Guidelines for Data Publishing v1.0

IEP Data Publication Workgroup

General Recommendations

- Publish in non-proprietary data formats (e.g. csv, txt, NOT .docx, .xlsx).
- Use non-proprietary software to store your data, or have a copy stored on non-proprietary software, in the case that your software goes obsolete.
- Use R or other repeatable method to prepare your data for publishing (QA/QC, data manipulation).
- Use descriptive file and variable names, without spaces (snake case or underscore)
- Each column should include only one data type (not multiple units).
- For IEP, use EDI to publish data if possible, unless your agency has a requirement to use a different data repository. For instructions on using EDI and CNRA, see The IEP GitHUB page.
    - Note: USGS publishes to Science Base instead.
- For ongoing datasets, recommend updating data annually.

Data Table Format

- To minimize the repetition of data, we recommend using the tidy format. This means organizing your data into separate tables for each scale of your data collection, and connecting tables with an identifying key. This is also known as the relational data model.
    - Benefits include reducing confusion and error for replicated data, reducing file size and processing time from duplicated data, and reducing wasted space (e.g. NAs for data types only relevant in certain cases).
    - See Borer et al. 2009 and scroll to number 10 for more benefits associated with a relational data model.
    - See example of recommended table structure (Data Publication Table Structure)
- We do recognize that the relational data model can make it difficult for data users without knowledge of R to access the entire dataset. Thus, if possible, we recommend including a joined dataset (flat file) along with a diagram of your relational tables and clear documentation on how the tables have been joined and how to interpret the data.

Database table relationships

- Key creation
    - Each table will need a key to join with the next level table.
    - An informative key is preferred over an auto-generated number (e.g. including date, station information in key) – see examples below, but up to database manager:

        - VisitID: 20200101_LIS
        - EventID: 20200101_LIS_1030 or 20200101_LIS_Tow1
        - SampleID: 20200101_LIS_1030_001 or 20200101_LIS_1030_LMB
        - IndividualID: 2020CHNS_001
    - Metadata should clarify what each row of the table represents and what each key represents
    - Each primary key should be unique (not repeated across rows)
- Provide code and description for data users to combine tables.
- Include a diagram of your database structure (see example).

Other recommendations

- Keep taxa caught in the same sample/tow in the same dataset
- Keep sampling done at the same time in the same data publication, even if there are different sampling types

Documentation and synthesis

- Where possible, fill in zeroes for species not caught to make synthesis and analysis easier; if not, include code to correctly fill in zeroes for your dataset
- If water quality is collected, but fish sampling cannot be conducted, fill NAs for species, or leave species blank and advise users to left join water quality to species so that NAs are created for species.
- Where possible, record when each taxa starts getting enumerated, or when there are changes to how taxa are recorded.
- Code for processing should be public. It can be published in EDI with each data version or on GitHub. If on GitHub, provide link in EDI publication abstract or metadata, and if there is available version information, include on EDI.
- Where possible, use IEP standardized codes for variable naming, and include a column of IEP fish codes along with taxonomy table to make synthesis easier:

Resources:

- NCEAS Data Modeling Essentials
- Borer et al. Some Simple Guidelines for Data Management
- IEP Data Publication Git Hub Page

Version number	Date created	Description of changes	Justification for change	Version editor(s)	Contact info
1.0	6/20/2022	Created document - DPWG		DPWG	trinh.nguyen@wildlife.ca.gov

Data Publication Table Structure

Example

Notes

- Highlighted indicates primary key (this table has one key per row).
- Arrows indicate direction of linkage.
- **environment** and **sample event** table may be combined if there is only one tow/sample event per water quality measurement.