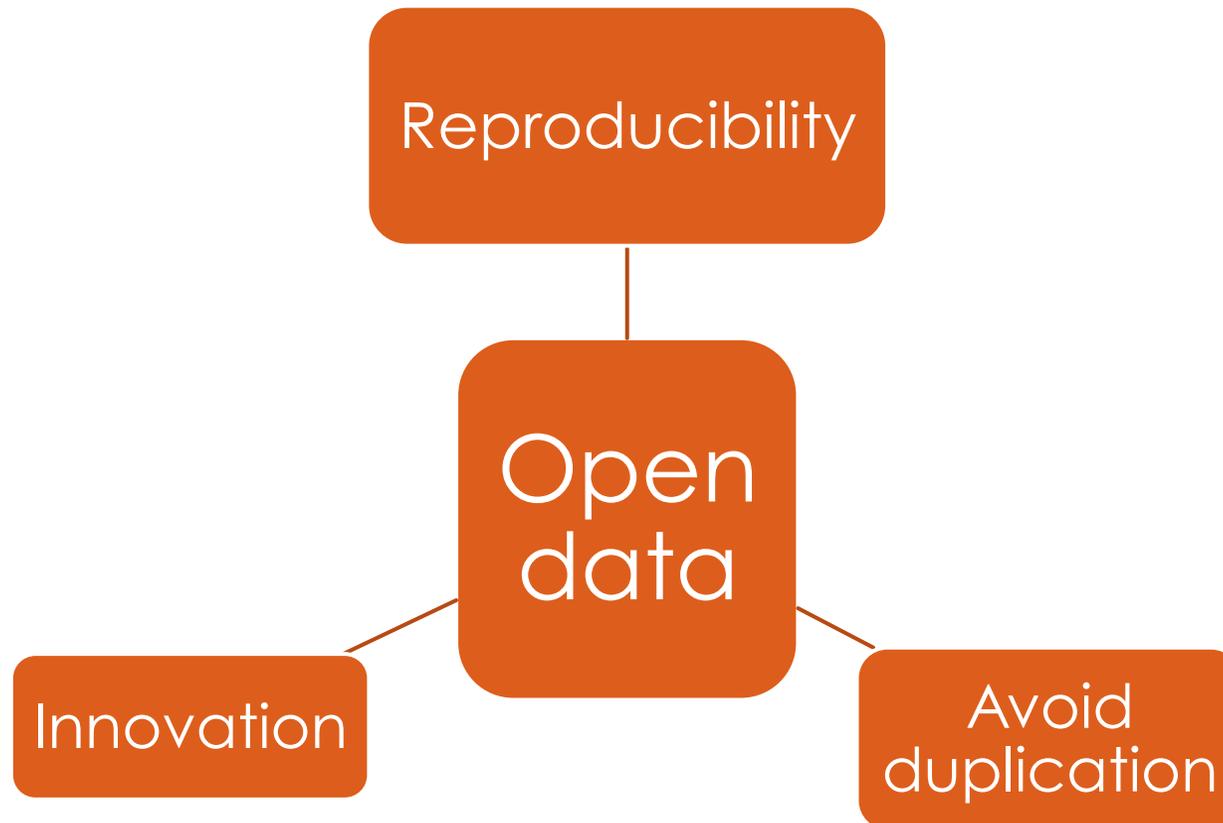


# Does Data Sharing Influence Data Reuse in Biodiversity? A Citation Analysis.

**NUSHRAT KHAN, MIKE THELWALL, KAYVAN KOUSHA**



# Open research data movement



# Why share research data?

In 2013, the Office of Science and Technology declared that most funding agencies will require policies on sharing open research data.

- ▶ Promote reproducibility, innovation and new data uses
- ▶ New collaborations between data creators and data users
- ▶ Avoid duplication of spending research fund to collect the same data
- ▶ Improvement and validation of research methods
- ▶ Increase the impact and visibility of research

# Motivations for data sharing

**Researchers want to know how their data has been reused <sup>1</sup>**

**How do we measure impact?**

1. Kratz, J. E., & Strasser, C. (2015). Making data count. Scientific data, 2.

# How do we define a dataset?

## What is a dataset?

- ▶ Can be a file or multiple files
- ▶ Packaged with adequate metadata and documentation so that it can be reused by others.
- ▶ Has proper licensing to make it clear how the data can be accessed and used by others.

# How do we define a dataset? (cont'd)

Here is the dilemma ...

- ▶ Type of data and data sharing practice differ depending on fields
- ▶ Not all fields have similar culture and rate of data reuse
- ▶ Not all repositories for different types of data provide the same metrics to assess impact

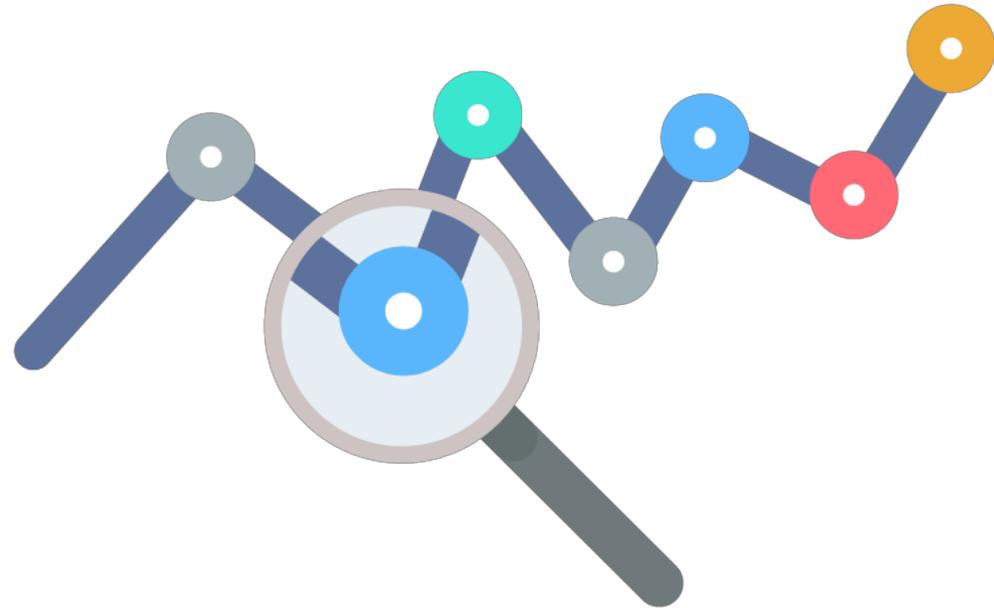
# Impact categories

Fear, K (2013) on measuring impact for social science datasets -

- ▶ Data reuse
- ▶ Quality of publications that reuse data
- ▶ Diversity of publications that reuse data
- ▶ Size of network stemming from a single dataset
- ▶ Number of unique individuals who download a dataset

# Tracking reuse

- ▶ **Data citation**
- ▶ **Altmetrics content**
- ▶ **Download counts**
- ▶ **?**



# Case of Biodiversity data

- ▶ Global Biodiversity Information Facility (GBIF) was used as the data source
- ▶ Total number of datasets (May 14, 2018) – 38,878
- ▶ Dataset types – Occurrence, Checklist, Metadata-only, Sampling-event
- ▶ Metadata fields – number of citations, DOI, type, title, description, language, date created, date modified
- ▶ Additionally download counts were manually collected and citation counts were updated for the random sample (October 25, 2018).

# Research Questions

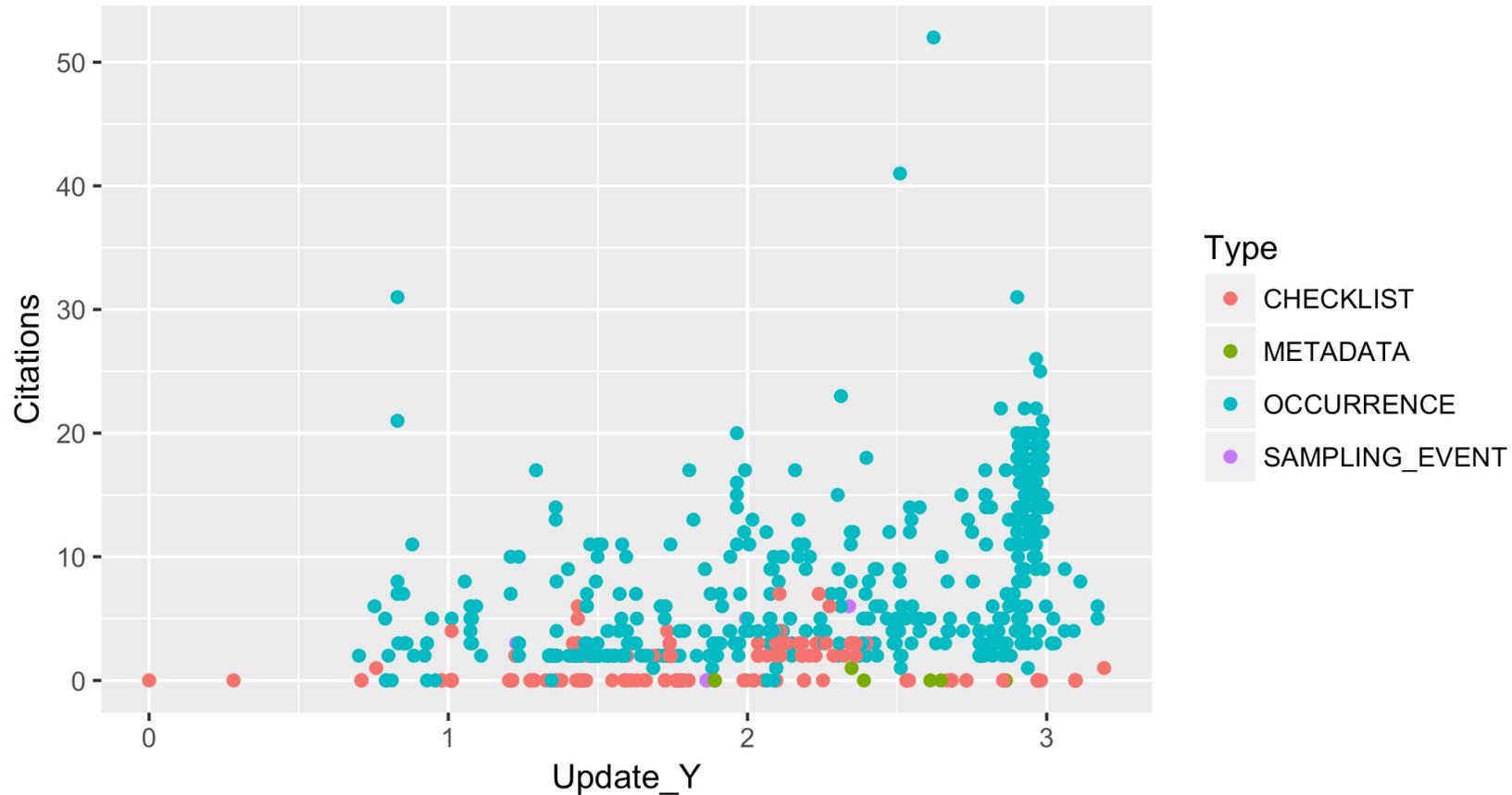
- ▶ Does the type of dataset or quality of information available affect citation rate?
- ▶ If a dataset is more recently updated (with the assumption that it is more frequently updated), does it have higher chance of being cited?
- ▶ How quickly dataset citations accrue?
- ▶ Does the citation count system in biodiversity result from coherent citation practice?
- ▶ Can we identify trending topics in biodiversity from the dataset types?

# Citation distribution

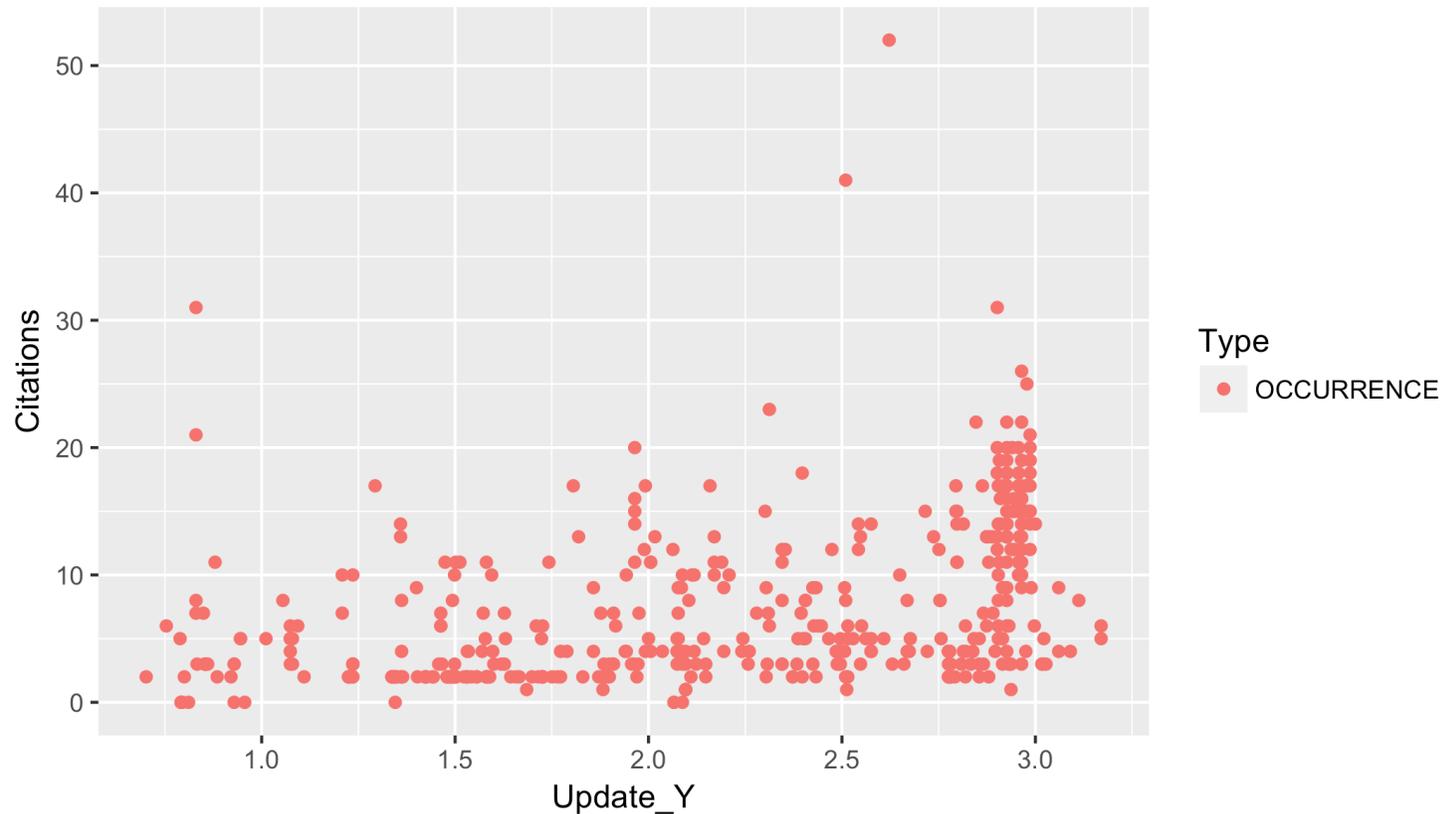
Type	Number of datasets	Average citations
Occurrence	14956	4.2967371
Checklist	23549	0.2082042
Metadata-only	204	0.1078431
Sampling-event	169	1.4023669

# Update frequency and citation rate of different datasets

Correlation between update time and citation rate for 2015 datasets



# Correlation test



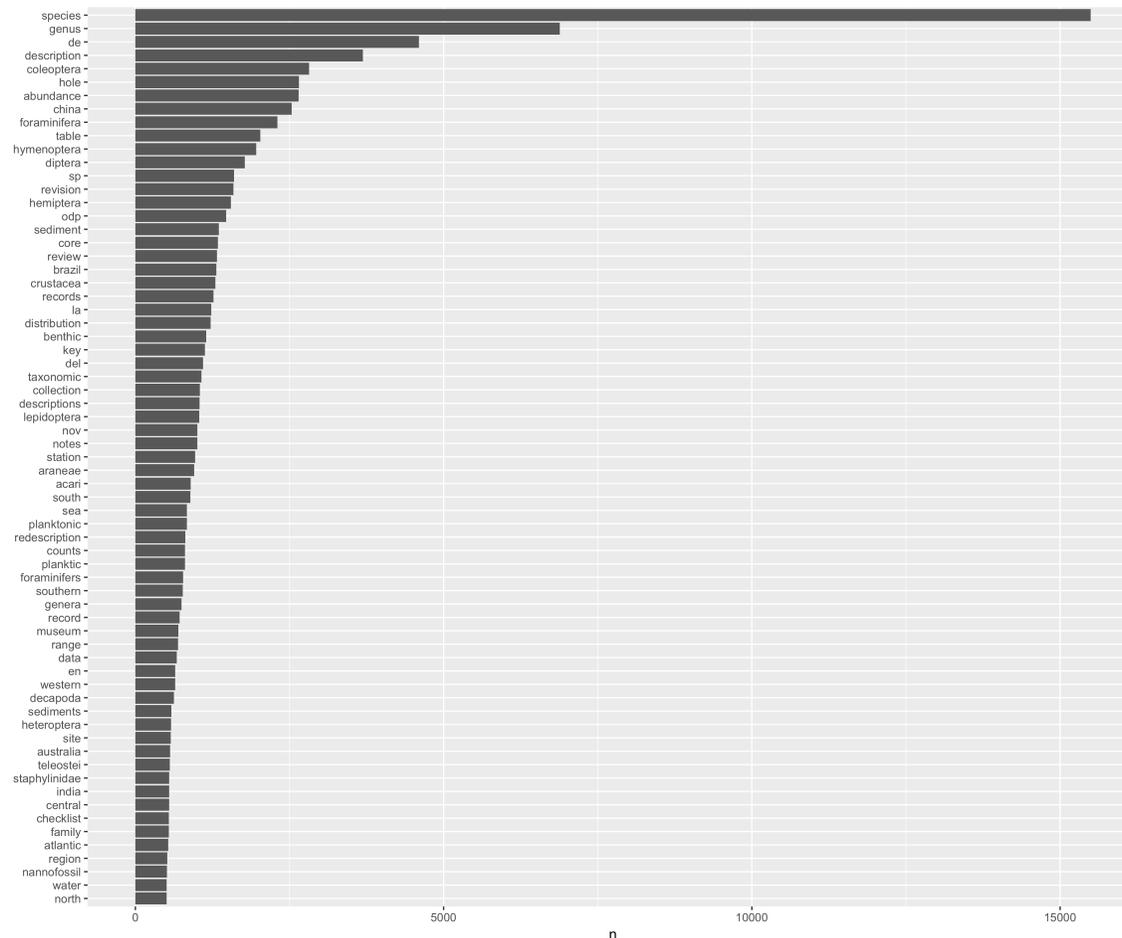
Correlation between dataset update time and citation rate;  
example of datasets published in 2015

## Spearman's rank correlation rho

**S** = 8443200, **p-value** < 2.2e-16  
alternative hypothesis: true rho is  
not equal to 0  
sample estimates:

**rho**  
0.5390476

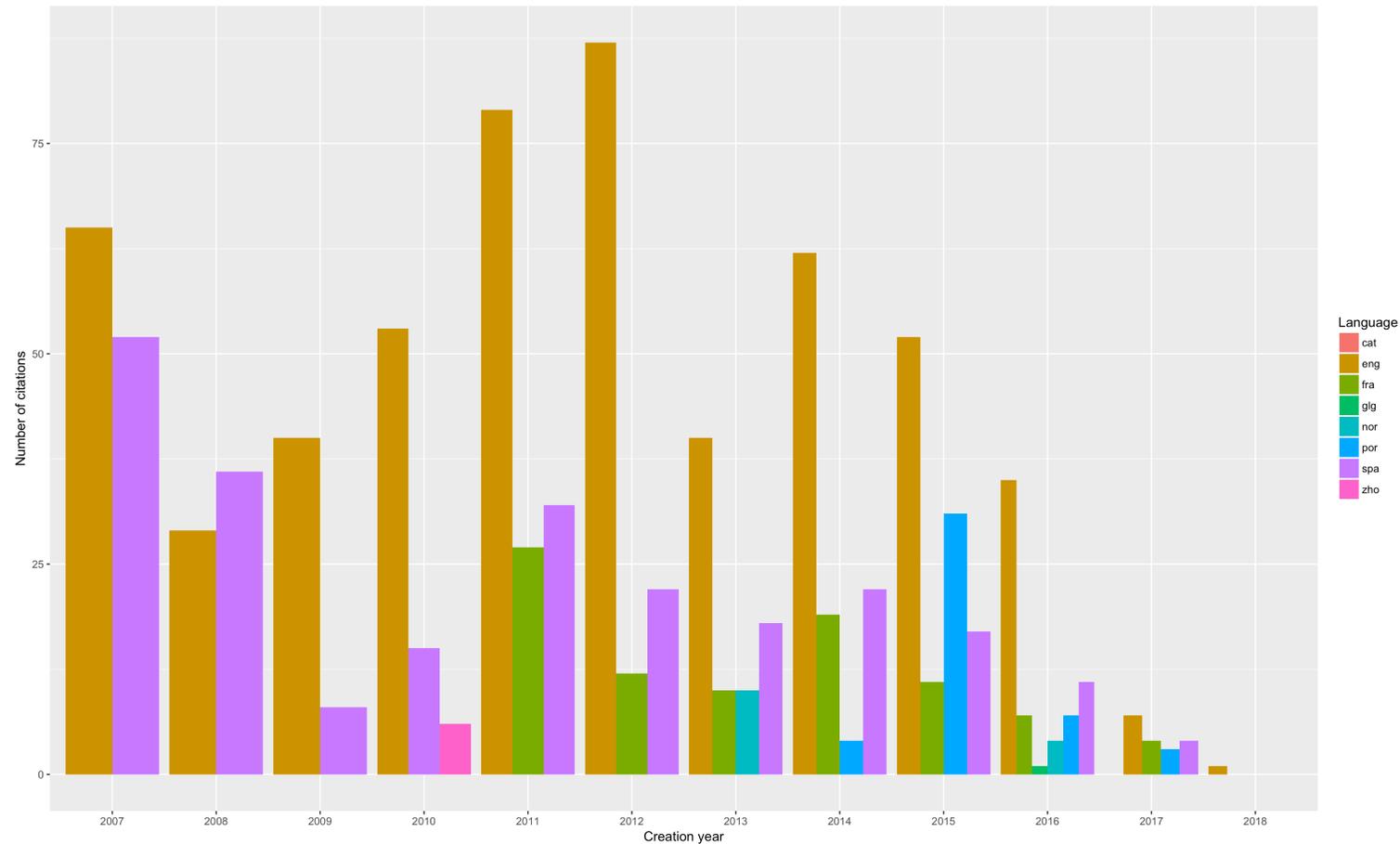
# Analysis of title texts



Some regions, including China, Brazil, Atlantic, Australia and India, appear more frequently than others

# Language and citation distribution

Number of datasets published in different languages between 2007-2018 and number of citations received



Number of datasets published in different languages between 2007-2018 and number of citations received

# Content analysis (work in progress)

- ▶ Random sample of 1000 out of 38,878 datasets
  - ▶ 438 datasets (43.8%) had at least one citation
- ▶ Collected download counts and updated citation counts to observe changes in past 6 months since initial data collection
- ▶ For datasets that received citations, selected one citing article randomly using random number generator
  - ▶ For each article (or book) collected DOI, citation location, data reuse case. (data for 100 cited datasets)

# Content analysis results (work in progress)

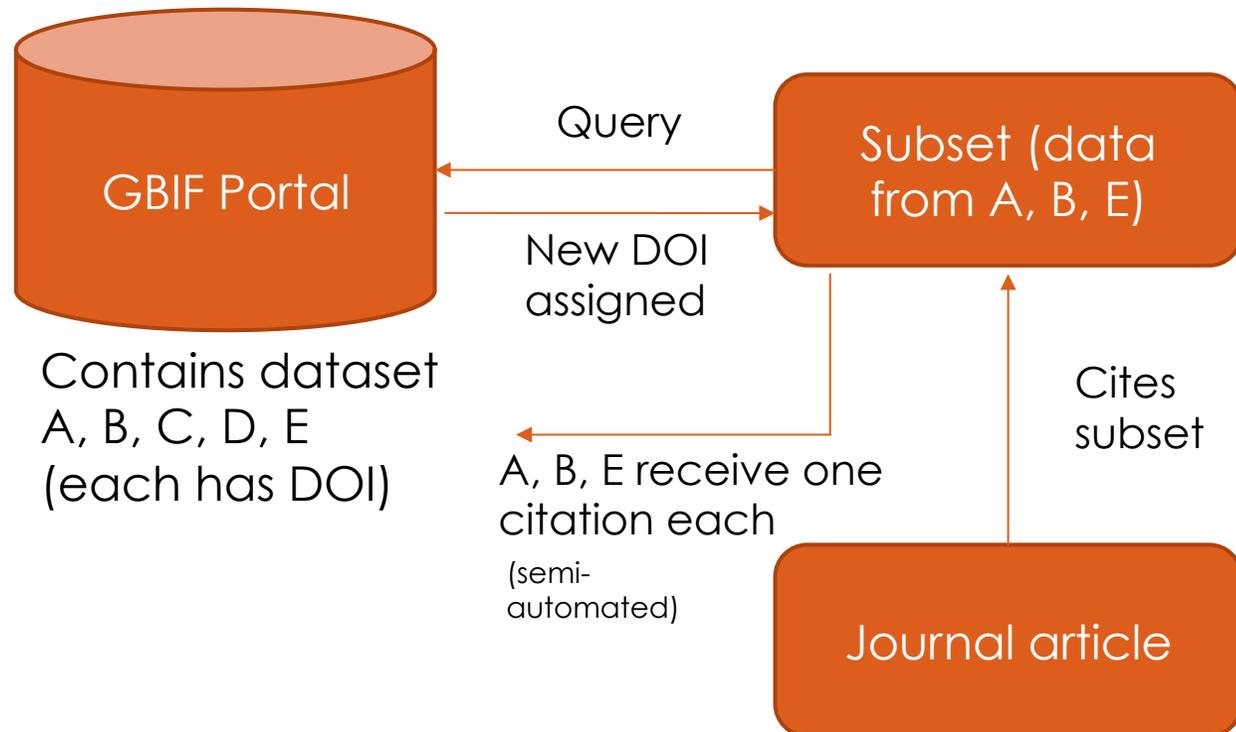
- ▶ Many datasets had accumulated citations over last 6 months
- ▶ 5 out of 100 datasets had high increase in citation (example: from 4 to 27)
- ▶ 25 out of 100 datasets had less citation than the initial record
- ▶ Almost all data reuse cases use multiple subsets of data
- ▶ Most of the times datasets are mentioned in methods and in some cases reference.

# Technical requirement

**Are we technically equipped to handle data citation count?**

**Why data citation count cannot always be the same as other citation counts?**

# How GBIF citation count works



# How GBIF citation count works

Citation example of GBIF subset download –

**Global Biodiversity Information Facility (GBIF). 2017b.**

[GBIF Occurrence Download](#). (accessed 20 December 2017).

<https://doi.org/10.15468/dl.qpjjdk>

Non-  
descriptive of  
the content

In case of reuse in  
multiple studies,  
doesn't indicate  
the original study

Total citation  
count for the  
original datasets  
don't reflect on  
regular  
databases

# Recommendations

D

Define

N

Normalize

D

Develop



# Google Dataset Search

Google Dataset Search

gbif



About



GBIF Occurrence Download

search.datacite.org

Updated Feb 6, 2017



GBIF Occurrence Download

search.datacite.org

Updated Feb 3, 2016



GBIF Occurrence Download

search.datacite.org

Updated Dec 29, 2015



GBIF Occurrence Download

search.datacite.org

Updated Feb 17, 2017



GBIF Occurrence Download

search.datacite.org

## Data from: GBIF - Global Biodiversity Information Facility

SCR\_005904, (GBIF - Global Biodiversity Information Facility, RRID:SCR\_005904), GBIF, Global Biodiversity Information Facility, GBIF Data Portal

Related Article

scicrunch.org

### Description

The Global Biodiversity Information Facility (GBIF) was established by governments in 2001 to encourage free and open access data, via the Internet. Through a global network of countries and organizations, GBIF promotes and facilitates the mobilization, discovery and use of information about the occurrence of organisms over time and across the planet. GBIF provides three core products: # An information infrastructure an Internet-based index of a globally distributed network of interoperable databases to provide primary biodiversity data information on museum specimens, field observations of plants and animals in nature, and results from so that data holders across the world can access and share them # Community-developed tools, standards and protocols that providers need to format and share their data # Capacity-building the training, access to international experts and mentoring programs national and regional institutions need to become part of a decentralized network of biodiversity information facilities. GBIF and partners work to mobilize the data, and to improve search mechanisms, data and metadata standards, web services, and the operation of an Internet-based information infrastructure for biodiversity. GBIF makes available data that are shared by hundreds of data publishers around the world. These data are shared according to the GBIF Data Use Agreement, which includes the provision that users of data accessed through or retrieved via the GBIF Portal will always give credit to the original data publishers. \* Explore Species: Find data on species or other group of organisms. Information on species and other groups of plants, animals, fungi and micro-organisms, including occurrence records, as well as classifications and scientific and common names. \* Explore Countries: Find data on the species recorded in each country, territory or island. Information on the species recorded in each country, including records shared by publishers throughout the GBIF network. \* Explore Datasets: Find data from a data publisher, dataset or data network. Information on the data

Tack!

Let's keep promoting (useful and curated) open research data!

Email: [n.j.khan@bath.ac.uk/](mailto:n.j.khan@bath.ac.uk)

[n.j.khan@wlv.ac.uk](mailto:n.j.khan@wlv.ac.uk)

Twitter: @brishti55