Pacific Northwest
NATIONAL LABORATORY

PNNL-31454

Deliverable D12 – Distributed Wind Data Catalog Development Guide and Instruction Manual

June 2021

Danielle Preziuso, Pacific Northwest National Laboratory
Anna Maria Sempreviva, Technical University of Denmark
Alice Orrell, Pacific Northwest National Laboratory

DTU

U.S. DEPARTMENT OF ENERGY

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

DISCLAIMER

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-31454

Deliverable D12 – Distributed Wind Data Catalog Development Guide and Instruction Manual

June 2021

Danielle Preziuso, Pacific Northwest National Laboratory
Anna Maria Sempreviva, Technical University of Denmark
Alice Orrell, Pacific Northwest National Laboratory

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

Pacific Northwest National Laboratory
Richland, Washington 99354

Acknowledgments

The authors wish to thank Bethel Tarekegne for her review of this document, as well as the participants of IEA Wind Task 41 for their contributions to the development of the data catalog. Th authors would also like to acknowledge the funding they have received to participate in IEA Wind Task 41: Pacific Northwest National Laboratory has been sponsored by the Wind Energy Technologies Office at the U.S. Department of Energy, and the Technical University of Denmark Wind Energy Department received funding from the Energy Technology Development and Demonstration Program (EUDP) 2019-II IEA Task 41 Journal Number 64019-0518.

Acronyms and Abbreviations

DTU	Technical University of Denmark
IEA Wind Task 41	International Energy Agency Wind Technology Collaboration Programme Task 41: Enabling Wind to Contribute to a Distributed Energy Future
IRPWind	Integrated Research Programme in Wind Energy
PNNL	Pacific Northwest National Laboratory
WP2	Work Package 2: Data Information Catalog for Distributed Wind Research

Contents

Figures

1.0 Introduction

Pacific Northwest National Laboratory (PNNL) and Technical University of Denmark (DTU) completed this deliverable as part of Work Package 2: Data Information Catalog for Distributed Wind Research (WP2) for the International Energy Agency Wind Technology Collaboration Programme Task 41: Enabling Wind to Contribute to a Distributed Energy Future (IEA Wind Task 41).

WP2 contained three deliverables, structured to help researchers understand the potential data contributors and users in IEA Wind Task 41, what relevant data are currently available, and what data are needed. Additionally, the deliverables enabled PNNL and DTU to document best practices around data collection, reporting, and storage. The culmination of these efforts is a distributed wind data catalog for IEA Wind Task 41 participants.

What is a data catalog?

A data catalog is a collection of metadata (information about data) for resources that are physically located at the owner's premises. They can contain information about data in academia, industry, or the public domain. The content and the detail of the information in a data catalog can vary depending on its purpose and the level of information needed from end users. It may include who owns the data, what information the data contain, where the data were collected, where the data entities are located, and how to access it. This metadata allows the data to be found by providing core information.

While a data catalog can take on a variety of formats, the IEA Wind Task 41 catalog is a formatted .xlsx file that aggregates all the metadata of the data that has been collected by Task 41 participants.

Why do we need a data catalog?

Data are the main component of a problem-solving cycle, whether it is scientific, social, technological, or innovative in nature. To optimize any cycle, data and analysis tools (i.e., digital objects) must be readily available. A data catalog of digital objects details each piece of data, providing access to those digital objects and helping individuals build insights, discover trends, and identify new products.

When data are distributed within an organization or among many organizations and individuals, databases are stored in various formats and on a range of media. Aggregating these through a data catalog enhances data sharing for reusability, both inter- and intra-organization. Data reusability depends on whether or not data are documented and catalogs are searchable.

To create the IEA Task 41 data catalog, a focus was placed on harvesting metadata rather than data itself, sidestepping issues around data sensitivity and intellectual property. The data catalog is not intended to be a repository in which to store data; it is instead intended to track available data through metadata. Metadata schemas help answer who, what, when, where, and how questions—what data exist? Who created the data, and who owns it? How can the data be

accessed? When was it created, and where is it applicable? The data catalog contains a series of metadata describing resources that task participants have identified and cataloged. In the specific case of the IEA Wind Task 41's data catalog, we envision that when a task participant needs specific data, they can consult the catalog, identify data sets, and then approach the data owner about conditions for use. The use of that data will be determined on a case-by-case basis between the interested party and the data owner; this will eliminate the need for IEA Wind Task 41 and WP2 to address individual use cases or intellectual property provisions and storage.

The final deliverable for WP2, Deliverable D12 includes a data instruction guide for the IEA Wind Task 41 distributed wind data catalog. As such, this document includes

- A step-by-step explanation of how the IEA Wind Task 41 data catalog was created, how it was populated, and how to use it in Section 2
- Future options for the IEA Wind Task 41 data catalog in Section 3
- Summary with recommendations for future work in Section 4.

2.0 Creating, Populating, and Using the Data Catalog

The resources allocated to WP2 allowed PNNL and DTU to generate a minimum viable product to ensure the findability of data relevant to the IEA Wind Task 41 community. The implementation of a searchable catalog on a public-facing platform was not possible during this initial phase.

Eight key tasks, shown below in Figure 1, were undertaken to create the IEA Wind Task 41 data catalog. These steps were defined by PNNL and DTU as a way of tracking the data catalog progress. While this process was largely linear, the tasks of selecting, refining, and implementing both the metadata schema and the controlled taxonomies were revisited as the metadata collection form was created. This allowed PNNL and DTU to practically implement the best available options, adjusting as needed to meet the specific needs of the IEA Wind Task 41 data catalog. The steps for using the data catalog and collecting and editing metadata are also iterative in nature. The first version of the data catalog is currently available for Task 41 participant use; however, more resources may be collected and previously collected metadata may be edited over time as the data catalog is used. Each of these tasks are described in the subsections that follow and can serve as a guideline for those seeking to create their own data catalog.

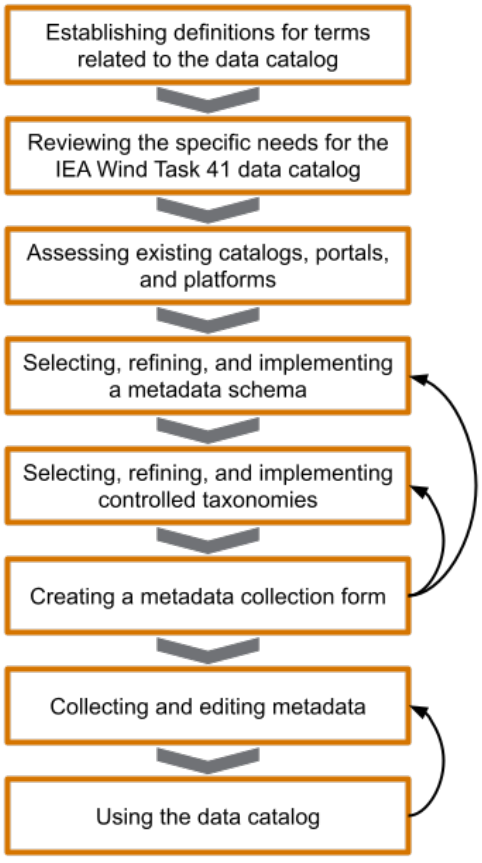

Figure 1. Steps to create the IEA Wind Task 41 data catalog.

2.1 Establishing Definitions for Terms Related to the Data Catalog

When beginning work on the data catalog, it was critical for PNNL and DTU to develop definitions of the different components relating to the data catalog and understand how those components would come together to create the final product. Primarily, the terms *portal* and *platform* were considered. With obvious connections yet clear distinctions between these terms, agreeing on definitions in the early stages of the project promoted a unified and cohesive approach to the catalog's development. The following were adopted:

- Portal: interface allowing a user to access the catalog of metadata. The portal is the entry point in a platform through which people are able to explore the collected metadata. IEA Wind Task 41 has chosen not to develop a formal portal at this time. Participants have direct access to the catalog of metadata in its .xlsx form through the task's members-only SharePoint site.
- Platform: website hosting the portal. The platform is the location at which users can access the portal and, subsequently, the information in the data catalog.

Although IEA Wind Task 41 has not developed a formal portal for the data catalog, the Task SharePoint site may still be considered the platform, since task participants can access the catalog in .xlsx form there.

Additional terms for shared understanding were also defined during this stage. This includes

- Data: detailed records of primary materials that comprise the basis for analysis.
- Primary material: any material that forms the basis of research (e.g., specimens, laboratory notebooks, interviews, texts and literature, digital raw data, recordings, and any other records, including computer code, necessary for the reconstruction and evaluation of reported results of research and the events and processes leading to those results). In the digitization era, primary material is generally called digital objects.
- Metadata: data about data or information about data. Each metadata can give information about administrative, structural, and descriptive matters.
- Metadata schema: a collection of metadata that give complete administrative, structural, and descriptive information about a resource.
- Taxonomies/vocabularies: lists of controlled terms available to use in descriptive metadata. A faceted search relies on those terms for filtering the resource in a catalog.

2.2 Reviewing the Specific Needs of the Data Catalog

Once definitions were established, the specific needs of the IEA Wind Task 41 data catalog were reviewed. This effort began with discussions at the IEA Wind Task 41 Fall 2019 meeting in Boston, Massachusetts, and was later finalized in Deliverable D10.[1] This review allowed PNNL and DTU to prioritize efforts up front and continued to inform decisions made throughout the catalog's development. The following needs, originally published in Deliverable D10, were identified for the IEA Wind Task 41 data catalog:

[1] Preziuso D.C. 2019. *IEA Wind Task 41: Enabling Wind to Contribute to a Distributed Energy Future – Work Package 2: Data Catalog Specification*. PNNL-29510. Richland, WA: Pacific Northwest National Laboratory. https://iea-wind.org/wp-content/uploads/2021/02/M5_milestone_deliverable_D10.pdf.

- A protocol for collecting the data
- Controlled taxonomies and formatting for specific metadata fields
- An understanding of the primary users, contributors, and content to inform the metadata collection process and catalog interface
- Assurance that the catalog embraces the FAIR—findable, accessible, interoperable, and reusable—principle to better guarantee catalog users can access the range of research and information contained within
- Content prioritization by the research needs of other IEA Wind Task 41 work packages to realize the benefits of the catalog in the near term
- An organizational construct that serves the broader distributed wind research community, if possible, to increase the impact of the work.

2.3 Assessing Existing Catalogs, Portals, and Platforms

As a result of the review of needs, it became apparent that many data catalogs, portals, and platforms already exist and could serve as reference points, opportunities for collaboration, or options upon which IEA Wind Task 41 could build. PNNL and DTU reviewed existing data catalogs, portals, and platforms related to IEA Wind Task 41 efforts, including Tethys Environmental,[2] the Data Archive and Portal,[3] Open Energy Information (OpenEI),[4] and ShareWind.[5] OpenEI and ShareWind emerged as the most relevant sources from this initial, high-level review, which initiated a more in-depth assessment of those two options. They were compared in detail by their capabilities, structure, cost estimates for collaboration or soliciting their services, metadata schemas, and implemented taxonomies. They were also compared to a least-cost option (i.e., hosting a .xlsx form of the data catalog on the IEA Wind Task 41 website), which was ultimately selected for the IEA Wind Task 41 data catalog at this time. However, the way in which the catalog has been developed leaves the possibility of coordinating with one of these options, or others, in the future.

A full account of the options that were surveyed can be found in Deliverable D10, and the in-depth comparison of OpenEI and ShareWind can be found in Deliverable D11.[6]

2.4 Selecting, Refining, and Implementing a Metadata Schema

The review of existing data catalogs, portals, and platforms directly informed the metadata schema selected for the IEA Wind Task 41 data catalog. The implemented metadata schema largely replicates the metadata schema adopted by ShareWind,[7] which draws upon the Dublin

[2] https://tethys.pnnl.gov/
[3] https://a2e.energy.gov/about/dap
[4] https://openei.org/wiki/Main_Page
[5] https://sharewind.eu
[6] Preziuso D.C., A. Sempreviva, and A.C. Orrell. 2021. *Deliverable D11 – Data Sharing, Storage, Security Protocols, and a Specification of a Potential Data Sharing Portal*. PNNL-30853. Richland, WA: Pacific Northwest National Laboratory. https://iea-wind.org/wp-content/uploads/2021/02/Task-41-Deliverable-D11.pdf.

[7] Sempreviva A.M., A. Vesth, C. Bak, D.R. Verelst, G. Giebel, H.K. Danielsen, L.P. Mikkelsen, M. Andersson, and N. Vasiljevic. 2017. *Taxonomy and Meta Data for Wind Energy R&D*. https://zenodo.org/record/1199489#.Xefe3ehKhdi.

Core Metadata Element Set[8] and adds metadata fields specific to wind energy data resources. Minor modifications to the ShareWind metadata schema were made to better serve the IEA Wind Task 41 data catalog needs. The following metadata fields, with their listed definitions, were implemented in the IEA Wind Task 41 data catalog from the Dublin Core Metadata Element set:

- Title: name given to the resource
- Creator: entity, or entities, primarily responsible for making the resource available; this might be a person, an organization, or a service
- Subject: key words describing the resource
- Description: account of the resource
- Publisher: entity responsible for making the resource available; this might be a person, an organization, or a service
- Contributor: entity responsible for making contributions to the resource; a secondary figure to the creator; this might be a person, an organization, or a service
- Date: date associated with the creation or availability of the resource, such as the publish or release date
- Type: nature or genre of the resource
- Format: file format or physical medium of the resource
- Identifier: unambiguous reference to the resource, such as a hyperlink, ISBN, or publisher number
- Language: language of the resource
- Coverage: spatial and/or temporal characteristics of the resource, such as country or time frame relevant to the resource
- Rights: information about rights held in and over the resource; this could include a statement about various property right associated with the resource, including intellectual property rights.

The following metadata fields specific to wind energy were developed for ShareWind through the Integrated Research Programme in Wind Energy (IRPWind), a project funded by the European Commission 7th Framework Programme. These are also included in the Task 41 data catalog.

- Variables: parameters measured or tracked within the resource
- External conditions: context in which the resource is relevant
- Activity: type of action conducted within the resource
- Instrument: device used within the resource
- Model: type of model used within the resource
- Material: material components of wind turbine.

[8] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

These metadata fields were previously discussed in Deliverable D10 and defined in Deliverable D11. Additional metadata were collected for the IEA Wind Task 41 data catalog, including the following:

- Notes: additional information about the source not yet cataloged through other entries
- Submitted by: the individual who submitted the metadata
- Country: the IEA Wind Task 41 country represented by the individual who submitted the metadata
- Key words: any other key words describing the resource that were not included in the subject list.

These additional metadata are the result of the review of needs and are unlikely to be published in any formal portal. They were collected to help task participants track who has submitted which data sets, connect with one another, and establish opportunities for collaboration. If the metadata were published, they would include all the information necessary for a potential data user to obtain the data from the data owner.

2.5 Selecting, Refining, and Implementing Controlled Taxonomies

The review of existing data catalogs, portals, and platforms also informed the controlled taxonomies that were implemented into the metadata collection form. These taxonomies help control the terms and formats of the metadata people submit, which ultimately allows the data catalog users to better search the catalog. Rather than generate new taxonomies, PNNL and DTU chose to work from existing taxonomies generated by experts in the field. Eight controlled taxonomies were implemented in the IEA Wind Task 41 data catalog. Of the eight that were implemented, seven were created by IRPWind in support of ShareWind. These include taxonomies for

- Subject (equivalent of topic)[9]
- Variables
- External conditions
- Activity
- Instrument
- Model
- Material.

The taxonomies for these metadata elements can be found at https://github.com/wind-energy/taxonomy-topics. The Dublin Core Metadata Initiative Type[10] was implemented for the "Type" metadata element. Additional taxonomies were considered but not implemented at this

[9] The subject taxonomy that was implemented differs slightly from the original taxonomy developed by IRPWind. It was updated to include additional terms that are relevant to the distributed wind community. Since the IRPWind subject taxonomy is community-based, coordination on the taxonomy will continue. PNNL and DTU recommend reviewing available metadata schemas and semantics when developing a data catalog. Engaging with the community to agree on a shared vocabulary rather than recreating existing efforts is advised.

[10] https://www.dublincore.org/specifications/dublin-core/dcmi-type-vocabulary/

time, since the data catalog is currently limited in size and scope, as well as to Task participant use only.

If a formal portal is implemented for the IEA Wind Task 41 data catalog, additional controlled taxonomies may be considered.

2.6 Creating the Metadata Collection Form

With the metadata schema and controlled taxonomies identified, the metadata collection form was developed. Creating the metadata collection form overlapped at times with selecting, refining, and implementing the metadata schema and controlled taxonomies as PNNL and DTU decided how to structure the form and determined which taxonomies were appropriate to implement at present. The metadata collection form is in a .xlsx file. The form contains instructions on how to fill it out, followed by a series of tabs into which the user can input metadata. Each of those subsequent tabs provides space for the metadata for a single resource.

The form contains a row of headers (i.e., the metadata elements defined above), followed by a row of definitions for those metadata elements, an example input for the metadata, and finally cells into which the user can input the metadata for their own resource. The cells into which a user can input metadata are either white or green in color. Green cells have dropdown menus containing terms from the controlled taxonomies that the user can select from, and white cells are open for the user to enter any information they would like. A partial screenshot of the form is shown below in Figure 2.

Metadata Element:	Title	Creator	Publisher	Contributor	Date	Type
Definition	A name given to the resource	An entity, or entities, primarily responsible for making the resource. This might be a person, an organization, or a service.	An entity responsible for making the resource available. This might be a person, an organization, or a service.	An entity responsible for making contributions to the resource, a secondary figure. This might be a person, an organization, or a service.	A date associated with the creation or availability of the resource such as the publish or release date.	The nature or genre of the resource
Example	Vestas V52 Wind Turbine - Dundalk Institute of Technology	Dundalk Institute of Technology	Dundalk Institute of Technology		2006-2019	Dataset
Your Resource						

Figure 2. Partial screenshot of the metadata collection form.

2.7 Collecting and Editing Metadata

Once the metadata form was created, PNNL began collecting metadata from other Task 41 participants. An initial email was sent to all IEA Wind Task 41 participants and included an explanation of the metadata collection efforts and the metadata collection form. Subsequent targeted emails were sent to individuals who expressed interest in the data catalog or had mentioned relevant data sets during task meetings.

In communication about the metadata collection form, emphasis was placed on two points: submitting an incomplete form was acceptable and sharing metadata does not promise anyone access to data, it simply shows that the data exist. At the time this report was written, 14 submissions from four countries have been collected for the catalog and subsequently edited for clarity. The collected metadata have been aggregated into a single .xlsx file that mirrors the format of the metadata collection form. This file serves as the IEA Wind Task 41 data catalog and is described below. As additional metadata are collected, they will also be edited for clarity and added to the catalog.

2.8 Using the Data Catalog

Using the data catalog can entail one of two things: adding new metadata to the catalog (done by someone who manages the data catalog) or searching the data catalog for needed resources (done by someone simply accessing the posted data catalog). To add new metadata to the catalog, it is important to note that each heading within the data catalog corresponds to a metadata field in the metadata collection form. The metadata provided by task participants can, therefore, be copied and pasted into the corresponding fields in the data catalog. There are some fields in the metadata collection form (i.e., subject, variables, conditions, activity, instrument, model, material) that limit the user to selecting a single term per cell. As such, those cells must be aggregated to fit the data catalog's one resource per line format. For consistency and streamlined opportunities to perform a batch upload of this file into any future portals, those terms are concatenated with a semicolon followed by a space ("; "). When an update is made to the data catalog file, the corresponding date in the file name is updated.

In comparison, to browse the catalog, a user can simply filter down the header columns through the built-in Excel features to search for the terms that are relevant to their needs.

If a formal portal is established on a platform in the future, the Excel file will serve as the backbone to that and provide a standard format for the metadata describing the resources that have been cataloged.

3.0 Further Recommended Features of a Data Catalog

While IEA Wind Task 41 created a minimal viable product for WP2, additional features can be added to this data catalog or included in the development of future data catalogs with supplemental funds. These options, explained below and adapted from *A Step-by-Step Guide to Build a Data Catalog*,[11] are in addition to supporting the data catalog on a platform with a formal interface.

- Web crawler: While the IEA Wind Task 41 data catalog was manually populated by task participants, a web crawler that automatically searches the web can also be implemented. Through such a mechanism, the data catalog can be regularly updated without significant labor hours and access a broader scope of data if it is documented in a compatible way.
- Data statistics: Statistics summarizing the data contained within a resource can help data users evaluate whether the data they have found matches their intended use of the data.
- Resource relationships: Enhancing relationships between data from different owners allows users to understand where related data exist within the network tracked by the data catalog. Including these relationships in a data catalog can increase the value of data for users. For instance, it can help a modeler in estimating uncertainties in models by having several resources available across multiple databases. Human knowledge, a metadata mining algorithm, and documented information from queries all present opportunities to better identify and mark data.
- Data lineage: If relationships between data are tracked within a data catalog, it becomes possible to build lineage, as well. Visual representations are particularly useful when depicting the path between data creator and data user.
- Accessibility and security: While a data catalog should be easy to find and use via a web browser or applications, security protocols with clear governance are key components for enabling long-term success. Implementing safeguards to protect the underlying data tracked within the catalog is critical, particularly when the data catalog is open for public use.

While these additional features offer increased capabilities for the data catalog, the amount of time and funding required to implement them will vary based on the capabilities of those involved and the amount of data that need to be cataloged. Some of these additional features may only take a few days, while others could take weeks.

It should be noted, however, that since any data catalog depends on a distributed community of data providers, it is critical for organizations to implement data management systems based on metadata. At the bottom of the chain, the responsibility of organizing data is on the shoulders of the researcher who collects the data. Many ways to achieve this practice can be implemented. The European Commission, for example, has established guidelines for each funded project to adopt a Data Management Plan. Specifically:

> *A Data Management Plan describes the data that form the basis of your research project. It contains details on how you want to collect, structure, analyze and publish the data, how you deal with external requirements and what value your data might have for*

[11] Varshney S. 2018. *A Step-by-Step Guide to Build a Data Catalog.* https://www.ovaledge.com/blog/step-step-guide-data-catalog.

other researchers and the public. It is an important part of any research project and covers all aspects of the Data Life Cycle.[12]

[12] Refn A. 2018. *Data Management Plans*. https://www.bibliotek.dtu.dk/english/servicemenu/publish/research-data/guide/before/dmp.

4.0 Summary and Future Work

This deliverable summarizes the eight key steps that PNNL and DTU followed to produce, use, and maintain the IEA Wind Task 41 data catalog. Considerations for future features of the IEA Wind Task 41 data catalog or other future data catalogs were also described. While this report is the final deliverable for WP2 in IEA Wind Task 41's current cycle, future work may include the development of a portal on a platform or the integration of the data catalog into another data catalog. Implementing additional features outlined in section 3 will also be considered as part of a future IEA Wind Task 41 Work Plan.

Pacific Northwest National Laboratory

902 Battelle Boulevard
P.O. Box 999
Richland, WA 99354
1-888-375-PNNL (7665)

www.pnnl.gov