How To Make Your Data Count

July 10, 2018

Alfred P. Sloan Foundation

University of California CDL California Digital Library

DataCite FIND, ACCESS, AND REUSE DATA

DataONE

Agenda

- Milestones in 2018
- Our first release
- Implementation at your repository
  - Code of Practice
  - Log Processing
  - Sending Usage Logs
  - Pulling & Displaying Usage & Citation Metrics
- Questions

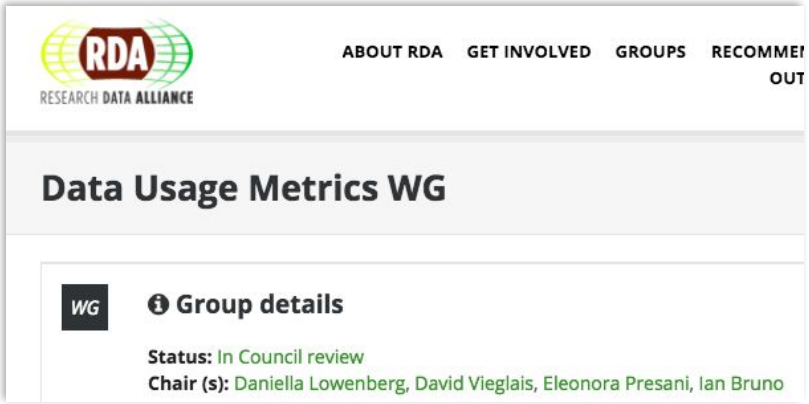
Milestones thus far

We created a data usage metrics standard

PeerJ Preprints NOT PEER-REVIEWED

Code of Practice for Research Data Usage Metrics Release 1

Martin Fenner[1], Daniella Lowenberg[2], Matt Jones[3], Paul Needham[4], Dave Vieglais[5], Stephen Abrams[2], Patricia Cruse[1] and John Chodacki[2] on behalf of the Make Date Count Project



RDA RESEARCH DATA ALLIANCE

ABOUT RDA GET INVOLVED GROUPS RECOMMEN OUT

Data Usage Metrics WG

WG

Group details

Status: In Council review

Chair (s): Daniella Lowenberg, David Vieglais, Eleonora Presani, Ian Bruno

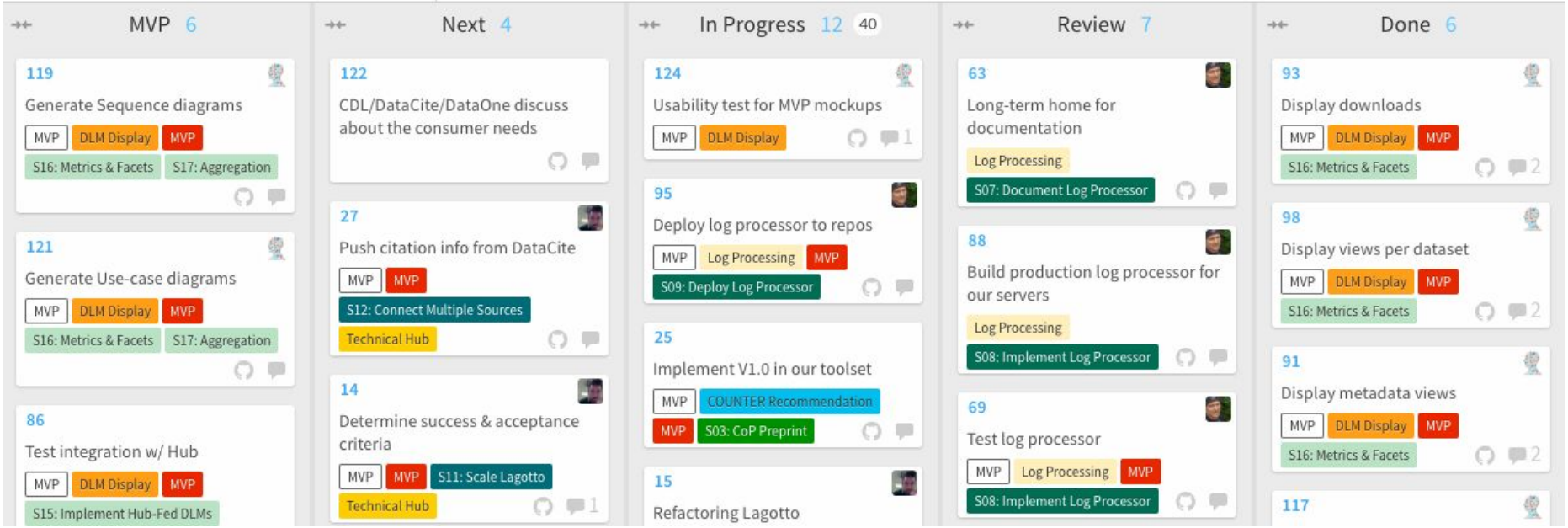
CDL DataCite DataONE

We gave MDC a narrative

We built out an open hub for usage metrics

https://api.datacite.org/events

We implemented at our own repositories



We began to address citations

Glad You Asked: A Snapshot of the Current State of Data Citation

June 1, 2018 by Kristian Garza and Martin Fenner • 7 min read

https://doi.org/10.5438/h16y-3d72

What does it look like?

European high-skilled mobility data and Scientific publication & collaboration data

Petersen, Alexander, UC Merced, https://orcid.org/0000-0002-0955-3483

Publication date: March 4, 2017

Publisher: UC Merced

https://doi.org/10.6071/M3RP49

Citation

Petersen, Alexander (2017), European high-skilled mobility data and Scientific publication & collaboration data, UC Merced Dash, Collection, https://doi.org/10.6071/M3RP49

Abstract

Raw data files corresponding to longitudinal country-level observations for two types of variable: (i) publication, citation, and international collaboration data from SCImago Journal & Country Rank (www.scimagojr.com/) and (ii) high-skilled mobility data, corresponding to headcounts by source and destination country, from the "Professionals moving abroad

Download the dataset ~ 3 MB

Download Data Publication (PDF)

Versions

March 4, 2017

Metrics

20 views

2 downloads

2 citations

CDL DataCite DataONE

European high-skilled mobility data and Scientific publication & collaboration data

Download the dataset ~ 3 MB

Download Data Publication (PDF)

Peterse

Publicat

Publishe

https://c

Citatic

Peterser

Merced

Works Referencing This Dataset

Petersen, Alexander M.; Puliga, Michelangelo (2017), High-skilled labour mobility in Europe before and after the 2004 enlargement, The Royal Society, Article-journal, https://doi.org/10.1098/rsif.2017.0030

Doria Arrieta, Omar A.; Pammolli, Fabio; Petersen, Alexander M. (2017), Quantifying the negative impact of brain drain on the integration of European science, American Association for the Advancement of Science (AAAS), Article-journal, https://doi.org/10.1126/sciadv.1602232

2 downloads

2 citations

Abstract

Raw data files corresponding to longitudinal country-level observations for two types of variable: (i) publication, citation, and international collaboration data from SCImago Journal & Country Rank (www.scimagojr.com/) and (ii) high-skilled mobility data, corresponding to headcounts by source and destination country, from the "Professionals moving abroad (Establishment)" data hosted by the European Union's "The EU Single Market Regulated professionals database".

Haff, Tonya M. and Magrath, Robert D.. 2013. **Data from: To call or not to call: parents assess the vulnerability of their young before warning them about predators.** Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.234d8?ver=2013-10-21T10:36:19.373-04:00.

Citations 20 Downloads 16 Views 38 Copy Citation

Files in this dataset

Name	File type	Size	Downloads (this version)	
Metadata: Data from: To call or not to call: parents assess the vulnerability of their young before warning them about predators	Dryad Metadata Application Profile Version 3.1	3 KB	4 views	Download

General

Id: http://dx.doi.org/10.5061/dryad.234d8?ver=2013-10-21T10:36:19.373-04:00

Abstract: Communication about predators can reveal the effect of both conspecific and heterospecific audiences on signalling strategy, providing insight into signal function and animal cognition. In species that alarm call to their young, parents face a fundamental dilemma: calling can silence noisy offspring and so make them less likely to be overheard, but can also alert predators that young are nearby. Parents could resolve this dilemma by being sensitive to the current vulnerability of offspring, and calling only when young are most at risk. Testing whether offspring vulnerability affects parental strategy has proved difficult, however, because more vulnerable broods are often also more valuable. We tested experimentally whether parent white-browed scrubwren, Sericornis frontalis, assessed brood noisiness when alarm calling. When a model predator was nearby, parents gave more alarm calls when playbacks simulated noisy broods, yet brood noisiness did not affect adult calling when only a control model was present. Parents were therefore sensitive to the tradeoff between silencing young and alerting predators to the presence of nests. Our study demonstrates that receiver vulnerability can affect signalling decisions in species other than primates.

Pub Date: 2013-10-21T14:36:15Z

Haff, Tonya M. and Magrath, Robert D.. 2013. **Data from: To call or not to call: parents assess the vulnerability of their young before warning them about predators.** Dryad Digital Repository. http://dx.doi.org/10.5061/dryad.234d8?ver=2013-10-21T10:36:19.373-04:00.

Citations 20 Downloads 16 Views 38 Copy Citation

Files in this dataset

Name	File type	Size	Downloads (this version)	
Metadata: Data from: To call or not to call: parents assess the vulnerability of their young before warning them about predators	Dryad Metadata Application Profile Version 3.1	3 KB	4 views	Download

General

Id	http://dx.doi.org/10.5061/dryad.234d8?ver=2013-10-21T10:36:19.373-04:00
Abstract	Communication about predators can reveal the effect of both conspecific and heterospecific audiences on signalling strategy, providing insight into signal function and animal cognition. In species that alarm call to their young, parents face a fundamental dilemma: calling can silence noisy offspring and so make them less likely to be overheard, but can also alert predators that young are nearby. Parents could resolve this dilemma by being sensitive to the current vulnerability of offspring, and calling only when young are most at risk. Testing whether offspring vulnerability affects parental strategy has proved difficult, however, because more vulnerable broods are often also more valuable. We tested experimentally whether parent white-browed scrubwren, Sericornis frontalis, assessed brood noisiness when alarm calling. When a model predator was nearby, parents gave more alarm calls when playbacks simulated noisy broods, yet brood noisiness did not affect adult calling when only a control model was present. Parents were therefore sensitive to the tradeoff between silencing young and alerting predators to the presence of nests. Our study demonstrates that receiver vulnerability can affect signalling decisions in species other than primates.
Pub Date	2013-10-21T14:36:15Z

Implementation at your repository

Why it is important

Community has long grappled with the problem of assessing and tracking the results of scholarship

- Researchers
- Repositories
- Funders
- Publishers

Five simple steps to Make YOUR Data Count

1. ***Read*** the data usage metrics standard “Code of Practice for Research Data”

2. ***Process*** your usage logs against this standard

3. ***Send*** processed and standardized usage logs to an open hub

4. ***Pull*** usage and citation metrics from an open hub

5. ***Display*** standardized usage and citation metrics on your repository interface

Getting Started

We have built a “Getting Started” guide walking through these steps as implemented in CDL’s Dash

https://github.com/CDLUC3/Make-Data-Count/blob/master/getting-started.md

1. Code of Practice for Research Data

Implementing the COUNTER Code of Practice for Research Data in Repositories

Getting Started

The COUNTER Code of Practice for Research Data (CoP) is a community developed recommendation for standardized data usage metrics. The below guide outlines how to implement this guide at the repository level, as done by the California Digital Library Dash platform. For a general overview on the project and recommendation:

- View this webinar and in particular the section where Martin Fenner talks about the CoP (starting at around 8 minutes).
- Read through the CoP preprint at https://peerj.com/preprints/26505/
- See examples of logging and log processing infrastructure at a large scale, check out documents from DataONE (Log Aggregation Overview, Event Logging and Reporting, DataONE Usage Statistics)

CDL DataCite DataONE

2. Log Processing

Standardized Logs

- Specialized logs that are processed against Code of Practice
- Views
- Downloads
- Users: at the country level, access during a session
- Session: de-duplicate access to page within 30 seconds

3. Sending Usage Reports

```
Calculating stats for doi:10.15146/R3RW23
Calculating stats for doi:10.7272/Q6057CV6
Calculating stats for doi:10.6086/D1H59V
Calculating stats for doi:10.7272/Q6154F00
Calculating stats for doi:10.7272/Q6SF2T3Q
Calculating stats for doi:10.7280/D1ZH39
Calculating stats for doi:10.6078/D1195K

Writing JSON report to tmp/test_out.json
submitted
```

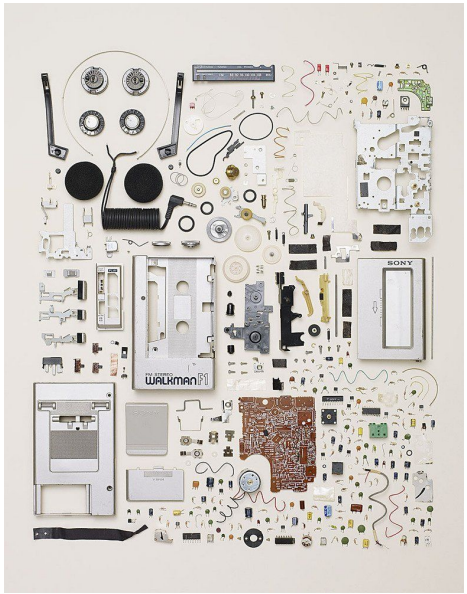
JSON Report - Body

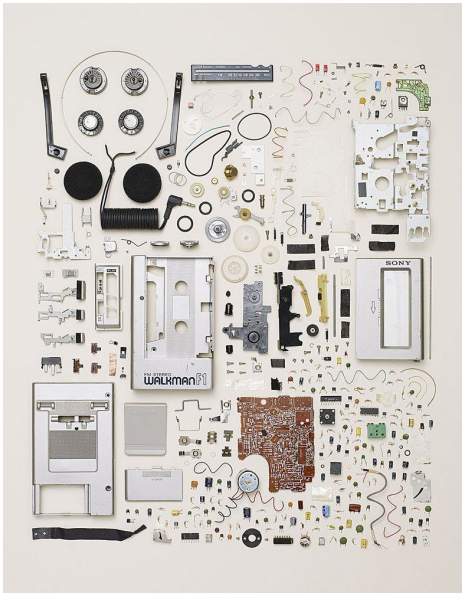

```
Performance: [
    {
        Period: {
            Begin-Date: "2018-03-01",
            End-Date: "2018-03-31"
        },
        Instance: [
            {
                Country: "US",
                Access-Method: "Regular",
                Metric-Type: "Total-Dataset-Requests",
                Count: 1
            },
            {
                Country: "US",
                Access-Method: "Regular",
                Metric-Type: "Unique-Dataset-Requests",
                Count: 1
            },
            {
                Country: "US",
                Access-Method: "Regular",
                Metric-Type: "Total-Dataset-Requests-Size",
                Count: 25262
            },
            {
                Country: "US",
                Access-Method: "Regular",
                Metric-Type: "Unique-Dataset-Requests-Size",
                Count: 25262
            },
            {
                Country: "US",
                Access-Method: "Regular",
                Metric-Type: "Total-Dataset-Investigations",
                Count: 1
            },
            {
                Country: "US",
                Access-Method: "Regular",
                Metric-Type: "Unique-Dataset-Investigations",
                Count: 1
            }
        ]
    }
]
},
```

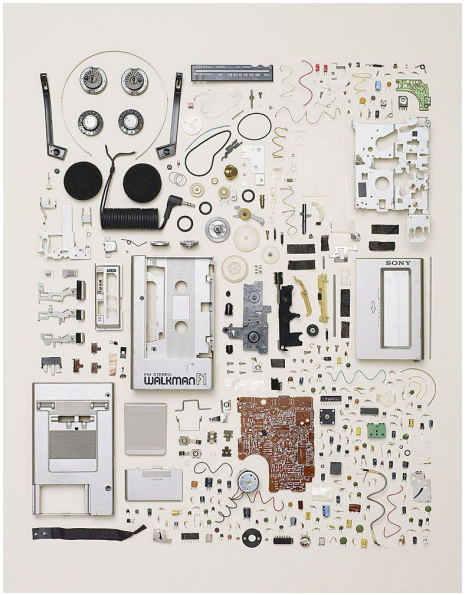

JSON Report - Header


```
{
    report_header: {
        report-name: "Dataset Report",
        report-id: "DSR",
        release: "RD1",
        created: "2018-03-09",
        created-by: "Dash",
        report-filters: [
            {
                Name: "Begin-Date",
                Value: "2018-03-01"
            },
            {
                Name: "End-Date",
                Value: "2018-03-31"
            }
        ],
        Exceptions: [
            {
                code: 3040,
                Severity: "Warning",
                Message: "Partial Data Returned",
                Help-URL: "String",
                Data: "Usage data has not been processed for the entire reporting period"
            }
        ]
    },
    report_datasets: [
        {…},
```

The Usage Metrics Hub







Usage Metrics Hub (hosted by DataCite)

- Aggregator of research data usage reports
- Usage reports are made available via API (in original JSON format) and soon web interface and CSV
- Usage reports are broken down by dataset (and request method), and can then be aggregated over time
- Information in usage reports can be combined with data citations and dataset metadata

4. Pulling Usage and Citations

Pulling Usage and Citations

- Data usage metrics and citations are made available via public API, with one “event” for each data citation or monthly usage count.

- Data citations are provided by DataCite metadata (i.e. come from data repositories) and Crossref, with more to come

- Currently separate APIs for usage and citations, and a third API for dataset metadata, will be combined into single API for easier retrieval of information

What’s next?

Looking Ahead

- Outreach and Adoption
  - Repository & Publishers
- Iterating on our implementation
  - Adding volume and usage by regions
  - Provide aggregation through DataCite
  - Beyond the DOI: metrics for other types of identifiers
  - Possible: altmetrics

Questions?

Alfred P. Sloan Foundation

University of California CDL California Digital Library

DataCite FIND, ACCESS, AND REUSE DATA

DataONE