Metadata for Reproducibility
24th Plenary Meeting Collaborative Notes

Group(s) name(s) organising the session: BoF - Alex Ball
Session link:
https://www.rd-alliance.org/members/alex-ball/plenary-participation/?application_id=174649
Session scheduled date/time/breakout session: *10.04.2025, Breakout 10, 14:00 UTC*

Key information:
This document: https://bit.ly/43T1NPd
Group webpages: Metadata IG & Reproducibility IG
Slides: RDA VP24 BoF Slides.pptx

Session summary (for Group co-chairs)

We will use the content in the table below to highlight your work to the RDA community as a report organised by the Technical Advisory Board.
Please complete ALL fields below by **Friday 25th April, close of business** to be included in the report & social media activities.

Summarise your session in three sentences:
We are a long way from the goal of full code/workflow reproducibility, and the journey towards it will consist of many steps. The first step will be to get good practices in place: ones that benefit researchers immediately in their work while laying the groundwork for reproducibility. Then we will need guidance/checklists/self-assessments for putting those pieces together so the work can be reproduced or replicated, and then we will be able to formalize the metadata "glue" for keeping the work reproducible across time and contexts.
Key outcomes/actions/takeaways
1. *Pursue creation of WG to produce recommendations on fundamentals of reproducibility* 2. *Pursue creation of WG to produce technical checklist for reproducibility, with full metadata scheme serialization as a stretch goal.*
Synergies and/or possible collaborations identified with RDA groups and other groups:
Primary group: Reproducibility IG Possible synergies/collaborations with ● Education and Training on Handling of Research Data IG ● FAIR Digital Object Fabric IG ● Metadata IG

- Skills and training curriculums to support FAIR research software IG

Get involved in **RDA Community**
This meeting will take place according to the **RDA Code of Conduct**

Attendee Check-in

Please complete this table to indicate your attendance (add rows as needed):

Full Name	Affiliation	Location	Email
Limor Peer	Yale University	USA	first.last@yale.edu
Lauren Cadwallader	PLOS	UK	lcadwallader@plos.org
Alex Ball	U Bath	UK	ab318@bath.ac.uk
Stephen Richard	CDIF/Astromat/SESAR	USA, Arizona	smrTucson@gmail.com
Chizuko Naoe	Nagoya University Library	Nagoya, Japan	naoe.chizuko.u6@f.mail.nagoya-u.ac.jp
Ramiro Bravo	University of Manchester	UK	ramiro.bravo@manchester.ac.uk
Hugh Shanahan	Royal Holloway, University of London	Egham, UK	hugh.shanahan@rhul.ac.uk
Leyla Jael Castro	ZB MED	Germany	ljgarcia@zbmed.de
Milica Tanasic	Clarivate	Belgrade,Serbia	milica.tanasic@clarivate.com
Chrys Wu	Invest in Open Infrastructure (IOI), home of Infra Finder	IOI is globally distributed. I'm joining from US East Coast Happy to connect on LinkedIn!	chrys@investinopen.org
Marina Chagas	SciELO	Brazil	marina.chagas@scielo.org
Lauriane Kuhn	SCIGNE, IPHC, CNRS, Université de Strasbourg	Strasbourg, France	lauriane.kuhn@iphc.cnrs.fr

Full Name	Affiliation	Location	Email
Myriam Chergui	BRGM	France	m.chergui@brgm.fr
Halle Burns	Princeton University	United States	halle.burns@princeton.edu
Diana Furcila	Spanish Foundation for Science and Technology (FECYT)	Madrid, Spain	diana.furcila@fecyt.es
Fanny Sébire	Institut Pasteur	Paris, France	fanny.sebire@pasteur.fr
Paty Buendia	Lifetime Omics	Miami, USA	paty@lifetimeomics.com
Alison Pamment	UKRI	United Kingdom	alison.pamment@stfc.ac.uk
Jérôme Colin	CNRS	Toulouse, France	Jerome.colin@cnrs.fr
Sarvi Ghafourian	Ocean Networks Canada	Canada	Sarvenazghbm@uvic.ca
Laurence El Khouri	CNRS	France	laurence.elkhouri@cnrs.fr
Adam Vials Moore	Jisc	UK	adam.vialsmooore@jisc.ac.uk
Josh Brown	MoreBrains Cooperative	UK	josh@morebrains.coop
Sabrina Granger	Inria/Software Heritage	France	sabrina.granger@inria.fr
Matthieu Pourieux	CNRS-CREM (UMR 6211), Université de Rennes	Rennes, France	matthieu.pourieux@univ-rennes.fr
Katherine Rial	Helmholtz-Zentrum Berlin für Materialien und Energie	Berlin, Germany	katherine.rial@helmholtz-berlin.de
Christian Pagé	CERFACS	Toulouse, France	christian.page@cerfacs.fr
Johannes Heinonen	Tampere University	Tampere, Finland	johannes.heinonen@tuni.fi

Full Name	Affiliation	Location	Email
Jože Hladnik	ZAG.si	Ljubljana, Slovenija	joze.hladnik@zag.si
Anne-Sophie Bage	INRAE	France	anne-sophie.bage@inrae.fr
Lara Ferrighi	MET Norway	Oslo, Norway	laraf@met.no
Mari Kleemola	TAU-FSD	Finland	mari.kleemola@tuni.fi
Lindsey Anderson	PNNL/MOMSI WG	United States	lindsey.anderson@pnnl.gov;lnanderscience@gmail.com
Claire Austin	Government of Canada	Gatineau, Quebec	claire.austin@canada.ca
Christine Le Bas	INRAE	FR	christine.le-bas@inrae.fr
Milan Ojsteršek	University of Maribor	Slovenia	milan.ojstersek@um.si

Collaborative Session Notes *(To be used by participants and chairs during the session)*

Stay connected

Indicate your interest in continuing the conversation:
https://forms.gle/B24YkKhDpnMz624c8

Agenda

Welcome & Introductions
Brief overview of the Metadata IG & Reproducibility IG
Problem statements: Social and technical
Topics: maFDO, maDMP, verifying reproducibility
Open discussion & wrap up

Notes

Introductions by Limor Peer

REview of Reproducibility IG activities
Review of Metadata IG activities

Lauren-- problem statement-- Social aspects:

- Researchers don't see need to share reproducible.
- why look at code?
    - understand results
    - reuse
    - quality assessment
- how? preferred activity is looking at code repository (e.g. GitHub)
- Researcher don't necessarily know how to share code that is reproducible (get different results, drops errors, doesn't run, dependencies not declared...)
- Skill Gaps (Milan)
- ethical considerations/informed consent (Milan)

Milan-- tech aspects:

- description of software/code, insufficient documentation/metadata, experiment protocol, instrument calibration, statistical methods
- dependency management, version control
- data management, access, licensing, data volume, format compatibility, preprocessing/transformation scripts. access to raw data.
- operating environment-- different hardware, cloud systems, OS, manual workflows; CPU, RAM GPU configurations
- random seeds in workflow..

Discussion:
H. Shanahan. need to identify situations where the need for reproducible documentation is greatest--notebooks? workflows? One other thought is about thinking how to give researchers when their research software is not reproducible. In an ideal world there is an automatic assessment which then says “your software is not reproducible for <X> reasons”.
-- J. Hladnik. problem is heterogeneity of software platforms/languages/approaches
-- HS Agreed - if one tries to look at it all for all possible languages then no progress is made, and hence the strategy suggests moving on one platform.

M. Pourieux. broader problem that use same data/code to get the same result? complexity is different views across disciplines
L. Peer. see need for a lot of metadata; unless this can be automated, can't expect researches to provide all that.
M. Ojstersek. domain experts don't have knowledge of software engineering.
S. Richard. When is it reasonable to expect that code will be reproducible? Requires a lot of work, time consuming, benefits to the researcher

RDA CURE-FAIR work mentioned: https://www.rd-alliance.org/groups/cure-fair-wg/outputs/

W. Akerstrom-- containerized code can help.
J. Castro. metadata is not sufficient, e.g. have to get licenses. use repo platforms that promote good practice
S. Granger Use metadata tools already out there - i.e. Codemeta (see also CodemetaR, CodemetaPy)
https://fair-impact.eu/events/fair-impact-events/meet-software-hash-identifier-do-one-thing-and-do-it-well)

Lightning Talks:
Milan -- Fair Digital Object.
provides tech foundation for reproducibility; need to be machine actionable. (see links in slides) (J. Castro-- aren't FDOs machine actionable by definition?)

Claire Austin. Machine actionable Data management plan (maDMP).
RDA maDMP is in maintenance mode; an extra-RDA group is working on an extension of the RDA maDMP standard see the easy to read preamble https://fairerdata.github.io/maDMP-Standard/ followed by the actual specification, and an even more friendly high level brief. The maDMP emphasizes structure, controlled vocabulary, and no redundancies for interoperability and integration with automated systems. It operationalizes FAIRER (FAIR + Ethical, Reproducible). The maDMP is the engine for collecting information about a project and its data in a discipline-agnostic standardized manner. An maDMP should be continuously updated through the data lifecycle, discipline agnostic (cross domain...). The extension minimizes mandatory fields (about 30 at this point). Many other optional fields are included that can be used when appropriate--applications create profile with specific requirements from this bundle of content items. An entity relationship diagram (ERD) will be posted on the github web page in a week or two. The recently released version is ready for implementation testing and we invite anyone who is interested to work with us for that. Contact claire.austin@canada.ca. I also want to acknowledge the contribution of Dominique Charles, Jennifer Cuffe, and the GCWG to this work.

- J. Castro. Any chance this extension is coordinated with the RDA maDMP group? I think they are reviewing their application profile at the moment, via issues in their GitHub.
- C. A. The maDMP extension that I presented is in a FAIRER Data Management GitHub maDMP repository that is forked from the RDA maDMP GitHub repository.

- C. A. Yes, we are coordinating with the RDA maDMP group. They are working on issues that have been identified to finalize the first batch of updates. After that, they/we are looking at what can be integrated from the extension into the core.

Limor Peer-- verifying reproducibility...
Metadata should indicate what aspects of reproducibility have been verified. Can claims be reproduced. What claims need to be verified; how does verification, how do users trust verification?
various groups have been working on problem (see links in slides)
TOP Guidelines (from Center for Open Science)
IEEE. is code available? is code reviewed (supports reproducibility). Journals give a badge.
ACM-- badges and processes. badges get published
DCAS standards (Data and Code Availability Standard) used by Am. Econ. Assoc.
Cornell Center for Social Science [M. Pourieux Clarification: DCAS are endorsed by a larger group of journals than those within the AEA (including from the very beginning) :)]
ISPS Yale; data in archive must be reproduced and verified. have 'review type' file, document, data, code...

L. Peer The AEA Data Editor checks, so the incentive is to be reproducible: "will assess compliance with this policy, including by conducting reproducibility checks"
https://www.aeaweb.org/journals/data/data-code-policy

checklists... (links in slides)

J. Castro. Some similar topics in relation to code metdata and reproducibility were discussed this in week in the SciCodes meeting https://scicodes.net/

J. Castro. within organization harmonize procedure across repositories. Start with assuring reproducibility for yourself and your immediate colleagues
L. P. agree-- start work towards reproducibility as early as possible in process. have to things about FAIR (and reproducible) from beginning.
L. C. -- would greatly assist publication process if reproducibility verification happened upstream (before pub submission...). There is also the question of scalability and sustainability for journals. It is difficult for high volume publishers to check all the manuscripts submitted to them for reproducibility
C. A.-- big issues are students need to be trained from outset to think about reproducibility/take for granted... have to normalize so happens from beginning to reduce workload on researcher, have to address costs

H. Shanahan to focus on baby steps there is checking that the software actually works and then there's checking that it gives the right numbers for a given data set. If we could even do the first step that would be a good start.

J. Castro-- sample data and expected output should be part of code package.

C. LeBas--How you manage the sensitive data in reproducibility?

-- M. Pourieux. Same question with confidential data. One solution: independent third-party can check that the code runs on the provided data and validates it (or not). See e.g.: https://www.cascad.tech/. Check code in advance with related/similar data; start reproducibility testing early in process, like software unit tests.

-- J. Castro Could the idea of federated Ml work there? The data stays with the owners, but they allow others to use it... some metadata has to be shared

-- L Peer. restricted access to data is a problem for reproducibility. ? provide same access to data for reviewers that researcher had. Simulated/synthetic data can be useful for work around (but more work...)

-- R. Bravo. I assume no sensitive data is published. You could anonymised the data for publication and share that data for reproducibility.

Alex Ball-- summary slide in deck

also Please complete this quick poll to indicate if you are interested in future work (or not!): https://forms.gle/B24YkKhDpnMz624c8

w. Akerstrom Might be interesting to also look at the Workflow Run RO-Crate profiles for some aspect of this, https://www.researchobject.org/workflow-run-crate/profiles/

*****At the end of the session, please fill in this form to indicate whether you are interested in taking any of the discussions further.*****

Chat

15:35:39 From Matthieu Pourieux To Everyone:

Maybe it is a bit naïve but it seems like the objective is a bit bigger than what some may have in mind as "reproducible research" (in my view whether applying the same code

on the same data generates the same results, which is different from replicability: same analysis on a different dataset).

15:37:44 From Hugh Shanahan To Everyone:
One other thought is about thinking how to give researchers when their research software is not reproducible. In an ideal world there is an automatic assessment which then says “your software is not reproducible for <X> reasons”.
Matthieu Pourieux: 👍

15:42:32 From Jože Hladnik (ZAG) To Everyone:
Replying to "One other thought is about thinking how to give re...":
We have Python, R, and many more "programin approaches"... Not possible to check it all.

15:46:18 From Limor Peer To Everyone:
RDA CURE-FAIR work mentioned: https://www.rd-alliance.org/groups/cure-fair-wg/outputs/
Wolmar Nyberg Åkerström: 👍

15:46:35 From Lauren Cadwallader To Everyone:
I agree baby steps are needed! I have been saying this for the past 10 years though 🥲
Limor Peer, Hugh Shanahan, Matthieu Pourieux, Milan Ojsteršek: 👍

15:46:52 From Hugh Shanahan To Everyone:
Replying to "One other thought is about thinking how to give re...":
Agreed - if one tries to look at it all for all possible languages then no progress is made, and hence the strategy suggests moving on one platform.

15:47:34 From Sabrina Granger To Everyone:
All my apologies, I have to go.
My 2 cents: could we build upon existing solutions such as Codemeta? (sse also CodemetaR, CodemetaPy)
Thanks for these interesting insights, looking forward to discovering the checklist.
(and maybe we'll meet again on April 29 to talk about the SoftWare Hash Identifier? https://fair-impact.eu/events/fair-impact-events/meet-software-hash-identifier-do-one-thing-and-do-it-well)
Hugh Shanahan: 👍

15:48:12 From Lauren Cadwallader To Everyone:
Replying to "All my apologies, I have to go. My 2 cents: could ...":

Thanks for sharing your thoughts before you have to go.

15:53:43 From Jael Castro To Everyone:
I thought the FDOs were machine-actionable by definition. What would be a non-machine-actionable FDO?

15:54:21 From Lauren Cadwallader To Everyone:
https://fairerdata.github.io/maDMP-Standard/

15:57:45 From Jael Castro To Everyone:
Any chance this extension is coordinated with RDA maDMP group? I think they are reviewing their application profile at the moment, via issues in their GitHub

15:58:26 From Milan Ojsteršek To Everyone:
Yes FDOs are machine actionable.
Please see EOSC minimum metadata set recommendation on https://docs.google.com/spreadsheets/d/19eJURTWjnrw16WnS_NeX-cOOm_lqEONF/edit?gid=74150701#gid=7415

16:01:52 From Claire Austin To Everyone:
The maDMP extension that I presented is in a GitHub repository that is forked from the RDA maDMP repository.
Jael Castro: 👍

16:02:33 From Chizuko Naoe To Everyone:
I believe that in order to promote interdisciplinary research, metadata should be easy to understand for researchers in other fields.
But this is a very difficult question at what level and how it should be described. Does anyone know of any guidelines or examples that would be helpful?
Milan Ojsteršek: 👍

16:02:47 From Claire Austin To Everyone:
So yes, we are coordinating with the RDA maDMP group.
Jael Castro: 👍

16:06:10 From Jael Castro To Everyone:
Some similar topics in relation to code metdata and reproducibility were discussed this in week in the SciCodes meeting https://scicodes.net/

16:08:23 From Matthieu Pourieux To Everyone:

Clarification: DCAS are endorsed by a larger group of journals than those within the AEA (including from the very beginning) :)

Limor Peer: 👍

16:13:41 From Anu Gururaj, NIAID To Everyone:

Are there are examples of publishers asking for reproducible work?

16:15:27 From Lauren Cadwallader To Everyone:

Replying to "Are there are examples of publishers asking for re...":

I'm sure lots of publishers are asking for it but not policing or enforcing it.

16:15:28 From Limor Peer To Everyone:

Replying to "Are there are examples of publishers asking for re...":

The AEA Data Editor checks, so the incentive is to be reproducible: "will assess compliance with this policy, including by conducting reproducibility checks" https://www.aeaweb.org/journals/data/data-code-policy

16:15:39 From Jael Castro To Everyone:

Replying to "Are there are examples of publishers asking for re...":

It would be great if journals pay more attention to reproducibility but I also think is something that the community needs to promote. For instance, reviewers can ask for the code and the data. It is more work though, for reviewers

16:16:00 From Hugh Shanahan To Everyone:

Again - to focus on baby steps there is checking that the software actually works and then there's checking that it gives the right numbers for a given data set. If we could even do the first step that would be a good start.

16:16:39 From Jael Castro To Everyone:

Replying to "Are there are examples of publishers asking for re...":

Maybe Gigbyte could be an example, I think they have adopted the Data, Optiomizatoina, Model and Evaluation recoommendations for ML in comptutatoinal biology

Lauren Cadwallader: ⬆️

16:16:42 From Anu Gururaj, NIAID To Everyone:

Replying to "Are there are examples of publishers asking for re...":

I would love to see if there are specific examples in the biomedical domain

16:16:58 From Lauren Cadwallader To Everyone:

Replying to "Are there are examples of publishers asking for re...":

There is also the question of scalability and sustainability for journals. It is difficult for high volume publishers to check all the manuscripts submitted to them for reproducibility

Jael Castro: 👍

16:17:41 From Christine LE BAS To Everyone:

How you manage the sensitive data in reproducibility?

16:19:19 From Lauren Cadwallader To Everyone:

Replying to "Are there are examples of publishers asking for re...":

@Anu Gururaj, NIAID, the article I cited by Samuel and Mietchen looked at biomedical articles and certain journals had a higher rate of reproducible articles.

16:19:36 From Matthieu Pourieux To Everyone:

Replying to "How you manage the sensitive data in reproducibili...":

Same question with confidential data. One solution: independant third-party can check that the code runs on the provided data and validates it (or not). See e.g.: https://www.cascad.tech/

Hugh Shanahan: 👍

16:20:11 From Lauren Cadwallader To Everyone:

Replying to "Are there are examples of publishers asking for re...":

Here is the DOI if you need it: 10.1093/gigascience/giad113

Jael Castro: 👍

16:20:48 From Anu Gururaj, NIAID To Everyone:

Replying to "Are there are examples of publishers asking for re...":

Thank you

16:20:56 From Jael Castro To Everyone:

Replying to "How you manage the sensitive data in reproducibili...":

Could the idea of federated Ml work there? The data stays with the owners, but they allow others to use it... some metadata has to be shared

16:21:17 From Ramiro Bravo To Everyone:

Replying to "How you manage the sensitive data in reproducibili...":

I assume no sensitive data is published. You could anonymised the data for publication and share that data for reproducibility.

16:21:38 From Jael Castro To Everyone:

Replying to "How you manage the sensitive data in reproducibili...":

Another option would be synthetic data, but then producing it is also additional work

16:23:17 From Hugh Shanahan To Everyone:

All - many apologies I have to step away at this point. This has been a fabbie session.

16:28:38 From Lauren Cadwallader To Everyone:

Please complete this quick poll to indicate if you are interested in future work (or not!): https://forms.gle/B24YkKhDpnMz624c8

16:29:27 From Wolmar Nyberg Åkerström To Everyone:

Might be interesting to also look at the Workflow Run RO-Crate profiles for some aspect of this, https://www.researchobject.org/workflow-run-crate/profiles/