DATA SCIENCE JOURNAL

Borycz, J and Carroll, B. 2018. Managing Digital Research Objects in an Expanding Science Ecosystem: 2017 Conference Summary. *Data Science Journal,* 17: 16, pp. 1–6, DOI: https://doi.org/10.5334/dsj-2018-016

ESSAY

Managing Digital Research Objects in an Expanding Science Ecosystem: 2017 Conference Summary

Joshua Borycz[1] and Bonnie Carroll[2]

[1] University of Tennessee – Knoxville, College of Communication and Information, 302 Communications, 1345 Circle Park Drive, Knoxville, TN, US

[2] CENDI, c/o Information International Associates (IIa), 104 Union Valley Road, TN, US

Corresponding author: Joshua Borycz (jborycz@vols.utk.edu)

Digital research objects are packets of information that scientists can use to organize and store their data. There are currently many different methods in use for optimizing digital objects for research purposes. These methods have been applied to many scientific disciplines but differ in architecture and approach. The goals of this joint digital research object (DRO) conference were to discuss the challenge of characterizing DROs at scale in volume and over time and possible organizing principles that might connect current DRO architectures. One of the primary challenges concerns convincing scientists that these tools and practices will actually make the research process easier and more fruitful. This conference included work from CENDI, the National Federal STI Managers Group, the National Federation of Advanced Information Services (NFAIS), the Research Data Alliance (RDA), and the National Academy of Science (NAS).

Keywords: Digital Research Objects; Data Market; FAIR Data; Digital Objects; Metadata; Digital Object Identifiers; Open Data; Open Source; Identifiers; Research Objects

Digital Objects – The Core of Our Complex Data Market

There are still many open questions about fundamental issues in the data management community that need to be addressed before any significant steps can be taken to increase the scalability and the management of digital research objects (DROs) and expand the user base and effective use of DROs by the scientific research community. There are still significant discussions related to the granularity, versioning, mutability, and lifecycle aspects of existing and developing DRO architectures. What should be the core model of software development for the evolving, global data market? Peter Wittenburg, Research Data Alliance (RDA) Europe Director, sought to address this issue by pointing towards the decades long experience of the modern data community with useful, transferable tools such as persistent identification codes (PIDs), metadata systems and standards, and data sharing principles (Australian National Data Service, 2017). With this experience, he thinks that it should be possible for data and information scientists to agree upon and adopt a core management model that can address DRO issues in what he describes as the open data market.

Peter Wittenburg's primary goal has been to use a digital object architecture as a starting point to address the many challenges of modern research. One of these issues is the exponential growth of information and connection through the web. This has led to the popularity of cloud technologies (Heilig & Voß, 2014), which have presented themselves as a more universal way of providing access to and organizing information. This is a fundamental need as approximately 75–80% of a scientist's time is lost in data management (CrowdFlower, 2017; Wittenburg, 2016; Wittenburg & Strawn, 2013). DOs[1] might be a general enough infrastructure to organize the influx of scientific data around, as they are based on simple concepts with a long history. Fundamentally, DOs are meaningful entities that exist in the digital world. They are collections of

[1] DO (Digital Object) is used here as a broader case of DROs but the points made are applicable to both. In parts of the Workshop, there is emphasis specifically on the research DOs.

data that are structured, have PIDs, are organized by metadata, and that represent a coalescence of work with a common purpose (Kahn & Wilensky, 2006). DOs are a potentially useful foundation for organizing research in a complex data market for several reasons. Namely, they are clearly identifiable objects that are potentially easy to search and access as long as consistent, detailed, rich metadata are used throughout the data market. To be widely utilized by the scientific community DOs must be Findable, Accessible, Interoperable, and Re-usable (FAIR) (Australian National Data Service, 2017), but the issues of security, differing community standards, and data collection practices still remain. Wittenburg's final point was that for DOs to bear fruit significantly more information must be collected, shared, and organized by members of the research community through the use of workflow tools (e.g. *WebLicht*) (CLARIN-D, 2017), data services (e.g. Research Data Alliance) (Research Data Alliance, 2017), and community standards. The current research data market does not encourage these practices and significant advancements in DO technologies must be made before the community practices that they depend upon will evolve.

Developing Frameworks

As mentioned by Wittenburg, one of the fundamental challenges of data management is developing a DRO architecture that is scalable, widespread, and user friendly, and there is a lot of work being done to develop the necessary technical and institutional frameworks. Under the leadership of Larry Lannom, Center for National Research Initiatives (CNRI), the RDA Data Fabric Group was formed to bridge between the many existing research and data management perspectives and identify the minimal set of common components upon which domain-specific workflows may be built (Research Data Alliance, 2018). The Data Fabric group has adopted the DO architecture in the form of the Cross-Continental Collection and Management Pilot (C2CAMP). This initiative addresses the issue of data abundance from the processing side by developing automation that can collect data and generate metadata by using the CNRI DO model. The infrastructure emerging from C2CAMP is built such that each DO has a globally unique and actionable identifier that is assigned a type with tightly associated metadata and a set of queriable operations that can be connected to any DOs. This is accomplished by utilizing the Global Digital Object Cloud (GDOC), which can be used to connect to any DO that has consistent metadata and a PID. Furthermore, C2CAMP is able to distinguish between data types, such as discovery, interpretation, and processing, which will allow this framework to extend to fields with different metadata requirements. By making use of these general, simple, and well-known architectures, Lannom hopes to ease the process for scientists who are open to a new social, data sharing infrastructure.

Brooks Hanson, Senior Vice President for publications at the American Geophysical Union (AGU) is also looking at institutional frameworks for DROs. Hanson's work with AGU involves incorporating FAIR data practices into the scholarly publishing process, which is the core motivator of most of the scientific community. With his work, Hanson hopes to bolster the modern movement away from the published article and towards DROs. It is increasingly becoming necessary to connect articles to other data sources that provide context, methods, software, funding information, author information, and reference information. Despite the obvious utility of connecting these data, few standards for metadata or linking have been adopted and researchers still almost exclusively use PDFs to build their knowledge base. In some cases, publishers require that relevant data and software be provided in the supplementary information by authors, but this practice is not widespread. Hanson works with the Coalition of Publishing Data in the Earth and Space Sciences (COPDESS), whose goal is to connect publishers, data repositories, and researchers in a mutually beneficial way that allows all members to address data sharing awareness by developing metadata standards that increase discoverability, connect publishers, and encourage researchers to start organizing and preparing data early within the workflow (COPDESS, 2017). If the publishing process is to include all of these new data sources, then it will present a significant organizational challenge for data scientists. The next challenge is to make sure that these new data sources can be connected and identified in a consistent and persistent way.

Incorporating Identifiers as Part of the Scholarly Record

Incorporating identifiers in the scholarly publishing process is a critical part of creating a DRO architecture and is a necessary step to connecting articles with data, authors, and institutions, and in increasing the findability of pertinent research across fields. In this newly developing data space, identifier best practices have an impact on how science gets done. Science that properly utilizes and connects databases, journals, and other research content from authoritative sources is more impactful (Gargouri et al., 2010; Goble, 2017; Lane, Owen-Smith, Rosen, & Weinberg, 2015; Murray, 2007; Piwowar, Day, & Fridsma, 2007). If a

shared identifier system is adopted then the findability of important research can vastly increase research efficiency, reproducibility, and speed. As such, there are a number of projects that are seeking to incorporate identifiers into publishing and data sharing. Julie McMurry, Oregon Health and Science University, advocates that the problem of developing robust identifiers for data that is on the web and for high impact research is non-trivial but is necessary since identifiers are the bedrock of scientific inquiry. The developed identifiers must accommodate the heterogeneity of research, be persistent, consistent, and connected to and informed by the data creator/discoverer. This might be achieved with a lighter touch that provides a small amount of top down control through generalizable standards.

The goal of Todd Carpenter from the National Information Standards Organization (NISO) is to use existing standards, e.g. Dublin Core (ASIS&T, 2018), Open Researcher and Contributor Identification (ORCID) (ORCID, 2018), and technologies to inform the development of new types of identifiers. These new identifiers must be applicable to each step in the research workflow (e.g. ideas, type of device, software, analytical model). Furthermore, these identifiers should do more than disambiguate objects, they should provide information about content through association with relevant metadata. The mission of *DataCite* (DataCite, 2018) discussed by Patricia Cruseis to provide tools for the research community that allow researchers to discover, access, use, connect, and cite data more confidently. *DataCite* hopes to provide the technical infrastructure that allows seamless connection with tools such as the ORCID code (ORCID, 2018) and *CrossRef* (CrossRef, 2018). *DataCite* has also developed a tool for connecting articles with repositories known as the Registry of Research Data Repositories (re3data), which has catalogued over 1,394 data repositories. Once the architecture and standards are in place, codes that scientists will actually want to use will have to be developed, but these architectures and standards are often discipline-specific. There is a lot of work that needs to be done on developing and interconnecting DRO architectures that are currently in development and in seeking out the scientific disciplines that are advancing them.

Research Objects: More than the Sum of Many Parts

DRO software has a lot of heavy lifting to do if it's going to be used for modern research, which is increasingly becoming data heavy even without the incorporation of the new data sources mentioned in the previous section. If data scientists are to suggest adding expanded experimental methods, computational codes, data, algorithms, workflows, Standard Operating Procedures, samples, etc. to the publication process in such a way that they can be reused and reproduced by other researchers within and across scientific fields, then significant breakthroughs have to be made in DRO technology development. These new all-inclusive DROs and their packaging must furthermore be continually tracked and vetted for accuracy to develop a user base. As multiple datasets and multiple models are often needed to support a study, each DRO must be able to be associated with datasets for construction, validation and prediction.

One such example of a DRO package is *Research Objects* (Bechhofer et al., 2013) developed in part by Carole Goble, professor of computer science at the University of Manchester. *Research Objects* is a framework by which the many nested and contributed components of research can be packaged together in a systematic way, and their context, provenance and relationships richly described. The DROs in *Research Objects* can contain a researcher's workflow, links to third-party codes, local files containing pertinent data, institutional information, links to a repository, and an in-progress or completed journal article. To start, a program called *FAIRDOM* (Wolstencroft et al., 2017) can both be used to assign a DOI to DROs that contain all of the important data relevant to a research project and can be used in conjunction with the workflow program, *MyExperiment* (De Roure, Goble, & Stevens, 2009), to help researchers gather these data. Lastly, these data can be packaged in a DRO using the *Research Object* software. DROs can be used to connect the individual steps involved in research projects, research projects into research objectives at an institutional level, and research objectives in fields at the global level. This general DRO tool presents a good concept for future development, but if such tools are to be incorporated into scientific research, then strong connections must be made with the potential user bases, which is a challenge of an entirely different kind. The involvement of the scientist as user was noted as a significant issue that needs to be addressed if the research data marketplace is ever to come to fruition.

Connecting Users with Digital Research Objects

Strong DRO drivers must maintain a user perspective in development and management in order to be closely connected with researchers and their work. Many DRO architectures have shown a significant ability to connect with individual researchers and institutions. In this session of the workshop, multiple initiatives were shared that presented possible examples for the data management community to build from

Lisa Kempler, David Vieglias, Danie Kinkade, and James Myers all discussed work on DRO packages that are growing in popularity in geoscience, oceanography, and environmental science.

Lisa Kempler works on *EarthCube* (Gil, Chan, Gomez, & Caron, 2014), which is a DRO architecture that is focused primarily on the geosciences. The goals of Kempler's project were to collect user data from geoscientists to determine their data needs and to provide access to tools and expertise from data scientists at *EarthCube* to help tune the architecture of the software to represent a wide range of geoscientific fields. The primary difficulty with this work is how to help geoscientists navigate the long list of data tools available for their fields. The end-user partnerships that this project hopes to generate require 1) interested communities, 2) a formalized process for interaction and use, 3) knowledge of use needs, and 4) creative and insightful responses from *EarthCube* technologists.

David Vieglais works with the Data Observation Network for Earth (DataONE) (Michener et al., 2011), which seeks to provide access to data across multiple repositories and to encourage data sharing and storage practices by providing education and easy to use tools. DataONE provides interoperability between 43 data repositories and provides tools such as safe storage and data cleaning. The architecture of DataONE allows flexible manipulation of data and metadata that is immutable, uniquely identifiable, resolvable, and retrievable. The data can also be annotated by users to help other scientists from the same field access and use it.

Danie Kinkade's data management work involves oceanographic data and is supported by the Biological and Chemical Oceanography Data Management Office (BCO-DMO) (Biological and Chemical Oceanography Data Management Office, 2018). BCO-DMO hopes to manage and curate research output for National Science Foundation (NSF) oceanographers and to provide help to researchers throughout the data life cycle. BCO-DMO scientists each have domain specific knowledge that provides insight into this research community. This helps them address the heterogeneity of the research in terms of output, specific repository curation, and research funding requirements. The strong community ties of BCO-DMO allow them to help oceanographers adopt data practices such metadata standards, controlled vocabularies, PIDs, and data links that best fit their needs.

James Myers' work addresses the reasons that researchers do not currently use DROs or other data management schemes and how this might be addressed through the DataNet Sustainable Environment – Actionable Data (SEAD) program (Myers et al., 2015). He suggests that the main issue with adopting new data practices is that research scientists are already very busy and are unfamiliar with the many different types of data software and data management jargon. The issue on the data management end is that the benefits of adopting new data management practices and software are outweighed by the costs to individual researchers. To correct this immediate, incremental value must be provided by engaging in these new practices. This means that they must have low operating costs, reliability, and be widely known by the scientific community. SEAD addresses these issues by creating a user-friendly interface that allows data files to be dragged and dropped into the system and annotated and linked easily. The DOIs within SEAD are persistent and can be assigned metadata that allows for easy discovery. The package is very lightweight and can be easily connected to a researcher's online research profile or institutional website. Currently this DRO project contains over 3,000,000 files, holds over 4 TB of data and has resulted in over 50 publications. SEAD's success stems in part from an ecosystem approach that not only provides data infrastructure and incremental value to individual researchers, but also supports the creation of new tools by third parties.

Conclusions

Taking the presentations and discussions of the full day DRO workshop into account as well as his expertise in data management, Jim Hendler, the Director of the Institute for Data Exploration and Applications, concluded that many different DO architectures, DRO tools, and data management strategies are available and funded, but what the scientific community really needs is one or two simple, stable, scalable solutions that can provide value in the short term. Hendler agrees that the steps necessary to develop these solutions involve choosing a basic architecture, framework, and set of principles as were discussed throughout the workshop but stated that there does not seem to be enough unity in the resulting software. Wittenburg's fundamental statement that the basic challenge of DRO implementation is scalability with respect to volume and time cannot be addressed until the data scientists themselves are able to agree on a product that fits the right criteria for research scientists. The plethora of currently available data tools presented here

could serve as a barrier to even those researchers that wish to be data conscious. Thus, for a data management tool to become widespread it must incentivize usage, be simple to adopt, provide incremental value to individuals, and support collaboration on a local and global level. The incorporation of well-funded technologies that are already in use as presented in Larry Lannom's work with C2CAMP may be a good place to start. James Myers suggestion that the ideas presented about incorporating DROs in the scholarly record and the research community at large will be hard address if researchers cannot be convinced that these new architectures will actually make their individual needs of primary importance. In the end, a simple and interoperable architecture should be built and made available so that the scientific community can see the benefits of DROs. Hendler's final suggestions was that the focus of data and information scientists should be on encouraging use rather than tweaking standards and tools. Many of the architectures presented during the workshop could serve as starting points, but Hendler's conclusion was the important step is to back an initiative and stick with it. Hendler's comments represented a general agreement that, in the end, there must be a focus on the utility for the scientist if there is to be significant advances in the management of DROs. It was recognized that there is a need to continue the development of the ecosystem and of the marketplace. Methods and standards may initially proliferate but must ultimately converge or interoperate to solve the biggest scientific challenges.

Competing Interests

The authors have no competing interests to declare.

References

ASIS&T. 2018. Dublin Core Metadata Initiative. Retrieved from: http://dublincore.org/.

Australian National Data Service. 2017. The FAIR Data Principles. *Working with Data*. Retrieved from: https://www.ands.org.au/working-with-data/fairdata.

Bechhofer, S, Buchan, I, De Roure, D, Missier, P, Ainsworth, J, Bhagat, J, Goble, C, et al. 2013. Why linked data is not enough for scientists. *Future Generation Computer Systems*, 29(2): 599–611. DOI: https://doi.org/10.1016/j.future.2011.08.004

Biological and Chemical Oceanography Data Management Office. 2018. Retrieved from: https://www.bco-dmo.org/.

CLARIN-D. 2017. Retrieved from: https://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page.

COPDESS. 2017. Coalition for Publishing Data in the Earth and Space Sciences. Retrieved from: http://www.copdess.org/.

CrossRef. 2018. CrossRef. Retrieved from: https://www.crossref.org/.

CrowdFlower. 2017. 2017 data scientists report. Retrieved from CrowdFlower: https://visit.crowdflower.com/WC-2017-Data-Science-Report_Thank-You.html.

DataCite. 2018. Welcome to DataCite. Retrieved from: https://www.datacite.org/.

De Roure, D, Goble, C and **Stevens, R.** 2009. The design and realisation of the Experimentmy Virtual Research Environment for social sharing of workflows. *Future Generation Computer Systems*, 25(5): 561–567. DOI: https://doi.org/10.1016/j.future.2008.06.010

Gargouri, Y, Hajjem, C, Larivière, V, Gingras, Y, Carr, L, Brody, T and **Harnad, S.** 2010. Self-Selected or Mandated, Open Access Increases Citation Impact for Higher Quality Research. *PLOS ONE*, 5(10), e13636. DOI: https://doi.org/10.1371/journal.pone.0013636

Gil, Y, Chan, M, Gomez, B and **Caron, B.** 2014. *Earth Cube: Past*, *Present*, *and Future*. Retrieved from: https://earthcube.org/document/2014/ec-past-present-future-excerpt.

Heilig, L and **Voß, S.** 2014. A Scientometric Analysis of Cloud Computing Literature. *IEEE Transactions on Cloud Computing*, 2(3), 266–278. DOI: https://doi.org/10.1109/TCC.2014.2321168

Kahn, R and **Wilensky, R.** 2006. A framework for distributed digital object services. *Int. J. Digit. Libr*, 6(2): 115–123. DOI: https://doi.org/10.1007/s00799-005-0128-x

Lane, J, Owen-Smith, J, Rosen, R and **Weinberg, B.** 2015. New linked data on research investments: scientific workforce, productivity, and public value. *Research Policy*, 44(9), 1659-1671. DOI: https://doi.org/10.1016/j.respol.2014.12.013

Michener, W, Vieglais, D, Vision, T, Kunze, J, Cruse, P and **Janée, G.** 2011. DataONE: Data observation network for earth — Preserving data and enabling innovation in the biological and environmental sciences. *D-Lib Magazine*, 17(1). DOI: https://doi.org/10.1045/january2011-michener

Murray, CJL. 2007. Towards good practice for health statistics: lessons from the Millennium Development Goal health indicators. *The Lancet*, 369(9564), 862-873. DOI: https://doi.org/10.1016/S0140-6736(07)60415-2

Myers, J, Hedstrom, M, Akmon, D, Payette, S, Plale, BA, Kouper, I, Marini, L, et al. 2015. (Aug. 31 2015–Sept. 4 2015). Towards Sustainable Curation and Preservation: The SEAD Project's Data Services Approach. *Paper presented at the 2015 IEEE 11th International Conference on e-Science*. DOI: https://doi.org/10.1109/eScience.2015.56

ORCID. 2018. Open Researcher and Contributor Identification. Retrieved from: https://orcid.org/.

Piwowar, HA, Day, RS and **Fridsma, DB.** 2007. Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLOS ONE*, 2(3), e308. DOI: https://doi.org/10.1371/journal.pone.0000308

Research Data Alliance. 2017. About RDA. Retrieved from: https://www.rd-alliance.org/about-rda.

Research Data Alliance. 2018. Data Fabric IG. Retrieved from: https://www.rd-alliance.org/group/data-fabric-ig.html.

Wittenburg, P. 2016. Group of European data experts in the research data alliance. *e-IRG Workshop*. e-IRG Website: e-IRG.

Wittenburg, P and **Strawn, G.** 2013. Common patterns in revolutionary infrastructures and data. Retrieved.

Wolstencroft, K, Krebs, O, Snoep, JL, Stanford, NJ, Bacall, F, Golebiewski, M, Goble, C, et al. 2017. FAIRDOMHub: A repository and collaboration environment for sharing systems biology research. *Nucleic Acids Research*, 45(D1): D404–D407. DOI: https://doi.org/10.1093/nar/gkw1032

How to cite this article: Borycz, J and Carroll, B. 2018. Managing Digital Research Objects in an Expanding Science Ecosystem: 2017 Conference Summary. *Data Science Journal,* 17: 16, pp. 1–6, DOI: https://doi.org/10.5334/dsj-2018-016

Submitted: 05 May 2018 **Accepted:** 12 June 2018 **Published:** 29 June 2018

]u[*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS