Implementing Effective Data Practices:

Stakeholder Recommendations for Collaborative Research Support

September 23, 2020

Association of American Universities
Association of Public & Land-grant Universities
Association of Research Libraries
University of California CDL California Digital Library

Table of Contents

Association of Research Libraries
21 Dupont Circle, NW
Suite 800
Washington, DC 20036
T 202.296.2296
F 202.872.0884
ARL.org
pubs@arl.org

Committee on Implementing Effective Data Practices: A Conference on Collaborative Research Support

John Chodacki
California Digital Library

Cynthia Hudson-Vitale
Pennsylvania State University

Natalie Meyers
University of Notre Dame

Jennifer Muilenburg
University of Washington

Maria Praetzellis
California Digital Library

Kacy Redd
Association of Public and Land-grant Universities

Judy Ruttenberg
Association of Research Libraries

Katie Steen
Association of American Universities

Additional report and conference contributors:

Joel Cutcher-Gershenfeld
Brandeis University

Maria Gould
California Digital Library

This conference was funded by the National Science Foundation under Grant No. 1945938.

Suggested citation
Chodacki, John, Cynthia Hudson-Vitale, Natalie Meyers, Jennifer Muilenburg, Maria Praetzellis, Kacy Redd, Judy Ruttenberg, Katie Steen, Joel Cutcher-Gershenfeld, and Maria Gould. *Implementing Effective Data Practices: Stakeholder Recommendations for Collaborative Research Support*. Washington, DC: Association of Research Libraries, September 2020. https://doi.org/10.29242/report.effectivedatapractices2020

Executive Summary

In December 2019 the National Science Foundation (NSF) sponsored an invitational conference on implementing effective data practices, convened by the Association of Research Libraries (ARL), the California Digital Library, the Association of American Universities (AAU), and the Association of Public and Land-grant Universities (APLU). Forty experts representing libraries, research offices, scientific communities, tool builders, and public and private funding agencies spent 1.5 days in a workshop environment designing guidelines for institutions to implement two specific data practices recommended by the NSF: (1) using persistent identifiers (PIDs) for data sets, and (2) creating machine-readable data management plans (DMPs). The conference agenda, as well as a participant list, are included as appendices to this report. The project team worked with Joel Cutcher-Gershenfeld, a professor in the Heller School for Social Policy and Management at Brandeis University, who led the workshop design and facilitated the event. Dr. Gershenfeld has extensive experience working with scientific societies in the FAIR data community, as well as the Campus Research Computing Consortium (CaRCC), and brought his academic and professional expertise in stakeholder alignment to the role of facilitator.

By gathering and synthesizing the valuable insights generated at the conference, the project team developed a set of recommendations for the broad adoption and implementation of NSF's recommended data practices within research institutions. The intent is to contribute the recommendations to the AAU-APLU Institutional Guide to Accelerating Public Access to Research Data (forthcoming, spring 2021). This report, however, also includes recommendations for stakeholder groups outside of research institutions—including publishers, tool builders, and professional associations—as well as considerations for funding agencies. By including recommendations for a wide range of stakeholder groups, the project team invites all of them to pursue how collaboration with others on the implementation of PIDs and machine-actionable DMPs (maDMPs) can advance public access to research for the benefit of the entire research enterprise. At the same time, the report sections are designed for conversation within stakeholder groups, so they can determine their unique contributions and leverage points in the research process.

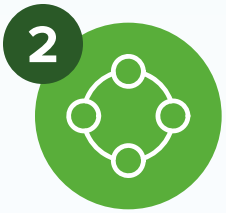
The recommendations and considerations in this report were circulated widely on social media for community review, and the project team held six consultative virtual focus groups with key stakeholders who were either invited to or participated in the conference.

Five key takeaways from this conference, validated in the web sessions and reflected in this report, include:

Center the researcher

Machine-actionable data management plans can serve as communication and collaboration vehicles for multiple units across an institution to form a more coherent research support environment. Active DMPs can also be core organizing tools within research labs, in support of good data management practices that drive good science. Tools, education, and services need to be built around data management practices in a way that accommodates the scholarly workflow, and not the other way around.

Researchers in the conference noted that in an ideal environment, there would not be a one-to-one relationship between grants and DMPs. Research projects extend beyond the life and boundary of individual funded projects. A researcher might have a data management plan for a project that spans several grants or funders.

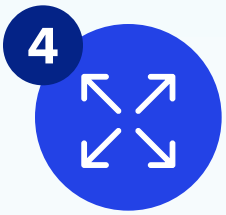


Create closer integration of library and scientific communities

Institutional offices of research, research computing, and academic and research libraries serve all disciplines within an institution. The conference focused on the need for greater alignment between disciplinary specialists (researchers and domain repository managers) and the research library community, which is committed to both data curation and stewardship across the research life cycle. Recommendations in this report also encourage more communication between library repositories and domain repositories, particularly at the point at which DMPs are finalized, and then when stewardship responsibility transfers from the researcher to the repository and identifiers are assigned in the process of data curation.

Open PID infrastructure is a core community asset

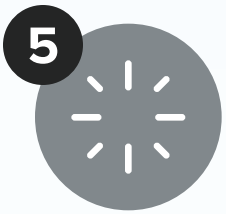
Persistent identifiers for people, organizations, and data and other outputs (instruments, code, and more) are essential to interlinking research across disciplines and domains. Some identifiers are domain-specific, while others—the ones recommended in this report—are universally recognized as valuable across the scholarly and research enterprise. Organizations that sustain identifier registries are essential pieces of scholarly infrastructure, and beyond adoption and use of PIDs, these organizations need the support of the research community. The research community is also best served by open licensing of metadata that enables interoperability across systems. Libraries, IT professionals, and research offices that develop or purchase research support systems can help accelerate the adoption of PIDs by requiring that these systems be designed to integrate with identifier registries, and by advocating for open metadata and open code.



Unbundle the DMP

While the value proposition for active or machine-readable DMPs is well demonstrated, there was a sense at the conference that the DMP as currently understood may be overloaded with too many expectations—that it would simultaneously be a tool within the lab, among campus resource units, and with repositories and funding agencies. Conference participants spent a lot of time in breakout groups working through the workflow in the pre-award and post-award context and discussing the advantages of unbundling the DMP.

Unbundling these distinct DMP functions, the groups observed, could help advance collaboration both within and among different stakeholder groups. For example, research institutions have repeatedly appealed to funding agencies to harmonize the requirements of a DMP for the purposes of building services to support compliance. Agencies, in turn, have articulated the need for program- and domain-specific elements in data management. There may be a core set of DMP elements that can be harmonized for the purposes of compliance. Within the institution, unbundling commitments to scientific practice in the lab from business processes like planning for computing and storage would enable a phased sequence of conversations, from pre- to post-award. The point at which the award is made to the institution might be when there is the greatest incentive to convene the multiple units affected by the DMP and finalize the internal budget allocation.



PIDs will unlock discovery

Finally, PIDs are part of the infrastructure necessary to connect metadata across systems and assemble diverse data to answer new questions. Highlighting tangible examples of data integration across repositories through PIDs, as well as tools and services designed to use them even without significant knowledge of them, will be key to their adoption by researchers.

The following report provides a summary of guidance and best practices for research stakeholders on how to coordinate and commit to the adoption and incorporation of PIDs and machine-actionable DMPs in service of realizing the goals and recommendations outlined in the May 2019 NSF Dear Colleague Letter (DCL)[1]. This vision requires active coordination among stakeholders—the core insight of both the National Academies' *Open Science by Design*[2] report, and the basis for new thinking generated at the conference.

[1] "Dear Colleague Letter: Effective Practices for Data," National Science Foundation, May 20, 2019, https://www.nsf.gov/pubs/2019/nsf19069/nsf19069.jsp.

[2] National Academies of Sciences, Engineering, and Medicine, Open Science by Design: Realizing a Vision for 21st Century Research (Washington, DC: The National Academies Press, 2018), https://doi.org/10.17226/25116.

Effective Data Practices Conference at a Glance

5 Questions for Conference Participants

1. What barriers remain to implement the widely recognized good practices in the NSF DCL?
2. What kinds of model workflows might address those barriers, while minimizing faculty burden?
3. What does the implementation of the NSF DCL mean for institutional data governance?
4. Which findings should be brought back to policymakers, funding agencies, and institutions so they can engage in a discussion about the next steps?
5. What are the recommendations for effective practices for grants offices, including guidance to their researchers?

5 Takeaways from the Conference

1. Center the researcher
2. Create closer integration of library and scientific communities
3. Open PID infrastructure is a core community asset
4. Unbundle the DMP
5. PIDs will unlock discovery

5 Core PIDs to Power Findability

1. Digital object identifiers (DOIs)
2. Open Researcher and Contributor (ORCID) iDs
3. Research Organization Registry (ROR) IDs
4. Crossref Funder Registry IDs
5. Crossref Grant IDs

5 Core Recommendations to Stakeholders

1. Design tools and services to support the use of PIDs.
2. Incorporate PIDs into policies.
3. Invest in infrastructure and initiatives that support the use of PIDs and maDMPs.
4. Minimize the burden on researchers.
5. Understand that PIDs and maDMPs are not static, and support the effective use of PIDs throughout the entire course of the research life cycle.

5 Core Incentives to Adopt These Data Practices

1. Get credit for sharing research.
2. Save time.
3. Identify key collaboration partners.
4. Facilitate data reuse.
5. Mitigate risk.

Introduction

In 2018, the National Academies of Science, Engineering, and Medicine (NASEM) published a consensus report vision for 21st-century research[3] that prompted many institutions, organizations, and faculty to assess their current practices and infrastructure to support a more open research ecosystem. To fully realize the vision for open science and scholarship, according to the NAS report, stakeholders need to work together to adopt key infrastructure, standards, and practices necessary to facilitate responsible research and data practices. In May 2019, the National Science Foundation issued a Dear Colleague Letter promoting two effective data practices: issuing persistent identifiers for data sets and creating machine-readable data management plans.[4]

In December 2019, the library community, represented by the Association of Research Libraries and the California Digital Library, in partnership with the Association of American Universities and the Association of Public and Land-grant Universities, and with funding provided by the NSF,[5] convened a small conference to discuss the current state of PIDs and machine-readable DMPs. The goal of the Implementing Effective Data Practices conference was to frame the suggested best practices in the DCL within the larger stated commitment by AAU and APLU institutions to expand public access to research data and to advance open science and scholarship within the framework of the NASEM report.[6] At this conference, approximately 40 experts came together with the aim to identify and determine:

1. What barriers remain to implement the widely recognized good practices in the NSF DCL?
2. What kinds of model workflows might address those barriers, while minimizing faculty burden?
3. What does the implementation of the NSF DCL mean for institutional data governance (e.g., sharing DMPs across campus units, between institutions, and publicly)?
4. Which findings should be brought back to policymakers (specifically NSF), funding agencies, and institutions so they can engage in a discussion about the next steps?
5. What are the recommendations for effective practices for grants offices, including guidance to their researchers?

[3] National Academies of Sciences, Engineering, and Medicine, *Open Science by Design.*

[4] "Dear Colleague Letter: Effective Practices for Data."

[5] "Award Abstract #1945938, Implementing Effective Data Practices," National Science Foundation, August 29, 2019, https://nsf.gov/awardsearch/showAward?AWD_ID=1945938&HistoricalAwards=false.

[6] "Implementing Effective Data Practices," Association of Research Libraries, accessed August 20, 2020, https://www.arl.org/implementing-effective-data-practices/.

Attendees of the workshop-style conference included US federal agency representatives, private funding organizations, research IT professionals, vice chancellors for research, professional societies, domain repository managers, tool builders, data librarians, and several active researchers. The conference provided an opportunity for fresh thinking on how the scientific community and the library community might partner for better data management, better stewardship, and better compliance with funders' requirements, without increasing researchers' administrative burden. The attendees pushed beyond the DCL in two key dimensions: 1) they promoted a core set of PIDs, not just for data sets, and 2) they defined the objective as machine-actionable DMPs, rather than machine-readable.

Research practices and culture are complex and influenced by disciplines as well as public and institutional policy and practice. Many conference participants raised the issue of incentives—either positive or negative—that are perceived to be the drivers of change. Since funding agency requirements often start off as recommendations, it behooves the research support community to build tools, services, and policies in alignment with them even before they need to for compliance. The recommendations in this report are offered in that spirit, and each section reflects the influence of that particular stakeholder group. On the first of the conference, a panel on Incentives and Policy generated a discussion about how PIDs and maDMPs contribute to a culture of credit for data as a research output—something that depends on a robust "socio-technical[7]" infrastructure of identifiers, curation and preservation repositories, discoverability, and citation.

Finally, a note on the present reality of scientific research, higher education, and the profound impact of COVID-19. Even pre-pandemic, many of the recommendations included here would be aspirational for either research domains or individual institutions. On one hand, the pandemic has presented a global case for open science practices, particularly rapid sharing of data and analysis. On the other, the abrupt closing of research labs in March 2020, along with the financial impact on research institutions, is dire. However, the coronavirus pandemic, by necessitating remote and distributed research teams and lab closures, also demonstrated the value of data management, access, and reusability under those conditions.[8] From that perspective, the recommendations advanced in this report, had they been widespread practices at the start of the pandemic, would have helped with institutional research continuity in the spring and summer of 2020.

[7] Carl Kesselman (panel presentation, "Implementing Effective Data Practices: A Conference on Collaborative Research Support," Washington, DC, December 11, 2019).

[8] *What Happens to Continuity and Future of the Research Enterprise? Report of a CNI Executive Roundtable Series Held April 2020* (Coalition for Networked Information, May 2020), https://www.cni.org/go/what-happens-to-continuity-and-future-of-research.

Defining PIDs and DMPs

Persistent Identifiers are essential components of the scholarly infrastructure. PIDs offer a way to both identify and describe researchers, institutions, equipment, protocols, and more. Common PIDs in scholarly communications include digital object identifiers (DOIs, used for journal articles and data sets), Open Researcher and Contributor (ORCID) iDs (used for researchers), and Research Organization Registry (ROR) IDs (used for research institutions). Each PID scheme has a defined metadata standard describing its objects, and that information can then be used by machines and people. The power of PIDs comes from the metadata that they capture and from how they can be connected to each other to advance the findability and interoperability of digital assets.[9]

While some PIDs are managed by nonprofit or community groups and made available under open licenses (for example, DOIs from Crossref and DataCite, ORCID iDs, and ROR IDs), others are run as commercial services with closed licenses that could be monetized and come at significant cost for the research community if widely adopted. In contrast, openly licensed PIDs interoperate and reference each other within research discovery and management systems, linking descriptive information to other objects, without access restrictions. As a result, simply assigning a PID for an object can open up a large set of information by connecting to other PIDs and other systems, in essence collaboratively completing a human- and machine-readable map of scholarly research.[10] The more comprehensively PIDs are used in declaring information about research, the more information there is available to leverage.

Data management plans (DMPs) offer a way for researchers to articulate how they will handle the data, software, and other research objects that are generated as part of their research project. Key types of information found in DMPs include: protocols for how researchers will conduct their research; equipment or reagents that will be used; descriptions of the size and type of data or software that will be generated; procedures for how the data will be packaged, preserved, secured, and shared internally with a team and externally to the public and other researchers; and the repository that will be used for long term access. DMPs are routinely required by funders as part of a grant proposal.

In the past, DMPs have been written as text documents and submitted as PDFs as part of a grant application. However, in recent years there has been a movement for the information included in a DMP to be tagged with PIDs and shared more broadly so that it is machine-readable as well as human-readable. Funders have also asked for updates to a DMP as a research project evolves and practices relating to research data are modified. Campus communities supporting research are looking for ways to help researchers comply with such requirements. The end goal of a machine-actionable DMP is to facilitate a research data management (RDM) system that allows data and information about research to be communicated and shared across stakeholders, linking metadata, repositories, and institutions, and allowing for notifications and verification, real-time reporting, and automated compliance. If leveraged appropriately, maDMPs have the ability to lessen the administrative burden on researchers and grant administrators.

[9] Findable, Accessible, Interoperable, Reusable (FAIR) data principles were displayed on large posters in the conference room.

[10] For more information about PIDs and work being done to link identifiers, follow the discussions on the PID Forum (https://www.pidforum.org/c/pid-graph/) and refer to the EU-funded FREYA Project (https://www.project-freya.eu/).

Core Principles and Recommendations

The recommendations of this report converge, broadly, on two core principles: openness and collaboration.

- **Openness**—The act of assigning persistent IDs for research data and creating machine-actionable DMPs will not advance open scholarship on its own. In order to be effective, openly licensed PIDs and maDMPs need to be shared and connected to outside systems, researcher communities, and to the larger knowledge graph.[11] This in turn, facilitates faster and broader dissemination of scientific knowledge and provides expanded opportunities for advancing scientific discovery.
- **Collaboration**—While machine-actionable DMPs and PIDs provide the technical means for data and information sharing, this is complemented by active collaboration between stakeholders. The stakeholder groups described in this report together drive the policy development, educational and training resources, and support and liaison relationships that contribute to the broader culture change necessary to realize the advantages of an open data ecosystem.

In the sections that follow, we provide recommendations for how individual stakeholder groups can adopt and incorporate PIDs and maDMPs into their work and communities. The following is a set of core PIDs that are fundamental and foundational to an open data ecosystem. Using these PIDs will ensure that basic metadata about research is standardized, networked, and discoverable in scholarly infrastructure.

- **Digital object identifiers (DOIs)** to identify research data, as well as publications and other outputs
- **Open Researcher and Contributor (ORCID) iDs** to identify researchers
- **Research Organization Registry (ROR) IDs** to identify research organization affiliations
- **Crossref Funder Registry IDs** to identifier research funders
- **Crossref Grant IDs** to identify grants and other types of research awards

[11] A knowledge graph represents a collection of interlinked resources or assets and articulates the relationship between those assets.

While the sections that follow are customized for each stakeholder group, there are continuities through and similarities across all of them. Below, we summarize and highlight these core principles and recommendations:

1. **Design tools and services to support the use of PIDs.** All research-related workflows and systems should be designed to enable the collection of PIDs, storage of PID metadata, and connections to PIDs in other systems.
2. **Incorporate PIDs into policies.** Research-related policies—for institutions, funders, publishers, professional societies, and other stakeholders—should require the use of PIDs as much as possible.
3. **Invest in infrastructure and initiatives that support the use of PIDs and maDMPs.** Actions include joining member organizations that promote open scholarly infrastructure, such as Crossref, DataCite, and ORCID; sponsoring institutional memberships with data repositories that follow best practices for FAIR data; supporting community-led initiatives such as the Research Organization Registry and EZDMP; and becoming an institutional member of the DMPTool.
4. **Minimize the burden on researchers.** Make it easy and seamless for researchers to use PIDs by designing workflows and systems to collect them automatically and by supporting institutional memberships for PID services or data repositories. For both PIDs and maDMPs, build training and guidance into library services and research consultations.
5. **Understand that PIDs and maDMPs are not static, and support the effective use of PIDs throughout the entire course of the research life cycle.** DMPs must be updated as research projects evolve. PID metadata must be updated over time to support long-term persistence and discoverability. Effective data practices require taking long-term responsibility.

Core Incentives for the Advancement of Research

For anyone involved in research data management, PIDs can make research tracking and associated analysis work easier, more efficient, and less prone to errors and/or duplication of effort. Relying on a core set of PIDs can also reduce costs by minimizing tedious data cleanup. For researchers themselves, the conference surfaced the following core incentives for adoption of PIDs and maDMPs:

1. **Get credit for sharing research.** The use of PIDs will allow researchers to track the citation of all of their research assets and get credit for sharing all of their research outputs.
2. **Save time.** By investing in PIDs and maDMPs, researchers will save valuable time in establishing connections for research support during the research period and at reporting time at the grant's end.
3. **Identify key collaboration partners.** With the broad adoption of PIDs, researchers could see researcher networks, including global networks—"who is working with whom?"—for funded and unfunded projects.
4. **Facilitate data reuse.** Leveraging PIDs enhances the reusability of data by precisely identifying core resources, instrumentation, and more. While reusability is always a priority, it is especially urgent during times of limited access to labs and travel, as has been the case during the COVID-19 pandemic.
5. **Mitigate risk.** Machine-actionable DMPs will establish notifications of administrative support and services, for example around personally identifiable information—improving the integrity of research.

The sections that follow provide recommendations on the implementation of PIDs and machine-actionable DMPs for individual stakeholder groups: researchers, academic and research libraries, research offices, institutional IT departments, scholarly publishers, tool builders, and professional associations and societies. The final section highlights key considerations for funding agencies.

Researchers

Researchers have many demands on their time and attention. Most researchers understand the general concepts and importance of identifiers, however many are unsure of the most applicable and useful identifiers in their disciplines. Additionally, although researchers have been required to produce DMPs for many years now, the DMP is generally seen as an administrative hurdle of the grant application process with little added value or utility to the overall project. The general assumption is that DMPs are a necessary step in the process, but researchers do not expect they will be substantively reviewed or used for compliance monitoring or resource allocation.

Much of the larger infrastructure and administration needed to transform research data management practices presented in the NSF DCL on Effective Practices for Data are beyond the purview of the individual researcher. The recommendations in this report have been designed to make it easier for researchers to follow best practices without requiring a detailed understanding of the intricacies of PIDs or maDMPs or overhauling their everyday work. With this in mind, there are several straightforward and relatively simple practices that researchers should do in the short term in order to follow the best practices as laid out in the DCL.

Incentives for adoption

To ensure the broad adoption of the recommendations and improvements to data management practices as laid out in the NSF DCL, the value and impact of these practices need to be clear. There also needs to be a variety of incentives and established infrastructure to insert these practices into the existing workflows of researchers. For example, a key goal of the maDMP workflow is to reduce the burden on researchers by generating automated updates to a plan and facilitating seamless integration with systems and groups that support research.

Key recommendations

Identifiers

- Obtain an ORCID iD and use this identifier wherever possible to get credit for work and improve the discoverability of research.

Data management plans

- Make data management a core component of all research activities. A DMP should be established at the beginning of all projects and used to define methods for collecting, managing, and sharing all research outputs (data, code, models, samples, and other research outputs).
- Use existing tools (such as the DMPTool or EZDMP) when creating a DMP in order to generate machine-actionable DMPs.
- Share machine-actionable versions of the DMP with the researcher's home institution as well as the intended data repository and any other data curation or preservation departments or staff.
- During the course of an award, bring any substantive changes to the DMP to the attention of the affiliated grant officer.

Data deposit and publication

- Publish all data sets underlying published works under a CC-0 or CC-BY license.
- Consult with the library regarding appropriate curation and preservation actions.
- Publish data sets in a data repository that will assign a persistent ID. Work with the university library or data repository (i.e., Dryad, Zenodo, etc.) or other research support staff to facilitate this process.
- In journal articles, cite all relevant data sets.

Academic and Research Libraries

Libraries are uniquely situated to play a central role in ensuring effective data management practices through instruction, outreach, technical systems development, and collaborations across campus. Additionally, libraries serve as stewards of research outputs by providing consultation and guidance for researchers to ensure the long-term discoverability and accessibility of these outputs in campus-based or domain repositories, as well as measurement and visibility of their impact.

While libraries have the expertise, strategic position, and professional mandate to play a key role in providing effective data management, some challenges exist that can limit their ability to achieve this goal. One key obstacle is that faculty and researchers are often unaware of the role libraries can play in supporting active research data management. The following set of recommendations can help guide the role of libraries in research data management practices and build upon libraries' critical role in making research data discoverable and reusable.

Incentives for adoption

PIDs and maDMPs are automated, efficient tools and scalable approaches to advance the mission of research libraries to provide effective stewardship of research outputs. Deploying these tools and approaches in concert with, and for the benefit of, researchers will result in optimizing the downstream stewardship and preservation of research outputs to magnify their impact.

Key recommendations

Integrate PIDs into existing research workflows, infrastructure, and policies

- Facilitate institutional membership in ORCID, DataCite, Crossref or other member-based identifier infrastructures providers. This can be done in partnership with the university research office or other centralized campus research support entities that are also invested in campus-based research support.
- Ensure that core PIDs, such as ORCIDs, are included for all deposits in institutional repositories (IRs), either by requiring their use at the point of submission or by providing metadata augmentation services post-deposit. IRs should provide the support and guidance to make PID usage seamless and easy for depositors.
- Offer the ability to assign DOIs for all data sets deposited in IRs, and make sure DOI metadata includes information about related works so that data sets can be linked to articles and other output.
- Work with campus colleagues to ensure that vended research management and support systems incorporate open PIDs.

Provide consultation and instruction

- Introduce PID support; provide tools, training, and advocacy; and help establish practices and technologies that extend the role of PIDs as critical research infrastructure. Start with a core set of PIDs (as mentioned above), including, for example, ORCID iDs, ROR IDs, Funder Registry IDs, grant IDs, and DOIs for data sets.
- After use of this core set of PIDs is established as a best practice for researchers, work with researchers to incorporate other disciplinary or specialized PIDs into their research activities.
- Encourage researchers producing DMPs to use platforms such as the DMPTool or EZDMP.
- Recommend that, at a minimum, DMPs should include the use of identifiers for people, institutions, and funders.
- Encourage the use of PIDs for ongoing project work and outcomes, such as publications, data sets, protocols, and other deliverables and findings.
- Customize institution-specific guidance for DMPs within the DMPTool (or other platform) to highlight library resources and best practices so that researchers are informed of services available to them in supporting their data throughout the research process.

Collaborate and advocate

- Work with the university research IT department and/or office of research to produce and disseminate clear guidance on available campus-specific resources for research data support that demonstrate the ways these resources increase research value and impact. Consistent and clear guidance will mitigate confusion across research projects and prioritize and showcase the role of the library in data stewardship.
- Pursue professional development opportunities to learn more about identifiers and RDM best practices.

Research Offices

Campus research offices advance the research based at their universities, reduce administrative burden on faculty, and ensure research projects comply with standards of ethics and integrity, as well as with funder mandates. While compliance includes activities such as meeting funder data availability policies and outputs as promised in DMPs, many research offices are struggling to develop the infrastructure required for meeting these requirements while simultaneously reducing the overall cost of research support systems.

Research offices can play an important role both in advancing the research outputs of their institutions and ensuring that their grants meet the requirements of frequently changing funder requirements by encouraging the use of machine-actionable DMPs combined with the tracking power of persistent identifiers. They can help researchers manage and track the outputs of their work by offering instruction and guidance on creating and updating DMPs and using interlinked PIDs.

Incentives for adoption

Research offices are in the valuable position of working collaboratively to support and ensure the integrity of the research on their campuses. Supporting machine-actionable DMPs and broadly adopting PIDs will help research offices automatically deploy services and project future resource allocation.

The following set of recommendations can help guide research offices in developing practices and policies to ensure that research projects are in compliance with funding mandates.

Key recommendations

Pre-award

- Require ORCIDs for all PIs, co-PIs, and collaborators included in a grant submission.
- Develop and promote institutional best practices for data management and data sharing and provide guidance documents, instructional resources, and examples of the essential elements of a good maDMP from successful awards.
- Direct researchers to campus service providers and tools such as the DMPTool that can help further develop maDMPs and assign PIDs to digital objects.

Upon award

- Instruct PIs to add award IDs to their ORCID profiles and/or to enable their ORCID profiles to be updated automatically.
- Produce a summary document with key identifiers to be used for tracking the award (ORCID iDs, ROR IDs, OpenFunder Registry IDs, grant IDs, and DOIs). Provide instruction for researchers on how and where to use these identifiers to facilitate tracking of the project, such as updating an ORCID profile or citing a data set.
- Review DMPs upon grant award with researchers and other relevant campus stakeholders. Review what they need to do to be in compliance with the DMP, paying particular attention to ensure all outputs can be tracked throughout the project. A maDMP will aid in this tracking.

Ongoing

- Work with other campus research support services and with researchers to keep DMPs up to date as the grant progresses.

Collaborations

- Form partnerships with various stakeholders (e.g., libraries, researchers, institutional review boards, offices of sponsored projects) to develop institutional data policies related to data management and sharing, as well as institutional expectations for use of DMPs and PIDs. Form partnerships with department chairs and faculty to assess the impact of good data-sharing practices and appropriately reward their adoption.
- Collaborate with the library to support PID and data repository services and memberships that provide open infrastructure in support of research and data publication, management, and sharing.
- Collaborate with the library in decision making regarding the tracking of research outputs for accessibility, reuse, impact tracking, and preservation.
- Establish institutional permissions for key campus offices (e.g., libraries, IT) to access the DMP, which can be streamlined by using maDMPs.

Institutional IT

Research IT and computing provide key services and support for the implementation of research at many institutions. Researchers often work with department, college, or central IT teams to establish critical data management needs, such as storage, backups, collaboration features, data transfers, and much more.

Incentives for adoption

The broad adoption of PIDs by the scholarly communications infrastructure has many benefits to institutional research computing and IT. By unbundling the DMP and breaking apart the administrative needs from the research needs, the flow of information within an institution can be improved. For example, institutional IT representatives could be alerted when a DMP and resulting research project are expecting large amounts of data to be created. Information from unbundled DMPs could be beneficial for projecting storage costs and estimating high-performance computing needs.

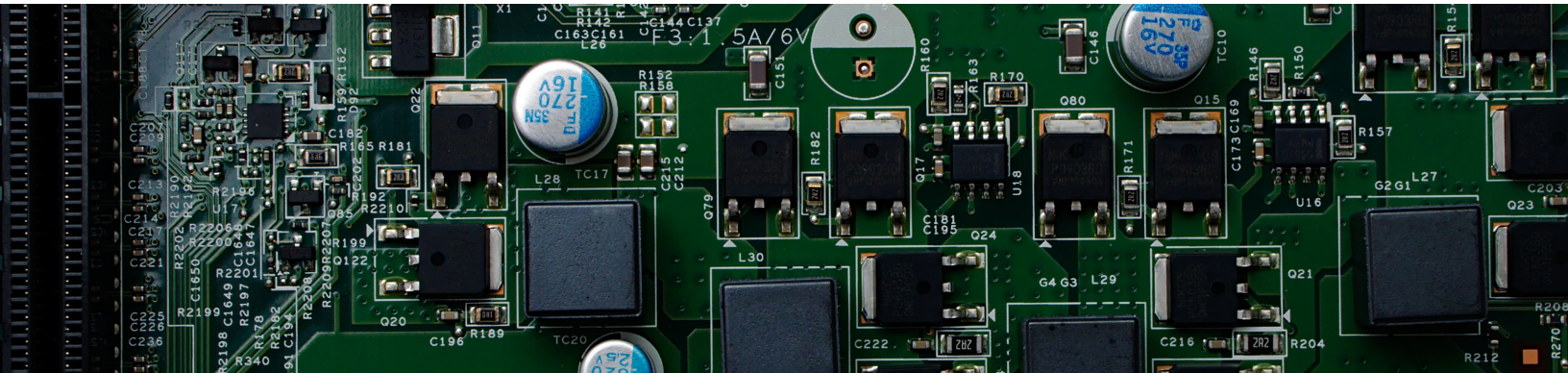
Key Recommendations

Collect and integrate PIDs into existing research workflows, equipment, infrastructure, and policies

- Work with campus colleagues to ensure that vended research management and support systems incorporate open PIDs.
- Create unique PIDs for research equipment to track use and project demand.

Collaborate and advocate

- Work with representatives from the libraries and research offices to build infrastructure to support the unbundled and machine-actionable DMP.
- Work with the university research IT department and/or office of research to produce and disseminate clear guidance on available campus-specific resources for research data support that demonstrate the ways these resources increase research value and impact.



Scholarly Publishers

Current academic publishing practices were born from print-based models, and those practices were baked into technical systems when publishing moved online. The shift to digital has opened up new opportunities for disseminating and connecting research. At the same time, many publishing workflows still struggle to transform from the legacy of print-centric frameworks.

Publishers are not a monolithic group. Regardless of type (commercial, nonprofit, university-based, or library-based), good publishing practices are part of good research practices, and publishers are important partners to engage in best practices around PIDs and DMPs.

PIDs collected in publishing workflows can be connected to each other in scholarly indexes and other downstream systems to network discrete pieces of data and information. Such networks can make connections between publications, preprints, data repositories, and funding systems, enabling key insights such as the relationship between the methods described in a DMP and those described in a related article, how research versions evolve over time, or how widely funders' open access mandates have been adopted.

As published articles are no longer the sole endpoint of a research endeavor, publishers need to retool and rethink how their world intersects with and communicates with other systems like data repositories and funder platforms. Inclusive of legacy systems and new publishing platforms and practices, the publishing industry as a whole must build PIDs and connections to DMPs into their processes to ensure that they and other stakeholders can leverage the networked knowledge that these connections enable.

Incentives for adoption

PIDs can play a central role in the publishing landscape by identifying the people, products, processes, and places involved in research. This includes identifiers not only for published articles, but also for other research-related outputs such as data, DMPs, and preprints. This also includes identifiers for the methods, materials, protocols, and facilities employed as part of the research. The use of these identifiers in research workflows and publications contributes to greater discoverability, access, citability, and preservation of knowledge, and supports the growing emphasis on openness, transparency, data sharing, and reproducibility across scholarly communities.

Key recommendations

Collect or provide PIDs wherever possible

- Join Crossref and assign Crossref DOIs for all published content.
- Establish editorial policies that:
  - Use PIDs instead of free text for funding information: OpenFunder Registry ID or ROR ID for funding organizations and grant IDs for grants and awards.
  - Require ORCID iDs for corresponding authors and coauthors.
  - Implement/adopt ROR IDs for affiliations of authors, editors, and reviewers.
- Establish author guidelines that:
  - Require authors to provide a DOI or other PID or data availability statement for underlying data associated with an article.
  - Require authors to cite all data directly referenced in articles and include DOIs or other PIDs in reference lists, data availability statements, and methods sections.
  - Encourage authors to obtain identifiers, as appropriate, for materials and methods, such as reagents, physical samples, code, etc.; to document their processes in platforms such as protocols.io; and to reference these identifiers in the article narrative. Relevant identifiers could include Research Resource Identifiers (RRIDs) to promote research resource identification, discovery, and reuse; International Geo Sample Numbers (IGSNs), DOIs; and more.

Support rich metadata and robust metadata connections wherever possible

- Allow authors to provide grant applications and/or published DMPs as related items with article submissions.
- Include metadata for related identifiers in DOI deposits so that articles can be linked to underlying data, grants, DMPs, and any other outputs.

Support downstream use and reuse

- Assign open licenses (CC0) for data and article metadata to maximize access by machines and reuse.
- Send citations (including properly indexed data citations) to Crossref and inform Crossref that the citations should be openly licensed.

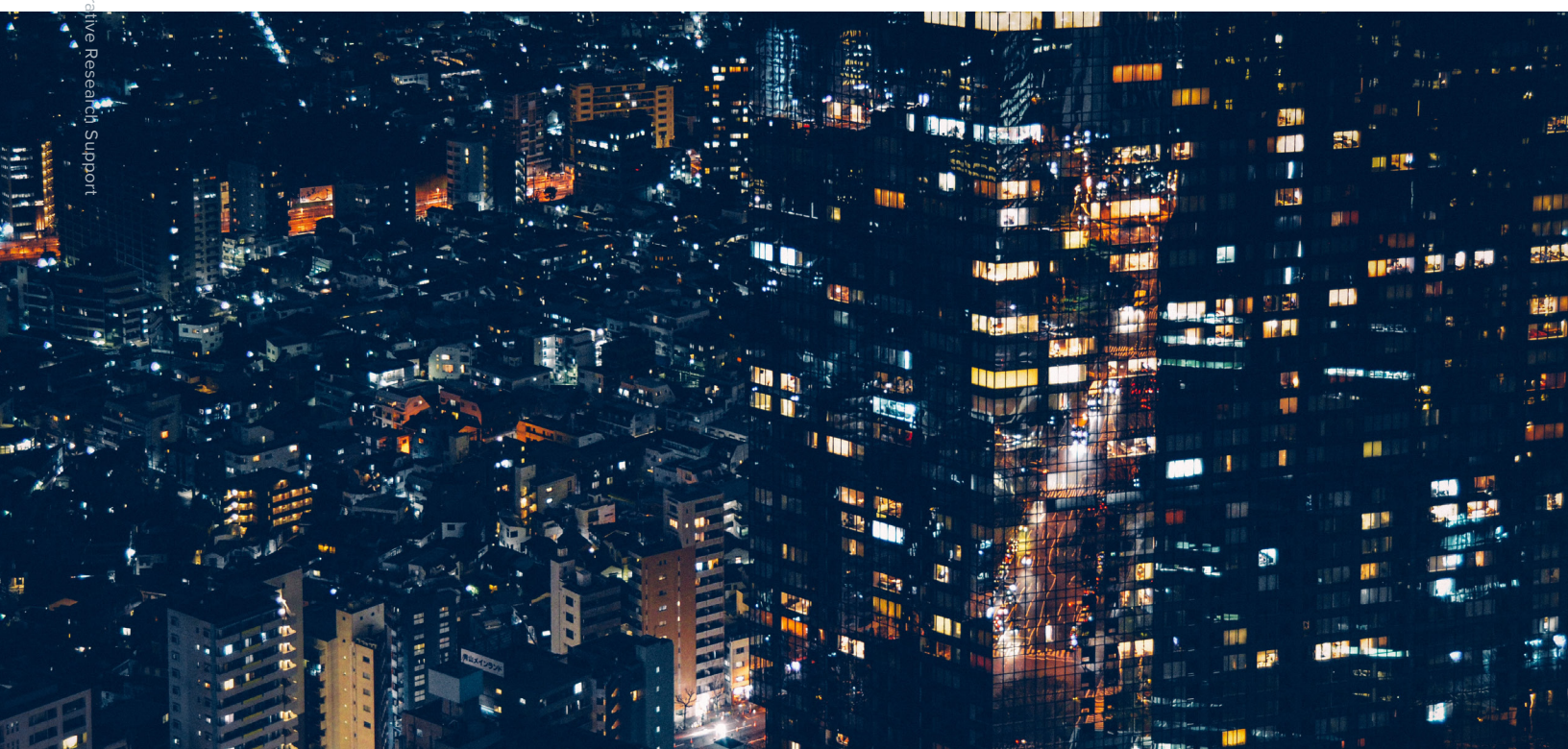
Tool Builders

Today's research landscape is marked by a plethora of tools and platforms. These offerings cover a wide swath of purposes, approaches, business models, and target communities. While this diversity can be an asset and is often necessary to meet the requirements of specific domains or user groups, the proliferation of options can also be overwhelming and/or confusing. The result can be platform fatigue, unclear or misinformed decision-making, or a misalignment of goals between tools and their intended users.

PID and DMP-related systems and services represent one layer of this landscape. Within the scholarly community, researchers and those who advise and support them must navigate an array of guidance on options and best practices when it comes to working with PIDs and DMPs. Those who build these tools have an opportunity—and, we would argue, an obligation—to facilitate best practices, minimize friction and fatigue, and address community needs.

It should be seamless and simple for tool users to work with PIDs, whether obtaining or entering an ORCID iD as part of an article submission workflow, providing an organizational affiliation or funder name while the corresponding ROR ID or Crossref funder ID is captured in the back-end system, or generating a DOI for a research output. However, the mere existence of PIDs in these systems is not sufficient on its own. Tools must submit this information to the relevant registration agencies or scholarly indexes to facilitate the tracking, discovery, and networkability of research through identifier infrastructure. The most prominent and powerful channels for doing this are through assertions against ORCID, Crossref, and DataCite, which support rich metadata and metadata connections that can especially enable the actionability and long-term tracking of DMPs over time.

Funders, publishers, librarians, researchers, and other stakeholders all rely on tools and platforms to publish, discover, and manage research outputs. Tool builders are uniquely positioned to enable broad and effective use of PIDs and to facilitate networked and dynamic research infrastructure to advance the goals of the 2019 NSF Dear Colleague Letter.



Incentives for adoption

In implementing the recommendations of this report, tool builders will be able to more easily support the integrity of institutional research and the dissemination of research throughout the scholarly communications ecosystem. By adopting PIDs, tool builders can help the knowledge graph of research become more of a reality.

Key recommendations

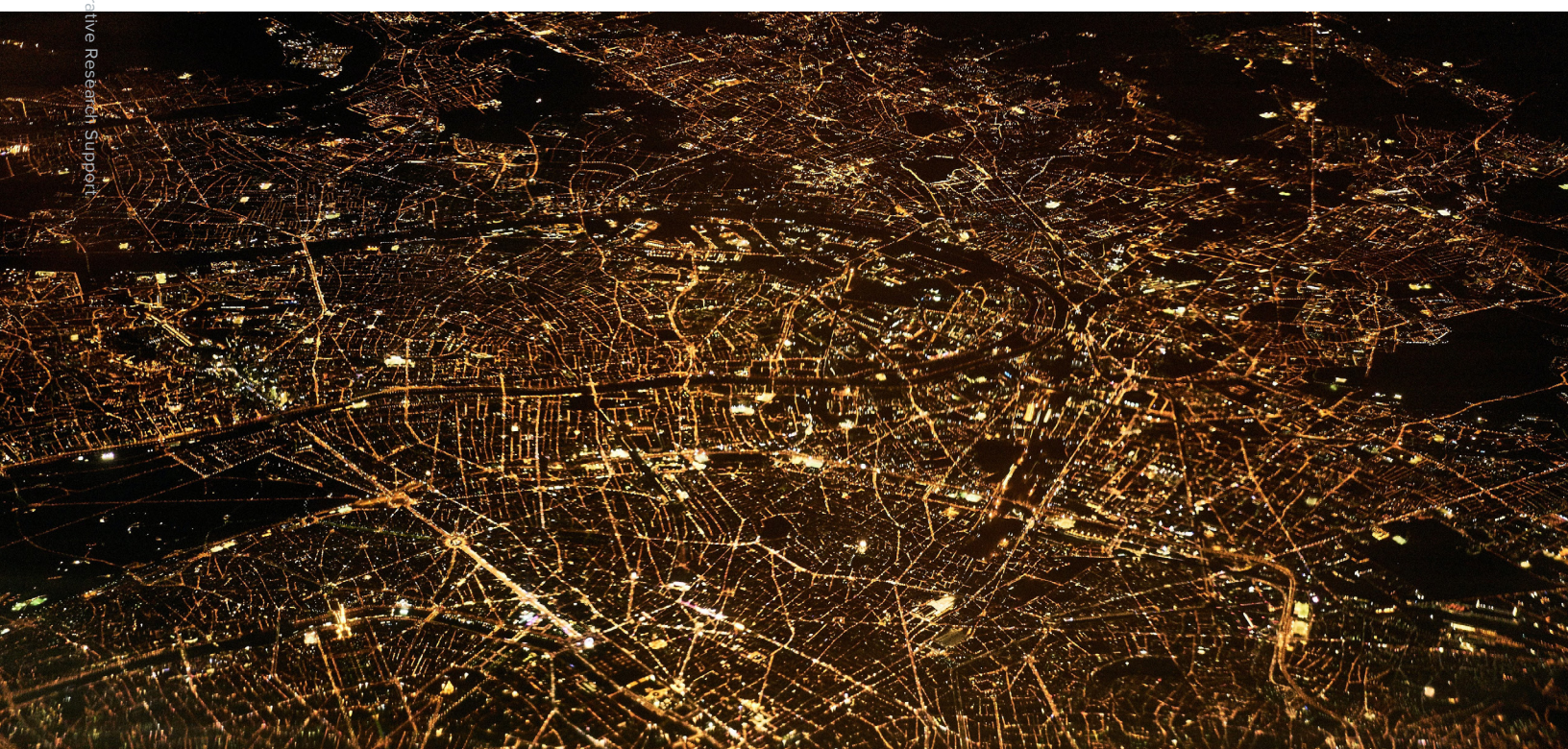
- Integrate PIDs that are openly licensed and free of reuse constraints.
- Automate and streamline PID aggregation and connection wherever possible and within the systems that researchers are already using.
- Design tools that can be integrated with those that researchers use in everyday work, such as GitHub and Jupyter Notebooks.
- Facilitate the exchange of information between research stakeholders by supporting open and secure APIs.
- Aggregate and connect PIDs into relevant registration agencies and scholarly infrastructure.

Professional Associations and Societies

Professional associations and societies convene, support, and advocate for scholars. They survey their members, collect data, develop and promote best practices, provide infrastructure for working groups, facilitate training and skills development, and advocate on behalf of their communities. They also enable the kind of informal networking that can accelerate adoption of effective data practices and provide ongoing knowledge sharing and feedback to those creating tools, standards, and policies.

Research data are as varied and diverse as the members of these associations and societies. Currently many disciplines lack clear definitions of the types of data being generated in the course of research and clear best practices for managing this data. Professional societies, with their cross-institutional engagement and informal knowledge-sharing mechanisms, and their practice- and implementation-oriented work, can play a key role in developing best practices in areas of managing research data. Their communities can provide discipline-specific guidance supporting PIDs for research data and help inform and promote best practices for machine-actionable DMPs. The diversity within these groups of professionals, as well as their field-tested approaches, make them ideal incubators for advancing PID and maDMP use across the community.

Many disciplinary and professional societies also act as publishers for their field by publishing society journals, conference proceedings, research reports, and more. Examples include the American Geophysical Union (AGU), which has played a leadership role in advancing FAIR data standards, and the Association of American Medical Colleges (AAMC), which has led the Credit for Data Sharing initiative. Disciplinary societies and associations can enact practices that embed their principles and missions for scholarly communication into their publishing practices, from peer review to their vendor-provided



platforms.[12] Societies that are involved in scholarly publishing have a unique and important role to play in shaping the publishing standards for their communities. Societies can implement and promote publishing best practices regarding the usage of PIDs in publications or citations in order to facilitate greater sharing of research data and research outcomes and impact.

Incentives for adoption

Integrating PIDs more completely into the scholarly communications infrastructure would allow professional associations and societies to see the impact of their memberships' contributions to the scholarly record.

The following set of recommendations can help guide professional associations and societies in facilitating and promoting the adoption of domain-specific standards and best practices for managing research data within their networks and members.

Key recommendations

Standards

- Build clear PID recommendations based on the needs and requirements informed by broad input of members.
- Develop discipline-specific components of a maDMP informed by broad input of members and their funders.
- Drive adoption of best practices in data sharing, formats, metadata standards, tools, and infrastructure.

Training and outreach

- Share exemplar DMPs and case studies that use PIDs for research data in order to demonstrate the benefits of effective data management practices.
- Produce written guidance for researchers on how to comply with funder requirements related to data management, with guidance tailored to funders and requirements specific to their disciplinary domain or area of research.
- Develop educational initiatives to raise awareness of research data management best practices and promote the use of existing standards and tools, such as data repositories, DMPTool, etc. Build this training into regular professional development offerings through a variety of instructional mechanisms.

[12] For instance, the Enabling FAIR Data Project from AGU was developed with funding from the Laura and John Arnold Foundation, resulting in a statement that guarantees specific commitments for multiple partners in the research life cycle. See "Commitment Statement in the Earth, Space, and Environmental Sciences," Enabling FAIR Data Project, accessed August 20, 2020, http://www.copdess.org/enabling-fair-data-project/commitment-to-enabling-fair-data-in-the-earth-space-and-environmental-sciences/.

Collaboration

- Partner with university libraries, institutional IT, and research offices to produce and share consistent guidance for researchers specific to research data management, PIDs, and DMPs.
- Create opportunities for researchers to discuss discipline-specific data sharing and management approaches, learn from illustrative examples, and promote successes.
- Share and cross-promote training for researchers on these practices with other associations and societies, campus peers, and other stakeholders.

Publishing practices

- In publications, include a statement of where and how the data are available. These statements should clearly state the PID of the data and describe how the data underlying the findings of the article can be found, accessed, and used.
- Require publications to include PIDs for all publicly published research data cited in the work.
- Implement the recommendations found in this report's Publishers section.

Key Considerations for Funding Agencies

Researchers would be well-served by clear and consistent requirements across funders for basic best practices for research data management.

Community-based organizations and initiatives can inform the standards, tools, software, and other infrastructure needed to develop maDMPs and PIDs. Strategic plans can help funders coordinate community-based discussions, invest in infrastructure that will support discoverability of research data, and provide incentives for the research community to adhere to and help inform these practices.

The following set of considerations can help guide funders in developing the policies and investments needed to ensure the discoverability and accessibility of the research data resulting from their funding.

Incentives for adoption

While tracking and sharing the outputs from funded projects and estimating the return on investment are concerns for funding organizations, gaps currently exist between broad policies supporting data availability and open access, and the clear guidance and compliance checking required for researcher adherence to such policies. By instructing researchers on exactly what information is required, and what new connections will be made available by the usage of persistent identifiers and machine-actionable DMPs, funders can ensure the outcomes of their supported projects are widely shared and their stated values of open access and support of open science are realized.

Considerations for funding agencies

Infrastructure

- Develop systems to generate automatic updates to a DMP using the PID knowledge graph.
- Provide model maDMPs for researchers that incorporate PID infrastructure as a means of demonstrating the new connections made possible with the addition of identifiers.

Policies

- Require researchers and/or research offices to enable appropriate sharing of the content of DMPs. While consideration for sensitive information and/or intellectual property would need to be established, this transparency facilitates compliance and adherence with best practices and could be based on adoption of the Research Data Alliance Recommendations for FAIR DMPs[13].
- Require publishers that receive funding and/or article processing charges to declare the funder ID and/or grant ID for the funding organization in all published articles. This connection will ensure that the results from the funding are publicly available.
- Require and facilitate the use of ORCID iDs, ROR IDs, grant IDs, DOIs, and funder IDs in all grant applications and reporting, as appropriate and to the extent possible.

Sustainability

- Implement pilot projects within the funding organization that use the connections made in maDMPs and PIDs. Through these exploratory projects all funders can learn how to build custom implementations and the community can further develop use cases.
- Invest in continuing support, including through partnerships with institutions and scientific societies, for open PID and DMP infrastructure for sustainability purposes.
- Join membership-supported organizations providing infrastructure, tools, and services essential to research activity, research data, and researchers themselves. Examples include Crossref, DataCite and ORCID.

[13] "Draft Recommendations," Research Data Alliance Exposing Data Management Plans Working Group, accessed August 20, 2020, https://forms.gle/2RuZCd2KeNMdH9CMA.

Appendix 1: December 2019 Conference Agenda

Implementing Effective Data Practices: a Conference on Collaborative Research Support

Omni Shoreham Hotel
2500 Calvert Street NW
Washington, DC 20008
Executive Room

AGENDA

Wednesday, December 11, 2019

8:30 a.m.–9:15 a.m. **Welcome and Introductions**

Pre- Conference Interviews: What we learned

Judy Ruttenberg, Director, Scholars and Scholarship, Association of Research Libraries

- Joel Cutcher-Gershenfeld, Professor, The Heller School for Social Policy and Management, Brandeis University

9:15 a.m.–10:00 a.m. **Panel 1: Current state and leading practices**

- Moderator: John Chodacki, Director of the University of California Curation Center (UC3) and Co-Chair FORCE11, Co-Chair Research Data Alliance Active Data Management Plans Interest Group
- Maryann Martone, Professor Emeritus Neuroscience, University of California, San Diego and FORCE11 Advisory Board Member
- Natalie Meyers, E-Research Librarian Navari Family Center for Digital Scholarship, University of Notre Dame and Co-Chair Research Data Alliance Exposing Data Management Plans Working Group
- Heather Pierce, Senior Director of Science Policy and Regulatory Counsel, Association of American Medical Colleges
- Maria Praetzellis, Product Manager for UC3/California Digital Library's Research Data Management Initiatives (including DMPTool, Support Your Data, and the NSF-funded machine-actionable DMP grant project)

10:00 a.m.–10:45 a.m. **Group work: Visions of success**

10:45 a.m.–11:00 a.m. **Break**

11:00 a.m.–11:45 a.m. **Panel 2: Incentives and policy**

	• Moderator: Cynthia Hudson-Vitale, Head of Research Informatics and Publishing, Pennsylvania State University Libraries and Visiting Program Officer, Association of Research Libraries • Greg Madden, Associate CIO for Research, Pennsylvania State University • Anurupa Dev, Lead Science Policy Analyst, Association of American Medical Colleges • Jason Gerson, Senior Program Officer for the Clinical Effectiveness and Decision Science, Patient-Centered Outcomes Research Institute • Jennifer Muilenburg, Research Data Services Librarian at the University of Washington and Visiting Program Officer, Association of Research Libraries • Carl Kesselman, Professor, Epstein Department of Industrial and Systems Engineering, University of Southern California
11:45 a.m.–12:30 p.m.	**Group work: Workflow mapping**
12:30 p.m.–1:15 p.m.	**Lunch and gallery walk of group work**
1:15 p.m.–2:15 p.m.	**Panel 3: Government and funding agency perspectives** • Moderator: Katie Steen, Federal Relations Officer, Association of American Universities • Beth Plale, Science Advisor for Public Access, National Science Foundation • Dina Paltoo, Assistant Director for Policy Development, National Library of Medicine • Carly Robinson, OSTI Assistant Director, Office of Information Products and Services, US Department of Energy • Benjamin Pierson, Senior Program Officer, Global Health, Bill & Melinda Gates Foundation
2:15 p.m.–3:00 p.m.	**Group work: Elements of implementation guidelines**
3:00 p.m.–3:15 p.m.	**Break**
3:15 p.m.–4:00 p.m.	**Group work: Elements of implementation guidelines, cont.**
4:00 p.m.–4:30 p.m.	**Small group reports**
4:30 p.m.–5:00 p.m.	**Panel 4: Closing reflections**

- Moderator: Joel Cutcher-Gershenfeld, Professor, The Heller School for Social Policy and Management, Brandeis University
- Zach Chandler, Director, Research Information Technology and Innovation, Vice Provost and Dean of Research, Stanford University
- Krisellen Maloney, Vice President for Information Services and University Librarian, Rutgers University
- Elaine Westbrooks, Vice Provost for University Libraries and University Librarian, UNC Chapel Hill

5:00 p.m. **Adjourn**

Thursday, December 12, 2019

8:30 a.m.–9:00 a.m. **Welcome and check-in**

9:00 a.m.–9:45 a.m. **Interactive exercise 1: Ecosystem design**

9:45 a.m.–10:30 a.m. **Interactive exercise 2: Use cases and value propositions**

10:30 a.m.–10:45 a.m. **Break**

10:45 a.m.–11:15 a.m. **Panel 1: Integration with broad public access to data initiatives, a look ahead at 2020**

- Katie Steen, Federal Relations Officer, Association of American Universities
- Kacy Redd, Associate Vice President, Research & STEM Education, Association of Public and Land-grant Universities
- Judy Ruttenberg, Director, Scholars and Scholarship, Association of Research Libraries

11:15 a.m.–12:00 p.m. **Group work: Communication plans**

12:00 p.m.–1:00 p.m. **Lunch: Communication and Transdisciplinarity**

- Gaetano Lotrecchiano, Associate Professor of Clinical Research and Leadership, The George Washington University

1:00 p.m.–2:30 p.m.	**Next steps: Action implications**
2:30 p.m.–3:00 p.m.	**Concluding comments**
3:00 p.m.	**Adjourn**

generously funded by the National Science Foundation

Appendix 2: Participant List

Kristin Antelman
UC Santa Barbara

Agnes Balla
University of California Office of the President

David Carlson
Texas A&M University

Zach Chandler
Stanford University

John Chodacki
California Digital Library

Sayeed Choudhury
Johns Hopkins University

Ryan Clough
Association of Research Libraries

Anurupa Dev
Association of American Medical Colleges

Briana Ezray
Pennsylvania State University

Joel Gershenfeld
Brandeis University

Jason Gerson
Patient-Centered Outcomes Research Institute

Randy Hall
University of Southern California

Bob Hanisch
National Institute of Standards and Technology

Michael Huerta
National Library of Medicine

Lisa Johnston
University of Minnesota

Heather Joseph
SPARC

Scott Kahn
Helmsley Charitable Trust

Mary Lee Kennedy
Association of Research Libraries

Carl Kesselman
University of Southern California

Margaret Levenstein
Inter-university Consortium for Political and Social Research

Cliff Lynch
Coalition for Networked Information

Greg Madden
Pennsylvania State University

Krisellen Maloney
Rutgers University

Maryann Martone
UC San Diego

Natalie Meyers
University of Notre Dame

David Minor
UC San Diego

Jeff Moon
Portage

Jennifer Muilenburg
University of Washington

Lisa Nichols
White House Office of Science and Technology Policy

Sarah Nusser
Iowa State University

Andi Ogier
Virginia Tech

Dina Paltoo
National Library of Medicine

Cynthia Parr
National Agricultural Library

Heather Pierce
Association of American Medical Colleges

Ben Pierson
Bill & Melinda Gates Foundation

Beth Plale
National Science Foundation

Maria Praetzellis
California Digital Library

Howard Ratner
CHORUS

Kacy Redd
Association of Public and Land-grant Universities

Carly Robinson
US Department of Energy

Judy Ruttenberg
Association of Research Libraries

Bob Samors
Association of American Universities Senior Scholar

Jerry Sheehan
National Institutes of Health

Amrit Singh
Bill & Melinda Gates Foundation

Ginny Steel
UCLA

Katie Steen
Association of American Universities

Cynthia Vitale
Pennsylvania State University

Guenter Waibel
California Digital Library

Nancy Weiss
Institute of Museum and Library Services

Elaine Westbrooks
University of North Carolina at Chapel Hill