### **ESIP/RDA Earth, Space, and Environmental Science Interest Group (ESES IG)**

**Coordinating Earth, Space, and Environmental Science Data Preservation and Scholarly Publication Processes Working Group Case Statement**

**Bitly:** [**https://bit.ly/3jXhkp8**](https://bit.ly/3jXhkp8)

**Email: repo2pub\_coord@rda-groups.org**

| V0 | Danie Kinkade, initial drafting pre-P19 |
| --- | --- |
|  |  |
|  |  |

### **Problem Statement:**

### The peer-review publication workflow is focused on making the submission process as easy as possible while still capturing the information necessary from the author team, specifically the corresponding author. For publishers that have policies on citing data, they struggle with how to recommend the best preservation repository for data, and commonly state something like use the “best possible community-accepted repository…and if that does not exist, select a generalist repository such as Zenodo, Figshare, or Dryad.” As long as a repository includes the data that supports the research, and can be cited, the publisher tends to accept the choice. Journal editors that understand the value of FAIR data, might encourage the use of discipline-specific repositories if they happen to know relevant resources.

To accomplish goals to make data as FAIR as possible, funders (in the EU, UK, and US for example) and repository managers desire researchers to work more collaboratively with discipline repositories and get the support for curation that makes data more FAIR.

Making data FAIR requires time to curate the data such that the needed metadata is documented and registered along with the dataset, the data are formatted in a community-accepted standard, and mapped with proper vocabularies. Curation can take several weeks depending on the repository and current workload. Working with the discipline repository during the research project is a necessary collaboration that is a common practice with most research teams. At the time the paper is submitted to a journal, the authors commonly have not considered publishing the data and seek the easiest solution, specifically the generalist repository -- which can register a dataset with little to no metadata in just a few minutes. Data that **should be** published in a discipline repository commonly ends up in a generalist repository because the timeline for curation is longer than the authors (or journal) has patience to pause the publication. Further, there is concern that during the journal peer review process, requests to update or gather more data might occur and there is confusion for how that request is resolved if the data were already published/preserved. Also, should the data be published prior the publication being accepted and published, could someone “scoop” the research.

Today, there is little to no communication between journals and data repositories. There are no best practices to use as guidance And more and more datasets are placed in generalist repositories where they have a higher likelihood of being poorly documented and not FAIR.

This working group is being established to address this problem.

### **Background and history:**

Increasing effort is being directed toward streamlining the activities surrounding research data publication, especially with respect to data that support the findings of scholarly publications. Yet, we still do not have a complete understanding of the highly variable data publication process, including the requirements and detailed workflows of individual stakeholders, such as the journal editor, author, and repository, and the touch points between them.

Bottlenecks are arising at these stakeholder touchpoints, as domain and institutional repositories are becoming increasingly challenged to fit into the scholarly publication workflow. This situation is exacerbated by a current lack of awareness and communication across these relevant stakeholders. Reaching consensus on stakeholder needs and processes is a first step in addressing the challenges associated with data sharing, publication and citation, and in better aligning these workflows to facilitate Open Science.

Within RDA, a number of working groups have touched upon various aspects of the data publication process through the course of achieving their respective working group objectives. One such group is the RDA/WDS Publishing Data Workflows Working Group, which sought to describe and document an idealized workflow for data publication workflow as a final step of the research process (Dallmeier-Tiessen, et al., 2016 <http://doi.org/10.15497/RDA00004>, and Austin, C.C., Bloom, T., Dallmeier-Tiessen, S. et al. Key components of data publishing: using current best practices to develop a reference model for data publishing. Int J Digit Libr 18, 77–92 (2017) <https://doi.org/10.1007/s00799-016-0178-2>). This output recommended an ideal joint workflow of publishers and repositories at a time when both workflows were separated. However it does not capture detail in touch points between stakeholder processes, such as the complexity in interaction between the researcher and the repository as the paper goes through the journal manuscript publishing process. Additionally, asynchronous adoption of any recommendations may have occurred, such that journal and repository workflows remained and continued to evolve separately.

The Data Policy Standardization and Implementation IG developed a recommended set of research data policy features (Hrynaszkiewicz et al., 2020 <http://doi.org/10.5334/dsj-2020-005>) for use by journals and publishers. This data policy framework contains features that relate to activities and elements of the data publication workflow on behalf of all stakeholders (i.e., journal or publisher, researcher and author, repository, and funder), however does not address the timing in provision of various policy features as data move through the respective scholarly and data publication workflows.

We propose undertaking a more detailed approach to document critical touch points between data publication stakeholder workflows, in addition to necessary requirements and timing of certain activities that will complement and support the prior RDA-endorsed workflows and policy features while providing additional, much needed guidance for journal editors, repositories, and research authors to help streamline research data sharing in support of the scholarly publication process.

### **Charter:** A concise articulation of what issues the WG will address within a 18 month time frame and what its “deliverables” or outcomes (including a Recommendation) will be.

*This working group aims to improve collaboration and coordination between repository and publisher workflows as they pertain to- and rely upon the data publication process.* The group proposes to refine and formally document a detailed workflow(s) between the journal, publication author, and Earth, Space, and Environmental Science disciplinary data repository and identify cross-stakeholder dependencies to recommend activities that are necessary for relevant processes to move forward in a more coordinated way. The WG intends to advance current progress on this effort by using the specific [use cases](https://docs.google.com/document/d/1Qig8O2cjuMkDrmG4e9_JEs64SHzsXSiVpgJ-YtK1yGs/edit?usp=drive_link) of data submission to a data repository for curation and preservation at, or near, the time the respective manuscript is submitted to the journal that will cite the data.

The WG plans to engage journal, repository, and research communities to solicit feedback that result in a set of recommendations for better transparency and coordination across both processes.

This work has been presented for discussion at prior working sessions of the Earth, Space and Environmental Science Interest Group at Research Data Alliance Plenaries (P17, P18, P19, P20, P21, P22) and the Earth Science Information Partners meetings over the past four years. The group experienced challenges in continuity during the COVID-19 Pandemic, however is now enjoying lively engagement with the RDA community and additional external stakeholders.

#### **Specific Objectives:**

* **Establish a shared understanding of ESES repository requirements** concerning relationship timeline with researchers, best practices for appraisal and deposition, and value of discipline-specific curation.
* **Establish a shared understanding of journal needs concerning data availability** and citation and preferred timing.
* **Educate and inform data publication stakeholders on the workflow of journals and the touchpoints with the author** in respect to ESES data that underpins a scholarly publication.
* **Document a set of use cases that demonstrate the current (somewhat disconnected) and preferred steps** in the publication processes (for repositories: soliciting World Data System members, and international repository members within RDA, US Council of Data Facilities, and ESIP Type I members; for publishers: starting with organizational RDA member publishers, noting to including editors here!)
* **Creation of a visual model of** [**publication workflow** **for journals**](https://drive.google.com/file/d/1d0-xUhp_rUOJ01SaQhamepPrdZiuUoRp/view) and discipline-specific [data workflow for repositories](https://drive.google.com/file/d/1qtRZmFl-sFoDtUgsLmWFQWgb4umnfIQg/view) based upon use cases.
* **Create a set of recommendations**
	+ Containing actions that support and facilitate data and scholarly publication workflows

### **Value Proposition:** A specific description of who will benefit from the adoption or implementation of the WG outcomes and what tangible impacts should result.

Increasing emphasis is being placed on research data publication, especially with respect to data that support the findings of scholarly publications. This new and growing pressure is creating bottlenecks at the touchpoints between data publication workflows of primary stakeholders such as journal editors, authors, and repositories. Yet, we still do not have a complete understanding of these highly variable data publication workflows, including their requirements and individual processes. Domain and institutional repositories are becoming increasingly challenged to fit into the scholarly publication process. This situation is exacerbated by a current lack of awareness and communication across these relevant stakeholders. Reaching consensus on stakeholder needs and processes is a first step in addressing the challenges associated with data sharing, publication and citation, and in better aligning these workflows to facilitate Open Science.

This group will: bring together key stakeholders to build on and extend earlier work, establish a common understanding of stakeholder workflows and needs, and reach consensus to develop recommendations necessary to streamline data publication in support of scholarly publication.

The intent of this work is to ease repository bottlenecks, facilitate data review, and ease effort and increase transparency for authors; ultimately streamlining the data and scholarly publication processes.

Primary adopters will include:

1. Repository managers: through repository policy, guidelines and education (as facilitated through the US Council of Data Facilities (CDF, <https://www.esipfed.org/get-involved/collaborate/council-of-data-facilities-cdf>), the Earth Science Information Partners (ESIP) Type I members (<https://www.esipfed.org/partners>), RDA, World Data System, and other participating repositories)
2. Journal publishers: (1) at the journal-level, through journal policy, guidelines and education for authors and editors (facilitated by societies and related entities such as AGU/EGU and the Coalition for Publishing Data in the Earth and Space Sciences (COPDESS, <https://copdess.org/>) and (2) at a publisher-level, through systems support, workflow design, staffing capacity, and third party service providers (e.g. of submission systems)
3. Researchers and other data producers: through raised awareness and behavioral change achieved through direct interactions with both repository and journal entities.

### **Engagement with Existing Work:** A brief review of related work and plan for engagement with any other activities in the area.

The group recognizes the work of RDA/WDS Publishing Data Workflows WG (Dallmeier-Tiessen, et al., 2016 <http://doi.org/10.15497/RDA00004>). RDA/WDS Publishing Data Workflow WG output represents an idealized workflow, but does not capture touch points between stakeholder processes. The results from this working group predominantly focus on “static” data publishing, i.e. data publishing as one of the final steps of the research process. To accommodate an emerging interest to connect the research workflow and data publishing, this second, more detailed workflow analysis has also been initiated to take into account interaction with the researcher and the repository as a manuscript goes through the review and publishing process and allows for the need to modify both the publication and/or the data during the peer review process. The new workflow also considers the possibility that the paper may be rejected. This more detailed work is intended to complement and extend the prior RDA-endorsed workflow by providing additional and much needed specific guidance for journal editors, repositories, and research authors.

This proposed group intends to engage with the RDA/WDS Certification of Digital Repositories IG and its working groups such as the Publishing Data Workflows WG, which cover various aspects of repository best practices. The World Data System (WDS) is an RDA membership organization covering different domain repositories and a variety of different workflows in data publishing, and is seen as a valuable resource to the proposed work described in this case statement. The working group will have overlapping membership with the RDA/WDS Certification of Digital Repositories Interest Group and plans to collaborate with the WDS on use case collection and adoption of our group’s outputs.

The RDA [Data Policy Standardization and Implementation IG](https://rd-alliance.org/groups/data-policy-standardisation-and-implementation-ig) developed a recommended set of research data policy features (Hrynaszkiewicz et al., 2020 <http://doi.org/10.5334/dsj-2020-005>) for use by journals and publishers. The policy framework describes features that relate to, and align with, certain activities and elements of the data publication workflow; for example, use of domain repositories, provision of data availability statements, and degree of data peer review performed. Certain framework features themselves may establish dependencies to the proposed work based upon the timing of their provision as data move through the respective scholarly and data publication workflows. This effort will carefully consider the framework features and the impacts of each on the timing of proposed workflow activities and adjust accordingly.

In spring 2022, four journals within the Springer Nature publisher portfolio (Nature Energy, Nature Chemistry, Nature Ecology and Evolution, and Nature Neuroscience) announced a pilot collaboration with the generalist repository, Figshare, to ensure authors were presented with an option for sharing the data associated with manuscripts submitted to the journal ([Nature Energy, 2022](https://doi.org/10.1038/s41560-022-01025-6); see also [Figshare Integration Guidance For Authors](https://www.springernature.com/gp/authors/research-data/figshare-integration)). This service will be offered to authors upon submission to the journals as part of the journal’s submitter workflow. This activity highlights a single data sharing use case that is driven by submission of the scholarly publication. Although it attempts to move consideration of data management earlier in this publication workflow (from after the manuscript has been submitted to the time an author begins their manuscript submission), the emphasis is still only at the time of scholarly publication, when much data curation and management may still need to be performed in order to maximize the understanding and reuse of the data. This use case also only considers the data relevant to the scholarly publication, not the full corpus of data potentially generated through the course of a research project (i.e., exploratory data collection, negative results, etc.). Perhaps most importantly, by collaborating and promoting a general repository, discipline-specific data curation may not be considered or promoted, even though domain repositories may exist and provide a better fit for the types of data supporting the manuscript and their curation needs prior to peer review, potentially making the review process less efficient and/or effective. The proposed here work will provide a set of recommendations that can address these considerations and ease the workflow challenges faced at the time of manuscript submission.

This effort also acknowledges work performed by publishers and journals collectively (and facilitated within the [RDA FAIRsharing WG](https://www.rd-alliance.org/group/fairsharing-registry-connecting-data-policies-standards-databases.html)) to create a set of criteria for selection of repositories through which authors can share their data ([Sansone et al., 2020](https://doi.org/10.5281/zenodo.4084763)) and plans to reference the document for journal requirements and needs during the work proposed in this case statement. Similarly, this effort will engage with the ongoing [RDA Data Repository Attributes WG](https://www.rd-alliance.org/groups/data-repository-attributes-wg), which is building on previous work by Sansone et al. 2020, to create a list of common attributes that describe a research data repository and providing examples of the current approaches that different data repositories are taking to express and expose these attributes.

Finally and as needed, this proposed effort can utilize existing resources (e.g. [re3data](https://www.re3data.org/) and the RDA-linked [FAIRsharing](https://fairsharing.org/)) that are designed to help researchers discover and choose the resources they need to curate and publish their data in alignment with FAIR.

### **Work Plan:** A specific and detailed description of how the WG will operate including:

* The form and description of the final Recommendation of the WG,
	+ A document outlining a set of recommendations for journals and repositories, containing actions that support each of these stakeholders’ workflows and requirements for manuscript and data publication processes respectively.
	+ A visual model of representative workflows for journals and repositories based on repository use cases
* The form and description of milestones and intermediate documents, code or other deliverables that will be developed during the course of the WG’s work
	+ Months 1-6: Document a set of repository use cases that demonstrate bottlenecks in publication processes.
	+ Months 6-12: Creation of a generalized visual model of publication workflow for journals and detailed data workflow for repositories based upon repository use cases.
	+ Months 12-18: Creation of a set of recommendations:
		1. For journals, containing actions that support and facilitate repository workflows.
		2. For repositories, containing actions to better support journal manuscript publication process needs.
* A description of the WG’s mode and frequency of operation (e.g. on-line and/or on-site, how frequently will the group meet, etc.),
	+ The group will plan to meet through a combination of monthly online meetings, ad hoc meetings as required, and RDA ESES IG Plenary sessions.
		1. Report outs and running minutes will be organized by chairs
		2. Progress will be shared at RDA Plenaries
	+ Asynchronous contributions to online documents will allow the group to make additional progress outside of meeting times. The group will use collaborative editing tools (e.g., Google Suite) to share meeting minutes, key documents and to solicit feedback as necessary.
* A description of how the WG plans to develop consensus, address conflicts, stay on track and within scope, and move forward during operation
	+ Consensus within the WG will be achieved through monthly meetings and email discussions. Any conflicts encountered will be resolved through facilitated discussion and adjudicated by the Co-chairs who have had extensive experience in managing and leading volunteer groups. WG milestones are tangible outputs by which the Co-chairs will gauge progress and ensure adherence to scope.
* A description of the WG’s planned approach to broader community engagement and participation.
	+ Engaging the relevant data publication stakeholders to build consensus on a path forward is intrinsically tied to a shared understanding of these individual groups’ needs and challenges. Much ground work has been laid by the ESES IG to engage these stakeholders through previous meetings and opportunities to share use cases and promote discussion. The WG will continue this practice through RDA Plenaries along with ESIP semi-annual meetings and leveraging relevant organizations such as COPDESS, AGU, and the US Council of Data Facilities.

### **Adoption Plan:** A specific plan for adoption or implementation of the WG Recommendation and other outcomes within the organizations and institutions represented by WG members, as well as plans for adoption more broadly within the community. Such adoption or implementation should start within the 18 month timeframe before the WG is complete.

* + Socialization and engagement through the organizations listed above (i.e., COPDESS, EGU, AGU, US CDF, ESIP, ARDC, WDS) is planned to target relevant stakeholders (i.e., repositories, researchers, and publishers) from which the WG will gather feedback on adoption and implementation feasibility. The WG will adjust and refine as necessary before the completion of the group’s 18 month term.
	+ Repositories
		1. World Data System [international members](https://worlddatasystem.org/members/member_directory/) (there will be overlap with the above)
		2. Members of the US Council of Data Facilities
	+ Publishers
		1. AGU
		2. Signatories of COPDESS

### **Initial Membership:** Need 2-4 Chairs, and broad international participation

| **Member Name** | **Affiliation** | **Role** |
| --- | --- | --- |
| Danie Kinkade | BCO-DMO, US | Co-chair |
| Lesley Wyborn | ANU, AU | Co-chair |
| Natalie Raia | University of AZ, US | Co-chair |
| Matthew Warke  | Taylor and Francis, UK | Co-chair |
| Shelley Stall | AGU, US | OA Representative |
| Martina Stockhause | DKRZ, DE | member |
| Allyson Lister | FairSharing | member (Liaison with Data Repository Attributes WG) |
| Chris Burbidge | iCRAG/ SFI, IE | member |
| Kristina Vrouwenvelder | AGU, US | member |
| Helen Glaves | BGS | member |
| Kerstin Lehnert | IEDA2; Astromaterials Data System (both US) | member (Co-chair of IG for Domain Repositories & Physical Samples) |
| Natalia Atkins | AODN, IMOS, AU | member |
| Milan Ojsteršek | University of Maribor, Slovenia | member |
| Meredith Goins | World Data System | member, (Co-Chair of [RDA/WDS Certification of Digital Repositories IG](https://www.rd-alliance.org/groups/rdawds-certification-digital-repositories-ig.html)) |
| Reyna Jenkyns | World Data System | member |
| Joan Damerow | ESS-DIVE data repository, Lawrence Berkeley National Lab, US | member |
| Maggie Hellstrom | ICOS Carbon Portal (hosted by Lund University, Sweden) | member (representing a community repository) |
| Sarah C Davidson | Movebank Data Repository, Max Planck Institute of Animal Behavior, University of Konstanz (Germany) | member |
| Liisa Lehtsalu | European Academy of Bolzano-Bozen (Eurac Research), Italy | member |
| Alice Fremand | British Antarctic Survey | member |
| Adam Mansur | National Museum of Natural History, Smithsonian Institution | member |
| Guillaume Wright | PLOS | member |
| Bruce Wilson | ORNL DAAC | member |
| Robert Downs | Columbia | member |
| Deb Agarwal | LBL, Ameriflux | member |
| Kyle Copas | GBIF | member |
| John VanDecar | Nature | member |
| Susana Barbosa | INESC TEC, | member |
| Elisha Wood-Charlson | Berkeley Lab, NMDC | member |
| Adam Leary | Oxford University Press, UK | member |
| Zhifang Tu | National Science Library, Chinese Academy of Sciences | member |
| Usharani K | Department of Computer Science, Sri Padmavati Mahila Visvavidyalayam  | member |
| Gustavo Adolfo Arteaga Botero | Director de Arquitectura en Pontificia Universidad Javeriana, Universidad del Valle (CO)Cali, Valle del Cauca, Colombia | member |
| Morufu Raimi | Federal University Otuoke, Nigeria | member |
| Daniela Cialfi | University of Chieti-Pescara, Italy | member |
| Tadele Sh.Gerasu | Jimma University, Ethiopia  | member |