PNNL-33552

Researcher Views on Water Data Accessibility, Usability, and Dissemination

October 2022

Kyle B Larson
James W Saulsbury
Evan R Margiotta

U.S. DEPARTMENT OF ENERGY

Prepared for the U.S. Department of Energy
under Contract DE-AC05-76RL01830

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor Battelle Memorial Institute, nor any of their employees, makes **any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights**. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof, or Battelle Memorial Institute. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

PACIFIC NORTHWEST NATIONAL LABORATORY
operated by
BATTELLE
for the
UNITED STATES DEPARTMENT OF ENERGY
under Contract DE-AC05-76RL01830

Printed in the United States of America

Available to DOE and DOE contractors from the
Office of Scientific and Technical Information,
P.O. Box 62, Oak Ridge, TN 37831-0062;
ph: (865) 576-8401
fax: (865) 576-5728
email: reports@adonis.osti.gov

Available to the public from the National Technical Information Service
5301 Shawnee Rd., Alexandria, VA 22312
ph: (800) 553-NTIS (6847)
email: orders@ntis.gov <https://www.ntis.gov/about>
Online ordering: http://www.ntis.gov

PNNL-33552

Researcher Views on Water Data Accessibility, Usability, and Dissemination

October 2022

Kyle B Larson
James W Saulsbury
Evan R Margiotta

Prepared for
the U.S. Department of Energy
under Contract DE-AC05-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99354

Summary

In October 2019, the U.S. Department of Energy's (DOE's) Water Power Technologies Office (WPTO) initiated a project to investigate pathways to improving the discovery, sharing, and use of water data. Through extensive stakeholder engagement in the first two phases of the project, we learned that opinions about these aspects vary widely depending on technical background, career, affiliation, and other factors. This report focuses on the third and final phase of the project, which sought to understand the challenges that researchers at Pacific Northwest National Laboratory (PNNL) face with finding, accessing, using, and disseminating water data and how those challenges affect their ability to do mission critical research for WPTO and other sponsors. We also take a closer look at approaches to disseminating data to help inform WPTO and lab researcher decisions about incorporating data dissemination into future projects.

We conducted structured interviews with 19 PNNL researchers, plus one collaborator at Idaho National Laboratory (INL), who manage or lead projects ranging across the spectrum of PNNL's WPTO hydropower portfolio. Many of the researchers also support other water-related programs for other DOE offices and federal sponsors. Interviews were one-hour long and followed a guiding set of questions inspired by the FAIR data principles: *Findability* – the ability for both humans and computers to search for, identify, and locate data; *Accessibility* – the ability of a user to retrieve the data once it is found; *Interoperability* – the need to be integrated with other data as well as the need for data to interoperate with applications or workflows for analysis, storage, and processing; and *Reusability* – the ability of a user to make use of the data once it is accessed. When applicable, we also asked researchers how they disseminate data that is produced through their work.

Collectively, the researchers we interviewed, and the authors of this report, have used all 14 major categories of water data that were identified and defined in the first phase of this project (i.e., Dams, Ecology, Flood Control, Hydroclimatology, Hydrography, Hydrology, Hydropower, Management Landscape, Meteorological, Migratory Barriers, Recreation & Aesthetic Importance, Socioeconomic, Water Quality Water Availability & Use). The most cited categories of water data (in order of most to less common) were Hydrology, Hydropower, Dams, and Water Quality. Nearly all of researchers said they depend almost entirely on data produced externally by other organizations, indicating that the FAIR-ness of external data sources is a critical factor in PNNL's ability to support WPTO and other water-related programs. Some researchers also noted that it is not easy to find potentially relevant data produced within PNNL or other national labs.

The findability and accessibility of data used by researchers we interviewed varies depending on the category and source of data. For example, much of the Hydrology, Hydroclimatology, and Water Quality data is easily findable and accessible on the internet, whereas key Dams, Hydropower, and Socioeconomic datasets are difficult to find and access because they are not publicly available and often require strict use agreements to obtain. Most researchers who work with Dams and Hydropower data said they find the data through professional connections and institutional knowledge of the project team and must obtain the data directly from dam owners/operators, Independent System Operators (ISOs), Balancing Authorities, or industry groups. Often, researchers must aggregate data from multiple sources to obtain the quantity and quality of data they need for their research. Consequently, the amount of time or resources that researchers said is required to find and access data ranged from 15-90% of the project duration or budget.

The interoperability and reusability of water data used by researchers we interviewed also varied depending on the category and source of data. Most researchers indicated that the bulk of data they work with is not “analysis ready” and usually requires moderate to substantial manipulation to be made suitable for their use. The quality of metadata (secondary data describing key information about the primary data) also varies by the type and source of data, which can make preparation and use of the data difficult. In general, researchers indicated the interoperability and reusability of water data that are publicly available and dispensed through well-established platforms (e.g., Hydrology, Hydrography, Hydroclimatology, Water Quality) is suitable to quite good, whereas data that are not publicly available or formally dispensed (e.g., Dams, Hydropower, Socioeconomic) is moderate to poor.

Only a few of the researchers we interviewed have incorporated data dissemination as a task or deliverable in their projects, either due to a desire or requirement by the sponsor, or the researcher’s own professional or personal motivating factors to share data. Researchers who use proprietary Dams or Hydropower data often cited that associated use restrictions forbid dissemination of the data, except in certain cases where it can be abstracted to protect proprietary aspects. This, however, reduces the usefulness of the data.

Approaches for disseminating data varied among researchers but can be grouped into two general categories: *available upon request* and *hosted* services. The available upon request approach is reactive and informal in nature and generally involves the researcher transmitting data directly to the requester via email, File Transfer Protocol (FTP), DropBox, Google Drive, etc. The hosted services approach is proactive and structured and involves exposing the data via standard internet protocols either behind (internal) or outside (external) the organization’s firewall. When comparing these approaches based on the FAIR principles, the hosted services approach is superior with respect to improving the findability, accessibility, interoperability, and reusability of data. It is also more appropriate for disseminating “larger” datasets, data with higher demand or applicability, or data that is regularly updated or incrementally increases in volume over time. The hosted services approach generally requires more time and financial investment on the part of the data provider compared to available upon request services, especially if the data provider chooses to build a custom hosted service. However, there is an increasing number of free or low-cost open-access data repositories, as well as discipline-specific data repositories, that have improved the feasibility and cost of using hosted services.

Acknowledgments

This project is funded by the U.S. DOE WPTO. The authors are grateful for the support and guidance of WPTO’s Hoyt Battey and Corey Vezina. The authors would also like to thank the following researchers from PNNL and INL (in alphabetical order by last name) whose candid input during interviews was greatly appreciated and informative: PNNL – Osman Ahmed, Casey Burleyson, Vishvas Chalishazar, Sohom Datta, Travis Douville, Juliet Homer, James Kershaw, Slaven Kincic, Kendall Mongird, Kostas Oikonomou, Feng Pan, Nader Samaan, Abishek Somani, Sean Turner, Nathalie Voisin, Todd Wall, Mark Weimar, Mark Wigmosta; INL – Shiloh Elliot.

Acronyms and Abbreviations

CUAHSI	Consortium of Universities for Advanced of Hydrologic Science
DOE	Department of Energy
DOI	Digital Object Identifier
EIA	Energy Information Administration
EPA	Environmental Protection Agency
FAIR	Findable Accessible Interoperable Reusable
FERC	Federal Energy Regulatory Commission
FTP	File Transfer Protocol
INL	Idaho National Laboratory
ISO	Independent System Operator
NDA	Non-Disclosure Agreement
NGO	Non-Government Organization
OASIS	Open-Access Same-Time Information System
OSTI	Office of Scientific and Technical Information
PM/PI	Project Manager/Principal Investigator
PNNL	Pacific Northwest National Laboratory
PSH	Pumped Storage Hydropower
USGS	US Geological Service
WPTO	Water Power Technologies Office

Contents

Tables

1.0 Introduction

In October 2019, the U.S. Department of Energy's (DOE's) Water Power Technologies Office (WPTO) initiated a project to investigate potential pathways to improving the discovery, sharing, and use of water data. The project consists of three phases. The first phase sought to characterize relevant categories of water data; describe the current state of accessing, using, and visualizing water data; and outline potential pathways and collaborations for future work (Larson et al. 2021). The second phase of the project focused on developing interactive, web-based data stories that highlight water data challenges and needs in several US river basins and holding a workshop on water data issues (Reicher et al. 2021). Both phases relied heavily on engagement with stakeholders from various federal, state, tribal, hydropower industry, and Non-Government Organizations (NGOs) across the US.

In the first two phases of the project, it became apparent that opinions about water data accessibility, usability, and sharing vary widely depending on technical background, career, affiliation, and other factors (Larson et al. 2021). In addition, it was clear that accessibility and usability of water data varies depending on the type of data. Based on these findings, we sought to understand how researchers that support WPTO projects, as well as water-related programs for other sponsors, view these matters and how it affects their research.

This became the focus of the third and final phase of the project, which we describe in this report. Ultimately, the challenges that researchers face with finding, accessing, using, and disseminating water data affect their ability to do mission critical research for WPTO and other sponsors. Hence, a goal of this paper is to relay those challenges to WPTO so it can continue its broader efforts to improve access and use of water data, especially those data that are more critical to its mission and the research performed on its behalf by national labs. In addition, we sought a basic understanding of how often data dissemination is factored into projects, as there is increased desire from WPTO and other sponsors to do so when applicable.

The paper is organized as follows: Section 2.0 describes how we structured and conducted interviews with researchers; Section 3.0 summarizes key findings from the interviews; and Section 4.0 summarizes general approaches for disseminating research data that we identified through the interviews.

2.0 Interview Structure

While there are many researchers in the DOE national laboratory complex that support WPTO projects and water-related programs for other sponsors that could be valuable sources of information for this effort, we limited our scope to interviewing researchers at Pacific Northwest National Laboratory (PNNL) because we could leverage internal resources and connections to identify them more readily compared to other national labs. We identified a diverse set of interview candidates by contacting project managers and principal investigators (PM/PIs) from across the spectrum of PNNL's WPTO hydropower portfolio and inviting them and their selected colleagues to participate. Ultimately, we interviewed 19 PNNL researchers, plus one collaborator at Idaho National Laboratory (INL), who support energy-water research funded by WPTO, and in many cases, other water-related programs for other sponsors.

We conducted structured interviews and limited them to one hour each. We sought to encourage candid responses by keeping interviewees' answers anonymous. Each interview began with getting a basic understanding of the researchers' projects and the types of water-related data they utilize. For the latter, we provided interviewees with a table of 14 broad categories of water data described by Larson et al. (2021) (Table 1) and asked them to identify those that apply to their research and provide specific examples.

Table 1. Fourteen basic categories of water-related data described by Larson et al. (2021).

Category	Description
Dams	Spatial and metadata information about dams, including dam uses, ownership, licensing, age, operations, risk/safety status, etc.
Ecology	Spatial and/or quantitative data about the presence of sensitive aquatic and terrestrial species, habitats, land cover, land use, conservation activities, etc.
Flood Control	Spatial and quantitative data about historical flood plains, flood control infrastructure, flood risk, etc.
Hydroclimatology	Spatial and/or quantitative data depicting current or future regional climate (temperature and precipitation) and climate-informed hydrology
Hydrography	Spatial and quantitative data about the physical characteristics of basins, sub-basins, watersheds, streams, stream order, canals, pipelines, etc.
Hydrology	Spatial and/or quantitative data about the quantity and timing of water in a basin, including stream gage data, modeled streamflow, reservoir storage, hydrologic alteration, aquifers, groundwater recharge, etc.
Hydropower	Spatial, quantitative, and metadata information about hydroelectric dams, generation, ownership, licensing, and associated infrastructure
Management Landscape	Spatial and metadata information about land ownership, protection status, zoning, etc.
Meteorological	Spatial and/or quantitative weather observation data
Migratory Barriers	Spatial and metadata information about anthropogenic barriers to fish migration and aquatic habitat use
Recreation & Aesthetic Importance	Spatial and metadata information about areas reserved for recreation or aesthetic purposes
Socioeconomic	Spatial and/or quantitative data related to the socioeconomic value of water resources, cultural resources, land, water uses, hydropower, fisheries, and other factors

Table 2 (continued). Fourteen basic categories of water-related data described by Larson et al. (2021).

Category	Description
Water Quality	Spatial and/or quantitative data about surface-water and groundwater quality (e.g., dissolved oxygen, nitrates, contaminants of concern)
Water Availability & Use	Spatial and quantitative data about surface-water and groundwater use/availability for public supply, domestic, irrigation, thermoelectric power, industrial, mining, livestock, and aquaculture, as well as water rights

We followed a guiding set of questions inspired by the FAIR (Findable Accessible Interoperable Reusable) data principles (Table 2), except when we sought to dive more deeply into certain responses. The FAIR principles are a set of guidelines put forth by the scientific data management community in 2016 to improve the *Findability*, *Accessibility*, *Interoperability*, and *Reuse* of digital assets (Wilkinson et al. 2016). *Findability* relates to the ability for both humans and computers to search for, identify, and locate data, whereas *Accessibility* relates to the ability of a user to retrieve the data once it is found. *Interoperability* pertains to the need to be integrated with other data as well as the need for data to interoperate with applications or workflows for analysis, storage, and processing. *Reusability* relates to the ability of a user to make use of the data once it is accessed. For simplicity, we collectively refer to *interoperability* and *reusability* as “usability”.

It should be noted, however, that our intent was not to assess the "FAIR-ness" of data used by the researchers as this was beyond our scope. Rather, our intent was to assess researcher opinions about these principles as they pertain to data sources used in their research, as well as how much time and effort is generally required to find, access, and make data usable. When applicable, we also asked researchers how they disseminate data that is produced through their work.

Table 2. Guiding questions (organized by general topic) used during researcher interviews about water data.

Research Context
• What type(s) of water-related research do you do for WPTO or other sponsors? • What types of water-related data do you utilize in your research? • What types of research activities do you use data for (e.g., modeling, statistical analysis, qualitative analysis, visualization)? • Are your project(s) dependent on external (outside of PNNL) data sources?
Findability and Accessibility
• How do you find data? • Are data difficult to find? • Do you have to aggregate the data from multiple sources? Is it clear who the contributors are? • How much time/effort (as proportion of a project) is generally required to find and access data? • Is the data registered or indexed in a searchable source? • Is the data publicly available and/or free? If not, why?

Table 2 (continued). Guiding questions (organized by general topic) used during researcher interviews about water data.

Interoperability and Reusability
• Is the data "analysis ready" or, at a minimum, have the sufficient metadata that enables you to use it? If not, what do have to do to make it so? • Is the vocabulary of the (meta)data domain-relevant and consistent? • Are there use restrictions with the data? If so, do they inhibit your desired use in any way?
Dissemination
• Is data dissemination "baked in" to your project(s), an afterthought, or typically not considered? • How do you disseminate data?

3.0 Interview Results

This section summarizes researchers' responses to questions about the findability, accessibility, interoperability, reusability, and dissemination of water data in their research. We first present responses to contextual questions about their research and the types of water data they use (Section 3.1). We grouped together responses pertaining to the findability and accessibility (Section 3.2), and interoperability and reusability (Section 3.3) due to the relatedness of those principles. Finally, we present responses regarding dissemination of research data (Section 3.4).

The material is presented in a Question & Answer format, where the questions are italicized, and answers are either bulleted or in tabular form. Answers have been edited for readability and synthesized to reflect common threads among multiple researchers who shared similar opinions.

3.1 Research Context

This section summarizes questions and answers pertaining to the nature of interviewee's research, types of water data they use, and general reliance on external/internal data sources. These questions are intended to help establish a general context for subsequent questions regarding factors affecting findability, accessibility, interoperability, reusability, and dissemination of water data.

What type(s) of water-related research do you do for WPTO or other sponsors?

- Techno-economic valuation and feasibility of pumped storage hydropower (PSH) development and operation.
- Digitalization of hydropower systems for resilience, sustainability, and decarbonization.
- Characterize differences in irrigation modernization pathways, designs, the role of hydropower, and outcomes based on the heterogeneous physical and jurisdictional conditions that exist.
- Improve representation of hydropower in production-cost models by informing with large scale simulation of hydrologic conditions, environmental constraints, and grid economics.
- Evaluate long-term effects of climate change on hydropower.
- Assess stakeholder needs for, and implement resilience modeling in, interdependent water-power systems.
- Assess gaps in modeling hydropower resource capabilities and system response to contingency events.
- Integration of renewable energy in the grid.

What types of water-related data in Table 1 do you utilize in your research?

Most of the researchers we interviewed use more than one category of water data. Therefore, we attempted to tally for each category the number of projects using those data, specific examples of the types of datasets that are used, and general notes about the accessibility of those data (Table 3). The intent is to illustrate the categories of water data that are essential to research performed for WPTO and other water-related programs. It should be noted that although none of the researchers we interviewed said they have used data pertaining to Migratory Barriers, we are aware of others (including the authors) who do use such data for WPTO-funded and other sponsor projects.

Table 3. Summary of water data categories used by PNNL researchers.

Category	No. of Projects	Example Data Types	Accessibility Notes
Dams	6	Spill; head; reservoir level; licensing	*Privately owned/operated facilities*: Generally proprietary, obtained from owners, other research collaborators. Sometimes requires non-disclosure agreement (NDA). *Federal facilities*: data are publicly available; HydroSource is a valuable resource.
Ecology	3	Fish species presence/absence; abundance; passage; creel census; soils; land use/land cover; bioenergetics; Leaf Area Index	Much of this data is publicly available, although not always easily accessible.
Flood Control	2		*See Accessibility Notes for Dams*
Hydroclimatology	4	Precipitation; temperature; runoff (historic/observed; historic/modeled; future/modeled)	Much of this data is publicly available and easily accessible. PNNL has regional downscaled hydroclimatological data that is available internally. Unsure if it is also available externally.
Hydrography	2	Stream network; canals; Digital Elevation Models	Much of this data is publicly available from US Geological Survey (USGS); e.g., National Hydrography Dataset, Watershed Boundary Dataset, 3DEM, National Elevation Dataset
Hydrology	10	Streamflow (historic, current, and forecasted); well depth/head	Much of this data is publicly available from USGS and state natural resource agencies.

Table 3 (continued). Summary of water data categories used by PNNL researchers.

Category	No. of Projects	Example Data Types	Accessibility Notes
Hydropower	6	Generation; sensor data; operational constraints; flow duration curves; equipment specs; power purchase agreements	*Privately owned/operated facilities*: Most data are proprietary and sometimes accessible with NDA. Some data is available through Hydropower Research Institute, Open-Access Same-Time Information System (OASIS), Independent System Operators (ISOs), and Balancing Authorities. *Federal facilities*: Some data is available through Energy Information Administration (EIA) and HydroSource.
Management Landscape	1	Crop types	Remotely sensed data on crop types (Cropland Data Layer) is publicly available from US Department of Agriculture National Agricultural Statistics Service
Meteorological	3	Weather station data	Most data are publicly available; sources include National Weather Service and National Center for Environmental Information, AgriMet, state and local networks.
Migratory Barriers [(a)]	0	Dams, culverts, diversions	Publicly available through National Anthropogenic Barriers Dataset, HydroSource, and various NGO data sources.
Recreation & Aesthetic Importance	1	Types of recreational water use	Generally, this data is publicly available, although not always easily accessible.
Socioeconomic	3	Hydro infrastructure and operating costs; electricity market data; population economic status	Generally proprietary, obtained from owner/operators, other research collaborators. Sometimes requires NDA.
Water Quality	5	Instream/reservoir water quality sensor data (e.g., DO, TDG); wastewater treatment facility data	Most data are publicly available and easily accessible from the Environmental Protection Agency (EPA).
Water Availability & Use	3	Water demand; irrigation use	Periodically summarized data publicly available from USGS Water Availability and Use Program

(a) Although none of the researchers we interviewed said they have used data pertaining to Migratory Barriers, we are aware of others (including the authors) who do use such data for WPTO-funded and other sponsor projects.

For what types of research activities do you use data?

- Modeling (e.g., hydropower generation, hydropower economics, project feasibility, hydroclimatological processes, evapotranspiration, crop suitability, canal seepage, production-cost modeling, resilience modeling, hydropower capabilities modeling, hydropower grid services, power storage and operational flexibility, renewable energy integration)

- Statistical analysis
- Qualitative analysis
- Visualization
- Decision-support tools
- Systems design

Do your projects depend primarily on external or internal sources, or a combination of both?

We attempted to categorize researchers' dependency on data produced externally by other organizations and data produced internally by PNNL or collaborators (Table 4). Overall, most researchers said they depend almost entirely on data produced by other organizations. While this confirms that the FAIR-ness of data produced by other organizations is a critical factor in PNNL's ability to support WPTO and other water-related programs, it does not dismiss the potential importance of data produced internally. Some researchers did note that it is not easy to find potentially relevant data produced by others within PNNL, or by other national labs for that matter.

Table 4. Summary of researcher's dependency on externally or internally produced data.

Use of external/internal data	No. of Researchers
Most or all the data we use is produced by an external source	11
Some of the data we use is produced by an external source and some is produced internally	2
Little of the data we use is produced by an external source; most is produced internally	0
All the data we use is produced internally	0

3.2 Findability and Accessibility

This section summarizes questions and researcher's answers regarding the findability and accessibility of water data. *Findability* refers to the ability for both humans and computers to search for, identify, and locate data. *Accessibility* refers to the ability of a user to retrieve the data once it is found. These attributes are affected by factors such as who produces the data, how data is stored and disseminated, and the sensitivity level of the data.

How do you find data?

- Through professional or interpersonal connections (e.g., word-of-mouth, call/email directly) with collaborators, sponsors, owner/operators, etc.
- Institutional knowledge of those working on the project and other colleagues.
- Internet research

Are the data difficult to find?

- Data specific to hydropower or other water infrastructure projects are generally difficult to find and must be obtained directly from the owner/operator, ISOs, Balancing Authorities, or industry groups. An exception, to some degree, is data for federally owned facilities. Some types of hydropower data for private and public facilities are accessible through the Hydropower Research Institute with a membership.

- Current and historic hydrological data (e.g., streamflow) is generally easy to get, although spatial and temporal coverage is not always as extensive as users might like.

- Socioeconomic data can be difficult to find due to an apparent lack of data or unfamiliarity with appropriate sources. Finding staff with appropriate expertise and institutional knowledge is most helpful for finding socioeconomic data.

- Hydroclimatological data generally are not difficult to find as most authoritative datasets are made available by governments or the scientific community that steward it.

Do you have to aggregate the data from multiple sources?

- Yes, "...more often than not" or "...almost always..." when it comes to hydropower data.

- For hydropower data, it's often a combination of owner/operators, ISOs, OASIS portals, and Federal Energy Regulatory Commission (FERC) eLibrary.

- In some cases, considerable time is spent searching for additional data to supplement other data sources that are less complete or applicable to the study area.

- When using federal data sources, generally the data provide adequate geographic coverage and do not require aggregating from other state or local resources.

- Usually not, but when multiple sources of the same data exist, we try to aggregate to cross-check the data and fill in spatial/temporal gaps.

Is it clear who the contributors are?

- Generally, hydropower owners/operators are the ones collecting data at/near the project site. Beyond that, it's a combination of federal, state, tribal, and NGO organizations. The fact that multiple entities collect the same type of data can make finding and accessing it difficult.

- For hydrology, water quality, and environmental data, it's generally well known who the primary providers are (e.g., USGS, EPA, federal/state natural resource agencies).

- Some researchers intentionally try to identify multiple sources of the data to vet which is the most appropriate, and to have a redundant source to fill in data gaps in other sources.

How much time/effort (as proportion of a project) is generally required to find and access data?

- 90%

- Varies by project and data type
- 35%
- 10-20%
- 15%
- Varies but is not trivial

Is the data registered or indexed in a searchable source?

- Federal data sources generally provide a search capability (e.g., USGS National Water Information System).
- Most hydropower data, except summarized generation data that must be reported to the EIA, is not findable/accessible in any searchable source and must be acquired from the owner/operator.
- Per FERC rule 18 CFR Part 37, "...each public utility (or its agent) that owns, controls, or operates facilities used for the transmission of electric energy in interstate commerce will be required to create or participate in an OASIS that will provide open access transmission customers and potential open access transmission customers with information, provided by electronic means, about available transmission capacity, prices, and other information that will enable them to obtain open access non-discriminatory transmission service."

Is the data publicly available and/or free? If not, why?

- Hydropower project data (e.g., generation, cost, spill) for privately owned facilities is generally not publicly available. Often, obtaining it requires signing an NDA or other use-limiting agreement.
- Some hydropower project data is available through organizations such as Hydropower Research Institute but requires membership and comes with use restrictions.
- Ecological and physical environment data are usually publicly available.
- Almost all the hydroclimatological data we use is openly accessible, and what little is not, we intentionally avoid.

3.3 Interoperability and Reusability

This section summarizes questions and researcher answers regarding the interoperability and reusability of water data. *Interoperability* refers to the need to be integrated with other data as well as the need for data to interoperate with applications or workflows for analysis, storage, and processing. *Reusability* refers to the ability of a user to make use of the data once it is accessed. For simplicity, we collectively refer to *interoperability* and *reusability* as “usability”. These attributes are affected by factors such as the presence/absence, quality, and clarity of metadata, provenance (or origin) of the data, and level of restrictions regarding data use.

Is the data "analysis ready" or, at a minimum, have the sufficient metadata that enables you to use it? If not, what do have to do to make it so?

- No. The data we receive come in multiple formats, some of which require substantial massaging to make analysis ready.

- Data received from collaborators often requires some data engineering to make it usable.

- Rarely. Often, we need to fill spatial or temporal gaps or reformat to be software readable.

- The data generally require some amount of massaging to make usable for project purposes. Sometimes this goes beyond correcting basic formatting issues into applying analyses (e.g., spatial or temporal interpolation, smoothing, metric derivation) to create derivative data.

Is the vocabulary of the (meta)data domain-relevant and consistent?

- Depends on the data source. The more specific it is, the easier it is to understand and utilize.

- Highly variable. Some are good, some are not even usable.

- Varies by the source. USGS metadata is generally very good.

- No. Often, we must speak to multiple people to get the correct understanding.

- The structure of and level of documentation for hydropower data provided by ISOs varies by ISO.

- It is generally pretty good for hydroclimatological and meteorological data.

Are there use restrictions with the data? If so, do they inhibit your desired use in any way?

- Yes, operational or economic data pertaining to hydropower projects typically comes with use restrictions. Other data derived from these either cannot be disseminated or must be anonymized before disseminating.

- Yes, hydropower data from private entities often require NDA to be signed that constrains how the data may be used or shared.

How much time/effort (as proportion of a project) is generally required to make data usable?

- Less time than the time spent finding and accessing the data.

- More time spent making usable than finding and accessing it.

- Depends on the size of the dataset; larger datasets require more time.

- Not an onerous amount of time.

3.4 Data Dissemination

This section summarizes questions and researcher's answers regarding dissemination of data that is produced through their work. Our goal was to get a basic understanding of how often data dissemination is a part of their projects and generally how they publish their data. Not all researchers we interviewed disseminate their data due to various reasons including use restrictions, lack of funding, not required by sponsor, or perceived inapplicability of their data to other users.

Is data dissemination integrated as a task on your project(s), an afterthought, or typically not considered?

- Case by case basis depending on the client, sensitivity of the data, and sponsor interest.
- Some sponsors (including WPTO) have expressed more desire to disseminate data and, in certain cases, made it a requirement for the project. However, use restrictions that come with proprietary hydropower data often limit our ability to disseminate. Certain types of derivative data are okay to disseminate so long as they are approved by the data provider and protect their interests.
- Data dissemination is typically integrated into the project when other deliverables include software or other tools. In cases when the data may be sensitive, either synthetic data are used, or data is not provided.
- Some sponsors strictly forbid sharing the data.
- No. Typically not required by sponsor or data comes with use constraints that prevent dissemination.
- Planning to create a webpage to make the project and certain outcomes findable and accessible, although not necessarily the data itself as some is considered sensitive or not useful in a general sense to the public.
- Some of the journals we publish in require the data also be published, either by journal or in an open-access repository.

How do you disseminate data?

- Unstructured forms such as tables or appendices in reports or peer-reviewed publications.
- Excel spreadsheets
- PNNL or sponsor website
- In some cases, the demand from our sponsors or user community is a tool that would allow them to generate the data for their area of interest, rather than us generate and distribute the data.

- Our objective is to disseminate a tool rather than data, although the tool will comply with FAIR data principles so that users can easily access, incorporate, use, and disseminate data.

- Engagement with industry and federal agencies.

- Create a custom web-based tool for a specific data dissemination purpose.

- Through free, online repositories such as Zenodo and GitHub.

4.0 Data Dissemination Approaches

As described in Section 3.4, researcher responses regarding whether data dissemination is incorporated into projects varied considerably. When data dissemination is included, it is either due to the desire of the sponsor, or the researcher's own professional or personal motivating factors to share data. Approaches for disseminating data also varied among researchers we interviewed.

In this section, we expand on some of the aforementioned approaches to data dissemination as well as others that we identified. These approaches can be grouped into two general categories: available upon request and hosted services. The primary distinction we draw between these approaches is that the first can be considered *reactive* whereas the latter can be considered *proactive*; i.e., data are disseminated reactively to a request (available upon request), or data are disseminated proactively in anticipation it is desired by others (hosted).

The following sections describe each approach with respect to FAIR principles, general reasonings to their selection, and key advantages and disadvantages. The purpose of this section is to provide WPTO a general understanding about these approaches to facilitate discussion with PIs/PMs about expectations and level-of-effort pertaining to data dissemination for future projects.

4.1 Available Upon Request Services

Researchers often share data on an *available upon request* basis when they receive a direct request (e.g., email, phone call, online request form, in-person request) from their sponsor, collaborators, or other interested researchers. This approach tends to be more informal and generally requires the least amount of effort on the part of the data provider. As such, it is often the preferred approach, especially for "smaller" datasets (i.e., tens to thousands of records). Likewise, it is an approach that many researchers are comfortable with and accustomed to.

This approach is also common for data that may have limited use restrictions, business, or security sensitives, or are perceived as not being generally applicable to a wider audience. Many of the researchers we spoke to that work with hydropower data said this approach is almost always the norm due to use restrictions and sensitivities that often come with that type of data.

The findability of data that are available upon request is generally poor. Often, knowledge of the data must be gained by direct communication with the data provider or word-of-mouth. Thus, socialization of the data's existence is a critical factor to its findability. In some cases, researchers may advertise the data is available upon request at a conference or in an associated journal article or other publication.

The accessibility of data that are available upon request is also generally poor due to poor findability. In best case scenarios, data is sent immediately after the initial request is received. However, this is rarely the case given that researchers providing the data are usually doing so as a courtesy and may require time to prepare the data for dissemination. In addition, it is common that the data provider and requester have some dialogue about the data to discuss its applicability, format, quality, etc., before it is delivered. Common methods for accessing the data once it is ready include email, file transfer protocols (FTP), or third-party file transfer systems such as Dropbox, Box, Google Drive, or Microsoft OneDrive.

The interoperability and reusability of data that are available upon request is highly variable. Data are often delivered “as is” and in unstructured forms such as digital spreadsheets, or tables or appendices in reports or peer-reviewed publications. Consequently, the data may require significant manipulation by recipients to be usable. Metadata, or a set of data describing key information about the data, are generally less detailed or likely to be included. These conditions, however, vary depending on the needs of the data for the project and researcher’s use of the data. For example, projects requiring large datasets or employing more sophisticated analytics generally require higher quality data; subsequently, derivative (meta)data is usually of higher quality as well.

4.2 Hosted Services

Hosted data services are generally the preferred option for researchers who are required or desire to take a proactive approach to disseminating their data. Here, we briefly define hosted services, discuss considerations for their use, applicability to FAIR principles, and describe examples of hosted services that we identified through our interviews.

Hosted services can be generically defined as servers that reside either behind (internal) or outside (external) the organization’s firewall that expose the data through standard internet protocols, enabling others to find and access it on their own. The data is either stored on the host server, an auxiliary (internal or external) server whose location is known by the host server, or in some cases both.

Using hosted services is appropriate when either the sponsor or researcher believes there is sufficient demand for the data that cannot otherwise be met feasibly by *available upon request* approaches. It is also appropriate for “larger” datasets (tens-of-thousands to tens-of-millions records) that are not easily transmitted via email or file transfer systems that are not optimized for large data storage and accessibility (e.g., FTP, DropBox, Box, Google Drive, Microsoft OneDrive). To clarify, “larger” datasets also include datasets that consist of many smaller sub-datasets (e.g., daily hydrologic data for the entire US) or are updated frequently (e.g., gridded climatological data).

Utilizing hosted services generally requires more time and financial investment on the part of the data provider because these services tend to have stricter requirements regarding the format, structure, and quality of the data, as well as accompanying metadata. Financial investment is further increased if the data provider chooses to build, host, and maintain the data service. Thus, it is vitally important incorporate data dissemination plans involving hosted services into early phases of project planning to avoid budget or execution shortfalls.

An advantage of hosted services is they generally enhance the findability of data because they expose data via standard internet protocols. Many hosted services usually also offer some search capability to improve the findability of data stored within the service. The findability of hosted services, however, varies depending on whether they are exposed internally or externally. Their findability is also affected by their interconnectedness to other websites or data services. This may include, for example, advertising the hosted service on websites of collaborators, sponsors, and online news and social media platforms. Most peer-reviewed journals require data sources be formally cited, which also enhances the findability of hosted services.

Hosted services also enhance the accessibility of data by providing the ability to obtain the data without a human in the loop. Generally, data can be downloaded directly by the requester from the host service or linked service. The level of accessibility, however, varies from open to varying degrees of limited access. Open data is data that is free to use, reuse, and redistribute without any legal, technological, or social barriers other than those inseparable from gaining access to the internet itself. Conversely, access may be limited by imposing authorization, financial, legal, or technological constraints on the data.

The interoperability and usability of data is generally improved using hosted services because data must be in a machine-readable format to be hosted. Consequently, hosted data generally requires less manipulation to be made usable. Most hosted services also require structured or semi-structured metadata to provide additional information about the data such as provenance, quality, collection methods, temporal and spatial domain, and other relevant characteristics that improve the usability of the data. The quality of hosted data, however, is ultimately dependent on the level of data standards that are required. The level of data standards varies among hosted services and is generally correlated with their level of sophistication, financial and technical support, and quality demands of end-users.

When choosing a hosted service option, there are a variety of existing externally hosted services researchers can choose from ranging from discipline-specific repositories to general-purpose repositories. By comparison, there are few internally hosted data service options that PNNL researchers can choose from. First, we briefly describe externally hosted services mentioned by researchers, which included Zenodo, GitHub, GitLab, DOE Office of Scientific and Technical Information (OSTI), and HydroShare and HydroClient data services hosted by the Consortium of Universities for Advanced of Hydrologic Science (CUAHSI).

Zenodo is general-purpose open repository operated by CERN, or European Organization for Nuclear Research, that allows researchers to deposit research papers, data sets, research software, reports, and any other research related digital artefacts. For each submission, a persistent digital object identifier (DOI) is minted, which makes the stored items easily citable. Zenodo is free to use and allows users to upload files up to 50 gigabytes in size.

GitHub and GitLab are general-purpose hosting services for software development and version control using Git. While they are designed and optimized primarily for software development, they are also used by some researchers for sharing data, or more commonly, software that obtains and processes data for scientific purposes.

OSTI fulfills DOE-wide responsibilities to collect, preserve, and disseminate both unclassified and classified scientific and technical information emanating from DOE-funded research and development activities at DOE national laboratories and facilities and at universities and other institutions nationwide. OSTI provides access to DOE data and information through a suite of web-based, searchable discovery tools and through other commonly used search engines. Like Zenodo, a persistent DOI is minted for all information submitted to OSTI.

CUAHSI offers to two data hosting services: HydroShare and HydroClient. The HydroShare data repository enables discovery of multiple types of water data and makes data available in a citable manner by minting a persistent DOI. The HydroClient data portal provides access to more than 100 data sources from federal agencies, university researchers, volunteer science groups, and others through a map interface. HydroClient emphasizes time series sensor data (e.g., streamflow measurements, meteorological data, repeated grab sample results, and soil moisture measurements).

Until recently, there were no internally hosted services at PNNL designed specifically to make PNNL data findable or accessible. There is a variety of internally hosted file storage services that can be used to share data with other researchers within the lab but are not designed or intended to be hosted data services (e.g., network share drives, cloud storage, cluster storage). Rather, these services function more as *available upon request* services and do not prescribe to FAIR principles the same way that hosted data services do.

More recently, PNNL has instituted a hosted data service called DataHub that helps researchers address the full data life cycle for their institutional projects and provides a path to creating FAIR data products. DataHub represents the first institutional effort at PNNL to create a hosted data service that enables researchers to proactively disseminate data. In addition to facilitating discovery of datasets, DataHub is also designed to facilitate discovery of research projects, data sources, and people at PNNL. DataHub has both internal- and external-facing nodes that are accessible at http://data.pnl.gov/ and http://data.pnnl.gov, respectively. Most researchers we interviewed were not aware of DataHub but expressed enthusiasm for its development and interest in learning more about it.

5.0 References

Larson, KB, M Wong, KC Mitchel, JD Tagestad, JW Saulsbury, BJ Bellgraph, and DW Reicher. 2021. Improving Discovery, Sharing, and Use of Water Data: Initial Findings and Suggested Future Work. Technical Report PNNL-29737. Pacific Northwest National Laboratory, Richland, WA.

Reicher, DW, J Seagrist, R Dhakal, KB Larson, BJ Bellgraph, JW Saulsbury, JD Tagestad, C Deciampa, E McCann, D Singh, T Ruggles, C Song, W Mo, D Hart, and B Barber. 2021. Improving the Discovery, Access, and Use of Water Data for Basin-Scale River Management: Phase 2 Workshop Report. Internal Report to U.S. DOE Water Power Technologies Office. Stanford Woods Institute for the Environment, Stanford, CA.

Wilkson, MD, M Dumontier, IJ Aalbersberg, G Appelton, M Axton, A Baak, N Blomberg, J-W Boiten, L Bonino da Silva Santos, PE Bourne, J Bouwman, AJ Brookes, T Clark, M Crosas, I Dillo, O Dumon, S Edmunds, CT Evelo, R Finkers, A Gonzalez-Beltran, AJG Gray, P Growth, C Coble, JS Grethe, J Heringa, PAC ‘t Hoen, R Hoof, T Kuhn, R Kok, J Kok, SJ Lusher, ME Martone, A Mons, AL Packer, B Persson, P Rocca-Serra, M Roos, R van Schaik, SA Sansone, E Schultes, T Sengstag, T Slater, G Strawn, MA Swertz, M Thompson, J van der Lei, E van Mulligen, J Velterop, A Waagmeester, P Wittenburg, K Wolstencroft, J Zhao, B Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3(1). https://doi.org/10.1038/sdata.2016.18.

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354

1-888-375-PNNL (7665)

www.pnnl.gov