Introduction to the Data Life Cycle, FAIR (**F**indable, **A**ccessible, **I**nteroperable, and **R**eusable) Data, and Data Citation

Presenter : Arctic Data Center, DataONE, Environmental Data Initiative, ESIP, GBIF, iDigBio, NEON

Amber Budden, Kyle Copas, and Kristin Vanderbilt

#datahelpdesk
Ecological Society of America 2019 ESAUSSEE
Career Fair Center in the Exhibit Hall https://esa.org/louisville/career-fair/
Monday 12 August 330-430 PM

Why Manage Data?

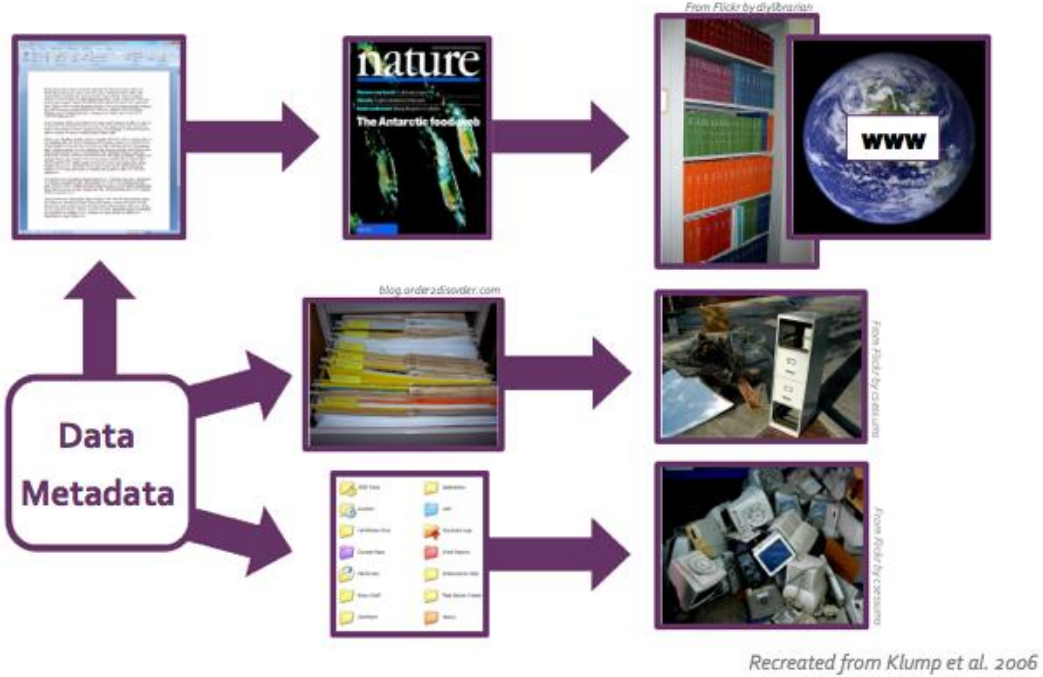
Carrots

- Data are being recognized as 1st class products of research
  - Data citation
  - Data publication
- Increased collaboration
- More efficient research
- Supports the advancement of science

Sticks

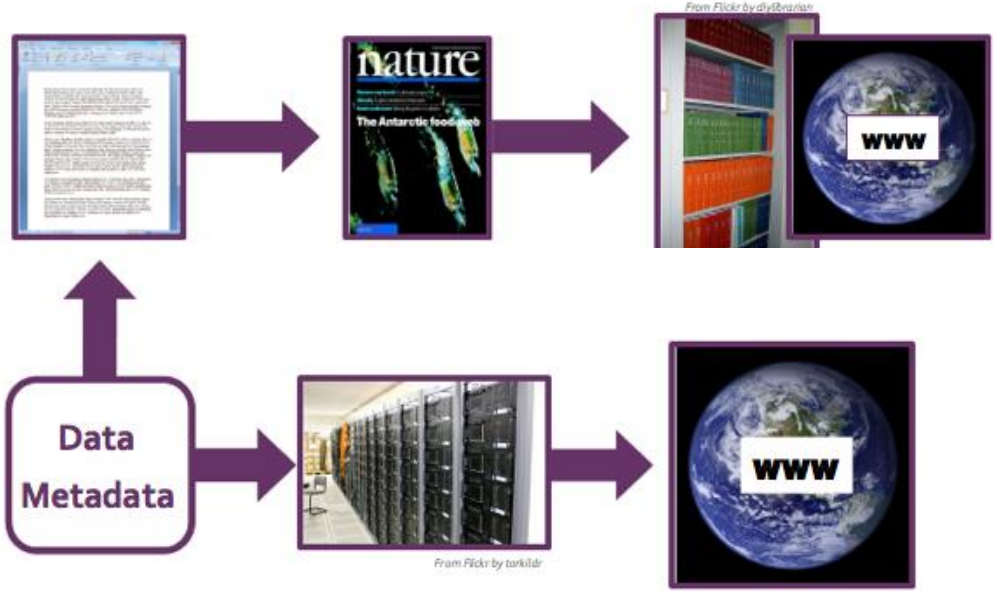
- Funders
- Publishers
- Peers

Data Life Span: Where majority of data ends up



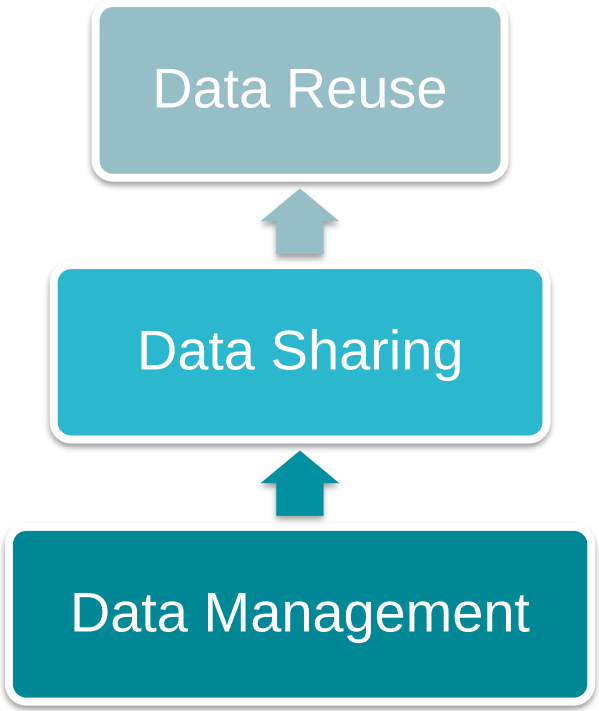
Recreated from Klump et al. 2006

An Alternative



Recreated from Klump et al. 2006

Good Data Management Facilitates Sharing & Reuse



Well-managed, shared, publicly accessible data is valuable

- Increases the impact and visibility of research
- Leads to new collaborations between data users and creators
- Supports open science
- Satisfies publishers and funders

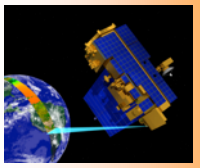
Well-Managed, Shared Data Can Result in Re-use, Integration, and New Science

eBird

Land Cover

Meteorology

MODIS – Remote sensing data



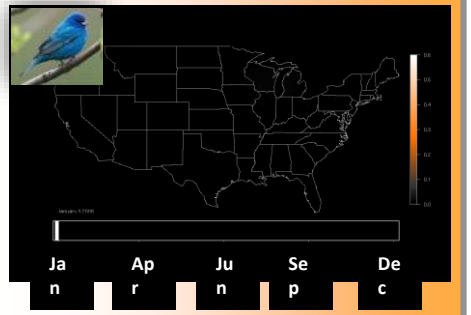
TeraGrid

$$F(X,s,t)=\frac{1}{n(s,t)}\sum_{i=1}^{m} f_i(X,s,t)I(s,t\in\theta_i)$$

Spatio-Temporal Exploratory Models predict the probability of occurrence of bird species across the United States at a 35 km x 35 km grid.

Model results

Occurrence of Indigo Bunting (2008)



Potential Uses-

- Examine patterns of migration
- Infer impacts of climate change
- Measure patterns of habitat usage
- Measure population trends

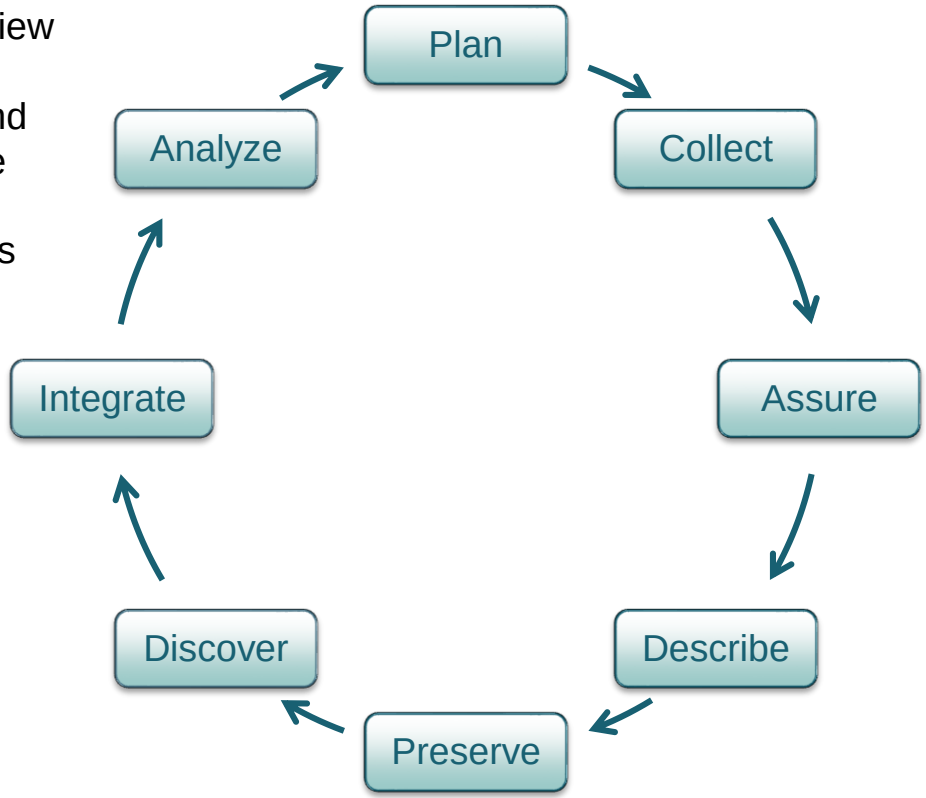
Slide courtesy of DataONE

Data Management

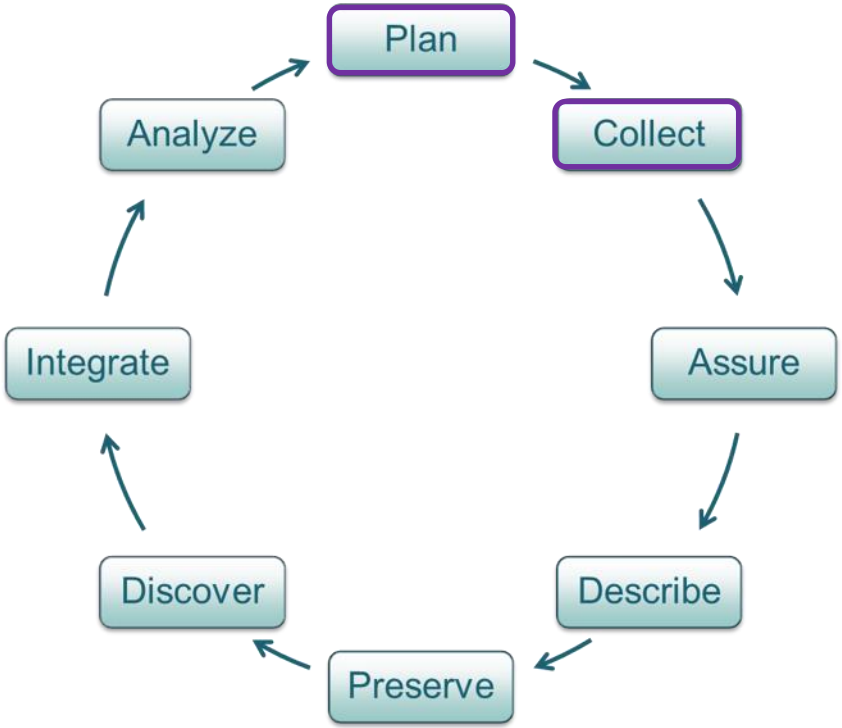
- Data Reuse <- Data Sharing <- Data Management
- A term that describes the organization, storage, documentation, and preservation of data collected and used in a research project
- Effective data management occurs throughout what we refer to as the *Data Life Cycle*.

Data Life Cycle

- Provides a high level overview of the stages involved in successful management and preservation of data for use and reuse.
- Helps define responsibilities
- Helps determine needed resources for data management (people, software, hardware)



DataONE Data Life Cycle

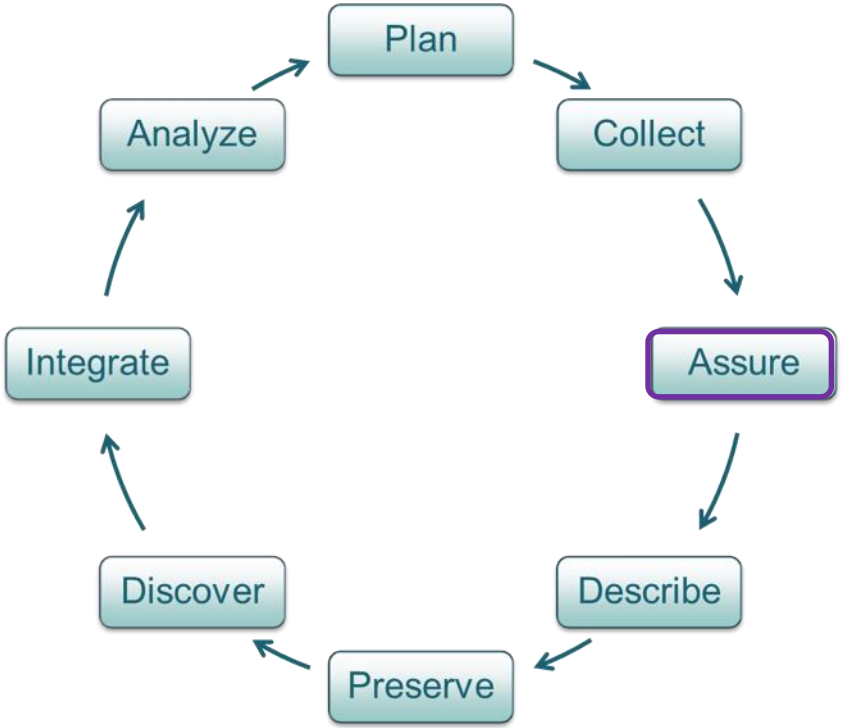


Plan: Write a Data Management Plan that describes the data to be acquired and how it will be documented, stored, quality assured, preserved and shared. DMP required for NSF proposals.

Collect: Data acquisition by hand or with sensors or other instruments.

- templates
- consistent file organization

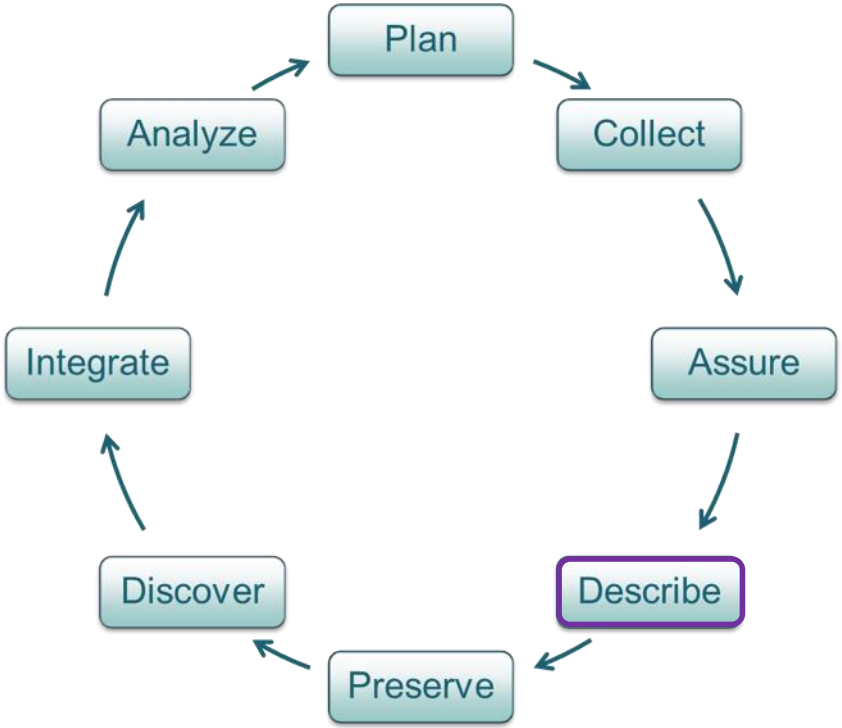
DataONE Data Life Cycle



Assure: Quality of data is assured through various checks and inspections.

- Perform statistical and graphical summaries
- Communicate quality

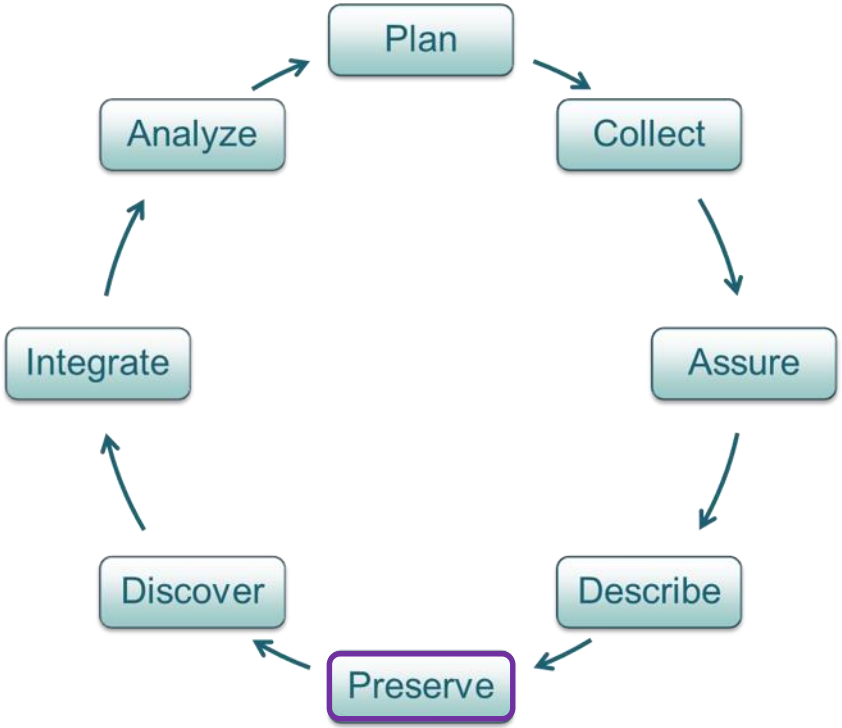
DataONE Data Life Cycle



Describe: Data are thoroughly documented using appropriate metadata standards. Key to understanding the data.

- Who? What? Where? When? Why?
- Structure
- Semantics
- License
- Standards: Ecological Metadata Language (EML); ISO19115; Darwin Core

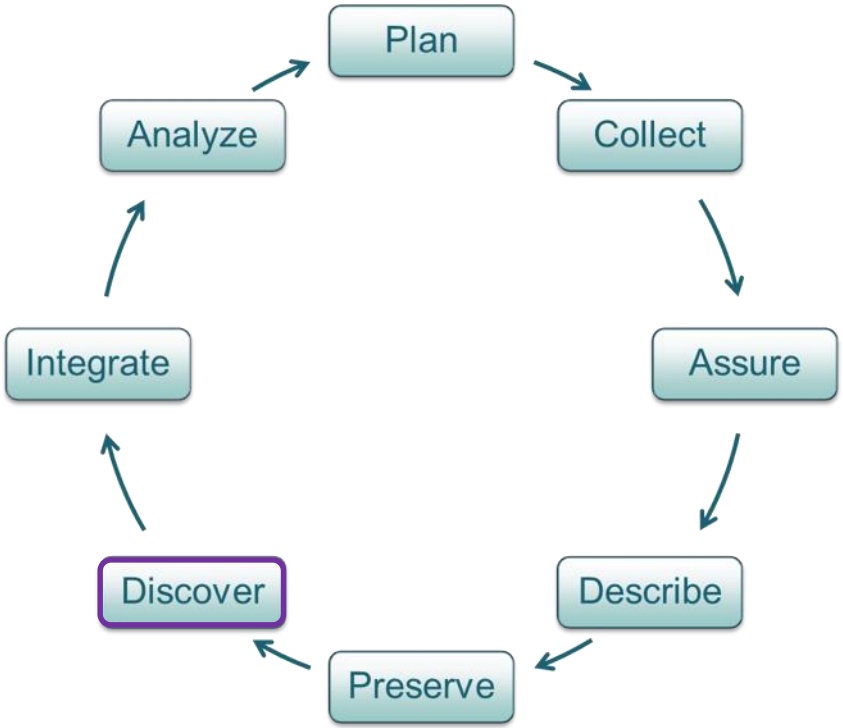
DataONE Data Life Cycle



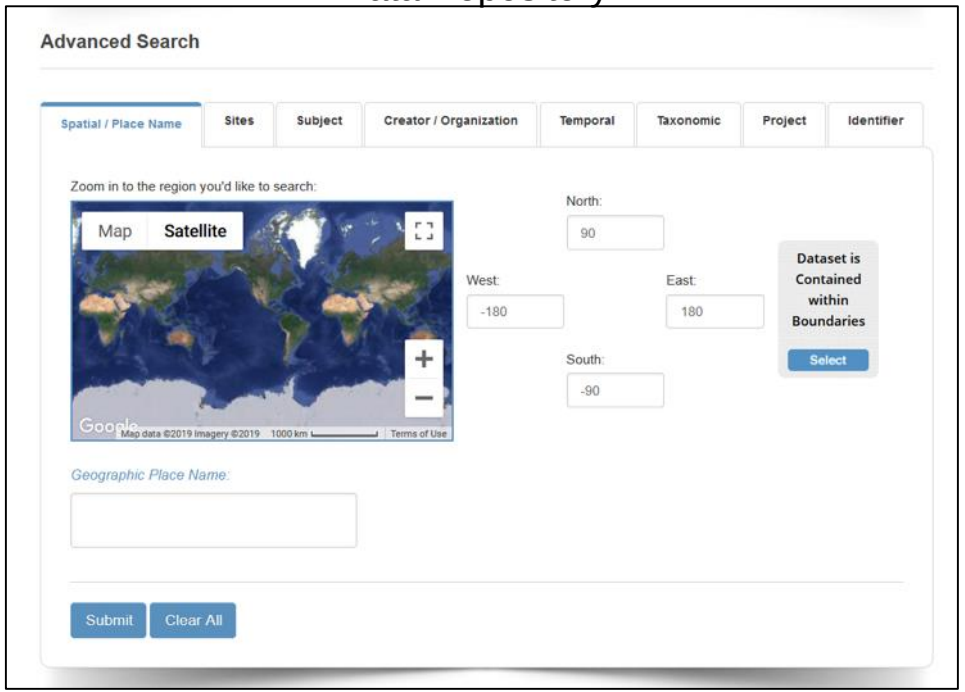
Preserve: Submit data to a data repository for long-term storage and to retain usability of data by someone other than the data creator.

- Data Repositories implement systematic stewardship practices to ensure data availability, dissemination and access.
- Data repositories are operated by universities, government entities, grant-funded organizations
- General and discipline-specific

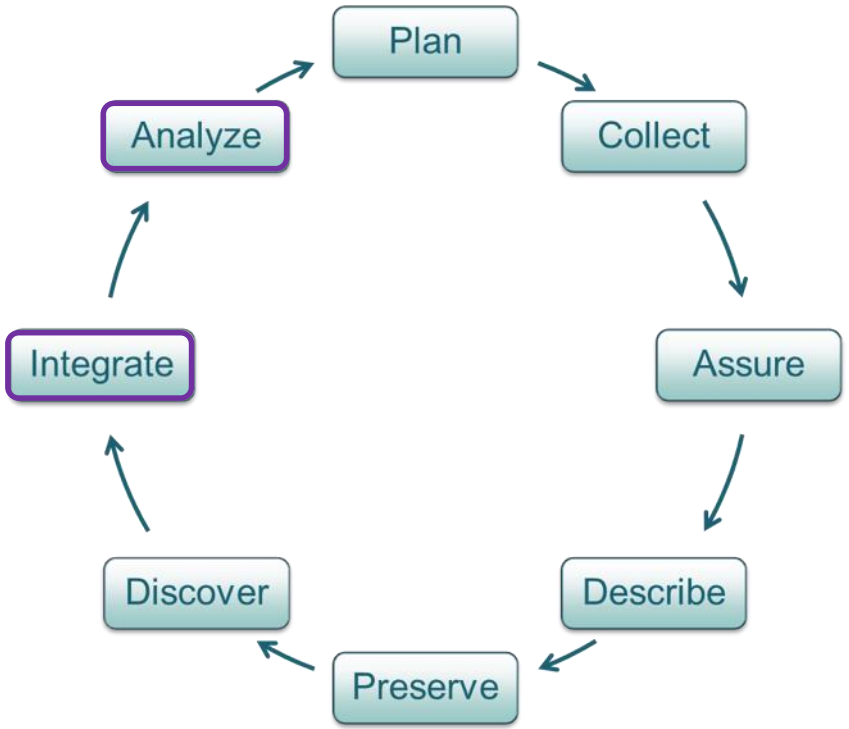
DataONE Data Life Cycle



EDI Data Repository



DataONE Data Life Cycle

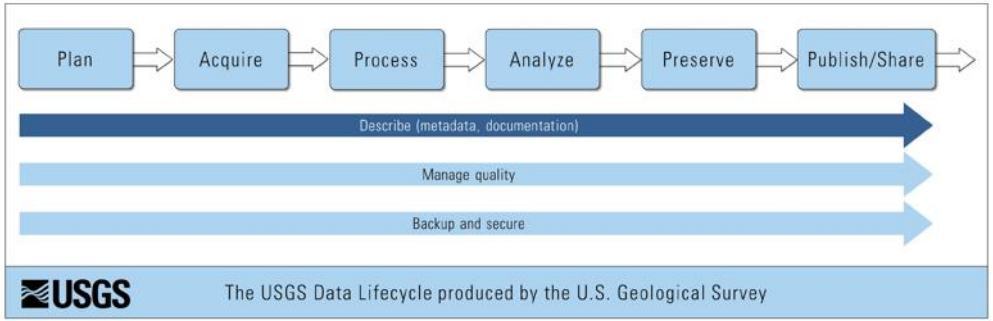


Integrate and Analyze: Data from disparate sources can be combined and analyzed. Papers are published and data are cited!

Other models of the Data Life Cycle

Data life cycle: Plan, Create, Process, Analyze, Preserve, Share, Reuse

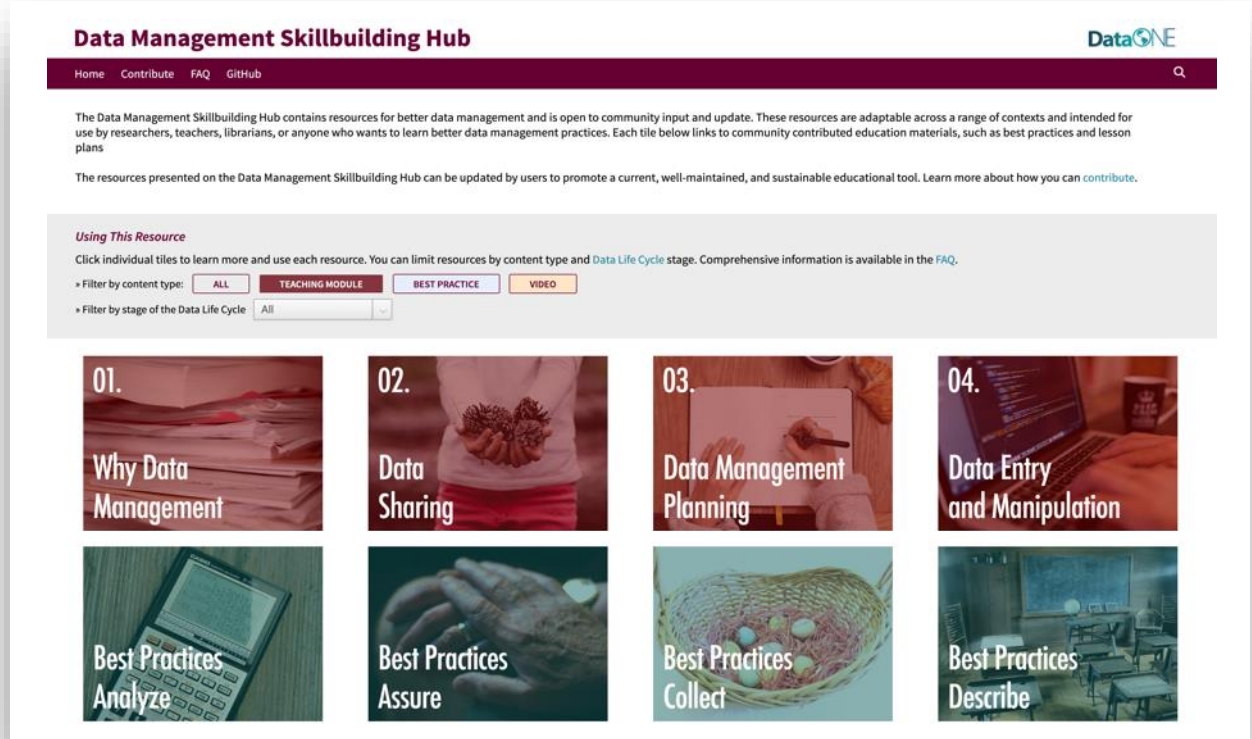
https://biblio.uottawa.ca/en/services/faculty/research-data-management/what-research-data-management



usgs.gov

Best Practices and FAIR Data

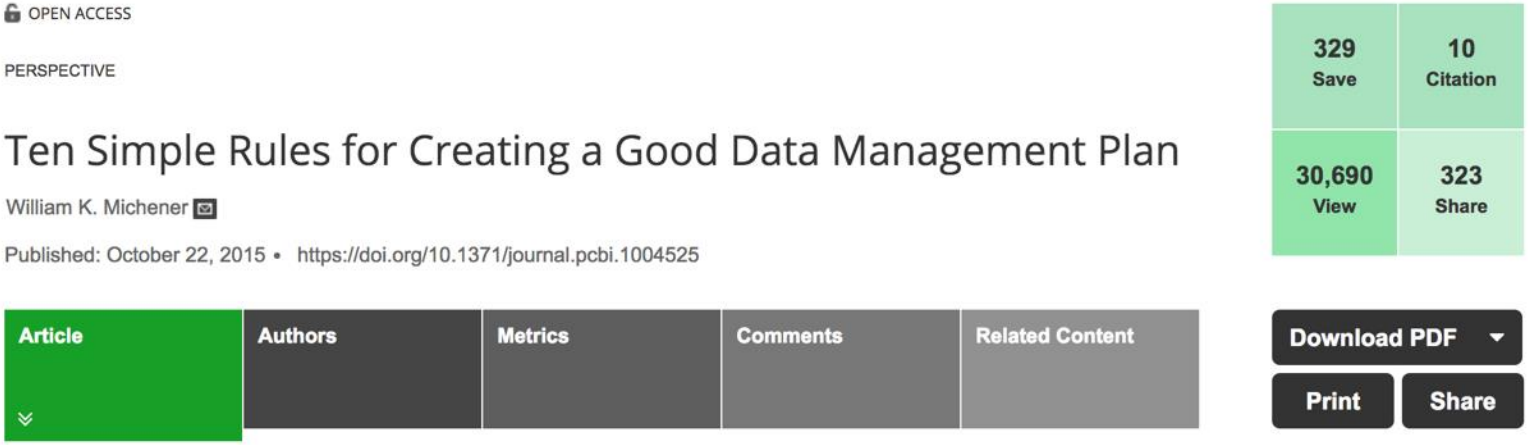
Best Practices: Skillbuilding Hub



dataone.github.io/Education

Plan

DMPTool



Arctic Data Center DataONE EDI SIP GBIF iDigBio NSF NEON Operated by Battelle

Collect

data.xls

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Site	Date	Plot	Species	Weight	Acult		Rodent Trapping 3/15/2010						
2	DeepWell	2/13/2010	1	DIPO	12.1	j		Site	Plot	Adult	RodentSp	Weight		
3	Deep Well	Feb-10	2	Pero	13.22	j		DW	1	y	Pero	12		
4	rioSalado	2/13/2010	1a	pero	16	N		RS	2	j	PERO	escaped <15		
5	riuSlado	"	1*	CleGap	18.92	gut away		RS	3	n	Clegap	91		
6				Mean1	15.06									
7														
8														
9														
10														
11														
12	Rodent Trapping		MJK & ALN	10-Apr-10										
13	Site	Plot	Adult	Species	grams	Comments								
14	deep well	1	y	woodrat	13									
15	riosalado	2	y	PERO	24.5									
16	riosalado	3	y	Clegap	91									
17														
18														
19														

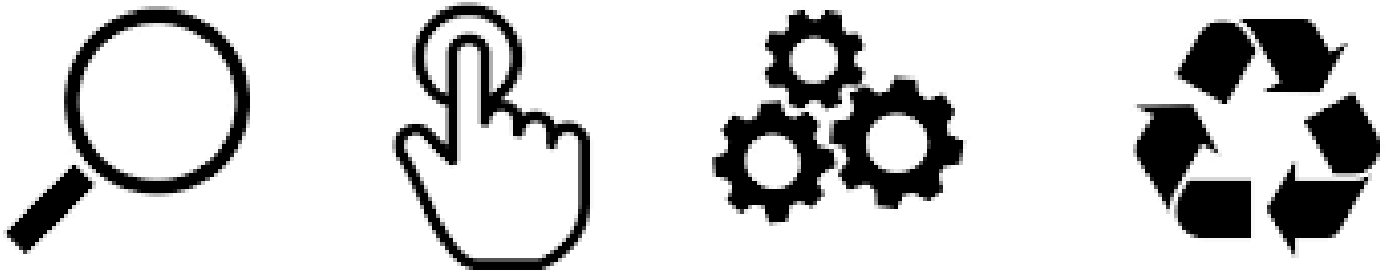
Sheet1

- Ensure column contents are consistent (don’t mix data and text)
- Create descriptive column names (no spaces or special characters)
- Use a descriptive file name
- Store data in a consistent format that can be well in to the future

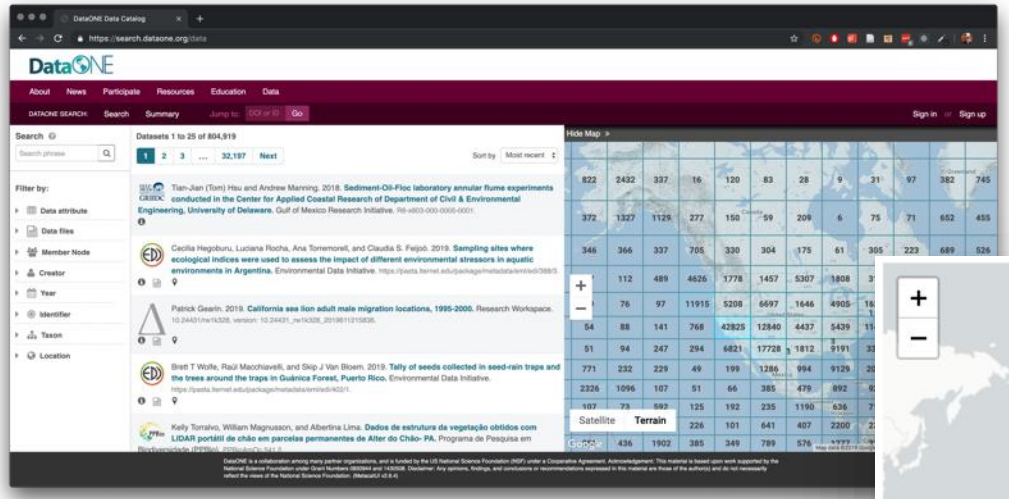
Preserve

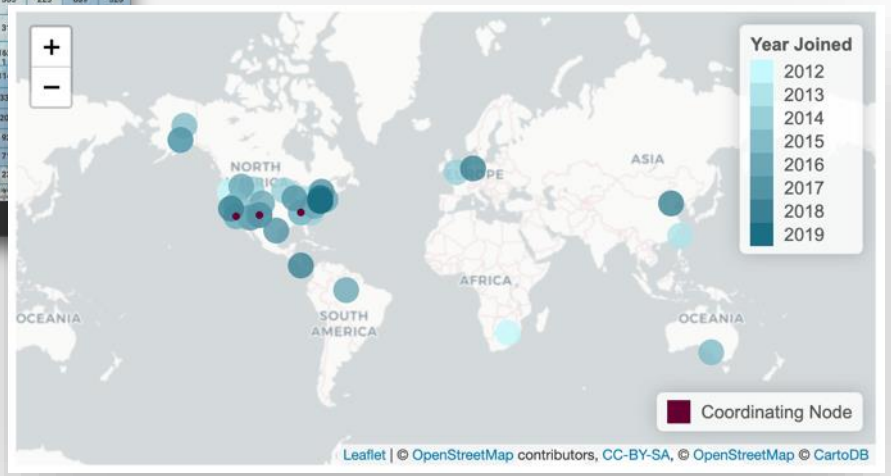
- Create a clear backup policy and review frequently
- Minimize or remove reliance on users
- Don’t assume backups are being performed
- Use non-proprietary, standard formats
- Check your backups manually
- Have multiple versions of backups on multiple formats in multiple places

Findable Accessible Interoperable Reusable



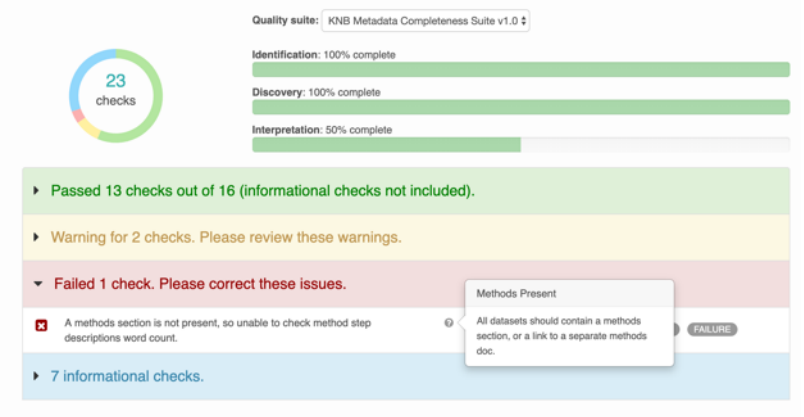
Finable Accessible

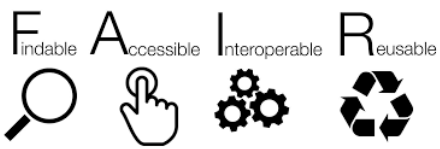


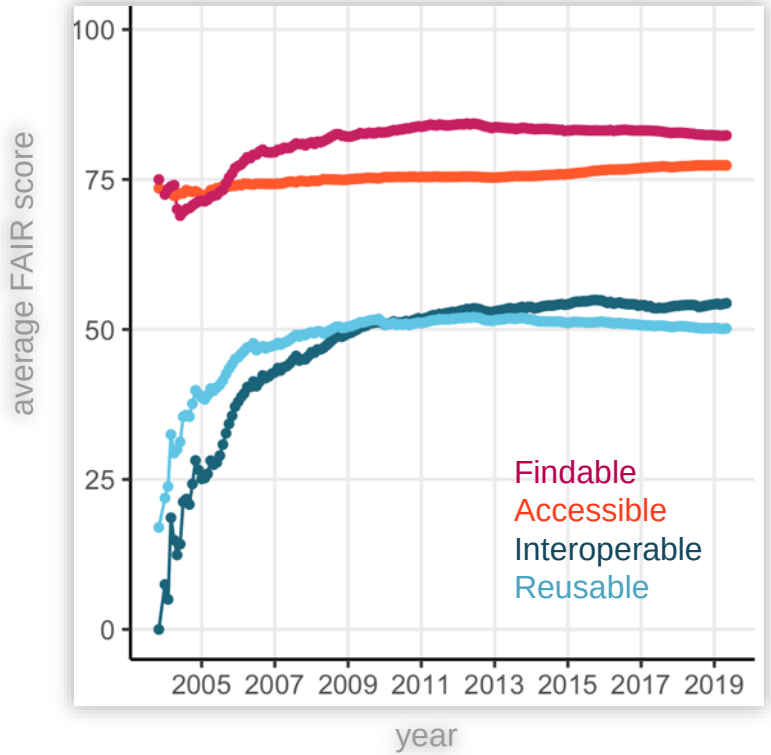


Interoperable Reusable

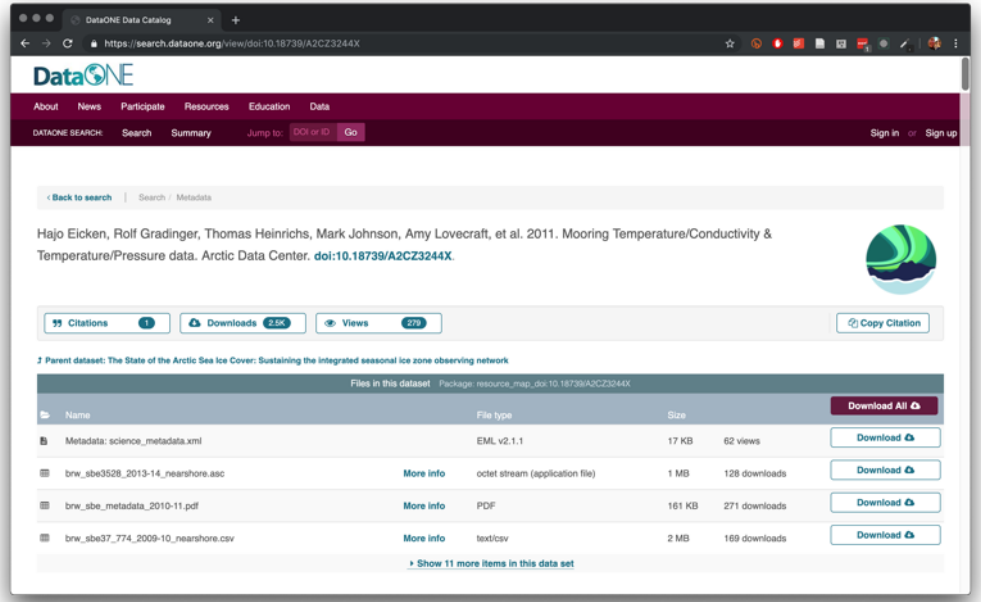
- Use open, reproducible workflows

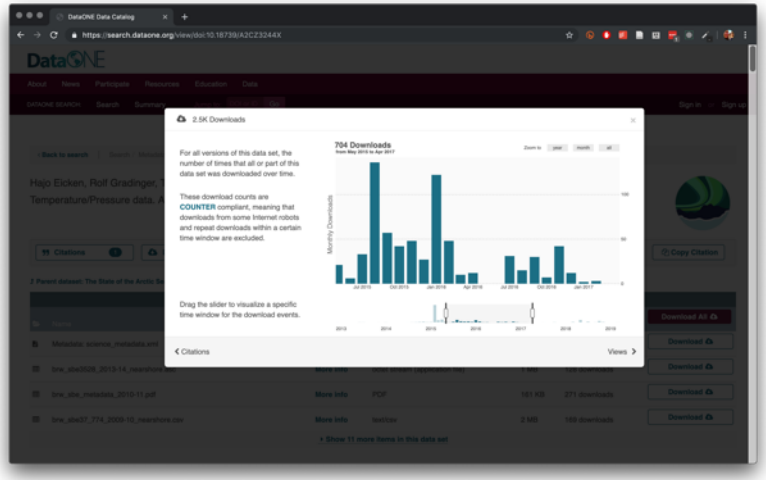






Getting Credit: Data Citation and Data Level Metrics





MAKE DATA COUNT

GBIF Global Biodiversity Information Facility

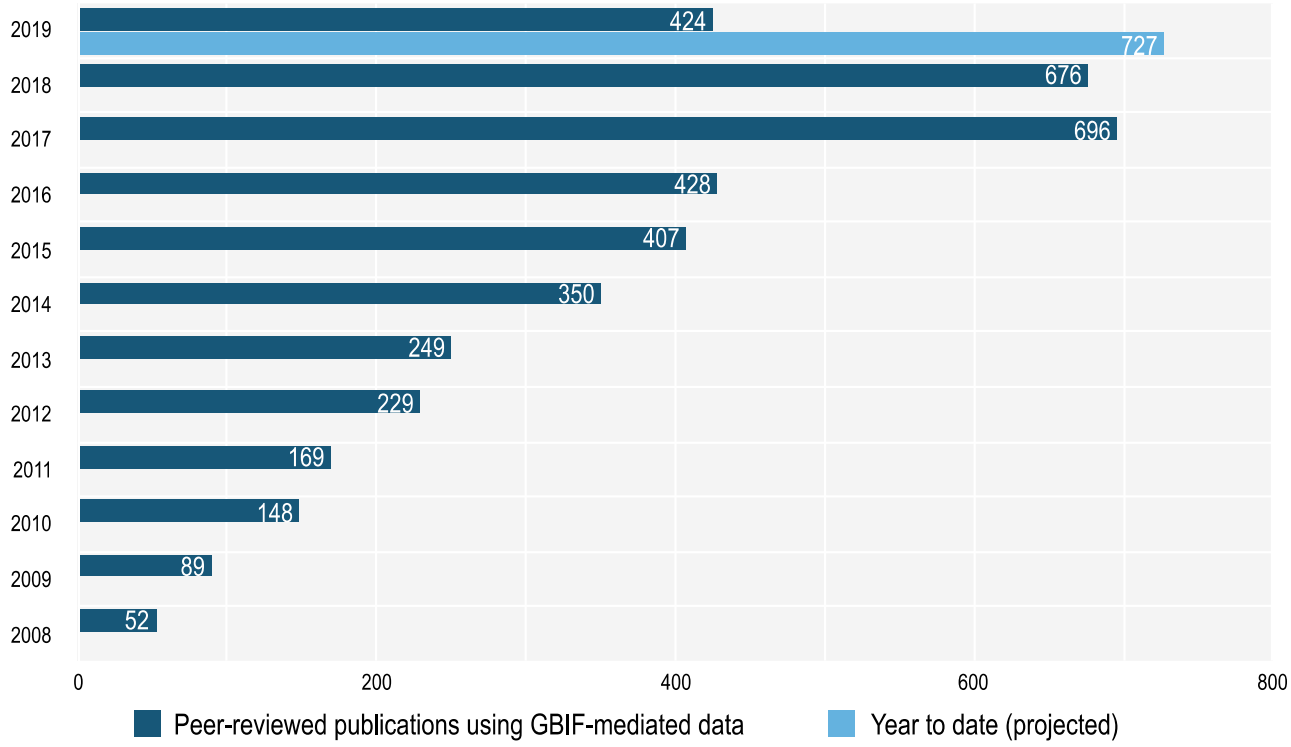
INTRO TO DATA CITATION

Kyle Copas | GBIF Secretariat

DATA DIALOGUES, ESA/USEEE ANNUAL MEETING
LOUISVILLE, KENTUCKY, USA | 12 AUGUST 2019

PEER-REVIEWED USES OF GBIF-MEDIATED DATA

through 31 July 2019



https://www.gbif.org/resource/search?contentType=literature&peerReview=true

GBIF

PRE-EXISTING (ENABLING) CONDITIONS

Data and supplementary materials have sufficiently rich metadata and a unique and persistent identifier.

FINDABLE

Metadata and data are understandable to humans and machines. Data is deposited in a trusted repository.

ACCESSIBLE

Metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.

INTEROPERABLE

Data and collections have a clear usage licenses and provide accurate information on provenance.

REUSABLE

Graphic by Association of European Research Libraries https://libereurope.eu/wp-content/uploads/2017/12/LIBER-FAIR-Data.pdf

GBIF

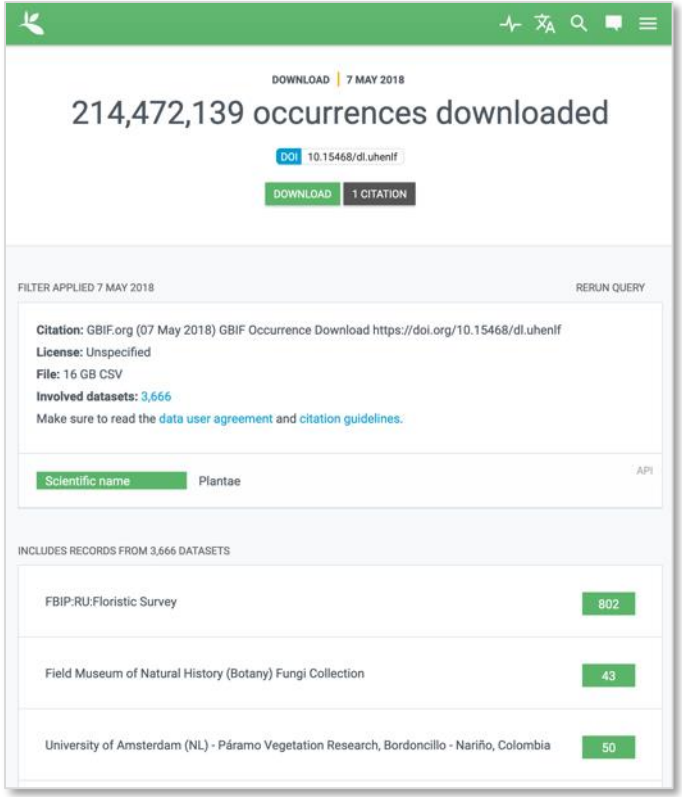
DOIs: DIGITAL OBJECT IDENTIFIERS

Four types of uses on GBIF.org

Use type	*Date of issue*	*Issued by*	*Function*
Mobilization	As close to publication event as possible	Data publisher, GBIF network host or GBIF.org	Institutional recognition of dataset through simple citation mechanism
Access	At time of user request for download from GBIF.org	GBIF	Simple citation, reproducibility, clarity of materials & methods
Bibliography	Publication date	Journal publisher	Stable identifier
Use	Data archiving before manuscript submission	Repository	Clarity of results, support reproducibility and discussion

DOWNLOAD DOIs

- GBIF stores each user download request and assigns a DOI resolving to a landing page that includes
  - Date and size of download
  - Terms/filters used in query
  - Links to download original file and to update query
  - Details of contributing datasets
  - Relationships modelled in metadata



DATA AS A FIRST-CLASS RESEARCH OBJECT

‘It is in the hands of researchers, who use such data, to properly cite the datasets or the associated data papers if they exist. **Failure to do so weakens the link between data and discovery** and reduces the probability of the latter by jeopardizing the availability of the former.’

Journal List › Database (Oxford) › v.2018; 2018 › PMC5892138

DATABASE
Alerts | Author Instructions | Submit | About
The Journal of Biological Databases and Curation

Database (Oxford). 2018; 2018: bay033. PMCID: PMC5892138
Published online 2018 Apr 9. doi: 10.1093/database/bay033 PMID: 29688384

The tragedy of the biodiversity data commons: a data impediment creeping nigher?

Nora Escribano, David Galicia, and Arturo H Ariño

‣ Author information ‣ Article notes ‣ Copyright and License information Disclaimer

Associated Data

‣ Supplementary Materials

Abstract Go to:

Researchers are embracing the open access movement to facilitate unrestricted availability of scientific results. One sign of this willingness is the steady increase in data freely shared online, which has prompted a corresponding increase in the number of papers using such data. Publishing datasets is a time-consuming process that is often seen as a courtesy, rather than a necessary step in the research process. Making data accessible allows further research, provides basic information for decision-making and contributes to transparency in science. Nevertheless, the ease of access to heaps of data carries a perception of ‘free lunch for all’, and the work of data publishers is largely going unnoticed. Acknowledging such a significant effort involving the creation, management and publication of a dataset remains a flimsy, not well established practice in the scientific community. In a meta-analysis of published literature, we have observed various dataset citation practices, but mostly (92%) consisting of merely citing the data repository rather than the data publisher. Failing to recognize the work of data publishers might lead to a decrease in the number of quality datasets shared online, compromising potential research that is dependent on the availability of such data. We make an urgent appeal to raise awareness about this issue.

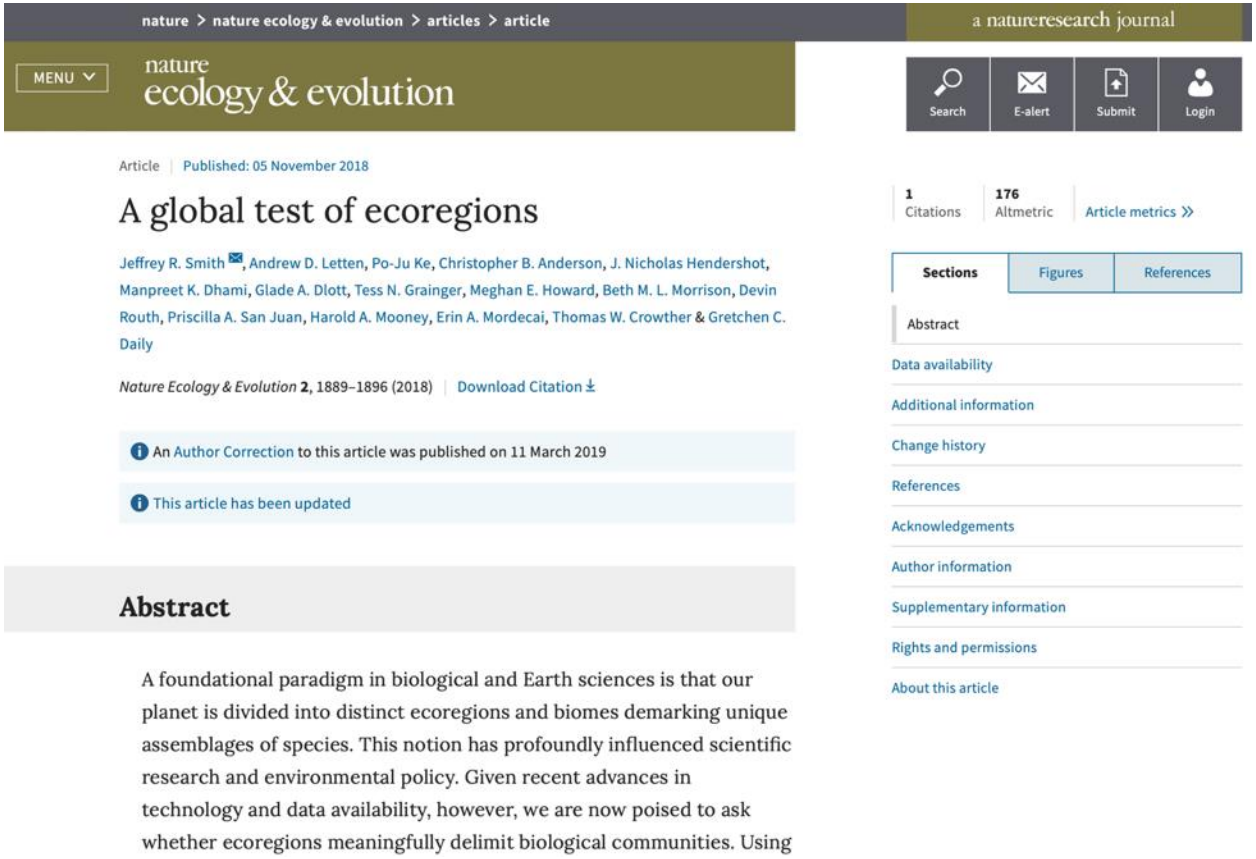
Database (Oxford)

Database (Oxford)

https://dx.doi.org/10.1093/database/bay033

GBIF

LITERATURE TRACKING



nature > nature ecology & evolution > articles > article

a natureresearch journal

MENU

nature ecology & evolution

Search | E-alert | Submit | Login

Article | Published: 05 November 2018

A global test of ecoregions

Jeffrey R. Smith, Andrew D. Letten, Po-Ju Ke, Christopher B. Anderson, J. Nicholas Hendershot, Manpreet K. Dhami, Glade A. Dlott, Tess N. Grainger, Meghan E. Howard, Beth M. L. Morrison, Devin Routh, Priscilla A. San Juan, Harold A. Mooney, Erin A. Mordecai, Thomas W. Crowther & Gretchen C. Daily

Nature Ecology & Evolution **2**, 1889–1896 (2018) | Download Citation

An Author Correction to this article was published on 11 March 2019

This article has been updated

1 Citations | 176 Altmetric | Article metrics

Sections | Figures | References

Abstract

Data availability

Additional information

Change history

References

Acknowledgements

Author information

Supplementary information

Rights and permissions

About this article

Abstract

A foundational paradigm in biological and Earth sciences is that our planet is divided into distinct ecoregions and biomes demarking unique assemblages of species. This notion has profoundly influenced scientific research and environmental policy. Given recent advances in technology and data availability, however, we are now poised to ask whether ecoregions meaningfully delimit biological communities. Using

https://doi.org/10.1038/s41559-018-0709-x

GBIF

Article | Published: 05 November 2018

A global test of ecoregions

Jeffrey R. Smith, Andrew D. Letten, Po-Ju Ke, Christopher B. Anderson, J. Nicholas Hendershot, Manpreet K. Dhami, Glade A. Dlott, Tess N. Grainger, Meghan E. Howard, Beth M. L. Morrison, Devin Routh, Priscilla A. San Juan, Harold A. Mooney, Erin A. Mordecai, Thomas W. Crowther & Gretchen C. Daily

The original article was published on 05 November 2018

Correction to: *Nature Ecology & Evolution* https://doi.org/10.1038/s41559-018-0709-x, published online 05 November 2018.

The original paper was published without unique DOIs for GBIF occurrence downloads. These have now been inserted as references 70–76, and the error has been corrected in the PDF and HTML versions of the article.

GBIF

Article | Published: 05 November 2018

A global test of ecoregions

Jeffrey R. Smith
Manpreet K. Dha
Routh, Priscilla
Daily

Nature Ecology

Abstract

A foundat
planet is
assembla
research
technolo
whether

70. *Plants*: GBIF.org (GBIF, 2018); https://doi.org/10.15468/dl.uhenlf

71. *Arthropods*: GBIF.org (GBIF, 2018); https://doi.org/10.15468/dl.uhumps

72. *Reptiles*: GBIF.org (GBIF, 2018); https://doi.org/10.15468/dl.pn1aw9

73. *Amphibians*: GBIF.org (GBIF, 2018); https://doi.org/10.15468/dl.as6uug

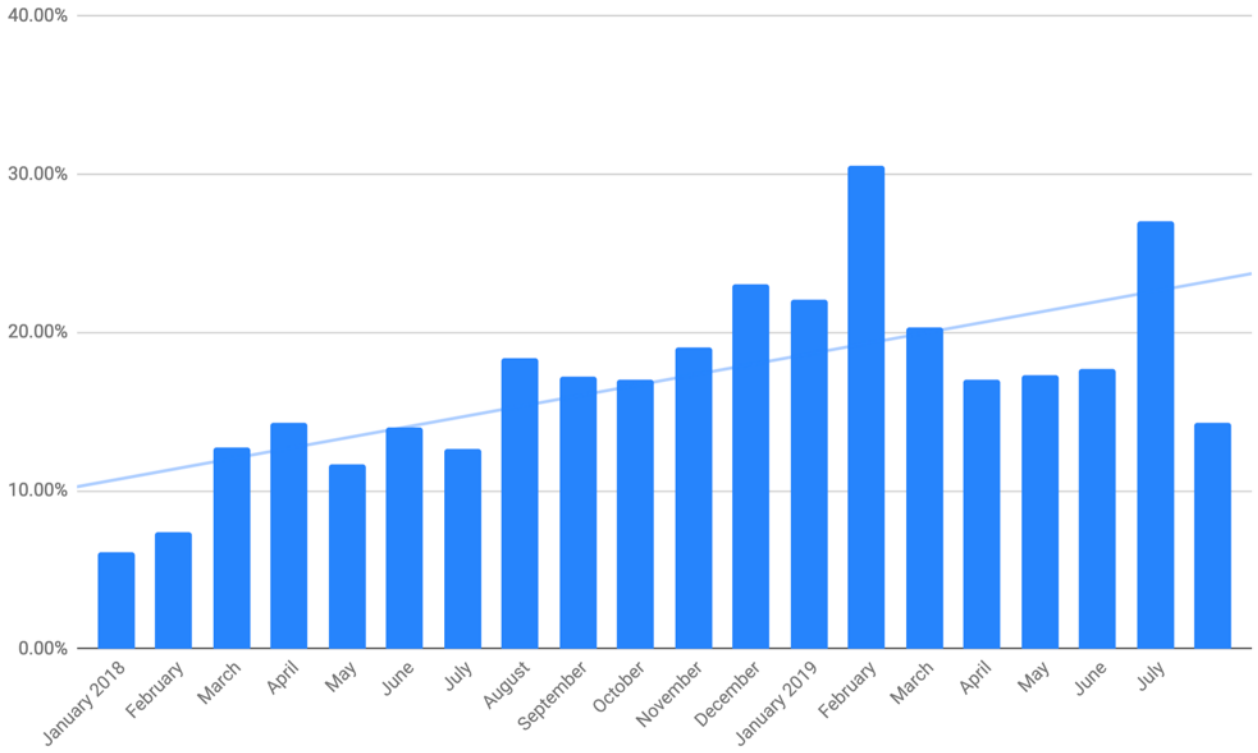
74. *Mammals*: GBIF.org (GBIF, 2018); https://doi.org/10.15468/dl.diqarx

75. *Birds*: GBIF.org (GBIF, 2018); https://doi.org/10.15468/dl.61rqcy

76. *Fungi*: GBIF.org (GBIF, 2018); https://doi.org/10.15468/dl.akftay

GBIF

DOI CITATION **IN PERSPECTIVE**



GBIF

UNPACKING A DATA CITATION

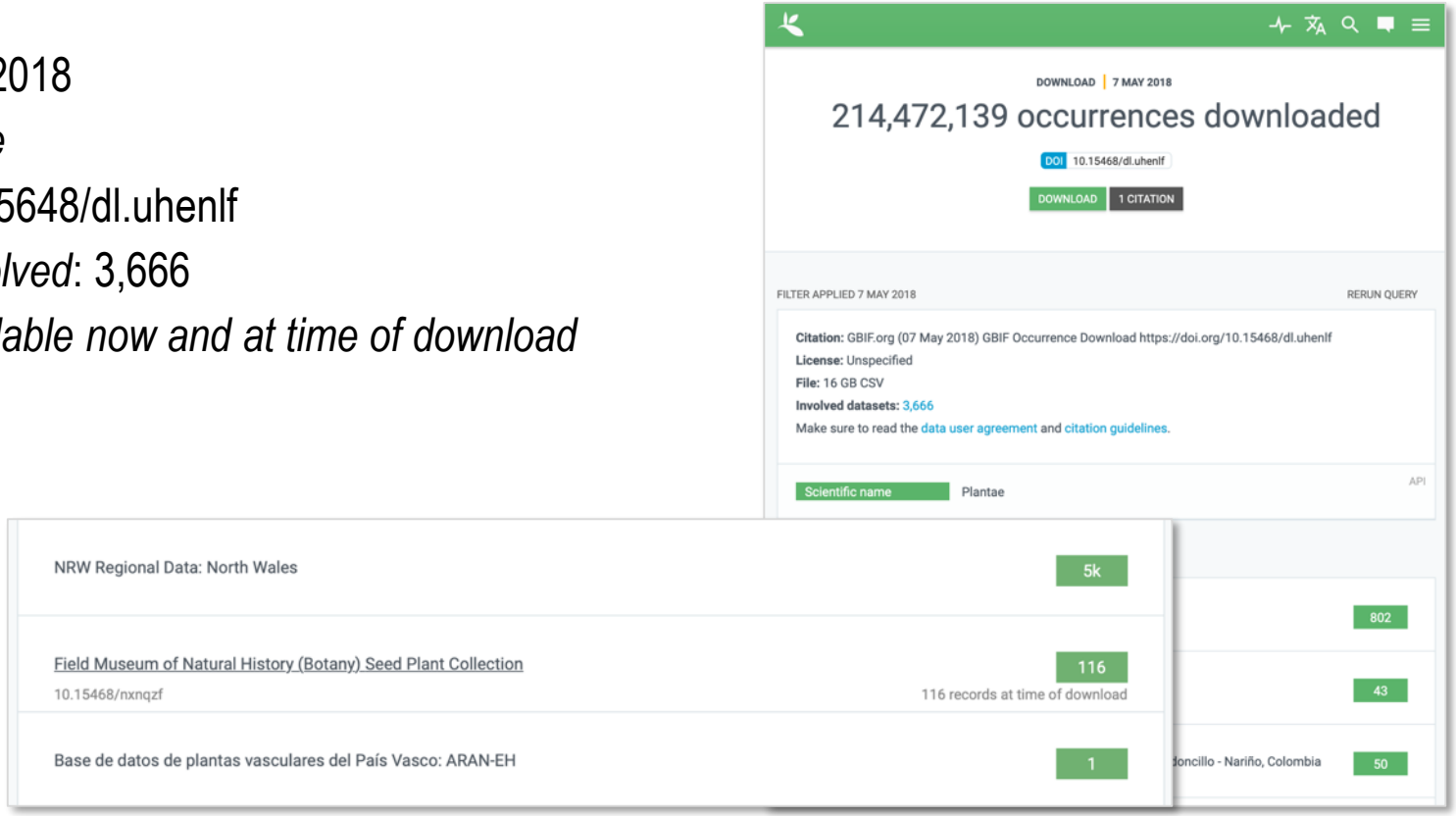
Date: 7 May 2018

Filter: *Plantae*

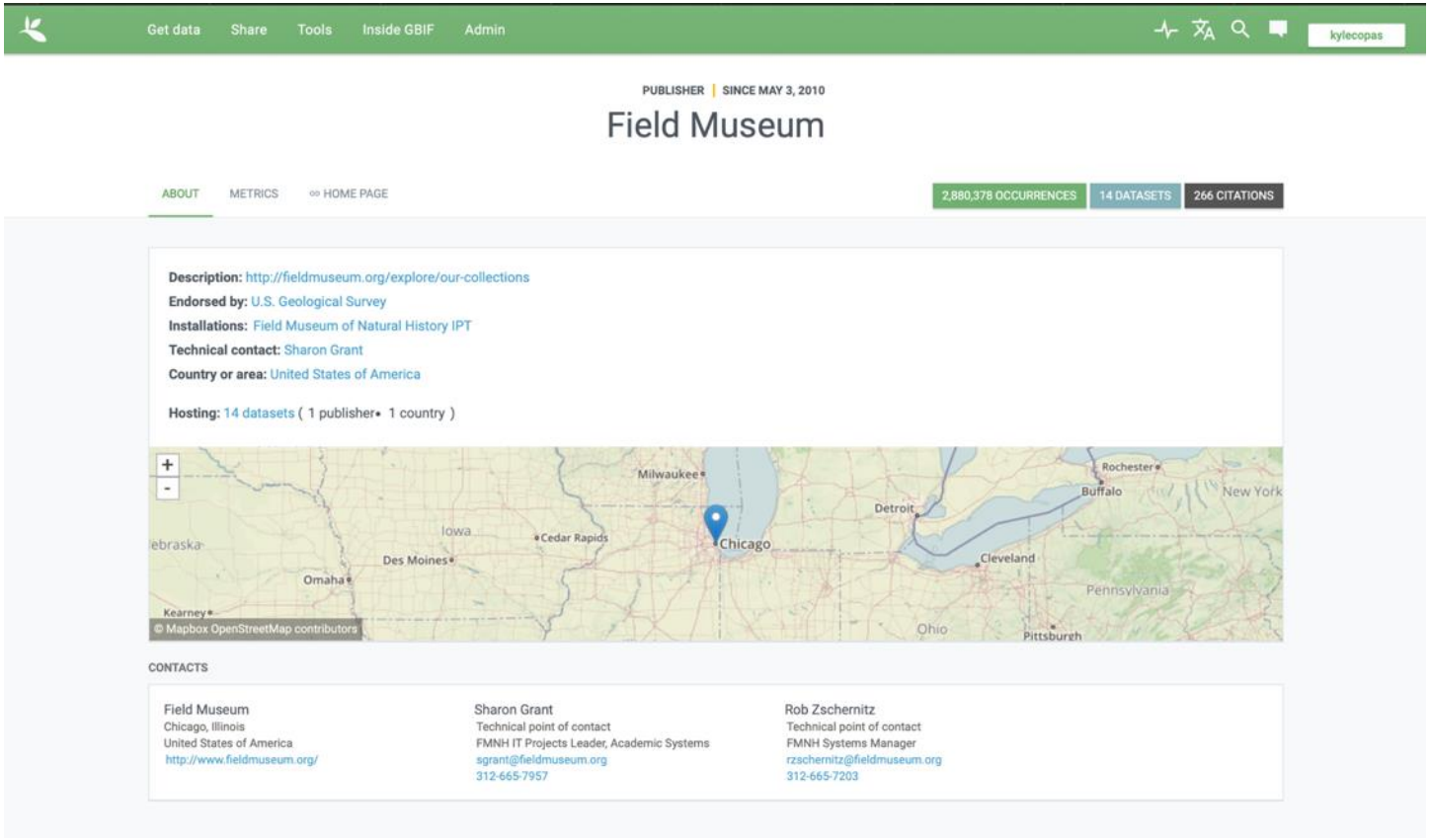
DOI: doi:10.15648/dl.uhenlf

Datasets involved: 3,666

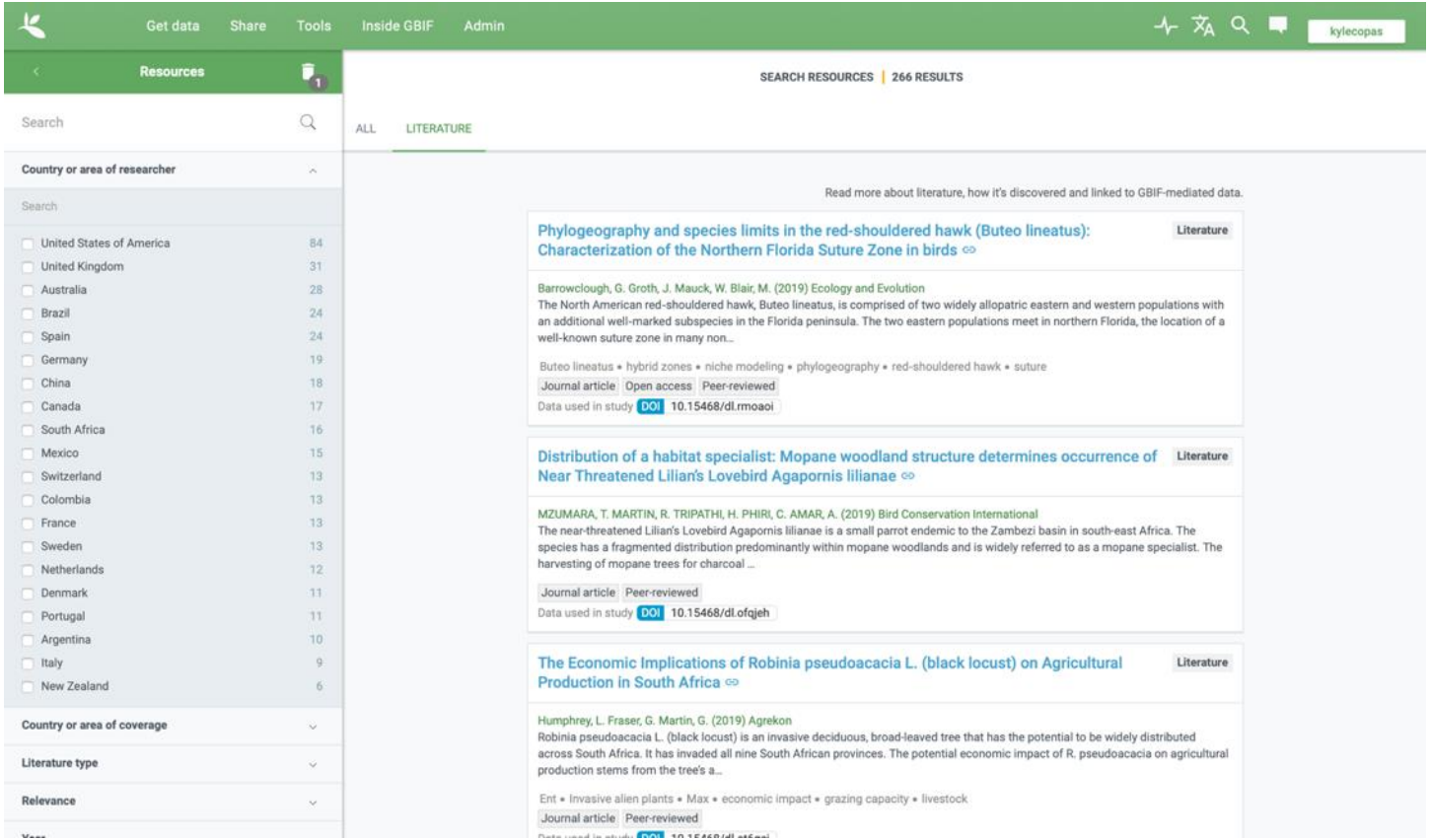
Records available now and at time of download



CREDITING INSTITUTIONS & COLLECTIONS



CITATION TRACKING



TOP TEN PUBLISHERS BY DOI CITATION

Rank	*Publisher*	*Citations*
1	National Museum of Natural History, Smithsonian	350
2	iNaturalist.org	328
3	Naturalis Biodiversity Center	312
4	Museum national d'Histoire naturelle	291
5	**Field Museum**	**285**
6	Natural History Museum (London)	283
7	California Academy of Sciences	258
8	GBIF Sweden	256
9	naturgucker.de	243
9	Senckenberg Gesellschaft für Naturforschung	243

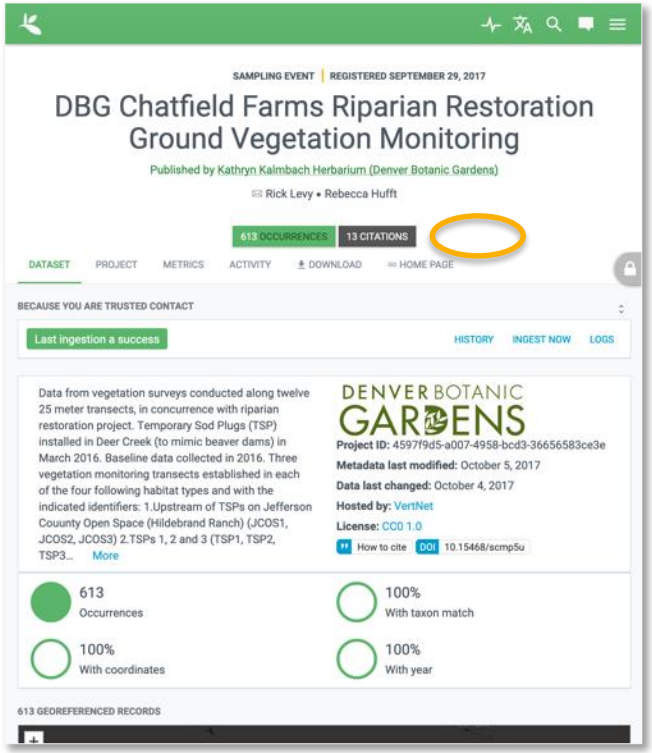
CREDITING INSTITUTIONS

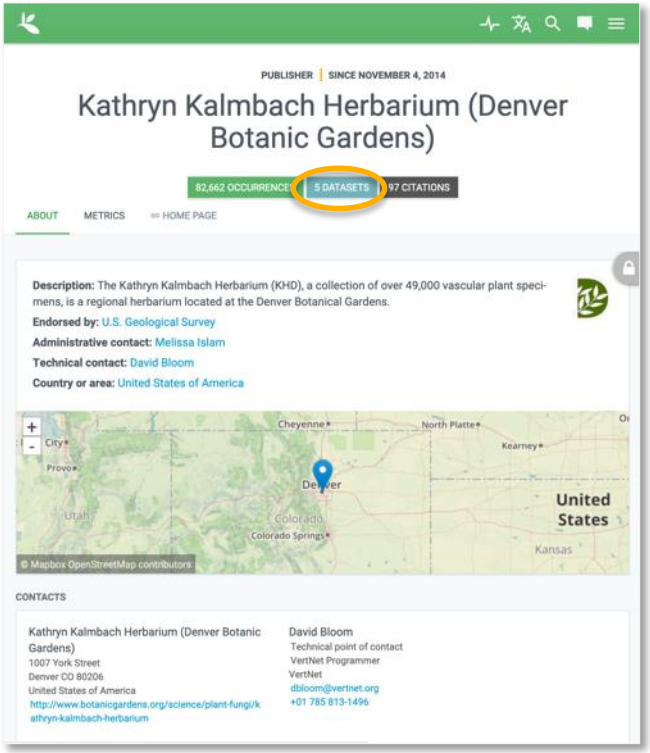
DBG Chatfield Farms Riparian Restoration Ground Vegetation Monitoring

10.15468/scmp5u

613

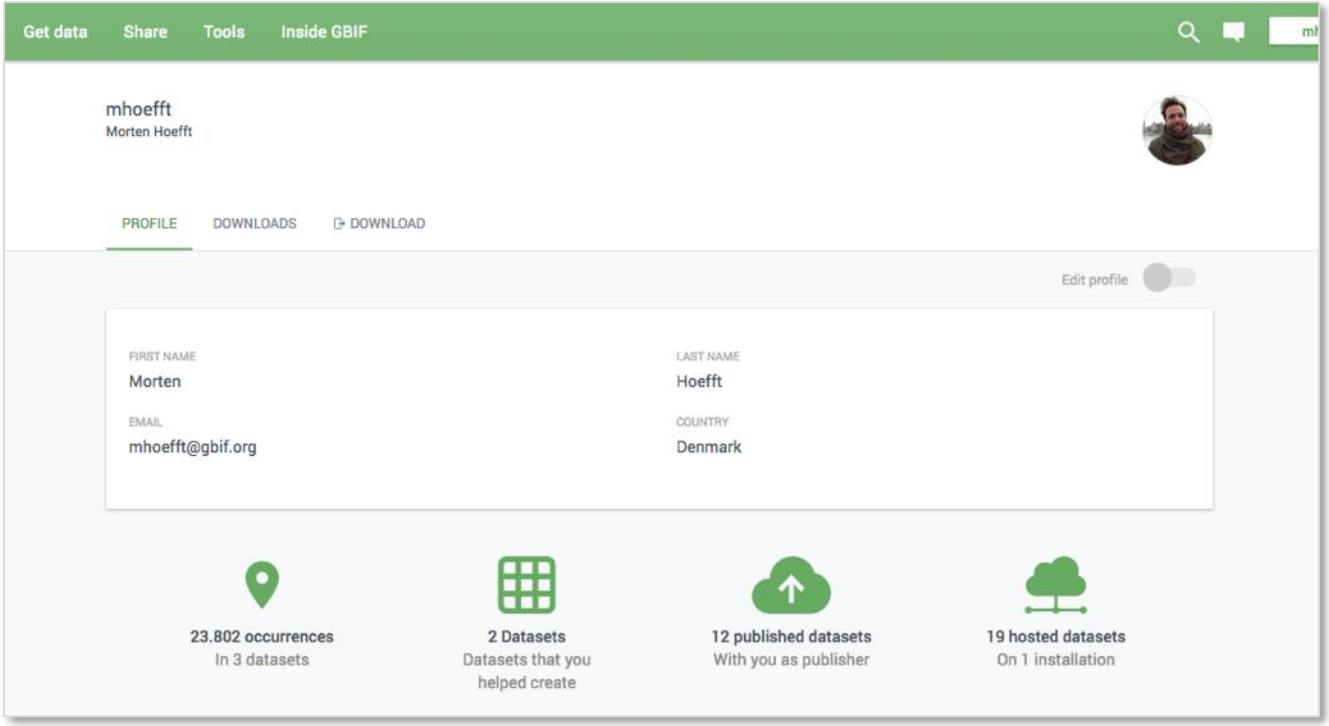
613 records at time of download





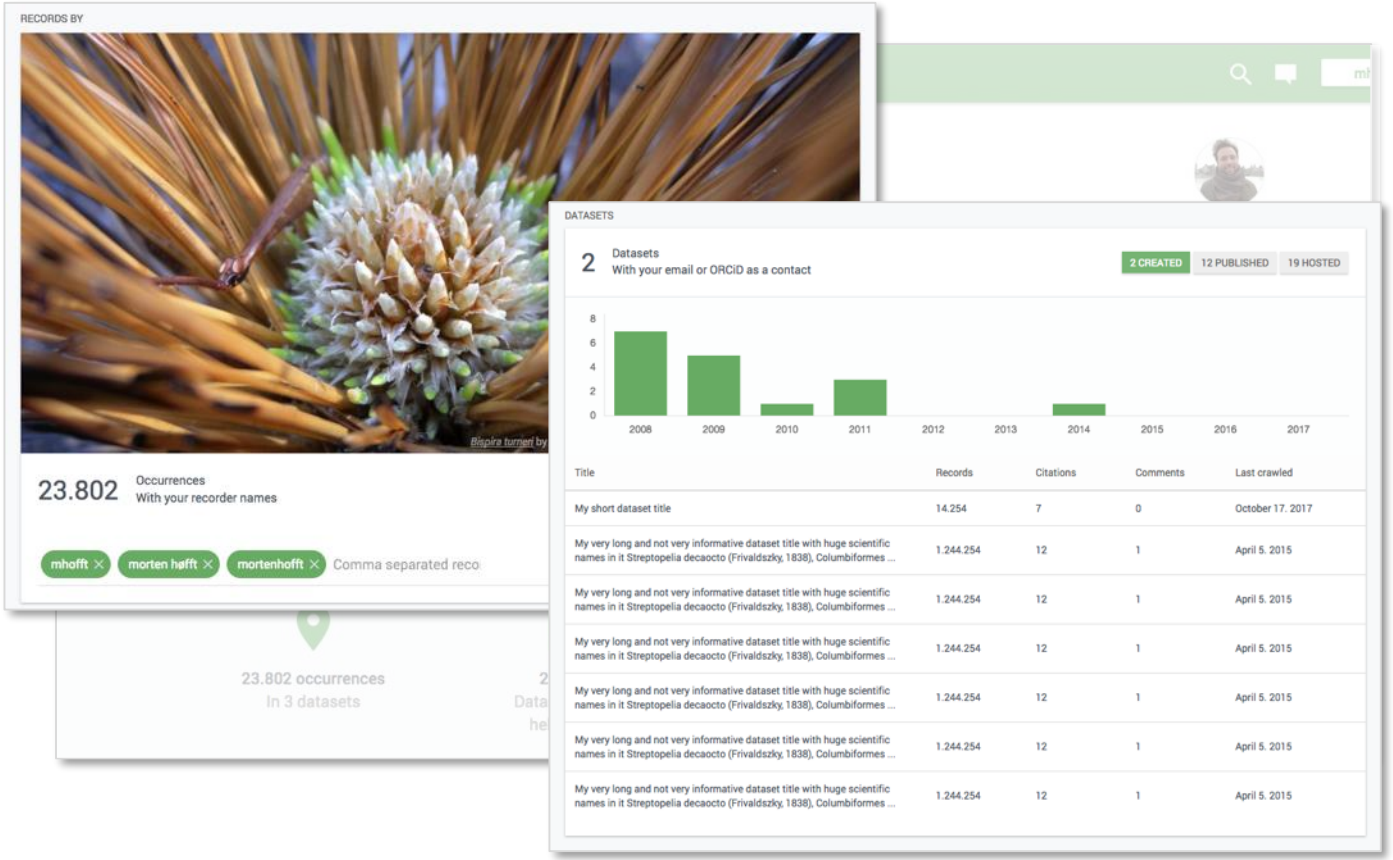
https://doi.org/10.15468/scmp5u | https://www.gbif.org/publisher/36488347-eceb-40ed-b687-406f3601d8cb

NEXT UP: **CREDITING INDIVIDUALS**



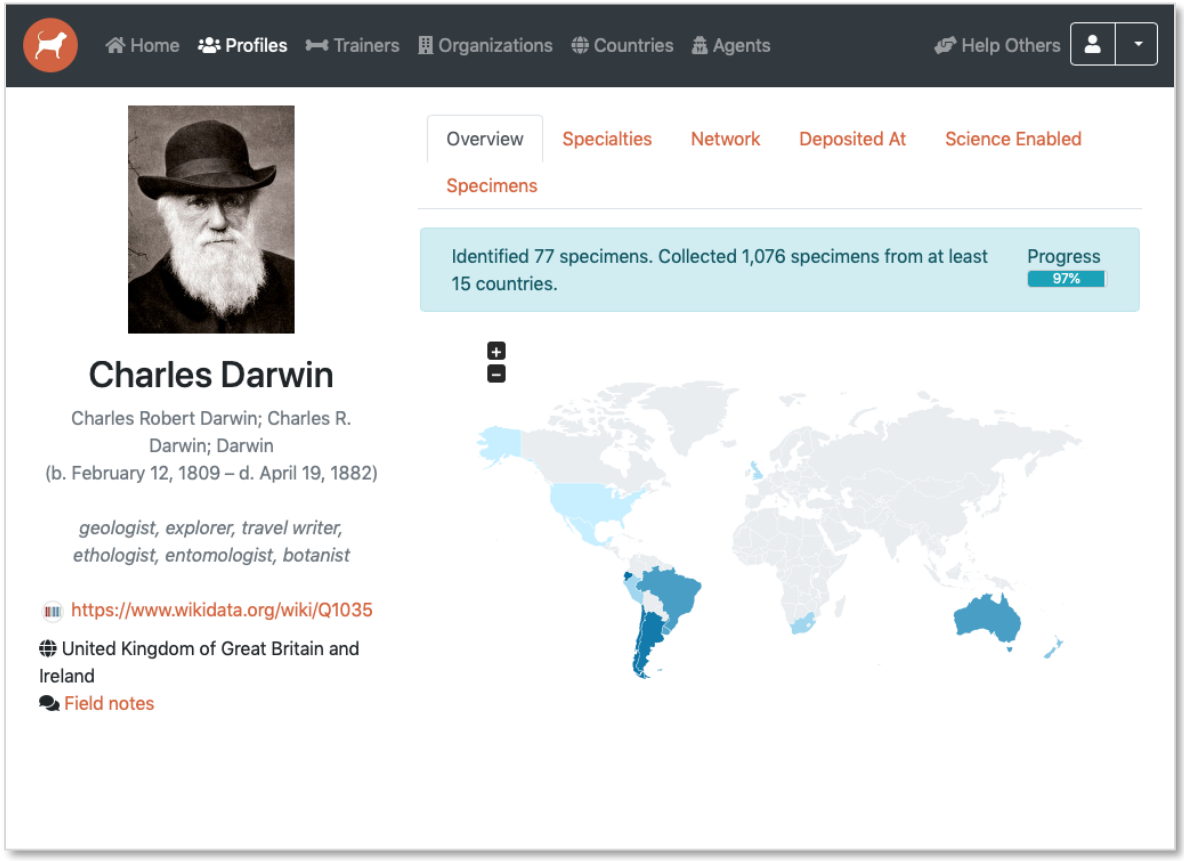
https://sketch.cloud/s/3lJ7D/a/4AJ753/play

NEXT UP: **CREDITING INDIVIDUALS**



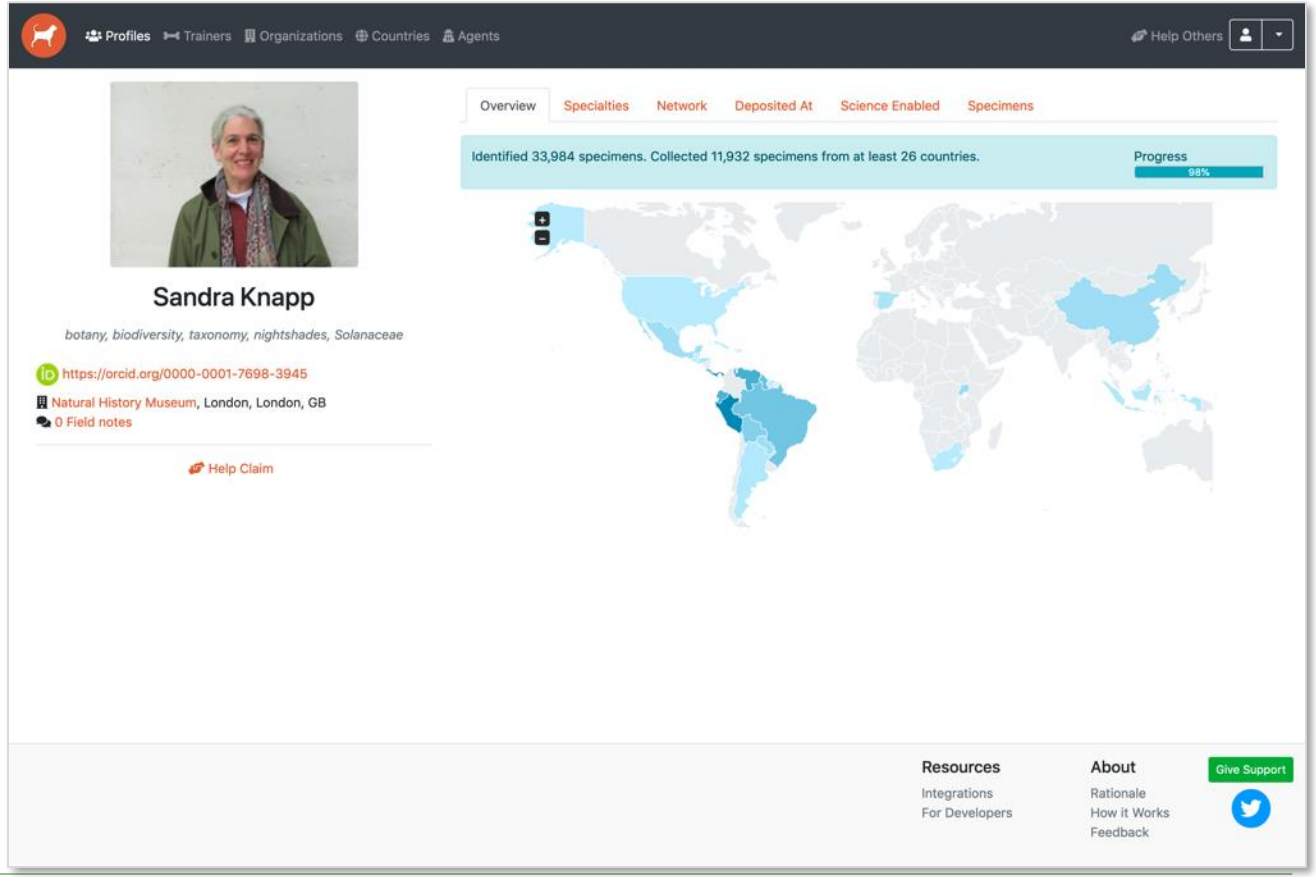
https://sketch.cloud/s/3lJ7D/a/4AJ753/play

CREDITING INDIVIDUALS: **BLOODHOUND**



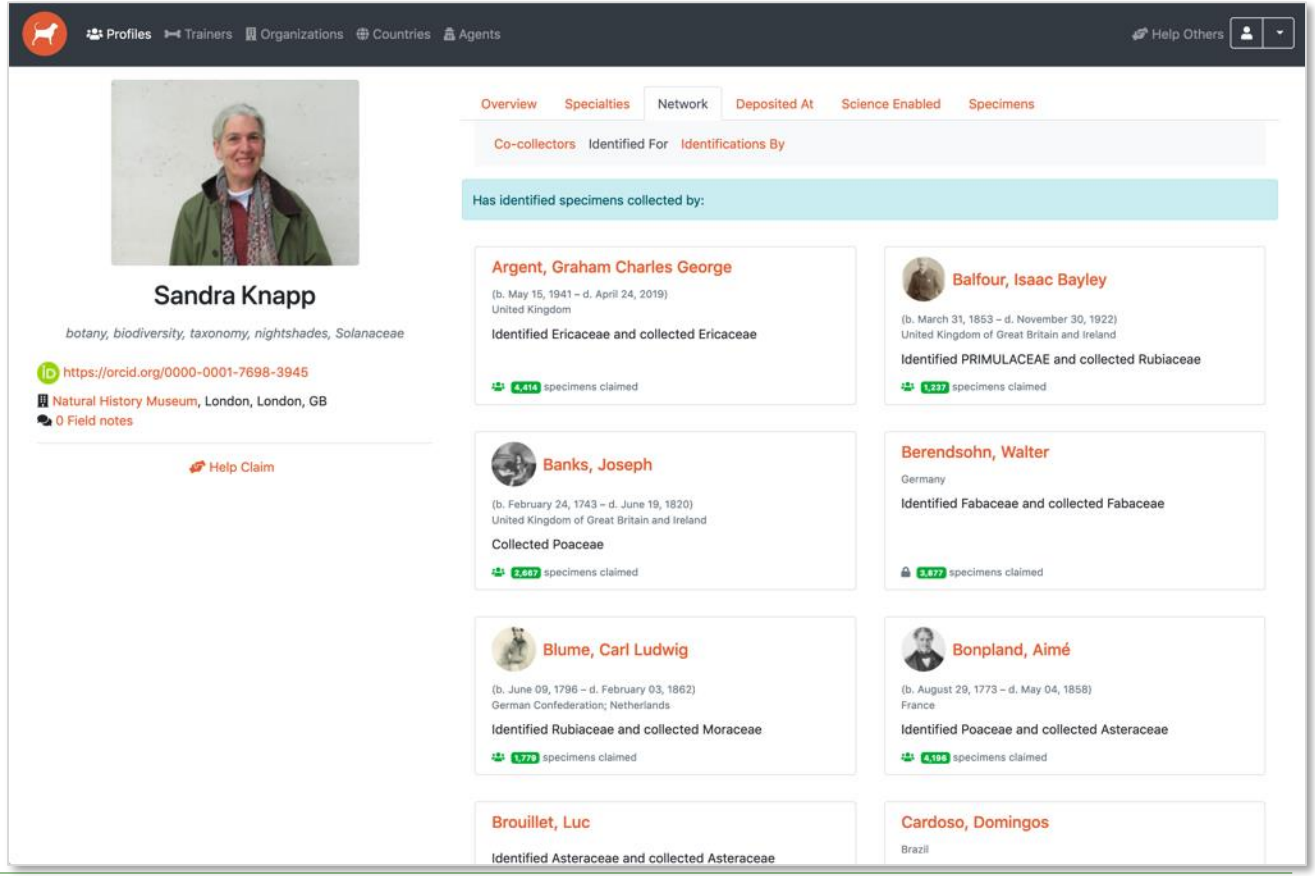
https://bloodhound-tracker.net/Q1035

CREDITING INDIVIDUALS: **BLOODHOUND**



https://bloodhound-tracker.net/0000-0001-7698-3945

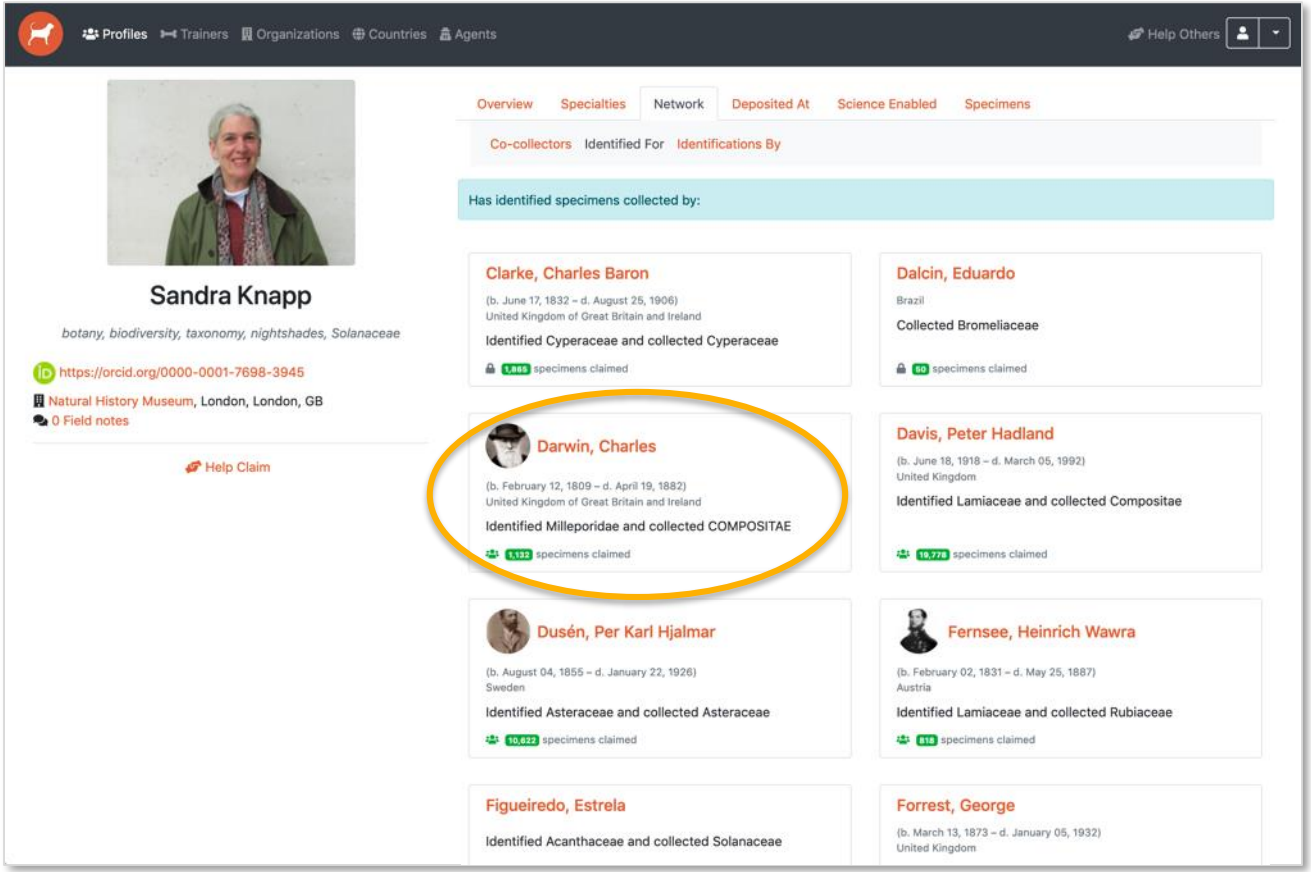
CREDITING INDIVIDUALS: **BLOODHOUND**



https://bloodhound-tracker.net/0000-0001-7698-3945/identified-for

GBIF

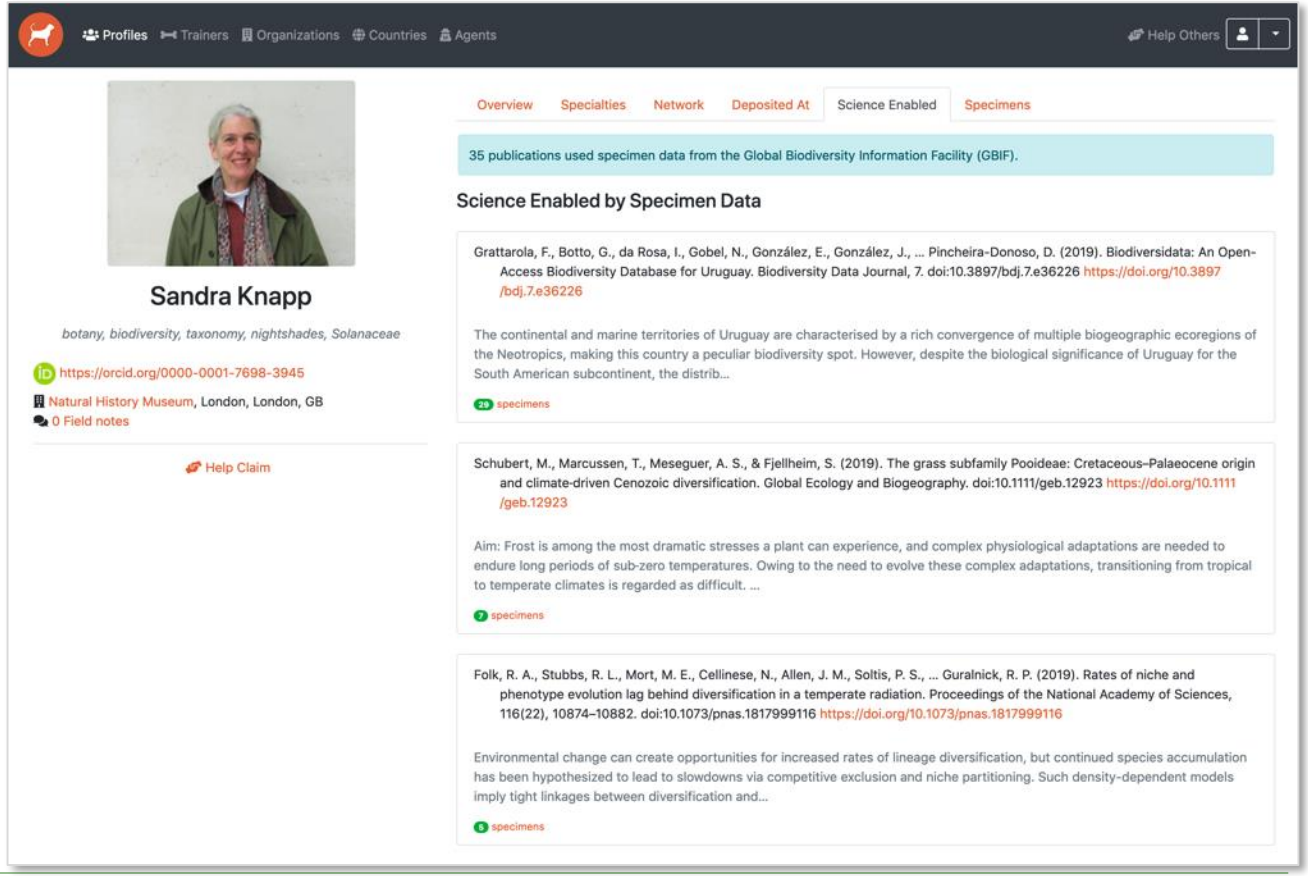
CREDITING INDIVIDUALS: **BLOODHOUND**



https://bloodhound-tracker.net/0000-0001-7698-3945/identified-for

GBIF

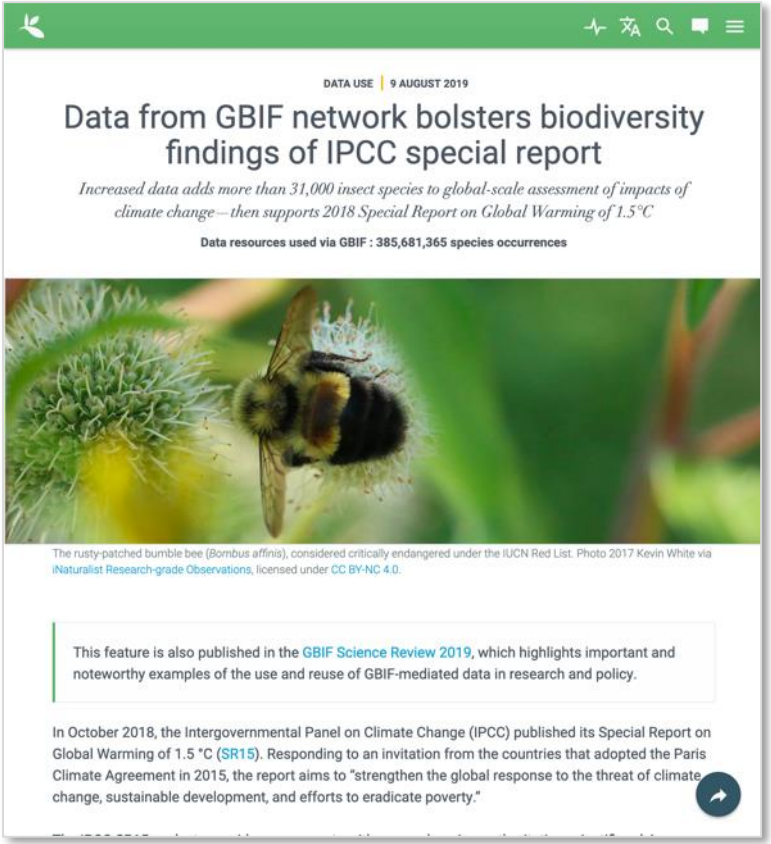
CREDITING INDIVIDUALS: **BLOODHOUND**



https://bloodhound-tracker.net/0000-0001-7698-3945/identified-for

POLICY-RELEVANT SCIENCE

'…when asked why they share their data, GBIF publishers of any of the 5,432 datasets cited in the download can unequivocally state: to **contribute to critical policy-relevant science** like the **first global-scale assessment of climate change impacts on insects** and the **IPCC's Special Report on Global Warming of 1.5 °C**.'



DATA USE | 9 AUGUST 2019

Data from GBIF network bolsters biodiversity findings of IPCC special report

Increased data adds more than 31,000 insect species to global-scale assessment of impacts of climate change—then supports 2018 Special Report on Global Warming of 1.5 °C

Data resources used via GBIF : 385,681,365 species occurrences

The rusty-patched bumble bee (*Bombus affinis*), considered critically endangered under the IUCN Red List. Photo 2017 Kevin White via iNaturalist Research-grade Observations, licensed under CC BY-NC 4.0.

This feature is also published in the GBIF Science Review 2019, which highlights important and noteworthy examples of the use and reuse of GBIF-mediated data in research and policy.

In October 2018, the Intergovernmental Panel on Climate Change (IPCC) published its Special Report on Global Warming of 1.5 °C (SR15). Responding to an invitation from the countries that adopted the Paris Climate Agreement in 2015, the report aims to "strengthen the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty."

https://www.gbif.org/data-use/51171KRV0EhwmHubpfDs3h

GBIF

Introduction to the Data Life Cycle, FAIR (**F**indable, **A**ccessible, **I**nteroperable, and **R**eusable) Data, and Data Citation

Presenter : Arctic Data Center, DataONE, Environmental Data Initiative, ESIP, GBIF, iDigBio, NEON

Amber Budden, Kyle Copas, and Kristin Vanderbilt

#datahelpdesk
Ecological Society of America 2019 ESAUSSEE
Career Fair Center in the Exhibit Hall https://esa.org/louisville/career-fair/
Monday 12 August 330-430 PM