Executable research compendia in geoscience research infrastructures

Summary
Researchers in "Computational X" and "X-informatics" across all geoscience disciplines collaborate and publish software relying on scripts, own source code, and libraries. They download data from domain specific or generic data repositories and deploy computations remotely. Their results are reviewed, published, archived, and compared with persistent identifiers, connected to other works via references, and listed in catalogues.

A single researcher, intentionally or not, interacts with all sub-systems of RIs and all building blocks of ERCS.

These interactions are vital for research and should be captured in a meaningful way for grasping the complete science and connecting key stakeholders, i.e. scientists, publishers, and librarians, to preserve our knowledge.

The ERC is developed by the DFG-funded project Opening Reproducible Research. It provides services for (a) semi-automatic creation of ERCS based on typical research workflows, (b) interactive manipulation of encapsulated analysis, and (c) deposition of complete ERCS with suitable metadata in repositories.

We are looking for RIs to collaborate on these ideas!

Environmental Research Infrastructure

- Interoperability
- Stakeholders & Viewpoints
- Distributed & Sub-systems
- data, APIs & services

Environmental Research Infrastructures (RI) provide advanced capabilities for data sharing, processing, and analysis to enable excellent research [...] in the environmental sciences [4]. They integrate large-scale sensor/observer networks with dedicated data curation services and analytical tools [2].

RIs and their requirements have been described in detail using the Open Distributed Processing (ODP) model [3,4] across different viewpoints of the participating stakeholders: science viewpoint, informational viewpoint, computational viewpoint, engineering viewpoint, and technology viewpoint. Derived from these viewpoints can be sub-systems of data acquisition, data curation, data access, data processing and community support [4]. These systems are a distributed infrastructure across stakeholders and domains.

RI interoperability concerns compatibility of data models, metadata standards, and service descriptions. They are tasked with semantic web technologies (linking, ontologies, vocabularies) and large scale interdisciplinary coordination projects bridging across domains (cf. [3]).

In the end, RIs are about making environmental data readily available for analysis using flexible, powerful and thus complex APIs and (web) services for large datasets and diverse user groups in the geosciences.

Non-comprehensive list of RIs [2, 3]:

- ACTRIS
- ARAGE
- ESCAT-IO
- EUKIR
- EMODnet
- ESMOS
- EPISO
- ESOTERI
- E km
- AP ano
- EUROFEETS
- EUROGEOSS
- EICKOS
- IAGOS
- ICOS
- INTERACT
- EL-ENES
- JERICO
- LifeWatch
- LITER
- SeaDataNet
- SIDS

ERC + RI

Exchange and Preservation
ERC as usable building blocks are a powerful item to be shared, e.g. for downloading from RIs (include the full pre-processing tool chain with data) or for archival. Even undocumentated knowledge is sure to be contained, ultimately in the source code of the ERC.

Documentation comprises both instructions (e.g. a README), the actual scientific publication, and metadata in standardized formats (licenses, discovery metadata, ...). The actual publication comes in a source format (i.e. based on literate programming) and a viewable format (e.g. an HTML document).

UI bindings open up the compendium. They allow reviewers to interact with diagrams and manipulate former hidden parameters for a comprehensive understanding of the underlying data and code.

A formal specification for ERC connects these building blocks in a meaningful way. It enables technical checking of computation outputs of an ERC and closes the gap of dependency preservation for computational scholarly works.

Self-consistency
ERCs intentionally remove all dependence on ephemeral sources, which RIs are due to their distributed nature and complex infrastructure. But an ERC could link to selected original sources, e.g. for data, or define selected trusted resources, e.g. an RI data processing system, which can be assumed to exist.

Metadata
ERCs connect the different parts of a piece of research in a meaningful way and facilitate discovery. By bridging to RI metadata models, metadata quality and richness can be improved.

Execution
ERC services create and execute a packaged analysis but integrate with existing platforms for storage (e.g. repositories or archives) and display (e.g. journal platforms). These services can also connect to be used by RIs.

Coherence
ERC services not only validate completeness and integrity of the contained building blocks but also check the consistency of results against the original outcome. They can improve research quality in RIs.