ESS-DIVE Package Level Metadata Review

Joan Damerow, Emily Robles, Zarine Kakalia, Charu Varadharajan

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

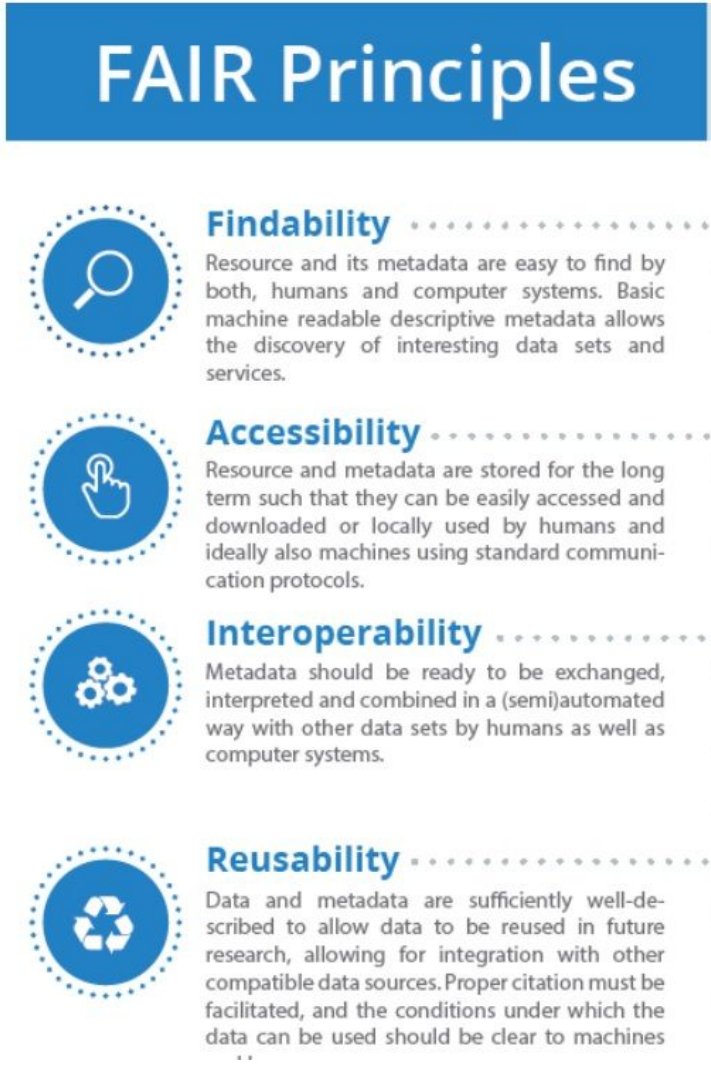
Motivation: Why spend time to create “FAIR” Metadata

Publishing, funding and scientific community moving towards FAIR

SBR Report - Open Watershed Science by Design

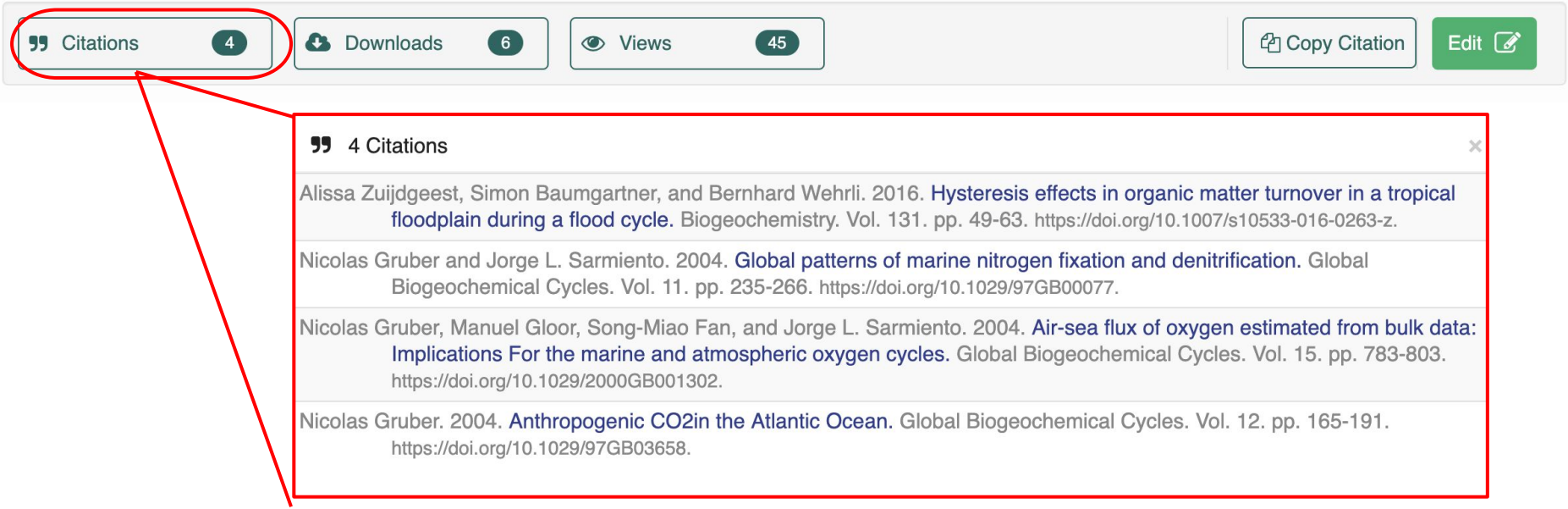
Maximize future value of data - support data REUSE

Citations, Contribute to H-Index

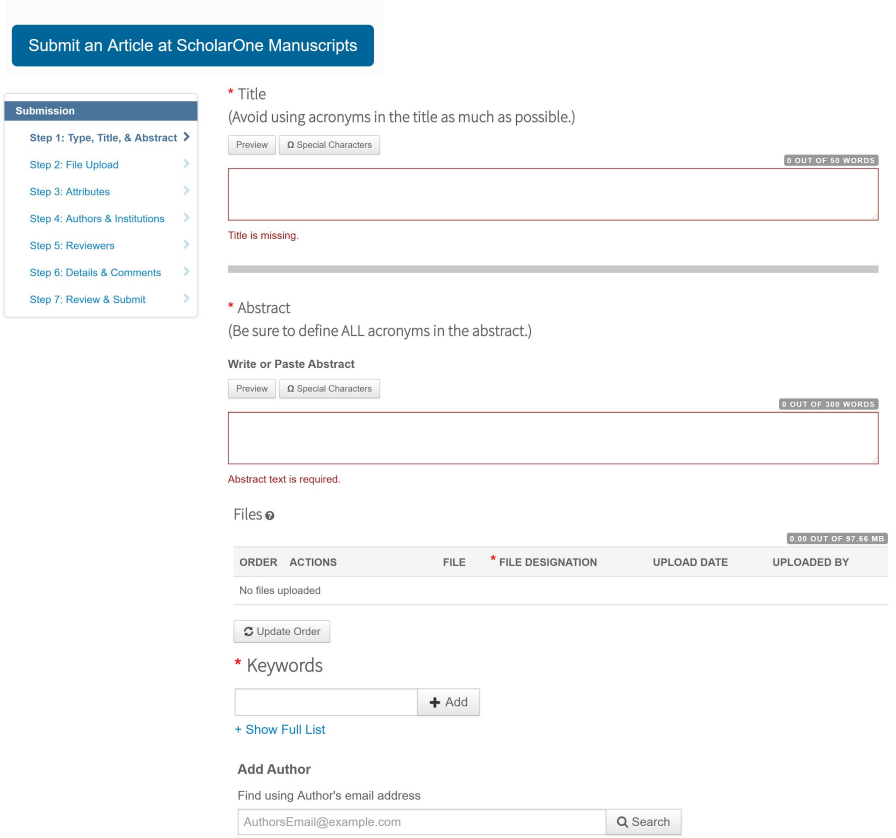


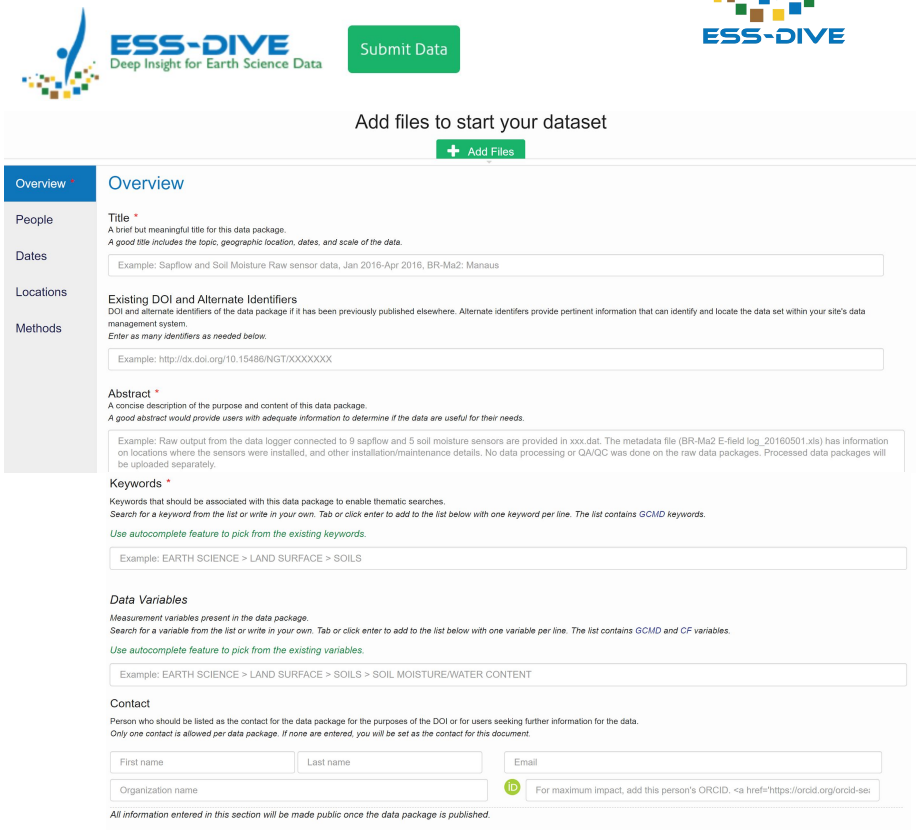
Data packages are publications

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

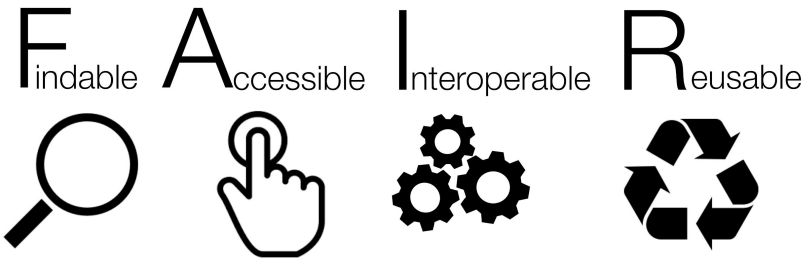


Familiar publication process



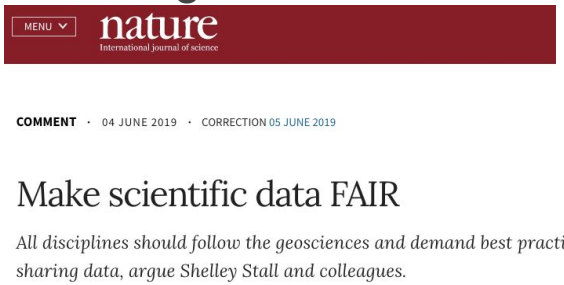


Descriptive Metadata Make Data More



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.”


MENU nature International journal of science

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

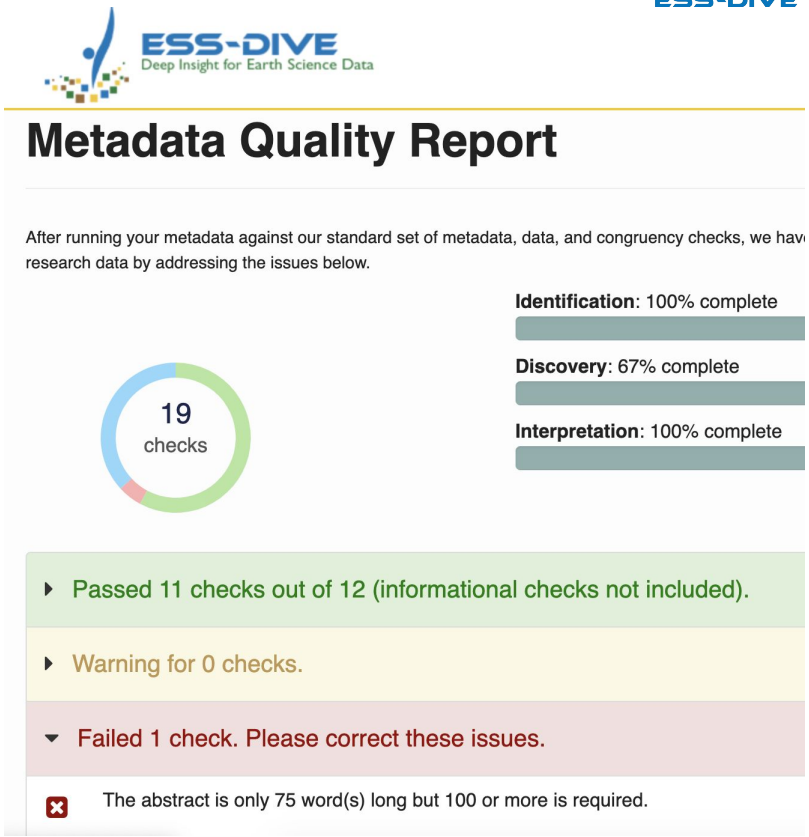
Existing Automated Metadata Quality Reports

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



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.

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

Date range

Clearly stated variable

Location

Pallardy S ; Gu L ; Wood J ; Hosman K P ; Sun Y (2018): Predawn Leaf Water Potential of Oak-Hickory Forest at Missouri Ozark (MOFLUX) Site: 2004-2017. Climate Change-Terrestrial Ecosystem Science SFA. doi:10.3334/CDIAC/ORNLSFA.004

Proposed check: **Keywords**

Related repository requirements:

- Keywords related to data type and geographic locations

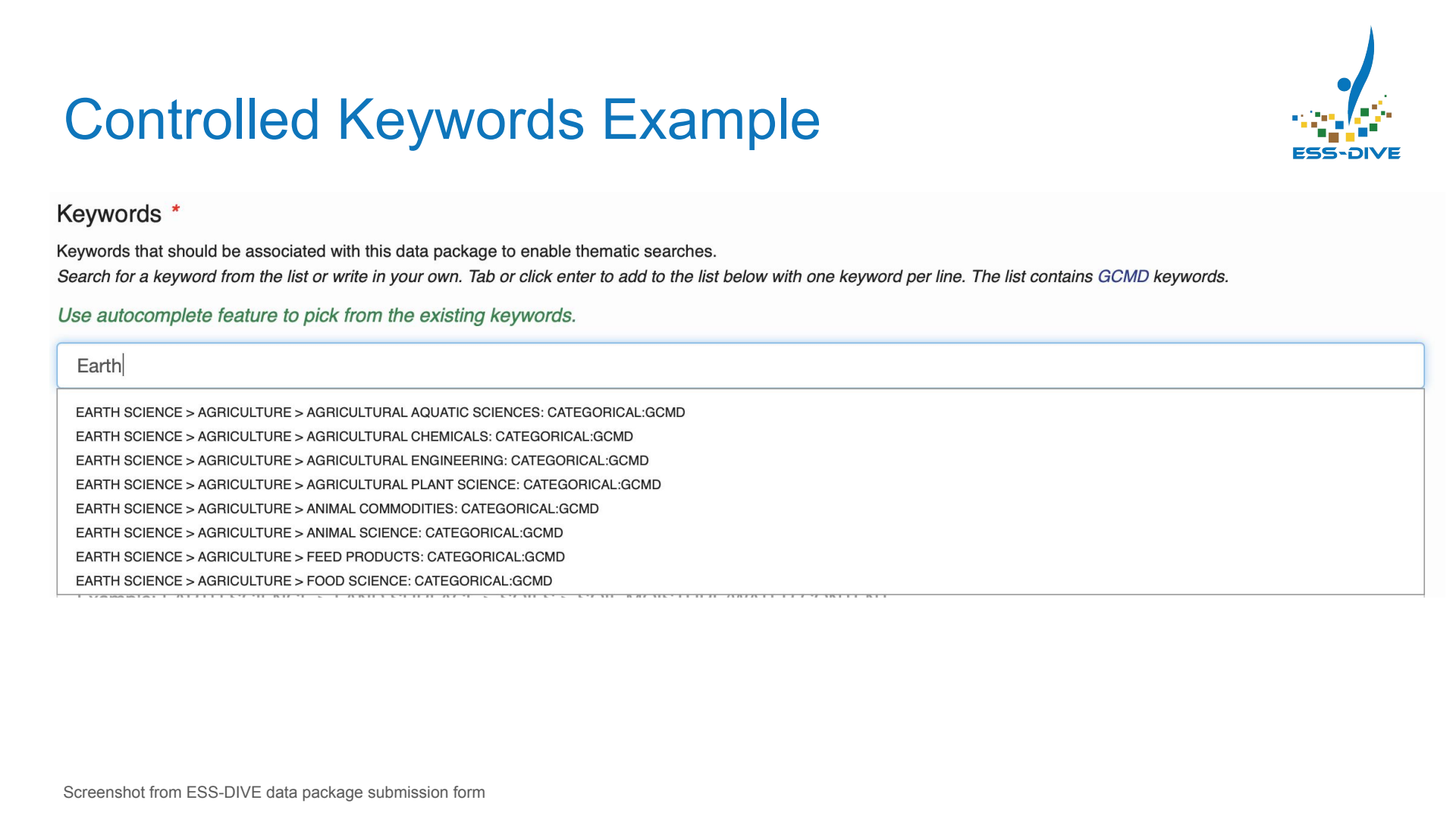
Journal requirements:

- 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



Screenshot from ESS-DIVE data package submission form

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

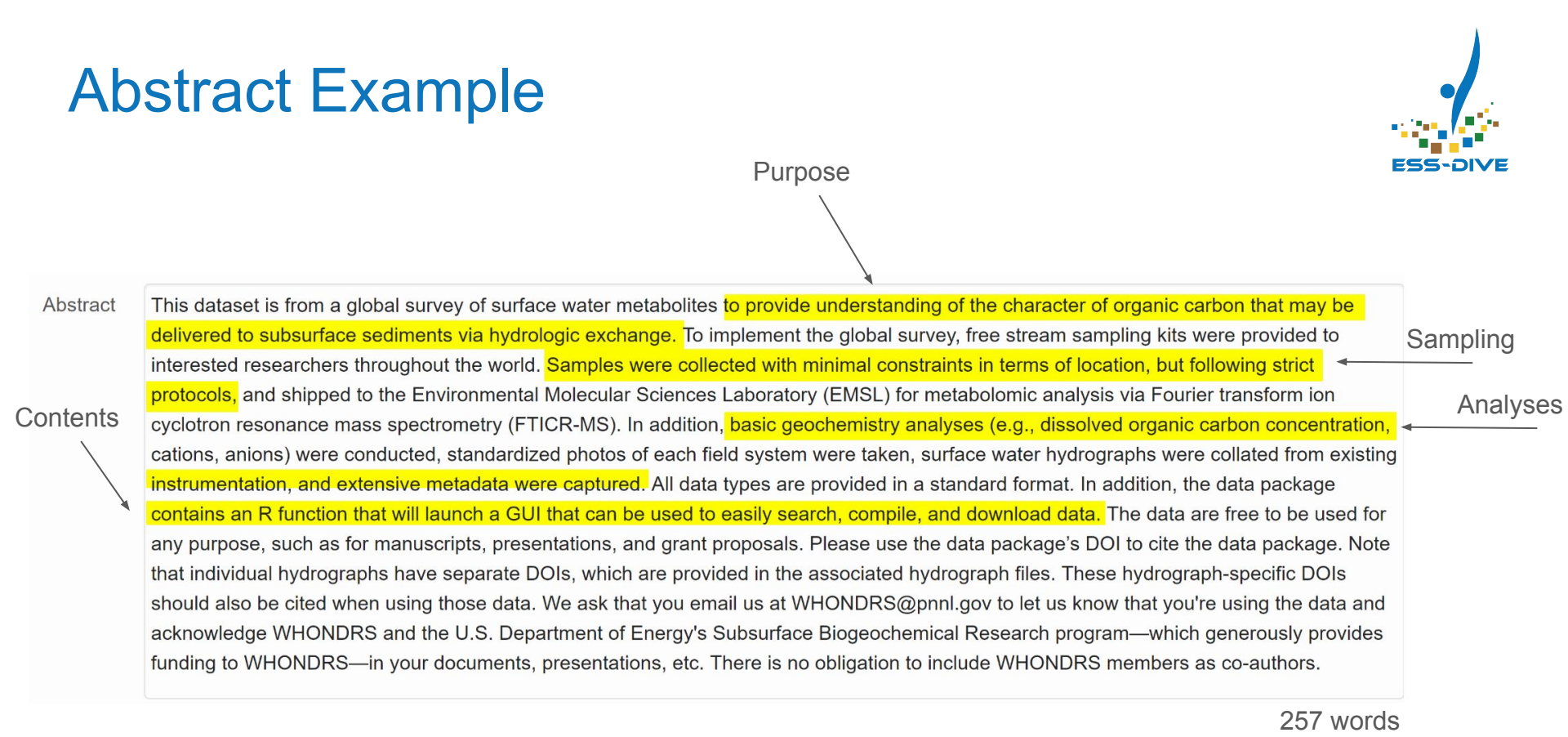
Journal requirements:

- 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



Stegen J C ; Goldman A E ; Blackburn S E ; Chu R K ; Danczak R E ; Garayburu-Caruso V A ; Graham E B ; Grieshauber C ; Lin X ; Morad J W ; Ren H ; Renteria L ; Resch C T ; Tfaily M ; Tolic N ; Toyoda J G ; Wells J R ; Znotinas K R (2018): WHONDRS Surface Water Sampling for Metabolite Biogeography. Worldwide Hydrobiogeochemistry Observation Network for Dynamic River Systems (WHONDRS). doi:10.15485/1484811

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

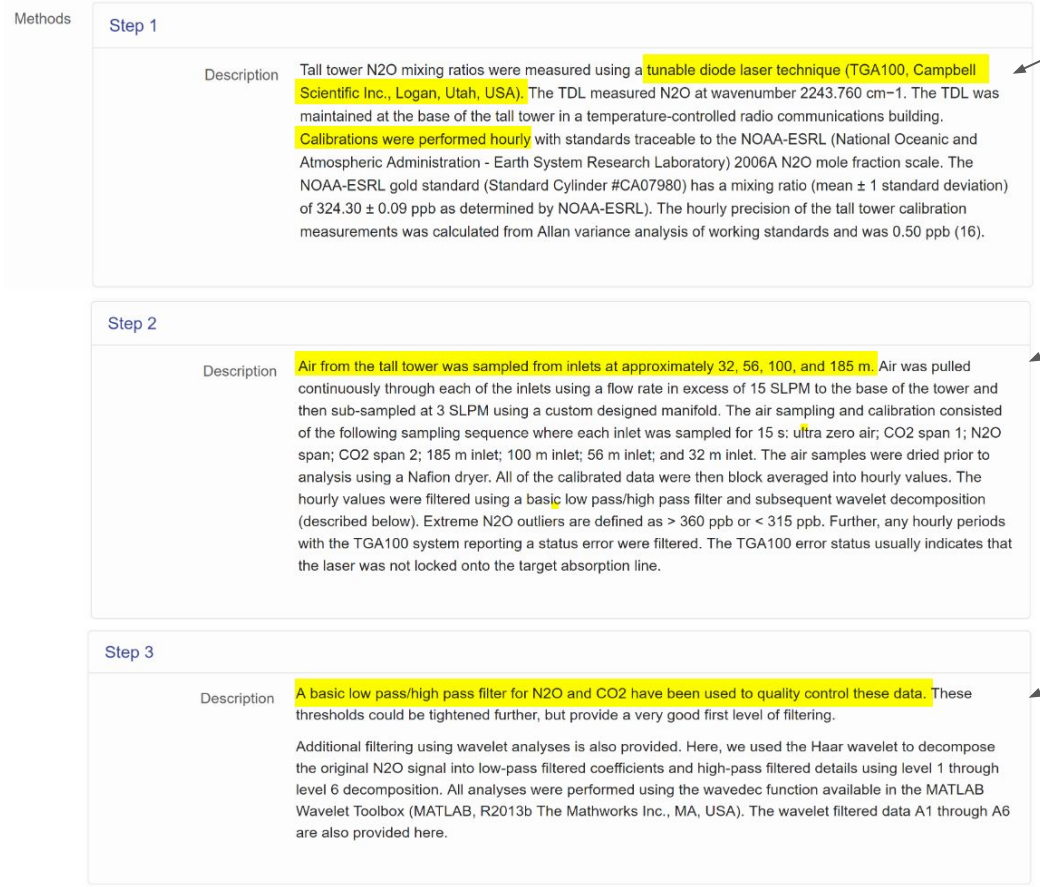
Journal requirements:

- 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

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).

Step 2

Description

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.

Step 3

Description

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.

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.

Technique and instrumentation

Sampling details

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

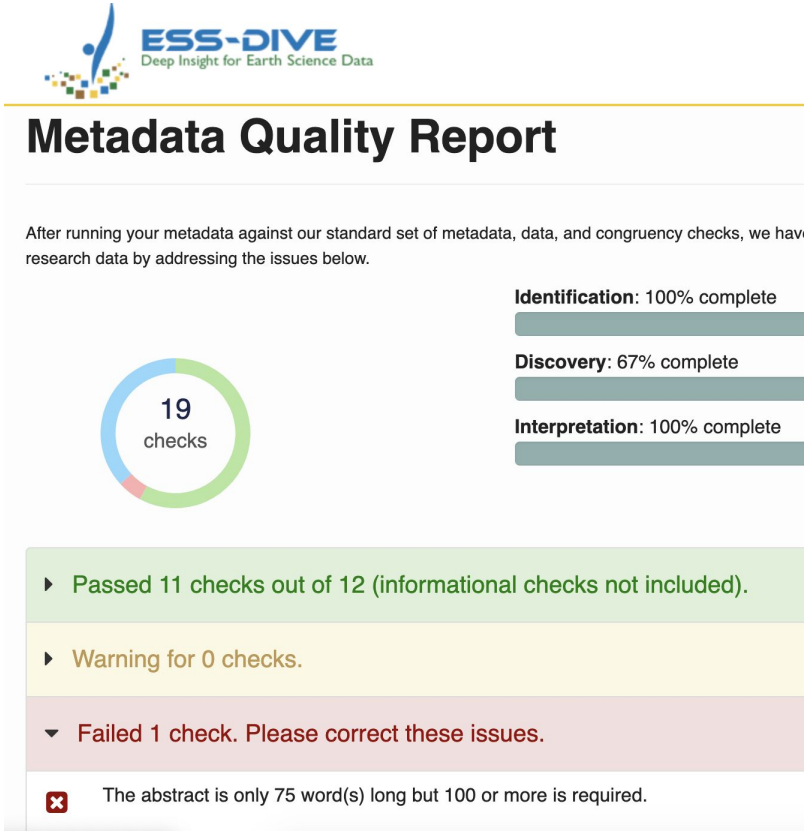
ESS-DIVE New Review Process

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



Existing Automated Metadata Quality Reports

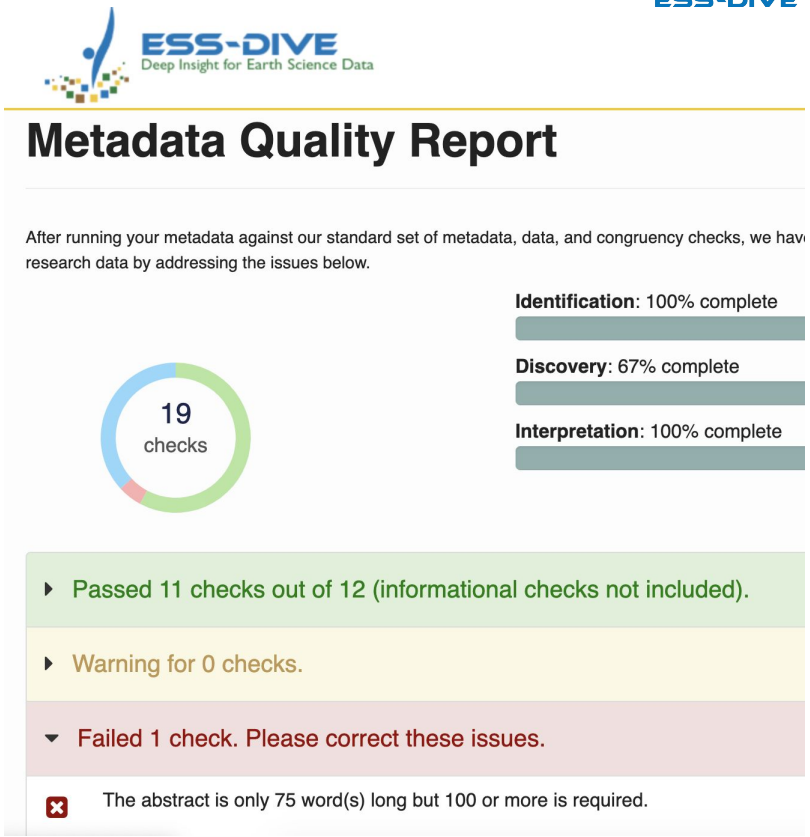
Developed by NCEAS and DataONE
Quality report shows percentage of checks that pass different categories

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



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/14v3hjPL9jDSgfSF6RCyDgwEKZIK4xwAzJRCb0zjTsfg/edit?usp=sharing