CURE-FAIR WG: Case Statement

Submission timeline: community draft completed by 24th of April, prepared for TAB submission by May 8th

CURE-FAIR WG Webpage

P14 BOF Session Notes

P14 BOF Session Slides

P15 CURE-FAIR + Reproducible Health Data Services Session Notes

P15 CURE-FAIR + Reproducible Health Data Services Session Slides

---

Case Statement Development and Review Process

In order to be considered by the RDA Council as a recognized RDA Working Group, groups should prepare a Case Statement to be assessed by the community, the Technical Advisory Board (TAB), and RDA Council. Please contact enquiries [at] rd-alliance.org if you want to submit a Case Statement. A Case Statement describes:

1. What is the research case (will the WG produce something useful)?
2. What is the business case (will people use it)?
3. Is there capacity (are the right people involved to adopt and implement).

---

1. Working Group Charter

The goal of the working group is to establish guidelines and standards for curating for reproducible and FAIR data and code (Wilkinson et al., 2016). The ultimate objective is to improve FAIR-ness and long-term usability of “reproducible file bundles” across domains.

When we think of specific research outputs, we might think of data, software, codebooks, etc. These individual outputs may have inherent value. For example, a set of observations that is very costly to produce, or that cannot be repeated, or a script that can be used by others for computation. Traditional curation has considered these outputs as its core objects. But in the context of empirical research, these outputs interact with each other, often to produce specific findings or results. Nowadays, the process by which results are generated is captured in computation. Our approach to curation takes into account this process and focuses on computational reproducibility.

Computational reproducibility is the ability to repeat the analysis and arrive at the same results (National Academies of Sciences, Engineering, and Medicine, 2019; Stodden, 2015). It requires using the data and code used in the original analysis, and additional information about study methods and computational environment. The reason to pursue computational reproducibility is to preserve a complete scientific record , to verify scientific claims, to do science and build upon the findings, and to teach (Elman, Kapieszewski, & Lupia, 2018; Resnik & Shamoo, 2017; Stodden, Bailey, & Borwein, 2013).

In this framework, the object of the curation is a “reproducible file bundle” and its component parts, including the files and their elements (e.g., variables), with the goal of enabling continued access and independent reuse of the bundle for the long term.

The CURE-FAIR WG is focused on the curation practices that support computational reproducibility and FAIR principles.

By curation we refer to the activities designed for “maintaining, preserving and adding value to digital research data throughout its lifecycle” (Digital Curation Center, n.d.).

The WG will deliver,

1) A snapshot of the current state of CURE-FAIR practices drawing upon community surveys and reviews of practice.

2) A synthesis of practices relating to curating for computational reproducibility and FAIR principles.

3) A final document outlining standards and guidelines for CURE-FAIR best practices in publishing and archiving computationally reproducible studies, including the associated computational methods and materials.

2. Value Proposition

The main outcome of this WG is a final document outlining standards and guidelines for CURE-FAIR best practices in publishing and archiving computationally reproducible studies, including the associated computational methods and materials. Bringing together a community of data curators, archivists, repository managers, researchers, scientific publishers, repository software developers, and others can help create consensus around CURE-FAIR standards and prevent fragmentation and development of siloed solutions and practices. Data archives, institutional repositories, scientific publishers, data curators and archivists, and researchers willing to adopt the guidelines and standards will benefit from a standardized approach to CURE-FAIR and what we hope will spur the development of widely-applicable solutions.

3. Engagement with Existing Work in the Area

3.1 Adjacent RDA Groups

The table below lists RDA groups with goals that complement those of the proposed CURE-FAIR WG.

RDA Interest/Working Group	Description	Status
Data Fabric IG	The goal of DFIG is to identify common components and define their characteristics and services that can be used across boundaries in such a way that they can be combined to solve a variety of data scenarios such as replicating data in federations, developing virtual	Recognised & Endorsed

	research environments, and automating regular data management tasks.	
Data Policy Standardisation and Implementation IG	This group continues previous work by JISC, which is committed to working with publishers in supporting the standardisation of journal data policies, with an end goal of supporting machine readable policies that would be easier for researchers and research support staff to utilize in selecting a suitable journal for publication, ensuring compliance with journal and funder data requirements.	Recognised & Endorsed
Data Versioning WG	A review of the recommendations by the RDA Data Versioning IG (the precursor to this group) concluded that systematic data versioning practices are currently not available. The Working Group will produce a white paper documenting use cases and recommended practices, and make recommendations for the versioning of research data.	Recognised & Endorsed; Wrapping up
Education and Training on Handling of Research Data IG	The objective of this IG is the exchange of information about existing developments and initiatives and promotion of training/education to manage research data throughout the data lifecycle. Concretely, it will make the case for creating taxonomies of the skills required by different group of data management specialists/professionals and elaborating reference models.	Recognised & Endorsed
FAIR Data Maturity Model WG	The RDA FAIR Data Maturity Model Working Group develops as an RDA Recommendation a common set of core assessment criteria for FAIRness and a generic and expandable self-assessment model for measuring the maturity level of a dataset.	Recognised & Endorsed
Preservation e-Infrastructure IG	The purpose of the PeIIG is to reach wide agreement on the e-Infrastructure services which are needed to help repositories to preserve their data holdings, to ensure the interoperability of service implementations, and to build trust of service providers.	Completed
RDA/FORCE11 Software Source Code Identification WG	The objective of this working group is to bring together a broad panel of stakeholders directly involved in software identification to develop concrete recommendations for the academic community to ensure that the solutions that will be adopted by the academic players are compatible with each other and especially with the software development practice of tens of millions of developers worldwide.	Recognised & Endorsed
RDA/WDS Certification of Digital Repositories IG	The Interest Group will build on previous work in the area of certification. It will deliver the global overview and the necessary recommendations and requirements that allow the effective implementation of certification	Recognised & Endorsed

	of digital repositories on a national, European and even global level.	
RDA/WDS Publishing Data IG	The Publishing Data Interest Group brings together all stakeholders involved in publishing research data including researchers, discipline specific and institutional data repositories, academic publishers, funders and service providers...to address the implementation of workflows for publishing data and therefore help establish appropriate supporting infrastructure.	Completed
Reproducible Health Data Services WG	The goal of the working group is to improve the reuse of health data by providing recommendations for reproducible data curation and brokerage workflow services.	In Group Revisions
Reproducibility IG	This interest group seeks to advance and enable reproducibility in research based on or producing datasets. Our goals are to provide community based recommendations and infrastructure solutions, doing so in coordination with the other RDA Working and Interest Groups where appropriate.	Completed
Research Data Provenance IG	This group focuses on the comparison and evaluation of models for data provenance. It is concerned with questions of data origins, maintenance of identity through the data lifecycle, and how we account for data modification.	Completed
Software Source Code IG	This interest group will provide a forum to discuss issues on management, sharing, discovery, archival and provenance of software source code. It will pay special attention to source code that generates research data and plays an important role in scientific publications.	Recognised & Endorsed
WDS/RDA Assessment of Data Fitness for Use WG	This working group was formed with the goals of defining criteria and procedures for assessment of data fitness for use, and developing a system of badges/labels communicating fitness for use of individual datasets.	Recognised & Endorsed
Digital Representation of Units of Measure (DRUM)		Proposed
Management of Computational Notebooks		Proposed

3.2 Plan for Engagement with Adjacent RDA Groups

The activities and final recommendations of the CURE-FAIR WG will complement the work of several RDA Working and Interest Groups that are focused on the standards and best practices that support computational reproducibility. Given the multitude of related activities ongoing within RDA working, interest groups, and affiliated organizations, our aim is not to duplicate efforts, but instead synthesize diverse recommended best practices and standards into an easy to use implementation guide. In addition, the majority of related Interest and Working groups address individual portions of the scientific workflow or technical infrastructure. Our aim will be to compile a broad range of surveyed practices and tested standards to address the reproducibility of the entirety of the scientific workflow, including related analysis software, data, and infrastructure components.

4. Work Plan

4.1 Final Recommendation

The final deliverable for the WG is a document outlining CURE-FAIR standards and guidelines for best practices in publishing and archiving computationally reproducible studies, including the associated computational methods and materials.

4.2 Milestones and Deliverables

The group already met at the RDA 14th Plenary as a Birds-of-a-Feather, during which over 100 people gathered to explore the issues concerning curating for FAIR (Findable, Accessible, Interoperable, and Reusable) and reproducible data and code, reflecting the need to confront the challenges of supporting the reproducibility of scientific research. Participants identified as researchers, data support professionals, repository managers, publishers, IT support, and software developers. The discussion at the session has helped determine the focus of the working group going forward.

The WG Case Statement, drafted by the WG co-chairs, was shared for feedback at the RDA P15 and submitted to RDA for official consideration. We hope the group can be formalized before P16 in the Fall of 2020.

From the endorsement of the WG, we will complete the following,

6 months: Milestone 1: CURE-FAIR survey of current state presented at the Plenary

During the first six months of operation, the WG will bring together various stakeholders across the landscape to collect and examine existing practices and standards across disciplines related to reproduction and replication of computational research methods. During our group’s first plenary session at p16 as an authorized working group we will present a draft of existing work describing the current state of curating for reproducibility and the working group’s mid-term objectives to build upon this initial survey. We will also focus our first plenary session and the associated lead-time upon coordinating existing working group membership and recruiting new membership.

The WG will collect use cases, stories, and interviews with researchers trying to reproduce computational workflows to learn about any pain points, especially across domains. The goal is to have a holistic view of the issue from multiple stakeholder points of view, including researchers, data professionals, and technical infrastructure experts across research- domain and geographical boundaries.

Discipline specific practices will be synthesized to analyze the essential elements of reproducibility. Stakeholders of interest include but are not limited to research funders developing and enforcing policies, journals instituting policies for bolstering dissemination and storage of materials (Hrynaszkiewicz et al., 2020), third-party repositories supporting linkage of multiple scientific artifacts, and researchers implementing practices throughout their workflow. These stakeholders may be categorized into sub-groups or working group task forces--an approach shown to be successful previously in other working and interest groups in supporting effective and efficient group coordination and timely progress.

12 months: Milestone 2: CURE-FAIR practices synthesis presented at the Plenary

In the period between p16 and p17, working group members-including sub-groups/stakeholder task forces-will solicit and synthesize input from their representative communities about current practices in curating for computational reproducibility. Best practices will be categorized along key dimensions, including:

- Content descriptions (e.g., reproducible file bundle for social science experiments)
- Formats of data and software accepted
- Workflow descriptions
- Quality certifications

The interim document and categorization of best-practices will clarify the specific curation activities that extend traditional data curation best practices and support computational reproducibility. To support ease of use by various audiences, presentation of these best practices will be in both long-text form describing each practice and strategies for implementation, as well as a matrix categorizing CURE-FAIR curation practices.

18 months: Milestone 3: CURE-FAIR standards and guidelines presented at the Plenary

This final document will outline standards and guidelines for CURE-FAIR best practices in publishing and archiving computationally reproducible studies. We will aim to have representative adopters from each stakeholder sub-group who will implement these recommendations within their own workflows to assess barriers to implementation, needs for further enhancement, coordinated alignment with related stakeholder groups, and overall impact to enhancing the computational reproducibility of the published work.

Milestone	**Deliverable(s)**	**Estimated Delivery**
Intermediate deliverable submission 1	CURE-FAIR current state summary	+6 months
Intermediate deliverable submission 2	CURE-FAIR practices synthesis document	+12 months
Final Deliverable Submission	CURE-FAIR standards and guidelines document	+ 18 months

4.3 Working Group Operations

In addition to meeting in person at plenaries, we will have two or more formal calls in between the plenaries and share information via a mailing list.

Documents will be created and made public through tools such as the Open Science Framework, Google docs, and Box. This allows for collaborative work and also serves as a form of communication. Those individuals actively working on outputs will have ad-hoc meetings as needed (e.g., Zoom). Trello and Github will be used for planning and tracking group deliverables.

4.4 Project Management

The WG's four co-chairs will share the responsibilities for exchanging ideas, updating the group on progress, and engaging with the community. Separate task groups will be formed to work on specific deliverables and will work on these between plenary sessions.

4.5 Community Engagement and Participation

Working group co-chairs and group members will conduct Interviews with active participants who are engaged in CURE and FAIR curation work. Our aim through these interviews is to both best understand community needs for best-practices, barriers to implementation, and increase public awareness of the RDA community and working group as a source for resources. We will aim to engage with aligned efforts across multiple domains, such as the ACM emerging interest group on Reproducibility and Independent Verification, the National Academies, National Libraries of Medicine, and the American Society of Clinical Oncology.

As described in the working group operations section above, project documentation and regular communication with WG members will be managed through open platforms and scheduled to accommodate multiple timezones. To increase openness and agreement of the final deliverable, working group members will be highly encouraged to participate in writing and editing of documents as well as reviewing feedback from adopters.

5. Adoption Plan

The individuals and institutions represented in this WG will endorse and adopt the CURE-FAIR standards and guidelines to spur broader adoption by the academic community. This will be the first document produced by the academic community to establish the guidelines and the standards for curating for reproducible and FAIR research outputs across domains. The WG will engage with various stakeholders, including journals, institutions, data repositories, research labs, and others to formalize the guidelines. At a time when concerns about reproducibility are ubiquitous, community-wide adoption of this document will be a step toward a standardized approach to CURE-FAIR.

6. Initial Membership

FirstName	LastName	Affiliation	Country	Member Type
Limor	Peer	Institution for Social and Policy Studies, Yale University	US	Co-chair
Anthony	Juehne	RDA U.S.	US	Co-chair
Florio	Arguillas	Cornell Institute for Social and Economic Research	US	Co-chair

Thu-Mai	Christian	Odum Institute, University of North Carolina at Chapel Hill	US	Co-chair
Tuomas J.	Alaterä	IASSIST, FSD Finland and RDA-OAB	FI	Member
Susana	Barbosa	INESC TEC, Porto	PT	Member
Ann	Borda	University of Melbourne	AU	Member
Sandor	Brockhauser	European XFEL	DE	Member
Erin	Clary	CARL, Portage Network	CA	Member
Robert	Downs	CIESIN, Columbia University	US	Member
Morane	Gruenpeter	Inria (Software Heritage)	FR	Member
Hilary	Hanahoe	Research Data Alliance	IT	Member
Rob	Hooft	Dutch Techcenter for Life Sciences & Health-RI	NL	Member
Jennifer	Huck	University of Virginia Library	US	Member
Susan	Ivey	NC State University	US	Member
Nick	Juty	ELIXIR	GB	Member
Stefanie	Kethers	Australian Research Data Commons	AU	Member
Henry	Lütcke	ETH Zurich	CH	Member
Dimitra	Mavraki	Hellenic Centre for Marine Research	GR	Member
Nadica	Miljković	University of Belgrade, School of Electrical Engineering	CS	Member
Deepti	Mittal	University of heidelberg	DE	Member
Hollydawn	Murray	F1000	GB	Member
Wolmar	Nyberg Åkerström	Uppsala University	SE	Member
André	Pacheco	University of Coimbra	PT	Member
Keith	Russell	Australian Research Data Commons	AU	Member
Rouven	Schabinger	Karlsruher Institut für Technologie	DE	Member
Sirko	Schindler	German Aerospace (DLR)	DE	Member
Frankie	Stevens	AARNet	AU	Member
Ana	Trisovic	Harvard University	US	Member

Mary	Uhlmansiek	RDA	US	Member
Huajin	Wang	Carnegie Mellon University	US	Member
Brian	Westra	University of Iowa	US	Member
Deborah	Wiltshire	UK Data Service	UK	Member
Lesley	Wyborn	National Computational Infrastructure	AU	Member
Yaqin	Yuan	National Space Science Center, Chinese Academy of Sciences	CN	Member
Qian	Zhang	University of Waterloo	CA	Member

References

Digital Curation Centre. (n.d.). *What is digital curation?* Digital Curation Centre (DCC). http://www.dcc.ac.uk/digital-curation/what-digital-curation

Elman, C., Kapiszewski, D., & Lupia, A. (2018). Transparent social inquiry: Implications for social science. *Annual Review of Political Science*, *21*, 29–47. https://doi.org/10.1146/annurev-polisci-091515-025429

Hrynaszkiewicz, I., Simons, N., Hussain, A., Grant, R., & Goudie, S. (2020). Developing a research data policy framework for all journals and publishers. *Data Science Journal*, *19*(1), 5. https://doi.org/10.5334/dsj-2020-005

National Academies of Sciences, Engineering, and Medicine. (2019). *Reproducibility and replicability in science*. National Academies Press. https://doi.org/10.17226/25303

Resnik, D. B., & Shamoo, A. E. (2017). Reproducibility and research integrity. *Accountability in Research*, *24*(2), 116–123. https://doi.org/10.1080/08989621.2016.1257387

Stodden, V., Bailey, D. H., & Borwein, J. M. (2013). Setting the default to reproducible in computational science research. *SIAM News*. https://sinews.siam.org/Details-Page/setting-the-default-to-reproducible-in-computational-science-research

Stodden, V. (2015). Reproducing statistical results. *Annual Review of Statistics and Its Application, 2*(1), 1–19. https://doi.org/10.1146/annurev-statistics-010814-020127

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*, 160018. https://doi.org/10.1038/sdata.2016.18