**Agenda**

1. A brief introduction to notebooks (Hugh Shanahan, RHUL).
2. Publishing notebooks (Martin Fenner, DataCite)
3. Notebooks and long-term preservation (Patricia Herterich, DCC)
4. Notebooks and FAIR digital objects (Christine Kirkpatrick, UCSD/US National Data Service)
5. Notebooks for Big Data & compute (Gergely Sipos, EGI)
6. Break-out groups to discuss topics 2-5.
7. Next steps.

Twitter tag #RDACompNotebooks

Notes for break out sessions

- Publishing notebooks
- Notebooks and long term preservation
- Notebooks and FAIR data objects
- Notebooks for big data & compute

Please complete the following table

| Name | Email | Institution | Interested in break out(s) |
|---|---|---|---|
| Fotis Psomopoulos | fpsom@certh.gr | Institute of Applied Biosciences (INAB\|CERTH) | |
| Amirpasha Mozaffari | a.mozaffari@fz-juelich.de | Forschungszentrum Juelich | |

| | | | |
|---|---|---|---|
| Leonardo Candela (by GoToMeeting) | leonardo.candela@isti.cnr.it | National Research Council of Italy, Istituto di Scienza e Tecnologie dell’Informazione | |
| Maria Sorokina | maria.sorokina@uni-jena.de | Friedrich-Schiller University, Jena, Germany | |
| Christopher Brown | christopher.brown@jisc.ac.uk | Jisc | |
| Daniel S. Katz | d.katz@ieee.org | University of Illinois | |
| Frances Madden | frances.madden@bl.uk | British LIbrary | |
| Patricia Herterich | p.herterich@ed.ac.uk | DCC, University of Edinburgh | |
| André Schaaff | andre.schaaff@astro.unistra.fr | Strasbourg astronomical Observatory / CDS / IVOA | |
| Thomas Jejkal | Thomas.jejkal@kit.edu | Karlsruhe Institute of Technology | |
| David Elbert | elbert@jhu.edu | Johns Hopkins University | |
| Niclas Jareborg | niclas.jareborg@nbis.se | NBIS/ELIXIR-SE | |
| Sirko Schindler | sirko.schindler@dlr.de | German Aerospace, DLR | |
| Eliane Blumer | eliane.blumer@epfl.ch | Ecole polytechnique fédérale Lausanne (remote) | |
| Raphael Ritz | raphael.ritz@mpcdf.mpg.de | Max Planck Computing and Data Facility | |
| Lesley Wyborn | lesley.wyborn@anu.edu.au | National Computational Infrastructure, ANU | |

| | | | |
|---|---|---|---|
| Janos Mohacsi | mohacsi.janos@kifu.gov.hu | KIFU/ Hungarian e-infra provider | |
| Tovo Rabemanantsoa | tovo.rabemanantsoa@inra.fr | INRA, France | |
| Wolmar Nyberg Åkerström | wolmar@ub.uu.se | Uppsala university, Sweden | Notebooks for big data & compute |
| Mingfang Wu | mingfang.wu@ardc.edu.au | Australian Research Data Commons | |
| Jez Cope (remote) | jez.cope@bl.uk | The British Library | |
| Alessandro Costantini | alessandro.costantini@cnaf.infn.it | INFN | |
| Fernando Aguilar | aguilarf@ifca.unican.es | IFCA-CSIC | |
| Rosie Higman (remote) | r.higman@sheffield.ac.uk | The University of Sheffield | |
| Joao Moreira | j.luizrebelomoreira@utwente.nl | VU Amsterdam / U.Twente | FAIR workflows |
| Henry Luetcke | hluetcke@ethz.ch | ETH Zurich | |
| Niek Van Wettere | niek.van.wettere@vub.be | Vrije Universiteit Brussel | |

# Overall notes

Link to GitHub repo: https://github.com/rdanotebooksbof/outline
Target hashtag: #RDACompNotebooks

## Introduction

Examples of notebooks: Jupyter notebooks (the most usual), but not the only. RNotebooks are also a very popular option

There are over 4M Jupyter notebooks on Github!!!

BoF is not about reproducing work done elsewhere.

## Publishing notebooks

Three challenges:

- How do you reference notebooks for specificity and/or credit?
- How do you find reusable notebooks
- How do you find notebooks linked to publication/data/funding?

Audience question: why should we treat notebooks as if they were software for the purpose of publication? It's clear that notebooks aren't the same as data, but it's not at all clear that they are the same as software.

What kind of data is a notebook? Is it a research output or a software

- Research object: contains all - data, software, process

Why do we want to publish the notebook?

- Need a stable version of the notebook

1. Understand the software
2. Get credit
3. Preservation

Where do they fit in the RDA community?

There is no other place they are discussed

Recommendations on how to store a notebook

Notebooks are needed to understand research. And can be a research output to get credit for.

What kind of data is a notebook?
- Research output
- Nearest to software
- All of above or something new?
- → it is not simply sofware
- Notebooks is a research object

Why we want to publish notebooks?
- Because the researcher need it for reusable research
- By introducing a identifier it is stable
- Share same version of a peace of work
- Everything in a notebook changes very fast (data, software, ..)
- In an article you want to cite to the right version
- 3 things: understand, credit, preservation
- Notebooks are not used for reusability more to understand the research software

Where do notebooks fit into RDA?
- Connection to virtual research environments?
- There is no other place
- The interesting part is generating data
- Make publishing of notebooks easier
- Archiving code and runtime environment
- Recommendations of how to build a notebook in order to be reusable
- Concept must be clear not the technology behind it

Conclusion;
- Notebooks are for understanding
- Publishing in an article
- Teaching and Reuse → Credit

## Notebooks and long-term preservation

Why should we preserve notebooks?

Increasingly citing

Becoming a primary research (and teaching) output

Discussion points:

Which comp notebooks should we collect and how to get there

Which elements should we preserve

Who needs to be involved in the discussion

Does this require additional skills?

Audience question: From the point-of-view of preservation, how are these challenges different from those for software preservation more generally?

Preservation:

Long-term: what exactly do we mean by that?

Usually this is 2years. Librarians aim for >10 years.

When to move from an actionable version to something more static

Discussion on the varieties of notebooks: scripts within notebooks instead of independent scripts. What is the vocabulary that can be used to describe these differences. RDA is an excellent place to discuss this.

There are a few Initiatives looking at this, so it might be worth interacting with them.

## Notebooks and FAIR digital objects

Identified a lot of opportunities / gaps (chopportunities) :)

What can this group/RDA community do on this:

1. Establish a WG. Sharing notebooks is different to sharing data/software.
2. How can a notebook provider offer so that it can ensure FAIRness
3. Give recommendations on how to ensure FAIRness.
4. Best practices / known configurations for notebooks
5. Researchers' perspective: what should you do when working with a notebook (workflows)
6. FAIR is supposed to be machine-compatible/readable. How can you ensure this?

7. Is there a limit on what you should (not could) do with a notebook?

Outputs:

Definitions, metrics for notebooks<br>
Making notebooks FAIR

## Notebooks for Big Data & compute

Run on EGI e-infra<br>
Start from JupyterHub

Scalable computing<br>
Scalable data access

Set of questions:

Scalable batch computing from Notebooks?<br>
Reproducibility from different JupyterHub installs?<br>
Communities best practices?

Discussion ( “→” indicates actions/outcomes)

- Composition of the group:
    - ~15 people
    - ⅓ of already using Notebooks for ‘big compute’, ⅔ are looking for examples/solutions to do this, 1 in both situations (how to help students move from local Notebook to big compute Notebooks)
- How to handle big data from Notebooks?
    - Gergely elaborated a bit how the EGI Notebooks does this (with a back-end data management system called DataHub)
    - Recognised connection point with the Data repository interfaces BoF → We could send Notebook use case(s) to that WG
- Batch computing:
    - EGI (and other cloud providers) started with interactive notebooks, now would like to add possibility of batch computing.
    - Supercomputers would like to do the opposite with Notebooks: Use Notebooks to make access more interactive
    - A topic of common interest and need for sharing good practices.
- Running long-running processes from notebooks will bring problems like time-out in browser, how to stay within ‘user quotas on the HPC/HTC site’, etc.

- We were debating/discussing why one wants/needs batch computation from Notebooks
    - Being able to scale up from test dataset to full dataset after the analytical code is working
    - Someone who can write a notebook may not be able to run batch jobs on large parameter space from command line.
    - How can Notebooks simplify this task from an interactive environment? Properties of Notebooks that can scale as the data grows. Condor (and a few other?) libraries exist to do batch computing from Notebook? → Who could check and report back?
- Another use case: A user with many notebooks, each requiring small capacity alone, but together can be a challenge.
    - Isn't this a question of having a powerful server under the Notebooks service?
- Portability of Notebooks across servers → Metadata about Notebooks
    - How much does Binder requirements.txt offer in this respect, and what more do we need?
- Examples of scaling up a Notebook from a local environment to a big compute machine → Do we have/can we have training for users? (the content depends on how the provider actually implements the scalability in its Notebook server)

## General questions

Do we need to talk about reproducibility of notebooks (or the lack of it) as another breakout topic? (or is this part of long-term preservation?)
Does it make sense to connect with similar discussions (such as within the ELIXIR Software Best Practices WG?)