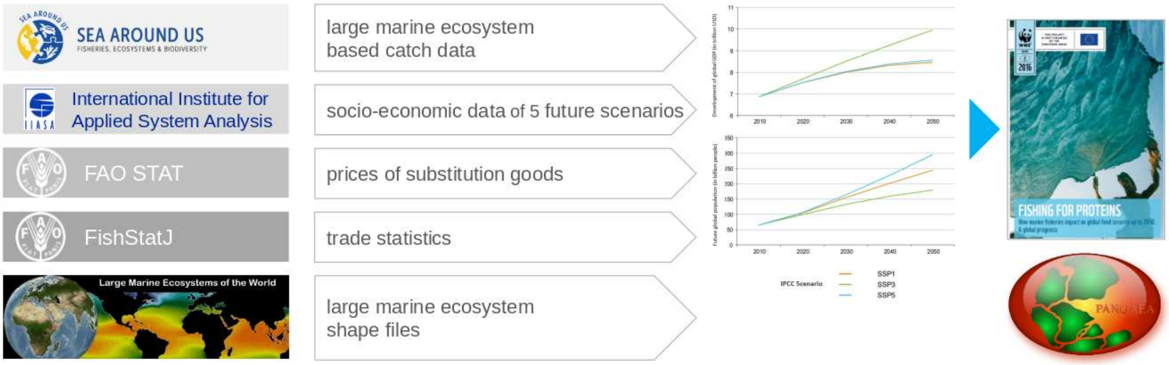
SEA AROUND US
FISHERIES, ECOSYSTEMS & BIODIVERSITY
International Institute for Applied System Analysis
IIASA
FAO STAT
FishStatJ
Large Marine Ecosystems of the World
large marine ecosystem based catch data
socio-economic data of 5 future scenarios
prices of substitution goods
trade statistics
large marine ecosystem shape files
Development of global GDP (in trillion USD)
11
10
9
8
7
6
2010
2020
2030
2040
2050
Future global population (in billion people)
350
300
250
200
150
100
50
0
2010
2020
2030
2040
2050
IPCC Scenario
SSP1
SSP3
SSP5
2016
FISHING FOR PROTEINS
PANGAEA

Figure 1: Dataflow for the creation of the WWF report

In this example, parameters for bio-economic fishery models are statistically estimated using catch and price data from three main sources: Sea Around Us [2], the FAO database FAOStat [3] and the FishStatJ fishery databases [4]. Information on the area of Exclusive Economic Zones (EEZ) and Large Marine Ecosystems (LMEs) are taken from the LME database [5]. The model is finally based on scenarios for total expenditures for protein-rich food, and the availability of protein-rich food other than wild capture fish using GDP and population data derived from model output from IIASA using the Shared Socioeconomic Pathways from IPCC [6].

Development of a generic research data infrastructure – driven by research questions

The GeRDI project [7] (www.gerdi-project.de) focuses on the development of a sustainable Generic Research Data Infrastructure. Its goal is to enable scientists to search, use and re-use external research data. In the current pilot phase, the software development is driven by research questions – including the exemplary one above. These questions originate from participating communities in various research disciplines – marine sciences, but also digital humanities, bioinformatics, and others.

The GeRDI services are implemented in a modular manner as microservices [8] as outlined in Figure 2. They communicate through well-defined protocols. Software and protocols are published as open-source. This offers the potential to "plug-in" and to replace parts with your own specialized services.

GeRDI offers an integrated web-interface to search repositories (for instance with ocean related data) whose metadata was previously harvested - preferably employing an open protocol (OAI-PMH [9]). An established metadata scheme (DataCite [10]) was adapted as a base for the search index. The faceted search – filtering using time, categories, etc. – can also be based on geolocation with the help of an interactive map. The results can be saved as a set of bookmarks that offer the download links and/or instructions for accessing the data. This set can be saved, modified and re-used, which supports repeatability and sharing of research workflows and experiments.

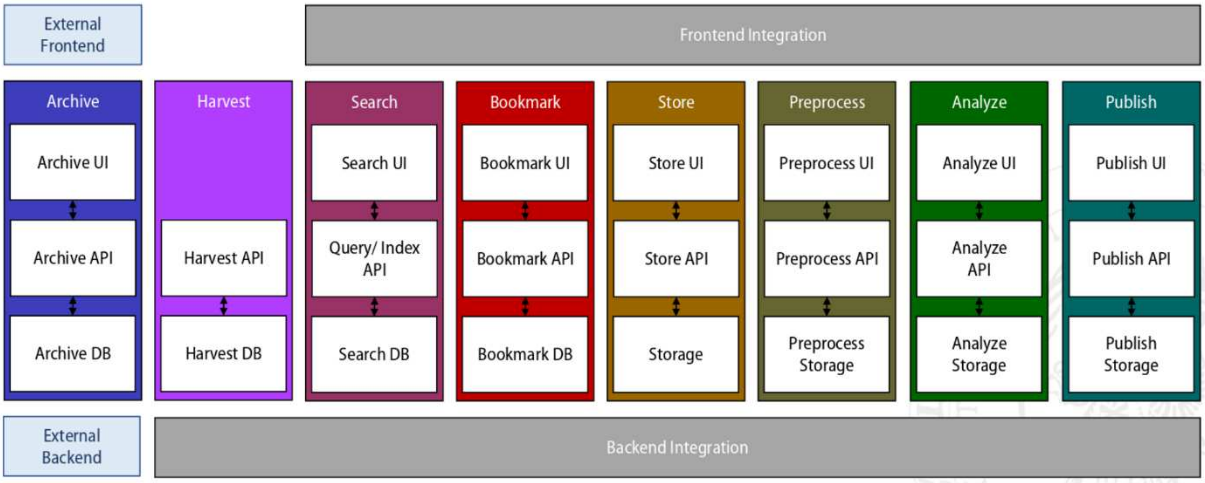

Figure 2: GeRDI Microservice-based software architecture [8]

The first three services in Figure 2 are based upon metadata: (1) **Harvest** existing repositories, (2) generic keyword-based **Search**, and (3) a persistent **Bookmark** of selected data sets. The example in Figure 1 indicates that the research results are published in the PANGAEA repository [11]. This leads to the extended services in GeRDI which handle the actual data (in addition to the metadata): **Store** bookmarked data sets (locally or in a cloud), **Preprocess** the data as preparation and then **Analyze** it ("gaining new insights") and finally **Publish** it. **Archive** represents the data repositories that offer upload facilities, thus closing the research data cycle.

References

[1] Quaas MF, Hoffmann J, Kamin K, Kleemann L and Schacht K (2016), "Fishing for Proteins. How marine fisheries impact on global food security up to 2050. A global prognosis", WWF, Hamburg

[2] Sea Around Us, University of British Columbia, www.seaaroundus.org (retrieved 2018)

[3] FAOSTAT - Food and Agriculture Organization Corporate Statistical Database, www.faostat.org (retrieved 2018)

[4] FishStatJ - software for fishery statistical time series. FAO Fisheries and Aquaculture Department, www.fao.org/fishery/statistics/software/fishstatj/ (retrieved 2018)

[5] NOAA - National Oceanic and Atmospheric Administration, www.st.nmfs.noaa.gov/ecosystems/lme (retrieved 2018)

[6] O'Neill BC, Kriegler E, et al. (2014), "A new scenario framework for climate change research. The concept of shared socioeconomic pathways." Climate Change 122(3):387–400.

[7] Grunzke R, Adolph T, et al. (2017), "Challenges in Creating a Sustainable Generic Research Data Infrastructure." Softwaretechnik-Trends, 37 (2). pp. 74-77.

[8] Tavares de Sousa N, Weber T, et. al. (2018), "Designing a Generic Research Data Infrastructure Architecture with Continuous Software Engineering", In: 3rd Workshop CSE 2018, pp. 85-88

[9] "Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)", www.openarchives.org/pmh/ (retrieved 2018)

[10] DataCite Metadata Working Group, "DataCite Metadata Schema 4.1", www.schema.datacite.org/meta/kernel-4.1/ (retrieved 2017)

[11] PANGAEA - Data Publisher for Earth & Environmental Science, www.pangaea.de (retrieved 2018)