

# Research compendia enable code review during peer review

with CODECHECK and Opening Reproducible Research (o2r)

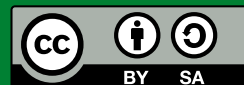
<https://codecheck.org.uk/> | <https://o2r.info>

Daniel Nüst @ Remote ReproHack, May 14, 2020

Institute for Geoinformatics, University of Münster | <http://nüst.de>

2020-05-14

CC-BY-SA 4.0



*The problem is that most modern science is so complicated, and most journal articles so brief, it's impossible for the article to include details of many important methods and decisions made by the researcher as he analyzed his data on his computer.*

Ben Marwick: How computers broke science – and what we can do to fix it

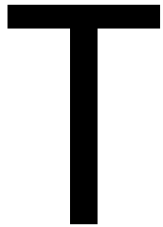
---

*"Claerbout's claim"* in Donoho (2010), An invitation to reproducible computational research

Claerbout & Karrenbach (1992), Electronic documents give reproducible research a new meaning

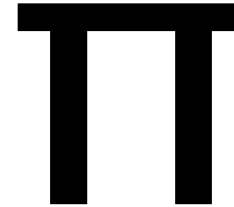
# Traditional and modern research(ers)

FROM



- broad knowledge: cross-discipline, collaboration
- deep knowledge: domain speciality (expertise and skills)

TO



- broad knowledge: cross-discipline, collaboration
- deep knowledge
  - domain speciality (expertise and skills)
  - stats/computing/reproducibility

# Research Compendium



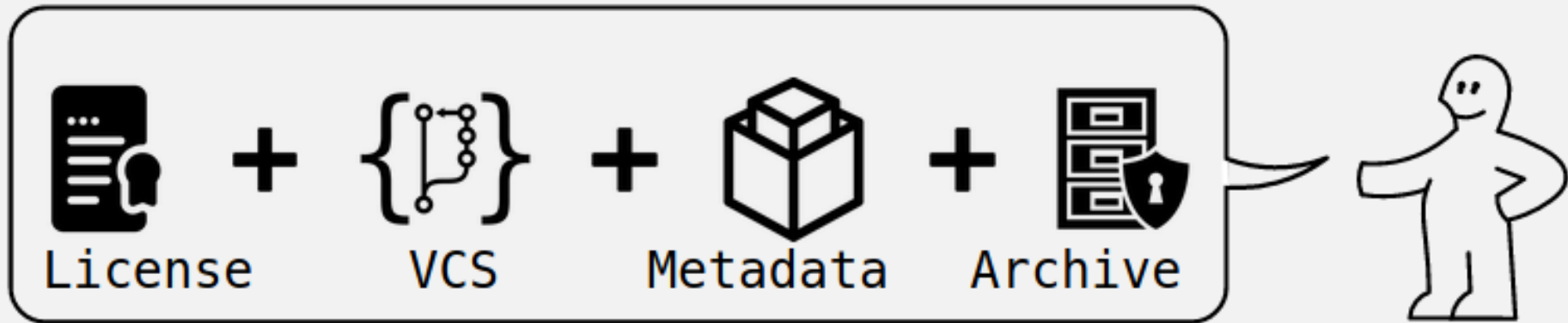
[research-compendium.science](https://research-compendium.science)

# Research compendia

...We introduce the concept of a *compendium* as both a *container* for the different elements that make up the document and its computations (i.e. *text, code, data, ...*), and as a means for *distributing, managing* and *updating* the collection.

Gentleman, Robert, and Duncan Temple Lang. 2007. "Statistical Analyses and Reproducible Research". Journal of Computational and Graphical Statistics 16 (1): 1–23. <https://doi.org/10.1198/106186007X178663>

# Key components you'll need for sharing a compendium



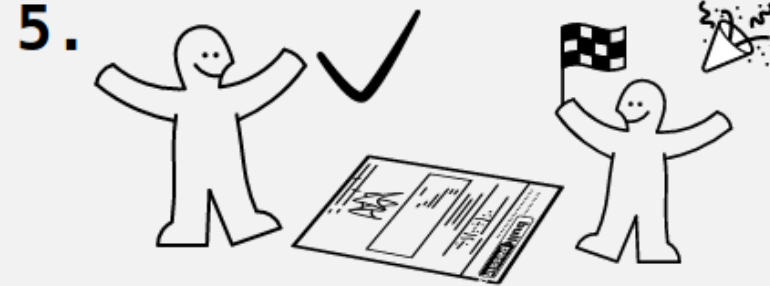
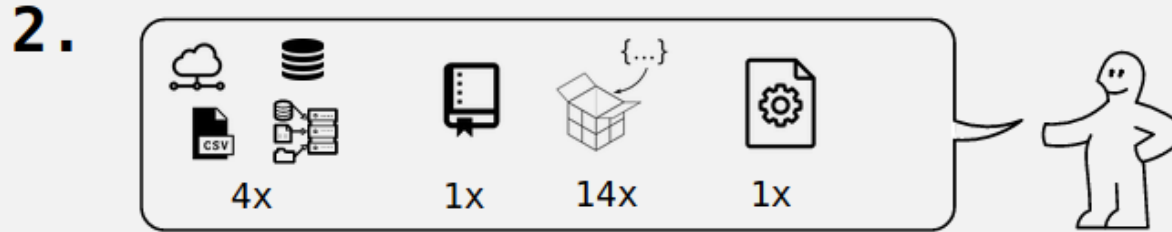
Source: Ram (2019), [How To Make Your Data Analysis Notebooks More Reproducible](#)

# Research compendia

Ram (2019), How To Make Your Data Analysis Notebooks More Reproducible

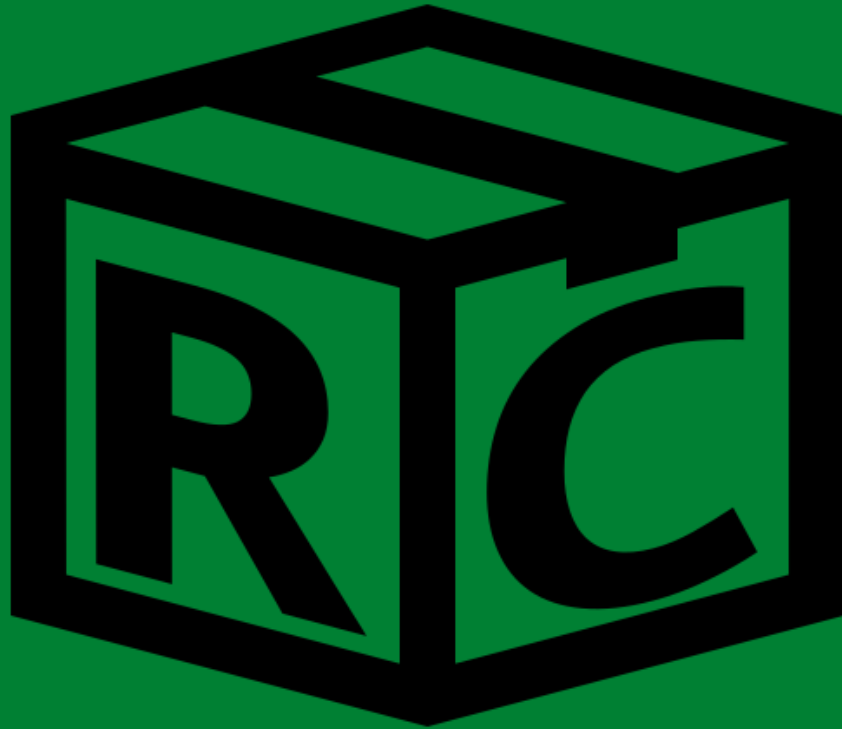
- small, medium, large compendia
- Stick with the conventions of your peers
- Keep data, methods and outputs separate
- Specify your computational environment as clearly as you can
- Leverage the R package structure and support tools/services as much as possible
- Use modern tools to make your compendia more accessible ([repo2docker](#), [containerit](#), [holepunch](#), [drake](#))
- Don't forget long-term archives and simpler formats ([Zenodo](#))

# KÖMPENDIUM



Source: Ram (2019), [How To Make Your Data Analysis Notebooks More Reproducible](#)





Transparent

Credit

Discover

Reuse

Colaborate

# Starting tomorrow, I will ...

(and you can recommend to the authors of the papers you reproduce)

- have a README ("all else is details")
- use good file names and relative paths (or Jenny Bryan will come and get you)
- use text-based, open file formats
- only work in containers (Rule 10)
- apply templates & follow community practices (e.g., `rrtools`)
- write and publish notebooks
- use only scripts, no point-and-click
- embrace openness & be ne fi ts
- document for future me
- work/review in the spirit of preproducibility



<https://gph.is/2JF2u2T>

# More on research compendia at research-compendium.science

---

You want to introduce changes in your community?

## *Reproducible Publications at AGILE Conferences*

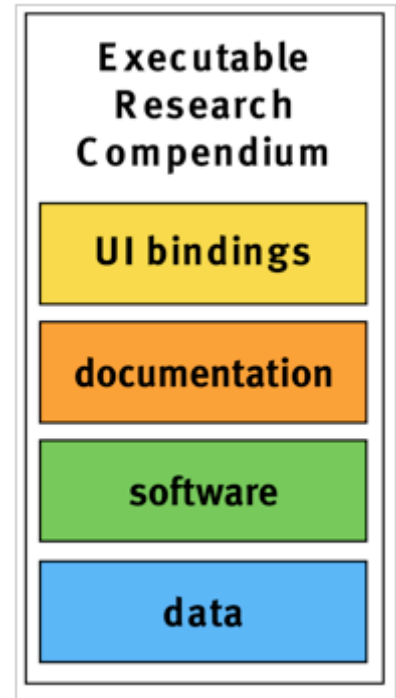
AGILE Reproducible Paper Guidelines

<https://reproducible-agile.github.io/>

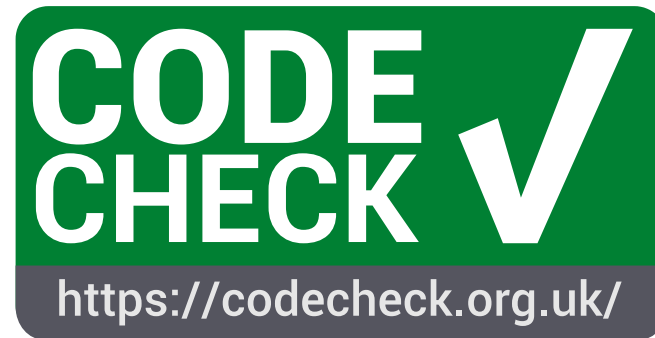
# Advanced control of computing environments



- Binder-ready research compendium
- Research compendium + container + bindings =  
Executable Research Compendium ([o2r.info/results/](https://o2r.info/results/))
- Ten Simple Rules for Writing Dockerfiles for Reproducible Data Science



How can research compendia enable code review during peer review?



*A process for independent reproduction of computations underlying research*

# The four CODECHECK principles

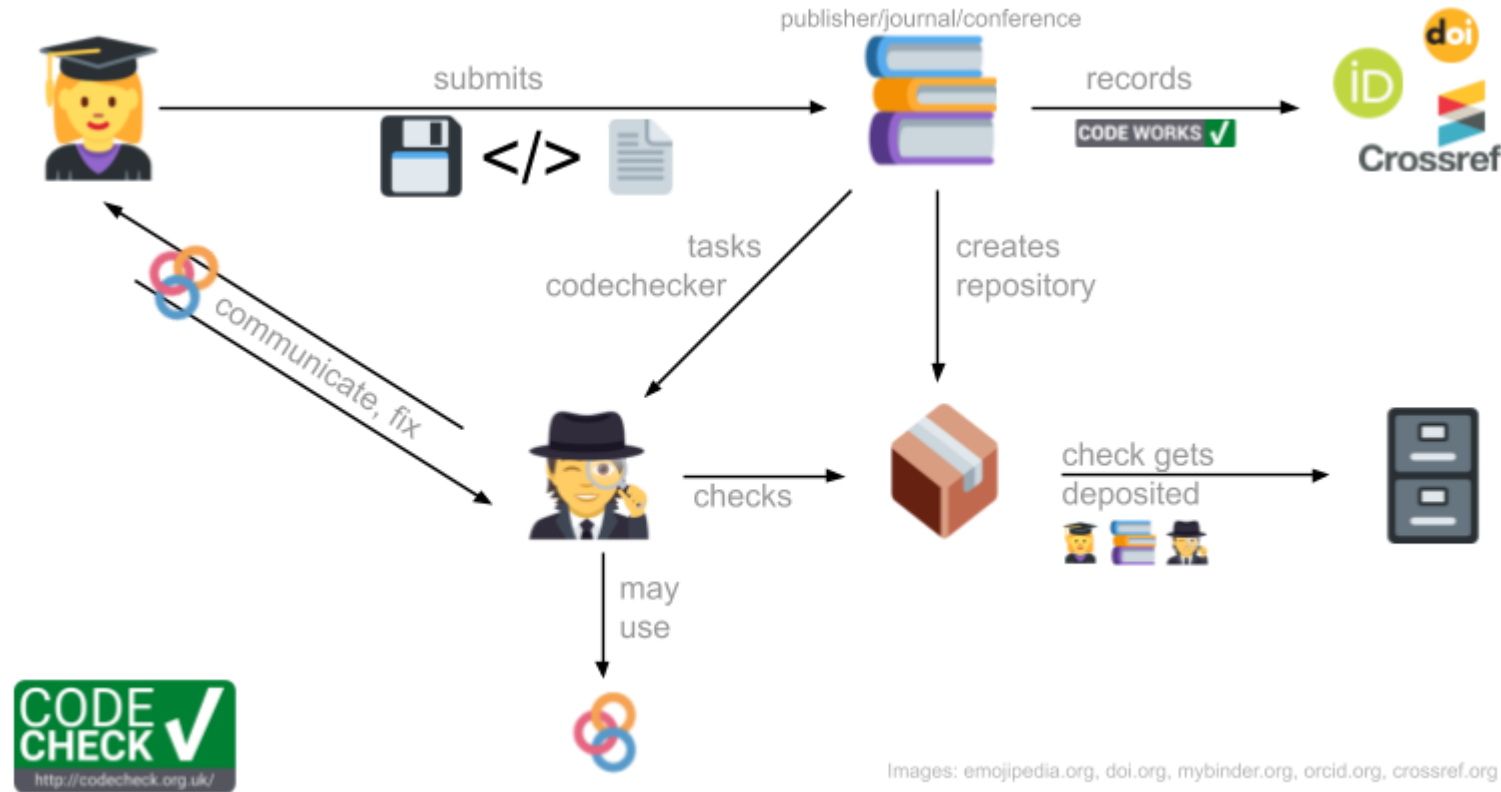
# The four CODECHECK principles

1. Codecheckers record but don't investigate or fix.
2. Communication between humans is key.
3. Credit is given to codecheckers.
4. Workflows must be auditable.





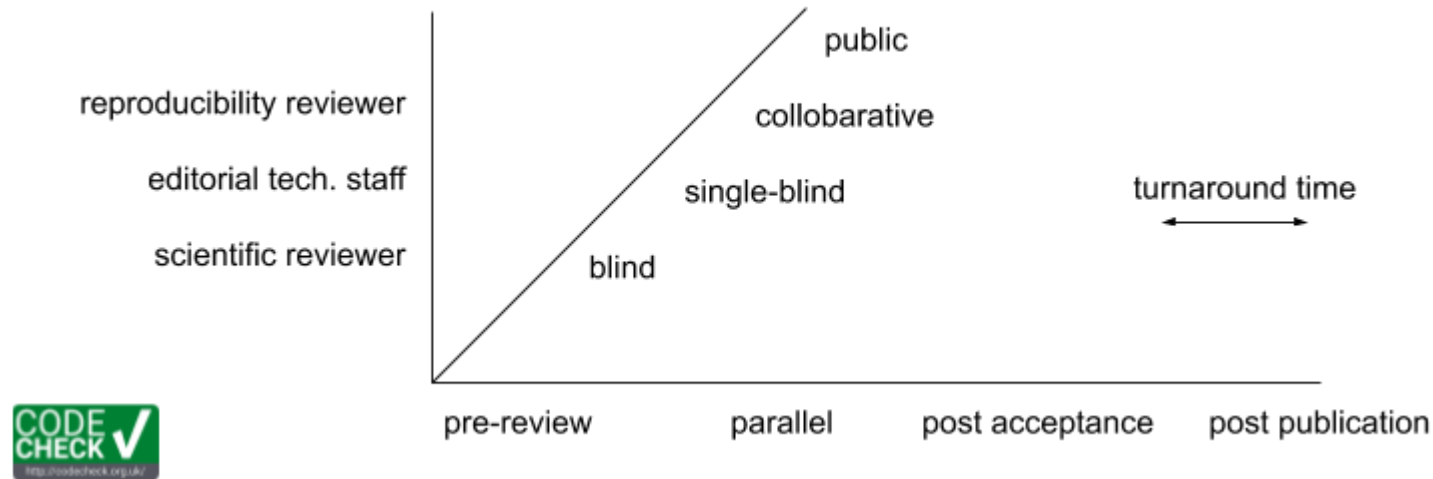
# Implementing the CODECHECK process



*Sketch of steps and responsibilities to implement a CODECHECK.*

<https://codecheck.org.uk/process/>

# Process variations



- Post review with extra role (ECR opportunity)
- Pre-review with extra role on staff
- Parallel reproducibility reviewer (flexible skill match)
- Regular reviewer (has not happend last 10 years  $\sim \backslash ( ' \vee ) \_ / \sim$  )
- Independent community pre-review at ReproHacks or for preprints

<https://codecheck.org.uk/process/>

# CODECHECKs

<https://codecheck.org.uk/register/>

<http://doi.org/10.5281/zenodo.3674056>

<https://doi.org/10.5281/zenodo.3741797>

# Get Involved

Go to a ReproHack to acquire the skills to become a codechecker ✓

<https://codecheck.org.uk/get-involved/>

- codechecker
- author
- reviewer
- editor
- publisher
- conference organiser

# More information: [codecheck.org.uk](https://codecheck.org.uk)

## Team

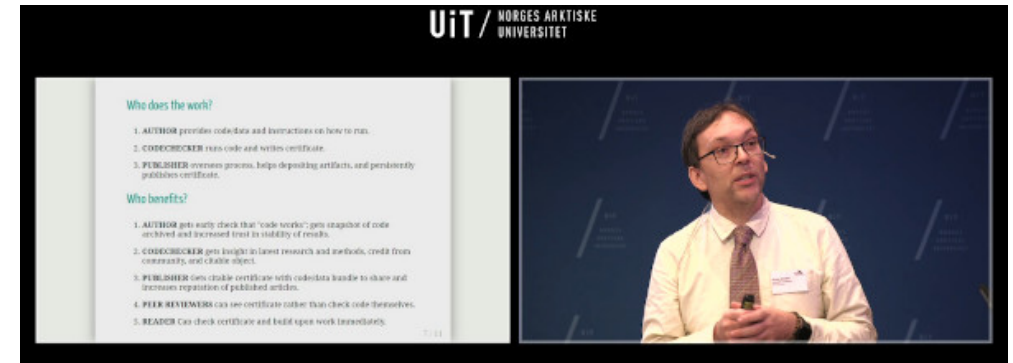
- [github.com/codecheckers](https://github.com/codecheckers)
- [StephenEglen](#)
- [nordholmen](#)
- [Stephen Eglen, University of Cambridge](#)
- [Daniel Nüst, ifgi, University of Münster](#)

The project is supported by a Mozilla Open Science Mini-Grant (see [official announcement](#)) from February 2019 to May 2020.



## Material

Watch Stephen's talk at The 14th Munin Conference on Scholarly Publishing 2019:



# ERCs in peer review



<https://o2r.info/pilots/>

- Collaboration pilots
- OJS pilot

# o2r goals and benefits

o2r

DISCOVER ERCCHRIS NIX | [orcid.org/0000-0001-6523-2935](https://orcid.org/0000-0001-6523-2935)LOGOUTHELP

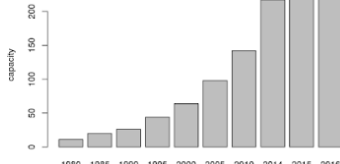
Check Results

o2r team

2017

Abstract

Capacity of container ships in seaborne trade of the world container ship fleet.



Year	Capacity (million metric tons deadweight)
1980	10
1985	20
1990	30
1995	40
2000	50
2005	70
2010	100
2014	150
2015	180
2016	244

(c) Statista 2017

This statistic portrays the capacity of the world container ship fleet from 1980 through 2016. In 2016, the world merchant container ship fleet had a capacity of around 244 million metric tons deadweight. As of January 2016, there were 5,239 container ships in the world's merchant fleet ([source](#)).

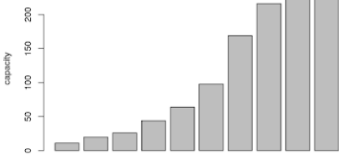
Sources: UNCTAD; Clarkson Research Services, via [statista](#).

o2r team

2017

Abstract

Capacity of container ships in seaborne trade of the world container ship fleet.



Year	Capacity (million metric tons deadweight)
1980	10
1985	20
1990	30
1995	40
2000	50
2005	70
2010	100
2014	150
2015	180
2016	244

(c) Statista 2017

This statistic portrays the capacity of the world container ship fleet from 1980 through 2016. In 2016, the world merchant container ship fleet had a capacity of around 244 million metric tons deadweight. As of January 2016, there were 5,239 container ships in the world's merchant fleet ([source](#)).

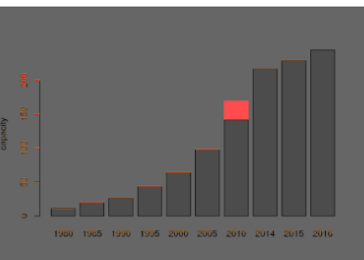
Sources: UNCTAD; Clarkson Research Services, via [statista](#).

o2r team

2017

Abstract

Capacity of container ships in seaborne trade of the world container ship fleet.



Year	Capacity (million metric tons deadweight)
1980	10
1985	20
1990	30
1995	40
2000	50
2005	70
2010	100
2014	150
2015	180
2016	244

(c) Statista 2017

This statistic portrays the capacity of the world container ship fleet from 1980 through 2016. In 2016, the world merchant container ship fleet had a capacity of around 244 million metric tons deadweight. As of January 2016, there were 5,239 container ships in the world's merchant fleet ([source](#)).

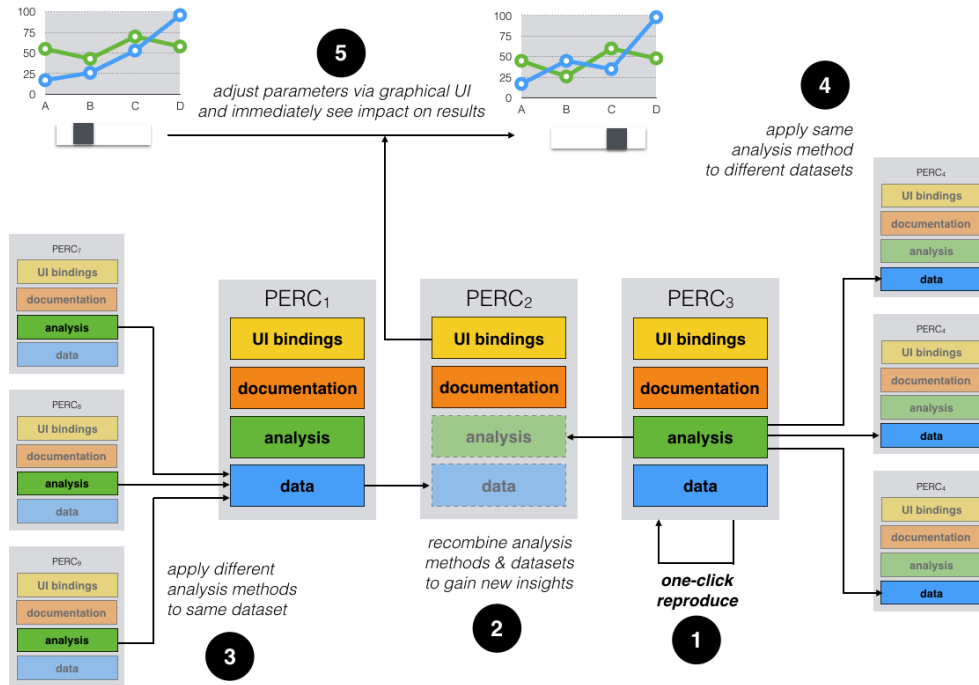
Sources: UNCTAD; Clarkson Research Services, via [statista](#).

Successfully built e031b03dfd2b  
Removing intermediate container 4e28a84808cd

<https://o2r.info/results> | <https://o2r.uni-muenster.de/>

23 / 28

# o2r goals and benefits



Nüst, Daniel, Markus Konkol, Edzer Pebesma, Christian Kray, Marc Schutzeichel, Holger Przibytzin, and Jörg Lorenz. 2017. Opening the Publication Process with Executable Research Compendia. D-Lib Magazine 23 (1/2).

<https://doi.org/10.1045/january2017-nuest>.

Kray, Christian, Pebesma, Edzer, Konkol, Markus, Nüst, Daniel (2019). Reproducible Research in Geoinformatics:

Concepts, Challenges and Benefits. <https://doi.org/10.4230/LIPloc-GISIT-2019-8>



# Take home messages

Rule 1: Have a README: all else is details.\*

*low tech & high tech* approaches exist (CODECHECK & o2r, and more), but we need individuals to be leaders and you can be one by adjusting your habits

*burden* mostly on author, but *benefits* clear and immediate > start early {in career, in project}

*marathon* to change habits and community practice > must push as individuals to achieve the needed *cultural change*

---

## THANK YOU!

Slides: <https://github.com/codecheckers/slides>



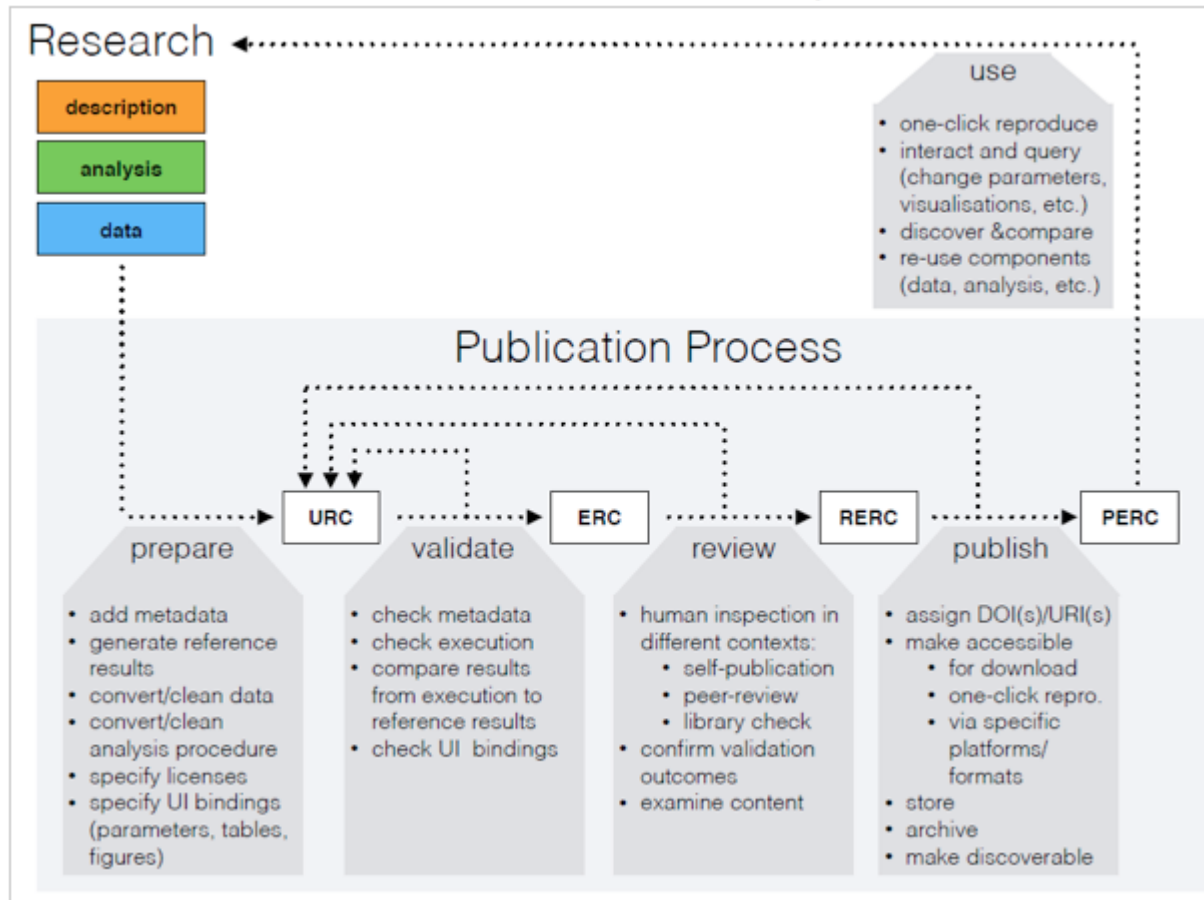
# Computational Research / Data Science 2020

## Challenges

- dependency hell
- FAIR
- licensing
- sensitive data
- big data

## Opportunities/Solutions

- version control
- containerisation
- openness (data, software, preprints)
- community
- self-education
- collaboration
- research integrity
- The Carpentries
- subsets/enclaves/domains



*The ERC-based publication process: a research workspace is prepared for a URC, which is validated to become an ERC, which turns into an RERC after peer review, and eventually published as a PERC to be used, e.g. for subsequent cycles.*

<https://doi.org/10.1045/january2017-nuest>