

ESS-DIVE Package Level Metadata Review

Joan Damerow, Emily Robles, Zarine Kakalia, Charu Varadharajan



U.S. DEPARTMENT OF
ENERGY

Office of
Science



Introduction

- **Introduction**
- Background Research
- Proposed Checks
 - Files
 - Title length and content
 - Keywords
 - Abstract length and content
 - Methods presence and content
- ESS-DIVE Metadata Review
 - Automated Checks
 - Quality Reports
 - Goals

Motivation: Why spend time to create “FAIR” Metadata

Publishing, funding and scientific community moving towards FAIR

SBR Report - Open Watershed Science by Design

Maximize future value of data - support data REUSE

Citations, Contribute to H-Index

FAIR Principles



Findability

Resource and its metadata are easy to find by both, humans and computer systems. Basic machine readable descriptive metadata allows the discovery of interesting data sets and services.



Accessibility

Resource and metadata are stored for the long term such that they can be easily accessed and downloaded or locally used by humans and ideally also machines using standard communication protocols.



Interoperability

Metadata should be ready to be exchanged, interpreted and combined in a (semi)automated way with other data sets by humans as well as computer systems.



Reusability

Data and metadata are sufficiently well-described to allow data to be reused in future research, allowing for integration with other compatible data sources. Proper citation must be facilitated, and the conditions under which the data can be used should be clear to machines

Data packages are publications



Chipman D ; Takahashi T ; Breger D ; Sutherland S (1996): CO₂, Hydrographic, and Chemical Data Obtained During the R/V Meteor Cruise 11/5 in the South Atlantic and Northern Weddell Sea Areas (WOCE sections A-12 and A-21). CDIAC.

doi:10.3334/CDIAC/OTG.NDP045

Metrics and actions for the data package:

- Citations: 4
- Downloads: 6
- Views: 45
- Copy Citation
- Edit

4 Citations

Alissa Zuijgeest, Simon Baumgartner, and Bernhard Wehrli. 2016. [Hysteresis effects in organic matter turnover in a tropical floodplain during a flood cycle](https://doi.org/10.1007/s10533-016-0263-z). *Biogeochemistry*. Vol. 131. pp. 49-63. <https://doi.org/10.1007/s10533-016-0263-z>.

Nicolas Gruber and Jorge L. Sarmiento. 2004. [Global patterns of marine nitrogen fixation and denitrification](https://doi.org/10.1029/97GB00077). *Global Biogeochemical Cycles*. Vol. 11. pp. 235-266. <https://doi.org/10.1029/97GB00077>.

Nicolas Gruber, Manuel Gloor, Song-Miao Fan, and Jorge L. Sarmiento. 2004. [Air-sea flux of oxygen estimated from bulk data: Implications For the marine and atmospheric oxygen cycles](https://doi.org/10.1029/2000GB001302). *Global Biogeochemical Cycles*. Vol. 15. pp. 783-803. <https://doi.org/10.1029/2000GB001302>.

Nicolas Gruber. 2004. [Anthropogenic CO₂ in the Atlantic Ocean](https://doi.org/10.1029/97GB03658). *Global Biogeochemical Cycles*. Vol. 12. pp. 165-191. <https://doi.org/10.1029/97GB03658>.

Familiar publication process

Submit an Article at ScholarOne Manuscripts

Submission

- Step 1: Type, Title, & Abstract >
- Step 2: File Upload >
- Step 3: Attributes >
- Step 4: Authors & Institutions >
- Step 5: Reviewers >
- Step 6: Details & Comments >
- Step 7: Review & Submit >

* Title

(Avoid using acronyms in the title as much as possible.)

Preview Special Characters

0 OUT OF 50 WORDS

Title is missing.

* Abstract

(Be sure to define ALL acronyms in the abstract.)

Write or Paste Abstract

Preview Special Characters

0 OUT OF 350 WORDS

Abstract text is required.

Files

0.00 OUT OF 37.56 MB

ORDER	ACTIONS	FILE	* FILE DESIGNATION	UPLOAD DATE	UPLOADED BY
No files uploaded					

Update Order

* Keywords

[+ Show Full List](#)

Add Author

Find using Author's email address



Submit Data



Add files to start your dataset

Overview

Overview

People

Title *

A brief but meaningful title for this data package.
A good title includes the topic, geographic location, dates, and scale of the data.

Example: Sapflow and Soil Moisture Raw sensor data, Jan 2016-Apr 2016, BR-Ma2: Manaua

Dates

Locations

Existing DOI and Alternate Identifiers

DOI and alternate identifiers of the data package if it has been previously published elsewhere. Alternate identifiers provide pertinent information that can identify and locate the data set within your site's data management system.

Enter as many identifiers as needed below.

Example: <http://dx.doi.org/10.15486/NGT/XXXXXX>

Methods

Abstract *

A concise description of the purpose and content of this data package.
A good abstract would provide users with adequate information to determine if the data are useful for their needs.

Example: Raw output from the data logger connected to 9 sapflow and 5 soil moisture sensors are provided in xxx.dat. The metadata file (BR-Ma2 E-field_log_20160501.xls) has information on locations where the sensors were installed, and other installation/maintenance details. No data processing or QA/QC was done on the raw data packages. Processed data packages will be uploaded separately.

Keywords *

Keywords that should be associated with this data package to enable thematic searches.
Search for a keyword from the list or write in your own. Tab or click enter to add to the list below with one keyword per line. The list contains GCMD keywords.

Use autocomplete feature to pick from the existing keywords.

Example: EARTH SCIENCE > LAND SURFACE > SOILS

Data Variables

Measurement variables present in the data package.
Search for a variable from the list or write in your own. Tab or click enter to add to the list below with one variable per line. The list contains GCMD and CF variables.

Use autocomplete feature to pick from the existing variables.

Example: EARTH SCIENCE > LAND SURFACE > SOILS > SOIL MOISTURE/WATER CONTENT

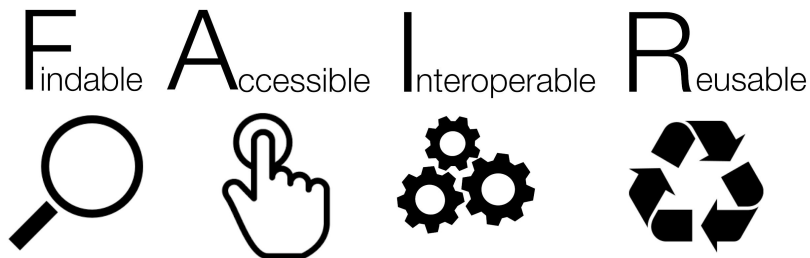
Contact

Person who should be listed as the contact for the data package for the purposes of the DOI or for users seeking further information for the data.
Only one contact is allowed per data package. If none are entered, you will be set as the contact for this document.

First name Last name Email
Organization name For maximum impact, add this person's ORCID. https://orcid.org/orcid-see

All information entered in this section will be made public once the data package is published.

Descriptive Metadata Make Data More



More journals and funders require data in repositories that support FAIR principles

Datasets are valuable research contributions, “not files that are shoved in as an afterthought.”



COMMENT · 04 JUNE 2019 · CORRECTION 05 JUNE 2019

Make scientific data FAIR

All disciplines should follow the geosciences and demand best practice for publishing and sharing data, argue Shelley Stall and colleagues.

Increasing calls for the entire scientific community to implement FAIR

Objectives of webinar



Synthesis of background research on metadata requirements and review

Get your **FEEDBACK**:



- Proposed manual metadata checks based on research
- Automated checks and existing quality reports
- DataONE FAIR checks and future quality reports

Background Research

- Introduction
- **Background Research**
- Proposed Checks
 - Files
 - Title length and content
 - Keywords
 - Abstract length and content
 - Methods presence and content
- ESS-DIVE Metadata Review
 - Automated Checks
 - Quality Reports
 - Goals

Background Research - Repositories

Similar data repositories and documented requirements for high-level fields

Contacted multiple repository representatives for additional information

Reviewed EML metadata schema to identify common required fields

Repositories:

- Arctic Data Center
- Environmental Data Initiative
- The Knowledge Network for Biocomplexity
- EarthData
- NOAA
- ORNL DAAC
- USGS
- NGEE Tropics Archive
- Pangaea



Background Research - Journals



Reviewed requirements for the same basic information required by earth and environmental science journals

Journals:

- Environmental Modelling & Software
- Science of the Total Environment
- IEEE Access
- ESA Journals
- Nature
- Science
- Environmental Science and Technology



Background Research - Datasets



Sampling - Whondrs

- Collection dates
- Sampling procedure (depth, location, instrumentation)
- Amount and frequency of medium collected
- Analyses done to samples

Field Campaign - NGEE Tropics

- Collection time
- Data collected for each sample
- Larger campaign for which these samples are a part of

Sensors and QA/QC - Ameriflux

- Installation of sensors and data loggers (height, instrumentation)
- Collection frequency
- Corrections and calculations to raw data
- Quality Control thresholds

Field Experiment - SPRUCE

- Field site
- Treatment/Manipulation procedure done
- Responses recorded

Laboratory Experiment - NGEE Arctic

- Sample retrieval site
- Treatment/Manipulation procedure done
- Responses recorded

Model Data - FACE

- Data and protocols necessary to simulate the experiments
- Major corrections to the original data

Existing Automated Metadata Quality Reports



Testing automated checks and reports developed by NCEAS/DataONE

Evaluate whether datasets pass/fail certain checks

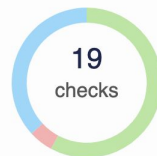
Generally based on some basic FAIR principles, but many will be upgraded

Demo later in presentation



Metadata Quality Report

After running your metadata against our standard set of metadata, data, and congruency checks, we have research data by addressing the issues below.



Identification: 100% complete

Discovery: 67% complete

Interpretation: 100% complete

▶ Passed 11 checks out of 12 (informational checks not included).

▶ Warning for 0 checks.

▼ Failed 1 check. Please correct these issues.



The abstract is only 75 word(s) long but 100 or more is required.

DataONE FAIR Checks

Participated in ESIP workshop

Provide feedback on DataONE FAIR checks (follow links to review and provide feedback:

- [Findable](#)
- [Accessible](#)
- [Interoperable](#)
- [Reusable](#)



Designed to quantify FAIR scores for datasets and entire repositories

- Some checks are more relevant at the repository level based on requirements
- **Required** directly affects FAIR score, **Optional** only applies to score if pass

Current checks either correspond to one of the FAIR principles or can be upgraded to a FAIR check when finalized

Proposed Checks

- Introduction
- Background Research
- **Proposed Checks**
 - Files
 - Title length and content
 - Keywords
 - Abstract length and content
 - Methods presence and content
- ESS-DIVE Metadata Review
 - Automated Checks
 - Quality Reports
 - Goals

Proposed checks: **Files**

Related repository requirements:

- Use of common file formats is required
- Any code used to process data included
- Each file has a short descriptive name



Proposed checks:

- **At least one associated file** - Accessible
- **Use common non-proprietary file formats where possible (e.g. csv, txt, pdf, png, jpeg, tiff, R or Python scripts, many others).** - Interoperable/Reusable
- **Software is specified, if necessary** - Reusable

**More extensive requirements coming soon with file metadata standards

Proposed check: Title length and content

Related repository requirements:

- Common range: 5 words minimum, 20 words maximum
- Include data package topic, geographic location, and dates at minimum
- Format similarly to a journal title

Journal requirements:

- Specific and informative
- Avoid abbreviations and acronyms
- Vary from **maximum of 96 to 120 characters**

Proposed checks:

- **Title length is 7- 20 or 40? words - Findable**
- **Title reflects data package specifically and may include information on what/where/when data was collected - Findable**
- **No unexplained acronyms or project-specific jargon**

Title Example



Title *

A brief but meaningful title for this data package.

A good title includes the topic, geographic location, dates, and scale of the data.

Date range

Predawn Leaf Water Potential of Oak-Hickory Forest at Missouri Ozark (MOFLUX) Site: 2004-2017

Clearly stated variable

Location

Proposed check: **Keywords**

Related repository requirements:

- Keywords related to data type and geographic locations

Journal requirements:

- Average 3-6 keywords
- Do not use words included in the title
- Can only include established acronyms

Proposed checks:

- **There are at least 3 keywords, differ from words in title - Findable**
- *Keywords from standardized controlled vocabularies - Findable*

Controlled Keywords Example



Keywords *

Keywords that should be associated with this data package to enable thematic searches.

Search for a keyword from the list or write in your own. Tab or click enter to add to the list below with one keyword per line. The list contains [GCMD](#) keywords.

Use autocomplete feature to pick from the existing keywords.

Earth|

EARTH SCIENCE > AGRICULTURE > AGRICULTURAL AQUATIC SCIENCES: CATEGORICAL:GCMD
EARTH SCIENCE > AGRICULTURE > AGRICULTURAL CHEMICALS: CATEGORICAL:GCMD
EARTH SCIENCE > AGRICULTURE > AGRICULTURAL ENGINEERING: CATEGORICAL:GCMD
EARTH SCIENCE > AGRICULTURE > AGRICULTURAL PLANT SCIENCE: CATEGORICAL:GCMD
EARTH SCIENCE > AGRICULTURE > ANIMAL COMMODITIES: CATEGORICAL:GCMD
EARTH SCIENCE > AGRICULTURE > ANIMAL SCIENCE: CATEGORICAL:GCMD
EARTH SCIENCE > AGRICULTURE > FEED PRODUCTS: CATEGORICAL:GCMD
EARTH SCIENCE > AGRICULTURE > FOOD SCIENCE: CATEGORICAL:GCMD

Proposed check: Abstract length and content



Related repository requirements:

- Summarizes the purpose and content of data
- Minimum required length varies between **20 and 100 words** depending on repository

Journal requirements:

- Contents of the dataset
- When and where the data were collected
- How to use the data
- **Purpose of collecting the data**
- Understandable to anyone in the scientific community

Proposed checks:

- **Abstract at least 100 words - Findable**
- **Include clear and concise description of the purpose and contents - Findable/Reusable**
- **Understandable to anyone who has not seen related manuscripts and contains no unexplained acronyms**

Abstract Example



Purpose



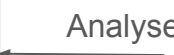
Abstract

This dataset is from a global survey of surface water metabolites to provide understanding of the character of organic carbon that may be delivered to subsurface sediments via hydrologic exchange. To implement the global survey, free stream sampling kits were provided to interested researchers throughout the world. Samples were collected with minimal constraints in terms of location, but following strict protocols, and shipped to the Environmental Molecular Sciences Laboratory (EMSL) for metabolomic analysis via Fourier transform ion cyclotron resonance mass spectrometry (FTICR-MS). In addition, basic geochemistry analyses (e.g., dissolved organic carbon concentration, cations, anions) were conducted, standardized photos of each field system were taken, surface water hydrographs were collated from existing instrumentation, and extensive metadata were captured. All data types are provided in a standard format. In addition, the data package contains an R function that will launch a GUI that can be used to easily search, compile, and download data. The data are free to be used for any purpose, such as for manuscripts, presentations, and grant proposals. Please use the data package's DOI to cite the data package. Note that individual hydrographs have separate DOIs, which are provided in the associated hydrograph files. These hydrograph-specific DOIs should also be cited when using those data. We ask that you email us at WHONDRS@pnnl.gov to let us know that you're using the data and acknowledge WHONDRS and the U.S. Department of Energy's Subsurface Biogeochemical Research program—which generously provides funding to WHONDRS—in your documents, presentations, etc. There is no obligation to include WHONDRS members as co-authors.

Sampling



Analyses



Contents



257 words

Proposed check: **Methods presence and content**



Related repository requirements:

- Descriptions of field and laboratory **sample collection**
- Details about hardware and software used to produce data
- Descriptions of **how the data were generated** and (if applicable) modified

Journal requirements:

- Sufficient information for a user to **understand and reproduce** your work
- Experimental design, sampling procedures, and QA/QC

Proposed checks:

- **Methods contain more than 7 words** - Reusable
- **Methods are included and clearly written, or at least refer to a previous publication** - Reusable
- **Encourage that methods document all data collection, processing, and/or QA/QC steps to produce the data** - Reusable

Methods Example



Methods	Step 1
Description	<p>Tall tower N₂O mixing ratios were measured using a tunable diode laser technique (TGA100, Campbell Scientific Inc., Logan, Utah, USA). The TDL measured N₂O at wavenumber 2243.760 cm⁻¹. The TDL was maintained at the base of the tall tower in a temperature-controlled radio communications building. Calibrations were performed hourly with standards traceable to the NOAA-ESRL (National Oceanic and Atmospheric Administration - Earth System Research Laboratory) 2006A N₂O mole fraction scale. The NOAA-ESRL gold standard (Standard Cylinder #CA07980) has a mixing ratio (mean ± 1 standard deviation) of 324.30 ± 0.09 ppb as determined by NOAA-ESRL). The hourly precision of the tall tower calibration measurements was calculated from Allan variance analysis of working standards and was 0.50 ppb (16).</p>

Technique and instrumentation

Methods	Step 2
Description	<p>Air from the tall tower was sampled from inlets at approximately 32, 56, 100, and 185 m. Air was pulled continuously through each of the inlets using a flow rate in excess of 15 SLPM to the base of the tower and then sub-sampled at 3 SLPM using a custom designed manifold. The air sampling and calibration consisted of the following sampling sequence where each inlet was sampled for 15 s: ultra zero air; CO₂ span 1; N₂O span; CO₂ span 2; 185 m inlet; 100 m inlet; 56 m inlet; and 32 m inlet. The air samples were dried prior to analysis using a Nafion dryer. All of the calibrated data were then block averaged into hourly values. The hourly values were filtered using a basic low pass/high pass filter and subsequent wavelet decomposition (described below). Extreme N₂O outliers are defined as > 360 ppb or < 315 ppb. Further, any hourly periods with the TGA100 system reporting a status error were filtered. The TGA100 error status usually indicates that the laser was not locked onto the target absorption line.</p>

Sampling details

Methods	Step 3
Description	<p>A basic low pass/high pass filter for N₂O and CO₂ have been used to quality control these data. These thresholds could be tightened further, but provide a very good first level of filtering.</p> <p>Additional filtering using wavelet analyses is also provided. Here, we used the Haar wavelet to decompose the original N₂O signal into low-pass filtered coefficients and high-pass filtered details using level 1 through level 6 decomposition. All analyses were performed using the wavedec function available in the MATLAB Wavelet Toolbox (MATLAB, R2013b The Mathworks Inc., MA, USA). The wavelet filtered data A1 through A6 are also provided here.</p>

QA/QC standards

Griffis T ; Baker J ; Millet D ; Chen Z ; Wood J ; Erickson M ; Lee X (2016): KCMP Minnesota Tall Tower Nitrous Oxide Inverse Modeling Dataset 2010-2015. AmeriFlux. doi:10.15485/1398272



ESS-DIVE Metadata Review

- Introduction
- Background Research
- Proposed Checks
 - Files
 - Title length and content
 - Keywords
 - Abstract length and content
 - Methods presence and content
- **ESS-DIVE Metadata Review**
 - Automated Checks
 - Quality Reports
 - Goals

ESS-DIVE New Review Process



Automated quality checks focus on presence of metadata and word counts - see Quality Report for instant feedback

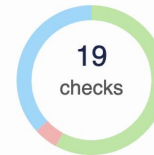
Manual metadata content reviews

Quality Reports from automated checks for each package, eventually projects, and ESS-DIVE

Review form for consistent checks and feedback
Data on quality before/after review and time

Metadata Quality Report

After running your metadata against our standard set of metadata, data, and congruency checks, we have research data by addressing the issues below.



Identification: 100% complete

Discovery: 67% complete

Interpretation: 100% complete

▶ Passed 11 checks out of 12 (informational checks not included).

▶ Warning for 0 checks.

▼ Failed 1 check. Please correct these issues.



The abstract is only 75 word(s) long but 100 or more is required.

Existing Automated Metadata Quality Reports



Developed by NCEAS and DataONE
Quality report shows percentage of checks
that pass different categories

Demo:

<https://data.ess-dive.lbl.gov/quality/ess-dive-28ef3e4a1360a48-20190815T182803220548>

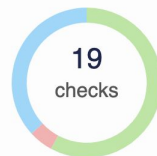
<https://data.ess-dive.lbl.gov/quality/ess-dive-ca7ea9922ea9aff-20181219T160938778966>

<https://data.ess-dive.lbl.gov/quality/ess-dive-3db5398a4a2fb59-20180704T200317625>



Metadata Quality Report

After running your metadata against our standard set of metadata, data, and congruency checks, we have research data by addressing the issues below.



Identification: 100% complete

Discovery: 67% complete

Interpretation: 100% complete

▶ Passed 11 checks out of 12 (informational checks not included).

▶ Warning for 0 checks.

▼ Failed 1 check. Please correct these issues.



The abstract is only 75 word(s) long but 100 or more is required.

Goals for Package Metadata Review



Clear guidance for each metadata element

Data and metadata curation done by data package authors

Automated checks and manual content review

Efficient (10-15 min) and standardized

Review feedback providing specific suggestions for metadata

Working towards FAIR data standards



Feedback on proposed automated and manual checks:



<https://docs.google.com/spreadsheets/d/14v3hjPL9jDSgfSF6RCyDgwEKZIK4xwAzJRCb0zjTsfq/edit?usp=sharing>