The DataCite MDC Stack

Martin Fenner, **Kristian Garza**

Blog post published June 22, 2020 in *DataCite Blog - DataCite*
https://doi.org/10.5438/v9pp-7a27

Abstract

In May, the Make Data Count team announced that we have received additional funding from the Alfred P. Sloan Foundation for work on the Make Data Count (MDC) initiative. This will enable DataCite to do additional work in two important areas: Implement a bibliometrics dashboard that enables bibliometricians – funded by a separate Sloan grant – to do quantitative studies around data usage and citation behaviors.

Keywords

Information Systems (Subfield), Make Data Count, Strategy, Bibliometrics, Data Citation, Data Usage

Recommended citation

Fenner, M., & Garza, K. (2020, June 22). The DataCite MDC Stack. *DataCite Blog - DataCite*. https://doi.org/10.5438/v9pp-7a27

Copyright

Copyright © Martin Fenner et al. 2020. Distributed under the terms of the Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

In May, the Make Data Count team announced that we have received additional funding from the Alfred P. Sloan Foundation for work on the Make Data Count (MDC) initiative. This will enable DataCite to do additional work in two important areas:

- Implement a bibliometrics dashboard that enables bibliometricians – funded by a separate Sloan grant – to do quantitative studies around data usage and citation behaviors.
- Increase adoption of standardized data usage across repositories by developing a log processing service that offloads much of the hard work from repositories.

In this blog post, we want to provide more technical details about the upcoming work on the bibliometrics dashboard; the log processing service will be the topic of a future blog post. The bibliometrics dashboard will be based on several important infrastructure pieces that DataCite has built over the past few years, and that are again briefly described below.

Image

DOI registration services

In the MDC initiative we track data citations in the scholarly literature, focussing on datasets registered with DataCite and publications registered with Crossref.

Event Data

We use the joint Crossref/DataCite Event Data service to exchange information about connections between publications and datasets, contributed via Crossref and DataCite members through the metadata they register. These connections are also made available via a Scholix-compliant REST API. In the previous MDC project, the Event Data service was expanded to include data usage stats and make retrieving information easier for DataCite members.

Data Usage Reports

DataCite members and repositories upload monthly reports about data usage to DataCite using a standard format (COUNTER Code of Practice for Research Data Usage Metrics and protocol (SUSHI). COUNTER Code of Practice for Research Data Usage Metrics and the DataCite usage reports API were developed in the previous MDC project.

GraphQL API

The DataCite GraphQL API [] built in the EC-funded FREYA project brings together all of the above information in a single API that supports the complex queries typically needed for retrieving aggregated data citation information.

Jupyter Notebooks

We use Jupyter notebooks to analyze and visualize the information made available in the GraphQL API [], and have developed documentation, demos, and training material with our partners in the FREYA project.

Common DOI Search

Common DOI Search, a service currently under development by DataCite with help from Crossref and others in the FREYA project, and with a first version planned to be released in August, will bring a single search interface for all scholarly DOIs, no matter from which DOI registration agency (DataCite, Crossref, etc.). All DOIs in Common DOI Search are in a single Elasticsearch search cluster using the same DataCite metadata schema.

Data Metrics Badge

The Data Metrics Badge – developed as part of the Parsec project – is an easy to installJavascript widget that displays up-to-date citations, views, and downloads for a single DOI, and links to the DataCite Search page for more detailed information.

Researcher Profile

Also as part of the PARSEC project, we have built the Researcher Profile that, using the researcher's ORCID ID, brings all academic outputs and their metrics for a given researcher into a single dashboard. This work serves as a blueprint for other aggregations (e.g. by research organization) in the bibliometrics dashboard.

Bibliometrics Dashboard

All the services described above are required building blocks for the bibliometrics dashboard we will start working on in August. What the dashboard will add is better insights into the data citation data we have collected, primarily helping the bibliometricians in the project, but also available to other users. We will use Jupyter notebooks for exploratory analyses and to address very specific research questions, and data visualizations in the bibliometrics dashboard that address the most common questions, such as the growth of data citations over time.

The bibliometrics dashboard will expand the common DOI search service that we are currently building, beyond FREYA, which ends in November. Common DOI search, and also the bibliometrics dashboard, are built using React, not only the most popular Javascript framework right now, but also integrating very nicely with GraphQL APIs. More specifically we are using next.js to run react on the server, helping with faster page loading and search engine optimization (SEO).

We have picked the popular Vega library for our data visualizations. Vega is not only widely used and very flexible, but also available in versions for Jupyter notebooks (Altair) and React (React-Vega).

Using the Bibliometrics Dashboard

DataCite members and the repositories they work with contribute to the bibliometrics dashboard in important ways, registering content with a DOI and standard metadata facilitating citation, inclusion of references in the metadata, and submission of data repository usage stats. The bibliometrics dashboard will increase our understanding of data citation and data usage stats through the bibliometrics work, but will also provide aggregations of information of interest to our members – for example data citations and data usage over time, by discipline or by repository – not available before. This information is displayed in the bibliometrics dashboard, and available via Jupyter notebooks and the GraphQL API.

References

The post The DataCite MDC Stack appeared first on DataCite.