RESEARCH PAPER

Playing Well on the Data FAIRground: Initiatives and Infrastructure in Research Data Management

Danielle Descoteaux[1†], Chiara Farinelli[2], Marina Soares e Silva[2] & Anita de Waard[1]

[1]Elsevier, Inc, 50 Hampshire St, Cambridge, MA 02139, USA

[2]Elsevier, B.V., Radarweg 29, 2043NX Amsterdam, The Netherlands

Keywords: Open data; Data sharing; Data citation; Open research

Citation: D. Descoteaux, C. Farinelli, M.S.e Silva & A. de Waard. Playing well on the data FAIRground: Initiatives and infrastructure in research data management. Data Intelligence 1(2019), 350-367. doi: 10.1162/dint_a_00020
Received: April 23, 2019; Revised: June 10, 2019; Accepted: June 28, 2019

ABSTRACT

Over the past five years, Elsevier has focused on implementing FAIR and best practices in data management, from data preservation through reuse. In this paper we describe a series of efforts undertaken in this time to support proper data management practices. In particular, we discuss our journal data policies and their implementation, the current status and future goals for the research data management platform Mendeley Data, and clear and persistent linkages to individual data sets stored on external data repositories from corresponding published papers through partnership with Scholix. Early analysis of our data policies implementation confirms significant disparities at the subject level regarding data sharing practices, with most uptake within disciplines of Physical Sciences. Future directions at Elsevier include implementing better discoverability of linked data within an article and incorporating research data usage metrics.

1. BACKGROUND AND MOTIVATION

The FAIR (*findable, accessible, interoperable* and *reusable*) Data Principles argue that standardized data management is "the key conduit leading to knowledge discovery and innovation" [1]. The storage, preservation, accessibility and citation of research data is an essential aspect of creating rigorous and reusable scholarly output. This means that a previously ancillary artefact – raw research data – is now becoming a usable and analyzable scholarly output in its own right. Data repositories, publishers, funders, institutions and scholars have been working through many venues to develop standards, goals for

† Corresponding author: Danielle Descoteaux (E-mail: d.descoteaux@elsevier.com).

interoperability and requirements for metadata and data permanence to allow storage and access to this growing body of publicly available research data, through such organizations as the Research Data Alliance (RDA)[①]. Defining, meeting, and raising the standards for open science, including best practices for research data management, is generally a community effort with global stakeholders. At the 2016 G20 Summit in Hangzhou, the G20 leaders declared their support to FAIR data principles being implemented to promote open science and to enable appropriate access to publicly funded research results [2]. Similarly, stakeholder groups such as CODATA and the European Open Science Cloud are actively engaged in enabling FAIR Data Principles throughout the scholarly workflow [3]. In specific domains, there are tailored efforts to focus the research data management (RDM) practices of an entire community around these standards. For instance, in the Earth and Space Sciences, a coalition of groups representing the international science community was convened by the American Geophysical Union (AGU), to develop standards to connect researchers, publishers and data repositories in these disciplines to enable FAIR data [4].

Despite these ambitious goals, research data management practices are still heterogeneous both geographically and across different areas of research. While most researchers agree that reusing data from others would benefit their research, data sharing is not widespread and researchers report having little experience with data sharing. According to the most recent Open Data Report [5], 73% of academics surveyed said that having access to published research data would benefit their own research, while only 64% are willing to allow others to access their research data. One of the reasons for this disconnect is that despite the growth of information on the importance of data sharing, most scholarly research is still aimed at publishing papers in reputable journals. Sharing and publishing data is not perceived by authors as a priority of their institutions ([5, 6]). It's for this reason we see a natural opportunity for scholarly publishers to take an active role. Manuscript submission, which prompts authors to provide information about their research, is a natural moment to bring research data together with an article: to require and enable data sharing, allow data annotation and connect RDM tools and standards to the publishing workflow. Creating these pathways to open data enables the raw data and the paper to be linked together, without extraneous and new workflows for researchers. We therefore also actively support and are enabling proper Data Citation Practices, as outlined by the Force11 Data Citation Guidelines [7] and have helped lead a convergence of science publishers on modes and systems of data citation [8]. Proper data citation practices can support citation counts, downloads and views of data sets, which can act as important metrics to establish review and reuse of data and serve to motivate the scholarly community to share and publish their data.

[①] NB: For the purposes of this article, data sharing will largely be defined as how data are saved, shared, cited and trusted, with each of these components incorporating several layers. Moreover, we will use "research data" interchangeably to encompass raw data, code, software and other research objects. We recognize that different communities will focus on the sharing and creation of different research objects and it is not our intention to impose a definition of those digital research output objects. There has been widespread agreement on standards that come from such discussions with the Research Data Alliance, Force 11, and FAIRsharing, with nuanced understanding about different *kinds* of data and the domain-specific repositories that might host them.

Below, we discuss a series of initiatives taken largely over the last few years to facilitate proper practices for data deposition, curation and discovery. This paper is organized as follows: first, we discuss the overall principles behind our RDM practices and tools (2.1); then, we discuss a series of efforts that we have engaged in, together with the community of stakeholders, over the past five years, and the practical outcomes we have seen from these efforts (2.2 – 2.6). Lastly, we discuss the implications of these efforts, and some thoughts on moving forward with this important challenge.

2. PROMOTING RESEARCH DATA MANAGEMENT AT ELSEVIER

2.1 Overall Vision on Research Data Management

Over the past five years, we have developed multiple initiatives aimed at promoting data management and sharing, discussed in the rest of this section. Throughout these efforts, we have been driven by an overarching idea of a "data Maslow hierarchy", as depicted in Figure 1 below from [9]. The idea behind this figure is that all components of data sharing support the "highest" goal (that of data reuse), but this goal cannot be obtained unless the "lower-level components" are in place, i.e. data must be stored, before it can be accessed; it must be accessible, to be reused. In our educational outreach (see e.g. Researcher Academy [10]), we consistently emphasize that good data management starts in the research planning phase, and an important role is played by a fruitful interaction with data librarians, data stewards and curators and others at the researchers' home institution or in their specific community of practice.

In the remainder of this section, we will discuss a series of efforts which we have undertaken to support this vision:

2.2. Data citation and deposition guidelines for our journals – supporting storage, preservation, access, discovery and citation;
2.3. Data Linking – supporting Data Discovery
2.4. Research Data Management Infrastructure – supporting storage, preservation, access, discovery and citation;
2.5. Data Journals – supporting the evaluation of Data Quality and Data Reproducibility.

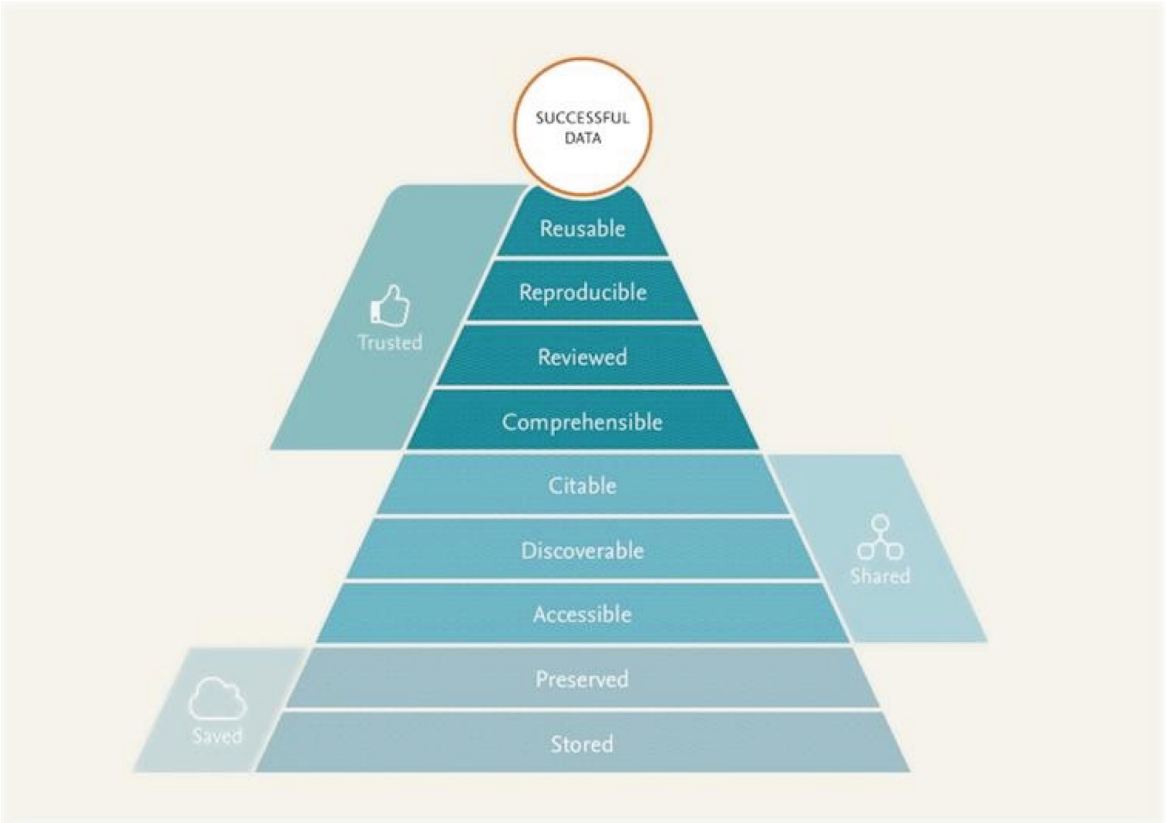

Figure 1. The "data Maslow hierarchy" visualizing the components of data sharing [11].

2.2 Research Data Deposition and Citation: The TOP Guidelines

As a first step to address the growing demand for guidance and tools to address calls for transparency, openness and reproducibility of research, we implemented a series of data citation guidelines and support to all applicable journals (approximately 2,200) using standard reference styles in 2016 [12]. In September 2017, we introduced a five-tiered data sharing policy across more than 1,700 of these titles. These policies were developed internally, in tandem with the Transparency and Openness (TOP) guidelines that were established by the Center for Open Science (to which Elsevier was a signatory [13]. The policy options are:

A) authors are **encouraged** to deposit their data in a repository, and cite it in their article;
B) Authors are **encouraged** to (i) deposit their data in a repository, cite **and** link to the data set in their article, **or** (ii) provide an Author Statement explaining why data cannot be shared;
C) Authors are **required** to (i) deposit their data in a repository, cite and link to the data set in the article, **or** (ii) provide an Author Statement explaining why data cannot be shared;

D) Authors are **required** to deposit their data in a repository, cite and link to the data set in the article (no option of an Author Statement about why data cannot be shared)
E) Authors are **required** to deposit their data in a relevant repository, cite and link to the data set in their article, and peer reviewers are asked to **review** the data prior to publication [14].

In our initial roll-out, of the journals eligible for the policy implementation, the majority were set with the default of Option B (Table 1). (Ineligible journals include case reports, which include little additional data or potentially sensitive patient data whose risk of exposure outweighs the sharing benefit, review journals, or journals not on a centralized editorial system; in this last example, off-system journals could and often do have data policies, but they are not enforceable, or trackable, through an editorial platform). Though Option B does not require data deposition, foregrounding data policy at the level of the article submission process is intended to heighten researcher awareness of best research data management practice. Moreover, this range of policies was designed with a range of communities and users in mind; publishers and editors were able to use these as starting points of discussion to apply to individual journals, so that a journal's data policy would be most informed by the existing practice within a specific research community. We had previously conducted a survey among 113 editors from a range of disciplines in August 2017, exploring attitudes and perceptions about data policies. Most editors considered their authors would be willing to share research data at time of publication (56 respondents answered "moderately willing", 20 respondents answered "very willing"; only 5 editors thought their authors would be "not at all" willing). Not surprisingly, most editors considered the most effective way to share data (prior to data policy implementation) was to either include them within a journal article (19 respondents) or as supplementary data to an article (56 respondents), in an appropriate data repository chosen by only 37, or 32.4% of respondents. Data sharing policies, then, also became an educational tool about evolving standards of data management. The policies were implemented on journals that were on one of our two editorial systems, EES or EVISE; however, this functionality was implemented on EES slightly later than it was on EVISE.

Table 1. Summary of results of implementation of data sharing policies at Elsevier, 2017–2018. Over 2,200 journals were eligible for data sharing roll out and their editors consulted for the advised policy to be instated.

	Policy A	Policy B	Policy C	Policy D	None/"Parked"	Opt Out
Life Sciences	0.2%	94.3%	1.6%	0.2%	2.8%	0.9%
Physical Sciences	0.3%	82.6%	9.1%	0%	7.6%	0.4%
Health and Medical Sciences	0.3%	55.8%	0.1%	0%	42.2%	1.6%

The majority of the "none/parked" column is due to journals that were ineligible for the policy rollout due to editorial platform transitions (e.g. moving from EES to EVISE). The opt-out policies were largely due to community sensitivities expressed by editors.

Data sharing can occur at multiple points of the research workflow process; often, it occurs outside of the publication workflow of a paper, meaning that data might be shared to a repository before or after publication of any corresponding research articles. Given our role, we looked to optimize data sharing at point of submission, implementing these data policies while also enabling our editorial systems to

accommodate their requirements. Our two main editorial platforms, EVISE and EES, were updated so that authors at point of submission could comply with the policies by providing either the DOI, PID, or accession number of their underlying data already stored on an repository, or by uploading data directly to Mendeley Data (on which, more below in Section 2.3) as a co-submission, or providing the Research Data Availability statement directly with their article submission. At present, this statement is explicitly oriented toward an explanation of why data cannot be shared; this is a potential area for further investigation to see if modifications to the requested author statement might lead to greater data sharing at point of submission.

In approximately 18 months, only 3% of all articles handled in EES included a link to shared data and 4% of those in EVISE did, out of nearly 220,000 articles handled on EES and over 1.6 million articles handled on EVISE (it is possible that this 1% discrepancy in uptake between the two editorial systems is due to EES's functionality for the data policy implementation coming on board later than EVISE). Exploring the results by subject area, we do see meaningful disparities among them. Over half of the papers that shared data at point of submission in both systems were in the Physical Sciences (55% EES, 52% EVISE). Health and Medical Sciences, for which data sets include "Clinical Trials", followed at 16% in both EES and EVISE. The areas of greatest uptake in the Physical Sciences include Energy and Earth Sciences, Environmental, Agricultural, and Aquatic Sciences, and Applied Bioscience. To build upon this *de facto* trend, the Energy and Earth Sciences portfolio in 2018 changed the default policy for the majority of the journals to Option C, which has also had some correlation with deposition to Mendeley Data (see following subsection).

In the months ahead we expect that more journals adopt stricter data policies. We are in discussion with communities that have already signaled interest in pushing toward more transparency and increase the rate of submission of research data up front. We will continue to offer communities the data sharing policies that are the best fit for them, with the guiding idea that transparency will continue to ramp up and become established across multiple disciplines. This means that our policies can scale from encouraging data deposition (or a statement) to eventually requiring deposition, based on continuing dialog with key data standards organizations, workshops, and attunement to funder policies.

2.3 Linking and Finding Data

To further support and improve data sharing practices Elsevier implemented Database Linking, working with a multidisciplinary range of repositories and guiding authors through the best practices for correctly citing data. We discuss three initiatives here: the Database Linking Tool, including a cross-stakeholder initiative called "Scholix", the ORCID Link, and the Data Search tool.

Database Linking and Scholix

Enabling links between a paper and a data set at submission is just one of several potential points at which a researcher might connect the two research objects. This allows for a more automated workflow, so that a paper in review and production always has a link to the relevant data set. Such link follows the

article through publication and enables bidirectional links for reader which makes it a preferred workflow. However, we are also able to support post-publication links to data sets where needed.

The Database Linking Tool [15] includes about 80 repositories, including examples such as DRYAD, PANGAEA and HEPData. Database Linking creates a bidirectional link between articles and data repositories, such that data can easily be discovered and accessed. To link to a database, when submitting an article the author can simply include a data DOI or PID or indicate in which repository the data have been deposited. We work with a multidisciplinary range of specific repositories and guiding authors through the best practices about correctly citing data on them; e.g. for repositories with accession numbers or identifiers instead of DOIs we provide summarized instructions by discipline. We require that repositories interested in linking with Elsevier must provide a description and link to the general information about their holdings, use, and policies, as well as links to their formatting information, citation information, XML scheme and coding, and that the repositories themselves are FAIR compliant even if not certified by CoreTrustSeal.

When an article with an associated data set is published on ScienceDirect, a link to the repository (and repository logo) is added to the article, making it easy for the reader to find and access the data. For journal articles to meet the minimum of findability (*F* of *FAIR)* a link pointing to available data must be provided by the authors (or the data availability statement in its stead).

To further improve linking between research literature and research data, a community and multi-stakeholder initiative was created under the name of Scholix Framework (SCHOlarly LInk eXchange) [16]. This effort, of which Elsevier has been one of the initiators, is a conceptual framework for interoperability developed in consensus between data centers, publishers, CrossRef, DataCite, OpenAIRE, among other stakeholders. It proposes a concrete standard approach to exchange data-literature links between established handlers of research objects, such as CrossRef and publishers [17]. Elsevier currently organizes bulk uploads of pairs of links between articles and associated data sets (independent of repository) to Scholix-hub CrossRef. We plan to automate these uploads so that these links to data are displayed on ScienceDirect at approximately the time of article publication. However, pairs of links between Mendeley Data (see 2.4) data sets and associated articles (independent of publisher) are currently sent to Scholix-hub DataCite at time of publication. Elsevier journal articles bear a "Research Data for this Article" section which is informed by a query from Science Direct to OpenAire asking for any linked research data to the article (data, software, accession numbers). Any member of the community is also able to autonomously retrieve these pairs of links article-research data and the reverse pairs by querying OpenAire [18] or directly DataCite [19].

Linking Data sets to ORCID Identifiers

As one of the founding sponsors of the ORCID project and its ancillaries, Elsevier has been invested in making ORCID a standard for the identification of scholars since its inception in 2012 [21]. Since 2012, Elsevier's editorial systems have supported the identification of authors with ORCID as part of the manuscript submission process. Linking to an ORCID is also available for data sets deposited in Mendeley Data Repository, albeit currently only indirectly via the Mendeley profile of the author being connected to the

data set. Mendeley users are invited to link their profiles to ORCID or another widely used researcher identifier to the data sets generated via Mendeley Data Repository.

The Mendeley Data Search engine

Driving discoverability of the data and facilitating linking between Elsevier-published articles and external repositories, as well as collaborating with subject specific initiatives to further increase transparency, is another key initiative. The Search function of Mendeley Data is a data search engine which initially went live as a standalone tool in June 2016. It is now integrated with the Mendeley Data [22] platform and it is openly accessible. It currently indexes over 10 million data sets from 35 supporting external repositories including Zenodo, PANGAEA and DRYAD, as well as Mendeley Data Repository itself. Its Push API allows any repository to push their data resulting in the latter appearing in Mendeley Data Search results. Furthermore, it continues to evolve to employ the latest advancements in search technology (e.g. relevancy of results is enhanced by deep indexing of data). Mendeley Data Search allows researchers to search for different data types and formats across a variety of domain-specific and cross-domain institutional data repositories and other data sources. The results retrieved are rendered with a preview functionality for quick inspection and can be filtered using different facets (repository name, data type, sources, etc.).

2.4 Infrastructures Supporting Research Data Sharing

Our hub for the complete research data management lifecycle, which further supports standardized data sharing, is the data repository and its suite of related functions in Mendeley called Mendeley Data. Working closely with partner institutions to understand what successful data management is, Mendeley Data provides a modular research data management ecosystem which integrates through open APIs with the global data ecosystem, including DANS, DataCite, OpenAIRE, ORCID and repositories. The product consists of five modules: Data Search, Notebook, Manager, Repository and Monitor. At present, the Repository, Search, and Manager are live. Notebook has been designed to become an Electronic Lab Notebook (ELN) integrated with the rest of the Mendeley Data platform and built upon the lessons learned from the standalone ELN Hivebench. Mendeley Data Monitor has been piloted with a number of development partner institutions and will be implemented in the near future [23].

Each module covers different aspects of the research data lifecycle. Crucially, researchers who currently do not share data, or who find it very difficult or labor intensive to do so, identify either legal issues (e.g. confidentiality/ethical issues), formatting (e.g. presenting data clearly), logistics (e.g. where to upload) or data cleaning (e.g. making the data usable) as their main obstacles [24]. The vision for the Mendeley Data platform is to provide researchers and institutions flexibility in meeting their RDM needs along the research data lifecycle. For example, researchers can set embargoes for their uploaded data sets so that they are only publicly available after a deferred date is reached. We also will offer institutions the ability to customize the metadata supplied for data sets in the repository, to supplement the standard metadata requirements and allow for greater detail in annotation to align with institution-specific data management policies.

Mendeley Data Repository is a general (not subject-specific) repository, with long-term and guaranteed preservation of data through a dark archiving agreement with DANS and which mints published data sets with DOIs and links to authors' ORCID identifiers. It is also a recipient of the Data Seal of Approval from CoreTrustSeal which assesses repositories on eighteen metrics in alignment with FAIR principles [25]. One key initiative over the course of 2017 was enabling our editorial systems (EES and EVISE) to directly connect to Mendeley Data Repository among other repositories and to allow authors to upload research data at point of submission of their articles. This prompt has proven to be a significant motivator for sharing data. From 2017 until December 2018, we have seen over 3,700 data sets across life sciences, physical sciences, and health sciences uploaded to the repository. Below are the subject areas which contributed data sets representing over 5% of the depositions.

Table 2. Deposition of data during manuscript submission to Mendeley Data Repository per subject category, 2017–2018.

Subject area	Percentage of data sets uploaded to Mendeley Data Repository
Energy and Earth Sciences	18.6%
Environmental, Aquatic and Agricultural Sciences	12.5%
Applied Biosciences	7.8%
Neuroscience and Psychology	7.6%
Social Sciences	6.8%
Engineering	6.6%
Chemistry	5.5%
Materials Science	5.2%

With caveats in interpreting this data, it does seem clear that one outcome of the Earth and Energy Sciences portfolio adopting a less open-ended data policy was the spike in data sets deposited to Mendeley Data. In Physics, all software associated with articles in *Computer Physics Communications* are published on Mendeley Data with an open license (about 400 computer programs since May 2016). In addition, the associated Program Library at Queen's University Belfast [26] is also being imported (more than 3,000 computer programs stored since 1969). The licenses of imported codes are converted to open ones, making the resulting library on Mendeley Data easily findable and freely available. Mendeley Data as a general repository, however, means we would not expect to see significant take-up in areas where established and familiar repositories exist; for example, in Chemistry, which accounts for just 5.5% of the data deposition, it is clear that these typically have data sets associated with them but are hosted on subject-specific repositories.

In addition to the repository element of Mendeley Data, the Manager module serves institutional users with a collaborative *Project* environment and workflow tool that enables researchers to share, organize and jointly annotate data in one place. This allows to prepare data to be published and shared in the form of a data set. Short term development of this module aims to not only provide researchers and institutions the opportunity to enrich their data sets with custom metadata but also to integrate with tools in the ecosystem, data sources and repositories both up and downstream of a data set creation.

Further development for Mendeley Data will focus on institutional customers who are looking to monitor data created by their researchers, e.g. tracking whether data exists on a local repository or on a third-party repository, in Mendeley Data Monitor. Providing a deeper understanding into where data live also enables tracking the citation and other metrics around their usage. This publisher-independent workflow toolbox will help librarians improve adoption of data sharing, and thus better comply with new mandates and funder regulation.

2.5 Role of Data Journals

Data and software journals, unknown a few years ago, have proven to be a valuable addition to the landscape, offering another route of findability for research data and including more detail and context than metadata alone generally covers. Data journals supplement the data held in repositories and offer another way to find the data – through A&I services, for example – while contextualizing the data themselves and oftentimes complementing full length research articles. Data journals are also particularly attractive as publication outlets for replication data, or negative results, as these can be outside the aims and scope of traditional field-specific journals, but important for other researchers' use. Because data journals offer their authors validation of their data via peer review, the data (negative or positive) are credentialed and the researcher has their research recognized by traditional metrics of output, i.e. publication in a peer reviewed journal. Data and software journals are generally also Open Access which increases the visibility of their publications.

These open access publications offer a significant incentive to researchers looking to share data. Generally, data journals promote data sharing in a way that can be aligned with institutional priorities, e.g. as formal, indexed publication outlets for research with the familiar metric of publication citations and altmetrics. In many cases they do not contain research data themselves but link to data repositories.

Our flagship data journal, *Data in Brief,* was launched in 2014 and has had a 40.7% CAGR (compound annual growth rate) between 2015–2018, with a CiteScore of .70. Our software journal, *SoftwareX*, was launched in 2015 and is growing rapidly (expected to exceed 100 publications in 2019); it aims to highlight the impact of software on today's research practice, and on new scientific discoveries in almost all research domains; it also emphasizes the contributions of software developers who are, in part, responsible for this shift in research trends. The validation provided by journals like *Data in Brief* and *SoftwareX* that review the data with their descriptors and the software, respectively, is an important tool for researchers seeking trustworthy data from outside their networks. To improve the peer-review process and provide researchers with an easy way to share, discover and run their published code, *SoftwareX* and several other software journals (including *Computer Physics Communications*, *Future Generation Computer Systems* and *Cell Systems*) have partnered with Code Ocean [27], a cloud-based computational reproducibility platform where researchers can upload their codes and data. Codes are privately shared with the editors and reviewers, and once a code is reviewed and accepted it receives a citable and permanent DOI, meaning that others will be able to access, download and replicate the code.

In parallel, within Cell Press we also launched the STAR Methods initiative, which is designed to improve communication and reproducibility through a structured approach for reporting of experimental materials and methods. STAR (Structured, Transparent, Accessible, Reporting) Methods summarizes the critical materials and approaches used in a paper by grouping key details (e.g. key reagents, resources used, including availability of data and software) under standardized headings, providing signposts of the article's contents for readers. This format has garnered positive feedback from researchers and core elements, particularly the Key Resources Table, are seeing broader adoption within Elsevier and beyond. Cell Press is now iterating STAR Methods to improve alignment with the TOP guidelines, and planning an extension to provide an additional level of information via *STAR Protocols*, a new open access platform that will incorporate video and images to provide clear practical guidance for researchers and further facilitate reproducibility.

Journals that allow researchers to elevate their research from raw data to a publication have twofold benefits: benefits to the reader who is able to find the relevant research quickly, and benefits to the author who is able to parlay important work into a DOI-minted and peer-reviewed publication that can draw attention to their research for replication or other uses.

3. CHALLENGES AND FUTURE DIRECTIONS

3.1 FAIR data at Elsevier

We will continue to work in concert with funding and institutional partners to raise the standards, tools and practices of research data management. To further support the creation of FAIR data, we have identified several elements in the research data lifecycle essential to enable FAIR data creation and sharing (Table 3). These constitute a high-level roadmap to implement FAIR Data at Elsevier which includes research data lifecycle events such as a) the creation and ownership of a data set, b) publishing a data set alongside or independent of a peer-reviewed article, c) the linking of deposited data sets to a journal article and d) enabling citation of one's own and others' data within the article.

Table 3. Roadmap to implement FAIR data support at Elsevier: high level overview of steps necessary to support FAIR data creation and sharing. Shaded cells green to red reflect if implementation is in the future (red) or already been initiated (yellow), or otherwise are live (green). Note that the status of these implementations is subject to change as we are continuously revising our implementations with input from all stakeholders in the research community.

	Research Data Life Cycle						
	Creating data/code: ownership		Publishing research data/object		Finding research object and/or related literature		Measuring research object impact
To implement at Elsevier	Link all own data sets published by an individual via ORCID		Deposit data set independent of or during peer-review		Link journal article data set to author data set that supports the article		Data sets (own and other authors) are cited in references and journal article body
How	ORCID linking available during submission.	ORCID linking to a data set.	Journal policies on research data sharing and editorial systems to support these.	Provide Elsevier authors a generic data repository option.	Journals require linking of research data to article or at least a data availability statement.	Scholix Framework: Data sets metadata sent to CrossRef together with article metadata. Mendeley Data sends article metadata together with data sets metadata to DataCite.	Links to data sets supporting the article or work of other authors are cited in article and list of references with active links.
Where do we stand	Available in editorial systems as an option.	Mendeley Data Repository data sets show author name and point to Mendeley profile, which supports ORCID. .	Editorial systems support data policies. If paper accepted data links become available on article page. Manual solution for data sets deposited post publication.	Researchers can use MD Repository to publish data sets.	To support FAIR there must be at least a data availability statement (even if data cannot be shared). Most of our journals encourage statement but do not require it.	Pairs data set-article links by MD sent to DataCite. Journal article pages updated for newly available data sets by querying OpenAire API.	An approach has been devised to create linked data references within the reference list and by adding accession numbers that link to reference list in article body.
Next steps		Increase level of FAIR by adding ORCID to data set as identifier for author.	Automate linking process for data sets deposited before publication.		Once other functions (in yellow) enabled, review which journals are ready to require a data availability statement.	Link to new OpenAire API to update links to data in article more swiftly and send metadata for data sets with article metadata going to CrossRef.	Implement data citation across all journals following approach listed above, and support authors with citing references in line with the Data Citation principles with clear guidelines.

This roadmap captures a birds-eye view of how Elsevier is making different aspects of data sharing and creation a reality. It is the result of a series of efforts undertaken by teams spanning Journals, Operations and Product divisions across the company.

Following the successful trials of our Earth and Energy Sciences journals, we must urge more of our journals to adopt data policies that require rather than encourage data deposition. As mentioned above, stakeholders in the Earth, Space and Environmental Sciences (Coalition for Publishing Data in the Earth and Space Sciences, or COPDESS) signed a Commitment Statement, to ensure that "research outputs, including data, software, and samples or standard information about them, are open, FAIR, and curated in trusted domain repositories whenever possible, and that other links and information related to scholarly publications follow leading practices for transparency and information" [4]. In addition Elsevier is actively pursuing a program to realign the data availability statements that authors provide in line with this Statement. The data statements are now generally encouraged rather than required; to align with FAIR principles, our guides to authors would need to change and our data policies would need to be reconsidered. The intended use of the Author Statement requires reexamination; it is currently promoted as the space for an author to describe why their data cannot be made publicly available. However, the optimal use of this statement would be for authors to describe how the data can be accessed and reused [28]. Our data policies can and should continue to evolve, recognizing the likelihood of successful adoption by subject area communities, and supported with technical implementation to facilitate seamless sharing, and remove unnecessary obstacles, for authors.

Next to building up on our data policies in a direction that increasingly supports FAIR we are also in the process of reviewing, refining and improving the infrastructures and workflows that enable the necessary research data and literature linking capabilities to enable FAIR data creation. Steps include: improving adoption of ORCID as a researcher identifier also within our RDM platform Mendeley Data, ensuring efficiency to our participation in the Scholix framework (as described in Section 2.3) as well as enabling data citation within an article body and bibliography.

This provides the current state on implementing the infrastructure to support FAIR data creation and sharing at Elsevier. A more detailed evaluation of each element of the research data lifecycle will be critical to decide the appropriate next steps to build further improvements toward FAIR.

3.2 Being part of and integrating with the Research Data ecosystem

Further education, in an ongoing dialog with all stakeholders (publishers, funders, repositories, and very much including researchers themselves) is a necessity for achieving more of the goals of open science. Elsevier (and other publishers, and data repositories) meet a real need by providing resources around data sharing best practice. Institutions can and do set RDM policies, but integrating RDM into curricula and established in lab practice tends to lag behind. Publishers and organizations like DataCite are often present now at scientific conferences leading field-specific workshops dedicated to Research Data Management, but feedback is often that finding time to integrate these workflows into institutional practice is challenging

at best, and that data management planning must happen at the stage of proposal writing. Moreover, such workshops are of course not scalable in affecting change at global data management practice.

We also will pursue further integration with the coalescing global research data management ecosystem with Mendeley Data. In addition to the planned integrations between our platform and external repositories and other data sources, we aim to integrate with machine-readable Data Management Plan, e.g. DMP Online, a tool developed by the Digital Curation Centre that enables researchers to create, review and share data management plans that meet institutional and funder requirements.

As data publication becomes increasingly normative across disciplines, it will also be important for publishers and other stakeholders to understand data reuse. Potentially, data usage and citation behave similarly to article usage and citation, meaning that reputation and pre-existing networks greatly impact the reuse of published data. Another speculation is that scientists are not looking at published data sets with the intention to mine them for novel research questions, but to replicate existing research [29].

Through our continued engagement with researcher communities in the context of publishing as well as with institutional users of Mendeley Data we have learned that one of the most common questions for researchers remains that of who owns their data. We acknowledge that the publishing and research data management landscape is complex and continued education and outreach to academic communities is essential. Elsevier is committed to not only supporting researchers and institutions to manage the data they own but also to provide them with flexibility beyond a single provider for their research data management needs. We view our participation in frameworks such as Force11, RDA or FAIR as essential in addressing such challenges, to promote interoperability and to continue to engender trust from the research community [5].

The next phase of our work will be to help define universal and open standards toward proper data management practices, and ensuring our own products and practices meet these standards. Data storage and preservation are being accelerated, and we want to support standards for ease of universal access and compliance, as well as make the interfaces for authors and institutions transparent and simple to use. Building tools that interact seamlessly with the existing research ecosystem is essential to enable a future where data are indeed findable, accessible, interoperable and truly reusable.

AUTHOR CONTRIBUTIONS

M.S.e. Silva (m.soaresesilva@elsevier.com) and A. de Waard (A.dewaard@elsevier.com) designed the framework for the paper. D. Descoteaux (d.descoteaux@elsevier.com) directed the overall planning of the article. D. Descoteaux, M.S.e. Silva, and C. Farinelli (c.fennell@elsevier.com) led the outreach to data collection from business units within Elsevier. A. de Waard, M.S.e. Silva, D. Descoteaux, and C. Farinelli all contributed to the writing of the paper.

ACKNOWLEDGEMENTS

The authors wish to thank and acknowledge IJsbrand Jan Aalbersberg (Senior Vice President Elsevier Research Integrity), Catriona Fennell (Director Elsevier Journal Services), and Adriaan Klinkenberg (Publishing Director, Elsevier Physics Journals) for their extended review of the manuscript. The authors also thank Maxim van Gisbergen (Senior Strategy Manager), Alberto Zigoni (Market Development Director, Research Data Management), Deborah Sweet (Vice President of Editorial at Cell Press), Lorenzo Feri (Director of Product Management) and Wouter Haak (Vice President, Research Data Management) for discussions on, among other topic relevant to this manuscript, the analytics regarding research data metrics at Elsevier.

REFERENCES

[1] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, … & B. Mons. The FAIR guiding principles for scientific data management and stewardship. Scientific Data 3(2016), Article No. 160018. doi: 10.1038/sdata.2016.18.

[2] G20. (2016). G20 Leaders' Communique Hangzhou Summit 2016. (September), 1–9. Available at: http://www.fsb.org/wp-content/uploads/G20-Antalya-Leaders-Summit-Communique.pdf.

[3] European Commission. Turning FAIRInto reality: Final report and action plan from the European commission expert group on FAIR data. (2018) doi: 10.2777/1524.

[4] COPDESS. FAIR Author Guidelines. Available at: http://www.copdess.org/enabling-fair-data-project/author-guidelines/.

[5] P. Wouters, & W. Haak. Open data: The researcher perspective. Elsevier – Open Science, 48(2017). doi: 10.17632/bwrnfb4bvh.1.

[6] W. Haak. 4 principles for unlocking the full potential of research data. Elsevier Connect. (2015). Available at: https://www.elsevier.com/connect/4-principles-for-unlocking-the-full-potential-of-research-data.

[7] M. Martone (ed.). Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. (2014). doi: 10.25490/a97f-egyk.

[8] H. Cousijn, A. Kenall, E. Ganley, M. Harrison, … & T. Clark. A data citation roadmap for scientific publishers. Scientific Data, 5, 1–11. (2018). doi: 10.1038/sdata.2018.259.

[9] A. de Waard, H. Cousijn, Ii. J. Aalbersberg. 10 Aspects of Highly Effective Research Data. Elsevier Connect, (December 2015), 1–6. (2015). Available at: https://www.elsevier.com/connect/10-aspects-of-highly-effective-research-data.

[10] Research Data Management. Available at: https://researcheracademy.elsevier.com/research-preparation/research-data-management.

[11] A. de Waard. The Mendeley Data management platform: Research data management from a publisher's perspective. (2017). Available at: https://www.elsevier.com/__data/assets/pdf_file/0005/504563/08212017144742_deWaard082117.pdf.

[12] Research Data-Data Citation. (n.d.). Available at: https://www.elsevier.com/about/open-science/research-data/data-citation.

[13] Transparency and Openness Promotion (TOP) Guidelines. (2014). Available at: https://cos.io/our-services/top-guidelines/.

[14] MSU. (n.d.). MSU – Research Data Guidelines. Available at: https://www.elsevier.com/authors/author-resources/research-data/data-guidelines.

[15] Database linking. Available at: http://www.elsevier.com/about/content-innovation/database-linking.

[16] A. Burton, H. Koers, P. Manghi, M. Stocker, … & U. Schindler. The SCHOLIX framework for interoperability in data-literature information exchange. D-Lib Magazine, 23(2017), 1–2. doi: 10.1045/january2017-burton.
[17] A. Burton, M. Fenner, W. Haak, P. Manghi, A. Burton, M. Fenner, … & P. Manghi (2017). SCHOLIXMetadata Schema for Exchange of Scholarly Communication Links. (2017). doi: 10.5281/zenodo.1120265.
[18] OpenAire. (n.d.). Available at: https://explore.openaire.eu/search/find.
[19] Datacite. (n.d.). Available at: https://search.datacite.org/.
[20] Datacite. Event Data. Available at: https://datacite.org/eventdata.html.
[21] B.M. Taylor & C. Shillum. New ORCID ID. (October), 1–5. (2012). Available at: https://www.elsevier.com/editors-update/story/practical-tips/new-orcid-id-aims-to-resolve-authorship-confusion.
[22] Mendeley Data. (n.d.). Available at: https://data.mendeley.com/.
[23] Mendeley Data roadmap. (n.d.). Available at: https://www.elsevier.com/solutions/mendeley-data-platform/releases/roadmap.
[24] P. WOuters & W. Haak. Open data: The researcher perspective. Elsevier – Open Science 48(2017). doi: 10.17632/bwrnfb4bvh.1.
[25] Data Seal of Approval. (n.d.). Available at: https://assessment.datasealofapproval.org/assessment_244/seal/html/.
[26] CPC Program Library. (n.d.). Available at: http://cpc.cs.qub.ac.uk/.
[27] Code Ocean. (n.d.). Available at: https://codeocean.com/.
[28] A. Stall, P. Cruse, H. Cousijn, J. Cutcher-Gershenfeld, A. de Waard, … & l. Yarmey. Data sharing and citations: New author guidelines promoting open and FAIR data in the earth. Science Editor 41(3), 83–87. (2018) Available at: https://www.csescienceeditor.org/wp-content/uploads/2018/11/CSEv41n3_text_83-87.pdf.
[29] I. Pasquetto. Do scientists reuse open data? Medium. (2019) Available at: https://medium.com/voices-from-the-open-science-movement/by-irene-pasquetto-bac8bc02afdc.

AUTHOR BIOGRAPHY

Danielle Descoteaux is Publisher of Ecology and Biodiversity Journals at Elsevier. Formerly, she was a Senior Editor of Linguistics and Psychology books at Wiley-Blackwell.

Anita De Waard is Vice President of Research Collaborations at Elsevier. Her work focuses on developing cross-disciplinary frameworks for sharing data and tools to store, share and search experimental outputs, in collaboration with academic and government groups, in the US and Europe.

Chiara Farinelli is a Product Manager for Pure and was formerly Publisher of the High Energy and Nuclear Physics journals at Elsevier.

Marina Soares e Silva is Product Manager for Research Data Management and Mendeley Data at Elsevier. Previously she was Publisher for the Elsevier Biomaterials journals portfolio.