Horsburgh Jeffery (Orcid ID: 0000-0002-0768-3196)

Shanley Lea (Orcid ID: 0000-0001-8449-4615)

Stall Shelley (Orcid ID: 0000-0003-2926-8353)

WIREs Wiley Interdisciplinary Reviews

Article Title: Assessing the State of Research Data Publication in Hydrology: A CUAHSI Perspective

Article Type: Overview

Authors:

First author
Jeffery S. Horsburgh*, ORCID: https://orcid.org/0000-0002-0768-3196, Department of Civil and Environmental Engineering and Utah Water Research Laboratory, Utah State University, Logan, UT, Email: jeff.horsburgh@usu.edu
This author has no conflicts of interest.

Second author
Richard P. Hooper,ORCID: https://orcid.org/0000-0002-3329-9622, Department of Civil and Environmental Engineering, Tufts University, Medford, MA, Email: richard.hooper@tufts.edu
This author has no conflicts of interest.

Third author
Jerad Bales, ORCID: https://orcid.org/0000-0001-8398-6984 , Consortium of Universities for the Advancement of Hydrologic Science, Inc., Cambridge, MA, Email: jdbales@cuahsi.org
This author has no conflicts of interest.

Fourth author
Margaret Hedstrom, ORCID: https://orcid.org/0000-0002-0356-6806, School of Information, University of Michigan, Ann Arbor, MI, Email: hedstrom@umich.edu
This author has no conflicts of interest.

Fifth author
Heidi J. Imker, ORCID: https://orcid.org/0000-0003-4748-7453, University Library, University of Illinois, Urbana, IL, Email: imker@illinois.edu
This author has no conflicts of interest.

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/wat2.1422

Sixth author
Kerstin A. Lehnert, ORCID: https://orcid.org/0000-0001-7036-1977, Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY, Email: lehnert@ldeo.columbia.edu
This author has no conflicts of interest.

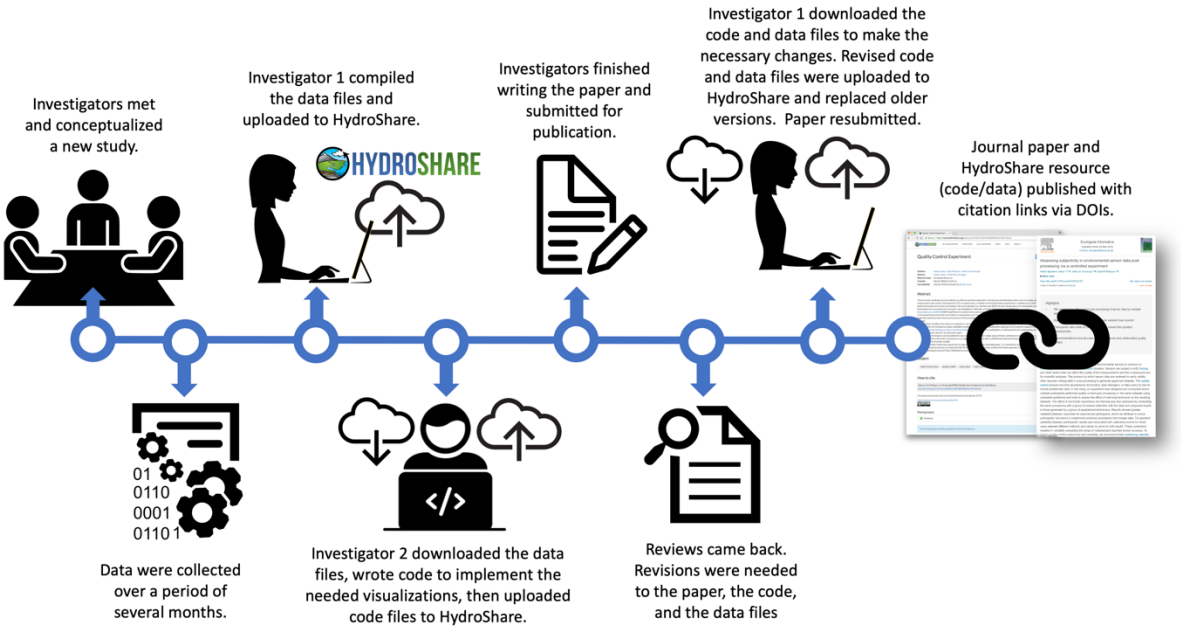
Seventh author
Lea A. Shanley, ORCID: https://orcid.org/0000-0001-8449-4615, Nelson Institute for Environmental Studies, University of Wisconsin-Madison, Madison, WI, Email: lshanley@wisc.edu
This author has no conflicts of interest.

Eighth author
Shelley Stall, ORCID: https://orcid.org/0000-0003-2926-8353, American Geophysical Union, Washington, DC, Email: sstall@agu.org
This author has no conflicts of interest.

Abstract

Many have argued that datasets resulting from scientific research should be part of the scholarly record as first class research products. Data sharing mandates from funding agencies and scientific journal publishers along with calls from the scientific community to better support transparency and reproducibility of scientific research have increased demand for tools and support for publishing datasets. Hydrology domain-specific data publication services have been developed alongside more general purpose and even commercial data repositories. Prominent among these are the Hydrologic Information System and HydroShare repositories developed by the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI). More broadly, however, multiple organizations have been involved in the practice of data publication in the hydrology domain, each having different roles that have shaped data publication and reuse. Bibliographic and archival approaches to data publication have been advanced, but both have limitations with respect to hydrologic data. Specific recommendations for improving data publication infrastructure, support, and practices to move beyond existing limitations and enable more effective data publication in support of scientific research in the hydrology domain include: improving support for journal article-based data access and data citation, considering the workflow for data publication, enhancing support for reproducible science, encouraging publication of curated reference data collections, advancing interoperability standards for sharing data and metadata among repositories, developing partnerships with university libraries offering data services, and developing more specific data management plans. While presented in the context of CUAHSI's data repositories and experience, these recommendations are broadly applicable to other domains.

Graphical/Visual Abstract and Caption



Caption: Depiction of an actual workflow from the HydroShare data and model repository demonstrating new capabilities for collaborative data publication that have been shaped by multiple years of experience in providing data publication services for the hydrology community.

Introduction

Scientific data publication mechanisms used by hydrologists have matured over the past decade with a diversity of data repositories coming online. These include domain-specific data centers, general purpose repositories, government sponsored repositories, data publication and archiving services offered by university libraries, and even private sector, commercial repositories (Table 1). The availability and use of these repositories augment past approaches, including making data available by request from the authors of journal articles, publishing datasets as supplementary information to journal articles, or sharing data via independent, peer-reviewed data papers.

Table 1. Example scientific data repositories.

Repository Name	Repository Type	Description
HydroShare	Domain specific	System operated by the Consortium of Universities for the Advancement of Hydrologic Science Inc. (CUAHSI) that enables sharing and publication of data and models in a citable and discoverable manner. URL: http://www.hydroshare.org
CUAHSI Hydrologic Information	Domain Specific	An Internet-based system for sharing time series of hydrologic data comprised of databases and servers connected through web services to client applications, allowing for the publication, discovery, and access of

System (HIS)		data. URL: https://www.cuahsi.org/data-models/discovery-and-analysis
Dryad	General Purpose	A general purpose repository for making research data discoverable, freely reusable, and citable. Dryad provides a general-purpose home for a wide diversity of data types. URL: https://datadryad.org/stash
Zenodo	General Purpose	General purpose repository hosted by the European Organization for Nuclear Research (CERN) for sharing research outputs from all fields of research. URL: https://zenodo.org/
European Open Science Cloud (EOSC)	Government Sponsored	A virtual environment with open services for storage, management, analysis, and re-use of research data, across borders and scientific disciplines for EU Member States. URL: https://eosc-portal.eu/
Figshare	Commercial	A commercially funded repository that allows users to upload any file format and research output for dissemination. URL: https://figshare.com/

Development of these repositories and publication services has been driven by requirements from funding agencies to publish data resulting from funded projects, by the publishers of scientific journals that require a statement of data availability and citations for datasets associated with published papers, by heightened demands for the reproducibility of research results, and by investigators seeking secure storage for their data and greater visibility of their data and research products. Indeed, there has been a major push within hydrology and across the community of scientific data producers and publishers to make data resulting from scientific research Findable, Accessible, Interoperable, and Reusable, resulting in a set of FAIR Data Principles focused on enhancing the ability of machines to automatically find and use the data as well as supporting its reuse by individuals (Wilkinson et al., 2016).

The increasing practice of Data Science in environmental applications – i.e., transforming data into understandable and actionable knowledge relevant for informed decision making (Gibert et al., 2018) – is also influencing hydrology, particularly with the application of machine learning and deep learning techniques to emerging large data sets generated by *in situ* sensors and by aerial and satellite remote sensing (Shen, 2018). Advancing and comparing these methods requires the availability of shared example, training, and benchmark datasets, a pattern that has been demonstrated across many domains where Data Science methods are employed (e.g., Deng et al., 2009; Wu et al., 2018). Recent investments in Big Data Regional Innovation Hubs by the U.S. National Science Foundation are encouraging the use of data science approaches to scientific and societal challenges in disciplines like hydrology that are just beginning to explore data science approaches.

In the related field of geochemistry, international collaboration among data providers is emerging to dramatically increase the volume of data available for advanced data mining, data analysis, and machine learning. Similarly, the National Institutes of Health (NIH) has developed a strategic plan for

data science, and is leading efforts to maximize the benefits from data and compute in the cloud by empowering broad and meaningful data sharing through initiatives like "data commons,"[1] and by fostering open science best practices and policies (e.g., Das et al., 2017; Kiar et al., 2017; Poldrack et al., 2019). Efforts have also been under way for some time within the climate science community to enhance access to large climate simulation data (e.g., Williams et al., 2009), and the FAIR data principles are driving new innovation in the communities, repositories, and services available to scientists working in many domains – e.g., biodiversity science and geoscience (Lannom et al., 2020), geosciences and chemistry (Stall et al., 2020), etc.

While significant progress has been made in both availability of tools and repositories for sharing and publishing scientific data and in the culture and attitudes of scientists regarding the practice of data publication, there are still several challenges to be met and improvements that can be made. This overview discusses those challenges, the current roles of different organizations involved in the practice of data publication in the hydrology domain, and how these roles have shaped data publication and reuse. We describe different fundamental approaches to data publication and provide perspective for how we might move beyond their existing limitations. Finally, we conclude with a set of specific recommendations that we believe will enable more effective data publication in support of scientific research. While we broadly discuss the state of research data publication in the hydrology domain and the different organizations and roles involved, we have included specific examples and discussion surrounding the data publication systems created and operated by CUAHSI for two reasons. First, there are still few hydrology domain-specific repositories that openly accept submission of research data products for sharing and publication. Although a search for hydrologic data repositories within the Registry of Research Data Repositories (re3data.org, 2020) using DataCite's Repository Finder tool[2] returned more than 60 repositories, only the CUAHSI repositories were identified as accepting open data submissions, whereas the rest were project, geographic area, or agency/organization specific. Second, the CUAHSI tools represent state of the practice systems that illustrate existing capabilities and highlight opportunities for improvement.

DATA PUBLICATION CHALLENGES

Several challenges remain that impact the effectiveness of existing systems that accept open submissions of data resulting from research projects, each of which may have technical and social aspects. For instance, choosing among the variety of available repositories can be difficult for researchers and is akin to choosing an appropriate journal to which their paper can be submitted. With the growing number of repository choices, inconsistency in how datasets are organized and packaged by different repositories can pose difficulties, with some imposing restrictions on the file formats and syntax for submitted datasets, while others impose no restrictions. Where restrictions

[1] Examples include https://www.braincommons.org/ and other data commons pilots like https://commonfund.nih.gov/commons

[2] https://repositoryfinder.datacite.org/

are not imposed, it is left to researchers to decide what should be deposited and how the content should be organized. This leads to another challenge involving how to address the quality of submitted data, which is dependent not only the methods and care used to produce the data, but also on the level of effort made to ensure the data are well curated and described. As data are collected, manipulated, and transformed, it can be easy to lose sight of (and potentially omit a description of) the many potential sources of error and bias that can accrue along the way (e.g., Wilby et al., 2017). Metadata accompanying published datasets rarely contain this level of data quality information.

Publishing data requires significant effort from researchers, and incentives are not always adequate to motivate participation (Bierer et al., 2017). The culture of academia still does not view the publication of high-quality datasets in the same way as publication of peer-reviewed journal articles or other formal research products that have a much longer history of being recognized as scholarly productivity for promotion and tenure decisions. Beyond academic credit issues, some data, such as social science data involving human subjects, involve sensitive information that complicates data sharing. Additional effort may be required during research planning stages and after data collection to ensure Institutional Review Board (IRB) protocols allow sensitive data to be released after they have been appropriately anonymized, aggregated, and/or summarized (Flint et al., 2017).

Sustainability and longevity of repositories is another major challenge. Some repositories grew out of research and development projects funded by agencies like the U.S. National Science Foundation, while others grew from commercial ventures. Neither scenario comes with guarantees of long-term funding support. To survive, repositories must develop sustainability/business models to ensure that archives are supported in the long term, making the case for which can be a difficult value proposition. Available resources must not only support the technical repository operation (e.g., maintaining websites, storage hardware, etc.) but also the provision of preservation and archival services (e.g., ensuring the integrity of artifacts over time).

ORGANIZATIONAL ROLES IN DATA PUBLICATION

In this discussion, we consider five different organizations involved in publication of hydrologic data that have similar but distinct objectives and may provide overlapping services (Table 2). Domain-specific repositories have a broad interest in serving their respective scientific communities in all aspects of data publication. They seek to be recognized by scientists within the domain by tailoring technologies to meet their needs and easing the burden of data publication and sharing. They may promote specific metadata standards and common data formats to promote data interoperability and to better enable value-added functionality for community members (e.g., data preview and automated validation of metadata completeness).

General purpose repositories provide services similar to those provided by domain-specific repositories for data publication and archival, but generally employ simpler and more general

purpose metadata standards like Dublin Core (DCMI, 2012) and rarely limit uploaded file/content types. Most general purpose repositories offer free deposition, but may limit the size or configuration of sets of files. University libraries operate much like general purpose repositories, but are focused on serving the needs of their own faculty, students, and researchers along with meeting legal mandates for sharing data produced by sponsored research on their campus. They often provide campus-based repositories and services for data publication and archiving that are, in many cases based on commercial software that may even be the same software used by general purpose repositories (e.g., Figshare for Institutions). They differ from general purpose repositories when they offer additional services such as advice on data management plans, assistance with curation of datasets, and metadata, data file, and supplemental documentation review.

Scientific journal publishers are increasingly playing an important role related to data publication as their policies evolve to encourage or even require authors to deposit the data supporting their published research in a trusted repository where it is preserved, well-documented, citable, and discoverable as an independent scholarly product. Many journals still provide authors with the ability to submit data as supplemental materials supporting papers, and some journal publishers even provide their own data repositories (e.g., the Mendeley Data[3] repository is owned by parent company Elsevier). Some journals are now requiring authors to include data availability statements and data citations with globally-resolvable, persistent identifiers that link to the actual dataset, supporting the integrity of the paper, transparency and reproducibility of the work, and ensuring appropriate credit to data authors (Stall et al., 2018).

Finally, research technology centers are building new Data Science and computational tools and capabilities that are being used by scientists to create new research products that then become artifacts that need to be preserved, shared, and published.

Table 2. Organizations involved in publication of hydrologic data.

Organization	**Examples**	**Data Publication Role/Objective**	**Description of Services Provided**
Domain-specific repositories	CUAHSI HIS, HydroShare, EarthChem Library[4]	Enable deposit, curation, and publication of research data	• Provision of pre-publication workspace (i.e., a place to put things while they are being worked on) • Formal data publication and digital object identifier (DOI) provision • Provision of post-publication data archiving (i.e., the final, published location of the data) • Promotion of specific metadata profiles • Support for dataset and file types commonly used by community members • Promoting interoperability among common data types • Functionality for data preview

[3] https://data.mendeley.com/
[4] http://www.earthchem.org/portal

General purpose repositories (including governmental and commercial)	Dryad, Figshare	Enable deposit, curation, and publication of research data	• Formal data publication and DOI provision • Provision of post-publication data archiving • Standard schemas for discovery • Usage license options for depositors • Metadata elements aligning with simplified or general purpose standards • May integrate with some scientific journals
University libraries	University of California Libraries, Utah State University Library	Enable deposit, curation, and publication of research data	• Formal publication and DOI provision • Provision of post-publication data archiving • Data management (e.g., advice on data management plans) • Curation, including metadata review and enhancement, file review, and supplemental documentation • Standardized metadata valuable to all datasets and digital objects – e.g., Dublin Core or DataCite (DataCite Metadata Working Group, 2019)
Scientific journal publishers	Elsevier, American Geophysical Union, Springer, Nature	Enable peer review and publication of research results based on data	• Formal publication and DOI provision for scientific papers based on data generated by research • Peer review and editorial support of primary paper content • Wide variation in review practices for supplemental material content. Data stored in supplements is usually not curated nor indexed for discovery • Promotion of policies related to data accompanying scientific papers
Research technology centers	NSF-funded Regional Big Data Innovation Hubs (BD Hubs) and virtual Extreme Science and Engineering Discovery Environment (XSEDE)	Assist scientists in creating data and other research results that need to be stored and published	• Seek to build capacity for new technologies (e.g., Data Science, Big Data) by providing access to cloud computing (e.g., Open Storage Network, Microsoft Azure, Amazon Web Services (AWS), Google Earth Engine), high performance computing, and trainings • Help scientists across a range of domains to generate, analyze, mine, and manipulate data sets (and model output) to generate finalized datasets and other research products

Each of these entities have developed independently, with coalitions being formed among some of them. The Research Data Alliance (RDA), for example, is an international forum for developing standards, tools and best practices for open sharing of research data.[5] As another example, the Data Curation Network[6], which is a partnership among university libraries, data repositories, and scholars, is building a network of human expertise across institutions to provide curation support that is

[5] https://www.rd-alliance.org
[6] https://datacurationnetwork.org/

discipline and/or format specific. Our experience has been that there remain significant opportunities for these entities to work together more closely to improve opportunities, available tools, and best practices surrounding scientific data publication. There is a particular need for domain-specific repositories, general purpose repositories, journal publishers, and university library-based data services to clarify their roles and unique contributions to data publication, preservation, and access. In the following section, we describe in more detail specific use cases for data publication and reuse, after which we discuss differences among the major approaches for providing this functionality. We conclude with specific recommendations for improving data publication practices.

USE CASES FOR DATA PUBLICATION AND REUSE

Enabling the reproducibility of scientific results is one purpose for data publication that seems relatively straightforward. A common practice has been to place data, computer code, and instructions describing the workflows necessary to reproduce an article's findings in supplementary material referenced by a journal article. Supplementary material may be included with the article in the journal's archive or it may be deposited in one or more separate repositories. Despite this being generally accepted as a common practice to enhance reproducibility, it rarely achieves that purpose. In a study of 360 of the 1,989 articles published by six hydrology and water resources journals in 2017, Stagge et al. (2019) were only able to reproduce the results of 1.6% of the articles they tested using their available artifacts. Other studies from different domains have found similar results (e.g., Aarts et al., 2015; Baker, 2016; Stodden et al., 2018). The study by Stagge et al. (2019) identified several factors that inhibited reproducibility, including complete inaccessibility of data, requirements to contact authors or a third party for access, lack of code used to generate results from data, and lack of instructions for using available artifacts, which clearly indicate significant opportunity for promotion of best practices in data/artifact publication to support reproducibility such as those suggested by Goodman et al. (2014). Nüst et al. (2017) studied this problem and suggested that lack of incentives and missing standardized infrastructure for providing research results such as data and source code along with a scientific paper are common causes. They suggested an "executable research compendium" as a new packaging mechanism for data, software, text, and a user interface description to better enable discovery, exploration, archival, and reuse of computer-based research.

Beyond the reproducibility use case, a prime purpose of data publication in hydrology is data re-use, particularly synthesis of existing data sets to develop new knowledge through reanalysis. Formal data publication seeks to make data more available to more people than informal peer-to-peer data sharing, but here the results have also been mixed (Pasquetto et al., 2017). Clarivate Analytics created the Data Citation Index in 2012 to index published data sources categorized as datasets, software, data studies, and repositories and now provides a search interface for over 380 data repositories worldwide (Clarivate Analytics, 2019). While more datasets are being published than ever before, formal citation metrics, if taken at face value, indicate that only a small number of them

are being reused. However, there are likely multiple factors at play. Poor data citation practice is one cause, with many informal data citations occurring intratextually (e.g., a mention in a paper's acknowledgements section) instead of as a formal citation in a paper's list of references that can more easily be tracked (Mayo et al., 2015). Furthermore, many scientists are more likely to cite a paper or report describing a dataset rather than including a formal citation to the data itself because they have been trained to cite publications, whereas they may not know how to cite the data directly. Indeed, proper data citation figures prominently in our recommendations for improving data publication practices near the end of this overview.

Initial Experience with CUAHSI HIS

An early hydrology domain-specific repository, the CUAHSI HIS (Tarboton et al., 2009; Horsburgh et al., 2009) warrants a closer look because it illustrates many of the challenges associated with data publication and reuse in the hydrology community. The HIS has experienced modest data access rates (on the order of hundreds of users per month) that have remained steady over the past several years. When the HIS was operationalized by CUAHSI it was anticipated that usage would grow as people learned of the provided services. While HIS has been successful in making data more available, it has not become the single, go-to repository for or source of data for hydrologists.

Why? Experience with HIS has shown that many data producers are not entirely aware of the services the HIS offers, they are faced with uncertainty about which repository they should use to publish their data, and conforming to the strict metadata standards and time-series structure of HIS are a significant barrier. Moreover, HIS supports only time series data, which is a very important class of data in hydrology, but HIS has no ability to handle other data types. Thus, it is only a partial solution to meet journal publication standards or grant data publication requirements

Beyond being unaware of its existence, another potential reason for limited use by data consumers is the lack of a critical mass of data – there must be enough data in a repository so that the chance of a scientist finding needed data is high enough to encourage repeated usage. The CUAHSI HIS attempted to overcome this problem for the hydrology community by providing proxy web services for access to and a central metadata catalog to support discovery of large government holdings of fixed-point time series data that are widely used by hydrologists (e.g., data from the U.S. Geological Survey's National Water Information System (NWIS) and from the National Aeronautical and Space Administration (NASA)), along with data contributions from university scientists.

As designed, a major advantage of CUAHSI HIS is the provisioning of data from multiple sources in a consistent format, using standardized metadata profiles, and providing access via standardized web services. For the first time, users could locate and access data from multiple agencies, organizations, and sources through a single map interface (Figure 1) and download the data into a simple tabular format with no programming. A library for the R Statistical Computing Environment (Kadlec et al., 2015) also allows data sets to be accessed via the HIS web services and downloaded into a data

frame object within R. However, the dominant data discovery use case provided by the main consumer-focused client application for the HIS —locating the data in a specific geographic area— seems not to be a dominant need of the scientific community as indicated by the modest usage of HIS. Additionally, as newer, general purpose data repositories became available after HIS was operationalized, some data producers have chosen to opt for the simpler data formatting and metadata requirements offered by these repositories, reducing the flow of data that might have otherwise been deposited in the HIS.

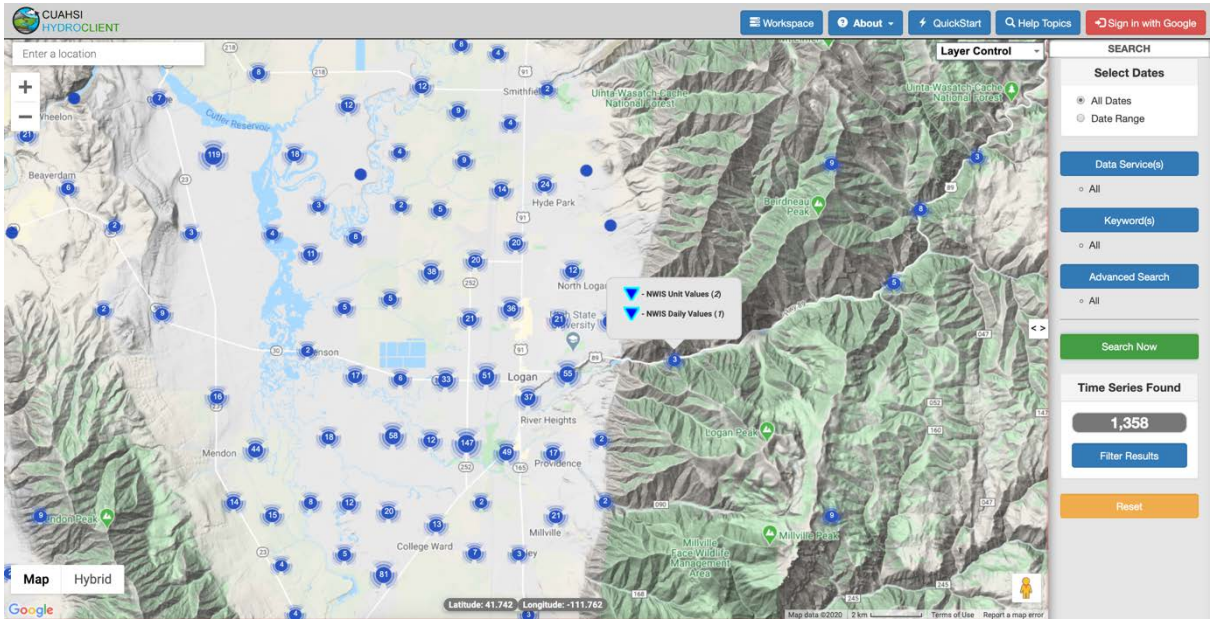


Figure 1. HydroClient web user interface for the CUAHSI HIS (http://data.cuahsi.org).

While the HIS was designed to make it easier to combine data from multiple sources, we are not aware of any major new data aggregations or synthesis products that have been created (e.g., new data collections with potentially different time or space domains derived from those contained within the HIS or new data products created by combining multiple datasets). In contrast, data synthesis in ecology and geochemistry seems to have been somewhat more successful than in hydrology. Synthesis centers, such as the National Center for Environmental Analysis and Synthesis (NCEAS), the National Socio-Environmental Synthesis Center (SESYNC), and the John Wesley Powell Center for Analysis and Synthesis have produced groundbreaking work in ecology and Earth science based on data re-use (e.g., Jackson et al., 2001; Thomas et al., 2004; Baron et al., 2017; Knox et al., 2019). Similarly, Hazen et al. (2019) see emerging large databases in mineralogy as enabling new data-driven discovery in that field. More research is needed to understand why these fields have seen greater success than hydrology in enabling data re-analysis. Synthesis centers, like NCEAS and the Powell Center, may be a critical ingredient in advancing data synthesis techniques, catalyzing the required collaborations, and facilitating data reuse. Hampton and Parker (2011) emphasize face-to-

face interaction, resident scientists at synthesis centers, multi-institutional collaboration, and participation in synthesis working groups as essential for leveraging synthesis to enhance scientific understanding, but make no mention of availability of data in a particular repository as a precursor for successful synthesis.

BIBLIOGRAPHIC VERSUS ARCHIVAL APPROACHES

While some of the reasons we provide for limited use of the CUAHSI HIS seem straightforward (e.g., only datasets consisting of time series of hydrologic observations from fixed monitoring sites are supported), two concepts from library and information science, ***bibliographic control*** versus ***archival control*** provide additional insight into why some traditional methods used in publication of other types of knowledge artifacts (e.g., books, journal articles, etc.) may not be as effective for data (Parsons and Fox, 2013; Kratz and Strasser, 2014). They also point to how data publication can be made more effective in light of the limitations discussed above.

Bibliographic control, which is focused on knowledge transfer and improving understanding beyond members of the original research team that produced a dataset, evolved out of library science where the goals of acquisition, organization, cataloging, and preservation revolved around acquiring books and other publications, creating a catalog entry for each item (i.e., author, title, publisher, date), and assigning one or more subject classification codes. The goals and practices of cataloging were established in the late 19th century to enable a reader to locate any library holding by author, title, subject, year, publisher, etc.

There are two salient points about the application of bibliographic techniques to data. First, the goal of bibliographic control is to allow users to find and retrieve an item by its title or author(s) and to find all items about a particular subject. Second, the bibliographic approach is applied to discrete items: a book, a sound recording, a map, etc., or a serial title supplemented by separate indexes and databases to provide access to subcomponents, such as journal articles. Bibliographic items are bounded and fixed, and there is an underlying assumption that each item's purpose and use is self-explanatory (although interpreting the contents may require domain expertise).

In contrast, *the archival approach*, which has more of a focus on reproducibility, is based on the concept of provenance, which tracks the production or assembly of a collection of documents or artifacts because they supported a particular function or served a particular purpose. The goal of archival control is to establish relationships between an archival collection and its context and is achieved through organization and description of collections. Archival collections are neither bounded nor fixed because materials can be added, deleted, and reorganized, and new relationships discovered both between material in a collection and among collections. However, archival collections rarely contain rich description at the item-level, making it challenging for users to know specific materials are held within a collection.

One of the challenges that arises conceptually from the notion of "data publication" is that it assumes that data are, or can be made into, publication-like entities that are amenable to bibliographic control. Each data publication would be bounded and fixed, discoverable by way of its author(s), title, dates, edition, subject matter, etc. This approach may work well for some static, well-structured data that are relatively self-explanatory, self-contained, or interpretable with a small amount of added metadata, such as a geospatial dataset, an image, or simple tabular data. However, data publication that relies strictly on basic bibliographic elements does not provide sufficient context for many other types of data: data that are dynamic (e.g., streaming data from environmental sensors that regularly change), pieces of a larger whole (e.g., data collections whose context is not captured by any individual element), or dependent on other pieces of data or code for meaningful use (e.g., packages whose data may be in one repository with related code in a different repository).

TOWARDS MORE EFFECTIVE DATA PUBLICATION

Considering these perspectives on publication approaches, some of the issues with data re-use become more apparent. Effective repurposing of data frequently requires context beyond simply a description of what was measured, where it was measured, and how it was measured. *Who* measured it? *Why* was it measured? *What* is it meant to represent? Providing sufficient context may explain why data re-use has been more successful in the geochemistry and ecology fields. Geochemical data are described within a mineralogical and lithological classification system that provides context. Ecological data described using Ecological Metadata Language (EML; Fegraus et al., 2005), which is used by Long-Term Ecological Research (LTER) sites, have rich metadata along with linkages to related data. Thus, climate data can be linked to soil metagenomics or soil biogeochemistry can be linked to plot biomass data. By itself, the CUAHSI HIS lacked this type of context information.

For published datasets that are linked to journal publications, contextual information may be captured within the linked publication. However, papers may only describe a subset of a larger dataset, and some datasets may have many associated papers (or none). Thus, we must consider whether and how archival systems can encode and preserve similar context with the data. This likely means promoting a culture of sharing data that goes beyond just making a minimal set of data files available. Because scientists are most likely to publish data in conjunction with publishing a paper or when completing a grant, data publication systems should focus on offering the ability to publish the collection of data (and potentially workflows) needed to support the paper or grant requirements. The data publication system should then enable access to the data at various levels of granularity – e.g., the entire collection or individual elements within that collection. The simplest way to discover the data would be to follow a formal citation of the dataset using its DOI from a journal article that cites it. However, for those datasets that do not have an associated journal paper, other discovery mechanisms could include geographic or keyword searches.

A hybrid approach is embedded into the DataCite metadata schema (DataCite Metadata Working Group, 2019), which can be used to register digital object identifiers (DOIs) to discrete datasets. The DataCite metadata schema contains elements (RelatedIdentifier, relatedIdentifierType, and relationType) to establish connections with related objects, such as datasets, articles, code, and others. Additionally, the relationType element contains dozens of options to indicate relationships, such as continuation, version, compilation, derivation, etc., which offer a machine-readable way to provide context and provenance since data is often not static. Example repositories that have incorporated DataCite relationship elements into their metadata schema include PANGAEA[7] in the earth and environmental sciences, the Inter-university Consortium for Political and Social Research (ICPSR)[8] in the social sciences, general purpose repositories Dataverse[9] and Zenodo, and the Illinois Data Bank[10] institutional repository. With elements implemented, connection between journal articles and data sets can be exposed as demonstrated by the SCHOlarly LInk eXchange (Scholix) framework, an output of a Research Data Alliance/World Data System working group (Cousijn et al., 2019). However, these elements are optional and complicated to implement cleanly (Stein and Dunham, 2018). As such, while some systems have been developed to take advantage of these elements, not all systems have done so.

The CUAHSI HydroShare repository (Tarboton et al., 2014; Horsburgh et al., 2015) provides many of these more advanced capabilities along with additional capabilities designed to incentivize its use. HydroShare casts hydrologic datasets and models as "social objects" that can be described with metadata, shared, collaborated around, annotated, discovered, accessed, and formally published. HydroShare's Resource Data Model (Horsburgh, 2015), which is an implementation of the Open Archives Initiative's Object Exchange and Reuse standard (OAI-ORE – Lagoze et al., 2008), recognizes that the data and models used by hydrologists are diverse in both file format and syntax and accounts for this by allowing users to assemble data and models within "resources" that may consist of individual files, groups of files, or even hierarchical file systems. All resources can be described using standard Dublin Core metadata elements along with custom, user-defined metadata elements (i.e., as key-value pairs) as well as through upload of a readme file that is rendered directly for potential data consumers to review directly on the resource's landing page. These mechanisms enable users to document the *who*, *why*, and *what* context of their data.

For known content types (e.g., hydrologic time series, multidimensional datasets stored using the Network Common Data Form (NetCDF), geographic feature datasets stored as shapefiles, geographic raster datasets stored as GeoTIFF, etc.), HydroShare provides more advanced metadata at the content level, some of which is automatically extracted from data files upon upload. Published resources receive a DOI and can be formally cited in linked publications, and HydroShare enables the

[7] https://www.pangaea.de/
[8] https://www.icpsr.umich.edu/icpsrweb/
[9] https://dataverse.org/
[10] https://databank.illinois.edu/

addition of "related resources" to the metadata for a resource, thus capturing the two-way linkage between a published dataset and any journal articles or other publications that use or describe it. HydroShare resources can be assembled into collections, and HydroShare's discovery interface enables geographic, temporal, keyword, or content type searches.

RECOMMENDATIONS FOR MORE EFFECTIVE DATA PUBLICATION

The authors' combined experience and perspective in providing data publication support services for scientific communities suggest that there are several recommendations related to the discussion in this overview that would improve data publication infrastructure, support, and practices. The recommendations below are based on our extensive experience over the past two decades, but also coalesced from the discussions and outcomes of a workshop that was held in May of 2019 among the authors and other experts from institutional repositories, journal publishers, organizations developing cyberinfrastructure for data management and publication, and providers of data support services for scientific communities.

Improve support for journal article-based data access and data citation. Journal publishers and some funders now require that data be published alongside a scientific journal article or deposited in an appropriate repository and then cited in the article. The journal article provides important context for the data that should enable more meaningful data re-use. This practice of linking datasets with the journal articles that describe them should be encouraged and promoted as a best practice. Furthermore, datasets should be specifically cited in the References section of the paper so that readers can follow those links just like they do to other referenced literature. Links to the published paper should be included in the metadata of the dataset(s) so that potential data users can traverse the opposite direction from dataset to paper. While these best practices are already encouraged, they are often not mandatory and are inconsistently applied, leaving it up to researchers' discretion as to whether or how they will comply. As data repositories and journal publishers evolve their data policies, there are opportunities to better enforce these best practices and to provide clear and consistent guidance to researcher on how to comply. Publishers, data repositories (and code repositories where data and code are shared in separate locations) could also facilitate this process through better coordination of the timed release and cross-referencing of peer-reviewed papers and associated datasets and code. An additional benefit of formalizing this practice will be that citations of datasets will become easier to track, enhancing ability to establish the impact of published datasets through formal citation metrics. Current work to develop data discovery interfaces should continue, but simple bibliographic retrieval (i.e., keyword, date, time, location) is not sufficient.

Consider the workflow for data publication. It may take several steps and iterations to arrive at the finished data products that scientists want to publish. Along the way, there is often a collaborative workflow that may include performing quality control on raw data, deriving aggregated or summarized products from original datasets, or advanced analyses of input datasets that result in

final data products. Additionally, there may be a period of time between when data products are finalized and when the authors are ready to formally publish them – e.g., an embargo period that provides data creators with an opportunity to finish their analyses and the journal paper describing them along with settling data authorship (e.g., Bierer et al., 2017) before the data are published. Providing functionality to support these workflow elements within data repositories may encourage data producers to deposit and curate their data earlier in the workflow, thus reducing the chance that the data are never published. The ability to first share data privately within a repository before formal publication – e.g., as implemented by the HydroShare repository – gives authors the flexibility to choose an appropriate embargo period while still enabling collaborative access to the data by a project team. This also allows the peer reviewers of the paper to confidentially access the data as they evaluate the research before the data is published.

Enhance support for reproducible science. Ensuring reproducibility and verifiability of scientific results is an important reason for publishing data and should be recognized as an essential form of data reuse. Repositories must acknowledge that reproducing results described in scientific articles may require more than just making data files available. Researchers have incentive to publish data when it is a condition of publishing their paper, and including review of submitted data and the reproducibility of the work as part of the peer-review process may help increase quality (Rosenberg et al., 2019). Inclusion of scripts and/or executable workflows (e.g., Jupyter Notebooks) along with instructions on how to use them along with data is an important step toward ensuring that potential data consumers can retrace the steps of the individual investigators and build upon their results. This may require repositories to develop additional functionality (e.g., the HydroShare repository has created a linked JupyterHub environment for online execution of Jupyter Notebooks contained within HydroShare resources), services (e.g., curation), and a commitment from data producers to invest the time and effort to create and share these assets along with the data. The work of Nüst et al. (2017) in developing the concept of an "executable research compendium" as a self-contained collection of data, code, and execution environment is relevant here as is the concept of a "Sciunit" developed by That et al. (2017) as a reusable research object that uses application virtualization to create a container of an executable application that could be integrated with a repository like HydroShare as demonstrated by Essawy et al. (2018) to enhance reproducibility.

Encourage publication of curated reference data collections rather than data publications. Some of the most widely (re)used and cited datasets are large-scale, long-term, curated data sets. For example, in the machine learning field, datasets like the ImageNet database (Deng et al., 2009), which is a large collection of labeled images designed for use with visual object recognition software research, have been used extensively and have driven many of the important developments in the field. Similarly, in the Earth sciences, spatially extensive climatological and hydrological data sets are widely used. Examples include the Model Parameter Estimation Experiment (MOPEX) dataset (Schaake et al., 2006) and the PRISM climate dataset (Daly et al., 1994; PRISM Climate Group, 2016). New reference data sets are being created and can now be more easily published. While some are

the result of large multi-agency/multi-institution endeavors that transcend the capabilites and resources of individuals, others result from the work of individuals (e.g., the MOPEX dataset) or smaller collaborative groups (e.g., within HydroShare, several collections of data associated with major Hurricane events, including Harvey (Arctur et al., 2018), Irma (Arctur, 2018), and Maria (Bandaragoda et al., 2019) have recently been published). Reference data sets are highly curated, have a specified context and use, and have been certified as sufficient for one or more specific purposes by experts. Like curated datasets have in other scientific fields, these types of curated collections may be essential for advancing the field of hydrology.

Advance interoperability standards for sharing data and metadata among repositories. Due to the many stakeholders and potential ways in which data could be packaged, shared, and reused, we need to move away from practice that requires potential data users to know which repository data resides in before they can determine whether it exists. Data users should be able to discover data, regardless of where they are hosted. Additionally, repositories should be able to catalog and/or reference data holdings initially deposited elsewhere, regardless of location. For example, domain-specific repositories may be interested in being able to represent data initially deposited in a general purpose or university-based repository to strengthen coverage within its own domain. Alternatively, to demonstrate their value the public, universities or federal agencies may be interested in cataloging data deposited in domain-specific or general-purpose repositories to strengthen their ability to track data related to their organization. Each of these use cases requires strong interoperability, socially and technically, among all repository types. The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH – Lagoze et al., 2015) is an early example of a protocol built for harvesting metadata descriptions from many archives. The DataONE project's model of creating a system to catalog metadata from more than 40 participating member nodes, each of which is a separate data repository, is another example of interoperability in action (Michener et al., 2011; https://www.dataone.org/). More recently, Google's Dataset Search (Noy, 2018) enables discovery of data wherever they are hosted through the use of an open standard schema and vocabulary for describing data (http://schema.org). Much can be learned from these hard-won experiences.

Develop partnerships with university libraries offering data services. In addition to maintaining their traditional role as provisioner and manager of scholarly collections, many university libraries have begun establishing data services. This evolution is not uncontested, as researchers, administrators, and traditional Library and Information Science (LIS) professionals debate what role the library can, or should, play. Regardless, libraries are often a common touch point for many researchers across all disciplines, and the LIS profession does have expertise in the standardized collection and dissemination of scholarly outputs. Trained data librarians help researchers develop data management plans prior to grant proposal submission and then again with data publication strategies, including dataset arrangement, creation of accompanying documentation, and selection of appropriate repositories for ongoing projects. Indeed, data librarians are well positioned to offer

advice to scientists related to technical best practices that promote the sharing of high-quality datasets (e.g., well-structured and accompanied by descriptive metadata). Additionally, several new library partnerships show promising potential. In one, Dryad has partnered with the California Digital Library to "…make it easier to integrate data publishing into researcher workflows and be focused on building a sustainable product that is a credible alternative to commercial offerings within the research data space" (Simms, 2018). In another example, the 10 institutions that established the Data Curation Network are committed to developing standardized curation practices and an accompanying workforce to improve quality and reuse potential of published datasets in a cost-effective and community-owned way (Johnston et al., 2018).

Develop more specific data management plans. Given the low success rates and time pressures on submitting grant proposals, data management plans are rarely considered with the same level of detail as other proposal materials and often use boilerplate language aimed at meeting funding agency requirements. Furthermore, they are often ignored when the grant is received. It may be more effective to require a basic data management plan at the time of grant proposal submission, followed by a more detailed plan after the funding decision is made, but prior to releasing the funds. The basic plan should be complete enough to provide the proposing institution with an understanding of what they are responsible for should the proposal be successful and should also prompt the proposal team to include the cost of data management and preservation in the proposed budget. A more detailed plan would include project-specific provisions for data management and might be subject to periodic review by program officers as part of regular grant reporting requirements. Awardees should be encouraged to consult their university libraries and, where available, domain-specific repositories on best practices so that a meaningful data management plan can be developed and executed.

Conclusion

More effective data publication requires sufficient context for enabling data re-use that typically goes beyond even an extensive metadata profile. Repositories must be structured to capture this context, by linking different kinds of data, scripts/code, and workflows. An initial objective of ensuring reproducibility of scientific analyses provides some guidance for the design of repositories. However, achieving the larger goal of data synthesis will require substantial effort on the part of scientists to document and organize their data. Hence, data must be considered and treated as a first-class research product to justify that investment of time. The common practice of sharing data in the supplemental materials associated with a journal paper accomplishes the goal of making data available, but does little to ensure that data are well organized, use formats familiar to scientists who might access the data, and are described with metadata that would help others interpret the data. Furthermore, supplemental materials have little context beyond the paper with which they are associated and may be hidden behind the same paywall that applies to the paper. The end result is that data are not widely discoverable and are unlikely to be reused.

In contrast, publishing data in a repository that supports FAIR Data Principles encourages a much higher level of curation, ensures data can be cited using persistent identifiers, and enables discovery either by reference from the citing paper or independently through repository or more general search functionality. The ability to properly cite data also makes it much easier to track and report the impact of research data using methods similar to those used to track the impact of research publications. Repositories are making steady progress here, but there are still challenges to be overcome. Packaging datasets into citable entities is useful for discovery and reuse of fixed and stable content that has already been used by someone for a particular purpose. However, it does not address all the rest of the data we collect every day, but that are not (yet) included in or described by a research paper, which may be most of the data we have. We know that more work is needed to build the social and cyberinfrastructure for bringing more of the data we collect into the scholarly record, and the recommendations provided above lay out potential next steps. While we are confident that these recommendations can improve data publication practices, additional work is also needed to quantify their impact.

Funding Information

This work was supported by the National Science Foundation under grant EAR-1838572. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Acknowledgments

Much of the content of this overview resulted from discussions held during a May 2019 workshop in Washington D.C. The authors would like to acknowledge the American Geophysical Union for hosting this meeting and the input of workshop participants who contributed to the ideas captured in this overview.

References

Aarts, A. A. et al. (2015). Estimating the reproducibility of psychological science, *Science*, 349(6251), https://doi.org/ 10.1126/science.aac4716.

Arctur, D. (2018). Hurricane Irma 2017 Collection, *HydroShare*, https://doi.org/10.4211/hs.f0740db7831d4096967faa84abdd95cf.

Arctur, D., Boghici, E., Tarboton, D., Maidment, D., Bales, J., Idaszak, R., Seul, M., Castronova, A. M. (2018). Hurricane Harvey 2017 Collection, *HydroShare*, https://doi.org/10.4211/hs.2836494ee75e43a9bfb647b37260e461.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility, *Nature*, 533, 452-454, https://doi.org/10.1038/533452a.

Bandaragoda, C., Phuong, J., Leon, M. (2019). Hurricane Maria 2017 Collection, *HydroShare*, http://www.hydroshare.org/resource/97a696e7202d4ca98349a0742a725451.

Baron, J. S. et al. (2017). Synthesis centers as critical research infrastructure, *BioScience*, 67: 750–759. https://doi.org/10.1093/biosci/bix053.

Bierer, B. E., Crosas, M., Pierce, H. H. (2017). Data authorship as an incentive to data sharing, The New England Journal of Medicine, 376, 1684-1687, https://doi.org/10.1056/NEJMsb1616595.

Clarivate Analytics (2019). Web of Science Platform: Data Citation Index, https://clarivate.libguides.com/webofscienceplatform/dci, accessed October 19, 2019.

Cousijn, H., Feeney, P., Lowenberg, D., Presani, E., & Simons, N. (2019). Bringing citations and usage metrics together to make data count, *Data Science Journal*, 18(1), 9, http://doi.org/10.5334/dsj-2019-009.

Daly, C., Neilson, R. P., Phillips, D. L. (1994). A statistical-topographic model for mapping climatological precipitation over mountainous terrain, *Journal of Applied Meteorology*, 33, 140-158, https://doi.org/10.1175/1520-0450(1994)033<0140:ASTMFM>2.0.CO;2.

Das, S., et al. (2017). Cyberinfrastructure for open science at the Montreal Neurological Institute, *Frontiers in Neuroinformatics*, 10:53, https://doi.org/10.3389/fninf.2016.00053.

DataCite Metadata Working Group (2019). DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. Version 4.3. DataCite e.V. https://doi.org/10.14454/7xq3-zf69.

DCMI (Dublin Core Metadata Initiative) (2012), DCMI Metadata Terms. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/, accessed October 13, 2019.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database, 2009 IEEE Conference on Computer Vision and Pattern Recognition, 20 – 25 June, Miami, FL, http://doi.org/10.1109/CVPR.2009.5206848.

Essawy, B. T., Goodall, J. L., Zell, W., Voce, D., Morsy, M. M., Sadler, J., Yuan, Z., Malik, T. (2018). Integrating scientific cyberinfrastructures to improve reproducibility in computational hydrology: Example for HydroShare and GeoTrust, *Environmental Modelling & Software*, 105, 217-229, https://doi.org/10.1016/j.envsoft.2018.03.025.

Fegraus, E. H., Andelman, S., Jones, M. B., Schildhauer, M. (2005). Maximizing the value of ecological data with structured metadata: An introduction of ecological metadata language (EML) and principles for metadata creation, *Bulletin of Ecological Society of America,* 86(3), 158-168, https://doi.org/10.1890/0012-9623(2005)86[158:MTVOED]2.0.CO;2.

Flint, C. G., Jones, A. S., Horsburgh, J. S. (2017). Data management dimensions of social water science: The iUTAH experience, *Journal of the American Water Resources Association (JAWRA)*, 1-9, https://doi.org/10.1111/1752-1688.12568.

Gibert, K., Horsburgh, J. S., Athanasiadis, I. A., Holmes G. (2018). Environmental Data Science, *Environmental Modelling & Software*, 106, 4-12, https://doi.org/10.1016/j.envsoft.2018.04.005.

Goodman, A., Pepe, A., Blocker, A. W., Borgman, C. L., Cranmer, K., Crosas, M., Di Stefano, R., Gil, Y., Groth, P., Hedstrom, M., Hogg, D. W., Kashyap, V., Mahabal, A., Siemiginowska, A., Slavkovic, A. (2014). Ten simple rules for the care and feeding of scientific data, PLoS Computational Biology, 10(4):e1003542, https://doi.org/10.1371/journal.pcbi.1003542.

Hazen, R. M., et al. (2019). Data-driven discovery in mineralogy: Recent advances in data resources, analysis, and visualization, *Engineering*, 5(3), 397-405, https://doi.org/10.1016/j.eng.2019.03.006.

Hampton, S. E., Parker, J. N. (2011). Collaboration and productivity in scientific synthesis, *BioScience*, 61(11), 900-910, https://doi.org/10.1525/bio.2011.61.11.9.

Horsburgh, J. S., Morsy, M. M., Castronova, A., Goodall, J. L., Gan, T., Yi, H., Stealey, M. J., Tarboton, D. G. (2015). HydroShare: Sharing diverse hydrologic data types and models as social objects within a Hydrologic Information System, *Journal of the American Water Resources Association (JAWRA)*, 52(4), 873-889, https://doi.org/10.1111/1752-1688.12363.

Horsburgh, J. S., Tarboton, D. G., Piasecki, M., Maidment, D. R., Zaslavsky, I., Valentine, D., Whitenack, T. (2009). An integrated system for publishing environmental observations data, *Environmental Modeling and Software*, 24, 879-888, https://doi.org/10.1016/j.envsoft.2009.01.002.

Jackson, J. B. C. et al. (2001). Historical overfishing and the recent collapse of coastal ecosystems, *Science*, 293:5530, 629-637, https://doi.org/10.1126/science.1059199.

Johnston, L. R., Carlson, J., Hudson-Vitale, C., Imker, H., Kozlowski, W., Olendorf, R., Stewart, C., Blake, M., Herndon, J., McGeary, T. M., Hull, E. (2018). Data Curation Network: A cross-institutional staffing model for curating research data, *International Journal of Digital Curation*, 13, 125–140. https://doi.org/10.2218/ijdc.v13i1.616.

Kadlec, J., StClair, B., Ames, D. P., Rill, R. A. (2015). WaterML R package for managing ecological experiment data on a CUAHSI HydroServer, *Ecological Informatics*, 28, 19-28, https://doi.org/10.1016/j.ecoinf.2015.05.002.

Kiar, G., Gorgolewski, K. J., Kleissas, D., Roncal, W. G., Litt, B., Wandell, B., Poldrack, R. A., Wiener, M., Vogelstein, R. J., Burns, R., Vogelstein, J. T. (2017). Science in the cloud (SIC): A use case in MRI connectomics, *Giga Science*, 6, 1-10, https://doi.org/10.1093/gigascience/gix013.

Knox, S. H. et al. (2019). FLUXNET-CH4 synthesis activity: Objectives, observations, and future directions, *Bulletin of the American Meteorological Society*, https://doi.org/10.1175/BAMS-D-18-0268.1.

Kratz, J., Strasser, C. (2014). Data publication consensus and controversies, F1000Research, 3, 94, https://doi.org/10.12688/f1000research.3979.3.

Lagoze, C., Van de Sompel, H., Johnston, P., Nelson, M., Sanderson, R., Warner, S. (2008). Open Archives Initiative Object Reuse and Exchange: ORE Specification – Abstract Data Model, available at http://www.openarchives.org/ore/1.0/datamodel.html, accessed October 19, 2019.

Lagoze, C., Van de Sompel, H., Nelson, M., Warner, S. (2015). The Open Archives Initiative Protocol for Metadata Harvesting, available at http://www.openarchives.org/OAI/openarchivesprotocol.html, accessed October 19, 2019.

Lannom, L., Koureas, D., Hardisty, A. R. (2020). FAIR data and services in biodiversity science and geoscience, *Data Intelligence*, 2(1-2), 122-130, https://doi.org/10.1162/dint_a_00034.

Mayo, C., Hull, E. A., Vision, T. J. (2015). The location of the citation: Changing practices in how publications cite original data in the Dryad Digital Repository, *Zenodo*, https://doi.org/10.5281/zenodo.32412.

Michener, W., Vieglais, D., Vision, T., Kunze, J., Cruse, P., Janée, G. (2011). DataONE: Data Observation Network for Earth – preserving data and enabling innovation in the biological and environmental sciences, *D-Lib Magazine*, 17(1/2), https://doi.org/10.1045/january2011-michener.

Noy, N. (2018). Making it easier to discover datasets, *The Keyword*, https://www.blog.google/products/search/making-it-easier-discover-datasets/, accessed November 9, 2019.

Nüst, D., Konkol, M., Schutzeichel, M., Pebesma, E., Kray, C., Przibytzin, H., Lorenz, J. (2017). Opening the publication process with executable research compendia, *D-Lib Magazine*, 23(1-2), https://doi.org/10.1045/january2017-nuest.

Parsons, M. A., Fox, P. A. (2013). Is data publication the right metaphor?, *Data Science Journal*, 12, 32-46, https://doi.org/10.2481/dsj.WDS-042.

Pasquetto, I. V. et al. (2017). On the Reuse of Scientific Data, *Data Science Journal*, 16: 8, 1–9, https://doi.org/10.5334/dsj- 2017-008.

PRISM Climate Group (2016). Descriptions of PRISM Spatial Climate Datasets for the Conterminous United States, http://www.prism.oregonstate.edu/documents/PRISM_datasets.pdf, accessed October 19, 2019.

Poldrack, R. A., Gorgolewski, K. J., Varoquaux, G. (2019). Computational and informatic advances for reproducible data analysis in neuroimaging, *Annual Review of Biomedical Data Science*, 2, 119-138, https://doi.org/10.1146/annurev-biodatasci-072018-021237.

re3data.org (2020). Registry of Research Data Repositories, https://doi.org/10.17616/R3D, accessed January 27, 2020.

Rosenberg, D. E., Fillion, Y., Teasley, R. L., Sandoval-Solis, S., Hecht, J. S., van Zyl, J. E., McMahon, G. F., Horsburgh, J. S., Kasprzyk, J. R., Tarboton, D. G. (2019). The next frontier: Making research more reproducible, *Journal of Water Resources Planning and Management*, 1-10, https://doi.org/10.1061/(ASCE)WR.1943-5452.0001215

Schaake, J., Cong, S., Duan, Q. (2006). The US MOPEX Data Set, in Large Sample Basin Experiments for Hydrological Model Parameterization: Results of the Model Parameter Experiment–MOPEX. IAHS Publication 307, https://iahs.info/uploads/dms/13600.04-9-28-SCHAAKE.pdf.

Shen, C. (2018). Deep learning: A next-generation big-data approach for hydrology, *EOS*, 99, https://doi.org/10.1029/2018EO095649.

Simms, S. (2018). Letter to the Community: CDL and Dryad Partnership, https://cdlib.org/cdlinfo/2018/05/30/letter-to-the-community-cdl-and-dryad-partnership/, accessed October 13, 2019.

Stagge, J. H., Rosenberg, D. E., Abdallah, A. M., Akbar, H., Attallah, N. A., James, R. (2019). Accessing data availability and research reproducibility in hydrology and water resources, Scientific Data, 6:190030, https://doi.org/10.1038/sdata.2019.30.

Stall, S., Cruse, P., Cousijn, H., Cutcher-Gershenfeld, J., deWaard, A., Hanson, B., Heber, J., Lehnert, K., Parsons, M., Robinson, E., Witt, M., Wyborn, L., Yarmey, L. (2018). Data sharing and citations: New author guidelines promoting open and FAIR data in the Earth, space, and environmental sciences, *Science Editor*, 41(3), 83-87, https://www.csescienceeditor.org/article/data-sharing-and-citations-new-author-guidelines-promoting-open-and-fair-data-in-the-earth-space-and-environmental-sciences/.

Stall, S., McEwen, L., Wyborn, L., Hoebelheinrich, N., Bruno, I. (2020). Growing the FAIR community at the intersection of the geosciences and pure and applied chemistry, *Data Intelligence*, 2(1-2), 139-150, https://doi.org/10.1162/dint_a_00036.

Stein, A., Dunham, E. (2018). Meaningful data sharing: Developing the Illinois Data Bank Metadata Framework, *Journal of Library Metadata*, 18:2, 59-83, https://doi.org/10.1080/19386389.2018.1488561.

Stodden, V., Seiler, J., Ma, Z. (2018). An empirical analysis of journal policy effectiveness for computational reproducibility, Proceedings of the National Academy of Sciences (PNAS), 115(11), 2584-2589, https://doi.org/10.1073/pnas.1708290115.

Tarboton, D. G., Horsburgh, J. S., Maidment, D. R., Whiteaker, T., Zaslavsky, I., Piasecki, M., Goodall, J., Valentine, D., Whitenack, T. (2009). Development of a community Hydrologic Information System, in Anderssen, R. S., R. D. Braddock, and L.T.H. Newham (eds) 18th World IMACS Congress and MODSIM09 International Congress on Modelling and Simulation, Modelling and Simulation Society of Australia and New Zealand and International Association for Mathematics and Computers in Simulation, July 2009, pp. 988-994, ISBN: 978-0-9758400-7-8.

Tarboton, D. G., Idaszak, R., Horsburgh, J. S., Heard, J., Ames, D., Goodall, J. L., Band, L., Merwade, V., Couch, A., Arrigo, J., Hooper, R., Valentine, D., Maidment, D. (2014). HydroShare: Advancing collaboration through hydrologic data and model sharing, in D. P. Ames, N. W. T. Quinn and A. E. Rizzoli (eds.), Proceedings of the 7th International Congress on Environmental Modelling and Software, San Diego, California, USA, International Environmental Modelling and Software Society (iEMSs), ISBN: 978-88-9035-744-2, https://scholarsarchive.byu.edu/iemssconference/2014/Stream-A/7/.

That, D. H. T, Fils, G., Yuan, Z., Malik, T. (2017). Sciunits: Reusable research objects, 2017 IEEE 13th International Conference on e-Science (e-Science), 24-27 October, https://doi.org/10.1109/eScience.2017.51.

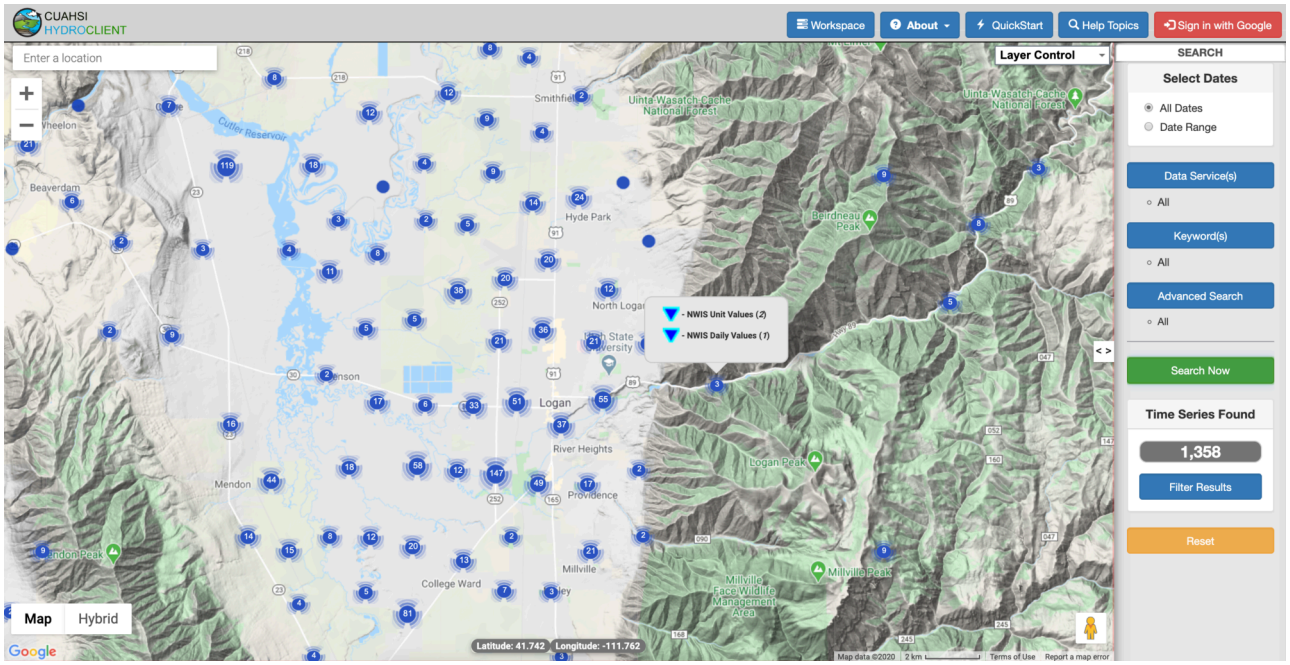
Thomas, C. D. et al. (2004). Extinction risk from climate change, *Nature*, 427, 145-148, https://doi.org/10.1038/nature02121.

Wilby, R. L., Clifford, N. J., De Luca, P., Harrigan, S. O., Hillier, J. K., Hodgkins, R., Johnson, M. F., Matthews, T. K. R., Murphy, C., Noone, S. J., Parry, S., Prudhomme, C., Rice, S. P., Slater, L. J., Smith, K. A., Wood, P. J. (2017) The "dirty dozen" of freshwater science: Detecting then reconciling hydrological data biases and errors, *WIREs Water*, 4, e1209. https://doi.org/10.1002/wat2.1209.

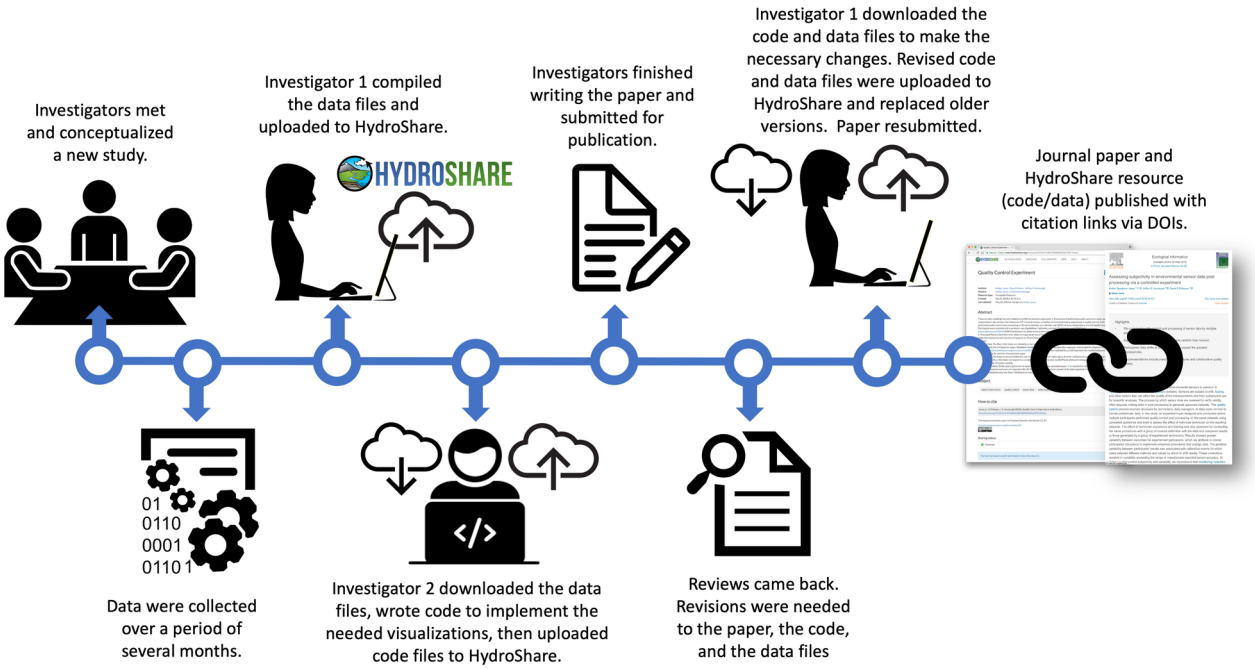
Williams, D. N., Ananthakrishnan, R., Bernholdt, D. E., Bharathi, S., Brown, D., Chen, M., Chervenak, A. L., Cinquini, L., Drach, R., Foster, I. T., Fox, P., Fraser, D., Garcia, J., Hankin, S., Jones, P., Middleton, D. E., Schwidder, J., Schweitzer, R., Schuler, R., Shoshani, A., Siebenlist, F., Sim, A., Strand, W. G., Su, M., Wilhelmi, N. (2009). The Earth System Grid: Enabling access to multimodel climate simulation data, *Bulletin of the American Meteorological Society*, https://doi.org/10.1175/2008BAMS2459.1.

Wilkinson, M. D. et al. (2016). The FAIR guiding principles for scientific data management and stewardship, Scientific Data, 3:160018, https://doi.org/10.1038/sdata.2016.18.

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., Pande, V. (2018). MoleculeNet: A benchmark for molecular machine learning, *Chemical Science*, 9, 513-530, https://doi.org/10.1039/C7SC02664A.



WAT2_1422_Figure1.tif



WAT2_1422_GraphicalAbstract.tif