Beyond Citations: New Metrics for Measuring the Impact of Research Data

Stacy Konkiel
Science Data Management Librarian

Digital Library Brown Bag Series
Indiana University Bloomington
10/16/2013

#dlbb

@skonkiel

INDIANA UNIVERSITY

Overview

- Definitions
- Current and changing research landscapes
- Metrics for data
  - Data citations
  - Metrics to measure repository & curator impact
  - Altmetrics for data
- Tracking your data's impact
- Discussion

Definitions

- Data
- Publication vs. publication
- Metrics
- Usage statistics
- Citations/cites
- Altmetrics
- Impact
- Repository

The Current Research Landscape

- Data aren't valued as standalone research outputs
- Data "hugging" a common practice
- Data often shared in non-machine readable forms
- Data is (sometimes) cited

The Changing Research Landscape

- Increased speed of research
- Expectation for data sharing
- Research “products” over publications
- “Big data,” e-Research, and networked science
- Data citation standards developing
- New means of measuring impact

Metrics for Data

Size dependent indicators

- Total performance indicators
- E.g. # citations, # tweets, etc

Size independent indicators

- Average performance indicators
- E.g. Journal Impact Factor

- Citation-based indicators
- Altmetrics-based indicators

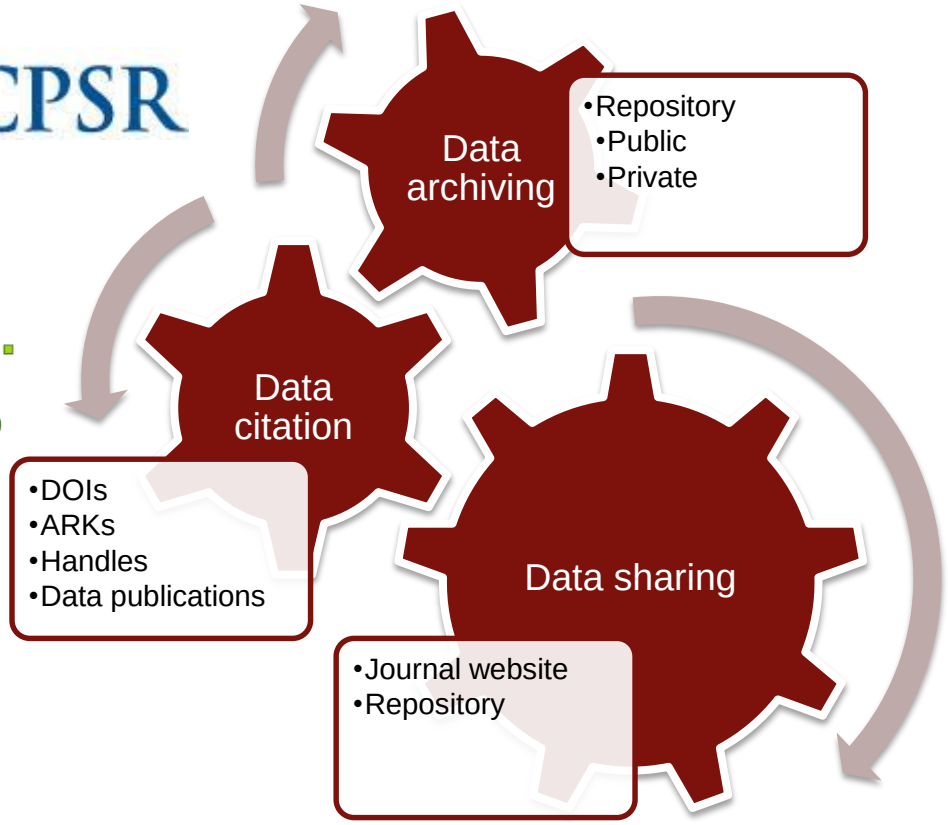
(Costas *et al*, 2013)

Metrics for Data: Challenges

- Culture
- Technical infrastructure
- Lack of standards for all disciplines
- Current practice does not address:
  - Granularity
  - Version control (except GitHub)
  - Microattribution
  - Contributor identifiers (i.e., ORCID)
  - Facilitation of reuse

(CODATA-ICSTI, 2013)

Data Citation: How it works



IUScholarWORKS Preserve and Share
research journals data

ICPSR

DRYAD

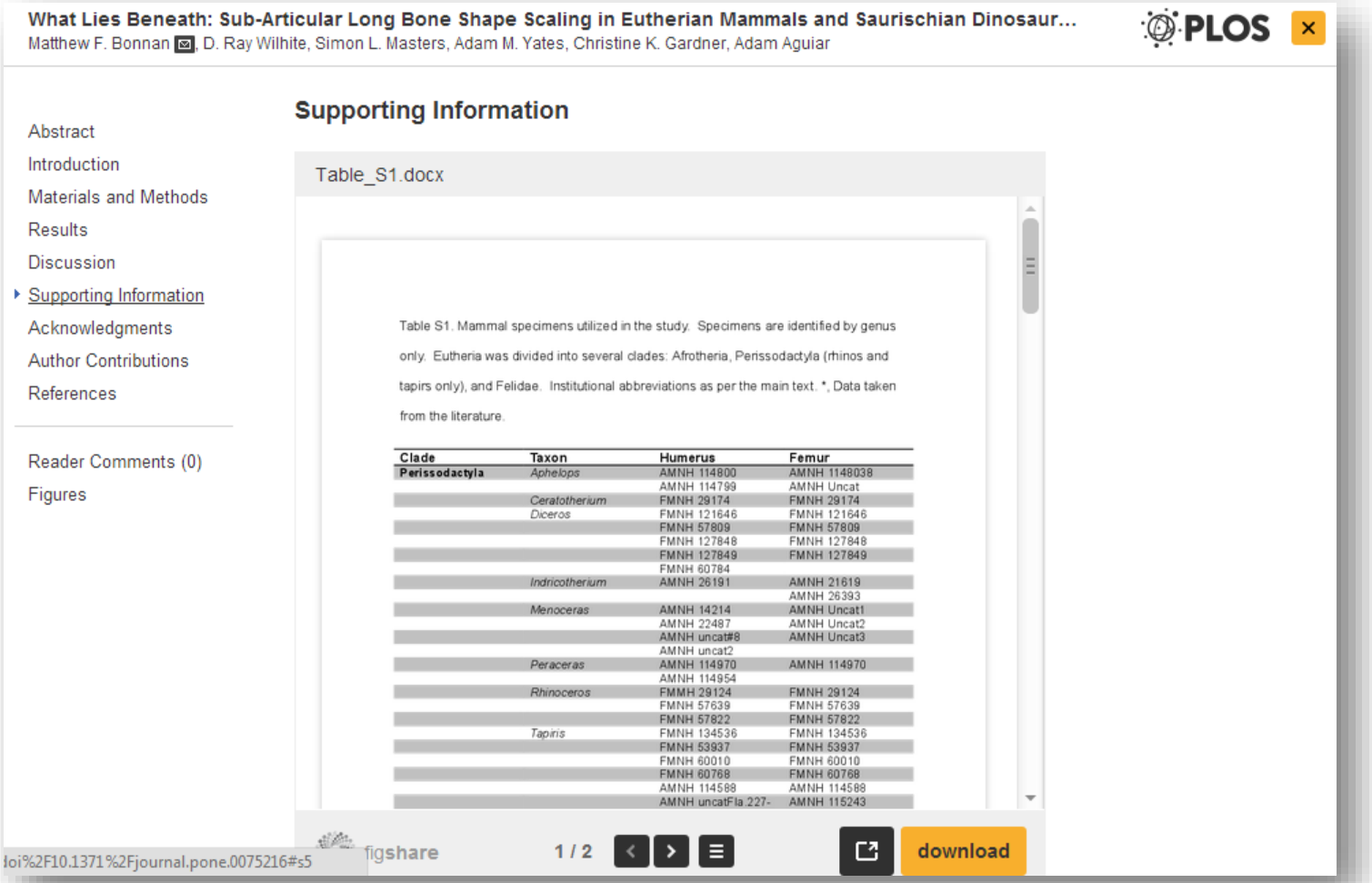
figshare
credit for all your research

DataCite

nature

PLOS ONE
A peer-reviewed, open access journal

Data Citation: How it works



What Lies Beneath: Sub-Articular Long Bone Shape Scaling in Eutherian Mammals and Saurischian Dinosaur...

Matthew F. Bonnan, D. Ray Wilhite, Simon L. Masters, Adam M. Yates, Christine K. Gardner, Adam Aguiar

PLOS

Abstract
Introduction
Materials and Methods
Results
Discussion
Supporting Information
Acknowledgments
Author Contributions
References

Reader Comments (0)
Figures

Supporting Information

Table_S1.docx

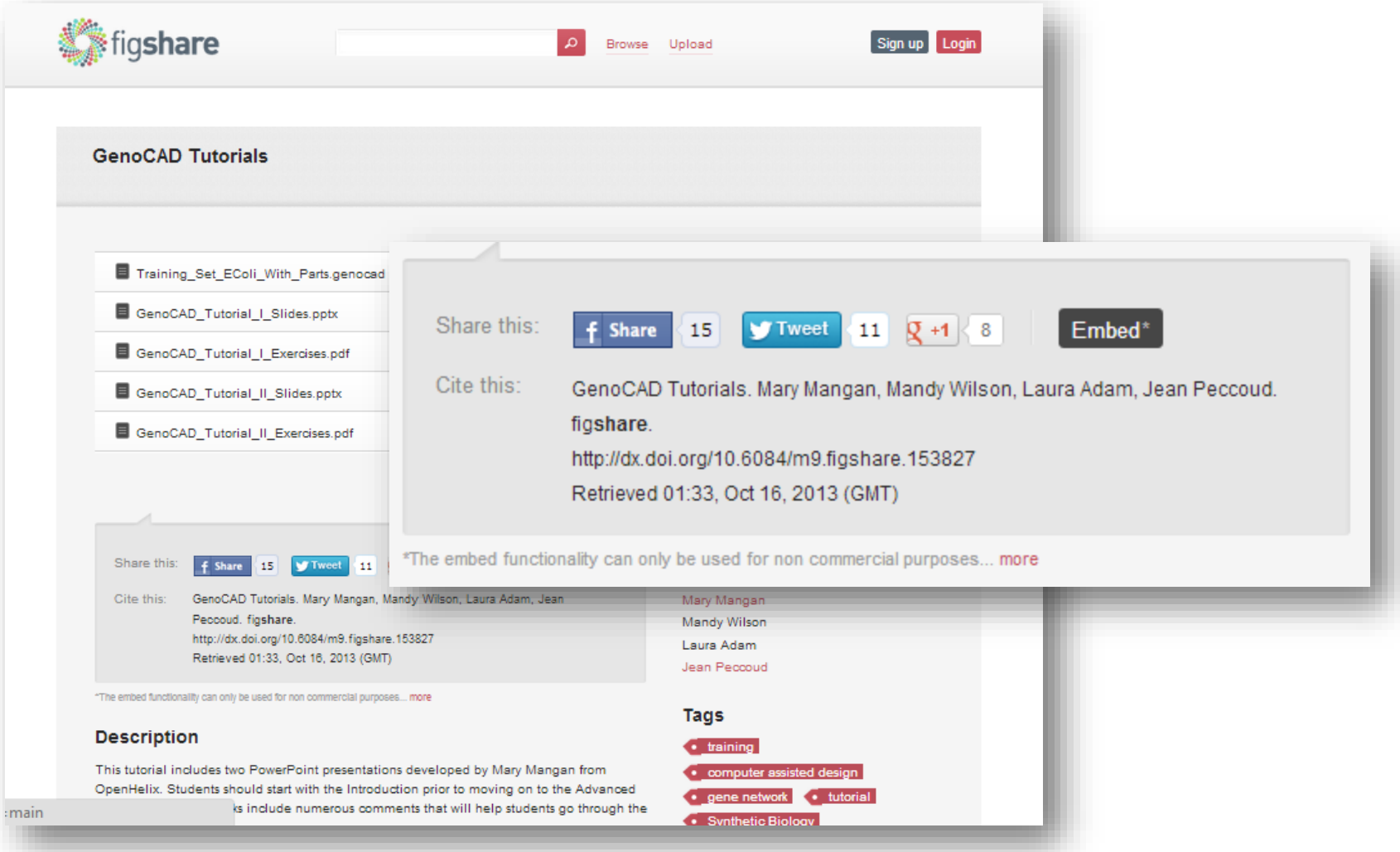
Table S1. Mammal specimens utilized in the study. Specimens are identified by genus only. Eutheria was divided into several clades: Afrotheria, Perissodactyla (rhinos and tapirs only), and Felidae. Institutional abbreviations as per the main text. *, Data taken from the literature.

Clade	Taxon	Humerus	Femur
Perissodactyla	*Aphelops*	AMNH 114800	AMNH 1148038
		AMNH 114799	AMNH Uncat
	Ceratotherium	FMNH 29174	FMNH 29174
	Diceros	FMNH 121646	FMNH 121646
		FMNH 57809	FMNH 57809
		FMNH 127848	FMNH 127848
		FMNH 127849	FMNH 127849
		FMNH 60784	
	Indricotherium	AMNH 26191	AMNH 21619
			AMNH 26393
	Menoceras	AMNH 14214	AMNH Uncat1
		AMNH 22487	AMNH Uncat2
		AMNH uncat#8	AMNH Uncat3
		AMNH uncat2	
	Peraceras	AMNH 114970	AMNH 114970
		AMNH 114954	
	Rhinoceros	FMMH 29124	FMNH 29124
		FMNH 57639	FMNH 57639
		FMNH 57822	FMNH 57822
	Tapiris	FMNH 134536	FMNH 134536
		FMNH 53937	FMNH 53937
		FMNH 60010	FMNH 60010
		FMNH 60768	FMNH 60768
		AMNH 114588	AMNH 114588
		AMNH uncatFla.227-	AMNH 115243

figshare 1 / 2 download

doi%2F10.1371%2Fjournal.pone.0075216#s5

Data Citation: How it works



Data Citation

- "First Principles"
  - Status of data
  - Attribution
  - Persistence
  - Access
  - Discovery
  - Provenance
  - Granularity
  - Verifiability
  - Metadata Standards
  - Flexibility

"Out of Cite, Out of Mind: The current state of practice, policy, and technology for the citation of data," CODATA-ICSTI Task Group, 2013. doi:10.2481/dsj.OSOM13-043

Data Citation

Elements

- Author
- Publication Date
- Title
- Publisher
- Resource Type
- Identifier
- Edition
- Location
- Feature Name & URI
- Verifier

"Out of Cite, Out of Mind: The current state of practice, policy, and technology for the citation of data," CODATA-ICSTI Task Group, 2013. doi:10.2481/dsj.OSOM13-043

Data Citation Elements

Core

- Creator(s)
- Date
- Title
- Publisher
- Identifier

Common

- Location
- Version
- Access Date
- Feature Name
- Verifier

"Data Citation Developments" [blog post], DataPub blog, John Kratz, 2013. Available at http://datapub.cdlib.org/2013/10/11/data-citation-developments/

Data Citation: Examples

Gu J.J., E.A. Smith, and H.J. Cooper. 2006. LBA-ECO CD-07 GOES-8 L4 Gridded Surface Ration and Rain Rate for Amazonia: 1999. Data Set. Available on-line [http://www.daac.ornl.gov] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A. doi:10.3334/ORNLDAAC/831

Attribution

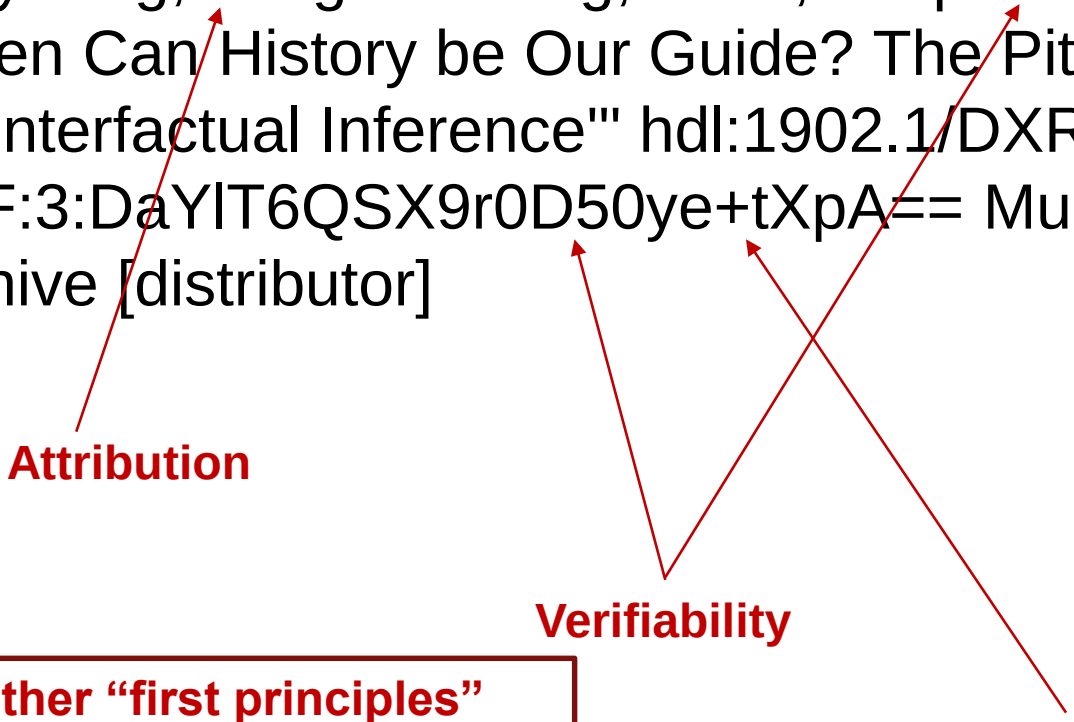
Verifiability

Persistence & Access

What other "first principles" are addressed in this citation? Core and common elements?

Data Citation: Examples

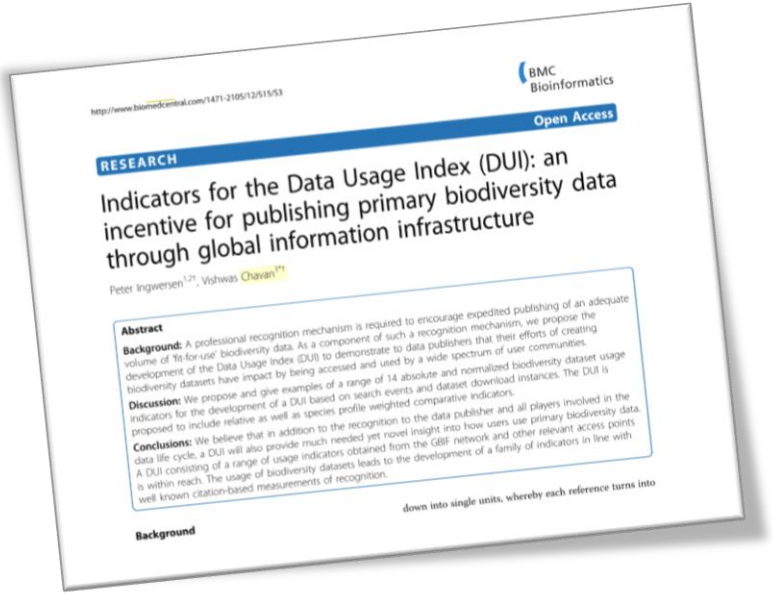
Gary King; Langche Zeng, 2006, "Replication Data Set for 'When Can History be Our Guide? The Pitfalls of Counterfactual Inference'" hdl:1902.1/DXRXCFAWPK UNF:3:DaYlT6QSX9r0D50ye+tXpA== Murray Research Archive [distributor]



What other "first principles" are addressed in this citation? Core and common elements?

Repository impact metrics

- Value that repository adds to a data set
  - Based on usage logs from Global Biodiversity Information Facility (GBIF) repository
  - 14 total absolute & relative measures
  - Not generalizable to other repositories



http://www.biomedcentral.com/1471-2105/12/S15/S3

BMC Bioinformatics

Open Access

RESEARCH

Indicators for the Data Usage Index (DUI): an incentive for publishing primary biodiversity data through global information infrastructure

Peter Ingwersen[1,2*], Vishwas Chavan[3*†]

Abstract

Background: A professional recognition mechanism is required to encourage expedited publishing of an adequate volume of 'fit-for-use' biodiversity data. As a component of such a recognition mechanism, we propose the development of the Data Usage Index (DUI) to demonstrate to data publishers that their efforts of creating biodiversity datasets have impact by being accessed and used by a wide spectrum of user communities.

Discussion: We propose and give examples of a range of 14 absolute and normalized biodiversity dataset usage indicators for the development of a DUI based on search events and dataset download instances. The DUI is proposed to include relative as well as species profile weighted comparative indicators.

Conclusions: We believe that in addition to the recognition to the data publisher and all players involved in the data life cycle, a DUI will also provide much needed yet novel insight into how users use primary biodiversity data. A DUI consisting of a range of usage indicators obtained from the GBIF network and other relevant access points is within reach. The usage of biodiversity datasets leads to the development of a family of indicators in line with well known citation-based measurements of recognition.

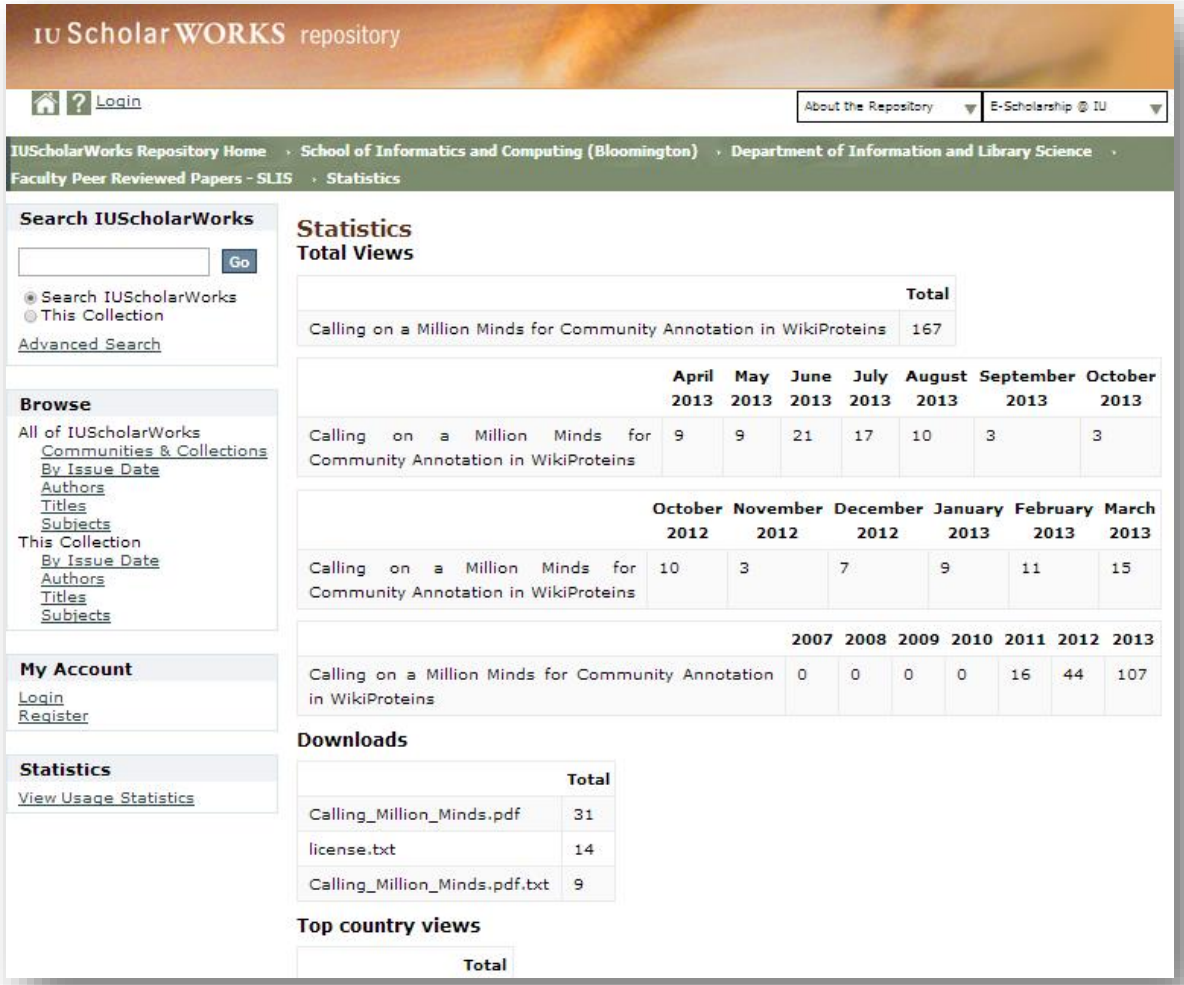
down into single units, whereby each reference turns into

Background

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* (Suppl 15), S3. doi:10.1186/1471-2105-12-S15-S3.

Indicator	Description
Searched records	Number of unit records searched/viewed (by IP address)
Download frequency	Number of downloaded records from unit
Record number	Number of records in (period; dataset(s); geographical and/or species unit)
Search events	Number of different searches (by IP address) in unit
Download events	Number of different downloads from unit
Dataset number	Number of datasets in (period, geographical etc)
Search density	Average number of searched records per search event
Download density	Average download frequency per download event
Usage impact	Download frequency per stored record per unit
Interest impact	Searched records per stored record per unit
Usage ratio	Ratio of download frequency to searched records in unit
Usage balance	Ratio of download events to search events for unit (in %)
Usage score	Ratio of unique downloaded records (U) to record number (in %)
Interest score	Ratio of unique searched records (I) to record number (in %)

Repository impact metrics



IU ScholarWORKS repository

Login

About the Repository | E-Scholarship @ IU

IUScholarWorks Repository Home › School of Informatics and Computing (Bloomington) › Department of Information and Library Science › Faculty Peer Reviewed Papers - SLIS › Statistics

Search IUScholarWorks

Go

Search IUScholarWorks
This Collection

Advanced Search

Browse

All of IUScholarWorks
Communities & Collections
By Issue Date
Authors
Titles
Subjects
This Collection
By Issue Date
Authors
Titles
Subjects

My Account

Login
Register

Statistics

View Usage Statistics

Statistics

Total Views

	Total
Calling on a Million Minds for Community Annotation in WikiProteins	167

	April 2013	May 2013	June 2013	July 2013	August 2013	September 2013	October 2013
Calling on a Million Minds for Community Annotation in WikiProteins	9	9	21	17	10	3	3

	October 2012	November 2012	December 2012	January 2013	February 2013	March 2013
Calling on a Million Minds for Community Annotation in WikiProteins	10	3	7	9	11	15

	2007	2008	2009	2010	2011	2012	2013
Calling on a Million Minds for Community Annotation in WikiProteins	0	0	0	0	16	44	107

Downloads

	Total
Calling_Million_Minds.pdf	31
license.txt	14
Calling_Million_Minds.pdf.txt	9

Top country views

	Total

Curator impact metrics

- Curation assessment
  - Data services
  - Archival content development
- Impact is based on the "specificities" of **systems** and **products**, which have more to do with the value of metrics that can be extracted than external factors

Specificities of Impact

8th International Digital Curation Conference
January 2013

The Product and System Specificities of Measuring Impact: Indicators of Use in Research Data Archives

Nicholas M. Weber, Andrea K. Thomer,
University of Illinois at Urbana-Champaign
Graduate School of Library and Information Science

Matthew S. Mayernik, Bob Dattore, Zaihua Ji, Steve Worley,
National Center for Atmospheric Research

Abstract

In this paper, we explore the construction of a suite of impact metrics that 'indicate' the use of research data. Our work builds upon a proposed data usage index (DUI) for biodiversity data, but here we consider the process of calculating impact and creating data-based metrics for the Research Data Archive (RDA) at the National Center for Atmospheric Research (NCAR). In adapting the DUI to a new research environment, we discuss the specificities of data-based metrics and their affect on calculating the impact of data curation activities in the RDA. We conclude with a discussion of

Weber, N. M., Thomer, A. K., Mayernik, M. S., Dattore, B., Ji, Z., & Worley, S. (2013). Indicators of use in research data archives. *8th International Digital Curation Conference (IDCC)*. Amsterdam, The Netherlands.

Curator impact metrics (Weber et al)

- Discovery Events
  - Homepage hits via direct links (bookmark, etc)
  - Homepage hits via query links (search engine)
    - Value of natural language description that drives web indexing
- Access Events
  - Programmatic
  - Assisted

Curator impact metrics (Weber et al)

	Code	Indicator	Explanation
1	uu(ds)	Unique Users	Unique users that downloaded data during a time window
1a	uu-p(ds)	Unique Users: Programmatic	Unique users that accessed data programmatically
1b	uu-as(ds)	Unique Users: Assisted	Unique users that accessed data via GUI or RDA Service
2	n(ds)	Number of Datasets	Number of Datasets assigned DS number by RDA
3	f(ds)	Files DS	Number of files in Dataset per time window
4	d(ds)	Download Frequency	Total number of files downloaded per time window
4a	d-p(ds)	Download Frequency: Programmatic	Files downloaded programmatically
4b	d-as(ds)	Download Frequency: Assisted	Files downloaded by Assisted users
5	hp(ds)	Homepage Hits	Home Page Hits of Data Set per time window
5a	hp-dl (ds)	Homepage Hits: Direct Link	Home Page Hits of Data Set per time window by users with direct link
5b	hp-q (ds)	Homepage Hits: Query	Home Page Hits of Data Set per time window by users link from an indexed list or retrieved by search

Table 1: RDA Absolute Indicators

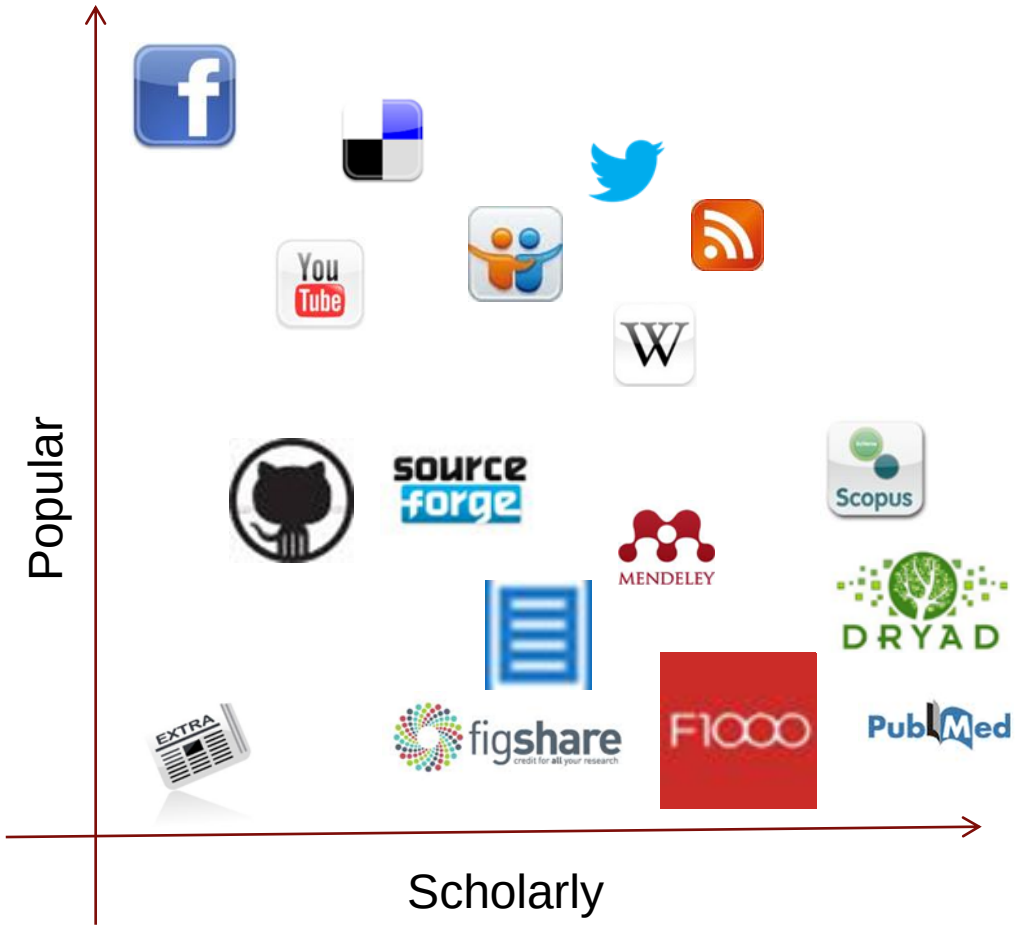
Curator impact metrics (Weber et al)

	Code	Indicator	Explanation
7	d(ds) /uu (ds)	Download Density	Average number of files downloaded per unique user
8	d(ds) / f (ds)	Usage Impact	Total number of downloaded files over total files in dataset
8a	d-ad(ds) / f(ds)	Usage Impact: Programmatic	"
8b	d-as(ds) / f(d)	Usage Impact: Assisted	"
9	d(ds) / hp(ds)	Usage Balance	Files downloaded by number of homepage hits per time window
10	hp(ds) / f(ds)	Interest impact	Total homepage hits per number of files in dataset
11	hp(ds) / uu(ds)	Secondary Interest Impact	Total Homepage hits over unique users
12	ss(ds) / d(ds)	Subset Ratio	Subset requests over total number files downloaded

Table 2: RDA Composite Indicators

Altmetrics

- Pageviews and downloads
- Citations
- Social bookmarks
- Social media
- Trackbacks
- Comments



(via Konkiel, 2012: http://hdl.handle.net/2022/14714)

Altmetrics

- (+) Capture information at the (output, researcher, institution, state, national) level
- (+) Quick to accumulate
- (+) Correlate with traditional metrics
- (+/-) Can give limited context

Altmetrics

Table 1. Types of altmetrics and examples

Altmetric Type	Description	Examples
Shares	Posted publicly in order to share news of research article or outputs	Twitter, Topsy, Facebook, reddit, news articles, blog posts, Google+, YouTube, Figshare, Mendeley
Saves	Saved on social bookmarking sites or favorited on social media and social coding websites	Mendeley, CiteULike, Delicious, Github, Twitter, Slideshare
Reviews	Discussed with additional commentary added	Faculty of 1000 (F1000), blog posts, article comments, Facebook comments
Adaptations	Creation of derivative works using an article or other output	Github
Social usage statistics	Downloads or views on web services and social media sites	Figshare, Slideshare, Dryad, Facebook, YouTube

Table 2. Altmetrics and their correlations to traditional measures of impact

Metric	Correlation to Traditional Impact
Twitter mentions	Citation counts [4] [12]
Facebook wall posts	Citation counts [12]
Mendeley & CiteULike saves	Citation counts [1] [7] [8]
F1000 Reviews	Citation counts [7]
Expert blog posts	Highly cited papers [11]; Journal Impact Factor [5]
News articles	Citation counts [12]
Wikipedia citations	Citation counts [3] [9]

via Konkiel, S. (2013.) “Altmetrics: A 21st Century Solution to Determining Research Quality.” *Online Searcher*, July/August 2013.

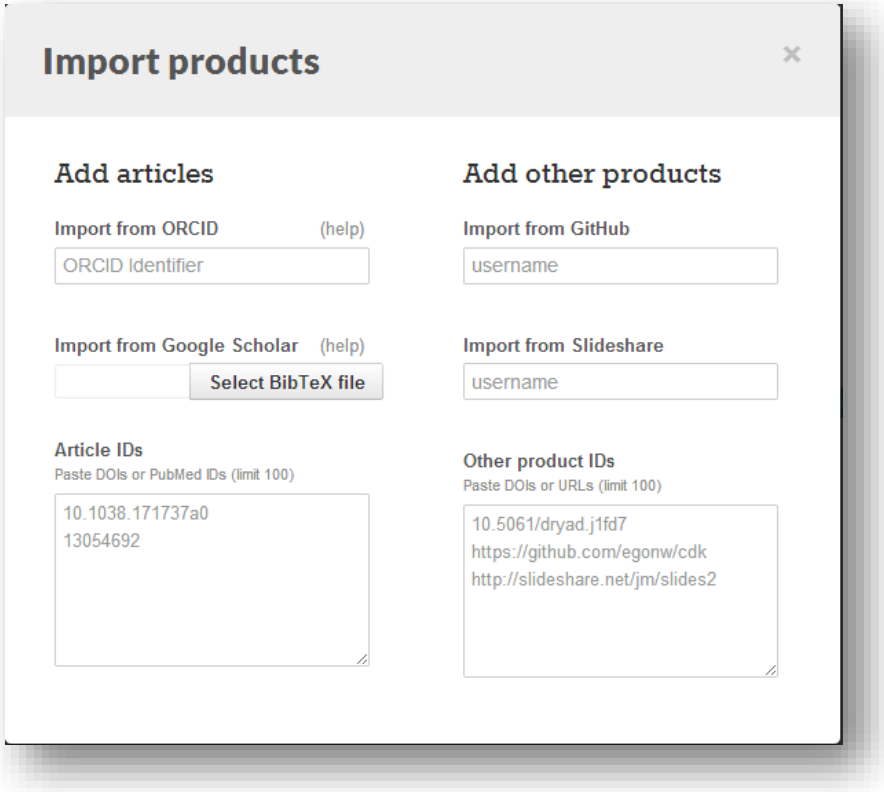
Altmetrics & Data

- Repository
  - Pageviews
  - Downloads
- Dryad
  - Total downloads
  - Package views
- Figshare
  - Downloads
  - Views
  - Shares
- Github
  - Forks
  - Stars
- PLOS
  - Figure views
  - Supp-data views
- PubMed Central
  - Suppdata views
  - Figure views

Source: ImpactStory

Tracking your data's impact

- Register DOIs for your data
- Upload to a repository
  (Or a journal, if you must)
- Create an ImpactStory report



Thank you!

Stacy: skonkiel@indiana.edu // @skonkiel

Download this presentation & handout:
https://scholarworks.iu.edu/dspace/handle/2022/15458