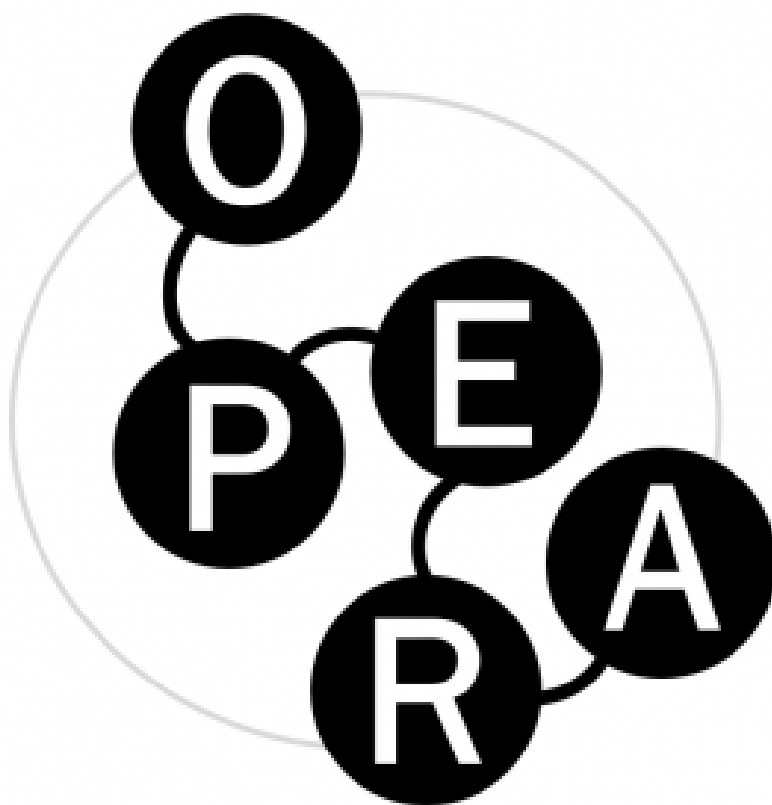


# **REVIEW OF EXISTING AND PROPOSED INDICATORS FOR OPEN SCIENCE ACTIVITIES**



**OPERA Project Report no. 1**

**November 4<sup>th</sup> 2020**

## **Contributor credits**

Specified using the [CRediT Taxonomy](#) contributor roles:

### **Conceptualization**

Ideas; formulation or evolution of overarching research goals and aims

### **Methodology**

Development or design of methodology; creation of models

### **Software**

Programming, software development; designing computer programs; implementation of the computer code and supporting algorithms; testing of existing code components

### **Validation**

Verification, whether as a part of the activity or separate, of the overall replication/ reproducibility of results/experiments and other research outputs

### **Formal Analysis**

Application of statistical, mathematical, computational, or other formal techniques to analyze or synthesize study data

### **Investigation**

Conducting a research and investigation process, specifically performing the experiments, or data/evidence collection

### **Resources**

Provision of study materials, reagents, materials, patients, laboratory samples, animals, instrumentation, computing resources, or other analysis tools

### **Data Curation**

Management activities to annotate (produce metadata), scrub data and maintain research data (including software code, where it is necessary for interpreting the data itself) for initial use and later reuse

### **Writing – Original Draft**

Preparation, creation and/or presentation of the published work, specifically writing the initial draft (including substantive translation)

### **Writing – Review & Editing**

Preparation, creation and/or presentation of the published work by those from the original research group, specifically critical review, commentary or revision – including pre- or postpublication stages

### **Visualization**

Preparation, creation and/or presentation of the published work, specifically visualization/ data presentation

### **Supervision**

Oversight and leadership responsibility for the research activity planning and execution, including mentorship external to the core team

### **Project Administration**

Management and coordination responsibility for the research activity planning and execution

### **Funding Acquisition**

Acquisition of the financial support for the project leading to this publication.

## **Contributors and roles**

**Birger Larsen** – <https://orcid.org/0000-0002-3622-2698> - [birger@hum.aau.dk](mailto:birger@hum.aau.dk)

ROLES: Conceptualization, Methodology, Investigation, Writing – Original Draft, Supervision

AFFILIATION: Aalborg University, Department of Communication and Psychology,

A. C. Meyers Vænge 15, 2450 København S, Denmark

# TABLE OF CONTENTS

<b>1</b>	<b>INTRODUCTION.....</b>	<b>1</b>
1.1	Executive summary .....	1
<b>2</b>	<b>DATA SHARING AND DATA CITATION ACROSS PLATFORMS.....</b>	<b>1</b>
2.1	NASA Open Data Portal .....	2
2.2	Global Biodiversity Information Facility (GBIF) .....	5
2.3	Mendeley Data .....	10
2.4	Google Dataset Search (beta version).....	13
<b>3</b>	<b>DATA SHARING AND OPEN SCIENCE INDICATORS .....</b>	<b>13</b>
3.1	Data Usage Index indicators (DUI).....	13
3.2	Altmetrics .....	15
<b>4</b>	<b>DISCUSSION AND CONCLUSION.....</b>	<b>23</b>
<b>5</b>	<b>ACKNOWLEDGEMENTS.....</b>	<b>25</b>
<b>6</b>	<b>REFERENCES .....</b>	<b>26</b>

# 1 INTRODUCTION

There is growing and widespread support for the Open Science movement across scientific fields. Manifestos like the *Amsterdam Call for Action on Open Science*<sup>1</sup> advocates for "full open access for all scientific publications", and endorses an environment where "data sharing and stewardship is the default approach for all publicly funded research", and the *FAIR Guiding Principles for Open Data*<sup>2</sup> stipulates that research data should be "Findable, Accessible, Interoperable and Reusable".

We aim to find and evaluate ways Open Science efforts may form part of research analytics, metrics and evaluation - and thus prepare the inclusion of some of these in analytics platforms and to contribute with practical experience and knowledge building in handling FAIR principles.

In this review we examine existing and proposed indicators for Open Science activities with a focus on data sharing in fields that have a long tradition for Open Data. We aim to select the most relevant and promising indicators for inclusion in Research Analytics Platforms and Research Information Systems.

**We first examine in Section 2 examples of platforms that facilitate data sharing and data citation. In Section 3 we analyse two examples of data citation and Open Data indicators – the suite of Data Usage Index (DUI) indicators proposed by Ingwersen and Chavan (2011) and the potentials of using altmetrics on datasets.**

## 1.1 EXECUTIVE SUMMARY

This report analyses examples of platforms that facilitates data sharing and data citation as well as examples of proposed data Open Data indicators.

# 2 DATA SHARING AND DATA CITATION ACROSS PLATFORMS

A prerequisite for making data sharing visible is an understanding how agencies, organisations, platforms and repositories facilitate data sharing, either as part of the Open Sciences movement

---

<sup>1</sup> <https://www.government.nl/documents/reports/2016/04/04/amsterdam-call-for-action-on-open-science>

<sup>2</sup> <https://www.nature.com/articles/sdata201618>

or as part of the traditions within their field. We therefore examine central examples of how existing data portals operate and how data sharing and data citation is facilitated in them.

Physics, astronomy, space and environment research are all datacentric fields of research. The *National Aeronautics and Space Administration (NASA)* was chosen as a representative of how research data are shared between researchers in a multifaceted scientific community. The *Global Biodiversity Information Facility (GBIF)* was selected because it illustrates how data collected by researchers across the world are created and shared in order to understand nature, and as it is a good example of the needs for standardisation of datasets and data citation practices. *Mendeley data* is a new initiative from Elsevier creating a data repository connected to their existing publishing and library platform. *Google Dataset Search (beta)* utilises the Google search engine to identify datasets across the web and the different existing data depositories making these datasets accessible from a single-entry point.

## 2.1 NASA OPEN DATA PORTAL

Starting from a White House Open Data Policy memorandum<sup>3</sup>, the NASA agency has developed an Open Data Portal providing access to publicly available datasets across NASA<sup>4</sup>. The exact number of datasets is not given but is stated as “tens of thousands”. The portal aggregates metadata of datasets and other open resources such as code across NASA organisations with standardised metadata with statistics of dataset views and downloads through the portal<sup>5</sup> - see Figure 1 and Figure 2. The exact metadata and options available depend on the individual NASA organisation. Some of the organisations include DOI and instructs any users of the datasets in how to correctly cite the dataset – see Figure 3.

**The NASA Open Data Portal is an interesting example of attempts to comply with FAIR principles. The number of dataset views and downloads as well as instructions on how to cite the datasets are interesting for the OPERA project. As an aggregator the NASA Open Data Portal is however dependent on the individual NASA organisation that provides datasets for consistent metadata and adherence to portal standards. This can cause problems with consistency and missing data – e.g. not all examined datasets have a DOI or instruct users how to cite datasets.**

---

<sup>3</sup> <https://obamawhitehouse.archives.gov/sites/default/files/omb/memoranda/2013/m-13-13.pdf>

<sup>4</sup> <http://data.nasa.gov>

<sup>5</sup> The number of downloads has been zero on all inspected datasets. This probably due to the fact that dataset information is aggregated through the portal whereas downloading takes place through the data centre at each individual NASA organisation.

The screenshot shows the NASA Open Data Portal search results for the query 'greenland'. The page displays 162 results, sorted by 'Most Accessed'. Two results are visible:

Category	Title	Description	Updated	Views
Earth Science	Land Ice: Greenland & Antarctic ice mass anomaly	Data from NASA's Grace satellites show that the land ice sheets in both Antarctica and Greenland are losing mass. The continent of Antarctica (left chart) has been losing more than 100 cubic kilometers (24 cubic miles) of ice per year since 2002.	January 29, 2020	54
Earth Science	Monthly snow/ice averages (ISCCP)	September Arctic sea ice is now declining at a rate of 11.5 percent per decade, relative to the 1979 to 2000 average. Data from NASA show that the land ice sheets in both Antarctica and Greenland are losing mass.	January 29, 2020	37

Figure 1: NASA Open Data Portal, search results for 'greenland', including number of dataset views.

The screenshot shows the NASA Open Data Portal interface. At the top, there is a navigation bar with the NASA logo, the text 'NASA's Open Data Portal', and links for 'Data Catalog', 'About', and 'Developer Resources'. Social media icons for Facebook, GitHub, Twitter, YouTube, Instagram, and Tumblr are also present, along with a search icon and a 'Sign In' button.

The main content area features a title 'Land Ice: Greenland & Antarctic ice mass anomaly' with a three-dot menu icon to its right. Below the title is a descriptive paragraph: 'Data from NASA's Grace satellites show that the land ice sheets in both Antarctica and Greenland are losing mass. The continent of Antarctica (left chart) has been losing more than 100 cubic kilometers (24 cubic miles) of ice per year since 2002.' To the right of this text, it says 'Updated January 29, 2020'.

Under the heading 'Access this Data', there is a button labeled 'TEXT/HTML'.

The 'About this Dataset' section is divided into two columns. The left column contains:
 

- Updated: **January 29, 2020**
- Metadata Last Updated: January 29, 2020
- Date Created: June 25, 2018
- Views: **54**
- Downloads: **0**
- Data Provided by: (none)
- Dataset Owner: NASA Open Data
- A blue button labeled 'Contact Dataset Owner'.

The right column contains 'Common Core' metadata:
 

- Publisher: National Aeronautics and Space Administration
- Contact Name: Felix Landerer
- Contact Email: <mailto:felix.w.landerer@jpl.nasa.gov>
- Public Access Level: public
- Geographic Coverage: regional
- Temporal Applicability: 2002-01-01/2013-01-01
- Update Frequency: irregular
- License: <http://www.usa.gov/publicdomain/label/1.0/>
- Unique Identifier: NASA-0000043

Below this is 'NASA Custom Metadata':
 

- Dataset Identifier: NASA-0000043

A 'Show More' link is located at the bottom right of the metadata section.

Figure 2: NASA Open Data Portal, example dataset, including number of dataset views and downloads.



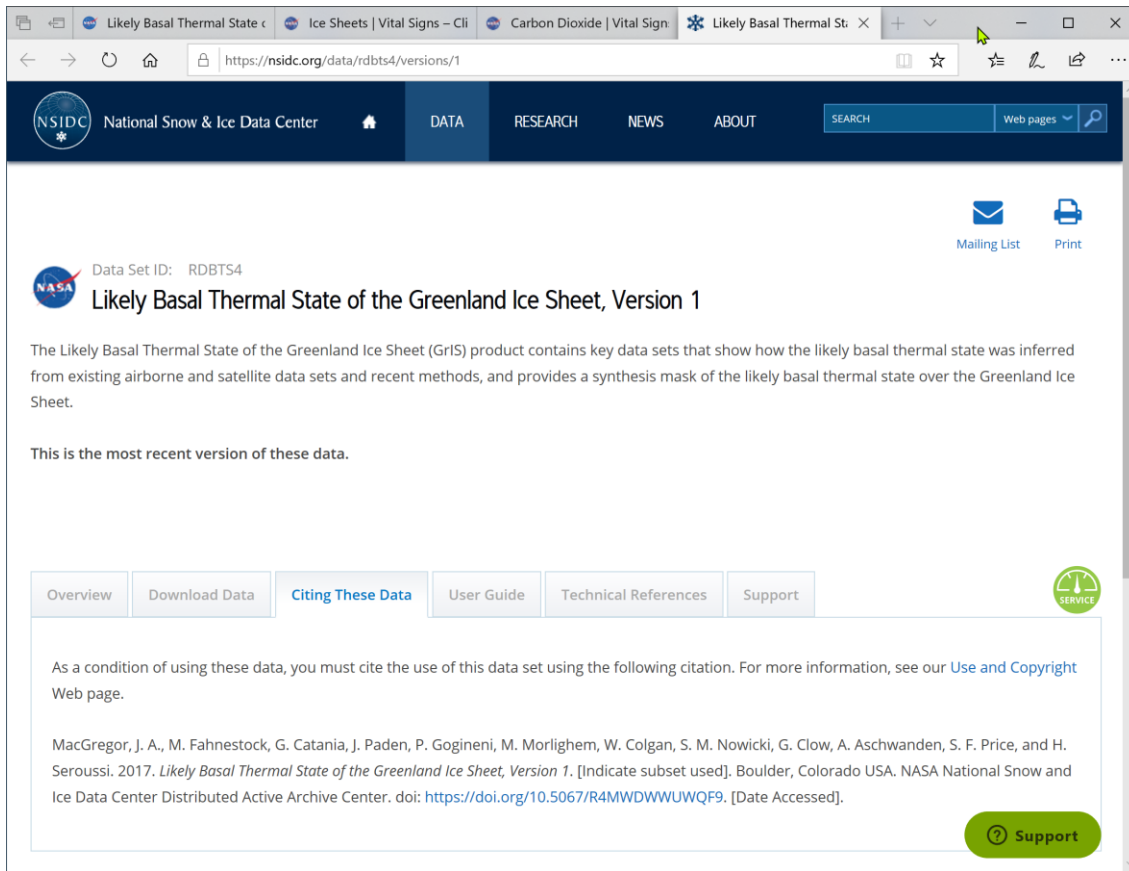


Figure 3: Example Dataset linked from NASA Open Data Portal at a NASA organisation - the National Snow & Ice Data Center). Includes instructions on how to cite the dataset including DOI.

## 2.2 GLOBAL BIODIVERSITY INFORMATION FACILITY (GBIF)

GBIF - the Global Biodiversity Information Facility – was established in 2001 based on an OECD memorandum of understanding<sup>6</sup>. GBIF is an international network and research infrastructure funded by the world's governments and aimed at providing anyone, anywhere, open access to data about all types of life on Earth. As such the GBIF repository was created so that the knowledge for the natural world could expand and dissemination in a manner that avoids duplication of effort and expenditure. GBIF acts as coordinator and provides institutions with the common standards and open-source tools which enable participants to engage with the natural scientific community. A typical dataset consists of counts of some species in certain locations<sup>7</sup>. The current number of datasets can be seen in the GBIF search engine: at the time of writing a total of 52,434 datasets, including 19,427 occurrence datasets, 31,237 checklist

<sup>6</sup> <https://www.gbif.org/what-is-gbif>

<sup>7</sup> See e.g. the "Great British Bee Count 2018 verified data" at <https://www.gbif.org/dataset/f794b231-42de-4008-ba8e-809e01ee7785>

datasets, 1,457 sampling events and 303 metadata datasets<sup>8</sup>. GBIF itself is more interested in the number of species included its data – which cannot easily be counted as a single number but lies somewhere between 1 and 2.3 million<sup>9</sup>. Also of interest is the number of occurrences of species, which is more than 1.4 billion in GBIF at present.

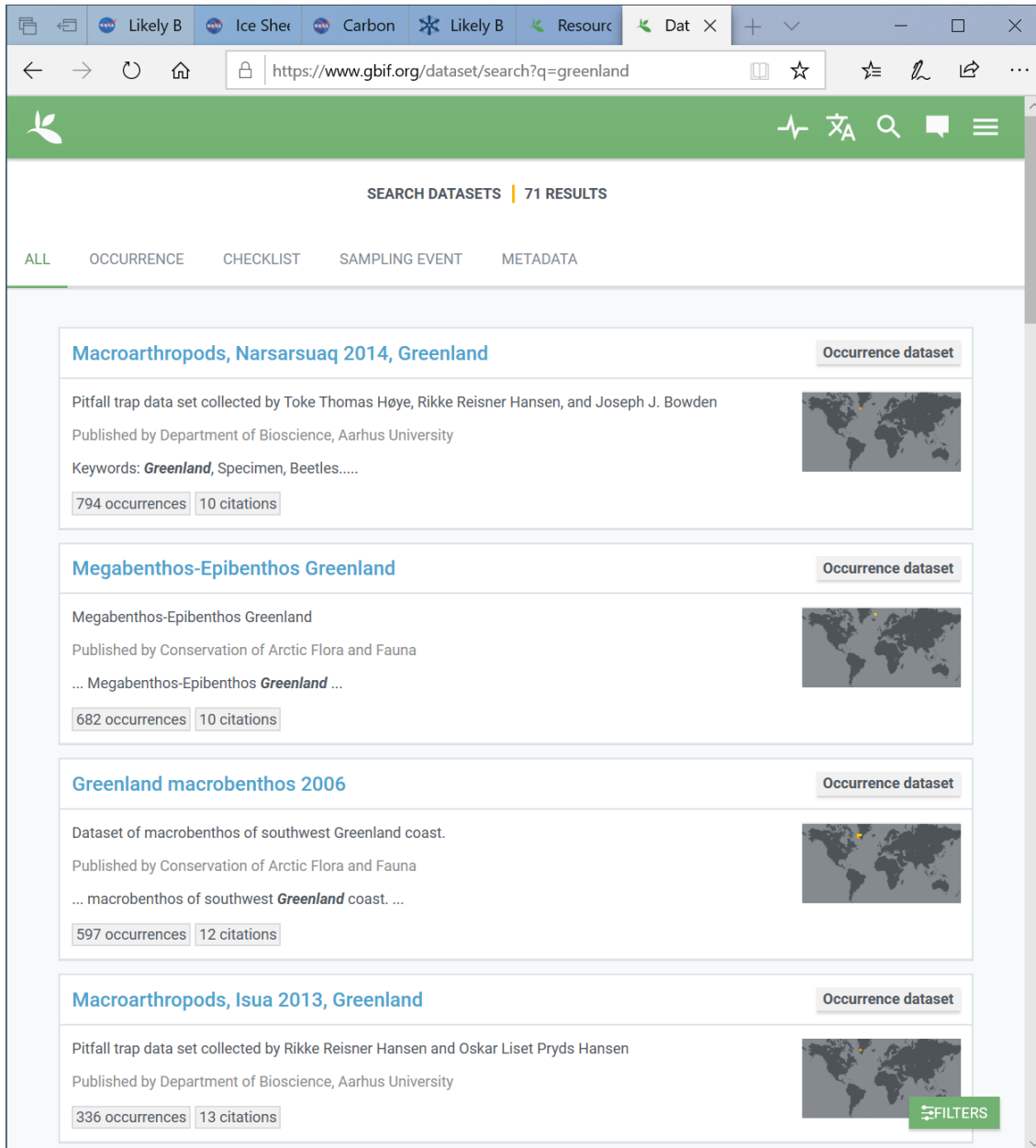


Figure 4: GBIF example dataset search results – including number of citing publications.

<sup>8</sup> Retrieved April 29, 2020 from <https://www.gbif.org/dataset/search>

<sup>9</sup> <https://www.gbif.org/about-species-counts>

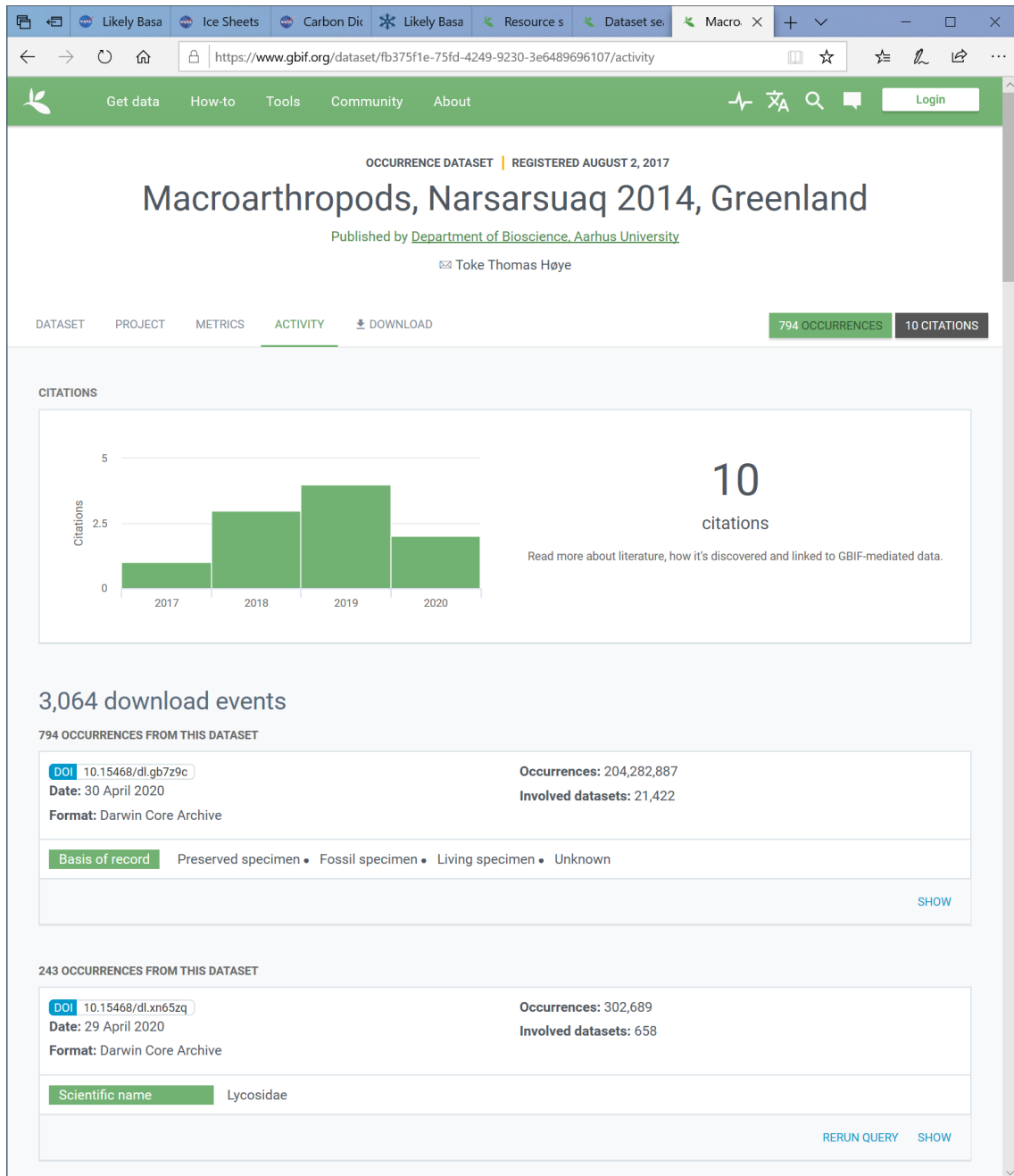


Figure 5: GBIF dataset example – with citation and download details. The dataset has 794 occurrences – in some cases all were included in the 3,064 download events, in other cases only some of the occurrences.

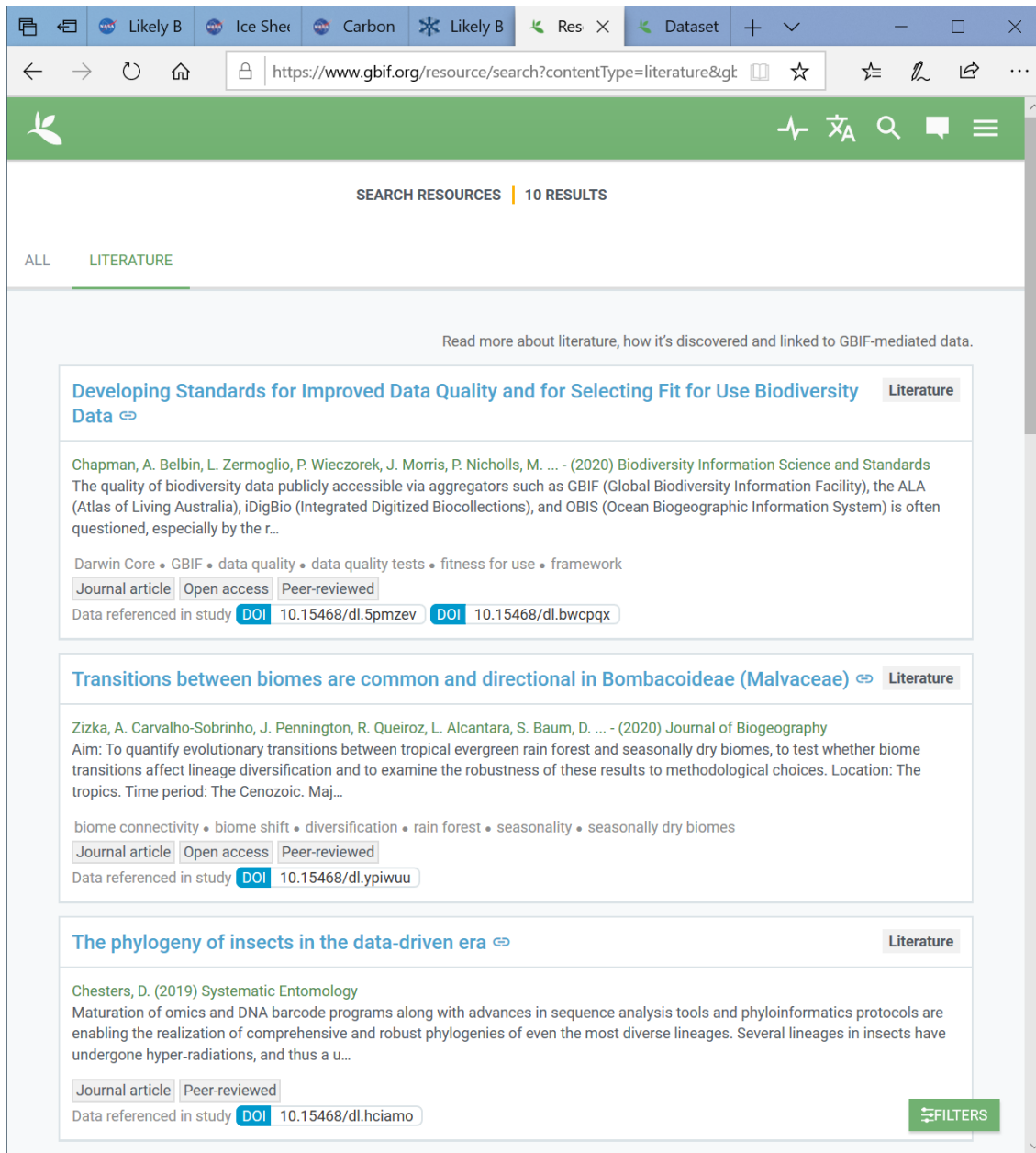


Figure 6: Example of metadata of publications citing a GBIF dataset. Where possible publications are linked to external fulltexts.

GBIF requires users who download individual datasets or search results and use them in research or policy to cite them using a DOI. Detailed citation guidelines are provided<sup>10</sup>, including instructions for how to cite downloads with multiple datasets, individual datasets, datasets accessed through third-party tools (such as python or R), as well as custom datasets exports. Users must be registered to download. To aid users an email with dataset specific citation

<sup>10</sup> <https://www.gbif.org/citation-guidelines>

instructions is sent every time a dataset is downloaded, and a list of all downloaded datasets are listed in each user's profile to further aid correct citation. Note that downloads often consist of data selected from multiple datasets, e.g. someone interested in bumblebees (*genus Bombus*) would get results for the over 250 species of bumblebee from datasets that include these. Such downloads with selected data from multiple datasets are assigned their own unique DOI. Figure 5 shows an example: A dataset uploaded in 2017 with 794 data occurrences on 'Macroarthropods, Narsarsuaq 2014, Greenland' has 10 citations in the literature and data from the dataset has been included in 3,064 download events – in some all of the 794 occurrences were included, in other only some – depending on the data requested.

GBIF also actively searches for research uses and citations of biodiversity information accessed through GBIF's global infrastructure<sup>11</sup>. Daily searches are carried out in Google Scholar, Scopus, Wiley Online Library, SpringerLink, NCBI Pubmed and bioRxiv, and the results are curated and added to a database from which citation statistics can be extracted. These are shown on the main <http://gbif.org> search page when searching for datasets (Figure 4) with details available on each dataset page (Figure 5) and can also be searched directly<sup>12</sup>.

**GBIF is an interesting example of an initiative to build an advanced portal to provide open access to an important datatype across the world. It has consistent standards, good support and seems to have strong backing and funding. GBIF has policies and support for correctly citing datasets including automatic assignments of DOIs including for custom downloads with data from multiple datasets. In addition, they are actively seeking and registering any citing publications that use GBIF datasets. GBIF is therefore is strongly placed for creating a culture of data citation within the field and is collecting data that can support advanced analysis of data usage. This data might pave the way for dataset creation and sharing becoming part of reward mechanisms. However, as a recent study by Kahn, Thellwall and Koucha (2019) shows, a best practise for data citation is yet to be established, with e.g. different practices across different journals, making it hard to ensure comprehensive data on data usage. GBIF also demonstrates the complexity of counting and analysing downloads and data usage when partial downloads of datasets is supported. In Section 3 we will examine the types of indicators that might be appropriate on such usage data as proposed by Ingwersen and Wishwas (2011).**

---

<sup>11</sup> <https://www.gbif.org/literature-tracking>

<sup>12</sup> See <https://www.gbif.org/resource/search>. At present more than 10,000 such citing resources are listed including 9,276 scientific publications.

### 2.3 MENDELEY DATA

Mendeley Data was announced in late 2018<sup>13</sup>. The Mendeley Data portal<sup>14</sup> imports a variety of data from different data depositories, journals and archives - and allows registered users to archive their own data. In addition to actual datasets, the portal also contains images, video, audio, software many of them extracted from articles. Currently Mendeley Data contains more than 21 million items. Table 1 shows the top-50 sources of these as well as the top-50 sources of the approximately 10 million datasets. For the datasets the ScienceDirect platform is a major provider; medium-sized providers include archives from various research fields, research institutes, figshare and arxiv.org.

Self-archived datasets are reviewed before they can be published on the platform – upon which they are assigned a DOI. Mendeley Data has version control, provides instructions on how to cite the data and shows the number of views and downloads through the platform – see Figure 7. Mendeley Data also contributes to Scholix, the Framework for Scholarly Link Exchange<sup>15</sup>, which creates an open global information ecosystem to collect and exchange links between research data and literature, as well as DataCite’s metadata index<sup>16</sup> (a comprehensive research datasets metadata index) and to the OpenAIRE portal<sup>17</sup>, the EU’s research portal which aims to make as much European-funded research output as possible available to all.

**Mendeley Data is a major effort from the Elsevier group. Resources have been put into the identification of possible sources, importing from these and to set up a platform with review procedures for self-archived data and consistent metadata. With more than 10 million datasets indexed and the possibility to self-archive datasets it is a major platform. Similarly to NASA, Mendeley Data instructs authors how to cite a dataset correctly, and shows the number of dataset view and downloads. Interestingly, despite the resources at Elsevier and Mendeley there was no indication at an attempt to identify or show the number of dataset citations in Mendeley Data.**

---

<sup>13</sup> <https://data.mendeley.com/faq>

<sup>14</sup> <https://data.mendeley.com/research-data/>

<sup>15</sup> <http://www.scholix.org/>

<sup>16</sup> <https://search.datacite.org/>

<sup>17</sup> <https://explore.openaire.eu/search/find/datasets>

**21,292,064 data items**  
**- total Mendeley Data content**

**10,511,094 data items**  
**- tabular Data, Datasets, Geospatial and Sequencing Data only**

<b>Top-50 Sources</b>		<b>Top-50 Sources</b>	
ScienceDirect	4,863,624 (23%)	ScienceDirect	3,325,890 (32%)
figshare Academic Research System	1,394,185 (7%)	USGS Mineral Res.	991,210 (9%)
Zenodo	1,383,381 (6%)	The Cambridge Structural Database	834,613 (8%)
USGS Mineral Res.	995,208 (5%)	Plutof. Data Manag.t & Publishing Platform	686,105 (7%)
Intl. Treaty on Plant Genetic Resources for Food & Agricult.	900,590 (4%)	GEOROC	476,421 (5%)
The Cambridge Structural Database	834,619 (4%)	Uni. of Southern California Digital Library	454,763 (4%)
E-Periodica	802,899 (4%)	DSMZ	399,956 (4%)
Plutof. Data Management and Publishing Platform	686,136 (3%)	PANGAEA	384,453 (4%)
arXiv	555,851 (3%)	ClinVar	376,028 (4%)
E-Pics Bildarchiv	523,628 (2%)	figshare Academic Research System	354,059 (3%)
GEOROC	478,309 (2%)	WSL Landesforstinventar	228,420 (2%)
University of Southern California Digital Library	454,764 (2%)	arXiv	184,609 (2%)
DSMZ	399,956 (2%)	NRCT Data Center	143,306 (1.4%)
PANGAEA	393,270 (2%)	RCSB-PDB	138,270 (1.3%)
ClinVar	376,028 (2%)	Leibniz-Institut für Astrophysik Potsdam (AIP)	134,789 (1.3%)
Columbia University Libraries	326,600 (2%)	Pitt Quantum Repository	106,099 (1.0%)
Data Planet	323,992 (2%)	UC Santa Barbara	104,016 (1.0%)
Apollo Cambridge	238,189 (1.1%)	NeuroMorpho	86,888 (0.8%)
WSL Landesforstinventar	228,420 (1.1%)	PetDB	84,194 (0.8%)
Leibniz-Institut für Astrophysik Potsdam (AIP)	219,880 (1.0%)	Environmental Data Initiative	72,795 (0.7%)
IPK Gatersleben	200,264 (0.9%)	Lawrence Berkeley National Laboratory (LBNL)	72,180 (0.7%)
University of British Columbia	174,853 (0.8%)	UCD James Joyce Library	71,740 (0.7%)
NRCT Data Center	163,146 (0.8%)	Zenodo	56,358 (0.5%)
AgEcon Search	139,472 (0.7%)	NAVDAT	53,522 (0.5%)
RCSB-PDB	138,270 (0.6%)	Biodiversity Institute of Ontario	50,332 (0.5%)
Pitt Quantum Repository	106,099 (0.5%)	UC San Diego	44,506 (0.4%)
UC Santa Barbara	104,339 (0.5%)	TOPMed	42,379 (0.4%)
Gene Expression Omnibus	97,540 (0.5%)	DANS	38,663 (0.4%)
Incorporated Research Institutions for Seismology	89,556 (0.4%)	ICPSR	36,096 (0.3%)
NeuroMorpho	86,892 (0.4%)	Neotoma Paleoeological Database	30,831 (0.3%)
Universität Zürich, ZORA	86,027 (0.4%)	Mendeley Data	29,395 (0.3%)
PetDB	84,938 (0.4%)	Dryad	29,223 (0.3%)
e-manuscripta	83,543 (0.4%)	ArrayExpress	26,980 (0.3%)
e-rara.ch	78,270 (0.4%)	GTEx	26,802 (0.3%)
DataSpace Princeton	75,889 (0.4%)	NeuroElectro	22,726 (0.2%)
Environmental Data Initiative	72,795 (0.3%)	Oxford University Library Service Databank	16,458 (0.2%)
Lawrence Berkeley National Laboratory (LBNL)	72,183 (0.3%)	Caltech High Throughput Experimentation	14,670 (0.1%)
E-Pics 4, Biosys	71,752 (0.3%)	Strasbourg Astronomical Data Center	12,609 (0.1%)
UCD James Joyce Library	71,740 (0.3%)	Biological Magnetic Resonance Data Bank	12,178 (0.1%)
ArrayExpress	70,654 (0.3%)	Harvard Dataverse	11,181 (0.1%)
figshare SAGE Publications	69,775 (0.3%)	ICPSR	10,514 (0.1%)
NAVDAT	66,781 (0.3%)	University of York	10,112 (0.1%)
DANS	65,621 (0.3%)	ThermoML NIST TRC	9,367 (0.1%)
University of Texas Libraries	64,954 (0.3%)	Digital CSIC	8,384 (0.1%)
TIB Hannover	63,016 (0.3%)	Prior Art Publishing GmbH	8,182 (0.1%)
University of Alberta Libraries	56,764 (0.3%)	Incorp. Research Institutions for Seismology	7,897 (0.1%)
ETH Zürich Research Collection	54,266 (0.3%)	Statistics Canada	7,817 (0.1%)
Biodiversity Institute of Ontario	50,332 (0.2%)	Geoscience Australia	7,521 (0.1%)
Michigan State University Libraries	48,526 (0.2%)	EC Joint Research Centre Directorate G	7,321 (0.1%)
UC San Diego	48,418 (0.2%)	UK Data Archive	6,982 (0.1%)
<b>TOTAL top-50</b>	<b>19,036,204 (89%)</b>	<b>TOTAL top-50</b>	<b>10,349,810 (98%)</b>

Table 1. Top sources in Mendeley Data – all and dataset specific ones.

# Study of the availability of nitrogen, carbon, and phosphorus in a blend of agro-industrial digestate and wood ashes under different acidification conditions

Published: 12 May 2020 | **Version 1** | DOI: 10.17632/n4c8d77224.1





Contributor(s): [Alejandro Moure Abelenda](#), [Kirk Semple](#), [Alfonso Lag Brotons](#), [Ben Herbert](#), [George Aggidis](#), [Farid Aiouache](#)

## Description of this data

Datasets and Supplementary Materials

## Experiment data files

[Download all files \(4\)](#)

	Conditions of the experiments carried out to trap the ammonia.docx	27 KB	<a href="#">Cite</a>	<a href="#">↓</a>
	Doses of the acids used in the treatment of slurry and manure.docx	35 KB	<a href="#">Cite</a>	<a href="#">↓</a>
	Further characterisation of the PVWD (determinations made by ... .docx	21 KB	<a href="#">Cite</a>	<a href="#">↓</a>
	Further characterisation of WBA.docx	21 KB	<a href="#">Cite</a>	<a href="#">↓</a>

## Latest version

**Version 1** 2020-05-12

Published: 2020-05-12

DOI: 10.17632/n4c8d77224.1

### Cite this dataset

Moure Abelenda, Alejandro; Semple, Kirk; Lag Brotons, Alfonso; Herbert, Ben; Aggidis, George; Aiouache, Farid (2020), "Study of the availability of nitrogen, carbon, and phosphorus in a blend of agro-industrial digestate and wood ashes under different acidification conditions", Mendeley Data, v1

<http://dx.doi.org/10.17632/n4c8d77224.1>

## Statistics

Views: **29** Downloads: **5**

## Institutions

Lancaster University

## Categories

Soil Science, Environmental Science, Air Quality, Waste, Anaerobic Digestion

## Licence

CC BY 4.0

[Learn more](#)

 Report

Figure 7. Example of dataset in Mendeley Data – with instructions on how to cite and statistics on views and downloads.



## 2.4 GOOGLE DATASET SEARCH (BETA VERSION)

Google Dataset Search<sup>18</sup> is a new dataset search function, providing access to datasets identified by Google on the open web. Datasets can be included if they have assigned correct schema.org metadata. Once metadata have been added, Google needs to be notified and the dataset metadata can be crawled. Google Dataset Search does not store the datasets themselves but acts as a platform that links to data providers. In case several providers provide access to the same dataset, Google attempts to deduplicate this and provides links from the dataset to all providers (Figure 15). In addition, if the dataset is cited in Google Scholar, the number of Google Scholar citations is shown (see Figure 8) - and links to an automatic Google Scholar search (Figure 16).

**Google Dataset Search is similar to other vertical Google search products in that Google aggregates a certain type of information crawled from the open web. A major difference is that correct metadata must be in place before content is crawled. While it is possible for individual scientists to add such metadata it is more likely that major archives and organisations will make the effort. The automated identification of citations to datasets from Google Scholar is interesting – although at present the citation counts do not appear to be updated frequently.**

## 3 DATA SHARING AND OPEN SCIENCE INDICATORS

Even though there are challenges in building strong dataset citation cultures and in identifying reliable statistics and view, downloads and datasets citations it is important to discuss which indicators might be useful in studying and visualising data sharing and open science activities.

We analyse two different examples of indicators below: those proposed by Ingwersen & Chavan (2011) for GBIF and those offered by Altmetrics.

### 3.1 DATA USAGE INDEX INDICATORS (DUI)

With GBIF being one of the major data portals giving open access to biodiversity data Ingwersen & Chavan (2011) aims to define useful indicators based on the usage of data from GBIF – inspired by traditional scientometric indicators. The overall goal is to encourage researchers to share their data by creating a mechanism that can ensure recognition for the effort put into

---

<sup>18</sup> <https://datasetsearch.research.google.com/>

creating and sharing data sets. Secondly, such indicators can also provide insights into the information behaviour of biodiversity scholars and their interaction with datasets.

It is worth noting that Ingwersen & Chavan (2011) do not use a citation-based approach because a data citation culture, standards and data citation indexes were not in existence at the time. As shown by Kahn, Thellwall and Koucha (2019) this is partially still the case - even for GBIF. Instead Ingwersen & Chavan base their Data Usage Index (DUI) indicators on interaction events associated to obtaining GBIF datasets: searches and downloads of datasets as registered in the GBIF platform. *Search events* are when data sets appear in a search result and are regarded as indication of *interest* in the data. *Download events* are when dataset or parts of them are downloaded as described in Section 2.2 – and as regarded as indication of *usage* of the data. The proposed indicators can be seen in Table 2. Indicators 1-3 are the basic units, e.g. #3 number of data records in a dataset. Indicators 4-6 are aggregated for use in relative indicators, e.g. #4 number of different downloads from a dataset. Indicators 7-14 are relative indicators that signify average interest and usage, e.g. #9 that indicates the average number of records in a dataset that has been downloaded.

Table 2. Basic Data Usage Index indicators for primary biodiversity data published through the GBIF network. From Ingwersen & Chavan (2011, p. 5).

#	Formula	Indicator	Description
1	$s(u)$	Searched records	Number of records searched/viewed (by IP address) in unit
2	$d(u)$	Download frequency	Number of downloaded records from unit
3	$r(u)$	Record number	Number of records in (period; dataset(s); geographical and/or species unit)
4	$S(u)$	Search events	Number of different searches (by IP address) in unit
5	$D(u)$	Download events	Number of different downloads from unit
6	$N(u)$	Dataset number	Number of datasets in (period, geographical and/or species unit)
7	$s(u)/S(u)$	Search density	Average number of searched records per search event
8	$d(u)/D(u)$	Download density	Average download frequency per download event
9	$d(u)/r(u)$	Usage impact	Download frequency per stored record per unit
10	$s(u)/r(u)$	Interest impact	Searched records per stored record per unit
11	$d(u)/s(u)$	Usage ratio	Ratio of download frequency to searched records in unit
12	$D(u)/S(u)$	Usage balance	Ratio of download events to search events for unit (in %)
13	$U(u)/r(u)$	Usage score	Ratio of unique downloaded records ( $U$ ) to record number (in %)
14	$I(u)/r(u)$	Interest score	Ratio of unique searched records ( $I$ ) to record number (in %)

Ingwersen and Chavan provide data for these indicators for a sample number of dataset and data providers. Conclusions that can be drawn from the application of the indicators include that there are many searches for data, but few of them lead to downloads of data (between 1-2% in the examples). Ingwersen and Chavan then go on to propose a number of additional relative and weighted relative indicators. These are similar to the well-known crown indicators

in bibliometrics (REF), where the citation performance of e.g. a research group can be compared to the expected number of citations in the same research field or to national or world averages.

**In summary, the Data Usage Indicators proposed by Ingwersen & Chavan (2011) provide a wide range of indicators for the interest (= appearance in dataset searches) and usage (= download of data), both for parts of datasets, whole dataset and dataset providers. The indicators rely on the solid data produced by the GBIF platform and its particularities and for some indicator, calculation will only be possible if similar data is available. Many of the indicators are however based on searches and downloads – data that is available in many platforms as shown in Section 1.**

### 3.2 ALTMETRICS

*Altmetrics* (from ‘alternative metrics’) are metrics that go beyond classical citation data and illustrate how scientific output, including datasets, are cited or mentioned outside the academic literature - mainly on internet platforms<sup>19</sup>. A well-known example is altmetrics.com where mentions on a wide range of social media and other internet-based platforms are aggregated and visualised (see Figure 8). Similar data form part of PlumX Metrics, now part of Elsevier<sup>20</sup>. Output from altmetrics.com includes a ‘donut badge’ where different colours indicate different types of sources, and an ‘Altmetric Attention Score’, a weighted indicator across mentions that gives higher weights to some types of sources, e.g. news outlets, blog posts, or policy documents<sup>21</sup>. Altmetric mentions are harvested by various identifiers and DOIs looking for references to academic work.

---

<sup>19</sup> See the Altmetrics Manifesto: <http://altmetrics.org/manifesto/>

<sup>20</sup> See <https://plumanalytics.com/>

<sup>21</sup> For details about the calculation Altmetric Attention Scores see:

<https://help.altmetric.com/support/solutions/articles/6000060969-how-is-the-altmetric-attention-score-calculated->

The colors of the Altmetric donut each represent a different source of attention:

### The Colors of the Donut

- |                                 |                               |
|---------------------------------|-------------------------------|
| ● Policy documents              | ● Google+                     |
| ● News                          | ● LinkedIn                    |
| ● Blogs                         | ● Reddit                      |
| ● Twitter                       | ● Research highlight platform |
| ● Post-publication peer-reviews | ● Q&A (Stack Overflow)        |
| ● Facebook                      | ● Youtube                     |
| ● Sina Weibo                    | ● Pinterest                   |
| ● Syllabi                       | ● Patents                     |
| ● Wikipedia                     |                               |



*Figure 8: Illustration of Altmetric Donut and Attention score with explanation of donut colours. From altmetric.com.*

altmetric.com is at present tracking some 38,000 datasets, of which more than 33,000 have been mentioned in at least one of their sources. Figure 9 to Figure 11 illustrate datasets with different altmetric profiles: The dataset in Figure 11 mainly has mentions on twitter, the one in Figure 12 mainly mentions from news outlets, with the one in Figure 13 having a more balanced profile with mentions on twitter as well as other sources. Figure 12 to Figure 14 show examples of mentions in news outlets, twitter and facebook for this dataset.

**Altmetric data offers a different view on the impact of datasets. The wide variety of sources are probably instrumental in the efforts put into easily readable visualisations by providers such as altmetrics.com and PlumX Metrics. The providers seem to rely mainly on DOIs and similar IDs to identify mentions, which can be a challenge for fields that do not use these in their datasets and dataset citation culture. It is also of great value that many the sources are linked so that it is possible in one place to see e.g. which tweets or news outlets mention a dataset. However, it remains somewhat unclear which sources are being crawled and what the coverage of altmetric products are. Also, as for scientific publications, it is not clear to what degree altmetric scores for dataset can be gamed (Eysenbach, 2011).**

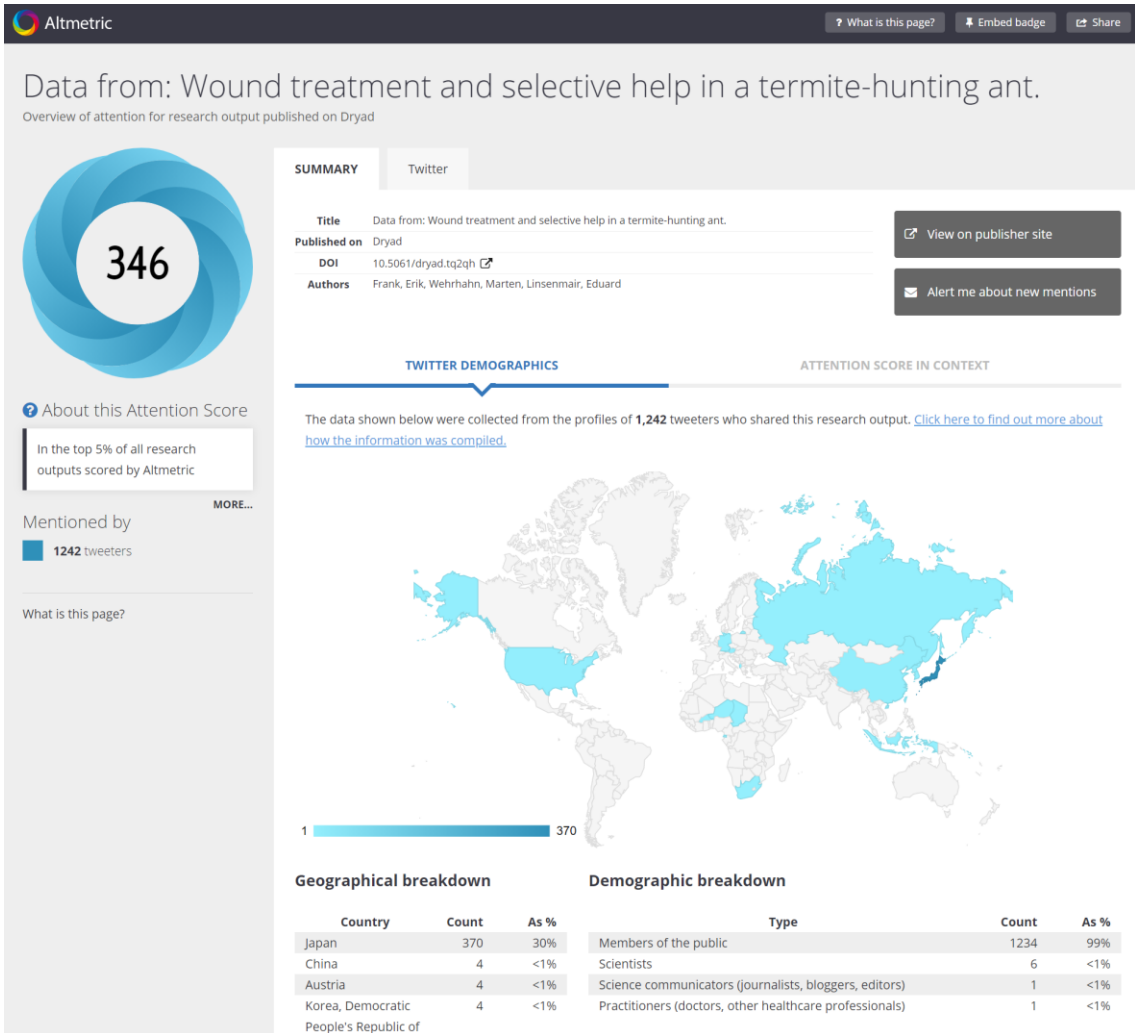


Figure 9. altmetric.com data on sample dataset, with a profile with mentions mainly on twitter (<https://www.altmetric.com/details/33183494>)



Figure 10. altmetric.com data on sample dataset (in the form of a figure), with a profile with mentions mainly from news outlets (<https://figshare.altmetric.com/details/75025367>).

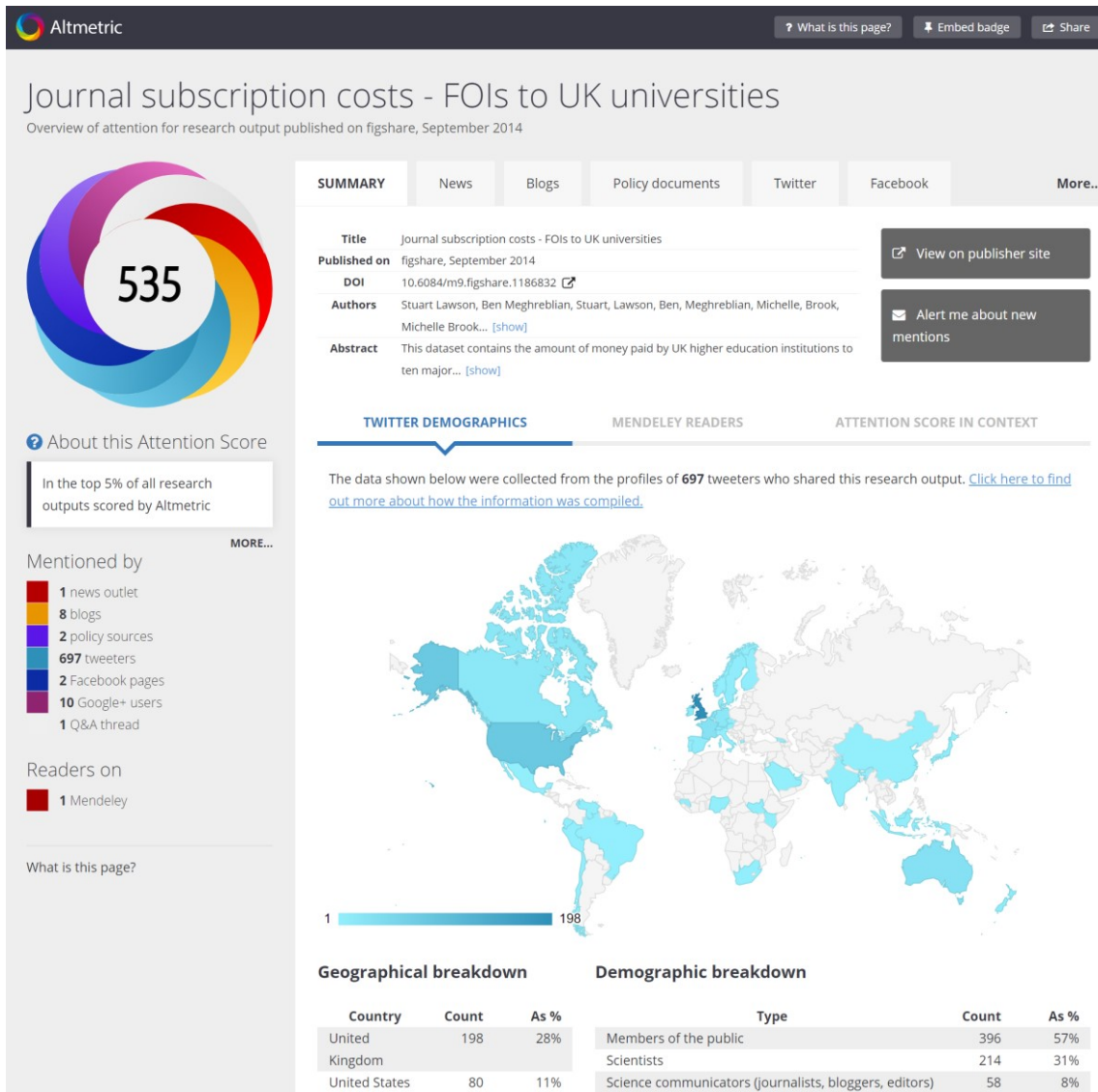


Figure 11. altmetric.com data on sample dataset, with a profile with mentions from several sources (<https://figshare.altmetric.com/details/2726745>)

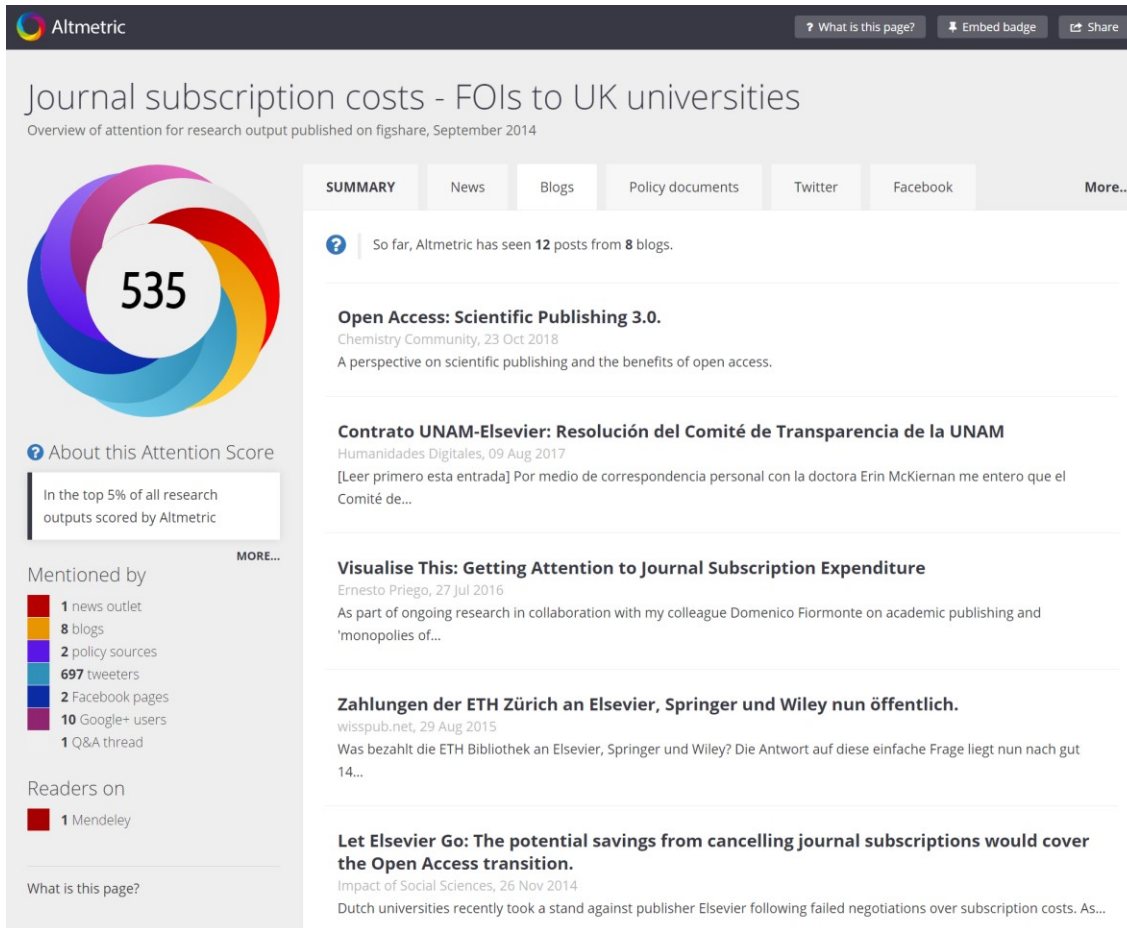


Figure 12. altmetric.com data on sample dataset, with examples of mentions from Blogs.



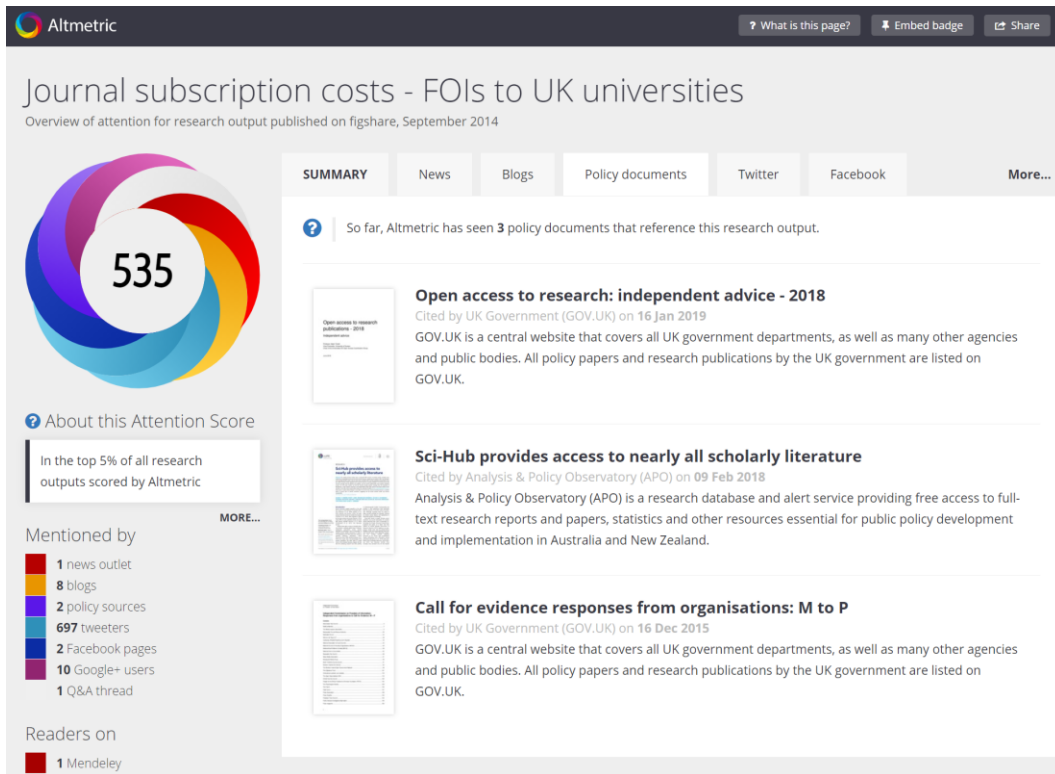


Figure 13. altmetric.com data on sample dataset, with examples of mentions in policy documents.

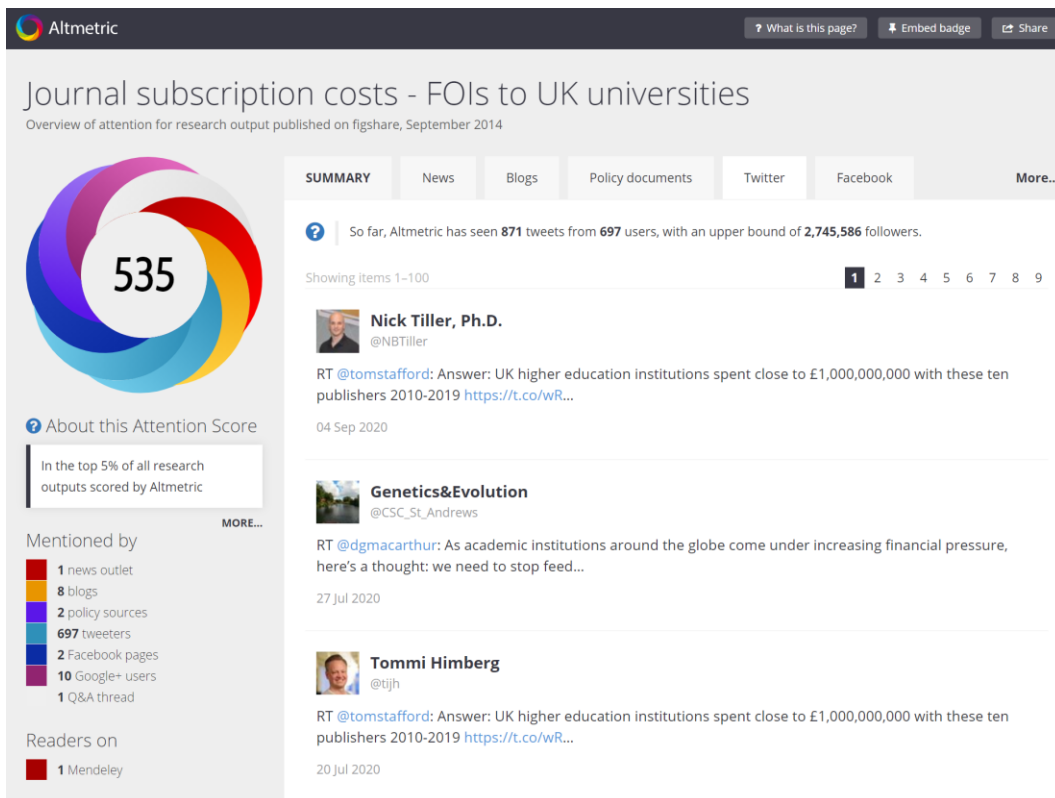


Figure 14. altmetric.com data on sample dataset, with examples of mentions on twitter.

datasetsearch.research.google.com/search?query=greenland&docid=wjfgT7t9zHZsu3WAAAA...

Google greenland

Updated Apr 22, 2020

**T** Greenland Internet Speed  
tradingeconomics.com  
Updated Oct 25, 2017

**statista** Emigration from Greenland 2019, by destination country  
www.statista.com  
Updated Apr 28, 2020

**T** Iceland exports from Greenland  
tradingeconomics.com  
Updated Jun 7, 2017

**P** Greenland geothermal heat flux distribution and estimated...  
doi.pangaea.de  
pangaea.figshare.com  
+1more  
tsv, html  
Updated Aug 13, 2018

**Google Earth Engine** Greenland Ice & Ocean Mask - Greenland Mapping Project...  
developers.google.com

**Google Earth Engine** 2000 Greenland Mosaic - Greenland Ice Mapping Project...  
developers.google.com

**statista** Most popular Facebook pages in Greenland 2020, by number...  
www.statista.com  
Updated Feb 28, 2020

**statista** Average number of employees in Greenland 2018, by industry

Greenland geothermal heat flux distribution and estimated Curie Depths, links to gridded files

Explore at PANGAEA Explore at pangaea.figshare.com  
Explore at search.datacite.org

2 scholarly articles cite this data set (View in Google Scholar)

tsv, html

**Unique identifier**  
<https://doi.org/10.1594/PANGAEA.892973>

**Data set updated** Aug 13, 2018

**Data set provided by**  
PANGAEA

**Authors**  
Yasmina M Martos

**Licence**  
[Attribution 3.0 \(CC BY 3.0\)](#)  
Licence information was derived automatically

**Area covered**

**Variables measured**  
File content, File format, File name, File size, Uniform resource locator/link to file

**Description**  
Curie depths beneath Greenland are revealed by spectral analysis of data from the World Digital Magnetic Anomaly Map2. A thermal model of the lithosphere then provides a corresponding geothermal heat flux map. This new map exhibits significantly higher frequency but lower amplitude variation than earlier heat flux maps, and provides an important boundary condition for numerical ice-sheet models and interpretation of borehole temperature profiles. In addition, it reveals new geologically significant features. Notably, we identify a prominent quasi-linear elevated geothermal heat flux anomaly running northwest-southeast across Greenland. We interpret this feature to be the relic of the passage of the Iceland hotspot from 80 to 50 Ma. The expected partial melting of the lithosphere and magmatic underplating or intrusion into the lower crust is compatible with models of observed satellite gravity data and recent seismic observations. Our geological interpretation has potentially significant implications for the geodynamic evolution of Greenland.

Figure 15. Example dataset in Google Dataset Search – with links to data providers and to citing articles in Google Scholar.

Articles 4 results (0,03 sec) My profile My library

**Any time**  
 Since 2020  
 Since 2019  
 Since 2016  
 Custom range...

**Sort by relevance**  
 Sort by date

include patents  
 include citations

Create alert

**Geothermal heat flux reveals the Iceland hotspot track underneath Greenland** [PDF] wiley.com  
 YM Martos, TA Jordan, M Catalán... - Geophysical ..., 2018 - Wiley Online Library  
 Abstract Curie depths beneath Greenland are revealed by spectral analysis of data from the World Digital Magnetic Anomaly Map 2. A thermal model of the lithosphere then provides a corresponding geo...  
 ☆ 18 Cited by 18 Related articles All 5 versions

**Surface expression of basal and englacial features, properties, and processes of the Greenland Ice Sheet** [PDF] wiley.com  
 MA Cooper, TM Jordan, MJ Siegert... - Geophysical Research ..., 2019 - Wiley Online Library  
 Abstract Radar-sounding surveys measuring ice thickness in Greenland have enabled an increasingly "complete" knowledge of basal topography and glaciological processes. Where such observations are s...  
 ☆ 8 Related articles All 8 versions

**Sensitivity of the Northeast Greenland Ice Stream to Geothermal Heat** [PDF] wiley.com  
 S Smith-Johnsen, NJ Schlegel... - Journal of ..., 2020 - Wiley Online Library  
 Page 1. manuscript submitted to JGR: Earth Surface Sensitivity of the Northeast Greenland Ice Stream to 1 Geothermal Heat 2 S. Smith-Johnsen1, NJ. Schlegel2, B. de Fleurian1and KH Nisancioglu1,3 3 1Department of Earth ...  
 ☆ 2 Cited by 2 Related articles

**A constraint upon the basal water distribution and thermal state of the Greenland Ice Sheet from radar bed echoes** [PDF] whiterose.ac.uk  
 TM Jordan, CN Williams, DM Schroeder... - ..., 2018 - eprints.whiterose.ac.uk  
 Page 1. This is a repository copy of A constraint upon the basal water distribution and thermal state of the Greenland Ice Sheet from radar bed echoes. White Rose Research Online URL for this paper: http://eprints.whiterose.ac.uk/150981/ Version: Published Version ...  
 ☆ 10 Cited by 10 Related articles All 16 versions

Figure 16. Automated search in Google Scholar from Google Dataset Search (see Figure 15). Note that number of citations in Google Dataset Search does not appear to be recently updated.

## 4 DISCUSSION AND CONCLUSION

Overall, the analysis of the existing portals shows that there are several different initiatives that facilitate open data sharing – both field specific and generic, both commercial and sponsored by governments or research organisations. Some of these function as **aggregators of metadata** (and do not offer any archiving of data themselves), some publish data from certain platforms or organisations, and others **facilitate self-archiving of datasets**.

Most of the examined examples attempt to give statistics on **the number of dataset views and dataset downloads**. However, as the same dataset can be discovered in several aggregators the views downloads statistics are also distributed and are hard to aggregate and analyse. Thus getting an overview and correct total for these figures is difficult. This situation is not unlike that of citation counts for publications where the same article has different citation counts in Web of Science, Scopus, Google Scholar and ResearchGate. Most aggregators do a fairly good job of presenting consistent metadata, e.g. preserving titles, author information, and DOIs and pointing back to the original source. However, different metadata levels and metadata specific

to some sources can be a challenge – with some fields being empty in an aggregator, and some information from the original source that does not fit in in the aggregator scheme.

Table 3. Broadly applicable Open Data indicators

Indicator	Strengths	Weaknesses	OPERA recommendation
Number of datasets published	The simplest indicator; many data sources	What defines a dataset? Does dataset size matter?	Include
Dataset size?	May indicate effort and importance.	Not clear how to measure size across fields.	Do not include
Number of dataset views	Can indicate visibility and potential interest.	Many datasets that are irrelevant to users can be viewed. Could be gamed.	Include
Number of dataset downloads	Can indicate strong interest.	Not certain that the dataset will be used. Could be gamed.	Include
Number of dataset citations	Strong indication of interest/utility.	Not clear what a dataset citation means.	Include
Advanced relative and weighted indicators	Allows comparisons to be made.	Data not good enough at present.	Do not include

In addition to views and downloads, actual **usage of data that leads to a dataset citation** in new publications is interesting and important to monitor. Google Dataset Search reports the number of citations in Google Scholar - automatically identified via a search on DOIs and archive name. GBIF does daily automated searches in a number of sources, and manually curates these. Identifying dataset citations is made difficult because a data citation culture is still to be established in most fields. This means that **many citations to datasets may be missed** because 1) many different ways of citing datasets is being used with little consistency (e.g. referring to the dataset in the main text, vs. in a footnote or in the reference list), 2) some may not be used to citing data, but cites the article describing the data instead or not at all. To counter this, several aggregators and dataset repositories give detailed instructions on how to cite the dataset, e.g. by posting a reference that can be readily copied in a manuscript, e.g. NSIDC under NASA (Figure 3) and Mendeley data (Figure 7). GBIF has the most advanced solution where not only each dataset can be cited, but also subsets and aggregates receive their own citable DOI. The disadvantage of this is that the same data can be cited with several different DOIs. Even

with such elaborate support in place example studies show that the data citation culture is still weak – see Kahn, Thellwall and Koucha (2019) for GBIF.

As to the possibility to use any Open Data indicators the opportunities to publish and share Open Data are now becoming more accessible – with institutional repositories offering this as well as field specific initiatives, such as the NASA Open Data Portal and GBIF. Thus, **the number of datasets published** is becoming a viable indicator. What constitutes a dataset and how to measure its size is less clear, and often field dependent, but important to investigate further. Most dataset portals give statistics on number of **dataset views and downloads**, and some the **number of dataset citations**. From these basic statistics and by correlation with other data relative and weighted indicators can be constructed, e.g. share of downloads to dataset views, share of cited datasets published, citations weighted relative to other datasets in the same field/year/country etc. However, at present with a weak dataset citation culture and incomplete dataset citation statistics advanced relative and weighted indicators will be hard to implement and interpret. Even simple indicators, like the number of dataset views and downloads remain hard to interpret: What does it *mean* that a dataset was viewed or downloaded many times? What is the relative importance of different kinds of altmetric metrics? To what degree can the statistics be gamed, and are they? **Regardless of these challenges, for the Open Science and Open Data movement to succeed it is important that we gain experience in collecting, publishing and using indicators of Open Data in order to learn more about how they can support these movements and aid in reaping their benefits for science.** Table 3 summarises some of the indicators discussed, their strengths and weaknesses, and recommendation whether to include them in the OPERA project RAP.

## 5 ACKNOWLEDGEMENTS

We wish to thank Pelle Annfeldt Israelsson, who worked as an assistant on this report initially and provided much valuable input. We also wish to thank Brian Kirkegaard Lunn, Senior Metadata Manager for Dimensions at Digital Science for help in identifying good example datasets in altmetrics.com. Finally, we want to thank the internal OPERA reviewers.

## 6 REFERENCES

- Eysenbach G. (2011). Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact. *Journal of Medical Internet Research*, 13(4), e123. <https://doi.org/10.2196/jmir.2012>
- Ingwersen, P., & Chavan, V. (2011). Indicators for the Data Usage Index (DUI): an incentive for publishing primary biodiversity data through global information infrastructure. *BMC Bioinformatics*, 12, S3. <https://doi.org/10.1186/1471-2105-12-S15-S3>
- Kahn, N., Thellwall, M. & Kousha, K. (2019). Data Citation and Reuse Practice in Biodiversity - Challenges of Adopting a Standard Citation Model. In: Catalano, G., Daraio, C. Gregori, M., Moed, H. F. & Ruocco, G. (eds). *Proceedings of the 17th Conference of the International Society for Scientometrics and Informetrics (ISSI'2019)*, volume 1, p. 1220-1225. [http://issisociety.org/proceedings/issi\\_2019/ISSI%202019%20-%20Proceedings%20VOLUME%20I.pdf](http://issisociety.org/proceedings/issi_2019/ISSI%202019%20-%20Proceedings%20VOLUME%20I.pdf)