Opening the Publication Process with Executable Research Compendia

by Daniel Nüst, Markus Konkol et al.

o2r opening reproducible research

http://o2r.info

@o2r_project
https://github.com/o2r-project

Funded by: DFG

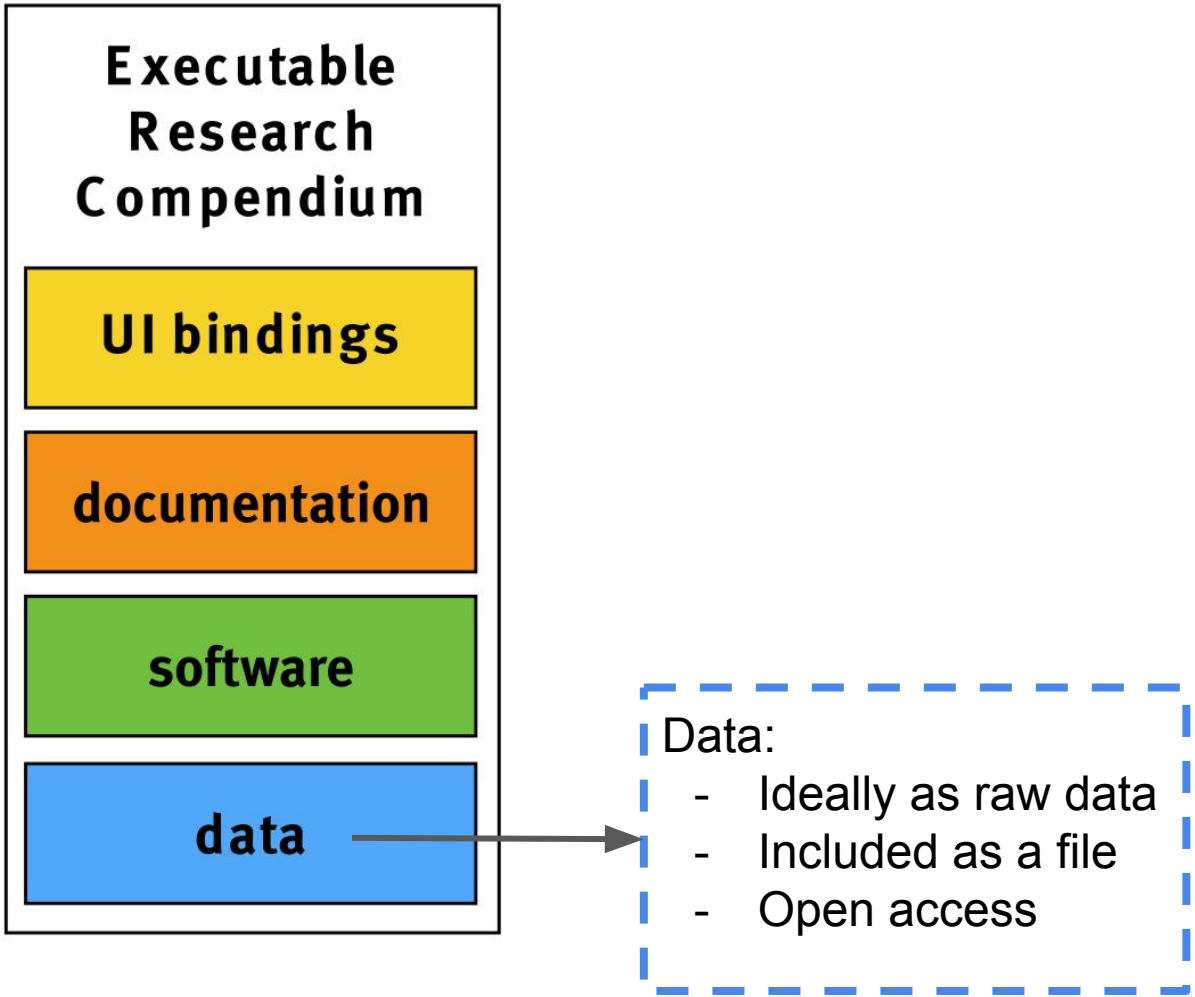

Reproducible Research?
Methods
Data
Source
Reviews
Access
O
P
E
N

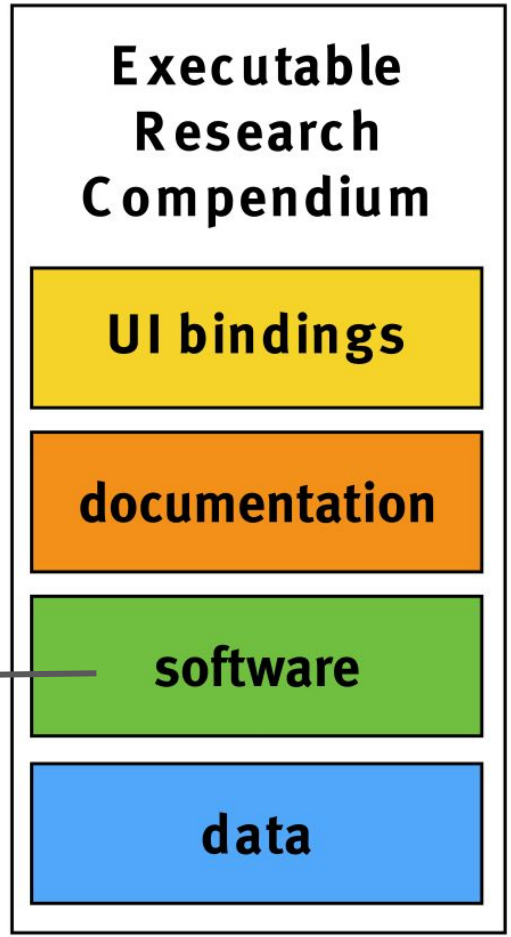
Goals

1. How can we **simplify** computational reproducibility?
2. How can we create **incentives**?

Our key contributions:

Executable Research Compendium (ERC) + Requirements + Publication process


Executable
Research
Compendium
UI bindings
documentation
software
data
Data:
- Ideally as raw data
- Included as a file
- Open access

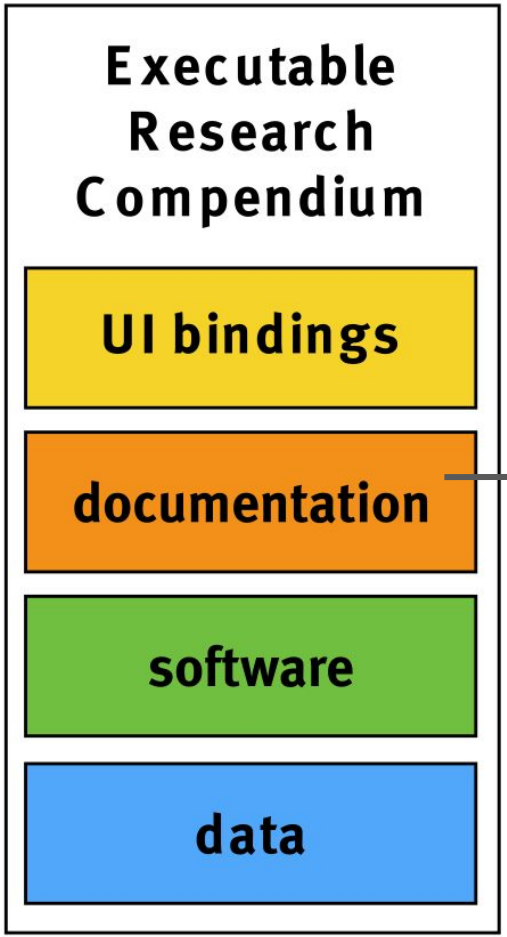


Software:

- Entire **runtime environment**
- Versions + dependency hell
- Open source
- Execution

Data:

- Ideally as raw data
- Included as a file
- Open access

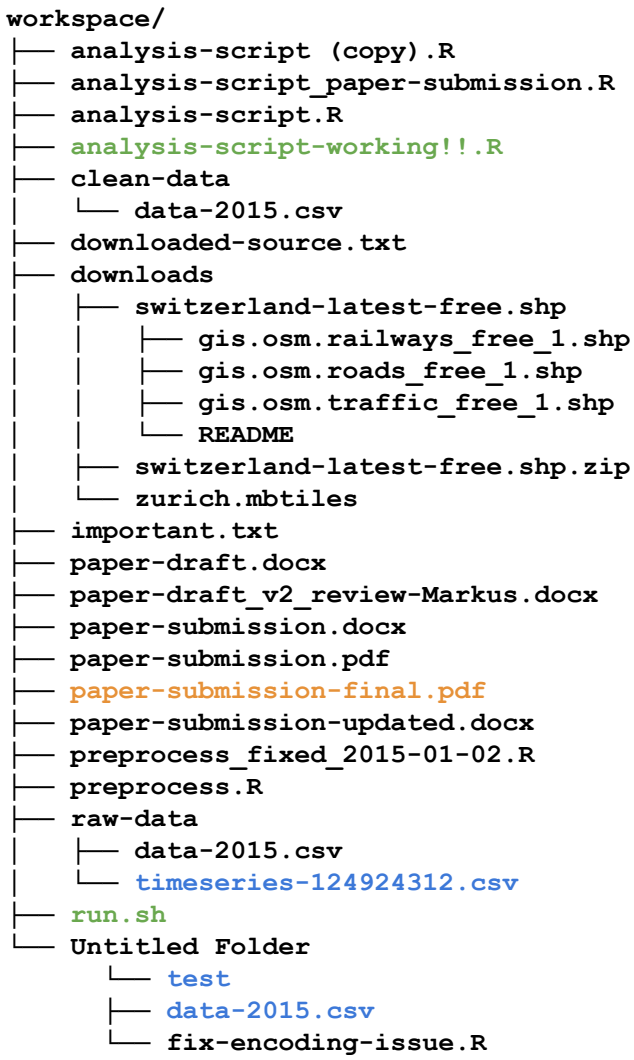


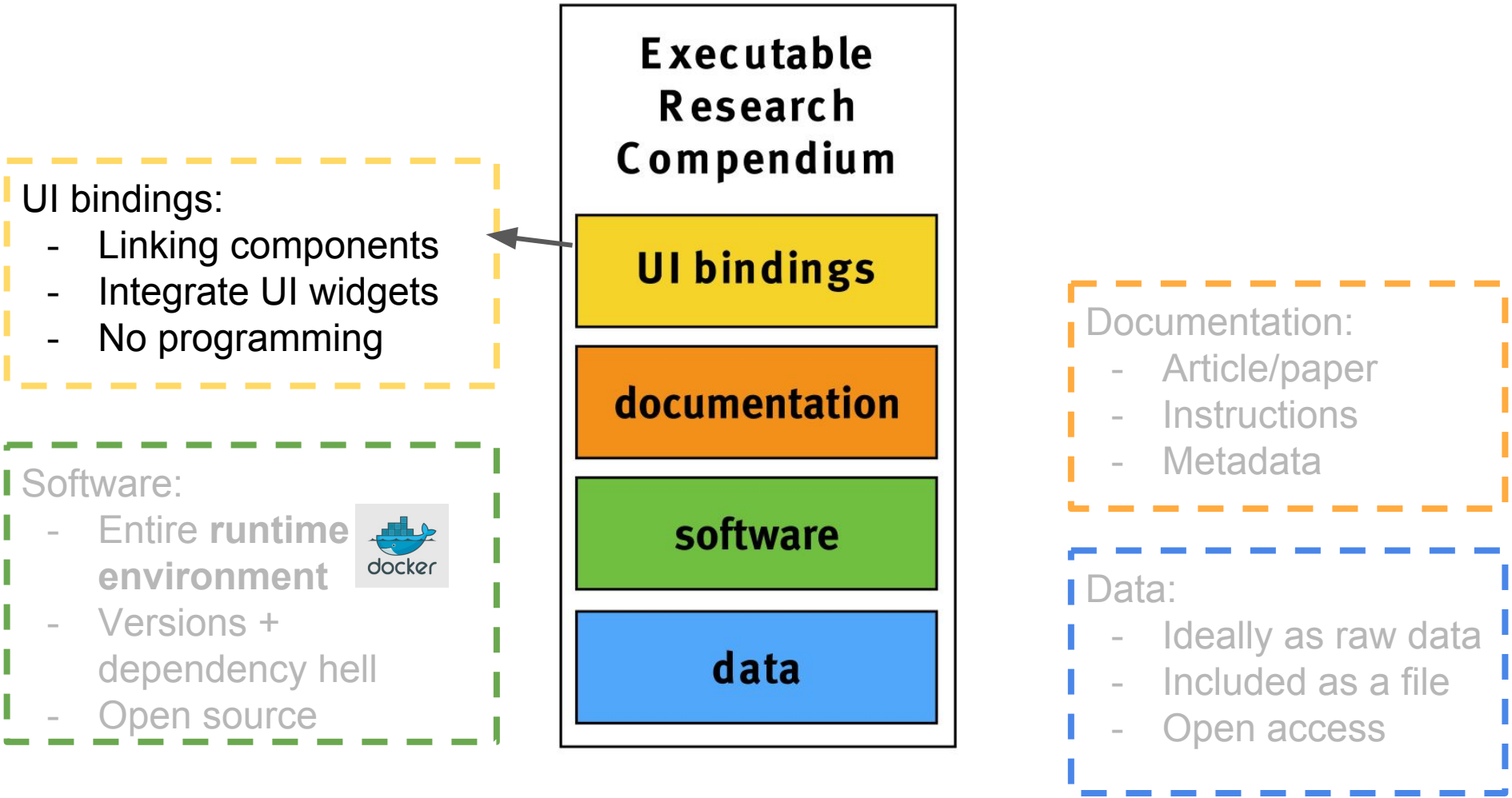
Documentation:
- Article
- Instructions
- Metadata

Software:
- Entire **runtime environment**
- Versions + dependency hell
- Open source
- Execution

Data:
- Ideally as raw data
- Included as a file
- Open access

```
workspace/
├── analysis-script (copy).R
├── analysis-script_paper-submission.R
├── analysis-script.R
├── analysis-script-working!!.R
├── clean-data
│   └── data-2015.csv
├── downloaded-source.txt
├── downloads
│   ├── switzerland-latest-free.shp
│   │   ├── gis.osm.railways_free_1.shp
│   │   ├── gis.osm.roads_free_1.shp
│   │   ├── gis.osm.traffic_free_1.shp
│   │   └── README
│   ├── switzerland-latest-free.shp.zip
│   └── zurich.mbtiles
├── important.txt
├── paper-draft.docx
├── paper-draft_v2_review-Markus.docx
├── paper-submission.docx
├── paper-submission.pdf
├── paper-submission-final.pdf
├── paper-submission-updated.docx
├── preprocess_fixed_2015-01-02.R
├── preprocess.R
├── raw-data
│   ├── data-2015.csv
│   └── timeseries-124924312.csv
├── run.sh
└── Untitled Folder
    └── test
    ├── data-2015.csv
    └── fix-encoding-issue.R
```



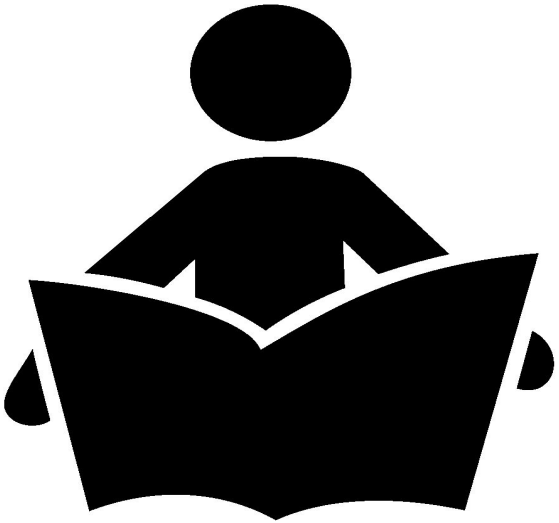

Executable Research Compendium
UI bindings
documentation
software
data
UI bindings:
- Linking components
- Integrate UI widgets
- No programming
Software:
- Entire runtime environment
- Versions + dependency hell
- Open source
docker
Documentation:
- Article/paper
- Instructions
- Metadata
Data:
- Ideally as raw data
- Included as a file
- Open access

Core requirements - Author

Create ERC based on existing workspace + **check** metadata

Core requirements - Reader

Find + **Rerun** with one click + **examine** + **manipulate**



Core requirements - Reviewer

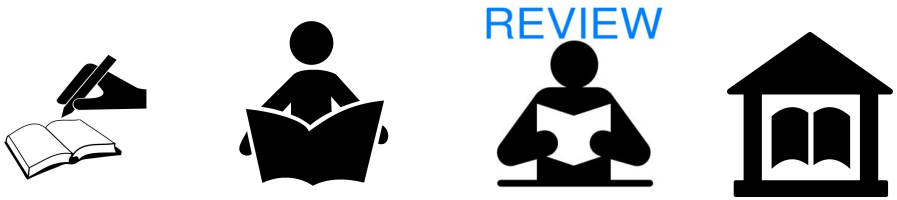
Verify results + scrutinize

Core requirements - Libraries and publishers

Integrate procedures, data + results for **publishing and discovery**

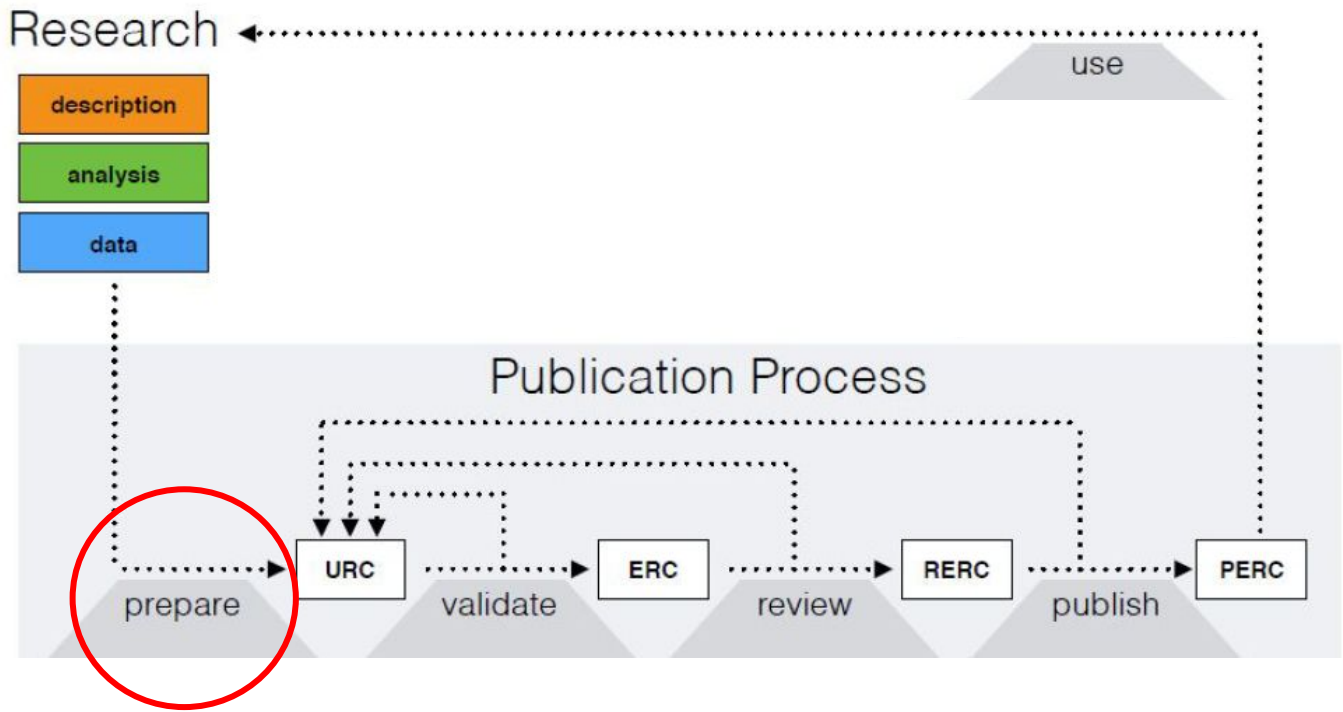
Core requirements - Curator and preservationist

Check deep integrity regularly + care for **metadata** + long-term archival



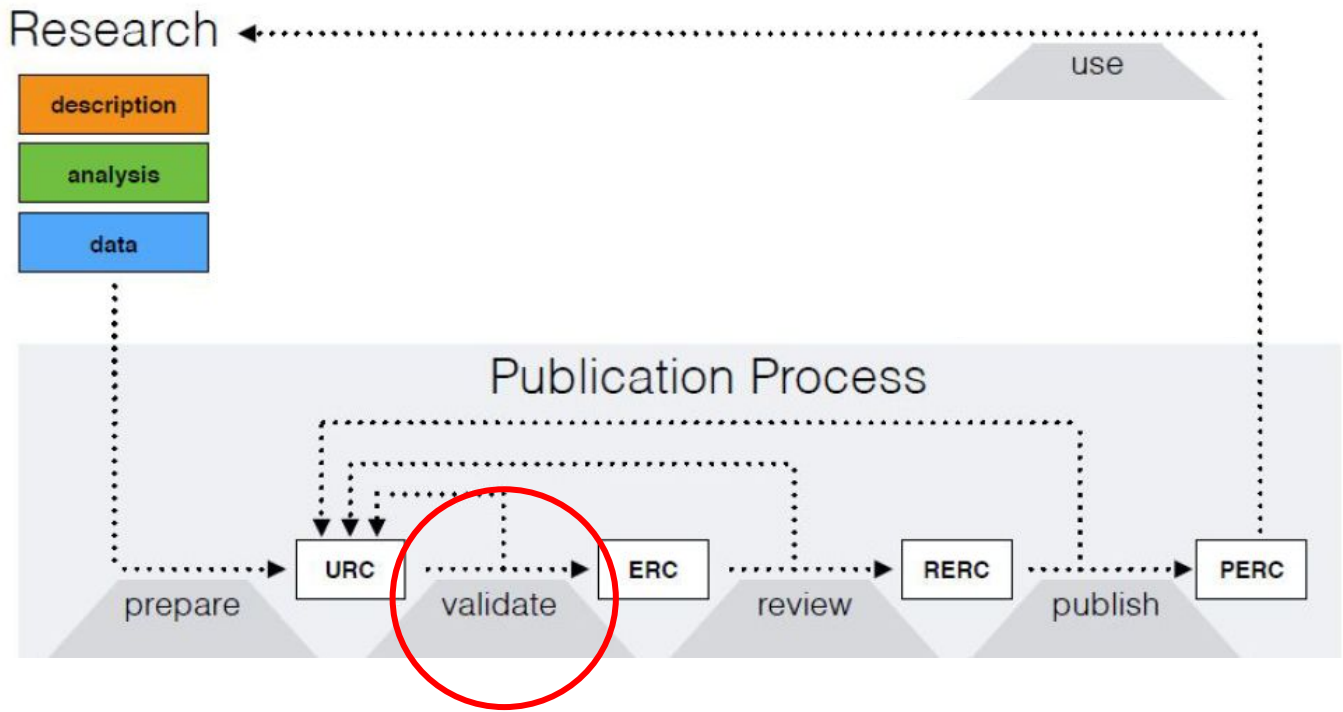
Publication process

URC: Unvalidated RC



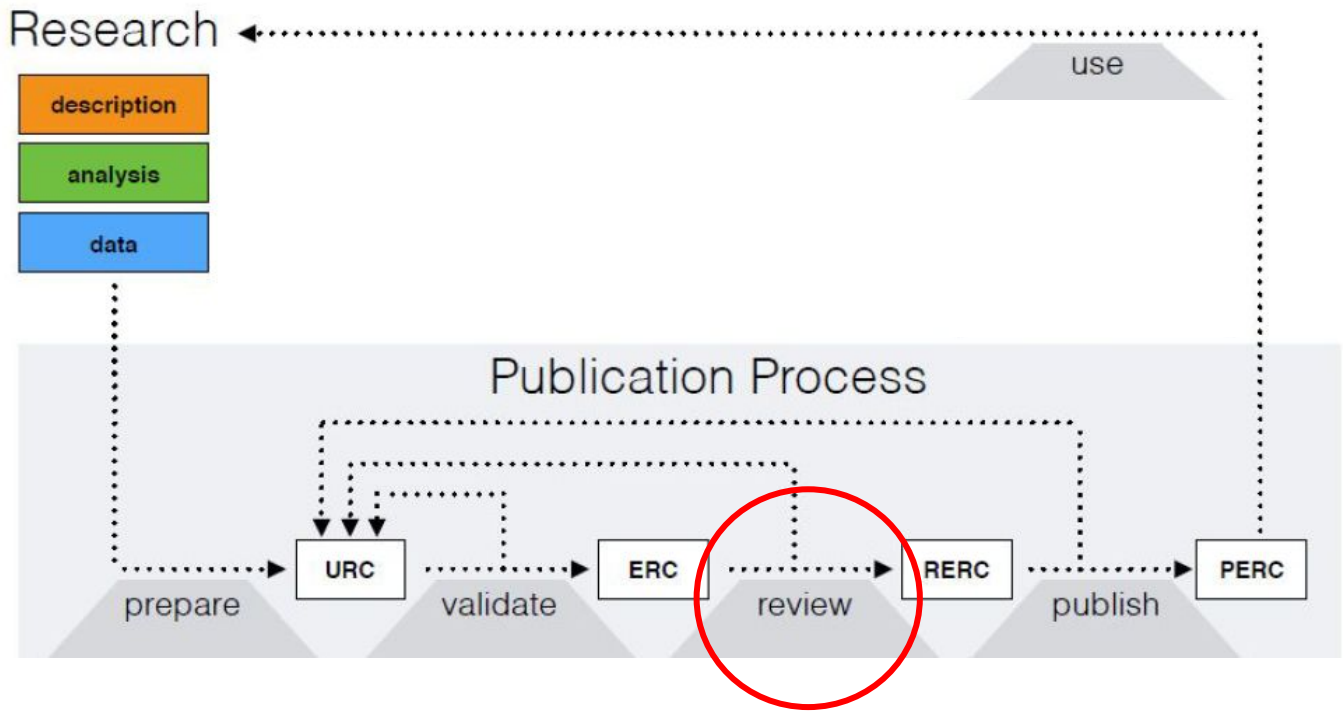
Publication process

URC: Unvalidated RC
ERC: Executable RC



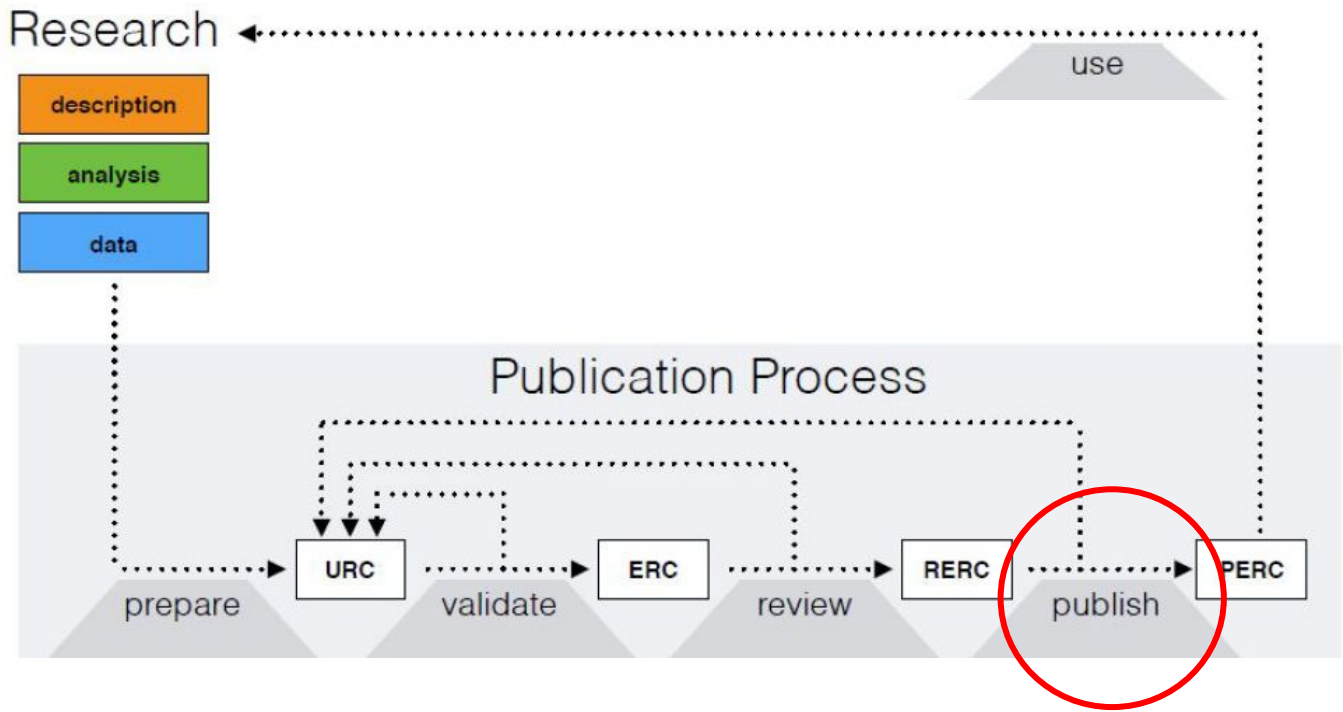
Publication process

URC: Unvalidated RC
ERC: Executable RC
RERC: Reviewed ERC



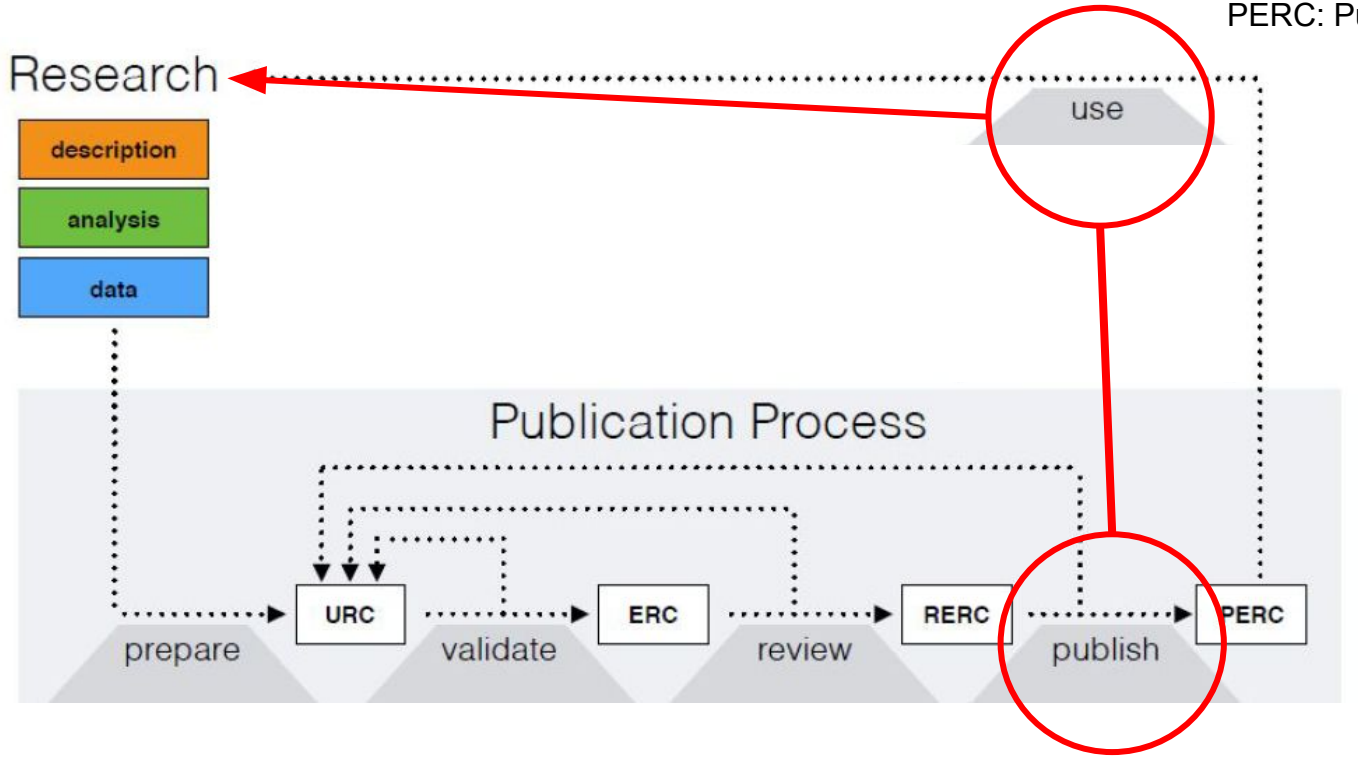
Publication process

URC: Unvalidated RC
ERC: Executable RC
RERC: Reviewed ERC
PERC: Published ERC

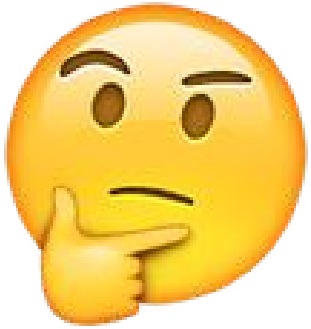


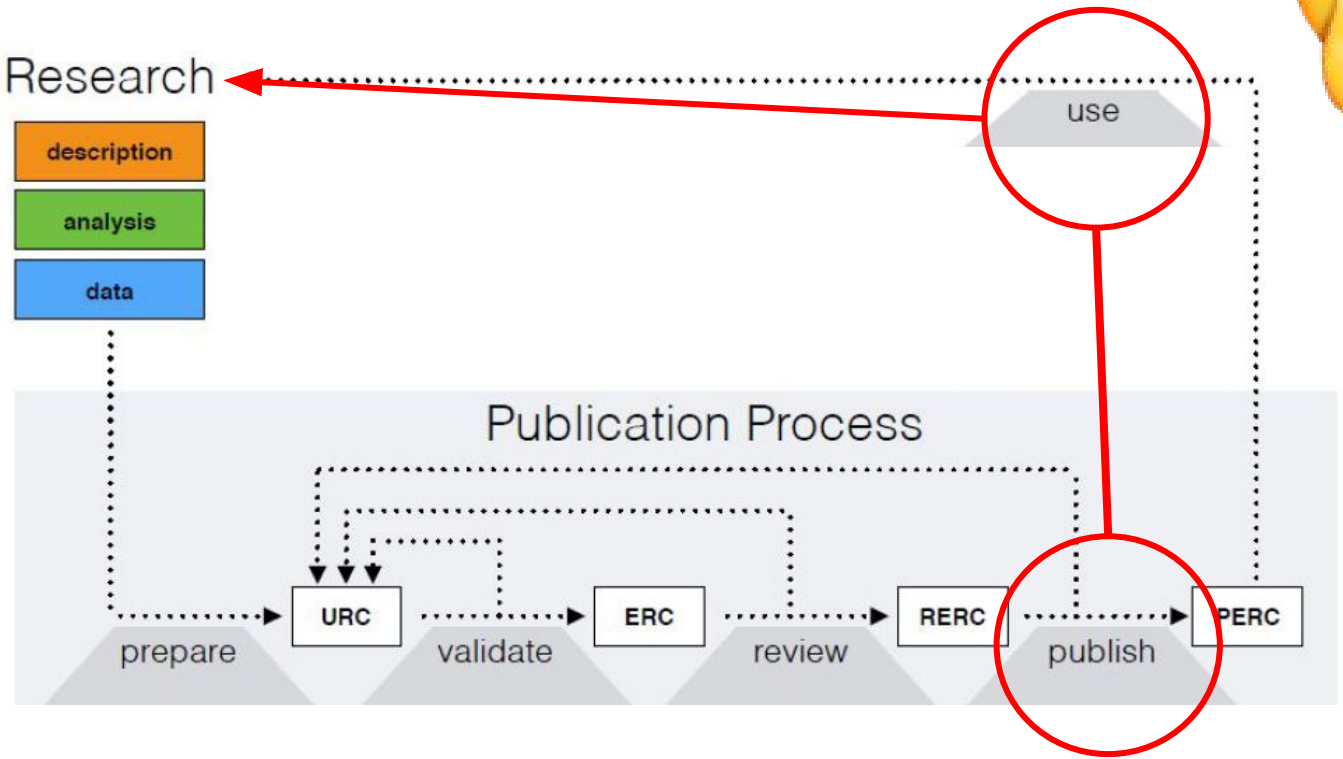
Publication process

URC: Unvalidated RC
ERC: Executable RC
RERC: Reviewed ERC
PERC: Published ERC



Publication process





Sources

Nüst, D. et al. "Opening Reproducible Research." *EGU General Assembly Conference Abstracts*. Vol. 18. 2016.

Nüst, D., Konkol, M. et al. “Opening the Publication Process with Executable Research Compendia”. *First international workshop on reproducible open science*. 2016.

Images:
http://www.pd4pic.com/images/writer-writing-author-book-hand-pen-pencil.png
http://www.ssdesign.com/librarypr/download/odds_and_ends/reader.bmp
http://ijrest.net/review-md.png
https://pixabay.com/p-160871/?no_redirect
http://mmtuploadfiles.s3.amazonaws.com/MUSEUM_CURATOR.png
http://emojipedia-us.s3.amazonaws.com/cache/db/d5/dbd57bcbb3fff7245025a39061012200.png
https://www.docker.com/sites/default/files/legal/small_v.png

o2r team

Dr. Stephanie Klötgen (ULB), Markus Konkol (ifgi), Prof. Dr. Christian Kray (ifgi), Jörg Lorenz (ULB), Daniel Nüst (ifgi), Prof. Dr. Edzer Pebesma (ifgi), Holger Przibytzin (ULB), Dr. Marc Schutzeichel (ULB), Dr. Beate Tröger (ULB)

Selected Programme Groups: IE **Search:** Submit

IE – Interdisciplinary Events

- Programme Committee login

- IE1 – Arctic Environmental Change: Threats and Opportunities
- IE2 – Citizen-Science
- IE3 – Big Data in the Geosciences
- IE4 – Communicating and Mitigating Natural Risks

Suggested session

Reproducible research in practice

Scientific leader ESSI

Session description

Reproducibility is important for science. The topic gains more attention each year with prominent papers, editorials and blog posts. Journals, researchers and funders drive forward the agenda on many aspects of openness: open access, open science, open data, open source. But still the vast majority of papers analyzing any kind of data is not accompanied by data, code and documentation that let you easily reproduce the calculations that underly the paper. This session will showcase papers that do. It is open to members of all programme groups and scientific disciplines to present how they conduct data-based research in a reproducible way. They are welcome to share practical advice, lessons learned, practical challenges of reproducibility, and report on the application of tools and software that support computational reproducibility. Computational reproducibility is especially important in the context of big data. Readers of articles must be able to trust the applied methods and computations because the huge amount of observations that are collected not only create a huge volume of data, but very often these data are also unique, observed by a single entity, or synthetic and simulated. Contributions based on small datasets are of special interest to demonstrate the variety in big data. Topics may include, but are not limited to, reproducibility reports and packages for previously published computational research, practical evaluations of reproducibility solutions for a specific research use case, best practices towards reproducibility in a specific domain such as publishing guidelines for data and code, or experiences from teaching methods for computational reproducibility.

Keywords

Statistical methods (Geostatistical methods)

o2r project - goals

ERC *technical specification* and *prototypes* tailored to geosciences

R
spatial and spatiotemporal data
map and timeseries **interaction**
submission of “plain workspace”

Usability *evaluation* of prototypes

Improvements by ERC

(i) reviewers obtain tools for easily validating results in scientific publications submitted by researchers

(ii) results are well-grounded since the research steps described in the text are shared

(iii) other researchers benefit as they obtain tools for accessing and reusing research components

(iv) scientists can easier build upon previous research findings.

Challenges in open reproducible research

(i) the creation of ERCs must be easy for authors

(ii) ERC--based interaction, discovery, exploration and reuse must provide sufficient benefits for scientists to result in broad uptake of the concept

(iii) ERCs must handle diverse workspaces and integrate requirements from all stakeholders.

Reviews (2):

From research compendia to executable research compendia:

- Runtime environment & specified execution != potential (!) for code and instruction > rephrase

Reviews (2):

What about papers which make use of randomness, such as simulations, numerical analysis, some machine learning methods?

- “Seed must be used.” > mention as “limit”?

I don't feel that a red or green button to show whether an experiment was successfully reproduced, as described later, are quite enough.

    - Deutlicher darstellen worauf “grün vs. rot” beruht/entschieden wird

Reviews (2):

How is the ERC going to support reproducibility of experiments requiring very large data (e.g. physics), computationally/memory expensive tasks, tasks requiring some specific e.g. distributed infrastructure (multi-agent systems), tasks experimenting with black-boxes or interacting with 3rd party services not under the researcher's control? It would be good to at least discuss these limitations/issues.

Reviews (2)

Is your work reprodicble?

Reviews: (3)

However, it has not been implemented, neither evaluated

Evaluate usability with examples

Evaluate publishing process in journal platform

Reviews (2)

- The paper mentions on page 3 (slightly critically) the Gold OA route (payments to publishers). It would be good to mention also the Green OA route that is achieved freely through self-archiving, especially as more publications are available as OA through Green OA than through Gold OA.

Reviews: (1)

- "The ERC defines machine-readable conventions that allow computer systems to control and evaluate the embedded containers" -> which kind of machine-readable convention? Docker recipes? Semantic models?

Make clearer: ERC Metadaten müssen das Kommando enthalten um den container laufen zu lassen; “conventions” = Konzeptebene; konkrete = Implementierungsebene
“More concretely, command line instructions to start the Docker image and retrieve the execution results.”

Reviews: (1)

- "It seems trivial to build a web service that downloads a research compendium from a repository, unpacks it, executes it, and returns the results created." -> I would not say that it is trivial. Not sure how easy this could be in general, taking into account how diverse software solutions are depending on the scientific domain.

“(Given a erc spec & prototype, der rest ist trivial), ERC abstracts from diversity

Reviews (1):

- UI bindings -> not clear how they would be represented and how they fit in the context of a Docker containers-based approach.

Representation: stored in ERC metadata;

Context: ERC specifies how the changed parameters are passed _into_ the container to the actual process;

What’s so difficult about reproducible research?

Problems:

1. Discovery of errors
2. Loss of competitive advantage
3. Lack of tangible benefits or rewards

What’s so difficult about reproducible research?

1. How can we make reproducibility **easy**?
2. How can we make reproducible papers **interactive**?
3. How can we help researchers to long-term **archive** their research?

Our key contributions:

Publication process + ERC + Requirements