Title
Balancing the needs of consumers and producers for scientific data collections

Authors
Agarwal, Deborah A
Damerow, Joan
Varadharajan, Charuleka
et al.

Publication Date
2021-05-01

DOI
10.1016/j.ecoinf.2021.101251

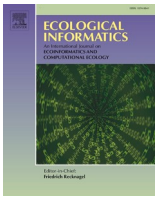
Peer reviewed

Ecological Informatics 62 (2021) 101251

Contents lists available at ScienceDirect

Ecological Informatics

journal homepage: www.elsevier.com/locate/ecolinf



Balancing the needs of consumers and producers for scientific data collections

Deborah A. Agarwal *, Joan Damerow, Charuleka Varadharajan, Danielle S. Christianson, Gilberto Z. Pastorello, You-Wei Cheah, Lavanya Ramakrishnan

Ernest Orlando Lawrence Berkeley National Laboratory, Berkeley, CA, USA

A B S T R A C T

Recent emphasis and requirements for open data publication have led to significant increases in data availability in the Earth sciences, which is critical to long-tail data integration. Currently, data are often published in a repository with an identifier and citation, similar to those for papers. Subsequent publications that use the data are expected to provide a citation in the reference section of the paper. However, the format of the data citation is still evolving, particularly with regards to citing dynamic data, subsets, and collections of data. Considering the motivations of both data producers and consumers, the most pressing need is to create user-friendly solutions that provide credit for data producers and enable accurate citation of data, particularly integrated data. Providing easy-to-use data citations is a critical foundation that is required to address the socio-technical challenges around data integration. Studies that integrate data from dozens or hundreds of datasets must often include data citations in supplementary material due to page limits. However, citations in the supplementary material are not indexed, making it difficult to track citations and thus giving credit to the data producer. In this paper, we discuss our experiences and the challenges we have encountered with current citation guidance. We also review the relative merits of the currently available mechanisms designed to enable compact citation of collections of data, such as data collections, data papers, and dynamic data citations. We consider these options for three data producer scenarios: a domain-specific data collection, a data repository, and a large-scale, multidisciplinary project. We posit that a new mechanism is also needed to enable citation of multiple datasets and credit to data producers.

1. Introduction

Funders of Earth science projects and academic publishers increasingly require scientists to publish data in an open-access data repository (Cousijn et al., 2018; Data Citation Synthesis Group, 2014; Office of Science, 2013; Stall et al., 2019). Many data repositories are available to store the data, with varying levels of service for data quality review, and guidance or helpful tools for data citation and tracking. These range from general purpose repositories that accept all data to domain, project, and/or sponsor specific repositories (Witt et al., 2019) that are critical for data integration. Data supporting a published paper may therefore be dispersed across multiple repositories that have different practices for data citation and linking to related data. In particular, citation approaches that address complexities of collections of data and dynamic data, represent an active area of research (Parsons et al., 2019; Silvello, 2018). Library, data, and information science communities have largely driven the development of principles and approaches to data citation (Parsons et al., 2019). However, more work is needed that explores specific use cases and the needs of scientific data producers and consumers to ensure a strong foundation for citation practice that is necessary to support data integration.

Many related organizations, standards, and best practices for citation have been established and are evolving (Parsons et al., 2019). For example, the DataCite organization maintains a preferred schema and registration service for metadata required to obtain a Digital Object Identifier (DOI) for data (DataCite Metadata Working Group, 2019) and best practices for data citations. The Research Data Alliance (RDA; https://www.rd-alliance.org/) working groups are developing recommendations for data citation, versioning, and collection standards (Rauber et al., 2016a, 2016b; Weigel et al., 2017). More recently, the Google dataset search has expanded opportunities to search for and access data, using JSON-LD as the required metadata (https://datasetsearch.research.google.com/). The ESIP Data Preservation and Stewardship Committee 2019 guidelines also lay out citation recommendations (ESIP Data Preservation and Stewardship Committee, 2019).

There is a growing movement to make data Findable, Accessible, Interoperable, and Reusable (FAIR) (Gries et al., 2019; Stall et al., 2019; Wilkinson et al., 2016, 2019). The key to supporting FAIR data principles and providing the ability to automatically find and using data is having a unique identifier, typically a DOI, for published data that data consumers can cite and an accurate citation text that enhances the machine readability (Wilkinson et al., 2016). Many organizations are

* Corresponding author.
E-mail address: daagarwal@lbl.gov (D.A. Agarwal).

https://doi.org/10.1016/j.ecoinf.2021.101251
Received 16 November 2020; Received in revised form 4 February 2021; Accepted 5 February 2021
Available online 15 February 2021

now working to make data FAIR and to define more precisely what that means, including FORCE11 (http://force11.org/), GO FAIR (https://www.go-fair.org/), and FAIRsFAIR (https://www.fairsfair.eu/). In addition, efforts are under way to assess a repository's adherence to Transparency, Responsibility, User Focus, Sustainability, and Technology (TRUST) principles (Lin et al., 2020). The Core Trust Seal has a certification that repositories can obtain to become Core Trustworthy Data Repositories (https://www.coretrustseal.org/). As expected in an evolving system, the adoption of emerging standards and best practices is lagging and citation practices in published scientific articles remains highly variable (Escribano et al., 2018; Khan et al., 2020).

The practice of data publication to date had been modeled using the paper publication processes as a guide (Borgman, 2016; Silvello, 2018). Data producers are typically also the authors of the data publication and make decisions on how to package the data using a variety of considerations, including: order of authors who need to be included, theme of the data, time range of the data, type of data, expected usage needs, project goals, sponsoring organization, and/or scale of the data. Once the data have been submitted to a repository with appropriate metadata describing the data, they are ideally reviewed by the repository and published (e.g., Kakalia et al., 2019; O'Brien et al., 2016). Typically, the repository assigns a DOI to each published dataset and provides an automated citation using DOI schema metadata fields (e.g., authors, title, publication year, publisher) (Fenner et al., 2019). Another important component of the data publication process is specification of data usage rights/license by the authors. In this paper, we will refer to a published dataset as a *data package*. A data package contains one or more data files and the associated metadata needed to find, interpret, and use the data.

Although data publication has many similarities to paper and software publication, important differences must be considered in future development of data publication systems (Borgman, 2016; Silvello, 2018). In this paper, we consider the perspectives of data producers and consumers serving multidisciplinary Earth and environmental science projects that focus on generating data, using data, and/or operating data services. In particular, we discuss the citation challenges encountered when integrating data from a large number of data packages as well as some of the emerging solutions.

2. Example earth science data producers and consumers

We present a perspective informed by working closely over many years with several US Department of Energy (DOE) projects, including the Environmental Systems Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE) data repository (Varadharajan et al., 2018; https://ess-dive.lbl.gov/), the AmeriFlux carbon flux network (Novick et al., 2018), the Watershed Function Science Focus Area (WFSFA; http://watershed.lbl.gov) (Varadharajan et al., 2019), and Next Generation Ecosystem Experiments - Tropics (NGEE-Tropics; https://ngee-tropics.lbl.gov/). These projects range from data synthesis studies to collaborations that generate data products using workflows, tools, and resources for data management. Our experiences also capture hosting domain specific repositories and a general purpose data repository.

2.1. ESS-DIVE

The ESS-DIVE repository stores data from research funded by the DOE's Biological and Environmental Research (BER) Environmental Systems Science (ESS) program. The ESS-DIVE data repository provides long-term data storage and access, and is working toward supporting FAIR data principles (Varadharajan et al., 2018). Each data package has a DOI either assigned or obtained by the data producer before it is stored in ESS-DIVE. There are important features in development in ESS-DIVE - project portals and the Fusion Database. Project portals are a feature, allowing a set of data packages to be grouped together with explanatory web pages and data search/access on that set. The Fusion Database enables advanced search of data across the data packages in ESS-DIVE and retrieval of the specific data found through the user's search query. The projects listed below are funded under the ESS program, and are required to submit data to ESS-DIVE for long-term stewardship.

2.2. AmeriFlux

The AmeriFlux network is a collection of independently-managed carbon flux measurement sites in the Americas. Participating AmeriFlux sites have been collecting data for one to over 28 years. Team members at each site typically change every one to five years with the principal scientists changing at a much lower frequency. There are currently 374 AmeriFlux sites with data published through the AmeriFlux central data system (https://ameriflux.lbl.gov/). The AmeriFlux data system provides access to over 2400 site years of data from 10 countries. Each AmeriFlux site's data product is assigned a DOI and is made available for download from the central data server under the AmeriFlux data usage policy that requires contacting the site's principal scientists and citing the data used, as well as offering the potential for co-authorship to the site team members. The data produced by a site include 40–120 time series and 10s - 1000s of sporadically-measured data. The data are updated at the central system regularly by the sites. An update will typically add new data and may also include corrections to prior data. Since 2015, there have been over 17,000 downloads of AmeriFlux data packages by over 3000 unique data consumers, and the data have contributed to over 950 published papers in the last five years. These papers range from studies of a small number of sites to large scale data analyses using all sites in the network. Example uses by consumers of the AmeriFlux data include: calibration of regional and global climate models, carbon dynamics analyses, and disturbance impact. AmeriFlux is also part of a global network of networks where we combine the data from across the regional networks into a global dataset called FLUXNET (https://fluxnet.org/).

2.3. WFSFA

The WFSFA project has over 70 researchers and 30 affiliated projects studying the East River watershed in the Upper Colorado River Basin to quantify the long-term impact of perturbations such as drought and early snowmelt on water availability and quality (Hubbard et al., 2018). The project has many teams studying different aspects of the watershed, such as bedrock-soil-plant-microbiome interactions in the hillslope and meander-river interfaces in the floodplain. In the first five years, the project teams generated over 96 data packages in its repository. Some of these packages are continually updated, long-term multivariate datasets spanning multiple years of data collections, while some are smaller and associated with field campaigns or publications (Kakalia et al., 2020). Affiliated projects are typically funded independently and may have data in other repositories. Thus, research publications from the project may utilize data from across multiple institutions and data repositories. The WFSFA data are made available for public use and issued a DOI, through publication to the ESS-DIVE data repository under a Creative Commons By Attribution Version 4.0 (CC BY 4.0; https://creativecommons.org/licenses/by/4.0/) usage policy. Example uses by consumers of the WFSFA data include: study of watershed dynamics, analysis of nutrient dynamics of microbial populations, and calibration of vegetation models.

2.4. NGEE–Tropics

The NGEE–Tropics project has an international research team with over 120 researchers from institutes including: US National Laboratories, the Smithsonian Tropical Research Institute, the U.S. Forest Service, and many institutions in the Tropics. Together, the team is modeling how the carbon cycle of tropical forests will respond to rising CO_2 concentrations and a changing climate. The data span

ecohydrological measurements like leaf temperature, sap-flux measurements, soil moisture, tree demography and mortality. The data packages for the project are first curated in a project data server and then transferred to the ESS-DIVE repository where they will be published for public use under the CC BY 4.0 usage license. Over the first five years, the project generated over 123 data packages. Example uses by consumers of the NGEE-Tropics data include: tropical forest succession dynamics, climate forcings in tropical regions, and human impacts on tropical forests.

3. Socio-technical challenges of data citations

The data in a data package from any of the projects mentioned above is the result of a significant scientific endeavor requiring time spent planning, collecting, cleaning, processing, and managing the data. Many people are often involved in collecting the data, typically over many years, enabling numerous future scientific publications by the team collecting data. The extent of an individual data package is generally determined by a combination of the data time period, spatial extent, authorship, and expected data usage. The decision regarding which data to include in a particular data package is usually made by the data producers. The time and effort required to make these decisions and prepare the data and metadata for the data package is usually significant and easy to under-estimate.

Publishing the data and making it publicly available in a repository before the authors have completed their associated publications comes with a perceived risk (Tenopir et al., 2015). Longitudinal surveys of researchers have found that the top concern for open sharing of data has been potential misuse of data, such as misinterpretation (Fane et al., 2019; Borrelli et al., 2020). Ideally, the rewards of publishing data will outweigh the risks. Current incentives that motivate researchers to publish data include meeting funding or journal requirements, getting proper credit for sharing data, and expanded use of data that contributes to advancing scientific outcomes (Piwowar and Vision, 2013; Fane et al., 2019; Borrelli et al., 2020).

If publishing a data package yielded equivalent or greater credit and prestige to publishing a paper, this would also incentivize early data publication. Subsequent citations and download metrics can be one way to demonstrate the scientific value of a data package (e.g. Mooney and Newton, 2012; Belter, 2014; Fenner et al., 2018). Such practices are being developed in the community, e.g., the COUNTER Code of Practice for Research Data has developed standard usage metrics as an important measure of impact for data (https://www.projectcounter.org/code-of-practice-five-sections/abstract/; Fenner et al., 2018). However, many scientists prioritize citation of one of their own papers describing the data over the data citation. This makes it difficult for the paper reader to find the original data. We need mandatory linkages between the data package and any associated paper(s), enforced across journals and repositories, where citation metrics for data and the associated papers are also explicit and linked.

A challenge in promoting data citation as an incentive for making data public is that many authors neglect to cite data sources in the references section of a paper (Ball-Damerow et al., 2019; Escribano et al., 2018; Mooney and Newton, 2012), and it is unfortunately difficult to obtain accurate data citation metrics as citation search today requires elaborate infrastructure and subscription fees to search all the possible sources. (Cook et al., 2016; Cousijn et al., 2019; Parsons et al., 2019). Scientific funders, journals, and data publishers must better support and incentivize researchers to create FAIR data packages, and contribute to improvements in data use and citation metrics that portray meaningful measures of impact (Nosek et al., 2015).

Another incentive to publish data is the scenario where collaborators share authorship of both data and associated journal publications resulting from data use. Co-authorship on subsequent publications is recommended in cases where the specific data is integral to the final paper (Duke and Porter, 2013). In the past, the only way to find data was through a personal connection to the data producers (Wallis et al., 2013). These connections often led to collaborations and co-authorship offers. However, broadly accessible data has led to more opportunities for scientific analyses that integrate diverse datasets from across a region or globally. In these cases, only a citation of the data (and associated papers) on which the resulting publication relies is expected (Duke and Porter, 2013; Kratz and Strasser, 2015).

As data consumers, some of our project scientists integrate and use data from other government agencies that generate large-scale datasets from sensors and satellites that only require a single citation or acknowledgement (e.g. United States Geological Survey National Water Information System: https://help.waterdata.usgs.gov/faq/miscellaneous/how-to-cite-usgs-water-data-for-the-nation-waterdata.usgs.gov-in-a-publication, NOAA National Climate Data Center: https://www.ncdc.noaa.gov/, National Ecological Observation Network: https://data.neonscience.org/, and NASA: https://lpdaac.usgs.gov/data/data-citation-and-policies/). Data consumers accustomed to citing these large agency datasets prefer to avoid using data that require multiple citations or other time-consuming usage terms such as data producer notification. For example, in our experience most data consumers prefer that offering co-authorship to data producers be optional. The FLUXNET data (https://fluxnet.org/) which include AmeriFlux data have often been released with sites divided between two usage policies, one with a requirement for co-authorship opportunity and one without. Over time, our experience indicates that the number of downloads of datasets from sites without the co-authorship requirement has been roughly double those requiring co-authorship opportunities.

Projects like AmeriFlux, WFSFA, and NGEE-Tropics each have many different teams working in the field and producing large numbers of high-quality, multi-disciplinary data packages. Integrating these together as one product can only be achieved if the data usage policies are compatible and the data producers can receive appropriate credit for their data. For example, when individual network datasets are included in the FLUXNET dataset they will be released under a FLUXNET data usage policy, the agreed data usage policy is a factor determining which regional datasets will be able to be included. When the FLUXNET data usage policy is incompatible with the network policy, often data can't be included. In cases where the data are considered high-value the data producers are reluctant to lose the direct credit to their data. In addition, new data versions are uploaded regularly due to collection of new data, new calibrations, and/or corrections. In our experience, the data producer would also like the version of the data to be tracked and indicated in the citation.

ESS-DIVE illustrates the socio-technical challenges that arise between data producers and data consumers. The data producers from DOE ESS-funded projects are explicitly required to store their data in the repository. When a data producer makes changes to a data package, they can decide whether this change creates a new data package or is a new version. However, it is often difficult for the data producer to make that decision without sufficient knowledge of the impact of the changes on the data consumer(s). On the other hand, the primary motivation of a data consumer is to find and use the data to address a science challenge. They would prefer to be oblivious to individual data packages and instead search for and download precisely the data needed to answer the question as will be provided in the ESS-DIVE Fusion Database. They would also like to cite the data they used in developing a paper but need an easy way to do it, such as automatically generated citations.

4. Data citation

A citation for data used is now generally recognized by data repositories and data producers as the most appropriate means of referencing data used in a paper. Surveys of researchers have shown that a full data citation within a research paper is one of the strongest motivators for sharing data, along with increasing the impact and visibility of their work and public benefit (Borrelli et al., 2020; Fane et al., 2019),

This highlights the need for the citation format and the ability to count citations of data packages. One way that data repositories can help is to provide easy-to-use data citations for downloaded data. Another way would be for paper publishers to make data citations in papers easy to search and collect statistics on.

Metadata fields to describe a data package (e.g. authors, title, publisher, keywords, spatial and temporal extent) are used to generate the citation and are important for data search and making data FAIR. The format of a dataset citation is relatively well defined (DataCite Metadata Working Group, 2019; ESIP Data Preservation and Stewardship Committee, 2019), but the ways to complete some of the metadata fields that contribute to the citation are still evolving. These fields can also impact the recommendations for citations for integrated data.

The publisher field is defined by DataCite as "The name of the entity that holds, archives, publishes, prints, distributes, releases, issues, or produces the resource" (DataCite Metadata Working Group, 2019). The ESIP repository field has a similar definition (ESIP Data Preservation and Stewardship Committee, 2019). Both the DataCite and ESIP guidelines recognize that entities other than the repository might be appropriate for this field, but when multiple entities perform the functions listed or when relocating the data to a different repository, the guidance is less clear. The publisher/repository field might contain any of the following: the repository where the data was first published; the entity that curated and prepared the data; the institution/agency/project that produced the data; or the repository where the data are currently available. In ESS-DIVE, the project that processed, curated, and packaged the data such as WFSFA or NGEE-Tropics is the publisher. These projects are also more likely than the repository to be used as a search term to find the data. We think of the data package as the long-term entity and the data repository as where the data packages are held currently. It is our expectation that data repositories come and go over the years.

The guidelines on including the data version in the citation also continues to evolve. The 2019 ESIP data citation guidelines now specify to add data version to the citation (ESIP Data Preservation and Stewardship Committee, 2019). The version field allows specification of the version of data cited without requiring a new DOI and is the approach we take. Versioning enables reproducibility without requiring a new DOI for each version and enabling easier tracking of citations of the data. Data that are frequently updated, such as continuous sensor streams, can still be challenging to version if the data are updated frequently (e.g. daily or hourly). In these cases, the date of the last update can be used to indicate the version. Including the correct citation with each download helps the data consumer to cite the correct version. Providing methods for retrieving past versions of the data also aids in reproducibility.

5. Collective data citation

As data becomes more accessible, data consumers are able to incorporate data from many data packages into an analysis. However, when a large number of data packages have been used, properly citing the data in the reference section without going over page limits can be difficult. For example, a paper based on the data from all the AmeriFlux sites, all of the NGEE-Tropics data packages, or all of the WFSFA data packages would need to include 350, 70, and 96 citations, respectively, in the references section. This large number of required citations often leads to the citations for the data being provided in the supplementary materials, within data/metadata files, or as a table of identifiers. In all of these cases, it is difficult to programmatically find the citation of the data and thus is unlikely to be counted in the data's citation metrics. A solution to this challenge would be to provide a single citation that can represent the many underlying data citations. There are three primary options for enabling this approach that have been recommended to date. We discuss the advantages and disadvantages of these options in the context of the four example Earth science projects.

5.1. Data collections

Several groups have developed recommendations for metadata and tools to support collections of digital objects on the web, such as OAI-ORE (https://www.openarchives.org/ore/), BagIT (https://tools.ietf.org/html/draft-kunze-bagit-08), Portland Data Model (https://github.com/duraspace/pcdm), ESIP Data Citation Guidelines for Earth Science Data (ESIP Data Preservation and Stewardship Committee, 2019), and most relevant to data collections in repositories is the RDA Recommendation on Research Data Collections (Weigel et al., 2017). A data collection consists of a persistent identifier (typically a DOI) for the collection, metadata describing the collection, and clear linkages to all datasets included in the collection; each member dataset has its own DOI and specific relationship to the collection that is tracked by the repository (e.g. "hasPart"/"isPartOf" from DataCite relation types). Data collections have been used to aggregate data products into larger groupings that can include datasets, other data collections, papers or other digital objects that may or may not be stored in the same repository. An advantage of data collections is that they can be used to group data into a thematic set to help data consumers to find and download related data packages easily.

Most data repositories do not have tools that explicitly allow users to create their own data collections. However, the basic metadata describing a data collection is similar to that of data packages and include authors, title, description/abstract, date created and published, and data access license. The related references/identifiers metadata field can allow the specification of all datasets included in the collection (DataCite Metadata Working Group, 2019); metadata describing each dataset's role in the collection could provide additional useful context for the collection (Weigel et al., 2017; ESIP Data Preservation and Stewardship Committee, 2019). We could use the existing data package submission interface to create collections, with the addition of specific related identifiers for datasets included and associated roles of each dataset in the collection. Or, the ESS-DIVE Project Portals could be turned into a more user-friendly interface to create data collections for data within the repository, if the required structured metadata were incorporated into the portals and a DOI was assigned.

The potential advantages of data collections are 1) relatively easy creation (for the repository and users) of a citation representing dozens of datasets that are linked by a theme, project, etc. 2) components of the data collection can easily come from other repositories, and 3) data collection producers and repositories can control exactly what data are grouped together into a data collection, without constraints of repository query capabilities. Creating a collection allows authors to provide useful groupings and contextualization for data, with descriptive metadata and links to the underlying datasets. For instance, the WFSFA project could build data collections using ESS-DIVE with all of the groundwater level data from the East River study site into one collection and the meteorology data into another, or they could build an updated data collection each year of all the data available from the project. The resulting citation for the data collection can then be used to cite all of the resources in the collection.

The most significant challenge in implementing data collections is ensuring that the datasets included in a collection receive appropriate credit for papers citing the collection (primary rather than secondary citations). Ideally, paper(s) citing the collection should also propagate to the citation counts of underlying datasets. There is little or no support for data collection creation and usage in the current data repository and publisher infrastructure. Other challenges involve deciding authorship for a data collection (the collection creators, the authors of the underlying datasets, both, or nobody), and ensuring that usage licenses of underlying datasets allow that they be compiled into a collection (e.g. compliance with dataset attribution requirements). Potential solutions to the issue of collection authorship may include implementing the CRediT (https://casrai.org/credit/) roles for authors involved in compiling the data collection. Another drawback of a data collection is

that a citation of a data collection does not convey information about specific subsets of the data used.

5.2. Data papers

Data papers are an emerging method of increasing the prestige and usability of a data package or set of data packages through the peer review process (Costello et al., 2013; Kratz and Strasser, 2015). A data paper usually provides information about a dataset (consisting of one or more data packages). The data paper can provide context and in-depth discussion regarding the data, including how and why they were collected, processed, and analyzed. It also typically provides details regarding formats, measurement method limitations, and other information needed to understand the data. A data paper can also provide overview statistics and plots describing the data. Some journals have requirements that can improve or hinder publication of data via the data paper approach. An example of a requirement for Nature's Scientific Data is that the data paper accompanies publication of new data (not previously widely available to consumers). Another typical requirement is that the data be publicly available without restrictions in an open data repository (e.g. CC BY 4.0). The data paper is not automatically a citation for the data that it covers. However, at the discretion of the data paper's authors and the data package publishers, the citation of the data paper can be used as a citation for the data. However, citation guidance generally advocates citing both the data paper and the underlying data.

A major advantage of a data paper is that the publication generally increases the prestige and visibility of the data, so authors of the member datasets do not care if the underlying datasets receive citation credit. An example of a recent data paper that was used as the citation for multiple data packages is the FLUXNET2015 paper published in 2020 (Pastorello et al., 2020). In this case, the data paper describes the FLUXNET2015 version of data from global carbon flux sites (including AmeriFlux sites). The development of this paper required on the order of eight months, a change of data usage policy to CC BY 4.0, and collection of significant additional data to meet the requirement that the data be different from what was previously available. Another example is a data paper describing the data collected by the WFSFA and collaborating institutions in the East River (Kakalia et al., 2020). The paper lists and cites more than 60 datasets spread across six repositories (including ESS-DIVE) and provides an integrated view of the diversity of data collected across various investigators. A conclusion that we drew from our experiences with these papers is that data papers are an excellent means of providing a reference and/or additional information for a collection of data packages that are 'final' and not expected to change. A data paper is also more likely to result in appropriate credit for authors, and accurate citation metrics.

One of the downsides of a data paper is that it is a static view of the data. This raises problems for any effort that involves continuous data collection such as Ameriflux or WFSFA. For example, the data paper for the East River will become quickly outdated because new data is continuing to be published after the paper was submitted. In addition, the methods used to collect and process the AmeriFlux and WFSFA data are not usually changing significantly year to year, and thus the newness requirement of some journals would be hard to satisfy.

5.3. Scalable dynamic data citations

Repositories that enable a search and custom download of data (i.e., a download matching a query), particularly on continually growing and changing data, needs to provide a citation for the specific data that were downloaded. The RDA Dynamic Data Citation Working Group recommends using a 'Dynamic Data Citation' (Rauber et al., 2016a), which creates a citation for the query and timestamp enabling the same data and version to be retrieved in the future. A unique identifier, usually a DOI, is assigned as well as a landing page with the query and timestamp that resolves to a list of the data packages, including their version and subset contributing to the result of the query. Tracking citation statistics for a dynamic data citation requires propagating data use and citation metrics from the dynamic data DOI to all components within the dynamic data citation.

An example of an international network and data infrastructure utilizing dynamic data citations is the Global Biodiversity Information Facility (GBIF). GBIF enables global search and integration of standardized biodiversity records and provides dynamic data citations with DOIs for each consumer data download, referred to as an occurrence. The consumer is requested to cite the occurrence DOI in any resulting papers (https://www.gbif.org/, https://www.gbif.org/citation-guidelines). A data consumer could still have multiple citations if they use multiple queries/downloads or repositories to compile relevant data. Several additional repositories have implemented versions of the dynamic data citations recommendations as part of the RDA Working Group recommendation acceptance process (https://www.rd-alliance.org/recommendations-outputs/adoption-stories). The ESS-DIVE Fusion Database will likely also need to implement dynamic data citations to provide accurate citations.

Currently, dynamic data citations are most advantageous to support query-based data downloads within a single repository or data aggregator and specifying subsets of data. It is particularly useful when data is standardized and specific data can be readily integrated within and across datasets from a search query, such as integrated species occurrence records from hundreds of data sources in GBIF. The data consumer only needs to reference one citation for the custom query and download instead of all the underlying data packages. It also helps support repeatability and accurate credit since it enumerates the exact subset of the data retrieved from the repository (Rauber et al., 2016b).

A primary disadvantage is that many repositories do not have the resources required to maintain DOI landing pages for each query, and to generate and track data use and citation metrics within the dynamic data citation for all underlying data packages. To date, citation tracking information has not been easy to obtain. Once crossref (https://www.crossref.org/) or another citation tracking service such as Scholix (http://www.scholix.org/) becomes reasonably comprehensive, citations will be easier to track and resolve. At present, the tracking overhead is high. According to GBIF statistics, millions of occurrence DOIs have translated into thousands of resulting publications citing those DOIs. The entity that created the dynamic data citation must also take on the long-term responsibility for providing a landing page for the issued DOIs.

Similar to data collections, the authorship of the dynamic data citation is open to interpretation. Some repositories have listed all the authors of the underlying data packages (Hunter and Hsu, 2015) and others have not specified any authors (e.g. https://www.gbif.org/). One advantage of adding all the authors is that it might reduce the need to explicitly track all citations of the dynamic data citations if there are not too many authors since authors can do this tracking themselves (although this would be impractical for large data collections like AmeriFlux, WFSFA, and NGEE-Tropics).

Dynamic data citations are also generally constrained to repository query capabilities. The way dynamic data citations are currently implemented at repositories assumes that the data can be resolved through a single or small number of queries that the consumer formulates at a single repository. One challenge is that the data consumer likely does not know exactly what data will contribute to the paper until the paper is mostly written.

An alternate approach would be to have a tool that allows the data consumer to visit a site and specify the dynamic data citations used in their paper and create a custom, synthesized data citation. This would limit the citations of datasets that were not used in the paper but had a dynamic citation generated and could allow the combination of data from many repositories into the dynamic data citation. Such a service would be of high value to the data consumer.

6. Discussion

There are many published guidelines that help in defining the metadata and format to support data citations (DataCite Metadata Working Group, 2019; Fenner et al., 2019; Mooney and Newton, 2012; Rauber et al., 2016a, 2016b; Weigel et al., 2017). Although FAIR principles focus on important concerns of data consumers, they do not adequately prioritize data producers' needs or challenges with integrated data. Changing the priority to focus on both the producers and consumers leads to differences in how the citation metadata fields are used and the format of the citation. Additionally, there remain many decisions for repositories and data producers on the path to generating, supporting, and building the citation for a data packaging system. A missing component in the data citation guidelines is recommendations for appropriately acknowledging the personnel who contributed to the tasks like processing, cleaning, and curation of the data (the ESIP Research Object Citation Cluster is addressing this now). The importance of acknowledging these roles increases when data are synthesized and/or transformed, as the integration process requires extensive intellectual effort.

The various constituencies in data publication have different needs in the processes of data publication and citation, and these needs lead to different choices in both the formulation of the citation and the management of the scaling challenge when citing multiple data packages. The current guidance on citations leaves flexibility needed to serve different constituencies. However, determining what approach is best for a particular community requires data producers and repository teams to study the options for data package granularity, versioning, and citation formulation, and provide tools that help consumers to generate, track, and use citations efficiently. The needs of the data producers and the expected usage scenarios all need to be taken into account, because the decisions will have long-term impact. The first time we were faced with this challenge was for AmeriFlux, and it took eighteen months to determine an approach to data packaging and citation. Examples of decisions needed included: granularity of data to assign individual DOIs, metadata fields, dataset titles, author criteria, and registering agency for DOI assignment. The final decisions of a single DOI per site and initial author list based on site PIs were made after several rounds of consultations with data producers, data consumers, and digital library experts. For subsequent projects, we were able to make these decisions in closer to three to six months due to our experience with the AmeriFlux project DOIs.

Data consumers, particularly modelers and machine learning developers, are utilizing data from many data packages and sources at a greater pace than ever before. Citation of large collections of data packages or specific integrated data within and across data packages is an under-appreciated challenge. We described some of the options currently available to repositories for providing integrated citations as well as their pros and cons. Which option is best depends on the particular use case. In all cases, these options have the potential to enable a compact citation for a paper. However, both data collections and especially dynamic data citations require significant specialized infrastructure and interfaces to be built at the repository to create them and to ensure appropriate credit to data producers, and none of the options have seen broad universal adoption. Data papers are currently the best way for data producers to receive credit but are time consuming and result in a static view of the data. Challenging decisions such as defining authorship and following usage policy requirements such as attribution are not directly solved. In cases where usage rules for the integrated data are not the same, a method for unifying the data usage policies is needed. In our experience, treating all the data as having the most restrictive policy can work if the policies are compatible. That is an approach we used with FLUXNET which often has two different data usage policies.

In ESS-DIVE, we expect to offer our communities the option to use data collections (possibly in the form of project portals) or data papers to provide collective citations. The AmeriFlux collaboration has also been leaning toward using data collections and writing data papers describing processing methods and statistics about the data to include in the collection. Tracking citations of a small number of data collections or data papers has the potential to make the implementation of these feasible using manual or semi-automated tracking processes. Our near-term strategy will include the ability to help our communities to move to a more open data usage policy such as CC BY 4.0 to enable easier creation of data collections and publication of data papers and other papers. This will also help remove a barrier to data integration.

Although there are many benefits of data collections, dynamic data citations, and data papers, no one approach solves all problems across use cases to enable citing and giving proper credit to a large number of datasets from across many repositories in a space limited paper. The solution may involve a hybrid approach with a 'container' that includes well-structured, machine readable citations to data packages, data collections, dynamic data citations, and data papers, along with other documentation and metadata needed to understand the purpose of the container. The container would need a unique identifier (maybe a DOI) and citation format that can be used to reference it in papers. Then we would need to have methods for managing the citation counts for the contents of the container that give the items in the container credit as if they were directly cited in the reference section of the paper.

Ideally, these containers could then be used by the author(s) of a paper to create a compact citation for a large number of data citations or by a group wishing to make a thematic container from a set of underlying data packages (e.g. an AmeriFlux data release, the datasets describing a watershed, etc). Scientific authors sometimes do this through writing a data paper that is then cited by the research paper. Today that means that the dataset becomes a secondary rather than a primary citation in this approach, and ultimately receives no direct citation. Instead of a data paper scientific authors could create a data container to reference. An early example of a capability similar to this 'container' system is where a machine readable, well-formatted supplementary information section for a scientific paper is required. However, there is still much work needed to define and standardize containers, build the support infrastructure, and achieve broad acceptance. In addition, the credit counting mechanisms would also need to be included in this process. It is our hope that leveraging all the work on collective data citation methods that has already occurred will enable rapid development of something like this container capability.

7. Summary

In recent years, scientific communities have used the paper publication paradigm to capture the data publication process. While this has provided a way for the community to leap forward, our experiences as outlined in this paper further expose the myth that data publication is like publishing a paper (Horsburgh et al., 2020).

The paper publication process has evolved to be a static entity that captures the research process and the end results. However, data publication is a more complex, iterative, and dynamic process where producers have raw data, data from QA/QC practices, derivatives of the data, and versions of the data that evolve over the life cycle of the scientific process. By using the paper publication paradigm for data, citations only capture a snapshot of the data at a particular time, often losing valuable information including context of the data, relationship to other data in the collection, etc.

A key element of data integration is enabling the use and citation of that integrated data. In this paper, we have considered several approaches (Scalable Dynamic Data Citation, data collections, data papers) that attempt to address some of the challenges. However, there is still a need to systematically evaluate the complex socio-technical issues around data producers and data consumers. Data producers are quickly becoming an important fabric of the complex scientific ecosystem enabling new modeling and machine learning/artificial intelligence

capabilities. There is a need to address the scalability and autonomy needs of data producers in creating and managing data citations while also addressing the diverse use cases and usability needs of data consumers.

Declaration of Competing Interest

None.

Acknowledgements

We want to thank the members of the ESS-DIVE, AmeriFlux, FLUX-NET, Watershed Function SFA, and NGEE-Tropics for the many useful discussions about citations and credit. We also thank all those who contributed to a variety of community discussions on data citations through a variety of means and community meetings. We want to particularly thank Shelley Stall, Peter Fox, and Martin Fenner for being willing to act as sounding boards for citation topics. This work was funded through the AmeriFlux Management Project and the ESS-DIVE repository by the U.S. DOE's Office of Science Biological and Environmental Research under contract number DE-AC02-05CH11231 to LBNL as part of its Earth and Environmental Systems Science Division Data Management program.

References

Ball-Damerow, J.E., Brenskelle, L., Barve, N., Soltis, P.S., Sierwald, P., Bieler, R., LaFrance, R., Ariño, A.H., Guralnick, R.P., 2019. Research applications of primary biodiversity databases in the digital age. PLoS One 14, e0215794. https://doi.org/10.1371/journal.pone.0215794.

Belter, C.W., 2014. Measuring the value of research data: a citation analysis of oceanographic data sets. PLoS One 9, e92590. https://doi.org/10.1371/journal.pone.0092590.

Borgman, C.L., 2016. Data citation as a bibliometric oxymoron. In: Sugimoto, C.R. (Ed.), Theories of Informetrics and Scholarly Communication. Walter de Gruyter GmbH & Co KG, Berlin, Boston, pp. 93–115.

Borrelli, L, van Selm, M, Hahnel, M, Crosas, M, Nosek, B, Shearer, K, Goodey, G, Hyndman, A, Baynes, G, 2020. The State of Open Data 2020: The longest-running longitudinal survey and analysis on open data. Digital Science, Springer Nature. https://doi.org/10.6084/m9.figshare.13227875.v1.

Cook, R.B., Vannan, S.K.S., McMurry, B.F., Wright, D.M., Wei, Y., Boyer, A.G., Kidder, J. H., 2016. Implementation of data citations and persistent identifiers at the ORNL DAAC. Ecol. Inform. 33, 10–16. https://doi.org/10.1016/j.ecoinf.2016.03.003.

Costello, M.J., Michener, W.K., Gahegan, M., Zhang, Z.-Q., Bourne, P.E., 2013. Biodiversity data should be published, cited, and peer reviewed. Trends Ecol. Evol. 28, 454–461. https://doi.org/10.1016/j.tree.2013.05.002.

Cousijn, H., Kenall, A., Ganley, E., Harrison, M., Kernohan, D., Lemberger, T., Murphy, F., Polischuk, P., Taylor, S., Martone, M., Clark, T., 2018. A data citation roadmap for scientific publishers. Sci. Data 5, 1–11. https://doi.org/10.1038/sdata.2018.259.

Cousijn, H., Feeney, P., Lowenberg, D., Presani, E., Simons, N., 2019. Bringing citations and usage metrics together to make data count. Data Sci. J. 18, 9. https://doi.org/10.5334/dsj-2019-009.

Data Citation Synthesis Group, 2014. Joint Declaration of Data Citation Principles. FORCE 11. https://doi.org/10.25490/a97f-egyk.

DataCite Metadata Working Group, 2019. DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. DataCite e.V. https://doi.org/10.14454/7xq3-zf69

Duke, C.S., Porter, J.H., 2013. The ethics of data sharing and reuse in biology. Bioscience 63, 483–489. https://doi.org/10.1525/bio.2013.63.6.10.

Escribano, N., Galicia, D., Ariño, A.H., 2018. The tragedy of the biodiversity data commons: a data impediment creeping nigher? Database 2018. https://doi.org/10.1093/database/bay033.

ESIP Data Preservation and Stewardship Committee, 2019. Data Citation Guidelines for Earth Science Data, Version 2. Earth Science Information Partners. https://doi.org/10.6084/m9.figshare.8441816.

Fane, B, Ayris, P, Hahnel, M, Hrynaszkiewicz, I, Baynes, G, Farrell, E, 2019. The State of Open Data 2019: A selection of analyses and articles about open data, curated by Figshare. Digital Science, Springer Nature. https://doi.org/10.6084/m9.figshare.9980783.v1.

Fenner, M., Crosas, M., Grethe, J.S., Kennedy, D., Hermjakob, H., Rocca-Serra, P., Durand, G., Berjon, R., Karcher, S., Martone, M., Clark, T., 2019. A data citation roadmap for scholarly data repositories. Sci. Data 6, 28. https://doi.org/10.1038/s41597-019-0031-8.

Fenner, M, Lowenberg, D, Jones, M, Needham, P, Vieglais, D, Abrams, S, Cruse, P, Chodacki, J, 2018. Code of practice for research data usage metrics release 1. PeerJ Preprints. https://doi.org/10.7287/peerj.preprints.26505v1.

Gries, C., Servilla, M., O'Brien, M., Vanderbilt, K., Smith, C., Costa, D., Grossman-Clarke, S., 2019. Achieving FAIR data principles at the environmental data initiative, the US-LTER data repository. Biodiversity Inform. Sci. Stand. 3, e37047.

Horsburgh, J.S., Hooper, R.P., Bales, J., Hedstrom, M., Imker, H.J., Lehnert, K.A., Shanley, L.A., Stall, S., 2020. Assessing the state of research data publication in hydrology: a perspective from the Consortium of Universities for the Advancement of Hydrologic Science, Incorporated. WIREs Water 7, 1. https://doi.org/10.1002/wat2.1422.

Hubbard, S.S., Williams, K.H., Agarwal, D., Banfield, J., Beller, H., Bouskill, N., Brodie, E., Carroll, R., Dafflon, B., Dwivedi, D., Falco, N., Faybishenko, B., Maxwell, R., Nico, P., Steefel, C., Steltzer, H., Tokunaga, T., Tran, P.A., Wainwright, H., Varadharajan, C., 2018. The East River, Colorado, Watershed: a mountainous community testbed for improving predictive understanding of multiscale hydrological-biogeochemical dynamics. Vadose Zone J. 17, 180061. https://doi.org/10.2136/vzj2018.03.0061.

Hunter, J., Hsu, C.-H., 2015. Formal acknowledgement of citizen scientists' contributions via dynamic data citations. In: Digital Libraries: Providing Quality Information. Springer International Publishing, pp. 64–75. https://doi.org/10.1007/978-3-319-27974-9_7.

Kakalia, Z., Damerow, J., Agarwal, D., Cholia, S., Elbashandy, H., Hendrix, V., Jones, C. S., Jones, M.B., Slaughter, P., O'Brien, F., Robles, E., Snavely, C., Whitenack, K., Varadharajan, C., 2019. Standardizing metadata quality review for an environmental data repository. In: AGU Fall 2019 Meeting. IN14B–09.

Kakalia, Z., Varadharajan, C., Alper, E., Brodie, E., Burrus, M., Carroll, R., Christianson, D., Hendrix, V., Henderson, M., Hubbard, S., Others, 2020. The East River Community Observatory Data Collection: Diverse, Multiscale Data from a Mountainous Watershed in the East River. Authorea Preprints, Colorado.

Khan, N., Thelwall, M., Kousha, K., 2020. Data citation and reuse practice in biodiversity - Challenges of adopting a standard citation model. 17th International Conference on Scientometrics & Informetrics, http://hdl.handle.net/2436/623005.

Kratz, J.E., Strasser, C., 2015. Researcher perspectives on publication and peer review of data. PLoS One 10, e0117619. https://doi.org/10.1371/journal.pone.0117619.

Lin, D., Crabtree, J., Dillo, I., Downs, R.R., Edmunds, R., Giaretta, D., De Giusti, M., L'Hours, H., Hugo, W., Jenkyns, R., Khodiyar, V., Martone, M.E., Mokrane, M., Navale, V., Petters, J., Sierman, B., Sokolova, D.V., Stockhause, M., Westbrook, J., 2020. The TRUST principles for digital repositories. Sci. Data 7, 144. https://doi.org/10.1038/s41597-020-0486-7.

Mooney, H., Newton, M.P., 2012. The anatomy of a data citation: discovery, reuse, and credit. J. Librariansh. Sch. Commun. 1, 1035. https://doi.org/10.7710/2162-3309.1035.

Nosek, B.A., Alter, G., Banks, G.C., Borsboom, D., Bowman, S.D., Breckler, S.J., Buck, S., Chambers, C.D., Chin, G., Christensen, G., Contestabile, M., Dafoe, A., Eich, E., Freese, J., Glennerster, R., Goroff, D., Green, D.P., Hesse, B., Humphreys, M., Ishiyama, J., Karlan, D., Kraut, A., Lupia, A., Mabry, P., Madon, T.A., Malhotra, N., Mayo-Wilson, E., McNutt, M., Miguel, E., Paluck, E.L., Simonsohn, U., Soderberg, C., Spellman, B.A., Turitto, J., VandenBos, G., Vazire, S., Wagenmakers, E.J., Wilson, R., Yarkoni, T., 2015. Scientific standards. Promoting an open research culture. Science 348, 1422–1425. https://doi.org/10.1126/science.aab2374.

Novick, K.A., Biederman, J.A., Desai, A.R., Litvak, M.E., Moore, D.J.P., Scott, R.L., Torn, M.S., 2018. The AmeriFlux network: a coalition of the willing. Agric. For. Meteorol. 249, 444–456. https://doi.org/10.1016/j.agrformet.2017.10.009.

O'Brien, M., Costa, D., Servilla, M., 2016. Ensuring the quality of data packages in the LTER network data management system. Ecol. Inform. 36, 237–246. https://doi.org/10.1016/j.ecoinf.2016.08.001.

Office of Science, 2013. Statement on Digital Data Manage [WWW Document]. URL. https://science.osti.gov/Funding-Opportunities/Digital-Data-Management (accessed 11.4.20).

Parsons, M.A., Duerr, R.E., Jones, M.B., 2019. The history and future of data citation in practice. Data Sci. J. 18, 52. https://doi.org/10.5334/dsj-2019-052.

Pastorello, G., Trotta, C., Canfora, E., Chu, H., Christianson, D., Cheah, Y.-W., Poindexter, C., Chen, J., Elbashandy, A., Humphrey, M., Isaac, P., Polidori, D., Ribeca, A., van Ingen, C., Zhang, L., Amiro, B., Ammann, C., Arain, M.A., Ardö, J., Arkebauer, T., Arndt, S.K., Arriga, N., Aubinet, M., Aurela, M., Baldocchi, D., Barr, A., Beamesderfer, E., Marchesini, L.B., Bergeron, O., Beringer, J., Bernhofer, C., Berveiller, D., Billesbach, D., Black, T.A., Blanken, P.D., Bohrer, G., Boike, J., Bolstad, P.V., Bonal, D., Bonnefond, J.-M., Bowling, D.R., Bracho, R., Brodeur, J., Brümmer, C., Buchmann, N., Burban, B., Burns, S.P., Buysse, P., Cale, P., Cavagna, M., Cellier, P., Chen, S., Chini, I., Christensen, T.R., Cleverly, J., Collalti, A., Consalvo, C., Cook, B.D., Cook, D., Coursolle, C., Cremonese, E., Curtis, P.S., D'Andrea, E., da Rocha, H., Dai, X., Davis, K.J., De Cinti, B., de Grandcourt, A., De Ligne, A., De Oliveira, R.C., Delpierre, N., Desai, A.R., Di Bella, C. M., di Tommasi, P., Dolman, H., Domingo, F., Dong, G., Dore, S., Duce, P., Dufrêne, E., Dunn, A., Dušek, J., Eamus, D., Eichelmann, U., ElKhidir, H.A.M., Eugster, W., Ewenz, C.M., Ewers, B., Famulari, D., Fares, S., Feigenwinter, I., Feitz, A., Fensholt, R., Filippa, G., Fischer, M., Frank, J., Galvagno, M., Gharun, M., Gianelle, D., Gielen, B., Gioli, B., Gitelson, A., Goded, I., Goeckede, M., Goldstein, A. H., Gough, C.M., Goulden, M.L., Graf, A., Griebel, A., Gruening, C., Grünwald, T., Hammerle, A., Han, S., Han, X., Hansen, B.U., Hanson, C., Hatakka, J., He, Y., Hehn, M., Heinesch, B., Hinko-Najera, N., Hörtnagl, L., Hutley, L., Ibrom, A., Ikawa, H., Jackowicz-Korczynski, M., Janouš, D., Jans, W., Jassal, R., Jiang, S., Kato, T., Khomik, M., Klatt, J., Knohl, A., Knox, S., Kobayashi, H., Koerber, G., Kolle, O., Kosugi, Y., Kotani, A., Kowalski, A., Kruijt, B., Kurbatova, J., Kutsch, W.L., Kwon, H., Launiainen, S., Laurila, T., Law, B., Leuning, R., Li, Y., Liddell, M., Limousin, J.-M., Lion, M., Liska, A.J., Lohila, A., López-Ballesteros, A., López-Blanco, E., Loubet, B., Loustau, D., Lucas-Moffat, A., Lüers, J., Ma, S., Macfarlane, C., Magliulo, V., Maier, R., Mammarella, I., Manca, G., Marcolla, B., Margolis, H.A.,

Marras, S., Massman, W., Mastepanov, M., Matamala, R., Matthes, J.H., Mazzenga, F., McCaughey, H., McHugh, I., McMillan, A.M.S., Merbold, L., Meyer, W., Meyers, T., Miller, S.D., Minerbi, S., Moderow, U., Monson, R.K., Montagnani, L., Moore, C.E., Moors, E., Moreaux, V., Moureaux, C., Munger, J.W., Nakai, T., Neirynck, J., Nesic, Z., Nicolini, G., Noormets, A., Northwood, M., Nosetto, M., Nouvellon, Y., Novick, K., Oechel, W., Olesen, J.E., Ourcival, J.-M., Papuga, S.A., Parmentier, F.-J., Paul-Limoges, E., Pavelka, M., Peichl, M., Pendall, E., Phillips, R.P., Pilegaard, K., Pirk, N., Posse, G., Powell, T., Prasse, H., Prober, S.M., Rambal, S., Rannik, Ü., Raz-Yaseef, N., Reed, D., de Dios, V.R., Restrepo-Coupe, N., Reverter, B.R., Roland, M., Sabbatini, S., Sachs, T., Saleska, S.R., Sánchez-Cañete, E. P., Sanchez-Mejia, Z.M., Schmid, H.P., Schmidt, M., Schneider, K., Schrader, F., Schroder, I., Scott, R.L., Sedlák, P., Serrano-Ortíz, P., Shao, C., Shi, P., Shironya, I., Siebicke, L., Šigut, L., Silberstein, R., Sirca, C., Spano, D., Steinbrecher, R., Stevens, R.M., Sturtevant, C., Suyker, A., Tagesson, T., Takanashi, S., Tang, Y., Tapper, N., Thom, J., Tiedemann, F., Tomassucci, M., Tuovinen, J.-P., Urbanski, S., Valentini, R., van der Molen, M., van Gorsel, E., van Huissteden, K., Varlagin, A., Verfaillie, J., Vesala, T., Vincke, C., Vitale, D., Vygodskaya, N., Walker, J.P., Walter-Shea, E., Wang, H., Weber, R., Westermann, S., Wille, C., Wofsy, S., Wohlfahrt, G., Wolf, S., Woodgate, W., Li, Y., Zampedri, R., Zhang, J., Zhou, G., Zona, D., Agarwal, D., Biraud, S., Torn, M., Papale, D., 2020. The FLUXNET2015 dataset and the ONEFlux processing pipeline for eddy covariance data. Sci. Data 7, 225. https://doi.org/10.1038/s41597-020-0534-3.

Piwowar, H, Vision, T, 2013. Data reuse and the open data citation advantage. PeerJ. https://doi.org/10.7717/peerj.175.

Rauber, A., Asmi, A., Van Uytvanck, D., Proell, S., 2016a. Data citation of evolving data: recommendations of the RDA working group on data citation (WGDC). Res. Data Alliance. https://doi.org/10.15497/rda00016.

Rauber, A., Asmi, A., Van Uytvanck, D., Proell, S., 2016b. Identification of reproducible subsets for data citation, sharing and re-use. Bull. IEEE Tech. Commit. Digital Libraries 12, 6–15 (Special Issue on Data Citation).

Silvello, G., 2018. Theory and practice of data citation. J. Assoc. Inf. Sci. Technol. 69, 6–20. https://doi.org/10.1002/asi.23917.

Stall, S., Yarmey, L., Cutcher-Gershenfeld, J., Hanson, B., Lehnert, K., Nosek, B., Parsons, M., Robinson, E., Wyborn, L., 2019. Make scientific data FAIR. Nature 570, 27. https://doi.org/10.1038/d41586-019-01720-7.

Tenopir, C., Dalton, E.D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D., Dorsett, K., 2015. Changes in data sharing and data reuse practices and perceptions among scientists worldwide. PLoS One 10, e0134826. https://doi.org/10.1371/journal.pone.0134826.

Varadharajan, C., Cholia, S., Snavely, C., Hendrix, V., Procopiou, C., Riley, W., Agarwal, D.A., 2018. Launching an accessible archive of environmental data. Eos 100. https://doi.org/10.1029/2019EO111263.

Varadharajan, C., Agarwal, D.A., Brown, W., Burrus, M., Carroll, R.W.H., Christianson, D. S., Dafflon, B., Dwivedi, D., Enquist, B.J., Faybishenko, B., Others, 2019. Challenges in building an end-to-end system for acquisition, management, and integration of diverse data from sensor networks in watersheds: lessons from a mountainous community observatory in East River, Colorado. IEEE Access 7, 182796–182813.

Wallis, J.C., Rolando, E., Borgman, C.L., 2013. If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. PLoS One 8, e67332. https://doi.org/10.1371/journal.pone.0067332.

Weigel, T., Almas, B., Baumgardt, F., Zastrow, T., Schwardmann, U., Hellström, M., Quinteros, J., Fleischer, D., 2017. Recommendation on research data collections. Res. Data Alliance. https://doi.org/10.15497/RDA00022.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. The FAIR guiding principles for scientific data management and stewardship. Sci. Data 3, 160018. https://doi.org/10.1038/sdata.2016.18.

Wilkinson, M.D., Dumontier, M., Sansone, S.-A., Bonino da Silva Santos, L.O., Prieto, M., Batista, D., McQuilton, P., Kuhn, T., Rocca-Serra, P., Crosas, M., Schultes, E., 2019. Evaluating FAIR maturity through a scalable, automated, community-governed framework. Sci. Data 6, 1–12. https://doi.org/10.1038/s41597-019-0184-5.

Witt, M., Stall, S., Duerr, R., Plante, R., Fenner, M., Dasler, R., Cruse, P., Hou, S., Ulrich, R., Kinkade, D., 2019. Connecting researchers to data repositories in the earth, space, and environmental sciences. In: Digital Libraries: Supporting Open Science. Springer International Publishing, pp. 86–96. https://doi.org/10.1007/978-3-030-11226-4_7.