Executable research compendia in geoscience research infrastructures

This presentation participates in OSPP

Outstanding Student Poster & PICO Contest

Daniel Nüst

Institute for Geoinformatics
University of Münster, Germany
✉ **daniel.nuest@uni-muenster.de**

Summary

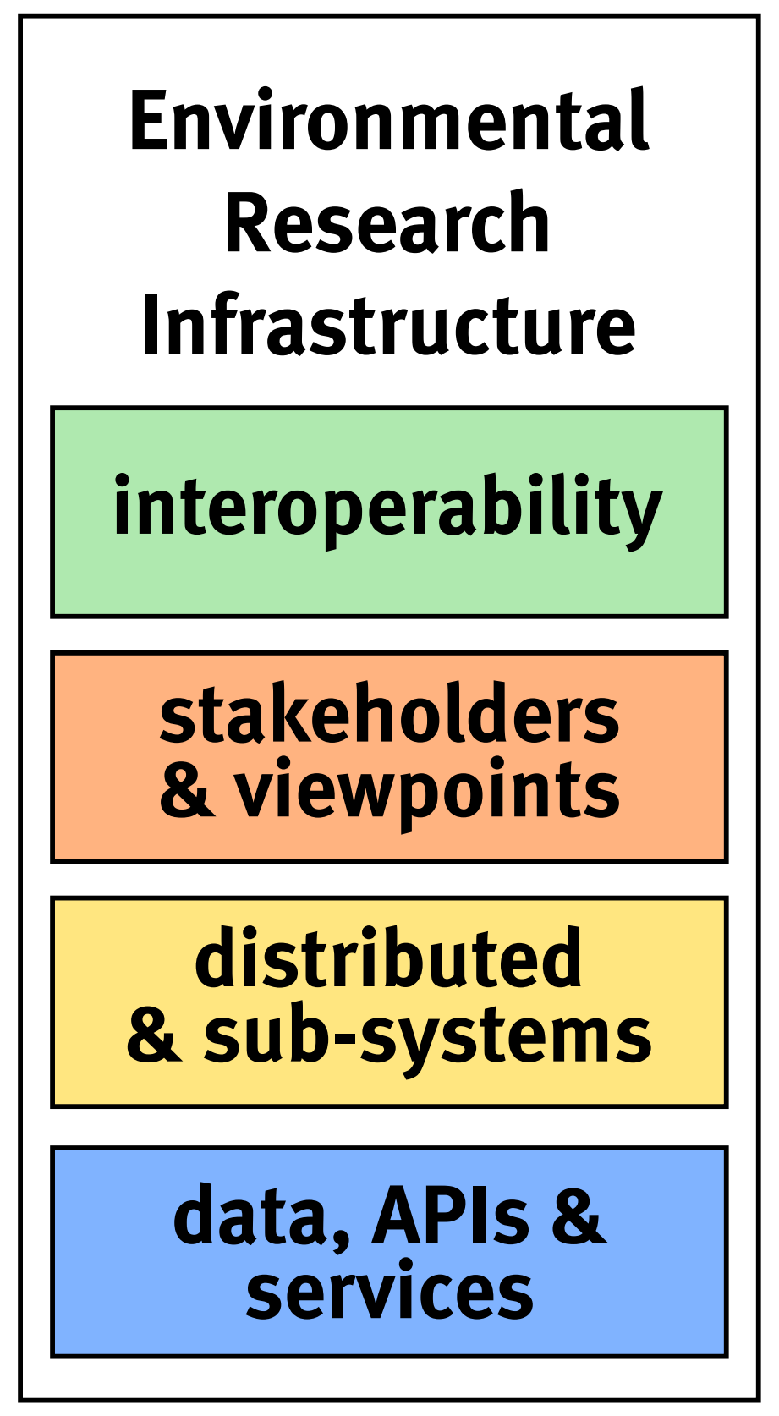
Researchers in "Computational X" and "X-informatics" across all geoscience disciplines collaboratively develop and publish software relying on scripts, own source code, and libraries. They download data from domain specific or generic data repositories and deploy computations remotely. Their results are reviewed, published, archived, tagged 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) interative 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 [3].

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, technology viewpoint. Derived from these viewpoints can be **sub-systems** 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 tackled with semantic web technologies (linking, ontologies, vocabularies) in large scale intersdisciplinary 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
ANAEE
EISCAT-3D
ELIXIR
EMBRC
EMSO
EPOS
ESONET
EURO-Argo
EUROFLEETS
EUROGOOS
FIXO3
IAGOS
ICOS
INTERACT
IS-ENES
JERICO
LifeWatch
LTER
SeaDataNet2
SIOS

[3] Open Information Linking for Environmental Research Infrastructures
Martin, P., Grosso, P. et al.
IEEE 11th International Conference on eScience, 2015
doi: 10.1109/eScience.2015.66

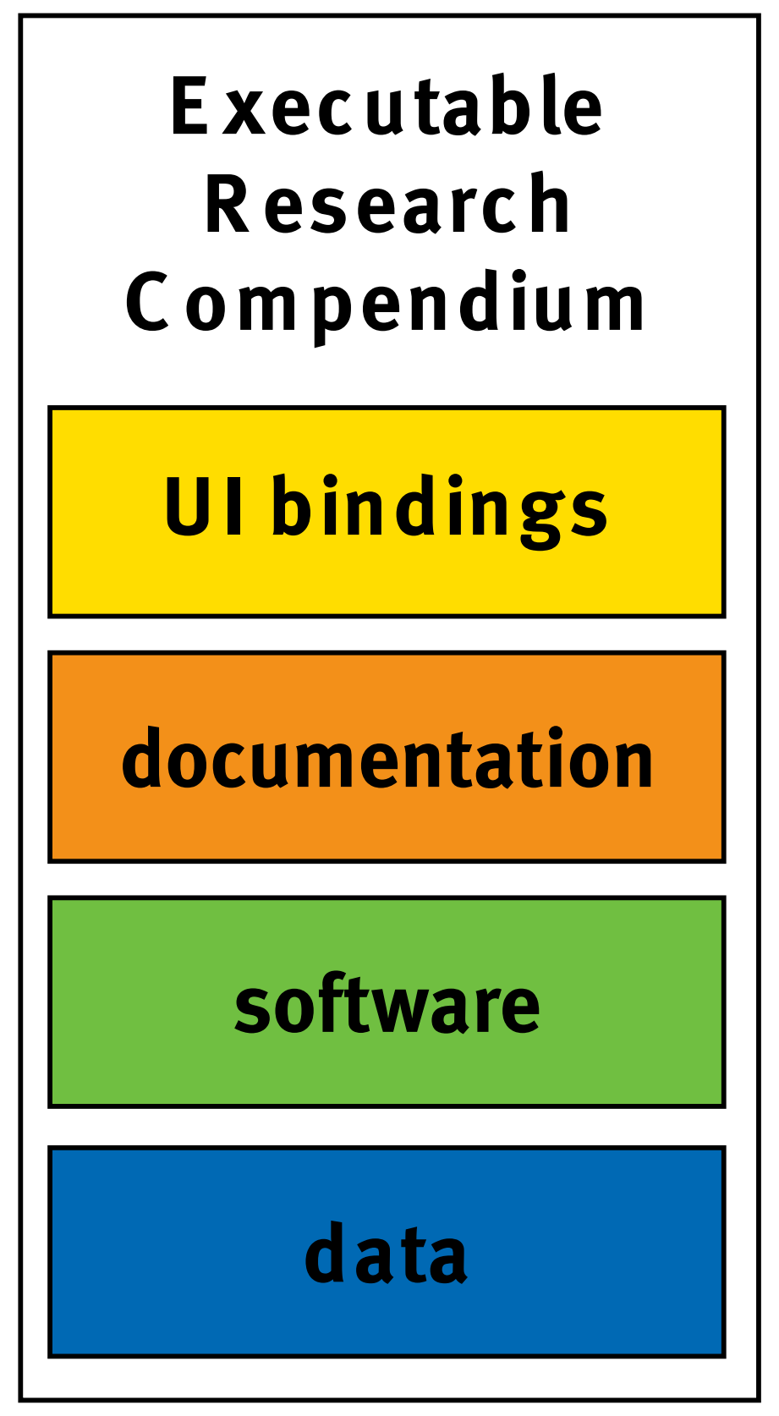
[4] Analysis of Common Requirements for Environmental Science Research Infrastructures
Chen, Y., Hardisty, A. et al.
The International Symposium on Grids and Clouds, Taipei, 2013
https://inspirehep.net/record/1291201/files/ISGC%202013_032.pdf

opening reproducible research

http://o2r.info

@o2r_project

Opening Reproducible Research is a joint project by the Institute for Geoinformatics and the University and Regional Library Münster funded by the German Research Foundation


Executable Research Compendium

UI bindings

documentation

software

data

Executable Research Compendia (ERC) support requirements of authors, readers, publishers, curators, as well as preservationists. They are a new way to package computational research combining data, software, text, and a user interface description and provide a novel potential to find, explore, reuse, and archive computer-based research. [1]

Data comprises all inputs for an analysis, ideally starting with raw measurements, in form of text files, or databases.

Software comprises analysis code/scripts created by a researcher and the complete runtime environment. In the first implementation, a Docker image container encapsulates all libraries and tools in an executable form and a Dockerfile provides a transparent manifest.

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).

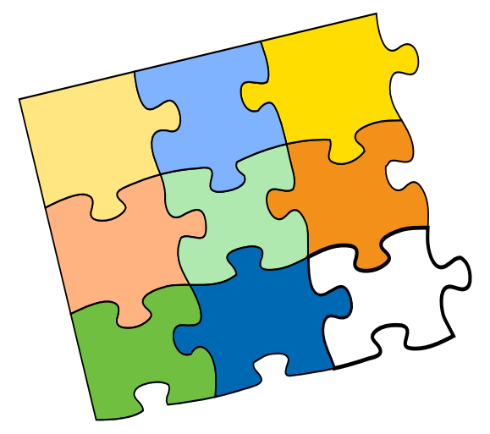
UI bindings open up the compendium. They allow reviewers to interact with diagrams and manipulate formerly hidden parameters for a comprehensive understanding of the underlying data and code.

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.

[1] Opening the Publication Process with Executable Research Compendia
Nüst, D., Konkol, M. et al.
D-Lib Magazine, 2017
doi: 10.1045/january2017-nuest

[2] Opening Reproducible Research
Nüst, D., Konkol, M. et al.
Geophysical Research Abstracts
Vol. 18, EGU2016-7396, 2016
http://meetingorganizer.copernicus.org/EGU2016/EGU2016-7396.pdf

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 undocumented knowledge is sure to be contained, ultimately in the source code of the ERC.

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 RIs 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 faciliate 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.

On-Poster Survey
What field of "Computational X" or "X-informatics" do you work in?

"Computational X" Buzzword Corner

Big Data
Cloud Computing
Open Data
Open Review
Open Science
Open Source
Reproducible Research