Research Tool Usage in Data Curation Workflows: Responsibilities & Collaboration

Plenary session link:
https://www.rd-alliance.org/research-tool-usage-data-curation-workflows-responsibilities-collaboration
Group(s) name(s) organising the session: N/A

Session summary (for Group co-chairs)

Please complete the table below by **7th June, close of business**. The information provided will be used to include your outcomes to the RDA community as a report organised by the Technical Advisory Board.

Summarise the session's key points and discussions in three sentences:
Key outcomes/actions/takeaways
1. *Example* 2. *Example* 3. *Example*
Synergies and/or possible collaborations identified with RDA groups and other groups:

*Get involved in **RDA Community***
*Check out **VP22 programme sessions***

This meeting will take place according to the **RDA Code of Conduct**

Shared presentation slides

- Dijkstra Haugstvedt, N. (2024, May 21). Establishing curation workflows. RDA 22nd Plenary Meeting (RDA VP22), Fully digital. Zenodo. https://doi.org/10.5281/zenodo.11235036
- Bloemen, D. (2024, May 22). Curation @ KU Leuven (Data curation workflows & tools) - BoF session RDA 22nd Plenary. RDA 22nd Plenary Meeting (RDA VP22), Online. Zenodo. https://doi.org/10.5281/zenodo.11243287

- Suchánek, M. & Plankyte, V. RSpace and DSW: Interoperable Tool Provider Joint Case Study. https://doi.org/10.5281/zenodo.11243644
- Migueles, O., & Patarcic, I. (2024, May 27). RDM and data curation support at the Max Delbrück Center. RDA Plenary 22 (RDA P22) , virtual, 14-23 May 2024 (Session Research Tool Usage in Data Curation Workflows: Responsibilities & Collaboration (BoF)). Zenodo. https://doi.org/10.5281/zenodo.11352206

Attendee Check-in from the first session (21 May 2024)

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

Name	Affiliation	Location	Email/social media
Noortje Haugstvedt	UiT The Arctic University of Norway	Tromsø, Norway	noortje.haugstvedt@uit.no
Marek Suchánek	CTU in Prague / DSW / FAIR Wizard	Prague, Czechia	marek.suchanek@ds-wizard.org
Matej Petrič	University of Ljubljana	Slovenia	matej.petric@ff.uni-lj.si
Dieuwertje Bloemen	KU Leuven	Belgium	dieuwertje.bloemen@kuleuven.be
Allyson Lister	University of Oxford	UK	allyson.lister@oerc.ox.ac.uk
Athina Papadopoulou	RDA Europe	Greece	athina.papadopoulou@rda-foundation.org
Carmen Reverté	IRTA (Institute of Agrifood Research and Technology)	Spain	carme.reverte@irta.cat
Tilo Mathes	Research Space	Berlin, Germany	tilo.mathes@researchspace.com / LI
Julia Gehrmann	University Hospital Cologne	Cologne, Germany	julia.gehrmann1@uk-koeln.de
Katherine Rial	Helmholtz-Zentrum Berlin für Materialien und Energie	Germany	katherine.rial@helmholtz-berlin.de

Stéphanie Cheviron	Université de Strasbourg	Strasbourg, France	scheviron@unistra.fr
Esther Collas	CDS, Observatoire astronomique de Strasbourg	Strasbourg, France	esther.collas@astro.unistra.fr
Tom Honeyman	Australian Research Data Commons		
Soizick Lesteven	CNRS, CDS, Observatoire astronomique de Strasbourg	Strasbourg, France	Soizick Lesteven@astro.unistra.fr
Roman Gerlach	Friedrich Schiller University Jena	Jena, Germany	roman.gerlach@uni-jena.de
Parul Tewatia	Scilifelab Data Centre	Uppsala, Sweden	parul.tewatia@scilifelab.se
Nina Grau	INRAE	Montpellier, France	nina.grau@inrae.fr,
Karen Ng Lee Peng	Universiti Malaya	Malaysia	karennlp@um.edu.my
Emma Devine	UKRI	UK	emma.devine@ukri.org
Vashti Galpin	University of Edinburgh	UK	Vashti.Galpin@ed.ac.uk
David Medyckyj-Scott	Manaaki Whenua Landcare Research	New Zealand	medyckyj-scottd@landcareresearch.co.nz
Alessa Gambardella	Leiden University	The Netherlands	a.a.gambardella@science.leidenuniv.nl
Isaí Ugalde Araya	National Center for High Technology	Costa Rica	isai.ugaldea@gmail.com

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

Q&A: add your questions for the presenters (please note down who the question is for if applicable)

- How could you collaborate with other Subject-RDA groups?
    - Ex. preservation tools techs and policies (https://www.rd-alliance.org/groups/preservation-tools-techniques-and-policies/) - perhaps that one, but not enough info there yet
    - Ex. research involvement in data management (FAIR) working group
    - Ex. some specific data groups, that deal with curation, domain and case studies groups
    - (from the zoom chat) there was a related session last week: Active disposal of research data: what you need to know and do before you push delete (https://docs.google.com/document/d/1NRD1dB2Kw0HKUgyzQx6AYZRlqQOg08QrelfjH-4SwfY/edit#heading=h.v0pplo32qg3m)
    -
- There was a related session last week: Active disposal of research data: what you need to know and do before you push delete (https://docs.google.com/document/d/1NRD1dB2Kw0HKUgyzQx6AYZRlqQOg08QrelfjH-4SwfY/edit#heading=h.v0pplo32qg3m)
- Data deletion is a tricky thing: who has the deciding power and what should be the philosophy behind what data should or shouldn't be archived 'forever' (researcher might want to keep everything forever, but is that desirable?) → also a type of curation
    - Comment: Our research institution, because we come under the New Zealand Public Records Act, requires us to keep research data for 20 years, about to rise to 25 years! We already have problems identifying data that is 20 years old to go through a delete/retain discussion.
    - → what data is in scope (the scope is broad and includes related materials)
    - → who's going to pay for it → something the funder leaves up in the air (an issue for institutions, as some research domains produce large amounts of data that are costly to store).
- A good resource for data curation is still the DCC guide "Whyte, A. & Wilson, A. (2010). "How to Appraise and Select Research Data for Curation". DCC How-to Guides.) https://www.dcc.ac.uk/guidance/how-guides/appraise-select-data
-

Questions for all participants: please provide your input on these questions

What does the term "data curation" mean to you? Are we missing an interpretation of "curation"?

- Could be: process of data handling (cleaning, description, documentation...) for data reuse and preservation. Or could be: all datasets submitted to the data repository/system receive consistent treatment to facilitate the reuse and data sharing
- Data curation can include dealing with changes in data, either because corrected data comes in (as with some Covid mortality figures that were updated), or through some curation process (community curation or the curation process that is required for a curated scientific database like https://www.guidetopharmacology.org/). It can be important to track these changes, rather than overwriting the old data with the new data, both to understand differences in analysis on the original data and the updated data, and also to provide reasons for the change (such as updated data, error in data entry, retraction of a journal paper, new evidence, improved data cleaning process, changes made by data cleaning).
- Software that modifies data would appear to be a crucial component of data curation.

What research tools have you found very beneficial for your curation work, and why?

- There are limited tools for curating data changes. Temporal databases provide a starting point for tracking data changes, and development of software to support data curation of data changes (dynamic data). See some examples in https://doi.org/10.1007/978-3-030-80960-7_19 and https://doi.org/10.21014/actaimeko.v12i1.1407
-

What features in those research tools have you found beneficial for curation?

-
-

Would an interest group around this topic be valuable to set up? What should be its focus? (copy and paste an emoji behind your choice in the table: 👍👎🤟🥳😬🧐)

Yes, an IG would be valuable	👍👍👍🤟
No, an IG wouldn't be valuable	

Perhaps, but not with the current scope/information	

In case support is wanted for the creation and running of a more focused RDA WG you can consider applying for RDA TIGER Working Group support services. More info on the TIGER services here and information on the open call here.

- Scoping the meaning of curation also needs to keep in mind different types of data (sensitive vs non-sensitive), what with researchers who say they don't have any data.

Second, Repeat Session (if applicable)

Attendee Check-in from the second, repeat session (22 May 2024)

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

Name	Affiliation	Location	Email/social media
Noortje Haugstvedt	UiT The Arctic University of Norway	Tromsø, Norway	noortje.haugstvedt@uit.no
Dieuwertje Bloemen	KU Leuven	Belgium	dieuwertje.bloemen@kuleuven.be
Marek Suchánek	CTU in Prague / DSW / FAIR Wizard	Czech Republic	marek.suchanek@ds-wizard.org
Coen Wilders	Rijksmuseum	Netherlands	c.wilders@rijksmuseum.nl
Monica Michel Rodriguez	University of Lille library	France	monica.michel-rodriguez@univ-lille.fr
Ana Inkret	Social Science Data Archives	Slovenia	ana.inkret@fdv.uni-lj.si
Limor Peer	Yale University	USA	limor.peer@yale.ed

			u
Andrea Budac	World Data System	Canada	abudac@oceannetworks.ca
Janey Hsiao	NIH	USA	janey.hsiao@nih.gov
Alessa Gambardella	Leiden University	The Netherlands	a.a.gambardella@science.leidenuniv.nl
Ramiro Bravo	University of Manchester	UK	ramiro.bravo@manchester.ac.uk

Collaborative Session Notes *from the second, repeat Session (To be used by participants and chairs during the session).*

Intro:

- Welcome & background of the session
- Focus of this session is on exploring the practical challenges in utilising data curation workflows in RDM tools.

Presentation 1: Case Study KU Leuven

- Different types of curation:
    - DMP reviews → monitoring tool (not more, because lack of machine-readability)
    - Data selection (what to keep, delete or publish?) → difficult to tool due to wide variety of research and use cases
    - Data publication reviews → a tool in place to track & automate feedback generation (review dashboard)
    - Long term preservation appraisal & selection → in the works to integrate in the review dashboard of the data repository
- Pain points/observation:
    - Tools can never run without human resources/investment for support, typically technically inclined personnel
    - Some tools aren't machine interoperable, so hooking curation tooling into it is difficult/impossible

Presentation 2: Case Study UiT - The Arctic University Norway

- Tools:
    - DMP

    - RSpace for documentation of data, assigning IGSNs (provide guidance & set-up support)
    - DataverseNO (provide guidance & dataset curation)
- DataverseNO
    - National curated data repository operated by UiT
    - Consists of 15 partner institutions
- What does curation mean to us?
    - Curation network for FAIR research data: often at the end of a research project, in dialogue with the research arrange the data in line with the FAIR principles and guidelines of the archive
    - Curation is necessary for each dataset that gets published in DataverseNO
    - There is a curation network to help support curators
- Curation responsibilities
    - Depositor/researcher: selection of data & data cleaning
    - Curator: curate and provide feedback
    - Depositor: revise and resubmit
- Curation workflow & tools @UiT
    - New dataset comes in and gets assigned by admin, communication via email or Teams
    - Dataset gets checked for compliance & write curation report (documentation pages, word, teams)
    - Dataset is returned & wait for resubmission (follow up is via email)
    - Publication of dataset & email to depositor & using social media to promote the publication
    - Uses tags in Dataverse to keep track of the dataset statuses
- How can we improve this curation process?
    - A curation tool integrated in the archive that contributes to transparency & traceability
    - For the depositor: curation/feedback while depositing data to make the curation process quicker
- In the perfect world: all the tools in the lifecycle are used an integrated with each other
- In reality: researchers typically use some tools, but don't use their integrations, likely due to a hesitancy to use new flows or new tools
    - Possible solution: give the sense that they're just interacting with one tool
- It's important to understand the user to improve tools and their uptake, though the tools themselves aren't the issue.

Presentation 3: Case Study Max Delbrück Center

- Research performed at the center is around molecular mechanisms of health and disease
- RDM as a centralized service to support researchers to achieve more effective user of resources and transparency across the research life cycle
- The RDM lifecyle isn't fully cyclical; it has internal cycles and returns or different order of steps in certain flows.
- RSpace as an elan solution is used for the collection documentation, description, analysis and sharing of information internally/inside a research group

    - Established in 2021 in the institute, over time the adoption has grown due to training, onboarding efforts.
    - Tools have been really useful in the onboarding of PhD students to improve the uptake of the tools
- OMERO plus for image data managing, specific for microscopic data used for visualization, annotation and analysis via other tools and can be used to store the data and share data internally
    - Established last year and looking at ways to make it attractive to users to get started
    - Working with IT to use the HPC to connect it to Jupyter notebooks to create more pipelines and therefore make it more attractive
- FAIRWizard (based on DSW) used for DMP, but with more expansive support
    - When software is introduced to students, they see the benefits of thinking about all the different RDM steps along the way
    - Usually DMPs are a great conversation starter to think more extensively about the different RDM flows and steps to take
- Future plans
    - DigTools: database of digital tools offered at the center
    - Looking at adopting protocols.io

Presentation 4: Tooling joint case study DSW & RSpace

- During the previous RDA plenary, discussions on integrating the two tools with each other were initiated
- RSpace
    - Electronic lab notebook & sample management system
    - Curation types:
        - Cleanup, selection, publishing, verification
    - In transition to open-source
    - Cloud or on-prem solution
- DSW
    - DMP platform with smart suggestions & advice
    - Covers entire data life cycle
    - Open-source with a cloud-based version FAIRWizard that uses DSW components
- Ideas for an integration, but what is best/most useful for users? Ideas:
    - Automatically update DMPs directly from ELN
    - Linking dynamic DMP from RSpace to DSW (hyperlink as single entry link)
    - Exchange PIDs
    - Storing documents generated by DSW in RSpace
    - Prepare directory structure and files in RSpace based on DMP
- Need to collaborate with institutions
    - Have a feedback loop
    - Understand the responsibilities within institutions, how tools are used, what tools are designed internally
    - User requirements
        - Different views on the word “curation”
        - Clarify the needs & tool focus

    - Seamless integrations are widely preferred, but this should abstract the technical complexity behind the UX
- Option to use out-of-the-box solutions
    - + professional approach, SLAs, support, training
    - + cover various use cases
    - + interoperable and more
    - - require customization by institution
    - - onboarding of users
    - - might not be able to integrate with niche institutional tools
- Option to build own custom solutions
    - + total fit & flexibility
    - + covers own use cases (know exactly what curation means in your use case)
    - + integration with other institutional tools
    - - internal Dev / DevOps cost
    - - still need domain experts to provide workflows & content
    - - reinventing the wheel & not reusable if too specific
- Balancing efforts on tools (nothing is “free”)
    - Always need experts
    - Different focus: technical focus vs customization focus
    - Evaluate whether they can allocate enough effort/resources (not just technical) for own development or to pick up cloud-based tools
    - Would contribute to an existing open-source tool be preferable, though risks of a lot of forking
- Ideal and interoperable design
    - Goal is to automate manual tasks, not to replace curators, as manual work will always remain necessary
    - Adjusting the tool is key for flexibility
    - Experts should transfer their knowledge into the tool
    - Set up standards around schema’s & APIs to optimize interoperability
- Looking for anyone who is interested in joining in on the effort and testing.

Q&A: add your questions for the presenters (please note down who the question is for if applicable)

- How to deal with the time conservation ? How to make the difference between the storage during the project, the tools for sharing like repositories, and archiving over 10-50 years ? Must the tools for curation be the same ?
    - Dieuwertje: I don’t think it’s desirable or possible to have one tool to do all types of curation. Ideally the curation integrated with the existing tools e.g. RSpace for sample management & curation to take away some of the manual work.
- How to deal during curation workflow with which data must be publish and share and what must be archived in the long term ?
    - Noortje: the decision on what to archive is something the research community typically knows best. The general guideline is the ability to reproduce

research results. Though other data might also be interesting to the research community. Though, data from e.g. failed experiments can more easily be removed.
    - Alessa: it could be a workflow tool, where the decision making is for the researcher, but executing it can be part of the tool. It's a human decision-making task. A mindset shift might be necessary to also focus on what data is available for future research instead of just the reproducibility focus.
- In terms of ITF to implement the tools like RSpace or FAIRWizard: how long does it take to implement to get to a first basic set-up?
    - Oscar: There is quite some work necessary behind the scenes. Some preliminary research (literature research, survey, interviews), make requirements analysis, pick one or two options to explore more in-depth the benefits to pick the final choice. It took about 2 years from start to the first test-phase. It depends on the bureaucracy of the institutions.
    - How out-of-the-box is the solution?
        - Oscar: once the tool was decided, it was a matter of months for IT to set it up
        - Marek: from the side of the tool provider: it's dependent on the institution and how technical they are. With FAIR Wizard as a cloud solution, it can be a couple of minutes to get a basic test set-up. But then the time is very dependent on how many people can dedicate their time to the implementation in the institution.
- For Marek: how do you handle researchers' fears around using a new tool?
    - Marek: it depends, we often don't get direct questions from researchers, we often get questions via the data stewards. So, it's mostly about meeting the requirements communicated by the data stewards. It's the choice of the institution to support just one tool in the long term or offer more tools with a possible shorter timespan. Changing tools with the same functionality frequently can be confusing for the researchers
    - Oscar: convincing researchers to use specific tools can be challenging. If the tool is centralised, you need to be able to explain the benefits. Showing the benefits in the long run can work well. Also emphasising the security/reliability of the centralised solution works well. It's up to the institution to ensure availability.
- More of a comment… Alongside tools, we've seen is guidelines or checklists, for example, CURATED by the Data Curation network and RDA's 10 CURE-FAIR Things (and https://curating4reproducibility.org/10things/)

Questions for all participants: please provide your input on these questions

What does the term "data curation" mean to you? Are we missing an interpretation of "curation"?

- [Dieuwertje] Dataset peer review (see yesterday's session "Data Review in Data Repositories to Facilitate Open Science")
- Documentation quality checking
    - Difficult due to many different research domains and domain-specific knowledge necessary to interpret it correctly
    - Limor Peer: Concept of data quality: not just on the data itself, but also on the research itself
        - 10 tips for cure-FAIR
        - The judgement is not so much on what is good science, but rather; are the digital artefacts usable, reproducible, interpretable? → a different perspective.
        - There is a possible connection between this possible IG with the existing reproducibility IG
    - Alessa: we can't control the research/data quality; can discussions around how we care for our data improve research quality. Look at data quality from the start, not just at the resulting data.
-

What research tools have you found very beneficial for your curation work, and why?

- Dataverse-Data-Curation-Tool: https://github.com/scholarsportal/Dataverse-Data-Curation-Tool
- At the Institution for Social and Policy Studies at Yale, we developed a tool, YARD, to curate data and code (and verify computational reproducibility) before publishing in our archive https://isps.yale.edu/research/data/approach
- Take a look also at https://www.bihealth.org/en/quest/service/service/automated-screening-tools

What features in those research tools have you found beneficial for curation?

- One location for deposit, curation, and publication with view available to several user roles
- Tracking and logging curation actions
- Ability to enhance metadata

Would an interest group around this topic be valuable to set up? What should be its focus? (copy and paste an emoji behind your choice in the table: 👍👎🤟🥳😬🧐)

Yes, an IG would be valuable	👍🧐🤟👍
No, an IG woudn't be valuable	
Perhaps, but not with the current scope/information	

- Perhaps initiating a survey to gauge the interest within RDA could prove useful.

Minutes from Q&A

- Dataverse curation tool – to enrich the metadata; an area of curation not discussed in the presentations (for Dieuwertje and Noortje)
- IT efforts - how long does it take to implement the tools? (for Oscar and Marek)
    - Takes quite some time from idea, more than a year…?
    - In case of the wizard, to do a demo is quick, but beyond may take week(s) to properly prepare guidance for users and adapt it in local workflows
- Reproducibility IG
    - Reviewing research output
        - More than just reviewing the data but also the code
- Other groups:
    - Reproducibility
        - reproducibility (and how it relates to data curation!), the Reproducibility IG is having a-synchronous conversations on slack during VP22: Login to RDA Reproducibility Slack Go to #rda-plenary-22 channel
    - PIDS