Apr 3rd, 12:00 AM

Panel Discussion presentation: "Value-based Indicators for Reuse & Their Implications for Data Curation"

Nic Weber
University of Illinois at Urbana-Champaign

Value-based indicators for reuse & their implications for data curation

03 April 2013
UMass + New England Area Librarians eScience Symposium

Nic Weber

Tiffany Chao, Karen Baker
Andrea Thomer & Dr. Carole Palmer

CIRSS

GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE
The iSchool at Illinois

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

Overview

I. The Data Practice Working group

-What we talk about when we talk about Value

II. Some research findings

-Qualitative Case : The Data Conservancy

-Quantitative Case: NCAR’s Research Data Archive

Data Practices Research Group

Dr. Melissa Cragin, Tiffany Chao, Karen Baker, Andrea Thomer & Dr. Carole Palmer

Qualitative & Quantitative Studies of Data production and use

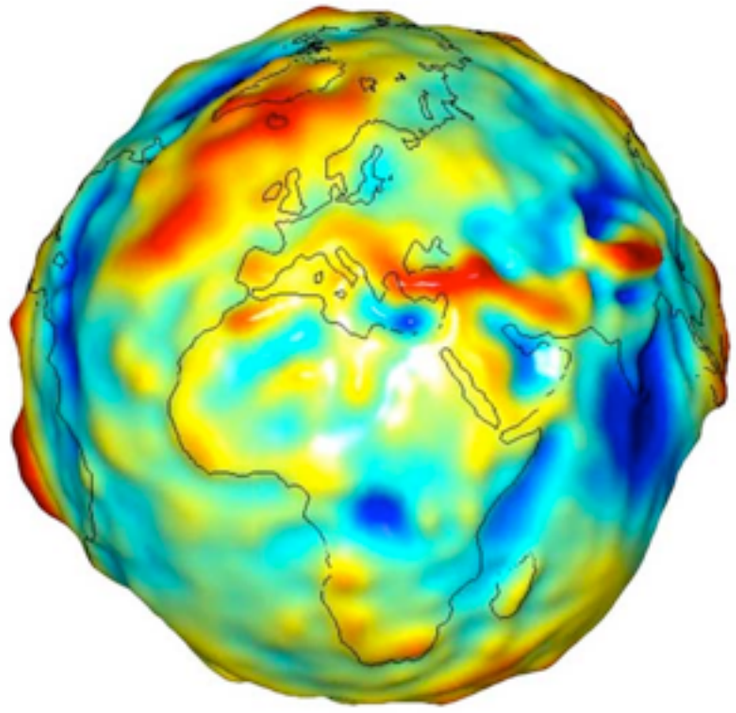
Aims: Inform development of curation services and systems

Focus: Long-tail, heterogeneous, ‘small’ data-intensive science

12,025 NSF grants awarded in 2007 = $2,865,388,605

Range	$300,000 - $38,131,952	$579 - $300,000
	20%	80%
Number of Grants	2405	9621
Total dollars	$1,747,957,451	$1,117,431,154

Value-based Indicators



(Some of our working assumptions)

If long-term preservation is the goal, research libraries and data centers want to make targeted investments in high value data collections.

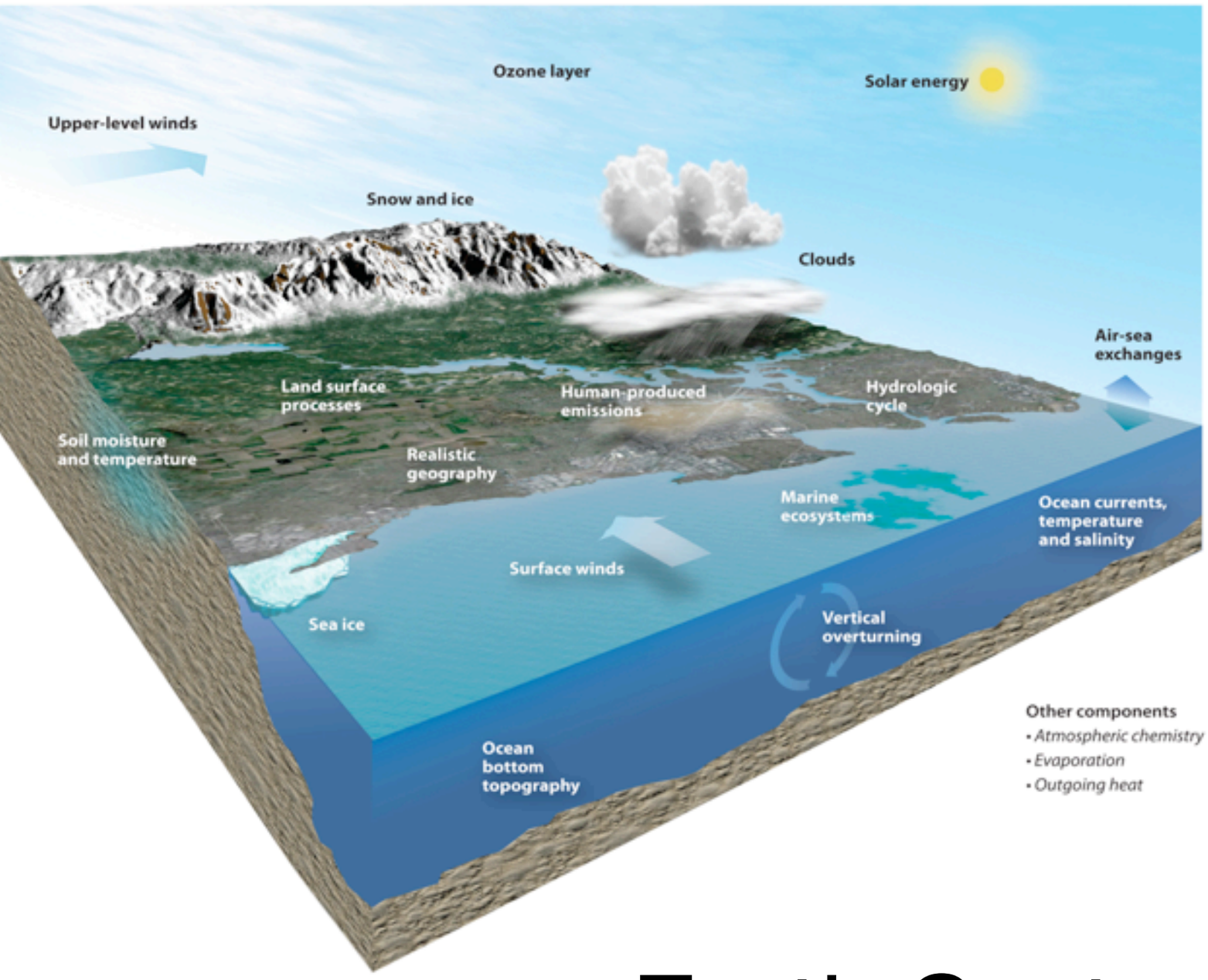
The value of data is a socio-technical phenomenon.

The view of data value is a relational one.

Value is not necessarily dependent on quality, size, scale, support, rarity or expense.

The value of data increases with use.

Qualitative Studies of Re-Use



"…the global earth environment can be understood only as an interactive system embracing the atmosphere, oceans, and sea ice, glaciers, and ice-sheets, as well as marine and terrestrial ecosystems"

Asrar, Kaye & Morel, 2001

Earth System Science (ESS)

Macro-perspective of Earth Science work:
Data Communities, Evidential Cultures.

Sub-disciplinary Profiles

	Soil Ecology	***Volcanology***	***Stratigraphy***	***RS Engineering***	***C/O Modeling***
Study approach	Biotic and abiotic properties of soil	Chemical and textural properties of rock samples combined with geospatial data	Range of signals compared to refine the geological time scale	Prototyping and designing field sensors to optimize field data collection	Computational or mathematical modeling of aquatic dynamics.
Kinds of data used	Physical soil samples, maps (paper & digital), biological species inventory, lab-based outputs	Whole rock samples; thin slices of rock samples on glass slides; chemical data; maps	Numerical data and graphs pulled from papers; physical samples; chemical, radioactive isotope, and astronomical cycle data	Autonomous field measurement of sensor and environmental data recorded on data loggers or transferred directly to a database	Water sample, meteorological, and remote sensing data downloaded; diverse models' output at many spatial & temporal scales
Patterns of data use	Systematic review of data for quality where values are checked against multiple sources	Iterative reference to & comparison of data sources, including chemical data, field notes, papers & maps	Highly iterative comparison of datasets and modeling of signals of time	Regular review of data for investigating various sensor configurations and contexts of data collection	Irregular patterns of use, based on need for model calibration or benchmarking for reliability
Norms of data re-use	Informal sharing of processed data and methods, though perceptions on re-use vary	High expectation of data re-use, particularly with physical samples and thin sections	Moderate expectation of re-use aiming to find new ways of determining geological time scales for re-use	Diverse, informal re-uses: optimizing sampling design; providing data to project researchers; or for public posting	Informal sharing of data inputs and software code; Informal and formal mechanisms for re-use and sharing of model

Value types: Frequent Data Re-Users

"…you have to go back to the data gatherer and ask them, "What's this (cell) value? This doesn't seem to be right.? Do you remember what happened? Did a shark hit your boat or something?" …the quality control doesn't exist really well. So one has to work back and forth with the data collector." Ocean Modeler

Value types observed : Verification, Depth of Description, Equivalence

Implications for Systems & Services Development:

Enable users and curators to trace provenance and context of production.

Data change in value based on the context of communities of practice- and participation in communities of practice are more dynamic than we often assume.

Identifying data producers (authorship ?) is burgeoning issue of importance for meaningful re-use. We have to come up with sound guidelines, and be able to establish persistent ways of tracking data producers (ORCID IDs!)

Value types: Data producers

"We have people who are participating in triathlons…and they want to know about the water temperature and want to know about patterns of temp change. We've had Search and Rescue teams download our data to be able to predict what will be going on…fishermen will request data to look at trends…We also have industry people, need to know what the typical water level will be so they can get their boat in there."
Sensor Engineer

Value types observed: Regenerative, Malleability

Implications for Systems & Services:

Design infrastructures to recapture secondary products, serve flexible / shifting client base. Discovery is still ad-hoc, back channel.

Find ways for signals of value to be consumed by both curators and re-users.

Types of Value

Re-users (How they describe valuable data for their own work)

Verification: This data helps me trust / refute existing data source

Depth of description: This data adds to basic understanding of existing data source

Equivalence: This dataset is the same (content) as that data source

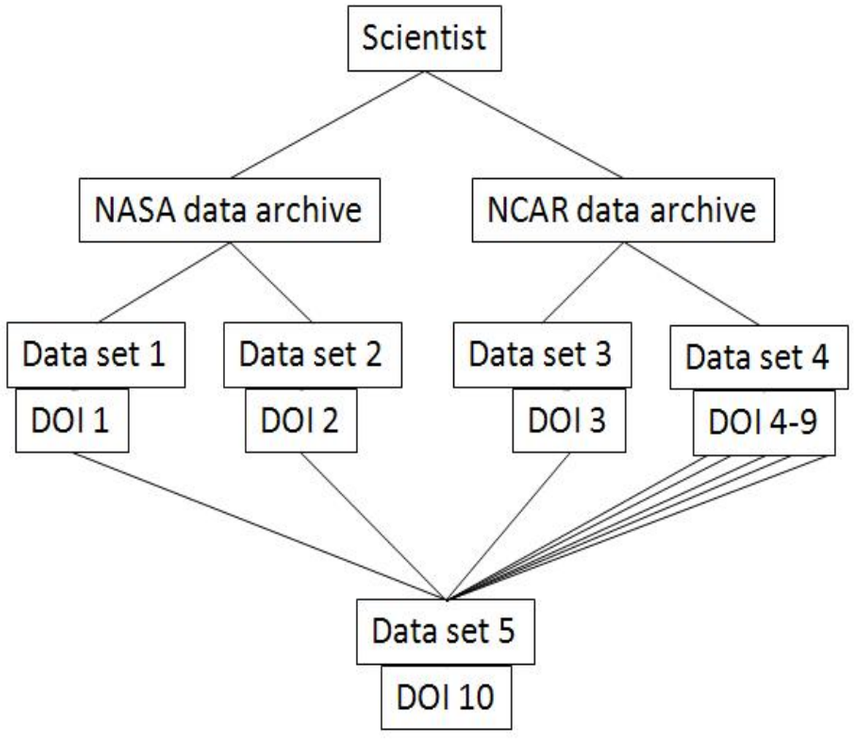
Producers (How they imagine their data having value)

Regenerative: Do data have “reach” beyond original intention or application?

Malleability: How flexible or fragile are data to new application, new domain or new method?

Quantitative Studies of Re-Use

Data Citation @ NCAR



Slide c/o Matt Mayernik

CISL Research Data Archive

Managed by NCAR's Data Support Section
Data for Atmospheric and Geosciences Research

RDA

Holdings > 1.3 PB , static and dynamic datasets including...

Atmospheric and oceanographic observational data, weather prediction model output, gridded analyses and reanalyses, climate model output, and satellite derived data

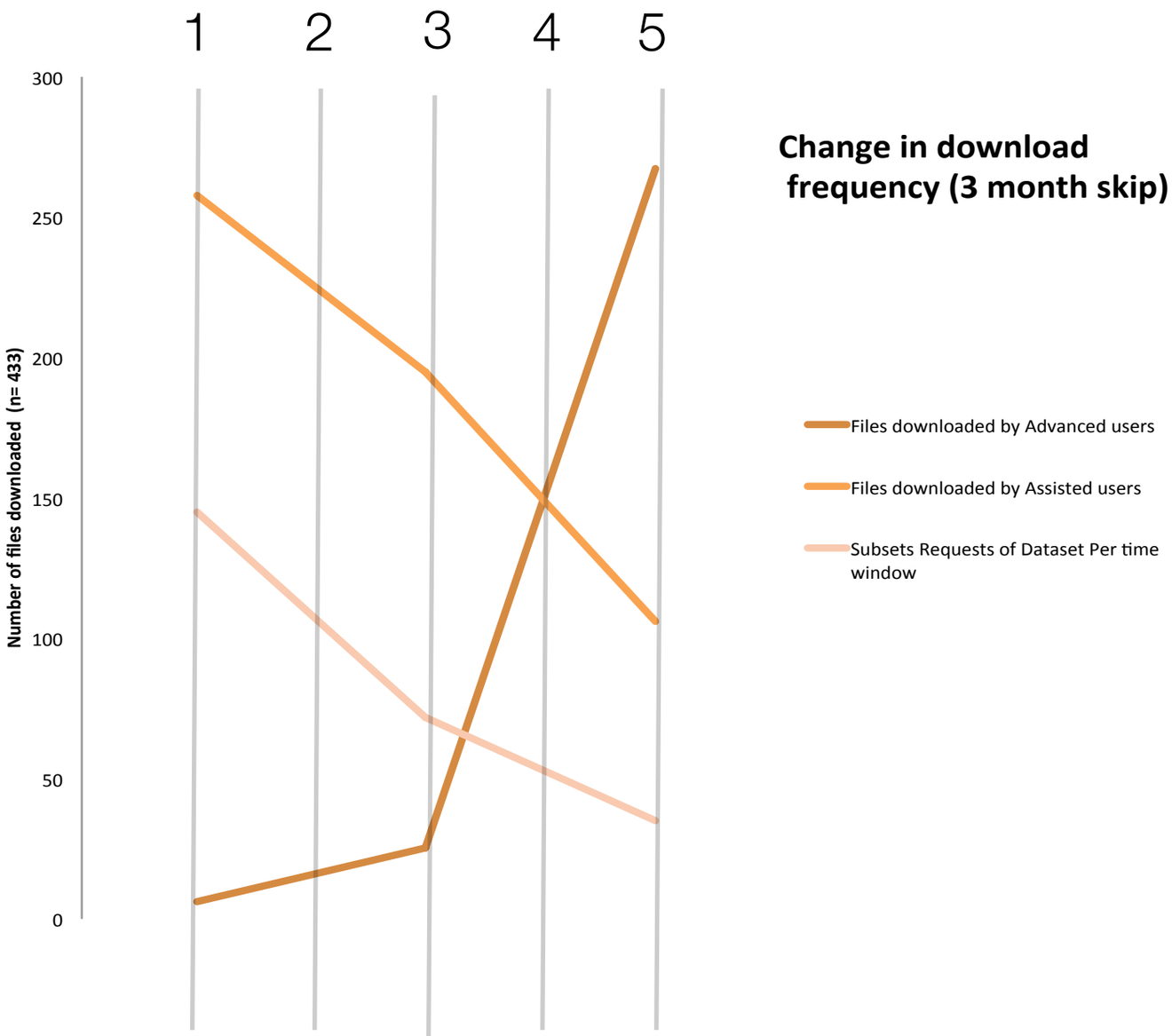
2012 served ~1 PB to ~1500 unique users from 127 different countries.

The Challenge

*"**2012** served ~1 PB to ~1500 unique users from 127 different countries"*

Impressive, but not very meaningful.

At best it's an incomplete picture of the RDA's curation work, and it's impact on Earth Science domain.



Is the RDA less helpful now?

*Subsets decreased

*Assisted users downloading *decreased*

*Advanced users downloading *increased*

The Data Usage Index

Originally developed for Biodiversity database (GBIF) by Ingwersen and Chavan (2011)

Takes suite of archive-user interaction metrics

Uses these as indicators or proxy of impact, that can be later combined in unique ways to demonstrate value

Indicators are standard across datasets, allows for comparisons across different data types, and time periods.

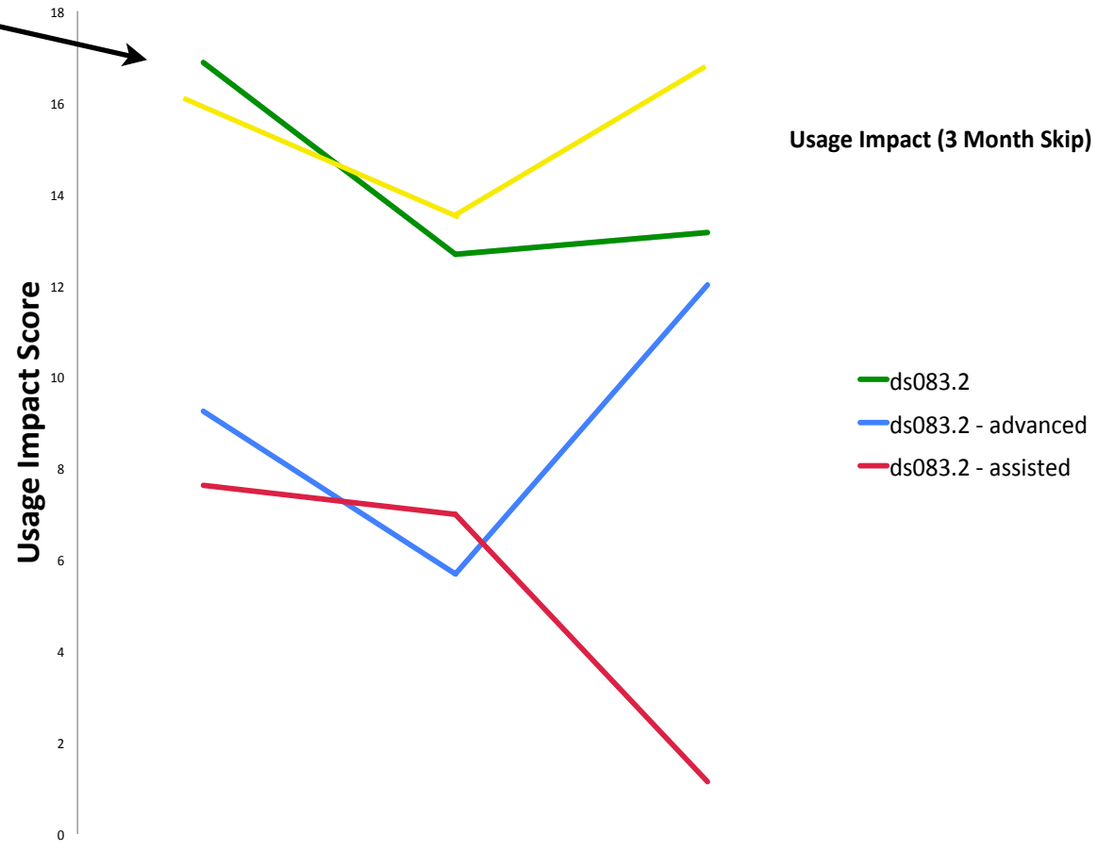
Indicator	Explanation	
Unique Users (UU)	Unique users that downloaded data during a time window	Coincident
Unique Users - Advanced	UUs that accessed data programmatically	
Unique Users - Assisted	UUs that accessed data via GUI or Service	
Number of Datasets	Number of Datasets assigned DS number	
Files DS	Number of files in Dataset per time window	Leading
Download Frequency	Total number of files downloaded per time window	
Download Frequency - Advanced	Files downloaded by Advanced users	
Download Frequency - Assisted	Files downloaded by Assisted users	
Homepage Hits	Dataset Homepage Hits per time window	
Homepage Hits - Direct Access	Dataset Homepage Hits per time window by users with direct access (link not indexed or retrieved by search)	
Homepage Hits - With Link	Dataset Homepage Hits per time window by users with link (from indexed list or retrieved by search)	
Subset Requests	Subsets Requests per time window	
Download Density	Average number of files downloaded per UU	Lagging
Usage Impact	Total number of downloaded files over total files in dataset	
Usage Impact - Advanced	“	
Usage Impact - Assisted	“	
Interest Impact	Total homepage hits per number of files in dataset	
Usage Balance	Files downloaded by number of homepage hits per time window	
Subset Ratio	Number of subset requests over total number files downloaded per time window	
Secondary Interest Impact	Homepage over UU	

Impact indicators

Tell a more complex story using combination of metrics

Allows us to convey to funding agencies long-term influence of introducing new services

And most important, can give long-term view of value.



Lessons learned

Metrics are a (painful!) craft process:

Start with a baseline (Data Usage Index)

Adapt for the specificities of your domain and your archive.

Find weird patterns, and explore (Science!)

Thank you.

nmweber@illinois.edu

@nniiicc

Papers where this work appears

Qualitative

Palmer, C. , Weber, N., and Cragin, M.. (2011). The analytic potential of scientific data: Understanding re-use value. Proceedings of the American Society of Information Science and Technology annual meeting. New Orleans, LA.

Weber, N., Baker, K. , Thomer, A. , Chao, T. , and Palmer, C. (2012) Context, Value and Data Use: Domain Analysis Revisited. Proceedings of the American Society of Information Science and Technology annual meeting. Baltimore, MD

Quantitative

Weber, N., Thomer, A., Mayernik, M., Worley, S., Dattore, R. and Hua, Z. (2013) The product and system specificities of measuring impact: Indicators of data use in research data archives. Proceedings of the 8th International Digital Curation Conference. Amsterdam, NL.

Weber, N. (2013) Lead, Lag or get out of the Index: Macroeconomic indicators of data use. Proceedings of the 2013 iConference. Ft. Worth, Texas.