Check for updates

CASE STUDY

REVISED Implementation and assessment of an end-to-end Open Science & Data Collaborations program [version 2; peer review: 2 approved]

Huajin Wang, Melanie Gainey, Patrick Campbell, Sarah Young, Katie Behrman

University Libraries, Carnegie Mellon University, Pittsburgh, Pennsylvania, 15213, USA

v2 **First published:** 05 May 2022, **11**:501
https://doi.org/10.12688/f1000research.110355.1

Latest published: 05 Dec 2022, **11**:501
https://doi.org/10.12688/f1000research.110355.2

Abstract

As research becomes more interdisciplinary, fast-paced, data-intensive, and collaborative, there is an increasing need to share data and other research products in accordance with Open Science principles. In response to this need, we created an Open Science & Data Collaborations (OSDC) program at the Carnegie Mellon University Libraries that provides Open Science tools, training, collaboration opportunities, and community-building events to support Open Research and Open Science adoption. This program presents a unique end-to-end model for Open Science programs because it extends open science support beyond open repositories and open access publishing to the entire research lifecycle. We developed a logic model and a preliminary assessment metrics framework to evaluate the impact of the program activities based on existing data collected through event and workshop registrations and platform usage. The combination of these evaluation instruments has provided initial insight into our service productivity and impact. It will further help to answer more in-depth questions regarding the program impact, launch targeted surveys, and identify priority service areas and interesting Open Science projects.

Keywords
Open Science, Metascience, Academic Libraries, Program Assessment, User Data

Open Peer Review

Approval Status ✓ ✓

	1	2
version 2 (revision) 05 Dec 2022	✓ view	✓ view
version 1 05 May 2022	? view	? view

1. **Guy A. Rouleau**, McGill University, Montreal, Canada
   Dylan Roskams-Edris, Tanenbaum Open Science Institute, The Neuro (Montreal Neurological Institute-Hospital), McGill University, Montreal, Canada
2. **Verena Heise**, Freelance Open Science Researcher, Gladbeck, Germany

Any reports and responses or comments on the article can be found at the end of the article.

This article is included in the Research on Research, Policy & Culture gateway.

Corresponding author: Huajin Wang (huajinw@cmu.edu)

Author roles: Wang H: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Supervision, Validation, Visualization, Writing – Original Draft Preparation; **Gainey M**: Conceptualization, Data Curation, Formal Analysis, Funding Acquisition, Investigation, Methodology, Project Administration, Resources, Supervision, Validation, Visualization, Writing – Original Draft Preparation; **Campbell P**: Data Curation, Formal Analysis, Investigation, Methodology, Validation, Visualization, Writing – Original Draft Preparation; **Young S**: Data Curation, Investigation, Methodology, Validation, Writing – Original Draft Preparation, Writing – Review & Editing; **Behrman K**: Data Curation, Investigation, Methodology, Writing – Original Draft Preparation

Competing interests: No competing interests were disclosed.

Grant information: Part of the program activities described in this paper were supported by an NSF grant (Award 1839014) and a grant from the David Scaife Family Charitable Foundation (DSF) through Mellon College of Science at Carnegie Mellon University.
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Wang H, Gainey M, Campbell P *et al.* **Implementation and assessment of an end-to-end Open Science & Data Collaborations program [version 2; peer review: 2 approved]** F1000Research 2022, **11**:501
https://doi.org/10.12688/f1000research.110355.2

First published: 05 May 2022, **11**:501 https://doi.org/10.12688/f1000research.110355.1

REVISED **Amendments from Version 1**

Minor revisions are made in the text and figure legends to address reviewers' comments. "Citizen Science" was removed from Figure 1 and Figure 2 to accommodate reviewer 2's comment. Additional references are also added.

Any further responses from the reviewers can be found at the end of the article

Introduction

The ways in which research is conducted are shifting toward more open, transparent and collaborative practices (McKiernan *et al.*, 2016). This trend has been a response to changes in the funding and publishing landscape (e.g., open access mandates and open access publishing models; Davidson, 2005; Fyfe *et al.*, 2017; Kozlov, 2022), the nature of research collaboration (e.g., availability of digital collaboration platforms, trends in interdisciplinarity of research teams; Heller *et al.*, 2014; Cummings and Kiesler, 2014), the emergence of digital research infrastructures (e.g., open data repositories, open peer review platforms; Ponte et al., 2017) and cultural shifts in scientific practice (e.g., toward more open and transparent practices, open innovation; Huizingh, 2011). The term 'Open Science' has been used as an umbrella term to describe these trends. In 2018, Vicente-Saez and Martinez-Fuentes arrived at the following formal definition of Open Science through an analysis of ten years of scholarly literature on the topic: "[T] ransparent and accessible knowledge that is shared and developed through collaborative networks" (Vicente-Saez and Martinez-Fuentes, 2018, p. 434). Similarly, Fecher and Friesike proposed five schools of thought that capture the breadth and complexity of the Open Science discourse, namely, schools focused on infrastructure, collaboration, public access to research, impact measurement and democratic principles (Fecher and Friesike, 2014, p. 20). More recently, increasing reference has been made to UNESCO's definition of open science, which was defined as part of their recently adopted open science recommendations to inform global science policy-making. In this document, open science is defined as "an inclusive construct that combines various movements and practices aiming to make multilingual scientific knowledge openly available, accessible and reusable for everyone, to increase scientific collaborations and sharing of information for the benefits of science and society, and to open the processes of scientific knowledge creation, evaluation and communication to societal actors beyond the traditional scientific community." (UNESCO, 2021, p. 7).

Academic libraries played an early important role in the open science movement, particularly around open access publishing, which emerged in the 1990s with the development of scholarly publishing on the internet. The Scholarly Publishing and Academic Resources Coalition (SPARC) was formed by the Association of Research Libraries in 1997 to advocate for and promote open alternatives to the status quo of scholarly publishing, which was leading to rising cost burdens placed on libraries, researchers and academic institutions, and inequitable access to scientific knowledge (Savenije, 2004). SPARC has since taken on issues of open data and open educational resources (SPARC, 2022).

Another driving force in the open science movement has been the reproducibility crisis, which arose in psychological science in the mid 2000s and early 2010s (Pashler & Wagenmakers, 2012). Questions arose at this time about the reproducibility of published research, thus calling into question the reliability of research findings, not just in psychology but across many scientific disciplines (Ioannidis, 2005). Since this time, a variety of approaches have been developed to address this problem, such as pre-registration of research protocols, open peer review processes and journal requirements for data sharing. Many of these practices have been codified in the TOP (Transparency and Openness Promotion) Guidelines and have helped to further the open science movement (Nosek *et al.*, 2015).

Continuing the shift towards more openness in research depends on many factors, including cultural and behavioral shifts amongst researchers, changes to incentive structures and publishing models, and infrastructural developments. While many research communities recognize the value of Open Science for furthering scientific knowledge, in actual practice, openness in research has been much more challenging to achieve (Nosek *et al.*, 2015). Funders, publishers, and the public all play key roles in moving research toward open practices, as do institutions of higher education where incentive structures may run counter to a culture of research transparency. Despite this, there are various stakeholders in higher education settings that can foster Open Science practices. Moreover, Open Science overlaps with different areas of support across a university. For example, entities dealing with research integrity may take on the promotion of Open Science through a research transparency lens (Bouter, 2018). Institutional research and analysis offices may have an interest in Open Science practices, as Open Science tools and platforms can assist with measuring and tracking research impact (De Castro, 2018). Open Science initiatives may sprout from disciplinary or cross-disciplinary projects or sit within computer or data science departments. Examples of such initiatives include Stanford's multi-school Center for Open and REproducible Science (CORES) and the Berkeley Initiative for Transparency in the Social Sciences (BITTS).

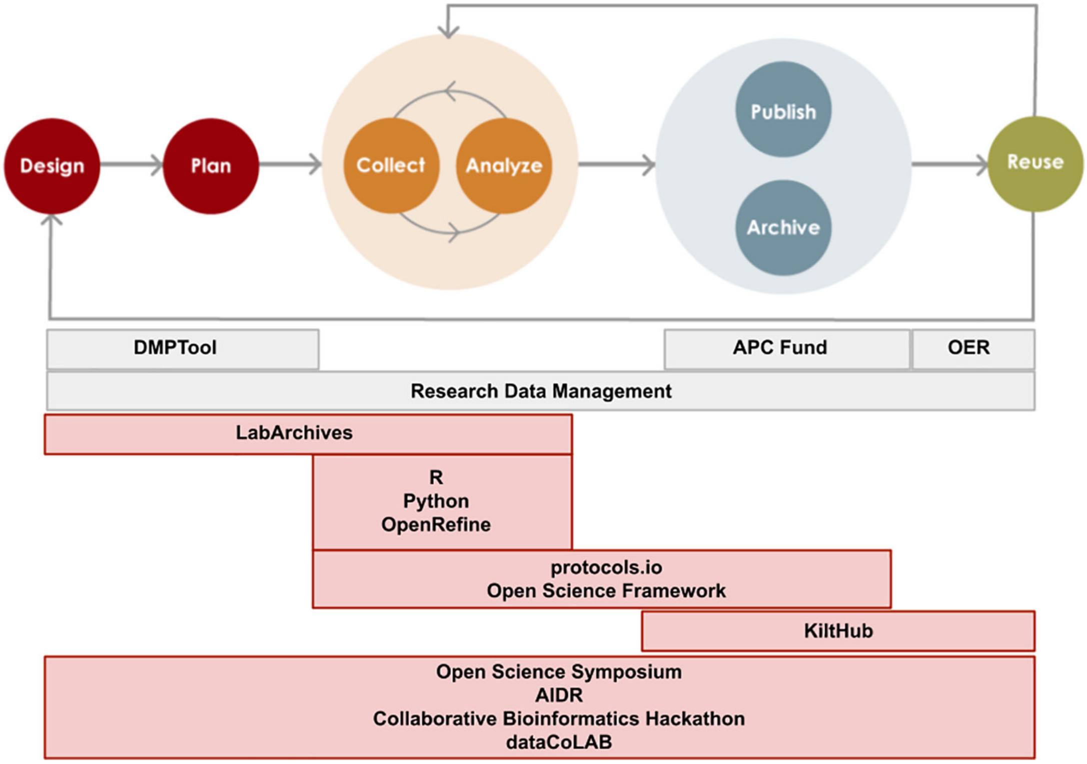

Figure 1. Open Science tools and services mapped to the research life cycle. Tools and services that OSDC supports with consultations, training opportunities, or licenses are mapped onto the phases of the research life cycle. Services and tools in gray boxes are supported by colleagues in the University Libraries that specialize in Open Access, Research Data Management, and Open Educational Resources. DMPTool: on online application that helps create Data Management Plans that meet funder mandates; OER: Open Educational Resources; APC: Article Processing Charge; OpenRefine: an open source digital tool for data cleaning and wrangling; KiltHub: CMU's institutional repository; AIDR: Artificial Intelligence for Data Discovery and Reuse Conference; dataCoLAB: Data Collaborations Lab, an initiative to foster partnerships on data science projects on real-world research data.

As Open Science has matured, academic libraries have leveraged the natural alignment to open science of existing services and principles related to information access and dissemination. For example, many libraries provide both infrastructure (e.g., institutional repositories) and funding (e.g., open access publishing funds) for sharing the products of research. Libraries also commonly provide training and support for managing research data, which relates to Open Science through the facilitation of practices that support data sharing and reuse. Recent literature suggests that libraries recognize their role in the Open Science movement, particularly in relation to repositories and open access publishing (Ogungbeni *et al.*, 2018). Ayris and Ignat discussed important roles for libraries in Open Science in Europe (Ayris and Ignat, 2018), and other research indicates a growing role for libraries in the Open Science landscape in Africa (Siyao *et al.*, 2017; Tapfuma and Hoskins, 2019). Nonetheless, to our knowledge, the formalization of these tools and services in the form of "Open Science programs" in academic libraries is rare. Moreover, most libraries are likely not yet building programs with goals of providing a suite of tools and services to support Open Science throughout the research lifecycle.

Here, we present the framework for a novel Open Science program established in 2018 at Carnegie Mellon University (CMU) Libraries. The program, called Open Science and Data Collaborations (OSDC), encompasses a range of activities, tool support and training addressing Open Science practices throughout the research lifecycle (Figure 1 and Table 1). Like other libraries, CMU Libraries also provide an institutional repository, a fund to partially cover author processing changes for open access publishing, and research data management services. While these services operate outside of the OSDC program umbrella, the programs and services work hand-in-hand to facilitate end-to-end Open Science practice, and much cross-team collaboration takes place. Therefore, we map these related services (gray boxes in Figure 1) together with those offered directly by OSDC to provide a bigger picture. The purpose of the current work is to present this model for a library-based Open Science program with a focus on program metrics and assessment. We begin with a brief environmental scan of Open Science activities at peer institutions. We follow with a logic model outlining our program activities, as well as short-, mid-, and long-term goals, and present examples of metrics that can be used and gathered to measure success. We conclude with a brief discussion of future implications for program planning and evaluation.

Table 1. Brief description of Open Science and Data Collaborations (OSDC) program components. Services in the OSDC program are composed of four major categories: tools, training, events, and collaboration.

Service	Description
Tools	
Open Science Framework (OSF)	Open Science Framework is an open-source web application for documenting and sharing project materials. OSDC provides an institutional license for OSF, as well as consultations and workshops to support use of it.
protocols.io	protocols.io is an open access repository for recording and sharing research methods and protocols. OSDC provides an institutional license for protocols.io, as well as consultations and workshops to support use of it.
LabArchives	LabArchives is a cloud-based Electronic Research Notebook (ERN) for documenting research. OSDC provides institutional licenses for the Education and Research editions of the platform, as well as consultations and workshops to support use of it.
KiltHub	Built on figshare and provided by CMU Libraries, KiltHub is CMU's comprehensive institutional repository. It can be used to make any research product publicly available and citable. CMU Libraries provides data management and light curation support for researchers using the platform.
Training	
Carpentries Workshops	OSDC maintains a membership with the non-profit The Carpentries. We organize 2-3 day hands-on workshops on foundational computing and coding skills with Python, R, shell, Git, or OpenRefine with instructors and lesson plans from The Carpentries. Our membership also allows us to provide Carpentries Instructor training to a handful of researchers at CMU each year.
Libraries Workshop Series	Short workshops on open science tools and research practices, including short Carpentries-style workshops on R.
Events	
Collaborative Bioinformatics Hackathon	Hosted 1-2 times a year in partnership with other academic partners and DNAexus, the hackathon is a multi-day event that brings together academic and industry researchers from around the world to collaboratively work on crucial problems and opportunities in clinical bioinformatics. OSDC provides support on data management and sharing the outputs of the event.
Open Science Symposium	An annual symposium organized by OSDC that brings together researchers, funders, publishers, and tool developers to discuss the challenges and opportunities of Open Research.
AIDR (Artificial Intelligence for Data Discovery and Reuse)	An annual symposium organized by OSDC that focuses on harnessing the power of AI to accelerate the dissemination and reuse of scientific data and building a healthy data ecosystem.
Collaboration opportunities	
dataCoLab (Data Collaborations Lab)	Matches up researchers who want help with their datasets with consultants who have data science skills. Through weekly office hours and project-based consultations, this creates opportunities for people with different technical and disciplinary backgrounds to work together, following best practices that enhance reproducibility.

Environmental scan

To evaluate the landscape of library Open Science programs, in the spring of 2021 we conducted an environmental scan of Carnegie Mellon University's peer institutions, a list of 13 institutions of common qualities and goals (as defined by the Office of Institutional Research and Analysis at the university) (Table 2). From the websites of each individual institution and each institution's library, we searched for the general terms "open science," "open scholarship," and "open research" to attempt to locate similar programming and services to those offered by OSDC at CMU. We also searched for traditional Open Access resources, such as an institutional repository and an institutional Open Access policy to benchmark the number of peers with general Open Research services that may not be specifically described as "Open Science." While

Table 2. Summary of Open Science programs at Carnegie Mellon University (CMU)'s peer institutions. CMU's peer institutions are California Institute of Technology, Cornell University, Duke University, Emory University, Georgia Institute of Technology, Massachusetts Institute of Technology, Northwestern University, Princeton University, Rensselaer Polytechnic Institute, Rice University, Stanford University, University of Pennsylvania, Washington University in St. Louis. The different levels of open science programming denoted in the table were defined as follows: Library Sponsored Open Science Programs: Full library-sponsored end-to-end Open Science programs similar to what CMU offers; Library Open Research Programming: Library-sponsored general open access/research/scholarship programs or units; Disciplinary Open Science Centers and Programs: Open Science programs and centers that are situated outside of or separate from the institution's library; Open Access Policies: Institutions with a policy or mandate for open access publishing; Institutional Repositories: Institutions with infrastructure for open sharing of research products and publications.

Library sponsored Open Science programs	Library Open Research programming	Disciplinary Open Science centers and programs	Open Access policies	Institutional repositories	Total peer institutions
0	4	5	10	12	13

the majority of peer institutions support open scholarship through open access policies and institutional and data repositories, dedicated open science centers and programming, either through the university library or through departmental structures, are less common (Table 2).

In addition to manually checking the websites of peer institutions and to identifying any related programs outside of peer institutions, we ran a Google search using the following search string, which queried sets of search terms within three words of other search terms and limited the results to websites of U.S.-based postsecondary institutions: "open|reproducible|reproducibility AROUND(3) research|science|scholarship AROUND(3) institute|center|program" site:.edu. We then reviewed the results of the search until no relevant results were found on five consecutive results pages. No additional dedicated Open Science programs were identified with the Google search.

Program implementation

In 2017, we began to develop services and initiatives to support open and reproducible research in response to the growing need for reliable infrastructure and training for Open Research practices (Mckiernan *et al.*, 2016; Nosek *et al.*, 2015; AAU-APLU Public Access Working Group Report and Recommendations, 2017). What began as an *ad hoc* collection of services and collaborations was formalized as the Open Science & Data Collaborations (OSDC) Program in 2018. This program within Carnegie Mellon University Libraries consists of a team of subject librarians with deep research expertise and specialists in research data management and Open Data. While we have adopted the name "Open Science" due to its common use in the community, we support all types of research and often use the term "Open Research" to describe our activities. The OSDC program provides training and support for tools and practices that can be mapped onto the phases of the research life cycle (Figure 1). Since our services together cover the entire life cycle, we describe the program as providing "end-to-end" support. The program has been in a phase of rapid expansion since its inception in 2018. We have leveraged our research experience, particularly in the life sciences, and our existing campus partnerships to develop new services that we believe will be of use and interest to the CMU community and help make research products open in accordance with the FAIR principles (Wilkinson *et al.*, 2016).

Prior to the development of the OSDC program, CMU Libraries already provided extensive support for some areas of scholarship that are typically defined as Open Science, such as Open Access publishing (Fecher and Friesike, 2014). Our comprehensive institutional repository, KiltHub, also predates the creation of OSDC. Currently, we collaborate with colleagues in the library that specialize in open access, research data management (RDM), and open educational resources (OER) to provide holistic support for open scholarship. These areas of Open Science that are outside of the purview of the OSDC program are not currently assessed by us (Figure 1). In spite of the fact that KiltHub existed prior to the development of OSDC, we currently help support the platform and assess its usage as an integral piece of infrastructure for data sharing.

Program assessment

As OSDC expands, one challenge has been getting structured and actionable feedback from the CMU research community, particularly from disciplines outside of the life and social sciences. To this end, we created a new arm of the program in 2021 that focuses on research and assessment. Our recent work has focused on developing a logic model and quantitative metrics on tool usage and event and training attendance. We will use this multi-pronged assessment approach to identify gaps in our service, shape the growth of the program in a data-driven and user-centered manner, identify future members for our Advisory Board, and create surveys designed for specific segments of our user

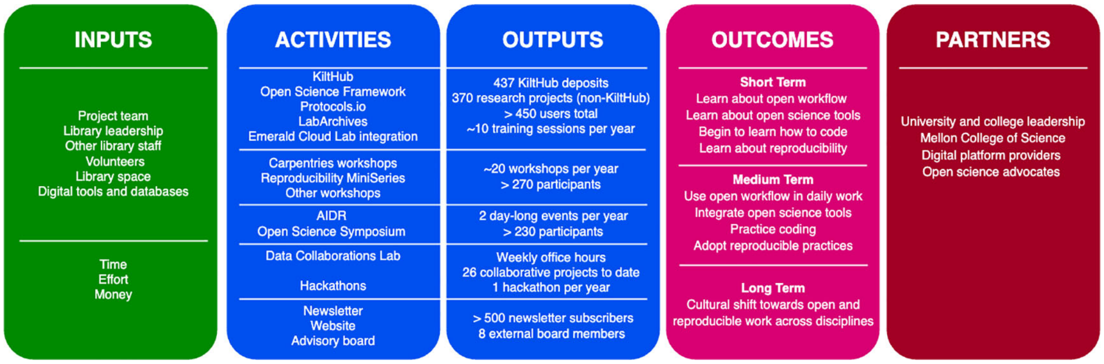
INPUTS
Project team
Library leadership
Other library staff
Volunteers
Library space
Digital tools and databases
Time
Effort
Money
ACTIVITIES
KiltHub
Open Science Framework
Protocols.io
LabArchives
Emerald Cloud Lab integration
Carpentries workshops
Reproducibility MiniSeries
Other workshops
AIDR
Open Science Symposium
Data Collaborations Lab
Hackathons
Newsletter
Website
Advisory board
OUTPUTS
437 KiltHub deposits
370 research projects (non-KiltHub)
> 450 users total
~10 training sessions per year
~20 workshops per year
> 270 participants
2 day-long events per year
> 230 participants
Weekly office hours
26 collaborative projects to date
1 hackathon per year
> 500 newsletter subscribers
8 external board members
OUTCOMES
Short Term
Learn about open workflow
Learn about open science tools
Begin to learn how to code
Learn about reproducibility
Medium Term
Use open workflow in daily work
Integrate open science tools
Practice coding
Adopt reproducible practices
Long Term
Cultural shift towards open and reproducible work across disciplines
PARTNERS
University and college leadership
Mellon College of Science
Digital platform providers
Open science advocates

Figure 2. Graphic summary of a logic model. A logic model was created by listing inputs, activities, outputs, outcomes, and partners for each activity and creating a narrative. A simplified graphic summary was created to represent essential elements of the logic model. Inputs: resources required for all activities. Activities: the five groups of activities in the OSDC program; from top to bottom: tools, workshop, events, collaboration, and outreach. Outputs: product of each activity. Outcomes: short-, medium-, and long-term goals. Partners: partnerships formed to date. Emerald Cloud Lab: a remote controlled, automated lab where equipment is run remotely and workflow, data and code are automatically recorded; Reproducibility MiniSeries: short format workshop series that currently include R and OpenRefine.

community. Keeping the user in mind will be critical as the needs of the research community continue to evolve against the dynamic backdrop of data sharing mandates and the increasing desire for transparency and reproducibility in the research community.

Logic model

The first component of our assessment strategy is a logic model (Newcomer *et al.*, 2015) that provides a snapshot of the activities offered by the OSDC program and their respective outputs, resources needed to run the program, short-, medium-, and long-term goals to achieve for our users, and a list of partnerships formed through the program (Figure 2). It provides a bird's-eye view of the activities of the program and guides operational decisions and strategic planning. Values in outputs are estimated and serve as a baseline for further assessment. It should be noted however that values are not comparable between tools due to different time frames used for component datasets. The logic model will be reevaluated yearly.

5W1H metrics framework

To find more quantitative ways to measure program impact, we developed the second component of our current assessment strategy, a "5W1H" (Who, What, When, Where, Why, How) framework. Using this framework, originally developed for communication action research (Yoshioka *et al.*, 2001), we developed metrics that use tool usage and event attendance data to help answer questions about our users and their use of our services.

First, we collected existing usage data across tool platforms. Specifically, we gathered usage data for the following tools: KiltHub, Open Science Framework, protocols.io, LabArchives. We also collected event registration data for Open Science-themed Libraries workshops, Carpentries workshops, Open Science Symposium, AIDR (Artificial Intelligence for Data Discovery and Reuse) Conference, and dataCoLAB (Data Collaborations Lab). We used event registration data as a proxy for event attendance since attendance data were not consistently collected. We expect, however, that registrations for events are higher than the actual attendance. Finally, engagement with the Open Science Newsletter, one of our core marketing tools, was also included in the assessment. The details of how data were collected for each of these services can be found in the Data Collection Methods section of this paper.

Data across platforms and events were cleaned and aggregated into a master dataset. The protocol we used to create the master dataset is published on protocols.io. We used Andrew IDs (CMU institutional emails) as unique identifiers for users of our services. Since the KiltHub dataset includes institutional and departmental affiliation data for all current CMU graduate students, staff, and faculty, we matched Andrew IDs for users of our other services to the KiltHub dataset. If users provided non-institutional email addresses, we queried their names in the CMU directory to determine their Andrew IDs, if possible. Undergraduates are represented in the dataset simply as "Undergraduates" since we could not consistently determine their departmental affiliation. We confirmed their status as undergraduates by querying their names in the CMU directory.

Table 3. Number and percentage of unique users by institution. In our analyses, we only included Carnegie Mellon University (CMU) users with known departamental affiliations (n=787). CMU users with unidentified or administrative affiliations (n=56), University of Pittsburgh users (n=60), or users at other or unidentified institutions were filtered out of the dataset (n=445).

Users	Count	Percent
Carnegie Mellon University (CMU)	787	58%
CMU with unidentified or administrative departments	56	4%
University of Pittsburgh (Pitt)	60	5%
Other or unidentified institution	445	33%
Total unique users	1348	100%

CMU and University of Pittsburgh (Pitt) have some joint centers and programs, and we noted that 60 users in the master dataset (5% of the total 1,348 unique users) have primary Pitt affiliations (Table 3). For our analyses, we filtered out Pitt users. We also filtered out users from other non-CMU institutions or unidentified institutions (33%) and CMU users if we could not determine their departmental affiliation or if they were affiliated with administrative units on campus (4%). The total remaining records represented in the Results section (n=787) represents 58% of the total unique records (n=1,348) with which we began.

Based on the master dataset and platform-specific data, we generated a list of meaningful questions within the 5W1H framework (Table 4). Metrics and their sub-variables were then defined to answer those questions. Currently, we are focusing on questions that we are able to answer readily with the data at hand, e.g.: who uses our tools and participates in our activities, which disciplines are the most engaged, and how do people use our tools and activities? Most of the metrics related to the questions require data collected from platform dashboards or provided by vendors. In other cases, the metric was derived from the dashboard or vendor data with simple calculations. For example, we can use data from the KiltHub

Table 4. List of current metrics and associated variables. Metrics being used to evaluate the performance of the Open Science and Data Collaborations (OSDC) program and the variable(s) that are used to calculate each one. Metrics are organized using a "5W1H" (Who, What, When, Where, Why, How) framework representing the major classes of query the dataset is designed to answer. Data for each metric can either be collected directly from dashboards, vendors, or registration records, or derived from the direct data with simple calculations.

Question	Metric	Variable(s)	Source of data
Who	User affiliation	Institution, Department	Dashboard
	Stage of career	User type (faculty, postdoc, etc.)	Derived
	Superusers	Counts, Number of projects and registrations (all tools/events)	Derived
What	Number of users per tool	User (T/F) - all tools/events	Dashboard, vendor
	Number of tools/events used per user	User (T/F) - all tools/events	Derived
	Number of registrations per event	Count (all events/workshops)	Dashboard
	Number of attendances per event	Count (all events/workshops)	Dashboard
	Number of event/ workshop registrations per user	Counts (all events/workshops)	Derived
	Departmental breakdown of users per tool/event	User (T/F), Institution, Department	Derived
	Career stage breakdown of users per tool/event	User (T/F), Career Stage	Derived

Table 4. *Continued*

Question	Metric	Variable(s)	Source of data
When	Growth rate (growth over time)	Number of users plus time/date field	Derived
	Activity over time	Output plus date/time fields	Derived
Why	User satisfaction* (qualitative and quantitative)	User comments/feedback	Advisory Board, surveys
	Financial metrics* (for users)	Cost savings	Vendors
How	Output (number of products, tasks completed, etc.)	Number of projects and registrations (OSF), number of notebooks (LabArchives), number of activities (LabArchives), number of protocols (protocols.io), count of events of each type attended (workshops, Carpentries, DataCoLAB, AIDR_OSS), Count_KiltHub (KiltHub)	Dashboard, vendors
	Reach	Open rate, Click rate (newsletter)	Dashboard

*Metric that we have partial data for and can be calculated in the future.

dashboard to determine the institutional and departmental affiliation of each user. We can then derive the stage of the career of the user by querying institutional email addresses in the CMU directory. It should be noted that while we know that our users are largely in the Pittsburgh area, we do not collect any other information, such as IP addresses, that could be used to answer "Where" questions. Additional data collection is required to answer more nuanced questions about the impact and value of our services for users. Questions we eventually hope to address include: why do people use our tools or activities, how much value did we provide to users, and what impact are we making in people's research process and in the whole research ecosystem?

Importantly, the metrics can be applied to the program as a whole or to specific tools and services. The user affiliation metric will indicate whether we are achieving broad coverage of disciplines with the program. The superuser metric will help us identify Open Science advocates in our campus community that can support our outreach efforts and provide valuable feedback. We can also track adoption of specific services over time with our Growth Rate metric to examine trends in Open Science research and help guide decision making.

Example applications of current metrics

Even though the framework is still a work in progress, limited by the state of existing data, it already allows us to ask simple questions. As a proof of concept, we provide a few examples of applying this framework to extract interesting patterns from existing data.

To obtain an overview of disciplinary engagement, we summarized the number of users for each department, based on their primary affiliations (Figure 3). The data came from the integrated dataset where usage of a given service or activity was represented as a "true/false" value. A user with a "true" in any of the services as counted as 1. These data show that the Heinz College of Information Systems and Public Policy has the highest number of users, followed by the Biological Sciences Department, University Libraries, and the Psychology department. We think this result can be partially explained by disciplinary culture as these disciplines are traditionally more engaged with library services and more active in the Open Science movement. Interestingly, some engineering and computer science departments also have high numbers of users, suggesting that we are starting to generate buy-in from these disciplines.

The number of users of each department does not necessarily reflect how active users from these departments are. Using KiltHub as an example, we further dissected the level of user activity for each individual tool hosted by the program. The reason we did not use the integrated master dataset for this purpose is that measurements between platforms, e.g., number of notebooks or number of registrations, are not comparable with each other. A breakdown of the number of users on KiltHub revealed that Software Engineering Institute (SEI), Psychology, and University Libraries, again, were among the departments or academic units that have the greatest number of KiltHub users (Figure 4A, blue bars). However, when looking at user activity levels, specifically public items owned by users, those from SEI collectively owned fewer items compared to those from Psychology and University Libraries (Figure 4A, red line). We further analyzed KiltHub usage at the level of individual users and saw different departments emerge when compared to the result from the total number of

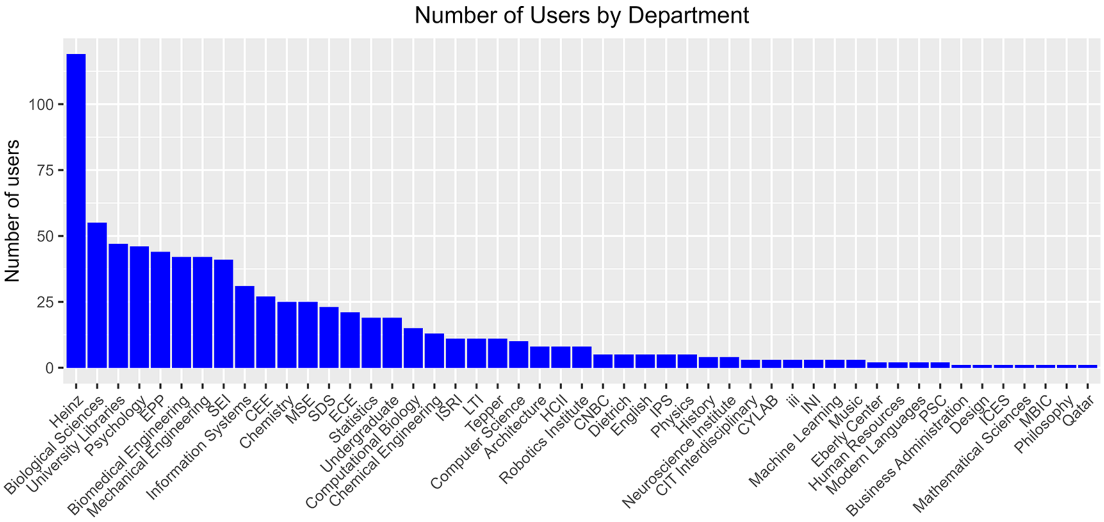

Figure 3. Departmental breakdown of all OSDC users. Number of users by department or academic unit, based on their primary affiliations. The main dataset integrating all usage data was used as input. Each user is counted only once even if they use multiple services. CNBC: Center for the Neural Basis of Cognition, CEE: Department of Civil and Environmental Engineering, CIT: College of Engineering, CYLAB: Security & Privacy Institute, ECE: Department of Electrical and Computer Engineering, EPP: Department of Engineering and Public Policy, MSE: Department of Materials Science and Engineering, Dietrich: Dietrich College of Humanities and Social Sciences, Heinz: Heinz College of Information Systems and Public Policy, HCII: Human Computer Interaction Institute, INI: Information Networking Institute, ICES: Institute for Complex Engineered Systems, IPS: Institute for Politics and Strategy, ISRI: Institute for Software Research, iii: Integrated Innovation Institute, LTI: Language Technologies Institute, MBIC: Molecular Biosensor and Imaging Center, PSC: Pittsburgh Supercomputing Center, SDS: Social and Decision Sciences, SEI: Software Engineering Institute, Tepper: Tepper School of Business.

users. Among the top 10 departments ranked by the median values of the number of public items owned by each user in a given department, the School of Business ranked the highest, followed by University Libraries and the Computer Science Department (Figure 4C). Even though the median values were relatively low overall – less than five items per user – some users owned much higher numbers of public items on KiltHub (Figure 4C, outliers). This trend was also reflected at the level of the individual user, with the most active users owning more than 20 public items on KiltHub while the majority of users owned less than five items (Figure 4B). We define users with more than 10 public items as "superusers." We were able to identify these users (anonymized in this manuscript) and their department affiliations (Figure 4D). The ability to identify superusers is especially valuable for collecting targeted feedback with interviews and surveys for service improvements in the future.

An important indicator of a program's success is its growth over time. We used tool usage over time as a proxy to explore this question. By examining the total number of users who deposited data on KiltHub or the total number of accounts on LabArchives and protocols.io over time (Figure 5), we found that there has been a steady increase in use every year since the inception of the OSDC program. This initial analysis establishes a useful baseline for future longitudinal studies.

Discussion and future directions

Data sharing has represented a massive paradigm shift for research in recent years (Gewin, 2016). This trend goes beyond data and applies to all research activities. In this paper, we use the term "data" loosely to refer to all research outputs including but not limited to data, code, and workflow. There are varying disciplinary norms and attitudes around data sharing and researchers often lack the training, time, infrastructure, or perceived incentives to openly share their research products. Fear that the data will be misused is another common concern (Fecher *et al.*, 2015; Tedersoo *et al.*, 2021; Tenopir *et al.*, 2015). To address these challenges, we have created one of the first end-to-end Open Science programs sponsored by a library, with services that map onto all phases of the research lifecycle. One of our guiding priorities for creating Open Science services is that they have an impact on fostering collaboration and a cultural change towards research transparency. It is important, however, that we remain mindful of the barriers that researchers face. We therefore support a full gradient of Open Science practices, ranging from sharing research products publicly to improving the reproducibility of private workflows. For example, for researchers that are unable to share data due to working with sensitive data types, or are simply uncomfortable with data sharing, we might improve the reproducibility of their workflow for their future selves and collaborators. These types of consultations provide us with valuable opportunities to

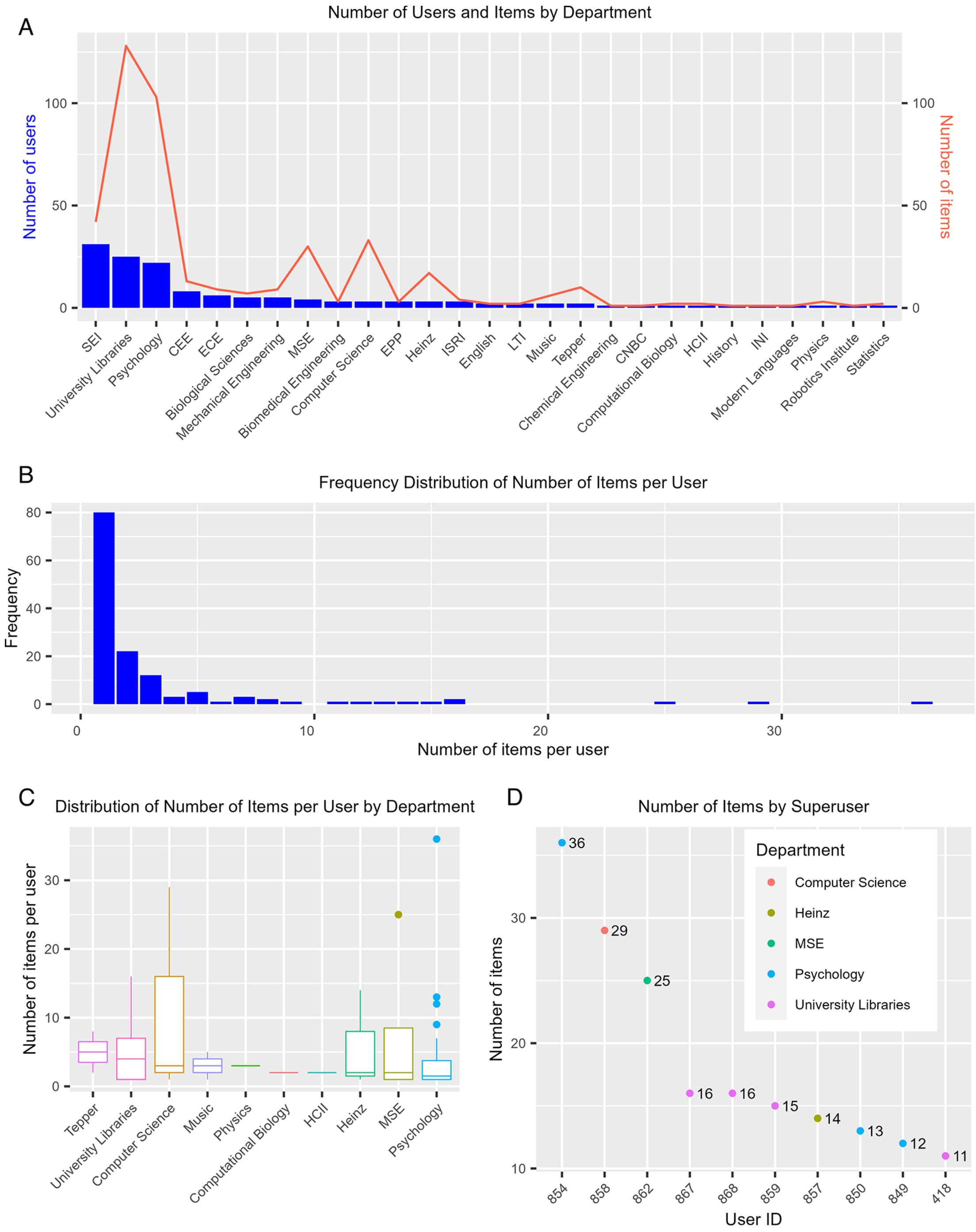

Figure 4. Summary of KiltHub use. (A) Departmental breakdown of number of KiltHub users (blue bars) and number of public items owned by users collectively in these departments (red line). (B) Frequency of number of public items owned per user (frequency is measured by the count of users with the specified number of public items owned). (C) Boxplot showing the distribution of the number of public items owned per user for the top 10 departments, measured by mean number of items per user. The boxed area represents the interquartile range (IQR), with the lower bar representing the first quartile (Q1) value, the intermediate bar representing the median value (Q2), and the top bar representing the third quartile (Q3) value. The lines, or "whiskers", extending above and below the boxed area represent the range of values contained within 1.5 times the interquartile range (1.5 x IQR). Points extending beyond the whiskers represent outlier values (> 1.5 x IQR). (D) Number of public items owned for the 10 most active users (items > 10). These users are identified by their User ID (autonumber value assigned by Excel) to conceal the users' identities. CNBC: Center for the Neural Basis of Cognition, CEE: Department of Civil and Environmental Engineering, ECE: Department of Electrical and Computer Engineering, MSE: Department of Materials Science and Engineering, Heinz: Heinz College of Information Systems and Public Policy, HCII: Human Computer Interaction Institute, INI: Information Networking Institute, ISRI: Institute for Software Research, LTI: Language Technologies Institute, SEI: Software Engineering Institute, Tepper: Tepper School of Business.

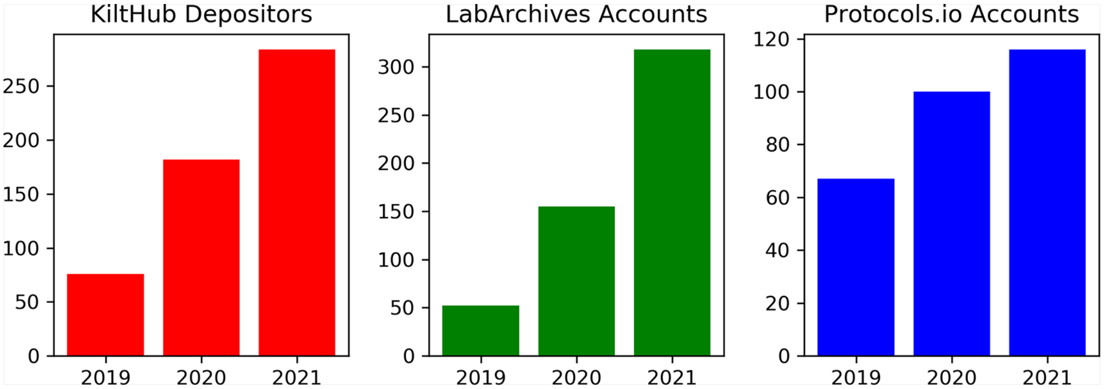

Figure 5. Tool usage over time (2019-2021). Number of total depositors on KiltHub, user accounts on LabArchives, and user accounts on protocols.io increased each year since the beginning of the program in 2018. Values presented are cumulative counts.

not only improve researcher experience around Open Science, but also discuss the benefits of publicly sharing research. Together with our community-building events, these types of interactions with researchers allow us to foster a culture of transparency.

As we continue to create services, we need to rely not only on conversations with researchers, but also on periodic quantitative assessments to understand their impact. The work presented here is the beginning of our program assessment and provides methods that we will update periodically and eventually supplement with additional metrics. This will allow us to focus our resources on priority areas, maximize the efforts of our small team, and guide our efforts to secure funding.

Limitations of current data sources and future user data management strategy

Most of the data currently in our possession focuses on event registration and tool usage. Registration data is useful primarily for providing insights into, e.g., the popularity of specific OSDC initiatives (number of registrants, frequency of use, etc.), the reach/coverage across CMU and broader research community, specifically with regard to user type (student, faculty, etc.), institutional and departmental affiliation, and potential superusers. Our current data also include several variables related to the effectiveness of our various initiatives, e.g., number of items on KiltHub, number of projects and registrations on OSF, number of notebooks or activities on LabArchives, event attendance, or open and click rate of the Newsletter. However, we are only scratching the surface about the effectiveness or impact of the various OSDC initiatives; many deeper questions, e.g., how many publications, grant applications, career opportunities that we help users to obtain, and how much time we save users in their daily research, cannot be answered with the existing metrics. Developing metrics that reflect researchers' productivity and success more directly would strengthen our value proposition to researchers and help them to rethink how productivity, efficiency, and impact are evaluated.

Despite these limitations, the current data and the 5W1H metrics framework will serve as a baseline to develop a strategy for user data management in the future to guide data collection, update, and analysis. A large part of our data collection process is limited by the platforms or tools that host the data. However, the usefulness of data can be improved by a few tweaks. To get the most out of our usage and registration data, a date field should be included for all relevant data tables, which will allow us to infer, for example, how the number of link clicks in a particular issue of our newsletter influences the number of registrations for specific events. Importantly, date information will help to control confounding factors when inferring whether the uptake in open science behavior is directly caused by the services we offer, or rather driven by important events and policy changes outside of the OSDC program. More generally, date information can reveal temporal patterns in the use of various tools/platforms and attendance at particular events, allowing us to better target our outreach efforts and workshops. Adding a date field will also allow us to track more meaningful changes in use after controlling for natural fluctuation patterns, which can in turn be used to estimate programmatic growth or decline.

To develop a more mature user data management system, metrics should be developed to provide insight into different stages of the research lifecycle (Figure 1), particularly around the issues of productivity, efficiency, and impact. The specific variables that are relevant in each case will depend on the particular stage of the research lifecycle we are considering. For example, usage of protocols.io would likely reflect the data collection and analysis stage, while KiltHub usage more likely reflects the publishing and sharing stage.

We would also like to develop a more systematic data collection strategy that allows regular updates to data and results. The current data collection, cleaning, and analysis process is highly manual, making it time-intensive, error prone, and difficult to update. Developing an automated or semi-automated workflow would help to ease the administrative overhead on data updates and enable us to ask more longitudinal questions.

Applications of the logic model and 5W1H framework

The combination of the logic model and the 5W1H framework provides complementary instruments to evaluate our program's impact and to inform decision making. The logic model provides a bird's-eye view of program activities and is an ideal tool for goal setting and communicating higher level ideas with leadership and stakeholders. The 5W1H framework, on the other hand, helps to evaluate and understand our activities and user engagement at a more granular level, making it possible to quantitatively assess our successes, identify areas for improvement, prioritize future work, and refine outreach strategies.

The most difficult thing in the metrics framework is the "why" question: what are the motivations for people to use our services? Is it to meet funder/publisher mandates, to get credit, or for other reasons? Developing such metrics would make it possible to quantitatively assess user motivation and productivity, evaluate the value and success of our services, and identify areas for improvement and prioritization in the future. For these types of questions, we would like to get direct feedback from users through surveys and interviews. To this end, the "superuser" metric (Figure 4D) in the 5W1H framework helps to identify the right users to reach out to. We had initial success applying this metric to form a OSDC Advisory Board from our users, composed of graduate students, postdoctoral fellows, and faculty who are Open Science advocates and practitioners, and represent a variety of disciplines. The group meets 3-4 times a year to provide feedback in the style of a focus group on service updates, outreach strategies, and disciplinary practices and challenges.

Our work on the implementation of an end-to-end Open Science program and the development of assessment instruments will serve as a model that can be adopted by Open Science programs at other institutions, or other service-oriented organizations that wish to evaluate their success and impact. With further enrichment and adoption, the combined logic model and 5W1H framework we developed has the potential to grow into a benchmarking tool for equivalent programs and products that require both qualitative and quantitative assessment.

Data collection methods

KiltHub. For the master dataset, profiles of all active users on or before 2 April 2021 were downloaded from the KiltHub Admin dashboard. We used the following data fields from the profiles for this study: ID, first and last name, email address, affiliation (department or center), and number of public items owned. Only data depositors with more than one public item owned were included in data analysis, while the names and email addresses of all users were used for data harmonization (see the published protocol for details). We filtered out private items since there are many reasons why a user might choose to keep their projects private. For usage over time, a separate dataset was downloaded from the dashboard that contains information about depositors.

protocols.io. Usage data including number of users, private, protocols, and public protocols were provided quarterly by the vendor and were collected for this study on 30 November 2021. Per protocols.io privacy policies, identifying information such as names, email addresses, or departmental affiliations were not shared. Therefore, these data were not included in the User Summary in Figure 3.

Open Science Framework (OSF). We collected user data from Open Science Framework (OSF) with our institutional OSF dashboard that includes first and last names and number of public projects and registrations on 19 January 2021. The number of public projects and registrations per user is the sum of these two metrics. Institutional emails were gathered by querying names in the Carnegie Mellon University Directory.

LabArchives. A Detailed Usage Report was downloaded from the Site Administrator dashboard. The report included first and last names, institutional email address, type of account (CE type), number of notebooks, and number of activities. For the purpose of this study, we were interested in Researcher and Instructor accounts. Student accounts were filtered out of the dataset. Data for the User Summary in Figure 3 were collected on 20 January 2021 and the usage over time data in Figure 4 were collected on 20 November 2021.

Newsletter. Newsletter data was accessed through Mailchimp. We were interested in users that routinely open the newsletter. To gather these data, we navigated to the Audience Dashboard and selected the Often segment under Engagement. This allowed us to collect data on our most engaged users, including first and last names and email addresses. We then searched user profiles in Mailchimp to gather data on Open Rate and Click Rate for each user. Newsletter data were collected on 20 January 2021.

Events. Event registration data for Open Science Symposium and AIDR were collected from the Indico, EasyChair, and EventBrite platforms. The data collected for each registrant included event name, first and last names, email address, and institution. Participation data from dataCoLAB were collected using a project intake form in Google Forms.

Workshops and training. Workshops and training related to Open Science at CMU are delivered primarily through two formats: (1) one- to two-hour workshops offered through the Libraries' workshop series on the following topics: OpenRefine, Jupyter Notebooks, Open Science Framework, Data Management, and R, and (2) Carpentries workshops, which are two-to-three-day training sessions organized and managed by the Libraries' Carpentries organizing team. Registration data were collected for Libraries' and Carpentries workshops from LibCal and Eventbrite, respectively. Registration data, including first and last names and email addresses, were collected for each occurrence of a workshop that had occurred by 1 January 2021. All Libraries' and Carpentries workshop data were combined for each workshop type (type defined by a combination of format and topic). Users were merged if they had used different emails for registration for different workshops, and it was clear from their name that they were the same person. For users that had registered for multiple occurrences of the same Libraries' workshop, it was assumed that they had only attended one. For Carpentries workshops, we assumed that registrants may have attended more than one workshop even if it covered the same content. Libraries' workshop data were then combined into a single dataset indicating whether or not a user had registered for a given workshop type and the total number of workshop types attended by each user.

Ethical approval

After extensive communication with the Institutional Review Board (IRB), it was advised that as this project is intended for evaluation and improvement of internal processes without making generalizing statements, did not fall under the definition of research, and therefore did not require IRB approval. Informed consent for collecting the original data hosted by the university and the libraries was obtained by the university's legal office. Data have been anonymized for this study before collection and analysis. Anonymizing the data does not change the scientific meaning of our findings.

Data availability

Because original data used to develop assessment methods contain identifiable user information, they are only for internal use. Deidentified and aggregated data are openly available in KiltHub, CMU's institutional repository (DOI: https://doi.org/10.1184/R1/19438586). Protocols used for data cleaning and processing are openly available on protocols.io (https://doi.org/10.17504/protocols.io.b29gqh3w).

Acknowledgments

We thank Hannah Gunderman and Emma Slayton for contributing workshop data. We thank Brian Mathews for helping to develop the logic model.

References

Association of American Universities (AAU) and Association of Public and Land-grant Universities (APLU): **AAU-APLU Public Access Working Group Report and Recommendations.** 2017.
Reference Source

Ayris P, Ignat T: **Defining the role of libraries in the Open Science landscape: A reflection on current European practice.** *Open Inf. Sci.* 2018; **2**(1): 1–22.
Reference Source

Bouter LM: **Fostering responsible research practices is a shared responsibility of multiple stakeholders.** *J. Clin. Epidemiol.* 2018; **96**: 143–146.
PubMed Abstract | Publisher Full Text | Reference Source

Campbell P, Wang W, Gainey M, *et al.*: **Cleaning, Aggregating, and Filtering CMU Libraries Open Science and Data Collaborations Program Data.** *protocols.io.* 2022.
Publisher Full Text

Cummings JN, Kiesler S: **Organization theory and the changing nature of science.** *J. Organ. Des.* 2014; **3**(3): 1–16.
Publisher Full Text

Davidson LA: **The End of Print: Digitization and Its Consequence—Revolutionary Changes in Scholarly and Social Communication and in Scientific Research.** *Int. J. Toxicol.* 2005; **24**(1): 25–34.
Publisher Full Text

De Castro P: **The Role of Current Research Information Systems (CRIS) in Supporting Open Science Implementation: The Case of Strathclyde.** *Informačné Technológie a Knižnice, Special Issue 2018.* 2018; 21–30.
Publisher Full Text

Fecher B, Friesike S: **Open science: One term, five schools of thought.** *Open. Sci.* 2014: 17–47.
Publisher Full Text | Reference Source

Fecher B, Friesike S, Hebing M: **What Drives Academic Data Sharing?.** *PloS One.* 2015; **10**(2): e0118053.
PubMed Abstract | Publisher Full Text

Fyfe A, Coate K, Curry S, *et al.*: **Untangling academic publishing: A history of the relationship between commercial interests, academic prestige and the circulation of research.** *Zenodo.* 2017.
Publisher Full Text

Gewin V: **Data sharing: An open mind on open data.** *Nature.* 2016; **529**(7584): 117–119.
Publisher Full Text

Heller L, The R, Bartling S: **Dynamic Publication Formats and Collaborative Authoring**. In: Bartling S, Friesike S (eds). *Opening Science.* Springer, Cham. 2014.
Publisher Full Text

Huizingh EKRE: **Open innovation: State of the art and future perspectives.** *Technovation.* 2011; **31**(1): 2–9.
Publisher Full Text

Ioannidis JPA: **Why Most Published Research Findings Are False.** *PLOS Med.* 2005; **2**(8): e124.
Publisher Full Text

Kozlov M: **NIH issues a seismic mandate: Share data publicly.** *Nature.* 2022; **602**(7898): 558–559.
Publisher Full Text

Mckiernan EC, Bourne PE, Brown CT, *et al.*: **How open science helps researchers succeed.** *elife.* 2016; **5**(JULY).
PubMed Abstract | Publisher Full Text

Newcomer KE, Hatry HP, Wholey JS: *Handbook of practical program evaluation.* 2015.
Reference Source

Nosek BA, Alter G, Banks GC, *et al.*: **Promoting an open research culture.** *Science (New York, N.Y.).* 2015; **348**(6242): 1422–1425.
PubMed Abstract | Publisher Full Text | Reference Source

Ogungbeni JI, Obiamalu AR, Ssemambo S, *et al.*: **The roles of academic libraries in propagating open science: A qualitative literature review.** *Inf. Dev.* 2018; **34**(2): 113–121.
Publisher Full Text

Pashler H, Wagenmakers E: **Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence?** *Perspect. Psychol. Sci.* 2012; **7**(6): 528–530.
Publisher Full Text

Ponte D, Mierzejewska BI, Klein S: **The transformation of the academic publishing market: multiple perspectives on innovation.** *Electron. Mark.* 2017; **27**(4): 97–100.
Publisher Full Text

Savenije B: **The SPARC initiative: A catalyst for change.** *LIBER Quarterly: The Journal of the Association of European Research Libraries*. 2004; **14**(3–4).
Reference Source

Siyao PO, Whong FM, Martin-Yeboah E, *et al.*: **Academic libraries in four Sub-Saharan Africa countries and their role in propagating open science.** *IFLA J.* 2017; **43**(3): 242–255.
Publisher Full Text | Reference Source

SPARC: SPARC-Who we are: About SPARC. 2022.
Reference Source

Tapfuma MM, Hoskins RG: **Open science disrupting the status quo in academic libraries: A perspective of Zimbabwe.** *J. Acad. Librariansh.* 2019; **45**(4): 406–412.
Publisher Full Text | Reference Source

Tedersoo L, Küngas R, Oras E, *et al.*: **Data sharing practices and data availability upon request differ across scientific disciplines.** *Sci. Data.* 2021; **8**(1): 192.
PubMed Abstract | Publisher Full Text

Tenopir C, Dalton ED, Allard S, *et al.*: **Changes in Data Sharing and Data Reuse Practices and Perceptions among Scientists Worldwide.** *PloS One.* 2015; **10**(8): e0134826.
PubMed Abstract | Publisher Full Text

UNESCO: *UNESCO Recommendation on Open Science*. 2021.
Publisher Full Text

Vicente-Saez R, Martinez-Fuentes C: **Open Science now: A systematic literature review for an integrated definition.** *J. Bus. Res.* 2018; **88**: 428–436.
Publisher Full Text | Reference Source

Wilkinson MD, Dumontier M, Aalbersberg IJ, *et al.*: **The FAIR Guiding Principles for scientific data management and stewardship.** *Sci. Data.* 2016; **3**(1): 160018.
PubMed Abstract | Publisher Full Text | Reference Source

Yoshioka T, Herman G, Yates J, *et al.*: **Genre taxonomy: A knowledge repository of communicative actions.** *ACM Trans. Inf. Syst.* 2001; **19**(4): 431–456.
Publisher Full Text | Reference Source

Open Peer Review

Current Peer Review Status: ✓ ✓

Version 2

Reviewer Report 13 December 2022

https://doi.org/10.5256/f1000research.140940.r157326

✓

Verena Heise
Freelance Open Science Researcher, Gladbeck, Germany

I would like to thank the authors for taking the time to address my comments and provide a revised version of the manuscript. There are still two minor issues that I see. They do not necessarily need to be addressed in a second revised version of the manuscript but it would be good to record them here:

1) The "end-to-end" model
In several places, the program is described as an end-to-end model for research projects and a slightly more cautious phrasing might be more appropriate. The program does offer very useful services across the stages of a research project but a) these services are not necessarily applicable to all types of research and only cover the "ends" of specific types of research, e.g. those that deal with mostly digital data, and b) an important stage of a research project is missing: the one where we develop hypotheses, e.g. through literature reviews.

2) Figure 5
I still wonder how useful the data analysis shown in Figure 5 is. I see two problems here: a) the number of depositors/ accounts will always increase over time if you look at cumulative figures. A better way to assess growth would be to just identify the number of new depositors/ accounts per year. b) The number of accounts is not necessarily related to the number of “users”, i.e. just because someone has an account does not mean that they actively deposit information/ use a service. Maybe there is a better way to determine active tool usage, e.g. via the number of deposits per year, and one would then want to see whether the number of deposits increase per year.

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Open Science, Reproducibility, Meta-research, Biomedical Research

I confirm that I have read this submission and believe that I have an appropriate level of

expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 12 December 2022

https://doi.org/10.5256/f1000research.140940.r157325

✓

Guy A. Rouleau
Department of Human Genetics, The Neuro (Montreal Neurological Institute-Hospital), Department of Neurology and Neurosurgery, McGill University, Montreal, Quebec, Canada
Dylan Roskams-Edris
Tanenbaum Open Science Institute, The Neuro (Montreal Neurological Institute-Hospital), McGill University, Montreal, Quebec, Canada

We are happy with the revised version.

Competing Interests: No competing interests were disclosed.

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Version 1

Reviewer Report 22 July 2022

https://doi.org/10.5256/f1000research.121952.r141852

?

Verena Heise
[1] Freelance Open Science Researcher, Gladbeck, Germany
[2] Freelance Open Science Researcher, Gladbeck, Germany

The article introduces an Open Science program developed at Carnegie Mellon University Libraries that supports researchers with tools, services and training to develop open and reproducible workflows. Additionally, it describes a framework for assessing the success of this program. Both of these aspects are very interesting for other institutions that are engaged in promoting open and reproducible research and would like to or are in the process of setting up similar services.

The article is very informative but nevertheless I have a couple of major and minor comments that I hope will be helpful to the authors:

Major items

1. **The "end-to-end" model**
   In several places, the program is described as an end-to-end model for research projects and I wonder whether a slightly more cautious phrasing might be more appropriate. The program does offer very useful services across the stages of a research project but a) these services are not necessarily applicable to all types of research and only cover the "ends" of specific types of research, e.g. those that deal with mostly digital data, and b) an important stage of a research project is missing: the one where we develop hypotheses, e.g. through literature reviews. The latter might be part of the services offered but it is currently not mentioned in the manuscript.

2. **Citizen Science**
   This is mentioned in a couple of places (e.g. Figure 1) as a service supported by the OSDC (Open Science & Data Collaborations) but there is very little information on it in the paper. Could you clarify what the role of OSDC is with respect to Citizen Science? And just out of curiosity: Citizen Science can be included in research projects at all stages, so I was wondering why it only seems to be applicable to the data collection and analysis stage in Figure 1?

3. **Figure 1**
   I was wondering whether it might make sense to remove the grey boxes since the article mostly focuses on the OSDC. Two items that are not currently explained in the legend or the following table are the DMPTool and OpenRefine, so it would be great to add some more info on these.

4. **Table 2**
   It would be quite helpful to explain the categories in the table in a bit more detail. For example, what is the difference between Library sponsored OS programs and Library Open Research programming?

5. **Figure 2**
   Activities: I'm not sure what the Emerald Cloud Lab integration and Reproducibility MiniSeries refer to.
   Outputs: I guess the training sessions are related to the tools themselves? Would it make sense to have these in the second category together with the other workshops?
   Weekly office hours: I was wondering why this is relevant. I guess the question would be how many people actually come to get support during the office hours?
   Outcomes (this is a conceptual point, so does not necessarily need to be addressed for this paper): I'm surprised they are defined as outcomes for users, not the OSDC. This might just be the way the strategic goals are defined for OSDC and it's absolutely fine if the figure reflects these. I wonder whether that's the best way to define outcomes for the OSDC because it's probably difficult to assess whether the OSDC has been successful with that selection of outcomes, especially the medium and long-term ones. For example, shifts in research culture require much more than just availability of training, tools and services, so it's not only down to the OSDC whether those cultural shifts actually materialise.

6. **Figure 3**
   As far as I can see the acronyms EPP, CIT, CYLAB are not covered in the figure legend. And another conceptual comment that does not necessarily need to be addressed here: it might be interesting to look at user numbers in relation to size of the institutes, e.g. percentage of faculty using services, so that institutes of different sizes can be compared quite easily.

7. **Figure 5**
   This is a very stupid question but it would be great to clarify that these are new items/ new registrations per year and not the total numbers (e.g. of accounts) in that year, which would be the sum of accounts from the previous year plus new accounts?

Minor items

Introduction

1. "This trend has been a response to changes in the funding and publishing landscape, the nature of research collaboration, the emergence of digital research infrastructures and cultural shifts in scientific practice" - I wonder whether a couple of references might help here or a some examples of the changes this refers to. I guess it refers to changes such as open access/ open data mandates but it would be helpful to see what the authors mean.

2. Definition of Open Science: I wonder whether it might be helpful to include the definition used in the UNESCO recommendation (https://en.unesco.org/science-sustainable-future/open-science/recommendation) because it's a definition of Open Science that is gaining quite a bit of traction, especially among policy makers.

Discussion

1. I was wondering why the first paragraph is mostly about data sharing and does not cover Open Science practices more broadly since OSDC is about more than just data sharing.

2. The sentence that ends with "prevents us from being able to make a clear value proposition to researchers for whom productivity, efficiency, and impact are the most important factors" might need rephrasing. I would hope that the impact of reproducible workflows and data management are obvious to many researchers because they have clear impacts on efficiency and impact (e.g. see higher citation rates of openly available papers and data). I understand that it can be difficult to quantify some of those aspects but nevertheless there is a clear value proposition for using open/ reproducible workflows and a number of papers have dealt with the selfish reasons for working in a more open/ reproducible way (e.g. McKiernan paper cited already
   and https://genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0850-7)

3. There are a couple of sentences about data management, e.g. this one: *"the current data and the 5W1H metrics framework will serve as a baseline to develop a strategy for data management in the future to guide data collection, update, and analysis"*. I think this refers to management of user data related to services and not research data itself and it would be great if that could be rephrased slightly to make the distinction more obvious.

References

1. Markowetz F: Five selfish reasons to work reproducibly.*Genome Biol*. 2015; **16**: 274 PubMed

Abstract | Publisher Full Text

Is the background of the case's history and progression described in sufficient detail?
Partly

Is the work clearly and accurately presented and does it cite the current literature?
Yes

If applicable, is the statistical analysis and its interpretation appropriate?
Yes

Are all the source data underlying the results available to ensure full reproducibility?
Partly

Are the conclusions drawn adequately supported by the results?
Yes

Is the case presented with sufficient detail to be useful for teaching or other practitioners?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Open Science, Reproducibility, Meta-research, Biomedical Research

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.

Author Response 30 Nov 2022

Huajin Wang

Thank you Dr. Verena Heise for your thorough review of our work. Please see our point-by-point response below.

Major items

1. "*The "end-to-end" model. In several places, the program is described as an end-to-end model for research projects and I wonder whether a slightly more cautious phrasing might be more appropriate. The program does offer very useful services across the stages of a research project but a) these services are not necessarily applicable to all types of research and only cover the "ends" of specific types of research, e.g. those that deal with mostly digital data, and b) an important stage of a research project is missing: the one where we develop hypotheses, e.g. through literature reviews. The latter might be part of the services offered but it is currently not mentioned in the manuscript.*"

We use "end-to-end" to describe a service model for open science with the intention of serving all stages of the research life cycle, but the implementation is a gradual process and the current composition reflects the demands we see at Carnegie Mellon and the tools and expertise that are readily available at the moment. We thank the reviewer for pointing out non-digital data and literature reviews as gaps in our service and will keep them in mind as our program matures. CMU Libraries indeed provide systematic review and evidence synthesis services to help researchers develop hypotheses from literature and increase research rigor and reproducibility, independent from the OSDC program. It would be important to think about how it overlaps and interacts with the OSDC program going forward.

2. "*Citizen Science. This is mentioned in a couple of places (e.g. Figure 1) as a service supported by the OSDC (Open Science & Data Collaborations) but there is very little information on it in the paper. Could you clarify what the role of OSDC is with respect to Citizen Science? And just out of curiosity: Citizen Science can be included in research projects at all stages, so I was wondering why it only seems to be applicable to the data collection and analysis stage in Figure 1?*"

Even though citizen science is a direction we aspire to support in the future, we have only started to explore how to support it with our services and so far have only offered a few workshops. In the revised manuscript we have now removed citizen science from Figure 1 and Figure 2 to avoid confusion.

3. "*Figure 1. I was wondering whether it might make sense to remove the grey boxes since the article mostly focuses on the OSDC. Two items that are not currently explained in the legend or the following table are the DMPTool and OpenRefine, so it would be great to add some more info on these.*"

We feel that keeping services that are within broader open science but not administered by OSDC in the gray boxes would help readers understand the interaction and relationship between related services in a larger context, as different libraries or universities may structure them differently. We now added additional text to explain this intention (page 3, 3rd paragraph). In addition, we have defined DMPTool and OpenRefine in the figure legend of Figure 1.

4. "*Table 2. It would be quite helpful to explain the categories in the table in a bit more detail. For example, what is the difference between Library sponsored OS programs and Library Open Research programming?*"

We have added more information to the Table 2 legend to define the column names more clearly.

5. "Figure 2.
Activities: I'm not sure what the Emerald Cloud Lab integration and Reproducibility MiniSeries refer to."

We have added to Figure Legend to briefly describe Emerald Cloud Lab and Reproducibility MiniSeries.

"Outputs: I guess the training sessions are related to the tools themselves? Would it make sense

to have these in the second category together with the other workshops?"

We'd like to keep tool-specific trainings separate from general skill-building workshops because these activities are led by different teams and often attract different audiences.

"Weekly office hours: I was wondering why this is relevant. I guess the question would be how many people actually come to get support during the office hours?"

We offer weekly office hours to support researchers who have data related questions. The reviewer brings out a great question about participation, unfortunately we haven't systematically tracked the number of participants especially during the pandemic when office hours become online. We will start tracking these numbers in the next iteration.

"Outcomes (this is a conceptual point, so does not necessarily need to be addressed for this paper): I'm surprised they are defined as outcomes for users, not the OSDC. This might just be the way the strategic goals are defined for OSDC and it's absolutely fine if the figure reflects these. I wonder whether that's the best way to define outcomes for the OSDC because it's probably difficult to assess whether the OSDC has been successful with that selection of outcomes, especially the medium and long-term ones. For example, shifts in research culture require much more than just availability of training, tools and services, so it's not only down to the OSDC whether those cultural shifts actually materialise."

Thanks for asking this conceptual question. We choose to use user behavior to define our outcomes because the goal of the OSDC program is to drive behavior change in researchers.

6. "*Figure 3. As far as I can see the acronyms EPP, CIT, CYLAB are not covered in the figure legend. And another conceptual comment that does not necessarily need to be addressed here: it might be interesting to look at user numbers in relation to size of the institutes, e.g. percentage of faculty using services, so that institutes of different sizes can be compared quite easily."*

The acronyms have now been defined in the Figure Legend. We thank the reviewer for suggesting using proportion to represent usage to normalize for department size. This is a great suggestion and would be something we'd like to implement in the future. However, we don't have data on the total number of students and faculty in each department at the moment. Additionally, our users are not only faculty but also staff, students, and postdocs, making it even harder to obtain an accurate total number per department.

7. "*Figure 5. This is a very stupid question but it would be great to clarify that these are new items/ new registrations per year and not the total numbers (e.g. of accounts) in that year, which would be the sum of accounts from the previous year plus new accounts?"*

The values presented in Figure 5 are cumulative numbers, i.e., the sum of accounts from the previous year plus new accounts. This is now clarified in the figure legend.

Minor Items

Introduction:

1. ""This trend has been a response to changes in the funding and publishing landscape, the nature of research collaboration, the emergence of digital research infrastructures and cultural shifts in scientific practice" - I wonder whether a couple of references might help here or a some examples of the changes this refers to. I guess it refers to changes such as open access/ open data mandates but it would be helpful to see what the authors mean."

Thank you for this suggestion. We have added parentheticals with examples of each of the changes noted in the sentence along with the following references.

1. Kozlov, M. (2022). NIH issues a seismic mandate: Share data publicly. Nature, 602(7898), 558–559. https://doi.org/10.1038/d41586-022-00402-1
2. Davidson LA. The End of Print: Digitization and Its Consequence—Revolutionary Changes in Scholarly and Social Communication and in Scientific Research. International Journal of Toxicology. 2005;24(1):25-34. doi:10.1080/10915810590921351
3. Ponte, D., Mierzejewska, B.I. & Klein, S. The transformation of the academic publishing market: multiple perspectives on innovation. Electron Markets 27, 97–100 (2017). https://doi.org/10.1007/s12525-017-0250-9
4. Heller, L., The, R., Bartling, S. (2014). Dynamic Publication Formats and Collaborative Authoring. In: Bartling, S., Friesike, S. (eds) Opening Science. Springer, Cham. https://doi.org/10.1007/978-3-319-00026-8_13
5. Fyfe, A., Coate, K., Curry, S., Lawson, S., Moxham, N., & Røstvik, C. M. (2017). Untangling academic publishing: A history of the relationship between commercial interests, academic prestige and the circulation of research.
6. Cummings, J. N., & Kiesler, S. (2014). Organization theory and the changing nature of science. Journal of Organization Design, 3(3), 1-16.
7. Huizingh, E. K. R. E. (2011). Open innovation: State of the art and future perspectives. Technovation, 31(1), 2–9. https://doi.org/10.1016/j.technovation.2010.10.002

2. "*Definition of Open Science: I wonder whether it might be helpful to include the definition used in the UNESCO recommendation (https://en.unesco.org/science-sustainable-future/open-science/recommendation) because it's a definition of Open Science that is gaining quite a bit of traction, especially among policy makers."*

We have added the UNESCO definition to the text.

Discussion

1. "I was wondering why the first paragraph is mostly about data sharing and does not cover Open Science practices more broadly since OSDC is about more than just data sharing."

Thanks for pointing out the terminology. We intend to use the term "data" broadly to refer to all research outputs, including data, code, workflow, and more. We added a sentence in the first paragraph of the Discussion and future directions section to clarify.

2. "*The sentence that ends with "prevents us from being able to make a clear value proposition to researchers for whom productivity, efficiency, and impact are the most important factors" might need rephrasing. I would hope that the impact of reproducible workflows and data management are obvious to many researchers because they have clear impacts on efficiency and impact (e.g.*

see higher citation rates of openly available papers and data). I understand that it can be difficult to quantify some of those aspects but nevertheless there is a clear value proposition for using open/ reproducible workflows and a number of papers have dealt with the selfish reasons for working in a more open/ reproducible way (e.g. McKiernan paper cited already and https://genomebiology.biomedcentral.com/articles/10.1186/s13059-015-0850-7)"

Even though the value of using open/ reproducible workflows have been shown to benefit researchers from a “selfish” perspective, many researchers are still unconvinced or unaware, and the mainstream research culture remains relying on quantifiable metrics to evaluate productivity and impact. We have revised the sentence that the reviewer refers to clarify our view on the value proposition.

3. "*There are a couple of sentences about data management, e.g. this one: "the current data and the 5W1H metrics framework will serve as a baseline to develop a strategy for data management in the future to guide data collection, update, and analysis”. I think this refers to management of user data related to services and not research data itself and it would be great if that could be rephrased slightly to make the distinction more obvious.*"

We agree with the reviewer that the terminology might be confused with “research data management”. We have now changed it to “user data management”.

Competing Interests: No competing interests were disclosed.

Reviewer Report 19 July 2022

https://doi.org/10.5256/f1000research.121952.r137000

?

Guy A. Rouleau
[1] Department of Human Genetics, The Neuro (Montreal Neurological Institute-Hospital), Department of Neurology and Neurosurgery, McGill University, Montreal, Quebec, Canada
[2] Department of Human Genetics, The Neuro (Montreal Neurological Institute-Hospital), Department of Neurology and Neurosurgery, McGill University, Montreal, Quebec, Canada
Dylan Roskams-Edris
[1] Tanenbaum Open Science Institute, The Neuro (Montreal Neurological Institute-Hospital), McGill University, Montreal, Quebec, Canada
[2] Tanenbaum Open Science Institute, The Neuro (Montreal Neurological Institute-Hospital), McGill University, Montreal, Quebec, Canada

This article is about the development of a program created by the Carnegie Mellon University Libraries to support Open Science practice at CMU, in particular by offering educational and support services around various platforms.

A good paper describing an important initiative. Their development of a unified program, as well as the generation of useful metrics, are an important contribution to the development of Open Science at institutions.

As discussed in the comments below, there is some information they could add about the history of open science/open research that would add clarity to the text, as well as a few minor revisions that should be made to supply necessary information in figures.

1. Greater discussion of the historical role of libraries in open research practices.

In the introduction the authors use the sentence *"As Open Science has matured, academic libraries have also entered this space leveraging the natural alignment with existing services and principles related to information access and dissemination."* This sentence somewhat mischaracterizes the historical role of libraries in the Open Science movement. To a significant extent open science emerged from the Open Access movement, and libraries have been heavily involved in that movement since at least 1997 when the Association of Research Libraries founded Scholarly Publishing and Academic Resources Coalition (SPARC). It would be more accurate to say that the role of libraries has always been important to the movement to open science, but that they have more recently started to play an instrumental role beyond open access to publications.

2. Earlier introduction to the relationship between reproducibility and open science.

The authors introduce the concept of reproducibility on page 5, both in the google search string they used to quality control their search of peer institutions and in the first sentence in the section on "Program Implementation." To readers versed in Open Science this connection will likely come as no surprise, but to the less versed reader the reference to reproducibility comes somewhat out of the blue. Why, for example, should that be included in a search for Open Science relevant programs? A brief explanation, perhaps in the introduction near where they address the "five schools of thought" would make their search strategy and the other references to reproducibility clearer. This observation is also relevant to the comment below on the history of the field of psychology and its relevance to the development of Open Science.

3. Further historical context around the field of psychology.

Similar to the comment on the history of libraries, and related to the comment on reproducibility, there are several ways a brief discussion of the history of the relationship between psychology and open science would add clarity to the paper. Unlike with libraries' advocacy for Open Access, which fed into Open Science and which were primarily concerned with rising subscription fees and the efficient dissemination of knowledge, psychology researchers have been a major force in the history of Open Science primarily as an aid to reproducibility and replicability.

Adding some content about this history would help both to make clear the relationship between reproducibility/replicability and open science—indeed psychologists were the first to identify it as a crisis (https://journals.sagepub.com/doi/10.1177/1745691612465253) – but would also contribute to why the CMU Psychology department has such a large number of users (see Fig. 3).

4. Adding some discussion of the difficulties in sharing sensitive human data and what this might imply for the data collected.

While the authors address the difficulties posed by ethical constrains on sharing sensitive human data in the section “Discussion and future directions”, acknowledging this particular difficulty in earlier sections where they discuss the data around departmental users would add important interpretive context for why departments that deal with this kind of data regularly (for example the Centre for the Neural Basis of Cognition) might show fewer users.

5. Acronyms used in Figure 3.

The “EPP” acronym on the x-axis in Figure 3 is not explained in the text below the figure, though it likely refers to the department of Engineering and Public Policy. Also, the use of “iii” rather than “III” for the Integrated Innovation Institute isn’t consistent with the capitalization of the other acronyms, though there may be a reason for this I am not aware of.

6. Note on important policy changes that may impact metrics

While it may not be possible in the context of this paper, it would be useful in future work for the authors to develop a way of assessing important events at the institutional, state, and national levels that impact the uptake of Open Science practices. The emergence of, for example, the NIH policy on data management and sharing may have a significant impact on their metrics in the absence of activity by the OSDC itself. They do address this to some extent in the proposed discussions with “superusers” in the section “Applications of the logic model and 5W1H framework”, and it will be very difficult to give a fully causal account, but at least acknowledging this confound further, or creating some automated system through google alerts or scraping relevant twitter hashtags/keywords (e.g., “Open Science” + “Policy” + filtering for CMU, state, and/or national users) could help provide such a timeline. If, for example, they see an increase in some metrics in the absence of events in the timeline, they can be more certain that the effect is endogenous.

References

1. Pashler H, Wagenmakers EJ: Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence?. *Perspect Psychol Sci*. 2012; **7** (6): 528-30 PubMed Abstract | Publisher Full Text

Is the background of the case’s history and progression described in sufficient detail?

Partly

Is the work clearly and accurately presented and does it cite the current literature?

Yes

If applicable, is the statistical analysis and its interpretation appropriate?

Yes

Are all the source data underlying the results available to ensure full reproducibility?

Partly

Are the conclusions drawn adequately supported by the results?
Yes

Is the case presented with sufficient detail to be useful for teaching or other practitioners?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Neuroscience, Genetics, Neurology, Open Science

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

Author Response 30 Nov 2022

Huajin Wang

We thank Drs. Rouleau and Roskams-Edris for your thoughtful comments and suggestions. Please see our point-by-point response below.

1. "*Greater discussion of the historical role of libraries in open research practices. In the introduction the authors use the sentence "As Open Science has matured, academic libraries have also entered this space leveraging the natural alignment with existing services and principles related to information access and dissemination." This sentence somewhat mischaracterizes the historical role of libraries in the Open Science movement. To a significant extent open science emerged from the Open Access movement, and libraries have been heavily involved in that movement since at least 1997 when the Association of Research Libraries founded Scholarly Publishing and Academic Resources Coalition (SPARC). It would be more accurate to say that the role of libraries has always been important to the movement to open science, but that they have more recently started to play an instrumental role beyond open access to publications.*"

Thank you for this insightful comment. We agree that the true characterization of libraries in the open science movement is more significant than previously stated. We have added a paragraph to the Introduction that discusses SPARC and the role of libraries and the open access movement as an early catalyst for open science.

2. "*Earlier introduction to the relationship between reproducibility and open science. The authors introduce the concept of reproducibility on page 5, both in the google search string they used to quality control their search of peer institutions and in the first sentence in the section on "Program Implementation." To readers versed in Open Science this connection will likely come as no surprise, but to the less versed reader the reference to reproducibility comes somewhat out of the blue. Why, for example, should that be included in a search for Open Science relevant programs? A brief explanation, perhaps in the introduction near where they address the "five schools of thought" would make their search strategy and the other references to reproducibility clearer. This observation is also relevant to the comment below on the history of the field of psychology and its relevance to the development of Open Science.*"

We have added a separate paragraph to address this and the following comment. The paragraph introduces the reproducibility crisis and highlights its role in facilitating new practices of open science.

3. "*Further historical context around the field of psychology. Similar to the comment on the history of libraries, and related to the comment on reproducibility, there are several ways a brief discussion of the history of the relationship between psychology and open science would add clarity to the paper. Unlike with libraries' advocacy for Open Access, which fed into Open Science and which were primarily concerned with rising subscription fees and the efficient dissemination of knowledge, psychology researchers have been a major force in the history of Open Science primarily as an aid to reproducibility and replicability. Adding some content about this history would help both to make clear the relationship between reproducibility/replicability and open science—indeed psychologists were the first to identify it as a crisis (https://journals.sagepub.com/doi/10.1177/1745691612465253) – but would also contribute to why the CMU Psychology department has such a large number of users (see Fig. 3).*"

See response above for Item 2, which also addresses this comment.

4. "*Adding some discussion of the difficulties in sharing sensitive human data and what this might imply for the data collected. While the authors address the difficulties posed by ethical constrains on sharing sensitive human data in the section "Discussion and future directions", acknowledging this particular difficulty in earlier sections where they discuss the data around departmental users would add important interpretive context for why departments that deal with this kind of data regularly (for example the Centre for the Neural Basis of Cognition) might show fewer users.*"

We have no evidence that ethical constraints are a factor that CNBC has fewer users. We think it's more likely due to the smaller department size. We'd like to eventually use proportion to represent user size in each department but at the moment we do not have data on department sizes. See also the response to Reviewer #2.

5. "*Acronyms used in Figure 3. The "EPP" acronym on the x-axis in Figure 3 is not explained in the text below the figure, though it likely refers to the department of Engineering and Public Policy. Also, the use of "iii" rather than "III" for the Integrated Innovation Institute isn't consistent with the capitalization of the other acronyms, though there may be a reason for this I am not aware of.*"

Thanks for pointing out the errors in acronyms. "EPP" indeed refers to Department of Engineering and Public Policy. This has now been added in the figure legend. As for the acronym of Integrated Innovation Institute (iii), lowercase letters were intentionally used by the department, presumably as a design choice.

6. "*Note on important policy changes that may impact metrics. While it may not be possible in the context of this paper, it would be useful in future work for the authors to develop a way of assessing important events at the institutional, state, and national levels that impact the uptake of Open Science practices. The emergence of, for example, the NIH policy on data management and sharing may have a significant impact on their metrics in the absence of activity by the OSDC*

itself. They do address this to some extent in the proposed discussions with "superusers" in the section "Applications of the logic model and 5W1H framework", and it will be very difficult to give a fully causal account, but at least acknowledging this confound further, or creating some automated system through google alerts or scraping relevant twitter hashtags/keywords (e.g., "Open Science" + "Policy" + filtering for CMU, state, and/or national users) could help provide such a timeline. If, for example, they see an increase in some metrics in the absence of events in the timeline, they can be more certain that the effect is endogenous."

It's a great point by the reviewer that in order to draw a conclusion on causal effect, one would need to factor in major confounding factors, such as external policy changes. We think that adding a "date" field across our data collection process will help to address this issue as it would enable us to associate changes in usage with external influencing factors. The lack of a "date" field was addressed in Limitations of current data sources and future data management strategy section; we now added a sentence to specifically draw attention to confounding from external signals.

Competing Interests: No competing interests were disclosed.