MEETINGS

Interoperable infrastructure for software and data publishing

STEPHAN DRUSKAT, KRISTI HOLMES, JOSE BENITO GONZALEZ LOPEZ, LARS HOLM NIELSEN, STEFANO IACUS, ADAM SHEPHERD, JOHN CHODACKI, DANIE KINKADE and GUSTAVO DURAND

March 28, 2023 . 7:50 AM — 6 min read

https://doi.org/10.54900/e21jg-1b369

Image via Pexels.com

Research data and software rely heavily on the technical and social infrastructure to disseminate, cultivate, and coordinate projects, priorities, and activities. The groups that have stepped forward to support these activities are often segmented by aspects of their identity - facets like discipline, for-profit versus academic orientation, and others. Siloes across the data and software publishing communities are even more splintered into those that are driven by altruism and collective advancement versus those motivated by ego and personal/project success. Roadblocks to progress are not limited to commercial interests, but rather defined by those who refuse to build on past achievements, the collective good, and opportunities for collaboration, insisting on reinventing the wheel and reinforcing siloes.

In the open infrastructure space, several community-led repositories have joined forces to collaborate on single integrations or grant projects (e.g. integrations with Frictionless Data, compliance with Make Data Count best practices, and common approaches to API development). While it is important to openly collaborate to fight against siloed tendencies, many of our systems are still not as interoperable as they could and should be. As a result, our aspirational goals for the community and open science are not being met with the pacing that modern research requires.

In November 2022, members of open, community-led projects that support research data and software came together during an NSF-funded workshop to address the above concerns and outline actionable next steps. As builders and educators that regularly work and collaborate together, we wanted to leverage our trusted relationships to take a step back and critically examine and understand the broader infrastructure, systems, and communities above and around us that influence our success and goals. Through this process, we identified and categorized key problem areas that require investment: lack of standards, false solutions, missing incentives, cultural barriers, the need for training and education, opportunities for investment in new/existing infrastructure, need for support for sensitive data, and lack of guidance/support for leadership in our communities.

From there, we built the first steps of a plan to collectively improve the scalability, interoperability and quality of data and software publishing.

If we did it right, what would it look like

Make data and software publishing as easy as possible – not just at one repository

Many data and software repositories market their submission processes to be seamless. But seamless workflows within a single repository are not enough. Researchers are not committed to a single repository, and data often need to be linked across repositories for successful reuse. Drastically disparate processes hinder us from meeting our goals. Making data and software publishing as easy as possible requires us to look upstream and invest in more accessible and interoperable tooling, as well as in education and training.

Investments in infrastructure should prioritize tools that focus on pipelines from data and software to repositories, while remaining platform agnostic and openly pluggable. Workflows differ across disciplines, but basic command line functions for API submission to repositories should be a baseline requirement.

There is also room to develop more educational investments in the form of training materials and experiences, beginning at the high school and undergraduate level or earlier, when students are discovering personal interests and eager to communicate and share their ideas. Students and non-students alike can nurture their excitement through hands-on exploration with data and code. The intersection of an interesting topic with critical questions, such as what is data and software publishing, how do I publish data and software, and basic computational skills for seamless publishing, can provide opportunities to nurture lifelong discovery and sharing. Groups like The Carpentries are optimal for building out needed modules and upskilling researchers earlier in their careers.

As we improve the processes for getting data and software into diverse repositories, we can begin to think about what it would look like for our repositories to expose standard sets of disciplinary metadata to allow for automated linkages. This type of work would drastically change search capabilities, required for use and reuse of all research outputs, including data and software.

Build scalable processes for ensuring quality and reusability

Repositories have varying levels of support for assessing the quality of hosted data and software, ranging from curation services and automated validation of files or metadata, to documenting and enforcing best practices. This work should be coordinated across repositories to ensure researchers can easily understand expectations and leverage standards; importantly, these expectations should be built into curriculum upstream from the data and software submission processes.

There are emerging standards for this work. They rely on building a clear discipline-specific understanding of the data and software and offer contextual support to create machine-readable information about the files. These standards and approaches must be coordinated and offered at scale across multiple repositories in order for them to be successfully adopted and iterated on at the rate researchers need.

Researchers reusing data should understand the usability and the fit for purpose of each dataset or software package. This cannot be adequately addressed by a mere badging scheme. To properly address this challenge and support trust and reuse, effective interfaces are needed to gauge the level of metadata and quality of data and software up front. Importantly, and unrelated to disciplinary metadata, there must be an emphasis on provenance. Data and software published without provenance information should be flagged as untrustworthy.

Examples of tools for this type of work are available. For software, the HERMES workflow tool can record basic software provenance - e.g. contributorship and authorship, versioning, dependencies - and records the provenance of the software metadata itself during the submission preparation process. For data, leveraging strategies such as data containerization facilitates the use of flexible open source tooling for manipulating and processing data during curation. Frictionless Data’s Data Package Pipelines employs reusable and structured declarative statements to execute routine actions on the behalf of the data curator; creating machine readable package manifests and provenance records, while decreasing human error and increasing efficiency in data processing. We know that investments into and adoption of these types of tools are essential to our greater success.

Launch a productive community for change

Broad coalitions across research data and software platforms exist and have a place in defining community benefit and approaches. However, they also stand in our way of action. We need a venue to openly discuss ideas, and we need to trust that collaborators will offer resources openly and productively, not just showing up for attendance’s sake but rather, be invested in building a fertile ecosystem together. While it may be unpopular in some circles, this will mean building an exclusionary space. One where the members have pledged to support the collective benefit over individual reward.

This type of community already exists. We just haven't formalized it. Now is the time to move quickly toward our common goals. This type of space is required for coordination across stakeholders to build clear examples of the ROI of our investments into data and software publishing, and better integrate leadership (across all stakeholders) into the conversation.

Committing resources and intentional work - and not just showing up

Achieving scalable, high-quality, interoperable data and software publishing is possible. There are already builders, some represented by the authorship of this article, that are on the right path, building tools that effectively meet the needs of researchers in an open and pluggable way. One example is InvenioRDM, a flexible and turn-key next-generation research data management repository built by CERN and more than 25 multi-disciplinary partners world-wide; InvenioRDM leverages community standards and supports FAIR practices out of the box. Another example of agnostic, pluggable tooling, in this case for software submission, are the submission workflow tools currently developed in the HERMES project. These allow researchers to automate the publication of software artifacts together with rich metadata, to create software publications following the FAIR Principles for Research Software.

Meaningful progress and lasting success requires people to do real work and commit real resources. Success requires community-led and community-serving projects across multiple scholarly and research communities to rally behind and support those driving progress in data and software publishing and adoption of best practices and community standards that enable a bright, interoperable, and function-forward scholarly ecosystem. Success will also depend upon transparency to shine a light on vanguards leading this journey, as well as exposure and an understanding of conflicting motivations and interests that prioritize good PR and drain energy and resources from the community. Success ultimately requires true collaboration, with a mindset of “for the community - long-term” as opposed to “for my project - right now”, and focused action to deliver results and solutions.

The time is now! We are highly committed to this vision by working together to build the community and technical structures needed to finally advance data and software publishing across research disciplines.

This post was cowritten by all attendess of the NSF- funded workshop:

John Chodacki, California Digital Library
Maria Praetzellis, California Digital Library
Gustavo Durand, Harvard
Jason Williams, Cold Spring Harbor Laboratory
Stefano Iacus, Harvard
Adam Shepherd, BCO-DMO
Danie Kinkade, BCO-DMO
Kristi Holmes, Northwestern/ InvenioRDM
Maria Gould, California Digital Library
Matt Carson, Northwestern
Stephan Druskat, German Aerospace Center (DLR)
Britta Dreyer, DataCite
Jose Benito Gonzalez, Zenodo/CERN
Kristian Garza, DataCite
Steve Diggs, California Digital Library
Lars Holm Nielsen, Zenodo/CERN