HydroShare Guide for Data Authors and Publishers

Amber Jones, Liza Brazil, September 2018

HydroShare is a collaborative environment for sharing and publishing data and models. Research data are interpreted broadly in a joint statement by The National Academies (2009):

> *It includes textual information, numeric information, instrumental readouts, equations, statistics, images (whether fixed or moving), diagrams, and audio recordings. It includes raw data, processed data, published data, and archived data. It includes the data generated by experiments, by models and simulations, and by observations of natural and social phenomena at specific times and locations. It includes data gathered specifically for research as well as information gathered for other purposes that is then used in research.*

The American Geophysical Union publications data policy (Adopted in 2013 and updated in 2016) data definition includes new protocols or methods used to generate the data and new code/computer software used to generate results or analyses. Thus, data and model sharing is increasingly mandated by funding entities and academic publishers and is recognized as a best practice for scientific research. Data and model sharing permits the reproducibility of science, including data evaluation, reuse, integration and comparative analysis. To facilitate science, data should be managed following established standards ensuring that data are made Findable, Accessible, Interoperable, and Reproducible (FAIR). This document is a guide to best practices for data creators and publishers using HydroShare to share their data and models following FAIR principles. Much of the content here was adapted from best practices for data publication by DataOne - refer to the DataOne best practices database for detailed guidelines. Additional, specific technical guidance is provided as part of the HydroShare help pages.

SHARING AND PUBLICATION

HydroShare users should treat datasets and models as first class research products - along with the papers that are written about them. There is opportunity to maximize the value of these research products by practicing good data management and creating descriptive metadata. HydroShare permits several levels for sharing and publication. A resource, which is the primary unit of digital content in HydroShare and may include data and models, may be made public at any point to provide discovery and access. Permanently publishing a resource should only be undertaken when authors are certain that it will not change.

PLANNING

Researchers should plan for sharing research products as part of research conceptualization. Planning helps address challenges and questions related to timing, organization, and authorship. Planning should include consideration of which products will be generated and

shared, whether there are sensitivities around any of the products, the number and type of files that will be generated, which variables the files will include, the organization and naming of files, and the person(s) responsible for content and metadata creation and curation.

TIMELINE FOR DATA SHARING AND PUBLICATION

There are a number of paths or data workflows that users might take to share and publish their research products using HydroShare. Planning for sharing should include timing considerations. For resources (i.e., data and/or models) closely tied to a research publication, we recommend sharing the resource publically when the paper is submitted, if not before. Doing so permits the resource to be cited in the paper, verified and examined by reviewers, and edited based on reviewer comments. When the paper is accepted for publication, the resource can then be permanently published as an immutable resource with a citable Digital Object Identifier (DOI). Below are steps to accomplish this. Order may change slightly based on the journal.

1. Create the resource in HydroShare, upload content files, and ensure that metadata are complete.
2. Make the resource public.
3. Cite the HydroShare resource in the references section of the paper using the HydroShare URL (e.g., Tarboton, D. (2017). Logan Specific Catchment Area, HydroShare, https://www.hydroshare.org/resource/877bf9ed9e66468cadddb229838a9ced/). This will ensure that reviewers can access the resource.
4. Submit the paper for review.
5. Based on reviewer suggestions and paper revisions, make any changes that are necessary to the HydroShare resource.
6. When the paper is accepted, in the final text submitted for typesetting cite the HydroShare resource using its DOI URL reference (e.g., Tarboton, D. (2017). Logan Specific Catchment Area, HydroShare, https://doi.org/10.4211/hs.877bf9ed9e66468cadddb229838a9ced). HydroShare DOIs use the pattern "https://doi.org/10.4211/hs." followed by the resource's unique HydroShare identifier. This anticipated DOI is also indicated on the resource landing page below the "How to cite" section.
7. When the publisher issues a DOI for the paper, add the full reference for the paper to the related resources metadata on the HydroShare resource. This ensures that the HydroShare resource has a link that points at the final paper. Then, finalize and permanently publish the HydroShare resource.

If you follow this process, you will have a citation to the paper in your HydroShare resource that uses a DOI and a citation to the HydroShare resource in the paper that also uses a DOI. Here's an example of a resource and associated paper for which this was done.

LINKING TO PUBLISHED PAPERS

Data publication should be considered a separate, though parallel, process from the publication of research results in papers, theses, etc., and datasets should stand alone without these

supplements. To the extent possible, the metadata of the resource should fully describe the content without needing to rely on references to a published paper. That being said, related papers might help explain the linkage between resources, innovative methods, and interpretation of results. In most cases, brief metadata descriptions can still provide the relevant details for a user to interpret the data/model without relying on a separate paper. Links and citations to related publications should be included as a Relation using the Related Resources HydroShare metadata.

RAW, INTERMEDIATE, AND FINALIZED DATA

Researchers need to consider the value of various potential data products. Determining the data resources to be shared early in the research process helps clarify significant data products. Researchers should consider how to make their research reproducible and data reusable for other purposes. In general, the preservation of raw data is essential to data reuse. If the data undergo multiple steps for analysis and processing, consider the value of including the results of intermediate steps; it may be better to include the raw data, finalized data, and code or algorithms with detailed description of the process rather than sharing each intermediate product. If raw data are large or complex, consider how they can best be organized for inclusion.

A published resource should be a finalized, clean, complete version of the data or model. There should not be errors, extraneous notes, highlighting, worksheets, or other components that are not pertinent to the data/modeling product. If they are relevant to the dataset and are included, they must be defined and described. It is recommended that another author or colleague proofread the resource for interpretability as well as spelling/grammar. Even if a dataset is not ready for permanent publication with a DOI, when it is made public, the data and metadata should be clear and free of errors as other HydroShare users can see and access datasets as soon as they are made public.

RESOURCE ORGANIZATION

For data to be unambiguous and interpreted properly, data publishers must consider whether another researcher can reasonably understand and replicate their work. Reusability and interpretability are objectives in determining how to structure and organize the content files within the resource as well as in determining what metadata and descriptions need to be included. More specific instructions are included throughout this document.

In HydroShare, a “resource” is the discrete unit of digital content. Persistent identifiers, access control, versioning, sharing, and discovery are all managed at the resource level. For more details see Horsburgh et al. (2016). Resources may contain one or many content files, and HydroShare permits the organization of those content files into nested folders. Users may also organize multiple resources into collections. Resources may also be versioned through the creation of a new resource that supersedes the old resource. Combined, this functionality supports considerable flexibility in the way content may be organized in HydroShare.

GROUPING/CHUNKING DATA

Most data and modeling efforts will generate multiple files and often have complexity that requires forethought toward resource and file organization. There is no single prescriptive practice for how resources and files should be grouped/separated. The content authors are the most familiar with the results and how they may be used. Consider how the content files might be used and how potential end users will want to interact with the resources and associated files. What structure will best facilitate access, interpretation, and reuse? Should the results be organized into multiple resources? If so, how many? Will they be organized by site, by variable, or by type of analysis?

There is not necessarily a correct answer for how to group/split a resource. It may help to consult with colleagues about organization to determine a plan. If there are commonalities and duplication between multiple resources, it might make sense to group the content more tightly to avoid this duplication. For a single research effort, 100 resources are likely too many, especially if the files in each resource are of the same type/format. How could these files be aggregated to be digestible by a user? Likewise, a single resource for a complex monitoring effort that included multiple components with many levels of folder organization should probably be separated into multiple resources - perhaps based on types of data that were collected/analyzed. Note that HydroShare's composite resource type permits the inclusion of metadata specific to each file in the resource so that different file types with distinct metadata can be included within the same resource. Relations can be used to provide linkages between resources, and multiple HydroShare resources can be associated with each other in a collection.

For example, research results from a broad effort were originally grouped as a single resource, but it was determined that three separate resources (1, 2, 3) were more appropriate given the complexity of the data and metadata required to adequately describe the contents.

This is an example of a dataset that was originally separated using a single resource for each file. It was determined that there was enough similarity and association between the files that they should be grouped into fewer resources organized by sampling site.

ONGOING DATA COLLECTION

For data that involve ongoing collection, resources should be created with full metadata and deliberate organization so that data may be appended either as additional files within the resource or as additional content in existing files. The Abstract should include details on the anticipated growth of the dataset.

RESOURCES WITH MULTIPLE FILES

Many research results are rich datasets, models, or model instances comprised of several data files. As such, they may not be straightforward to follow for someone unfamiliar with the work. Best practice is to describe the overall organization of the content, what information is contained in each file, and how the files relate to each other. This can be done in the abstract or in a README file that provides an index to the content files to describe what the files consist of and what variables they contain.

RESOURCES WITH MULTIPLE FILES AND ASSOCIATED FILE METADATA

HydroShare's composite resource type permits the inclusion of metadata specific to each file in the resource, so that different file types with distinct metadata can be included within the same resource. In HydroShare, the file metadata section provides additional metadata fields depending on the file type. Users are encouraged to complete the file-level metadata if any file(s) contains a specific set of metadata that is not logical to display in the resource-level metadata. For instance, for a composite resource that contains a shapefile of the boundary of a watershed and an Excel file with a water-level time series, it is encouraged to provide file metadata for each file in the file metadata tab and resource-level metadata describing the larger project or dataset as a whole. HydroShare provides automatic extraction of metadata for some supported file types, such as shapefiles and raster datasets. Be sure to review metadata automatically extracted from known file types for correctness, and manually provide metadata for other file types.

METHODS

The methods and techniques used to generate results should be defined for each variable. This emcompasses descriptions of the methods used to collect, process, and analyze samples, including the instrument used as well as calculations performed. These descriptions could be included in the abstract or in a separate methods document. Methods may also involve formulas, algorithms, and other analytical steps. We recommend specifying analytical steps/formulas in a separate methods document and including tables of measured data and calculated data. Methods used to perform quality control on data should be included as well. If duplicate or triplicate samples were collected and/or analyzed, define how they are represented or how the derivative data were determined.

FILE FORMATS

Best practice is to use file formats that are open and documented standards with widespread use in the research community and with multiple software options for opening and viewing the files as opposed to proprietary, closed formats that can only be accessed using specific software. It is acknowledged that there are cases where files that use proprietary software are used in research processes (e.g., Matlab code used for analyses or files associated with a closed model). HydroShare will permit sharing of these types of files. However, in some cases, even though the files were created by proprietary software, they may still be exported to a format that can be accessed using open software. As a best practice, researchers should carefully consider how to provide the best access to results and files.

TABULAR DATA

It is best practice to report tabular results in a csv (comma separated values) or similar flat file type rather than Excel or another proprietary format. Each column in a table should have a detailed description either in the Abstract, a separate README file, or in a descriptive header within the file that contains the table. This description should include the parameter or variable that the column in the table represents and address the questions: What does it represent? What are the units? How was it obtained?

Excel: Some researchers conduct much of their work in Excel with mixed data types, multiple sheets, numerous tables, and embedded formulas. Within the same file, it is common to have raw data, finished products, and the intermediate steps between the two, which may include multiple sheets and multiple blocks of data within each sheet. The workflow may be difficult to follow for someone unfamiliar with the exact process, and best practices recommend refining data to be accessible, interoperable, and reusable. If Excel must be used given specialized formatting, formulas, or workflows, the guidelines below should be followed:

1. Excel formulas: In general, workflows using spreadsheet formulas are fragile as they could be easily changed or lost. The formulas may also not be clear to all potential users. Best practice is to remove them and document the steps and equations used to move from one step to another. If formulas are left in a spreadsheet, they should be defined in file/data explanations.
2. Excel formatting: Any formatting needs to be defined. If there are colors or other formats that are meaningful, they need to be explained. Do not include embedded comments as they are often overlooked and can easily be lost.
3. Ensure that all cells have data types as expected. For example, Excel formulas create a "#VALUE!" result when there is any sort of a problem. It is best practice to not mix data types within a single tabular column (e.g., do not store text and numbers in the same column).
4. Ensure that all sheets/tabs are defined and described.
5. Do not use hidden columns.

This example is a dataset with results in Excel files with complex formatting. The resource includes a metadata document to provide descriptions for all files, sheets, and tables contained therein.

This example's results were determined by a complex Excel workflow. The resource includes a data dictionary and README file to explain the workflow and each component of the file.

MODELING

Modeling resources are diverse, and HydroShare has the flexibility to handle many types of modeling data - input, model components, and results. Users must determine the versions of output to share, which depends on model scenarios, objectives, and calibrations. In general, input files and calibration parameters should be included so that subsequent users may replicate the analysis. The metadata for modeling resources should address the following points: What scenarios were included and what do they represent? How were parameters determined? How was calibration undertaken/determined? One method for organizing modeling resources is to include:

1. One model program resource to describe the model itself with version information. This may include uploading source code or compiled binaries or linking to the source if it is already hosted/published elsewhere. If the model is proprietary, this may not be possible.

2. One or more model instance resources that includes the input data and the output data that is linked to the model program resource. A user can choose to create one model instance resource per scenario or put all of the scenarios into one model instance resource. The resource-specific metadata for this should be relatively straightforward. For more details see this publication.

See here for an example of a modeling resource with a well organized file structure and metadata and read me files to describe the workflow and file contents.

SECONDARY ANALYSIS AND SYNTHESIS

For secondary analyses that use existing data, the resource should include:

1. Details on the provenance of the data used. If data are proprietary, include a description of how to obtain the data. Note that files for original articles or materials that are authoritatively obtained elsewhere should not be included, but including a bibliographic reference list with full citation information is appropriate. This can be done through related resources metadata or in a separate content file.
2. The value-added data product.
3. A workflow or description of the processes or methods used to achieve the result derived from the analyses performed.

This example presents a secondary analysis of news media articles, with links to the articles analyzed included in spreadsheet files along with descriptions of procedures and the value-added product.

SOCIAL SCIENCE AND SENSITIVE DATA

There are sensitivities around some classes of social science datasets that limit the degree to which they can be shared, while other types of social science datasets do not contain sensitive information and can be shared freely. In many cases, dissemination is restricted because the data may contain personal information or are controlled by an Institutional Review Board (IRB) or other authoritative entity. In these cases, data sharing can and should still be undertaken, but may require additional steps to anonymize, aggregate, or summarize sensitive data. Advance planning for data sharing, in particular during the process of drafting an IRB protocol that details how data will be collected and shared, will facilitate this process. See this paper for additional information.

This example includes interview and focus group transcripts that have been carefully anonymized. This example consists of a summary report of interviews for which the transcripts could not effectively be anonymized (and, thus, could not be shared). This example reports the results of a broad household survey for which some geographic aggregations had to be made to preserve anonymity.

CODE AND SCRIPTS

Some efforts generate code or scripts as an end product or use code as an intermediate step in data processing. For HydroShare resources that include code or scripts, best practice is to include:

1. General descriptions of the code's operations and functions in a separate README file or in the metadata for the resource.
2. Input files needed to run the code.
3. Output files (if appropriate).

Inputs and outputs should be described, including formatting. More detailed documentation of the code should be included as comments within the files or within a separate README file.

This example includes separate folders within a single HydroShare resource for code and data, with a README file that describes the structure of each, along with instructions on how to execute the code on the data.

SPATIAL DATA

If your data describe anything with a spatial component, best practice is to include spatial metadata most appropriate for your data. This may consist of coordinates of sites, a geospatial file (e.g., a shapefile), and/or a map or schematic. Include the spatial metadata in only a single location within the resource, and use reference codes (e.g., site codes) to avoid repeating the information throughout the resource. If the spatial data are a geographic feature or raster, they can be designated as such using the file-specific metadata in a composite resource.

METADATA BEST PRACTICES

The following are best practices associated with individual metadata elements within a HydroShare resource:

Abstract: The Abstract should put the content of the resource in context, including the rationale for the data collection or modeling and should be a specific description of the content files included in the resource. Ensure that the results described in the Abstract correspond to the files contained in the resource. The Abstract should include a description of the resource content and organization to help a potential user navigate the various files included. Where the content and organization of a resource is complicated, the Abstract may reference a README file included as a content file in the resource that provides a more detailed description of the resource's content files. Additional details may be added to the Abstract to describe the methods of data creation - i.e., what you did, why you did it, how you did it, when and where you did it, etc. If text is used from the Abstract of a journal article related to the resource, add additional details focused on describing the actual content of the resource.

Keywords: Keywords facilitate data discovery and should be selected to be descriptive and thorough for the associated data. Consider what the data represent, the variables included in the dataset, how the data were generated, and the geographic area that they represent. Unique acronym-like keywords may also be used similar to the way hashtags are used in some social media. For example, the term "iUTAH" is unique to a specific Utah research project named

“iUTAH.” Use of this as a keyword facilitates discovery of resources associated with the iUTAH project.

Files: File names should be descriptive and meaningful. Where many files are included in a resource, a README file that describes the content and/or organization of each file or the resource as a whole can be very useful in helping others understand the content of the resource.

Spatial and Temporal Coverage: These are optional metadata and should be used if the data or model in the resource has a geospatial “footprint” or location or a temporal component/time window. Including appropriate spatial and temporal coverage metadata can help others discover your data if they are searching within a specific geographic location or over a specific time period.

Other Details: All abbreviations need to be defined somewhere within the resource. Even seemingly obvious cases should be spelled out (e.g., C=Carbon, N=Nitrogen). Unit symbols and abbreviations should also be defined (e.g., mg/L=milligrams per liter). Definitions should be handled wherever sub-file metadata is contained (e.g, in the resource Abstract, in a README file, or in file headers).