Open Energy Data Initiative (OEDI) Data Lake

Data Management and Submission Best Practices

April 2022

National Renewable Energy Laboratory (NREL)

Table of Contents

Introduction

This document is intended to offer advice to users of the Open Energy Data Initiative (OEDI) for organizing and storing data prior to submission for the purposes of simplifying the data submission process and maximizing data accessibility, usability, and longevity. This document also provides detailed information about the OEDI data submission process, including what data to submit, what metadata to include and where, and what happens to data following submission. Please also see the OEDI submission training videos available online.

Data Management and Sharing Best Practices

Managing data properly requires a comprehensive strategy outlined at the beginning of the project lifecycle. Early strategic decisions on file organization, naming conventions, and metadata collection can not only streamline data sharing and eventual dissemination, they can also help improve data access for project team members in increase the overall efficiency of the project as a whole.

Metadata Collection

It's important to collect all relevant metadata as data are being collected in preparation for eventual dissemination. Properly documenting metadata while the details are still fresh in the minds of researchers helps to avoid loss of this information and streamlines the data submission process. Researchers should document as much metadata as possible, including, but not limited to:

- where the data were collected (including maps of sample locations),
- resolution of the data, units, and any assumptions,
- survey spacing and timing,
- any problems encountered where data were lost or recorded improperly,
- anomalies or significant events (e.g. equipment malfunction),
- specific methods used to create the data, and
- any filtering or down-sampling methods applied to the data.

File Organization

Proper file organization is critical in allowing intuitive access to data by others. When files are disorganized, public usability of data is reduced. When storing data with the intention of uploading it to OEDI, it is generally best to upload files individually; however, groups of files that can only be used together, such as shapefiles, should be grouped together in archives. In addition, if you are uploading an overwhelming number of files related to the same dataset, it may be intuitive to store them in a directory rather than uploading them individually. When storing files in directories, they should be structured in an intuitive way, with readme files and data dictionaries to provide supplemental information where needed. During the data submission process, you will be asked to provide a unique description of each resource uploaded. A good rule of thumb is, if you cannot think of a unique quality to distinguish one data file from another, then they are suitable for combination in an archive. Any files that are describable or usable individually, should probably be uploaded as independent resources.

File Naming Conventions

File naming is also essential for ease of access to data by others. Files should be named using a consistent, concise, and meaningful naming convention. If naming conventions could be considered unclear to those outside of the project, either explain the naming convention or create a data dictionary. When naming files, please be mindful of potential future iterations on data, results, and even measurements. Consider the use of vintage and version information in the file name. For example, data collected in 2020 may undergo a revision in 2021 (or subsequent years) to address a recording anomaly, which would ideally result in a file named "data-2020-v2" or "data-2020-revised-2021".

File Formatting

Preferred formats are those that support the best reusability (e.g. non-proprietary formats); however, OEDI accepts a variety of file formats and will, in most cases, accept your submission in whatever format you wish to provide it. For data available in multiple formats, please consider the following guideline when choosing which format to submit. The tiers in Figure 1 are arranged in order of increasing inherent reusability.

Figure 1: The three tiers of data

Data Pipelines and Standards

Tier 3 data delivery is the preferred scheme, but because of the additional effort required to edit and review datasets for Tier 3 delivery, data standards and data pipelines are being created for non-tabular data and other datasets that are best represented in formats other than spreadsheets and/or XML. Such data include but are not limited to geospatial data, video files, audio files, picture libraries, 3D models, programs, code, and timeseries data or other data streams.

Geospatial Data

The preferred format for geospatial datasets is a shapefile. Be sure to upload all files necessary to use the shapefile, including the main file (.shp), the index file (.shx), and the dBASE file (.dbf). Include any optional files as well, such as the projection file (.prj), extensible markup language file (.xml), spatial index file (.sbn and .sbx), and code page file (.cpg). Make sure that geographic projection system is included either as a .prj file or in the submission's text. The more metadata you include, the more useful the data will be for the next user. All shapefile components should be archived together into a single file before upload, as many of the components are useless without the others (e.g. "GIS data shapefile.zip").

GIS software such as Quantum GIS, ESRI ArcMap, and MapInfo all define application-specific project file types that record information about structure, organization, and portrayal of workspace content in a single bundle (Guidelines for Provision and Interchange of Geothermal Data Assets, 2016). These GIS workspace packages are often preferred over shapefiles when the structure, organization, and portrayal of workspace content is important for understanding the data.

Large Datasets

OEDI does not have any limits on file size nor number of files per submission. However, larger files may be difficult to upload over some internet connections, especially shared connections. If you have concerns about your file size or are having trouble uploading a large file, please contact us.

As a general rule, datasets on the order of 10 TB or more or overly complex or multi-dimensional datasets (even if smaller in size) may be good candidates for submission to the Open Energy Data Initiative (OEDI) Data Lakes. The OEDI Data Lakes are centralized repositories of high-value DOE-funded research datasets that have been integrated with OEDI. Centralized data is advantageous for collaboration due to the provided ability for users to work with the data where it is stored, reducing duplication of effort. A centralized data location also reduces the overall cost of storage and analytics of large data sets. Figure 2 shows the OEDI Data Lake concept. Data stored in the lake is actionable and discoverable, increasing accessibility to large and complex datasets (Brodt-Giles and Rossol 2019). In this concept, data that are too large or complex to conveniently be downloaded via conventional means are made accessible in formats that support cloud-based high-performance compute and parallelization (such as hierarchical data formats, or hdf), allowing potential users of the data unrestricted access to massive datasets without the need for high bandwidth connections or high-performance computational solutions. The cloud-based nature of the data lakes allows users to access and manipulate large data without having to directly download data. If you think your data is a good fit for the OEDI Data Lakes, you can contact OpenEI.Webmaster@nrel.gov, or visit https://data.openei.org/data_lakes to learn more about the OEDI Data Lakes.

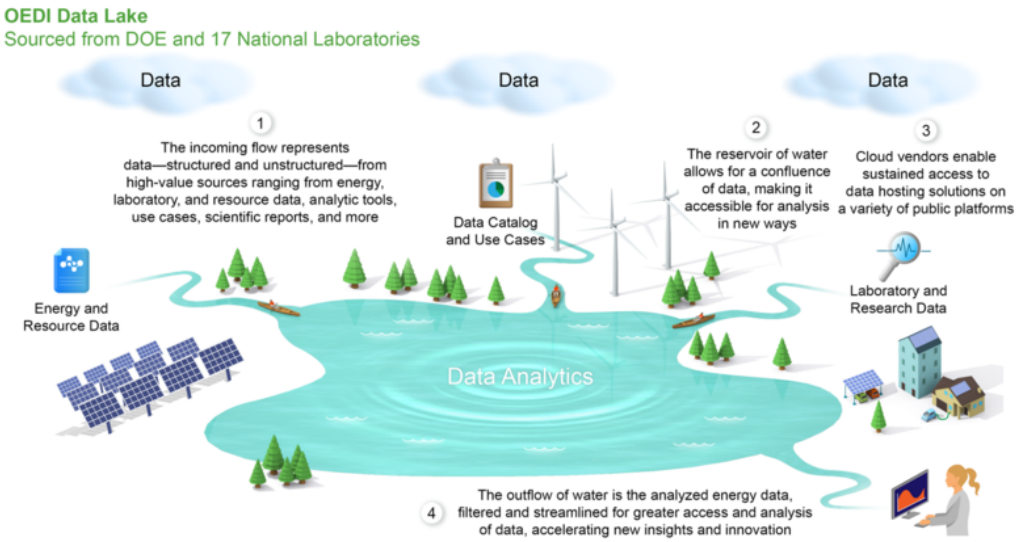

Figure 2: Diagram showing the OEDI Data Lake concept.

3D Models

Various software packages are in use to generate 3D models, but standard interchange formats for models have not been adopted. The National Archives indicates several 'preferred' or 'acceptable' formats for Computer Aided Design (CAD) vector graphics that are likely to be suitable. The degree of adoption and support for these formats by modeling software vendors is not known.

In the near term, the recommended course of action is to:

1. Upload models in the file format native to the software in use,
2. Upload a copy of the model in an export/interchange format if any such format is available in the software package
3. A viewable and manipulable version of the model should be uploaded if there is an easily accessible, free software package (with a link or description of how to access it) that can be used to view the model. This may utilize the format from items 1 or 2 above or be another format.
4. Metadata for model should contain references/links to each dataset included in the model along with description of how the data subset was selected (if applicable). The details of the file format versions and software necessary to use the files should be clearly documented.
5. High resolution (300 dpi) images should be included as well, providing visualization of the model from useful perspectives.

File Storage

It is important to choose a proper solution for file storage early on to avoid complications and confusion later. All files should be stored in an organized, secure, collaborative location to avoid data loss, simplify the process of submitting the data to OEDI, and to allow a convenient file sharing solution for teams.

While storing datasets on hard drives may seem like a viable solution, it is important to consider drive storage capacity, limits on the number of files, and hard drive failure rates (on average around 1.27%) (Weers and Huggins 2019). Cloud-based solutions such as OEDI and Data Foundry do not have these limits, and therefore provide more robust and reliable long-term file storage solutions.

Effective cloud storage requires high-speed connections to the cloud, which may not be possible in the field. This can require creative storage solutions both out in the field and back at the lab, or wherever data analyses, visualization or modeling activities are taking place. For collaborative projects or larger organizations, these could be distributed among multiple partners creating a need not only for multiple storage solutions, but also for an efficient transport mechanism capable of moving large amounts of data (Weers and Huggins 2019).

Data Foundry

The Data Foundry is one example of a file storage solution for DOE-funded research that is hosted by OpenEI, which provides access to open energy information, data and resources.

The Data Foundry provides secure, cloud-based storage and universal access to digital information, enabling the greater scientific community to collaborate seamlessly with government agencies, national labs, universities and private organizations. The Data Foundry includes a user-friendly file management interface that allows authorized users access to project data. It allows the creation of teams for organizing groups of people who may require access to multiple projects.

OEDI Submission Best Practices

What to Submit to OEDI

DOE's Open Energy Data Initiative (OEDI) was established to receive, manage and make available all high-value energy research datasets aggregated from the U.S. Department of Energy's (DOE's) Programs, Offices, and National Laboratories. OEDI facilitates access to a broad network of findings, including the data available in technology-specific catalogs like the Geothermal Data Repository and Marine Hydrokinetic Data Repository. While OEDI is funded and operated by the U.S DOE, non-DOE-funded data are also accepted into the repository. The only requirement is that the data are related to energy research. Also note that the metadata requirements are slightly different for non-DOE-funded data submissions.

In addition to submitting raw or processed datasets related to energy research, links to conference papers, journal articles, and/or final technical reports are useful to include to provide more detailed supporting information.

Certain data types and formats are better suited for reuse that others. Submitters to OEDI are encouraged to provide access to raw data along with the summarized, final data products typically associated with their project. This is because summary data is intrinsically biased towards a specific result, while raw data is unbiased and can be used in new, unforeseen ways.

What Not to Submit to OEDI

Any personally identifiable information, business proprietary information, or copyrighted material should NOT be submitted to OEDI.

Personally Identifiable Information (PII) is any piece of information or combination of pieces that could be used to compromise the identity of an individual. A person's name alone is not considered PII, especially in the case of attribution. Contact information, such as email and home addresses, should not appear any submitted data. A submitter's contact information is required but will only be used for questions about the data submission. Contact information for organizations is ok, including office email, the office address, coordinates, and phone and fax numbers. Personal information, such as home telephone numbers, email and home addresses, and birth dates is not allowed. Furthermore, private information, such as social security numbers, bank account numbers, passport and driver's license numbers, is expressly forbidden. All submissions should be purged of PII prior to submission.

Business Proprietary Information (limited rights) should also not be included in the data submitted. Data submitted will eventually be made available to the public. Data subject to copyright, business arrangement, publication or purchase agreement, and all data not authorized for eventual public release should not be uploaded.

Copyrighted Material of any kind, including published or pending journal articles, should not be uploaded to the catalog. When publicly available elsewhere, these data can, however, be linked to, if permanently hosted on other sites, using the Add Link button.

The Main Idea

When submitting your data to OEDI, you should aim to make your data as usable as possible for the next person interested. OEDI is a tool intended to facilitate communication and data sharing between members of the greater research community and those advancing energy-related research. OEDI is focused on interoperability, knowledge sharing, and communication of its data catalog with partner sites (see Figure 3). Preserving and allowing open access to data is not just about exposing them to the public to satisfy a requirement. It also entails opening them up to the possibility of reuse in new and exciting ways. To this end, data submissions should be formulated similarly to a conference paper intended to be presented to scientific peers. All data submitted to OEDI is ultimately disseminated to a larger network of scientific data repositories.

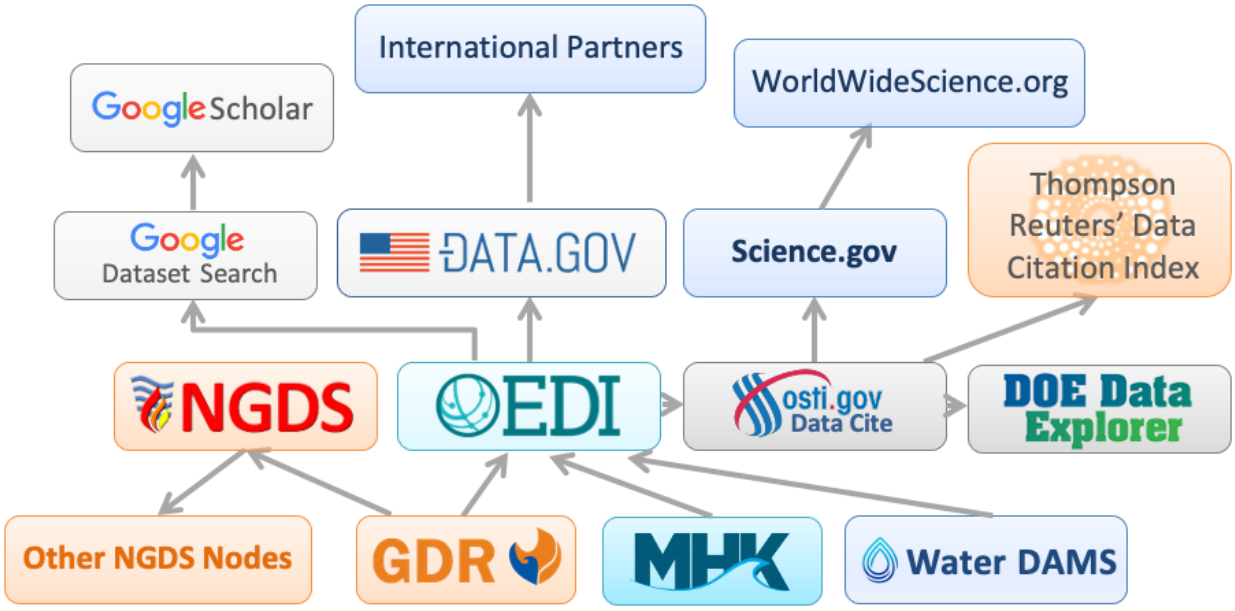

Figure 3: Propagation of metadata through the network of OEDI data-sharing partner sites.

How to Organize Submissions

Data can be submitted as a single, consolidated submission or in multiple submissions. An individual submission can contain an unlimited number of data resources (files and links), but each resource must have a unique name within the submission. Submissions should be grouped into logical sets, associating like data together so that elements necessary for the comprehension of a resource are not in a different submission. If needed, a previous submission may always be linked to from a newer submission as one of its resources.

Combining resources by zipping or archiving should only be done when the resources are of little use individually. For example, the zipping of individual shapefile components into a single shapefile resource is strongly encouraged. Zipping is also recommended when submitting large

quantities of files which are otherwise unable to be adequately organized. In this case, a separate file describing the structure and contents of the files should be included to allow ease of navigation.

Submission Name

Submission titles should be as descriptive as possible, without being overwhelmingly long. As a convention, OEDI submission names are to include the project name or abbreviation, the type of data, and the location that the data is applicable to. If the project name is long, it should be abbreviated using the standard abbreviation for the project. If the data are applicable no matter the location, then the location should be left out of the title. Figure 4 shows an example submission name associated with an existing OEDI submission. In this example, "Solar-to-Grid" is the project name, "Photovoltaics Generation, Capacity Credit, and Value" indicated the type of data, "2010-2020" indicates when the data were recorded, and "Utility-scale and Distributed" indicates the scale of the data. Notice that abbreviations are written out completely the first time they are used.

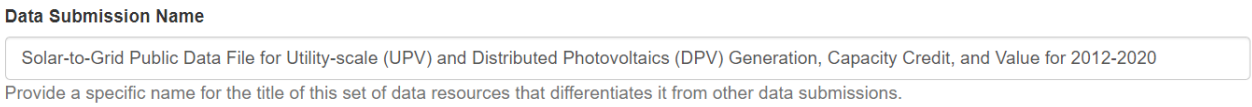

Figure 4: Example submission name from an existing OEDI submission

Abstract

The abstract should describe the submission as a whole, including information about when the experiment was carried out, the types of data that were collected, and any general nuances of the data. Consider the following questions, and whether or not their answers apply to the submission as a whole, when describing data files:

- What is in each data file?
- When, where, why and how was the data was captured/collected?
- Are the units for the data obviously and unambiguously labeled?
- What would someone need to know to use the data properly?
- Are there any assumptions, proprietary software requirements, or other prerequisites to using the data?

The answers to these questions that are not unique per data resource should be included in the submission abstract. If they are unique per data resource, the responses should be included in the data resource descriptions instead, which is discussed in the Resource Description section below.

An example abstract from an existing OEDI submission is shown in Figure 5. This example specifies the data included is hourly, project-level distributed PV generation data at both the utility and county scale covering the span of 2012 through 2020. The abstract mentions that the data are organized into seven wholesale markets and ten additional balancing areas, and are updated on an annual basis. Notice that the description points to the links in the submission's resources for additional information rather than including a link, because submission abstracts cannot include hyperlinks.

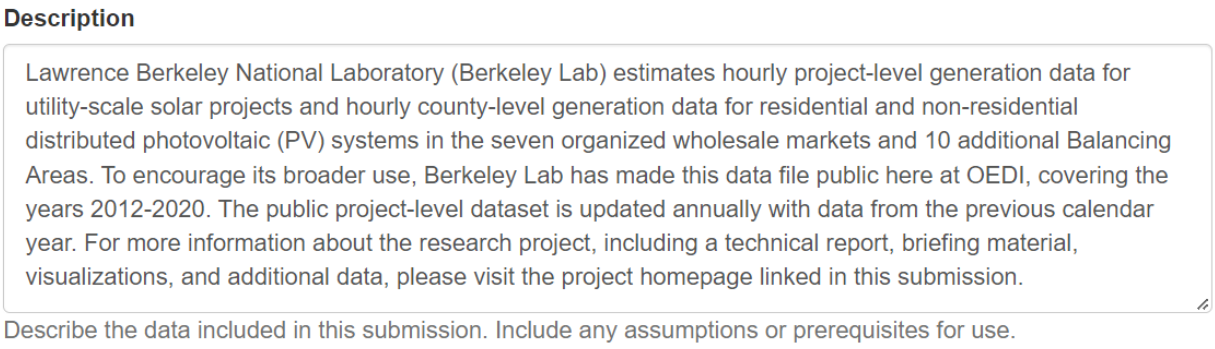

Figure 5: Example abstract or description from an existing OEDI submission

Submission Keywords

Keywords are added to OEDI submissions to help organize data and increase discoverability by making the dataset appear in searches. The keywords "energy" and "power" are automatically added to each submission to help datasets appear in broad search engine results but can be removed by hitting the small gray 'x' in either box or using the backspace key on your keyboard.

The OEDI search page has a faceted search option that allows the user to browse and select structured facets representing available data (Figure 6). Facets filter submissions using keywords along with the other metadata associated with a submission, and are grouped by research area, accessibility, data type, organization, and source. In the future, these search facets will be expanded to include additional keyword-populated categories.

In order to maximize discoverability and support population of future data facets, each submission should ideally be categorized as at least one research area, technology, data type, and project although this is not always the case. For example, an assessment of renewable energy potential within an area may not have a known associated technology yet, or in the example in Figure 7, there is no consistent data type. In such cases, it is okay to exclude some of the suggested keywords.

Tables 1-4 provide keyword suggestions for each type of submission. Keywords must align with "Strongly Suggested Keywords" to be sorted properly into search facets. "Additional Suggested Keywords" are meant to serve as ideas for further increasing the discoverability of each OEDI submission. The user should use these suggested keywords as a starting point, and then expand by adding other applicable keywords.

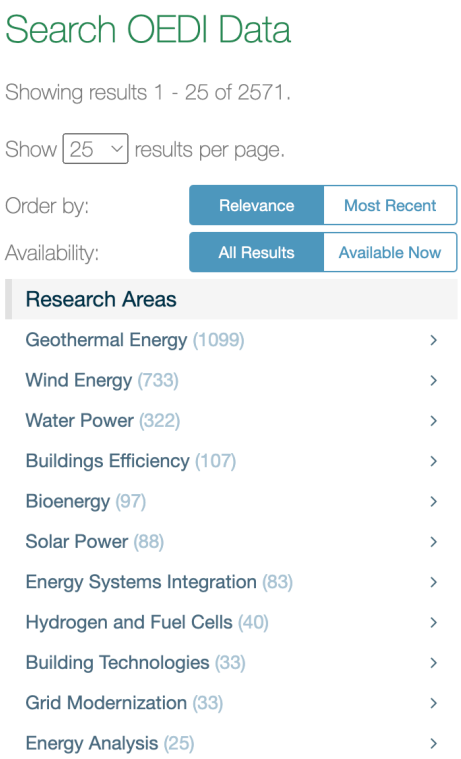

Figure 6: The Research Area portion of the OEDI search facets. Note that the rest of the search facets are in progress.

Figure 7 shows an example list of keywords associated with an existing OEDI submission. In this example, the project is “Solar-to-Wind,” the technology type is "PV" or “Photovoltaics,” and the associated topic is “hourly generation” data. The “solar” research area included in Table 3 is applicable to the data in this example submission.

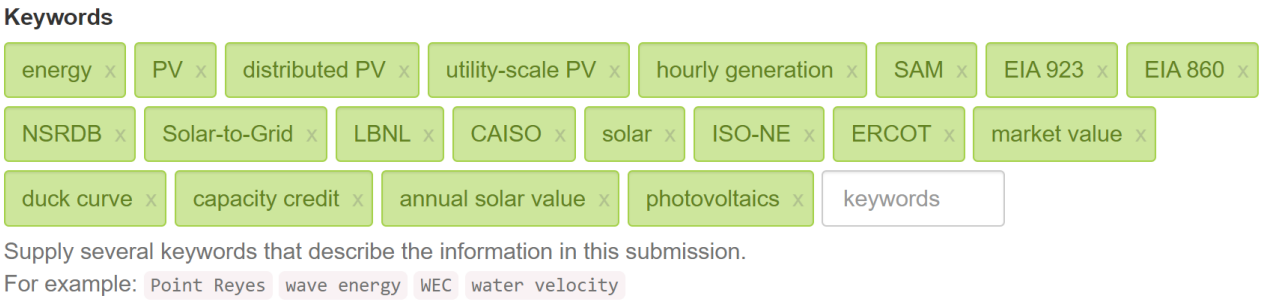

Figure 7: Automatic keywords, custom keywords, and keyword input on OEDI submission form

Research Areas

Research areas are defined as the categories that most OEDI energy research overlaps with that are not limited to a specific technology. Table 1 shows the suggested keywords for some of the most common research areas associated with OEDI submissions.

Table 1: Research areas		
Research Area	**Strongly Suggested Keywords**	**Additional Suggested Keywords**
Advanced Manufacturing	“advanced manufacturing”	Branch or sector of advanced manufacturing research. For example, “energy intensive technology” or “industrial process utilities”.
Agrivoltaics	“agrivoltaics”	Branch or sector of argivoltaics research. For example, “crop production” or “livestock grazing”
Bioenergy	“bioenergy”	Branch or sector of bioenergy research. For example, “biofuel” or “biomass”.
Buildings Efficiency	“buildings efficiency”	Branch or sector of building efficiency research. For example, “residential” or “commercial”.
Chemistry	“chemistry”	Branch or sector of chemistry research. For example, “geochemistry” or "organic chemistry”.
Computational Science	“computational science”	Branch or sector of computational science research. For example, “high performance computing” or “machine learning”.
Energy Analysis	“energy analysis”	Branch or sector of energy analysis research. For example, “utility rates” or “demand”.
Energy Systems Integration	“energy systems integration”	Branch or sector of energy systems integration research. For example, “utility” or “grid”.
Geothermal Energy	“geothermal”, “geothermal energy”	Type of geothermal technology. For example, “EGS” or “direct use”.
Grid Modernization	“grid”, “grid modernization”	Branch or sector of grid modernization research. For example, “combined heat and power” or “smart grid”.
Hydrogen and Fuel Cells	“hydrogen”, “fuel cells”	Branch or sector of hydrogen and fuel cell research. For example, “hydrogen

		compatibility" or "hydrogen demand".
Marine Hydrokinetic Energy	"marine hydrokinetic energy", "marine hydrokinetic", "MHK", "marine", "hydrokinetic"	Type of MHK technology. For example, "WEC" or "CEC".
Materials Science	"materials science"	Branch or sector of material science research. For example, "brittle ductile transformation" or "material physics".
Solar Power	"solar", "solar power"	Type of solar power research. For example, "PV" or "CSP"
Transportation	"transportation"	Branch or sector of transportation research. For example, "fuel stations" or "alternative fuel".
Water Power	"water", "water power"	Type of water power research. For example, "hydropower" or "pumped storage"
Wind Power	"wind", "wind power"	Type of wind power research. For example, "wind turbine" or "distributed wind".

Technologies

Technologies are defined as techniques for carrying out a specific energy-related goal, such as photovoltaics (PV), enhanced geothermal systems (EGS), or point absorber wave energy converter (WEC). Submissions may classify as more than one technology, for example municipal water treatment may encompass wastewater treatment, desalination, and reuse. Some technologies may be broken down into sub-technologies to further refine search results. Others are relatively new technologies and have not yet developed clearly defined sub-technologies. Table 2 shows the recommendation for determining keywords associated with technology types. This is subject to change as these technologies progress. Technologies are not always relevant to submissions. For example, submissions that convey information such as weather data or energy demand data may not have a relevant technology. In these scenarios, there may be a parallel category that is relevant, such as location that the data are relevant to or data collection method.

Table 2: Technologies

Strongly Suggested Keywords	**Additional Suggested Keywords**
The name of the technology (e.g. "PV", "EGS" or "point absorber")	Any words describing how this technology or experiment varies from the norm (e.g. "rooftop PV", "superhot EGS" or "extreme wave environments")

Data Types

Data types keywords are mostly based on the assigned resource type for each resource; however, some of these categories are rather vague. To reduce ambiguity, OEDI splits some of its data

types up further using keywords. While code resources may be faceted using assigned resource types, it is useful to include the associated languages and programs so that users may further refine results if needed. Table 3 shows suggested keywords associated with various data types.

Table 3: Data Types – Beyond those Automatically Assigned		
Resource Type	**Strongly Suggested Keywords**	**Additional Suggested Keywords**
Geospatial Data	"geospatial data"	type of geospatial data (e.g. "shapefile", "GeoTIFF", "raster", etc.)
Raw data	"raw data"	condition of raw data (e.g. "pre-processed" or "preprocessed", etc.)
Processed Data	"processed data"	condition of processed data (e.g. "reprocessed" or "re-processed", etc.)
Code	"code", "algorithm", "software package", or "application"	associated program(s) or language(s) (e.g. "MATLAB", etc.)

Projects

Even if the project name is already included in the title or description, or if the project is not well known, an abbreviated version of the project name should always be included in the keywords to enable project searches. Table 4 includes a list of suggested keywords related to projects in general.

Table 4: Projects – General (non-project-specific) Keyword Suggestions	
Strongly Suggested Keywords	**Additional Suggested Keywords**
Important, descriptive words within project name, project abbreviation or acronym, specific project location if more than one project location	Terms related to the project research area, specific location or associated aquifer

Research Areas

The Research Areas section asks that the user selects one or more research areas that are applicable to the data being submitted. Figure 8 shows the possible research areas, which include "Advanced Manufacturing," "Agrivoltaics," "Bioenergy," "Building Efficiency," "Chemistry," "Computational Science," "Energy Analysis," "Energy Systems Integration," "Geothermal Energy," "Grid Modernization," "Hydrogen and Fuel Cells," "Marine Hydrokinetic Energy," "Materials Science," "Solar Power," "Transportation," "Water Power," and "Wind Energy." Figure 8 also shows an example research area selection associated with an existing OEDI submission. In this example, "Energy Analysis", Energy System Integration", and "Solar Power" are the relevant research areas.

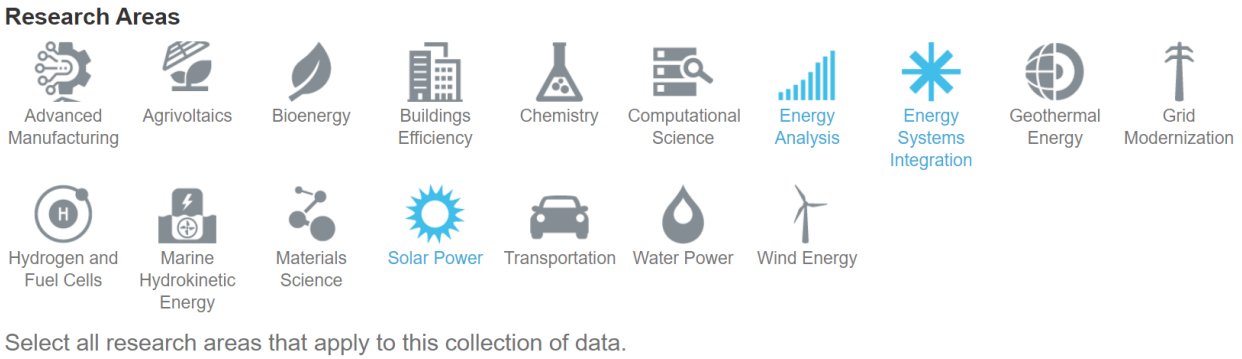

Figure 8: Research areas applicable to the data being submitted.

Organization and Contact Information, Authors, and DOE Project Information

Organization and Contact Information

The OEDI submission form requires an organization and contact information to supply this information to users who may have questions about the data contained in the submission. This information is also used by curators during the curation process if the curator needs clarification, additional information, or additional data. This section requires a contact name, email, and organization. The phone number field is optional. Also required is an origination date for the dataset. This should be the date that the most recent resource in the submission was finalized.

Authors

The authors section is intended to give contributors credit for their work in putting together the data and associated publications. All contributors should be included with a first name, last name, and organization. If available, ORCID iDs may be included as well.

DOE Project Information

DOE project information is required only for DOE-funded data submissions for the purpose of data organization and record-keeping. It is important that this metadata is correct because it's used for organizing submissions by project, allowing related data submissions to be displayed at the bottom of each submission. In addition, GTO project leads listed within this section of a OEDI submission receive an email upon submission and will soon also receive an email when the data submission is published. This allows GTO project leads to monitor individual projects' data submissions to ensure project teams are meeting their data submission requirements as laid out in their data management plans.

Data Resources

A user can add as many data resources as they like to a OEDI submission. For recommendations on how to organize files and data resources, see the File Organization section under Data Management Best Practices above.

Data resources may be either links or files, and require the user to add a display name, resource type, resource description, creation date, and location for each. All data resources also have an

associated size which is determined automatically by OEDI. Figure 9 shows an example list of data resources associated with an existing OEDI submission.

Links

Links to files or websites may be added if relevant data is permanently and publicly available elsewhere. This is done by clicking the 'Add Link' button. The link you submit must be a permanent URI (i.e. a URL that leads directly to a resource and does not pass through a search page or require more than one click to navigate to the data). Figure 10 shows examples of good and bad URLs.

Display Name

Each data resource is automatically given a display name that is equivalent to the file name. The display name should be changed if the file name is unclear about what is contained in the data resource or if the file names do not clearly separate each file from other files in the submission. Changing the display name does not change the file name, it only changes how file name is displayed within OEDI and the OEDI Data Partners. Figure 9 includes an example of how to name data resources.

Resource Type

The submission form will auto select a resource type based off the file extension, but this selection not always correct. Please double check and change resource type if necessary. Options in the resource type drop-down menu are determined by the file extension. If you don't see accurate resource type, select "Other." Resource type affects Data Type search facet population, but also determines whether or not the OEDI submission may receive a digital object identifier (DOI). DOI's are not assigned to Documents or Presentations because the U.S. DOE's Office of Scientific and Technical Information (OSTI) does not consider these resource types to be "data." Figure 9 includes some example resource types.

Resource Description

The resource descriptions should not be the same as the submission description. They should briefly describe the individual data resources, what makes them different from the other data resources in the submission, and their nuances. Consider the following questions, and whether or not their answers are unique per file, when describing data files:

- What is in each data file?
- When, where, why and how was the data was captured/collected?
- Are the units for the data obviously and unambiguously labeled?
- What would someone need to know to use the data properly?
- Are there any assumptions, proprietary software requirements, or other prerequisites to using the data?

If the answers to these questions are unique per data resource, this information should be included in the resource descriptions. Otherwise, the responses should be included in the submission abstract instead, as discussed in the Abstract section above. Figure 9 includes an example resource description.

Creation Date

A creation date should be added to each individual data file to describe when each file was created or last updated.

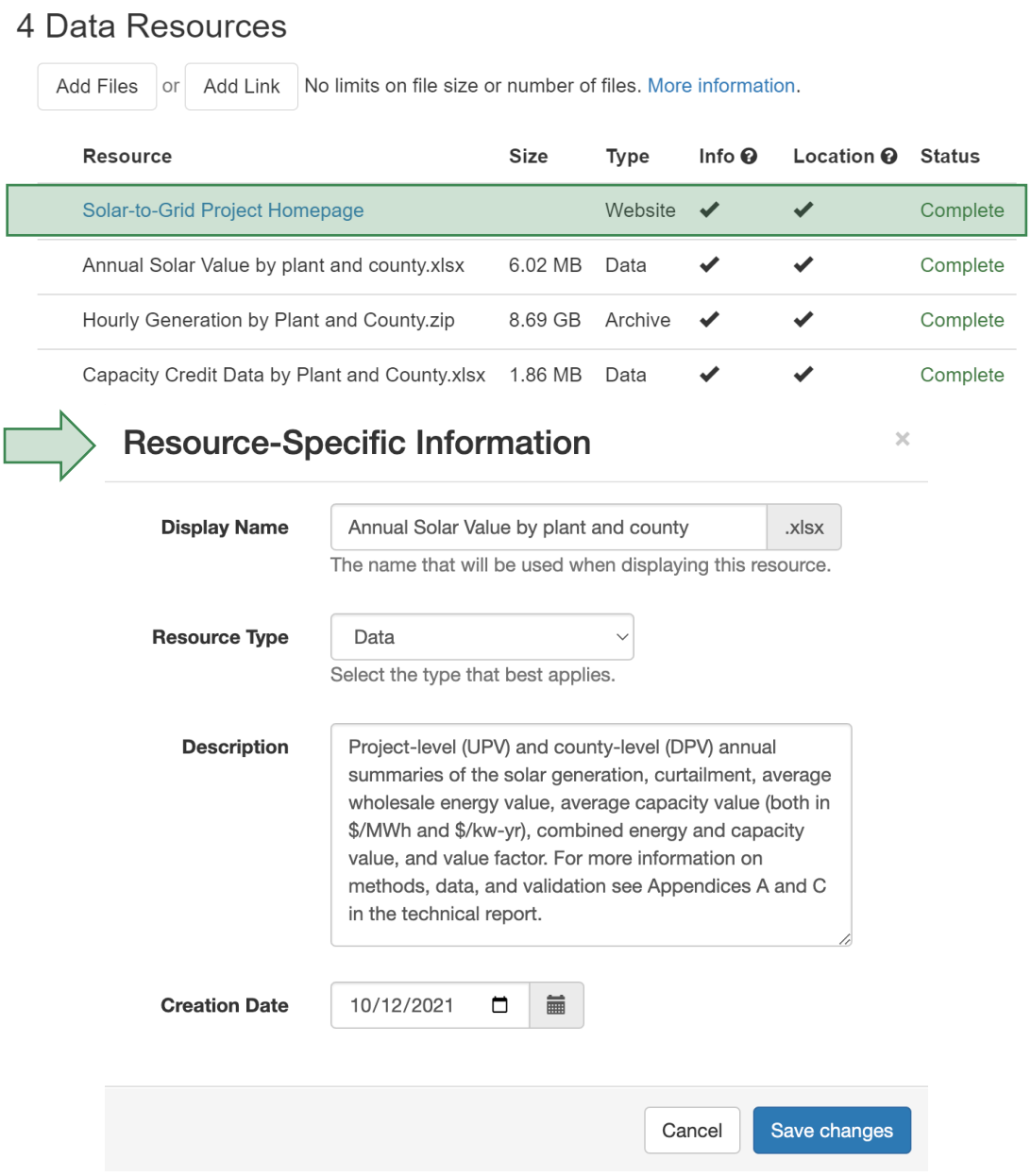

Figure 9: Example of data resources associated with an existing OEDI submission. Annual Solar Value by Plant and County.xlsx's resource-specific information is also shown.

Examples of good, permanent URLs:
http://goodsite.com/conference/paper-13.pdf
http://goodsite.com/the+title+of+the+paper.pdf

Examples of bad, temporary URLs:
http://badsite.com/search?conference=WorldScience&paper=13
http://badsite.com/node/13

Figure 10: Examples of good, permanent URLs and bad, temporary URLs.

Location

A location is required as a means of geotagging each dataset. The user may select a point or an area manually or may select from a list of locations. The pre-defined areas consist of "world," "North America," "contiguous US," "east coast," and "west coast". Figure 11 shows an example location associated with an existing OEDI submission. Notice that the pre-defined area for "contiguous united states" was used to create the polygon enclosing the location of interest.

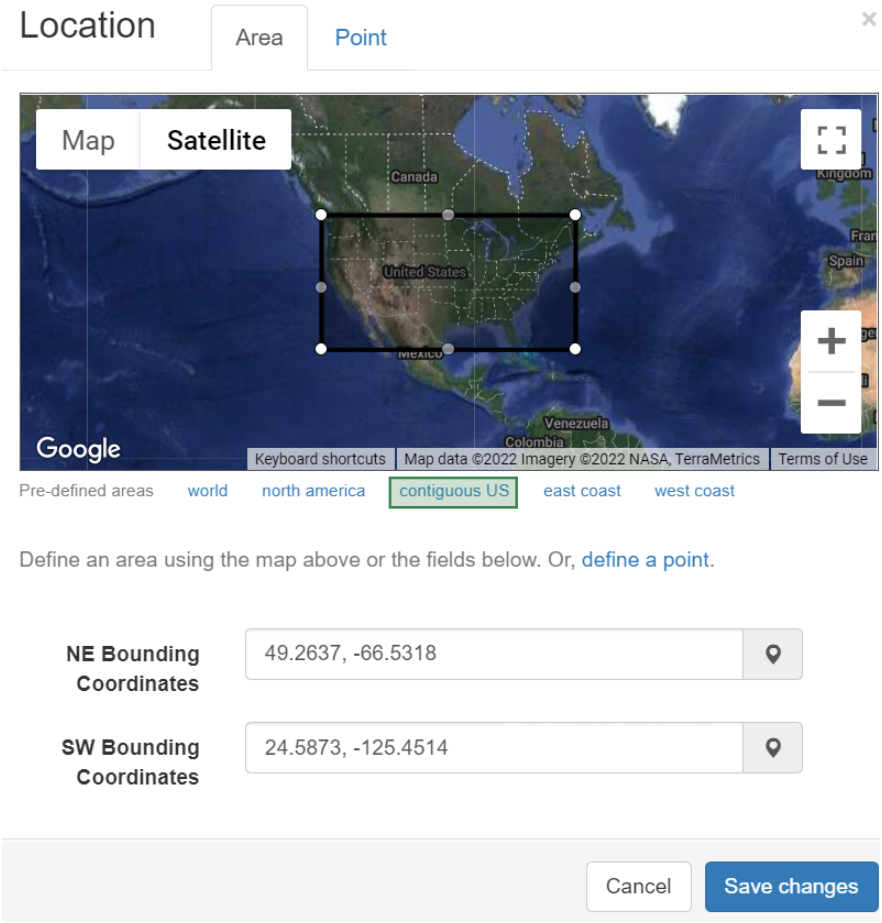

Figure 11: Example location associated with a resource in an existing OEDI submission. Note the pre-defined area for contiguous US was used to draw this polygon.

Version Control

Current practice for uploading new versions of files is to supplement the existing submission with an additional, updated resource, making sure the older version remains intact. This avoids getting rid of the older version, as it likely has a DOI and may be referenced elsewhere. It also ensures that if a user happens to discover the older version, that the updated version is prominently displayed and easily accessed.

Submission Statuses, Digital Object Identifiers, and Moratoriums

Submission Statuses

Each OEDI submission has an associated status. The statuses are intended to represent how far along a submission is in the process of publication. The statuses begin when a user starts and saves a new submission and end when the submission is publicly accessible. Figure 12 shows the progression of the statuses.

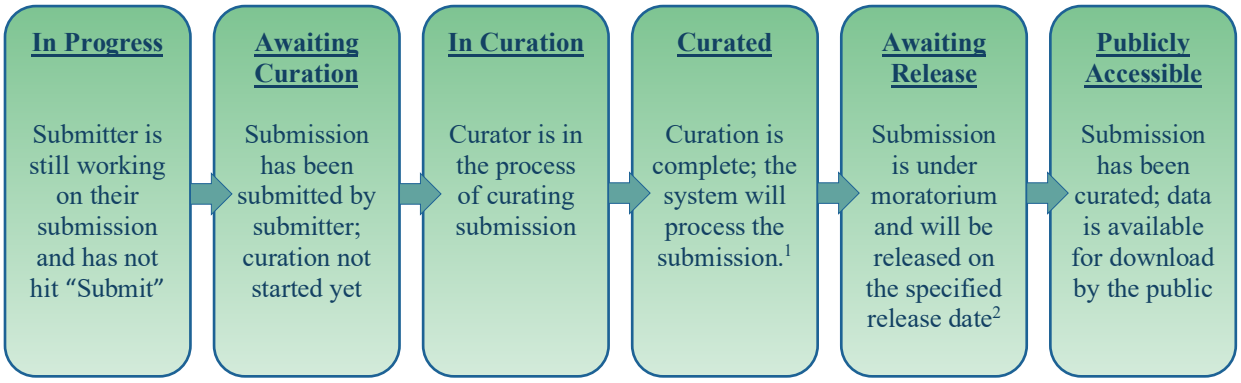

Figure 12: Flow of submission statuses with descriptions for each.

[1] This status only appears briefly while the system processes the action.

[2] This status only applies to submissions under moratorium. All other submissions will be published immediately.

Digital Object Identifiers

Digital Object Identifiers (DOIs) are unique resource identifiers (URI) with permanently resolvable links to the individual data resources and are added to submissions with resources that qualify as data by U.S. Department of Energy Office of Scientific and Technical Information's (OSTI's) standards. This includes submissions that contain resources labeled as any type other than 'Document' and 'Link.' In addition, only DOE-funded data submissions receive DOIs. See OSTI's Data ID Services for more information on what is considered data by OSTI.

When a DOI is added to a submission, the dataset is automatically added to OSTI's DOE Data Explorer and further propagated through the network of OEDI partners (Figure 3).

Moratoriums

If the data are required to be submitted to OEDI but are not ready to be made available to the public, the user may add a moratorium to the submission with a specified release date (Figure 13). When a moratorium is added to a submission, the submission metadata is made available to the public, but the data may not be downloaded by anyone beyond the OEDI curation team until the specified release date. This serves to protect the data until the specified date while making vital information about the data available to the scientific community. After all, there is an intrinsic value in knowing that data exist, even if they are not available yet. Contact information is made available for each dataset to allow interested parties to reach out to data owners in advance of highly anticipated data releases and inquire about the data or explore potential collaborations. Figure 14 shows an example of what is viewable to the public when a submission is under moratorium.

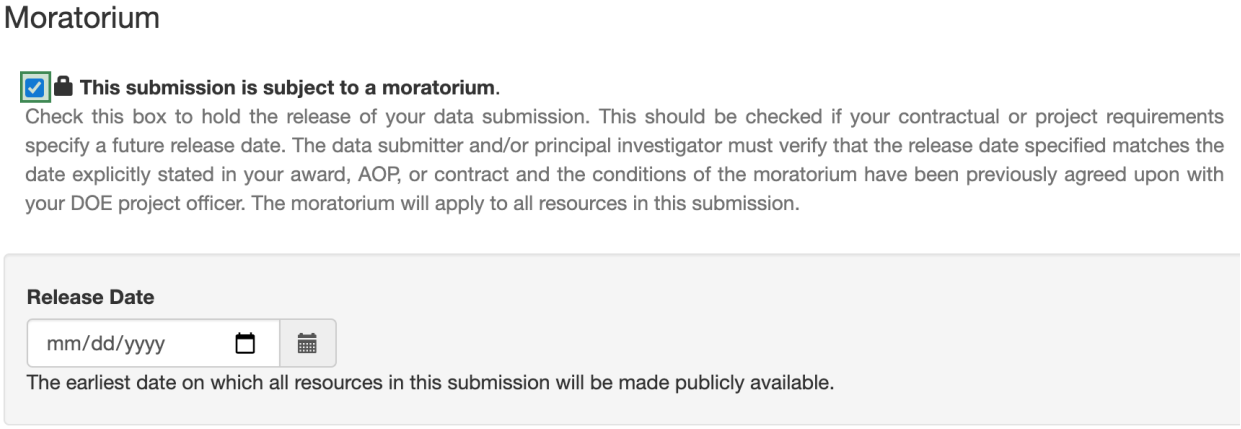

Figure 13: Screenshot showing moratorium section of OEDI submission form. Note that you must check the box next the lock for the release date portion to appear.

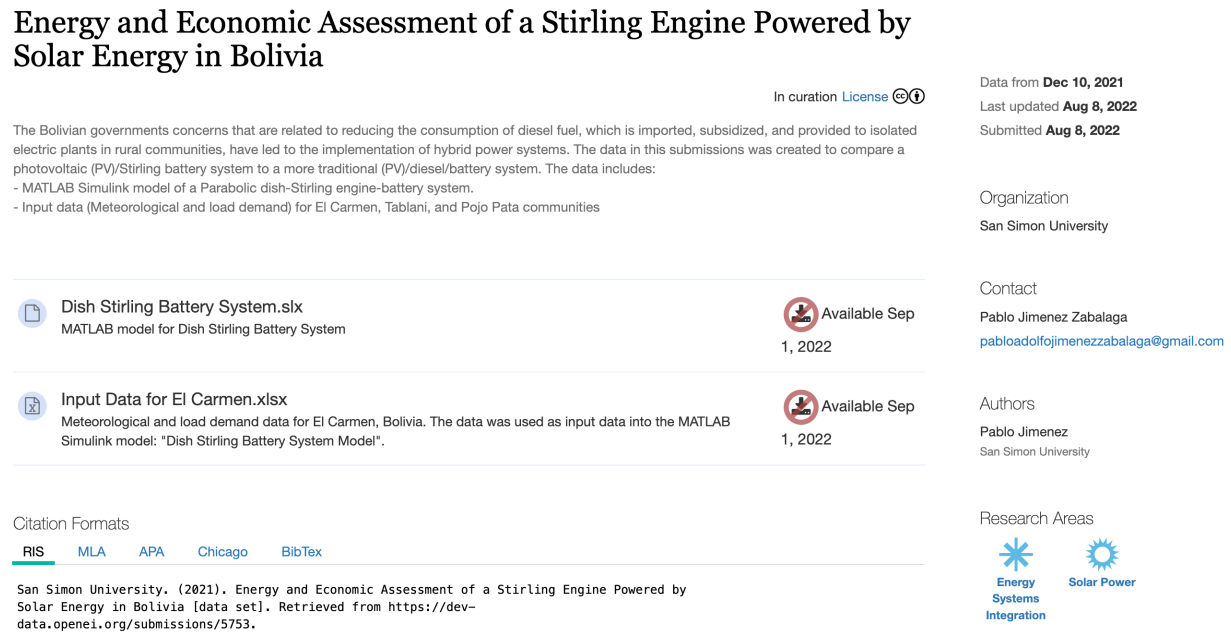

Figure 14: Screenshot depicting which parts of a OEDI submission are made available to the public when submission is under moratorium. Note that other information including contact information is also made available.

Additional Resources

For additional information, please see the OEDI's Frequently Asked Questions page, the OEDI Tutorial accessible from the Help drop-down on the data submission page, and the OEDI submission training videos.

Acknowledgements

This work was authored by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by U.S. Department of Energy Office of Energy Efficiency and Renewable Energy (EERE) Solar Energy Technology Office (SETO). The views expressed in the article do not necessarily represent the views of the DOE or the United States Government. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for United States Government purposes.

Appendix A: API Documentation

The metadata for all data stored in OEDI are available in JSON-LD through a *data.json* drop file located at: https://data.openei.org/data.json.

The JSON file contains the complete metadata records for all data submissions in the OEDI catalog and has been developed in cooperation with the U.S. Government's Project Open Data using the DCAT-US Schema, also known as the Project Open Data Metadata Schema, as a basis for the JSON-LD schema detailed below:

Catalog Fields

Field	Label	Definition	Required
@context	Metadata Context	URL or JSON object for the JSON-LD Context that defines the schema used.	No
@id	Metadata Catalog ID	IRI for the JSON-LD Node Identifier of the Catalog. This should be the URL of the data.json file itself.	No
@type	Metadata Type	IRI for the JSON-LD data type. This should be dcat:Catalog for the Catalog.	No
conformsTo	Schema Version	URI that identifies the version of the Project Open Data schema being used, i.e. "https://project-open-data.cio.gov/v1.1/schema"	Yes

describedBy	Data Dictionary	URL for the JSON Schema file that defines the schema used.	No
dataset	Dataset	A container for the array of Dataset objects. See Dataset Fields below for details.	Yes

Dataset Fields

Field	Label	Definition	Required
@type	Metadata Type	IRI for the JSON-LD data type. This should be dcat:Dataset for each Dataset.	No
title	Title	Human-readable name of the asset. Should be in plain English and include sufficient detail to facilitate search and discovery.	Yes
description	Description	Human-readable description (e.g., an abstract) with sufficient detail to enable a user to quickly understand whether the asset is of interest.	Yes
keyword	Tags	Tags (or keywords) help users discover your dataset; please include terms that would be used by technical and non-technical users.	Yes
modified	Last Update	Most recent date on which the dataset was changed, updated or modified.	Yes
publisher	Publisher	The publishing entity.	Yes
accessLevel	Public Access Level	The degree to which this dataset **could** be made publicly available, *regardless of whether it has been made available*. Choices: public (Data asset is or could be made publicly available to all without restrictions), restricted public (Data asset is available under certain use restrictions), or non-public (Data asset is not available to members of the public).	Yes
bureauCode	Bureau Code	Federal agencies, combined agency and bureau code from OMB Circular A-11, Appendix C (PDF, CSV in the format of "019:20").	Yes
sectors	Program Codes	Federal agencies, list the primary program related to this data asset, from the Federal Program Inventory. Use the format of "019:001".	Yes
license	License	The license or non-license (i.e. Public Domain) status with which the dataset or	If applicable

		API has been published. See Open Licenses for more information.	
distribution	Distribution	An array of Distribution objects. See Dataset Distribution Fields below for details.	If applicable
dataQuality	Data Quality	Whether the dataset meets the agency's Information Quality Guidelines (true/false).	No
issued	Release Date	Date of formal issuance.	No
landingPage	Homepage URL	This field is not intended for an agency's homepage (e.g. www.agency.gov), but rather if a dataset has a human-friendly hub or landing page that users can be directed to for all resources tied to the dataset.	No
DOI	DOI	Digital Object Identifier	If applicable
projectLead	Project Lead	The person at DOE directly reported to for this project.	If applicable
projectTitle	Project Title	The official title of this project, from the AOP or DOE award.	If applicable
projectNumber	Project Number	DOE Project Number, CPS Number, or AOP WBS number. Examples: *EE0012345*, *12345*, or *FY13 AOP 1.2.3.45*.	If applicable
fullName	Contact Name	Contact name, first and last, of the contact for this dataset.	Yes
email	Contact Email	Email address for contact.	Yes
phone	Contact Phone Number	Phone number for contact.	Yes
publisher	Publisher	Publishing organization	Yes
submitted	Submission Date	Date dataset was submitted	Yes
authors	Dataset Authors	Authors of the dataset. See Authors below for details.	Yes

Dataset Distribution Fields

Field	Label	Definition	Required
name	Title	Human-readable name of the distribution.	Yes
desc	Description	Human-readable description of the distribution.	Yes
size	Size	Size of resource if resourceType is file.	If applicable
resourceType	Resource Type	Either "file" or "link".	Yes

sampleDate	Sample Date	Sample or Creation date	Yes
URI	Download URL	URL providing direct access to a downloadable file of a dataset.	Yes
coordinates	Coordinates	Lat/lon for the distribution.	If applicable
extent	Bounding Coordinates	NE and SW coordinates of bounding box describing location of distribution.	If applicable
languages	Languages or Technologies	List of languages or technologies on which the dataset depends.	If applicable

Catalog Fields

Field	@context
Required	No
Accepted Values	String (URL)
Usage Notes	The URL or JSON object for the JSON-LD Context that defines the schema used. The URL for version 1.1 of the schema is https://project-open-data.cio.gov/v1.1/schema/catalog.jsonld
Example	{"@context": "https://project-open-data.cio.gov/v1.1/schema/catalog.jsonld"}

Field	@id
Required	No
Accepted Values	String (IRI)
Usage Notes	A unique identifier for the Catalog as defined by JSON-LD Node Identifiers. This should be the URL of the data.json file itself.
Example	{“@id”: https://data.nrel.gov/submissions/18}

Field	@type
Required	No
Accepted Values	String (IRI)
Usage Notes	The metadata type as defined by JSON-LD data types. This should be dcat:Catalog for the Catalog.
Example	{"@type": "dcat:Catalog"}

Field	conformsTo
Required	Yes
Accepted Values	String (URI)
Usage Notes	This is used to identify the schema version using a URI. The URI for version 1.1 of the schema is https://project-open-data.cio.gov/v1.1/schema.
Example	{"conformsTo": "https://project-open-data.cio.gov/v1.1/schema"}

Field	describedBy
Required	No
Accepted Values	String (URL)

Usage Notes	This is used to specify a JSON Schema file that defines all fields. By default, it is recommended that the canonical JSON Schema file is referenced (https://project-open-data.cio.gov/v1.1/schema/catalog.json) but if the schema had been extended, publishers may reference a file that defines those extensions.
Example	{"describedBy": "https://project-open-data.cio.gov/v1.1/schema/catalog.json"}

Field	dataset
Required	Yes
Accepted Values	Array of Objects
Usage Notes	This field is a container for an array of Dataset objects. See Dataset Fields below for details.
Example	{“dataset": [...]}

Dataset Fields

Field	@type
Required	No
Accepted Values	String (IRI)
Usage Notes	The metadata type as defined by JSON-LD data types. This should be dcat:Dataset for the Dataset
Example	{"@type": "dcat:Dataset"}

Field	title
Required	Yes
Accepted Values	String
Usage Notes	Acronyms should be avoided.
Example	{“title”:”Types of Vegetables”}

Field	description
Required	Yes
Accepted Values	String
Usage Notes	This should be human-readable and understandable to an average person.
Example	{"description":"This dataset contains a list of vegetables, including nutrition information and seasonality. Includes details on tomatoes, which are really fruit but considered a vegetable in this dataset."}

Field	keyword
Required	Yes
Accepted Values	Array of strings
Usage Notes	Surround each keyword with quotes. Separate keywords with commas. Avoid duplicate keywords in the same record.

Example	{"keyword":["vegetables","veggies","greens","leafy","spinach","kale", "nutrition"]}

Field	modified
Required	Yes
Accepted Values	ISO 8601 Date
Usage Notes	Dates should be ISO 8601 of highest resolution. In other words, as much of YYYY-MM-DDThh:mm:ssTZD as is relevant to this dataset.
Example	{"modified":"2021-01-15"} or {"modified":"2021-01-15T12:00:01Z"}

Field	publisher
Required	Yes
Accepted Values	String
Usage Notes	Publishing organization.
Example	{“publisher”: “United States Geological Survey”}

Field	accessLevel
Required	Yes
Accepted Values	Must be one of “public”, “restricted public”, or “non-public”.
Usage Notes	This field refers to the degree to which this dataset *could be made available* to the public, regardless of whether it is currently available to the public. For example, if a member of the public can walk into your agency and obtain a dataset, that entry is **public** even if there are no files online. A *restricted public* dataset is one only available under certain conditions or to certain audiences (such as researchers who sign a waiver). A *non-public* dataset is one that could never be made available to the public for privacy, security, or other reasons as determined by your agency.
Example	{"accessLevel":"public"}

Field	bureauCode
Required	Yes
Accepted Values	Array of Strings
Usage Notes	Represent each bureau responsible for the dataset according to the codes found in OMB Circular A-11, Appendix C (PDF, CSV). Start with the agency code, then a colon, then the bureau code.
Example	The Office of the Solicitor (86) at the Department of the Interior (010) would be: {"bureauCode":["010:86"]}. If a second bureau was also responsible, the format like this: {"bureauCode":["010:86","010:04"]}.

Field	sectors
Required	Yes
Accepted Values	Array of strings
Usage Notes	Provide an array of programs related to this data asset, from the Federal Program Inventory.

Example	{"programCode":["015:001"]} or if multiple programs, {"programCode":["015:001","015:002"]}

Field	license
Required	Yes, if applicable
Accepted Values	String (URL)
Usage Notes	See list of license-free declarations and licenses.
Example	{"license":"http://creativecommons.org/publicdomain/zero/1.0/"}

Field	distribution
Required	If applicable
Accepted Values	Array of objects
Usage Notes	This is a container for one or multiple distribution objects which group together the fields: name, desc, size, resourceType, sampleDate, URI, coordinates, extent, and languages.
Example	See below


```
"distribution": [
        {
          "name": "Building Characteristics for Residential Hourly Load Data.pdf",
          "size": "204463",
          "desc": "Documentation for the data index",
          "resourceType": "file",
          "sampleDate": "2014-10-14T06:00:00Z",
          "coordinates": [
            36.9255,
            -95.9916
          ],
          "extent": {
            "boundingCoordinatesNE": [
              49.2637,
              -66.5318
            ],
            "boundingCoordinatesSW": [
              24.5873,
              -125.4514
            ]
          },
          "URI":
"https://data.openei.org/files/153/buildingcharacteristicsforresidentialhourlyloaddata.pdf"
        },
        {
          "name": "TMY2 Residential Base.zip",
          "size": 111934369,
          "desc": "TMY2 Data",
          "resourceType": "file",
          "sampleDate": "2014-11-20T07:00:00Z",
          "coordinates": [
            36.9255,
            -95.9916
          ],
          "extent": {
            "boundingCoordinatesNE": [
              49.2637,
              -66.5318
            ],
            "boundingCoordinatesSW": [
              24.5873,
              -125.4514
            ]
          },
```

```
      "URI": "https://data.openei.org/files/153/EPLUS_TMY2_RESIDENTIAL_BASE.zip"
    }
  ]
```


Field	dataQuality
Required	No
Accepted Values	Boolean true or false
Usage Notes	Indicates whether a dataset conforms to the agency's information quality guidelines.
Example	{"dataQuality": true}

Field	issued
Required	No
Accepted Values	ISO 8601 Date
Usage Notes	Dates should be ISO 8601 of highest resolution. In other words, as much of YYYY-MM-DDThh:mm:ssTZD as is relevant to this dataset.
Example	{"issued":"2021-01-15"} or {"issued":"2021-01-15T12:00:01Z"}

Field	landingPage
Required	No
Accepted Values	String (URL)
Usage Notes	This field is not intended for an agency's homepage (e.g. www.agency.gov), but rather if a dataset has a human-friendly hub or landing page that users can be directed to for all resources tied to the dataset.
Example	{"landingPage": "https://data.openei.org/submissions/309"}

Field	DOI
Required	If applicable
Accepted Values	String
Usage Notes	Digital object identifier for the dataset, if one exists.
Example	{"DOI": "10.15121/1261909"}

Field	projectLead
Required	If applicable
Accepted Values	String
Usage Notes	Name of person at DOE directly reported to for this project.
Example	{"projectLead": "Mike Weathers"}

Field	projectTitle
Required	If applicable
Accepted Values	String
Usage Notes	The official title of this project, from the AOP or DOE award.
Example	{"projectTitle": "Wind and EERE-Solar 32307 and 1.2.5.401"}

Field	projectNumber
Required	If applicable
Accepted Values	String
Usage Notes	DOE Project Number, CPS Number, or AOP WBS number.
Example	{“projectNumber”: “EE0012345”} or {“projectNumber”: “FY21 AOP 1.2.3.45”}

Field	fullName
Required	Yes
Accepted Values	String
Usage Notes	Contact name, first and last, of the contact for this dataset.
Example	{“fullName”: "Ezra Zemach”}

Field	email
Required	Yes
Accepted Values	String (email address)
Usage Notes	Email address for this dataset’s contact.
Example	{“email”: “ted.jones@nrel.gov”}

Field	phone
Required	Yes
Accepted Values	String (phone number)
Usage Notes	Phone number for this dataset’s contact.
Example	{“phone”: “303-275-1234”}

Field	publisher
Required	Yes
Accepted Values	String
Usage Notes	Publishing organization
Example	{"publisher": "Davenport Newberry Holdings, LLC"}

Field	submitted
Required	Yes
Accepted Values	ISO 8601 Date
Usage Notes	Submission date. Dates should be ISO 8601 of highest resolution. In other words, as much of YYYY-MM-DDThh:mm:ssTZD as is relevant to this dataset.
Example	{"submitted":"2021-01-15"} or {"submitted":"2021-01-15T12:00:01Z"}

Field	authors
Required	Yes
Accepted Values	Array of objects
Usage Notes	Authors of the dataset

Example	See below.

```
"authors": [
        {
          "firstName": "John",
          "lastName": "Shervais",
          "affiliation": "Utah State University"
        },
        {
          "firstName": "James P.",
          "lastName": "Evans",
          "affiliation": "Utah State University"
        }
]
```

Dataset Distribution Fields

Field	name
Required	Yes
Accepted Values	String
Usage Notes	Human-readable name of the distribution, or resource.
Example	{"name": "Final Technical Report.pdf"}

Field	desc
Required	Yes
Accepted Values	String
Usage Notes	Human-readable description of the distribution, or resource.
Example	{"desc": "Data released under the Department of Energy's Open Energy Data Initiative (DOE).”}

Field	size
Required	Yes, if distribution resourceType is “file”.
Accepted Values	Integer
Usage Notes	Size of file resource in bytes.
Example	{"size": 300469}

Field	resourceType
Required	Yes
Accepted Values	String
Usage Notes	Either “file” or “link”.
Example	{“resourceType”: “file”} or {“resourceType”: “link”}

Field	sampleDate
Required	Yes
Accepted Values	ISO 8601 Date
Usage Notes	Sample or creation date. Dates should be ISO 8601 of highest resolution. In other words, as much of YYYY-MM-DDThh:mm:ssTZD as is relevant to this dataset.

Example	{"sampleDate":"2021-01-15"} or {"sampleDate":"2021-01-15T12:00:01Z"}

Field	URI
Required	Yes
Accepted Values	String (URL)
Usage Notes	URL providing direct access to a downloadable file of a dataset.
Example	{"URI": "https://registry.opendata.aws/oedi-data-lake/"}

Field	coordinates
Required	If applicable
Accepted Values	Array with two floats
Usage Notes	Lat/lon for the distribution.
Example	{"coordinates": [38.031672339667, -122.18298984375]}

Field	extent
Required	If applicable
Accepted Values	Object
Usage Notes	NE and SW coordinates of bounding box describing location of distribution.
Example	{"extent": {"boundingCoordinatesNE": [38.9203190,-120.7163], "boundingCoordinatesSW": [37.1430255,-123.64961]}}

Field	languages
Required	If applicable
Accepted Values	Array of strings
Usage Notes	List of languages or technologies on which the dataset depends.
Example	{“languages”: [“Jupyter Notebook”, “Python”]}

References

Cover Photo Image Source: https://www.wgu.edu/blog/neural-networks-deep-learning-explained2003.html

"Data ID Services." *OSTI.GOV*, www.osti.gov/data-services.

Brodt-Giles, Debbie and Michael Rossol. "Open Energy Data Initiative." US Department of Energy Office of Energy Efficiency and Renewable Energy, 2019. https://www.nrel.gov/docs/fy19osti/73632.pdf

Weers, Jon, and Jay Huggins. "Getting Data Out of the Ground: Modern Challenges Facing EGS Collab, the DOE Geothermal Data Repository, and the Geothermal Industry." PROCEEDINGS, 44th Workshop on Geothermal Reservoir Engineering. 2019, https://pangea.stanford.edu/ERE/pdf/IGAstandard/SGW/2019/Weers.pdf.

Weers, Jon, and Jay Huggins. "The Data Foundry: Secure Collaboration for the Geothermal Industry" PROCEEDINGS, 45th Workshop on Geothermal Reservoir Engineering. 2020, https://pangea.stanford.edu/ERE/db/GeoConf/papers/SGW/2020/Weers.pdf.