Data Deposit Guide for Iowa Research Online (IRO)

A. Pre-deposit considerations

1. **When possible, convert data to open, non-proprietary formats, which facilitate long-term access.** For instance, Excel files should be converted to .csv or .tsv format. More information about open formats: https://www.lib.uiowa.edu/data/manage/file-formats/ Questions? Contact us: lib-data@uiowa.edu or brian-westra@uiowa.edu
2. **For tabular data, review the data structure and revise as needed to facilitate reuse.** These steps make it easier to convert from Excel to comma separated values (.csv) or tab-separated values (.tsv), which enables others to more easily open, view and understand your data, and for software to read/use the data.

   For instance:

   - don't use colored text or backgrounds (they will be lost in conversion to .csv files)
   - use a single column header row;
   - avoid spaces, special characters, and punctuation in headings/variable names;
   - in general, use columns for variables, rows for observations;
   - exercise caution if you have date fields in an Excel file, and check them after conversion to .csv;
   - do not embed formulas in cells (they will not convert to .csv);
   - use a separate Excel tab for each table – don't combine tables into one large table or tab;
   - Save the .csv or .tsv with UTF-8 encoding.

   See these short guides for more information about best practices for tabular data:

     - https://www.tandfonline.com/doi/full/10.1080/00031305.2017.1375989
     - https://www.unmc.edu/publichealth/centers/ccorda/exceldata.html
     - more here: https://www.lib.uiowa.edu/data/manage/data-structure/

3. **If there are multiple files**, name and organize them to help the reader; list them in the order you want them displayed in the repository (put the readme and data dictionary first), and include that list in the readme.txt (see **E.1** below)
4. **Create a data dictionary for each tabular dataset, or codebook for surveys** (see **E.2** below)
5. **If article/peer reviewers need to have temporary access to the data before it is publicly accessible**, we can provide a temporary link to the dataset from IRO that for peer review access.

B. Obtain a DOI from us so you can include a citation for the data in your article submission

We can reserve a DOI for you before your data is published, so that you can include a reference/citation for the data set(s) in the manuscript, even before the data is made openly accessible*.

- See **C** and **D** below for adding a reference to your data in your article and creating a data availability statement to include in your article.

*The DOI will not work until it is activated. Once it is activated, the data is published, and this is not reversible.

Required information: To reserve a DOI for your data, we need the following:

1. **Title of the Data Set**: Not the same as the title of the article.
   Example: *PCB Emissions data from Paint Colorants.*

Title for the data/dataset:

2. **List all Creators (aka authors) of the dataset, in the order you want them to be listed in the data record and citation for the data.** Please provide first name (and middle initial if so desired), and last name.

   For UI researchers, please make sure you have an ORCID and that it is connected with the UI:

   - If you don't have an ORCID yet, start here: https://iam.uiowa.edu/planter/
   - If you have an ORCID, connect it to the University of Iowa: https://guides.lib.uiowa.edu/ORCID

   If there are Creators/Authors from other institutions, please provide their ORCIDs if you can

   Use this table to list the Creators, in the order that you want them to appear in the citation for the dataset:

Creator Name	**ORCID if known, for non-UI researchers, and UI grad students.**	**Institution (for non-UI researchers)**

3. **Publication Year**: we will use the current year, unless you tell us that the dataset has previously been made available elsewhere.

4. **Funder(s) and funder-supplied grant number(s)**

Funder	**Grant number (see the guidelines for NIH grant numbers)**

5. Indicate which license you would like to use for the data.

If you have multiple types of materials (i.e., data, code/software, instructional materials), you may need to select licenses appropriate for each type of material

For data: the links below will take you to descriptions of each license option. Please feel free to contact us if you aren't sure which license to choose.

- Open Data Commons Public Domain Dedication and License v1.0
- Open Data Commons Attribution License v1.0
- Open Data Commons Open Database License v1.0

Your license choice for data:

For software, use these guides to select an appropriate license:

- https://opensource.org/licenses
- https://spdx.org/licenses/

Your license choice for code/software, if applicable:

For things like instructional videos and other copyrightable works, select a Creative Commons license

- CC license chooser: https://creativecommons.org/choose/

Your license choice for instructional materials, videos, etc., if applicable:

C. Cite the data in the text of your article manuscript/thesis and in the references section

This will ensure that the citation to the data is displayed in both the pdf and online versions of your article, helping readers of online or print versions of the article to find the dataset. It also enables systems to generate citation statistics for your data, just as you can find for article citations.

Once you have the DOI from us, the citation would include the following elements, most of which are described above in **B**, above.

The exact order and punctuation will depend on the citation style of the journal in which you are publishing

Creator (PublicationYear). Title. Publisher. (resourcetype). Identifier

- **Publisher** is the University of Iowa, since it is being published in the UI's repository.
- **resourcetype** is usually 'dataset' but might also be 'collection,' 'model,' 'software,' etc., depending on the nature of the material being deposited
- **Identifier** is https://doi.org/ + DOI

Example:

> Jahnke, Jacob C. and Hornbuckle, Keri C. (2019): Dataset for PCB Emissions from Paint Colorants. University of Iowa. (dataset). https://doi.org/10.25820/vtd8-n771

D. Include a Data Availability Statement in your article manuscript

Note: This is in addition to citing the dataset as described in **C** above.

In the text of the document, include a statement describing how the data underlying the findings of your article can be accessed and reused. This should include a footnote or endnote to the citation for the dataset, in the article's references section.**

- See, for example: http://www.copdess.org/enabling-fair-data-project/enabling-fair-data-faqs/#3_Data_Availability_Statement_and_Data_Citation
- Your journal or publisher might include guidelines for the data availability statement. If they do not, others, such as Taylor & Francis, provide examples of data availability statements.

**Some publishers may also have an author submission process that includes a form or fields for a data availability statement. This should be filled out as well, but often this information is only displayed with the online version of the article and not in a pdf version of the article.

E. Before publication of the dataset, please provide the following details about the dataset:

The following information should be provided before the DOI is activated (we usually activate the DOI (making the data accessible) at the time when the article is published. Once the DOI is activated it can't be de-activated.

This information will help others find, understand, and reuse the data.

1. Create a Readme.txt file

The readme file provides context about the data, explains the methods, and is indexed by search engines for use in web searches. So the more detail, the better.

Cornell University's Research Data Management Service Group offers a great outline. For more information, see: https://www.lib.uiowa.edu/data/manage/documenting/readme/#readme

- Save the txt file with UTF-8 encoding.

Example from IRO:

- See the readme file in this record for an example: https://doi.org/10.25820/data.006135

2. For tabular data, create a Data Dictionary as a separate file (often as a .csv file, but a .txt file may also work) for each data file or set of data files.

A data dictionary is critical to making your research more reproducible because it allows others to understand your data. The purpose of a data dictionary is to explain what all the variable names and values in your spreadsheet really mean.[1]

- Variable names
- Readable variable name (may include a definition/description of the variable)
- Measurement units
- Allowed values, or range of values, if applicable
- Are null values allowed for the variable?
- Other codes for the variable (e.g., for missing data, data below limit of detection/quantitation)
- What data type is the variable (text, string, number, ISO 8601 date, etc.)
- Synonyms for the variable name (optional)
- Other resources

See examples from: Open Science Framework, the Smithsonian, and this USDA blank template.

3. For surveys, create a Codebook

Codebooks are a type of data dictionary that are more appropriate for survey and interview data.
See: https://www.lib.uiowa.edu/data/manage/documenting/readme/#codebooks

4. Create an abstract and a methods statement (if you did not already include these in the readme.txt file):

The abstract should be a brief description of the data and the context in which the data was collected or created.

[1] https://help.osf.io/hc/en-us/articles/360019739054-How-to-Make-a-Data-Dictionary

Focus on the data, rather than reusing the abstract from the related article. Abstracts help make your data more discoverable, and they provide context and information about the dataset for the researcher who finds your data.

An abstract might also be included in the readme.txt file; in fact, we recommend having these texts in both places. If it is in your readme file, we will use that for the Abstract when we create the data record.

Abstract, if it is not in the readme file:

Methods can also provide important information to researchers and others who may find and view your data. Here too, try to focus on the methods that are relevant to the data. This should describe the methodology employed for the study or research.

Methods, if they are not in the readme file:

5. Subjects (keywords, descriptors)

Please provide a list of descriptors or keywords describing what the data is about, etc. If there are terms from controlled vocabularies or other sources, please indicate (i.e.., gene names, species, chemical names)

Provide a list of subject(s) (or keywords, descriptors) here. To separate keywords, put a comma between each word or phrase.

Whenever possible, use terminology from established thesauri, taxonomies, etc. and make note of where the term is from.

For example:

For PCB 114, from https://pubchem.ncbi.nlm.nih.gov/compound/53036	
PubChem Compound CID:	53036
InChIKey:	SXZSFWHOSHAKMN-UHFFFAOYSA-N

OR

Rheumatic Fever	From Disease Ontology: https://disease-ontology.org/
DOID:	1586

6. Are any related works already published?

If other materials (software, data, articles) are about to be published or have already been published that are related to the data, provide the full citation (including DOI, ISBN, etc.), so we can add that to the record and link the two together.

For instance, if:

- the dataset is a subset of a published dataset, or incorporates data from other sources,
- code or software is published elsewhere and associated with the dataset
- another article has been published on this dataset

provide the citation(s), including DOIs, for those sources:

Related work (citation, with DOI)	Type of relationship: source of data; publication about the data; subset of the data, etc.)

F. Contact us when the article is published

Send us the DOI for the article (or other related materials) and we will update the data record so that it has a link to the article, and we will activate the dataset DOI, making it publicly accessible.

This will enable links in both directions between article and dataset.