METADATA CAPITAL: CONCEPTUAL UNDERSTANDING, PREDICTIVE VALUE

MESA - METADATA MADNESS
MARCH 31, 2015

Jane Greenberg, Professor
Director, Metadata Research Center
Interim Department Head, Information Science
College of Computing & Informatics

What is capital to you?

Write something down

How do you describe value?

Write something down


Capital?


Value?

YOUR DATA IS ONLY AS GOOD AS YOUR METADATA

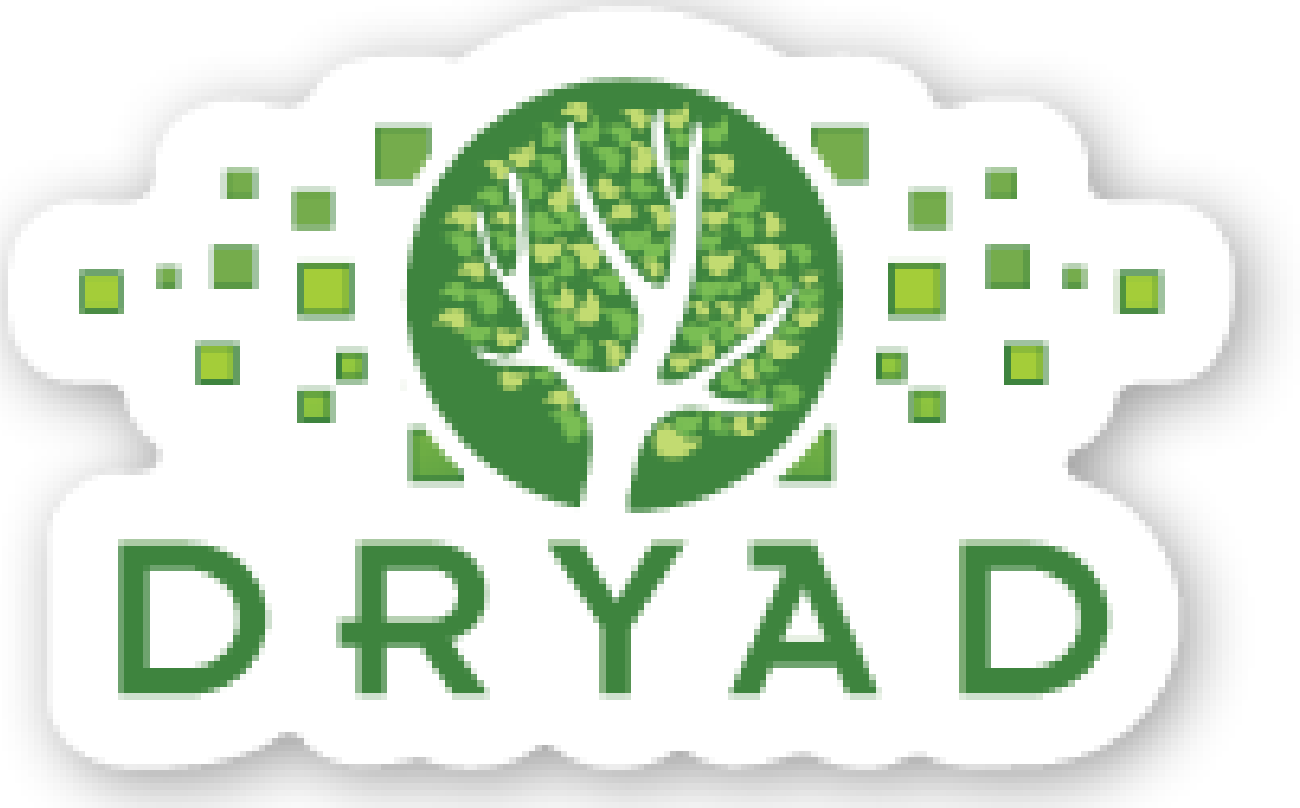
***Metadata** is a first class object*

DREXEL UNIVERSITY
Metadata Research Center
College of Computing & Informatics

THE TOPIC…

Motivation – Making metadata work harder (DRYAD)

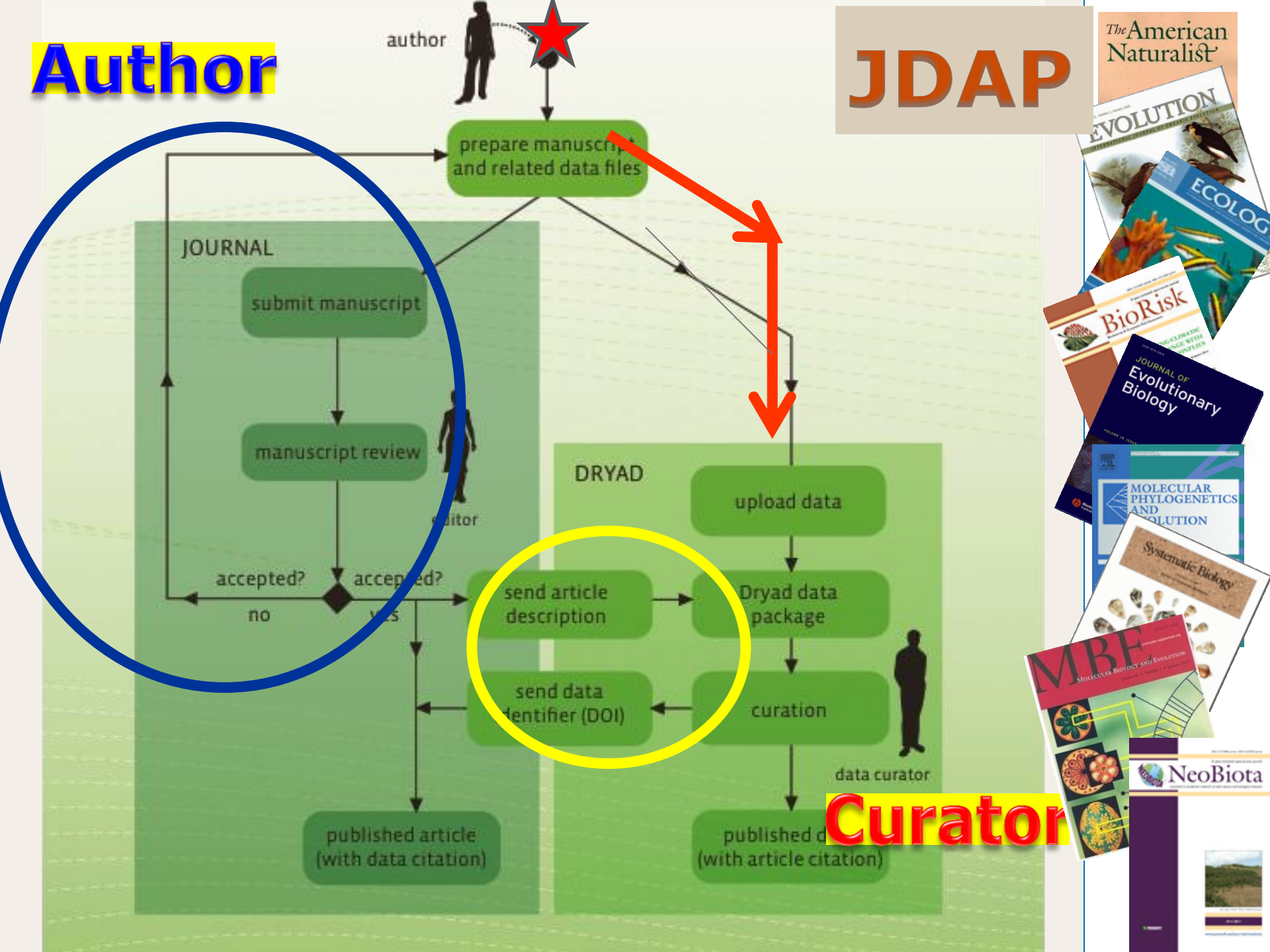
ROI – return on investment (CAPITAL)


DRYAD

Dryad...a curated general-purpose repository...makes data discoverable, freely reusable, and citable.

"...enables scientists to validate published findings, explore new analysis methodologies, repurpose data for research questions unanticipated by the original authors, and perform synthetic studies." (http://datadryad.org/)

Not this →


Author
author
prepare manuscript and related data files
JOURNAL
submit manuscript
manuscript review
editor
accepted?
no
accepted?
yes
published article (with data citation)
DRYAD
upload data
send article description
Dryad data package
send data identifier (DOI)
curation
data curator
published d
(with article citation)
Curator
JDAP
The American Naturalist
EVOLUTION
ECOLOG
BioRisk
JOURNAL OF Evolutionary Biology
MOLECULAR PHYLOGENETICS AND
Systematic Biology
MBE
NeoBiota

Describe publication

Submitting data to Dryad consists of three simple steps:

1. **Describe your publication**
2. Upload and describe your data files
3. Approve data for publication

Please describe your publication in as much detail as possible. Providing a detailed description will make it easier for othe
data in Dryad. Please describe the **publication only**. Do not enter information specific to your data files on this page.

Fields marked with an asterisk (*) are required. For more information on expected contents for a field, hold your mouse o
question.

Publication metadata

Title*: Adaptive responses and disruptive effects: how major wildfire

Authors*: Add

Last name, *e.g. Smith* First name + initial, *e.g. Donald F.*

- Banks, Sam
- Blyton, Michaela
- Blair, David
- McBurney, Lachlan
- Lindenmayer, David

Remove selected

Journal name*: Molecular Ecology

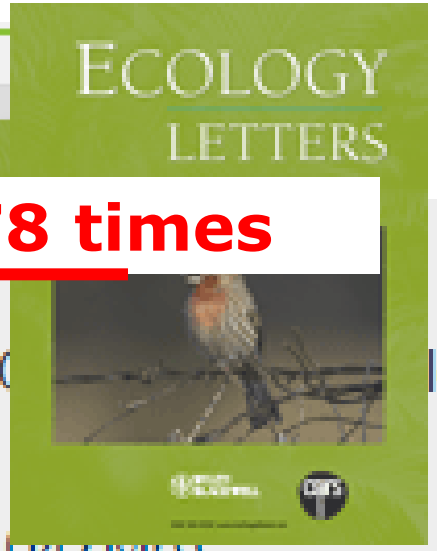
Abstract: Environmental disturbance is predicted to play a key role in the evolution of animal social behaviour. This is because disturbance affects key factors underlying

Pre-populated metadata field

Data from: Towards a worldwide wood economics

Downloaded	12378 times
Description	Please direct all correspondence to C Gonzalez@leeds.ac.uk>
Download	GlobalWoodDensityDatabase.xls (2.047Mb)
Details	View File Details

Download 12378 times



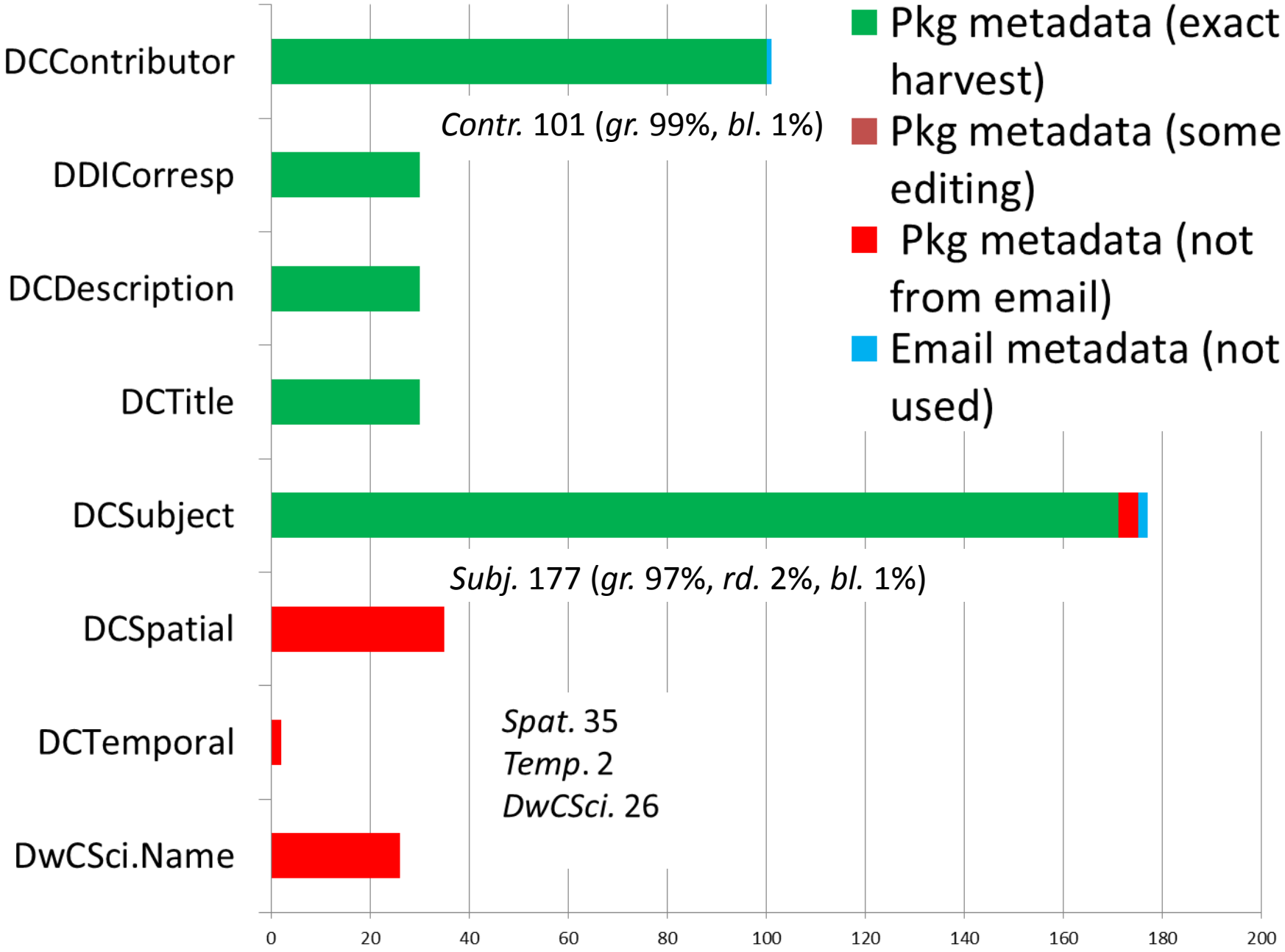
Observations, motivating study of metadata capital

1. Metadata generation costs money
2. Metadata reuse is **a BIG part** of Dryad's workflow
3. Metadata reuse via OAI
4. Metadata reuse via data sharing, reuse, and repurposing

Statistics

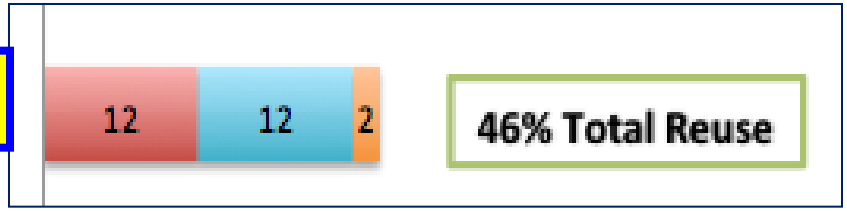
Type	Total	30 days
Data packages	7978	326
Data files	24903	1195
Journals	390	105
Authors	28714	4279
Downloads	772229	18043

PACKAGE METADATA HARVESTED FROM EMAIL

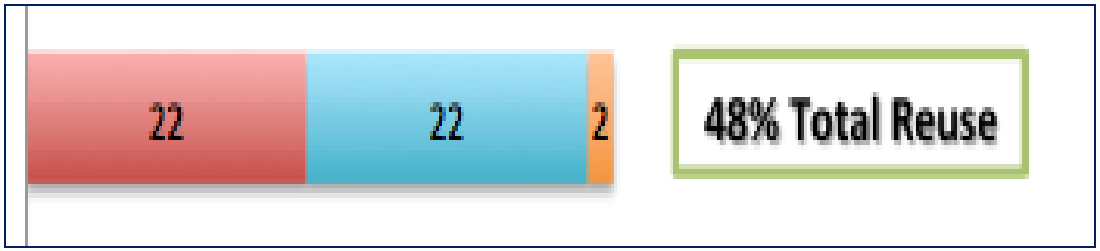


Author

Dcterms.spatial→



Subject



← DwC.ScientificName

MetaDataCAPT'L

ncds THE NATIONAL CONSORTIUM for DATA SCIENCE

$$R + \sum_{i=1}^{n} a_i = R + a_1 + a_2 + a_3 + \ldots a_n$$

R = value of the metadata record

i= number of usages

a = incremental increase in value

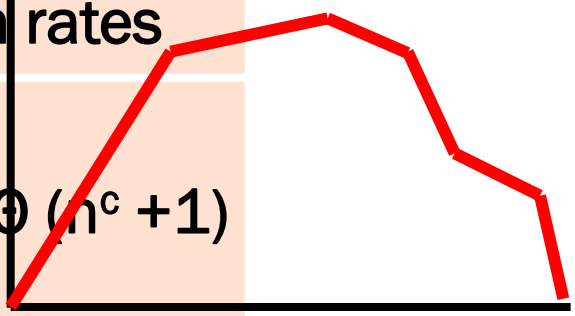
n = maximum number of reuse

Metadata duplication is inefficient, tedious

An economic concept (Weber, 1905; Smith, 1776)

Successive growth rates

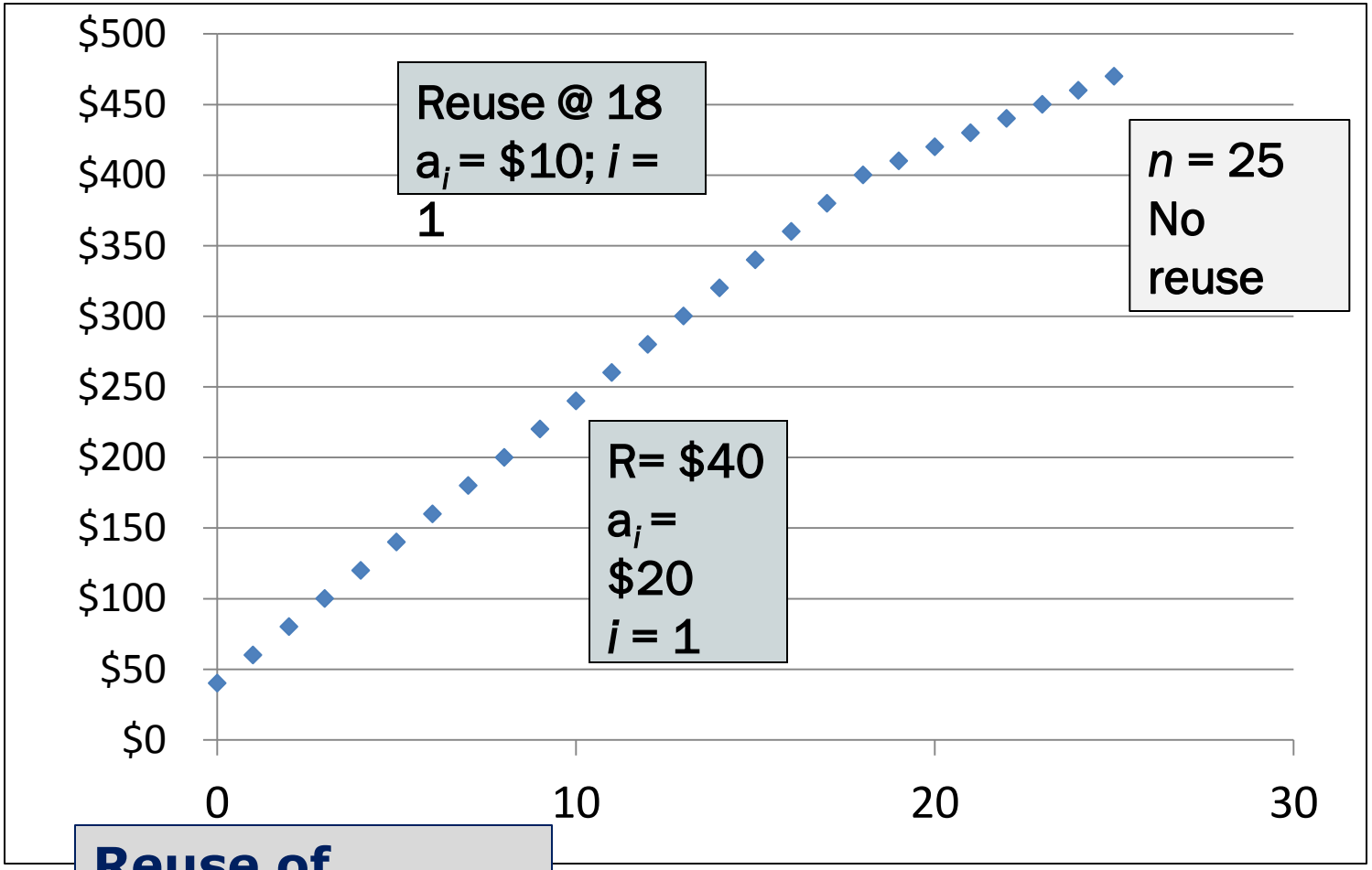
$$\sum_{i=1}^{N} i^c = \Theta\,(n^c + 1)$$



Cycles…

THE NATIONAL CONSORTIUM for DATA SCIENCE

METADATACAPT'L @ $440 FOR A DATA OBJECT

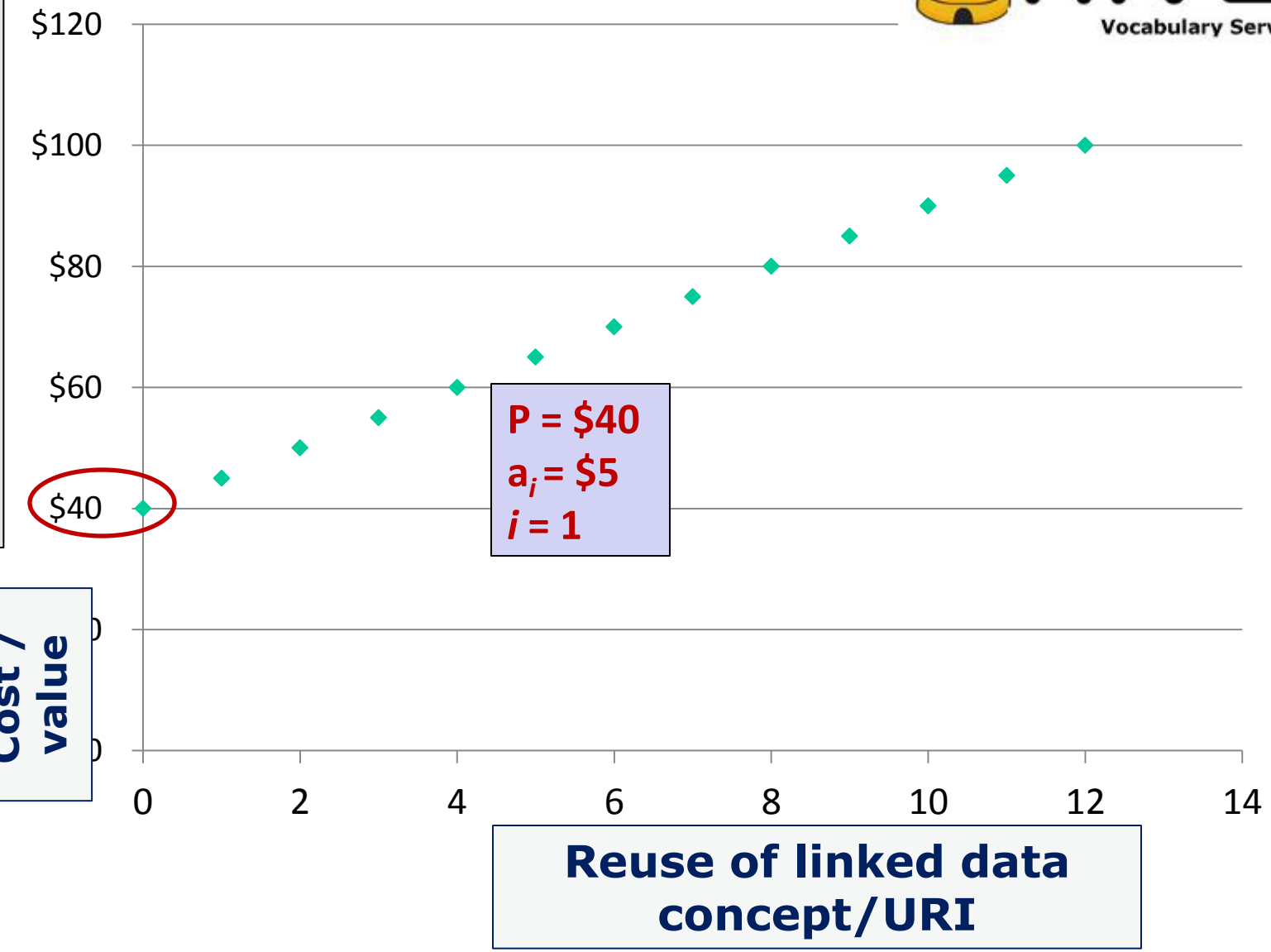


DRYAD

MODIFIED CAPITAL-SIGMA NOTATION FOR LINKED DATA

HIVE Vocabulary Server

P = Determined by the number of terms in an ontology, labor hours to generate, integrate, etc,



RTI INTERNATIONAL

SGHIx CONSORTIUM

	A	B	C	D	E	F	G	H
	Validic API Cate	Parameter in Us	BodyMedia	FatSecret	DailyMile	Fitbit	Fitbug	Fleetly
	Fitness	utc_offset	P	X	X	X	X	P
	Routine	timestamp	X	NA	NA	X	X	NA
	Routine	steps	X	NA	NA	X	X	NA
	Routine	distance	X	NA	NA	X	X	NA
	Routine	floors	X	NA	NA	NA	NA	NA
	Routine	elevation	NA	NA	NA	NA	NA	NA
	Routine	calories_burned	NA	NA	NA	X	X	NA
	Routine	utc_offset	X	NA	NA	X	X	NA
	Nutrition	timestamp	X	X	NA	X	X	NA
	Nutrition	calories	X	X	NA	X	X	NA
	Nutrition	carbohydrates	NA	X	NA	X	P	NA
	Nutrition	fat	NA	X	NA	X	P	NA
	Nutrition	fiber	NA	NA	NA	X	P	NA
	Nutrition	protein	NA	X	NA	X		NA
	Nutrition	sodium	NA	X	NA	X		NA
	Nutrition	water	NA	NA	NA	NA		NA
	Nutrition	meal	X	X	NA	X		NA
	Nutrition	utc_offset	X	X	NA	X		NA

Total Fields Referenced (FitBit), toward SGHIx

X Availale: 39

P (Pending): 3

NA (not available): 42

(Caruso & Ogletree)

metadata

Most popular
timestamp
type
start_time
distance
duration
calories
utc_offset

College of Computing & Informatics

CONCLUSION…OTHER VALUATION APPROACHES

Market cap of Facebook per user: $40 – $300

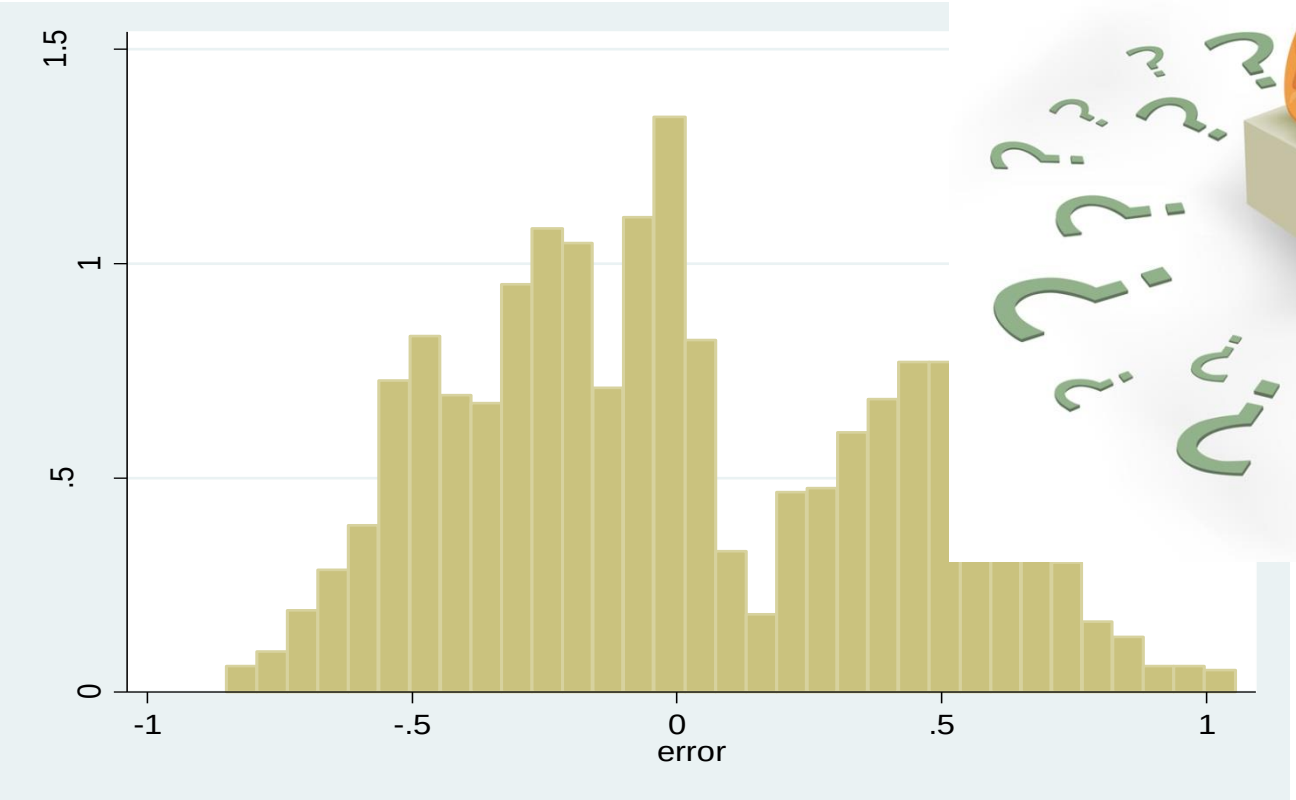
Revenues per record per user: $4 – $7 per year

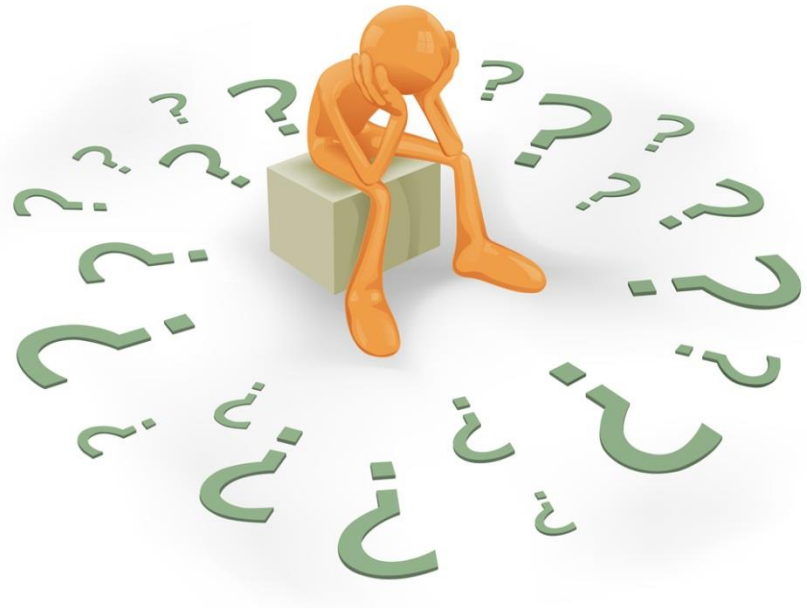
- Facebook
- Experian

Market prices of personal data:

- $0.50 for street address
- $2.00 for date of birth
- $8 for social security number
- $3 for driver's license number
- $35 for military record

SOURCE: OECD. *Exploring the Economics of Personal Data: A Survey of Methodologies for Measuring Monetary Value*. OECD Digital Economy Papers. Office for Economic Cooperation and Development Publishing, 2013.





IN THE FITBIT DATA SCENARIO, IF A PATIENT'S EXERCISE DATA AND ENVIRONMENTAL QUALITY DATA CAN BE COMBINED WITH ASTHMA CONDITION DATA, WE WILL GET A BETTER PREDICTION OF THE WAY IN WHICH ASTHMA EVOLVES.

WORKING ON METADATA CONNECTION…


NIEHS

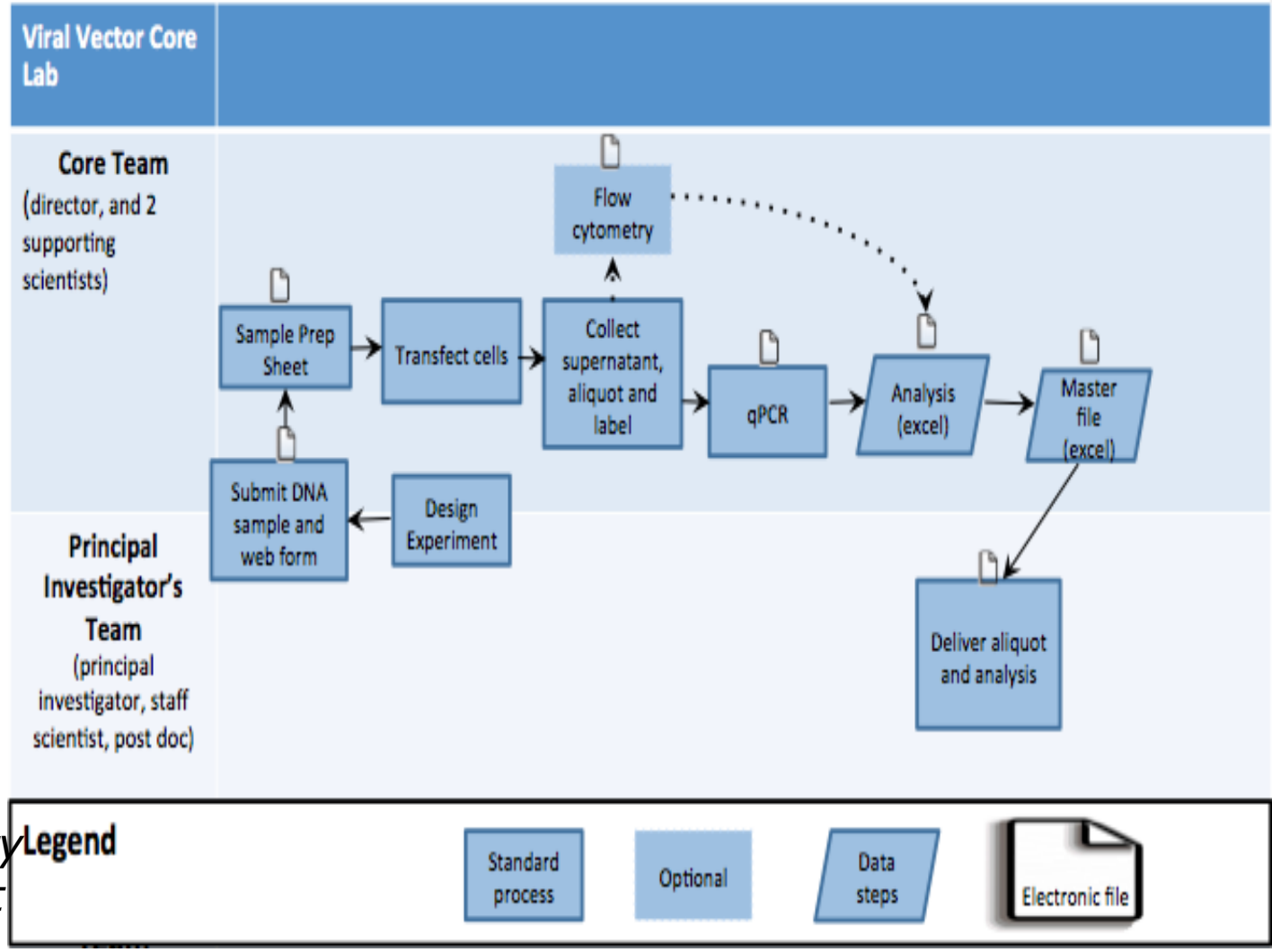
GOALS OF EXPLORATORY WORK

- Understand the Viral Vector Core Laboratory (VVCL) workflow.
- Map the VVCL metadata lifecycle.
- Explore machine-actionable rules that can support the VVCL metadata lifecycle.
- Create an iRODS prototype for the VVCL workflow, and explore the application of machine-actionable rules.

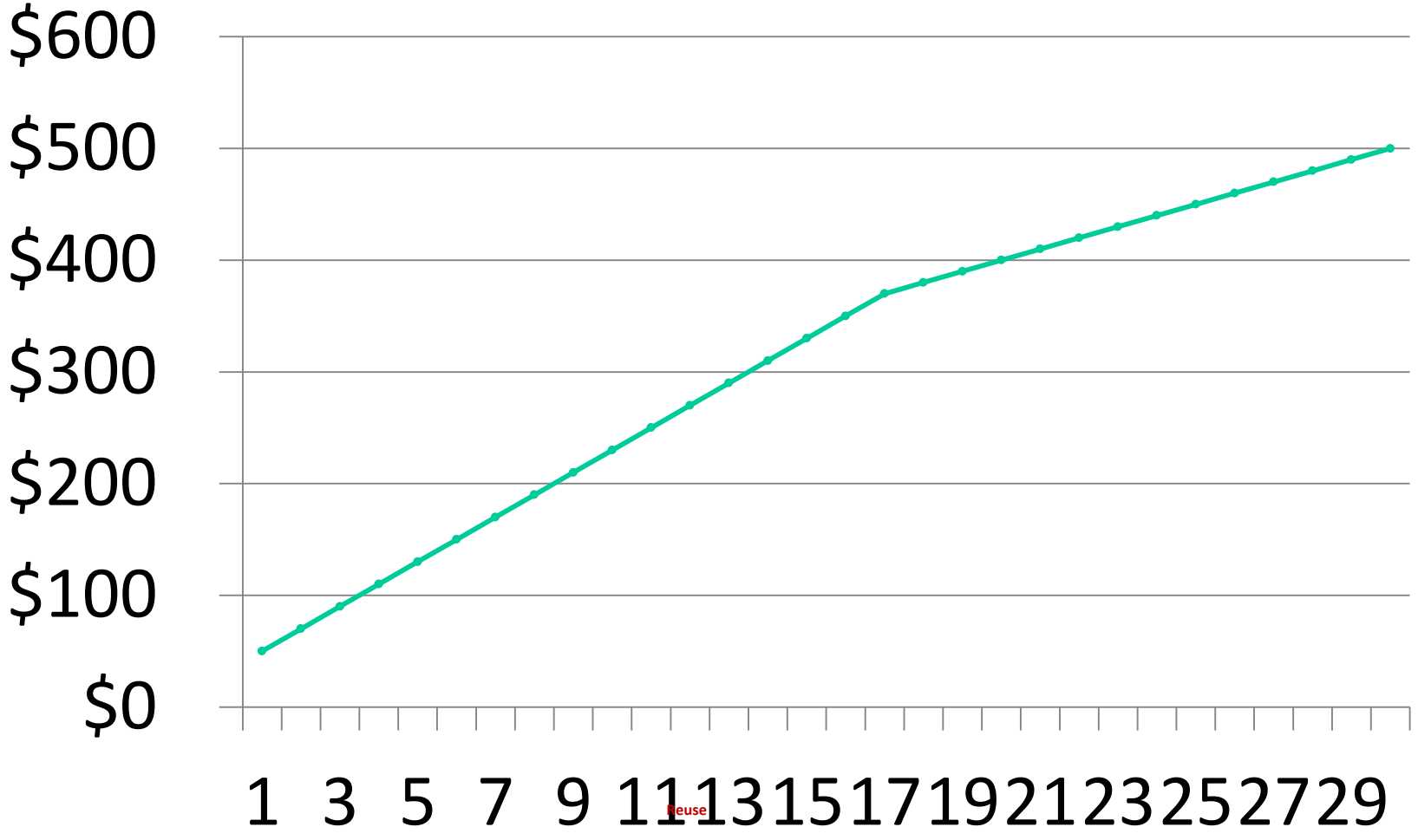
METHODS

- Collaborative workflow modeling was used to capture the day-to-day workflow.
- A metadata analysis was conducted to identify basic metadata generated and automatically propagated during each workflow stage in the VVCL process.
- Microservices articulation

- *Design experiment:*
- *Submit DNA sample and web form*
- *Transfect cells*
- *Collect the supernatant, aliquot, and label*
- *Present the qPCR*
- *Flow cytometry*
- *Deliver aliquot and analysis*



50 USD PER HOUR FOR AN EXPERIMENT



SOME CONCLUDING OBSERVATIONS

- Discover and advance the application of methods for quantifying the cost and value of metadata over time
- *Raise dialog*
- Advance nascent work on "metadata capital"

Information as an economic asset - Machlup's *The Production and Distribution of Knowledge in the United States*

- *Metadata experts emphasize the value of metadata for data lifecycle management (e.g., data capture, use/reuse, provenance tracking, etc.) (Lytras and Sicilia, 2007; Garoufallou, E., Papatheodorou, IJMSO, 2014)*

LIMITATIONS

- Modified capital-sigma is only one dimensional; all metadata properties/concept are not equal
- Also, we know cost/value relationship is not 1:1.
- Metadata is only as good as your data

  not always true

- What about successive growth rate may be the way to go


Capital?


ANTIQUES
Value?
& OLD STUFF
www.bigstock.com · 164410

Friendship, social network

Capital,

value?

Discussion…

Can we study cost?

How do we convey value?

Is there a connection between cost/value/quality?

How does this all fit with media and enterainment

YOUR DATA IS ONLY AS GOOD AS YOUR METADATA

***Metadata** is a first class object*

DREXEL UNIVERSITY
Metadata Research Center
College of Computing & Informatics

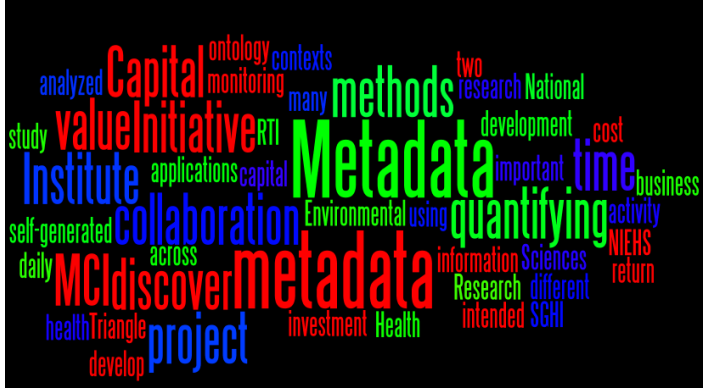
PETER FOX, Tetherless World Constellation Chair And Professor Of Earth And Environmental Science And Computer Science At Rensselaer Polytechnic

Get rid of the word 'metadata' (RDA Conference, Sweden, March 2013, keynote)

- Provenance data
- Descriptive data
- Authenticity data

DREXEL UNIVERSITY
Metadata Research Center
College of Computing & Informatics

THE TEAM / ACKNOWLEDGMENTS



Tom Caruso, Health Information

Liaison Research Associate, UNC-SILS/RTI

- Self-generated Health Information (SGHI)

Rebecca Boyles, Data Scientist, NIEHS

- Common Core Vocabulary

Jane Greenberg, SILS/UNC, MRC

Herbie Huang, Ph.D. student, Economics Dep. UNC

Austin Mathews, BSIS student, SILS/UNC, MRC

Angela Murillo, Ph.D. student, ,SILS/UNC, MRC

Adrian Ogletree, MSIS student, SILS/UNC, MRC

Erik Scott, Senior Software Dev./RENCI (Renaissance Computing Institute)

ncds THE NATIONAL CONSORTIUM for DATA SCIENCE

ACKNOWLEDGMENTS

Dryad Consortium Board, journal partners, and data authors

NESCent: Laura Wendell (Executive Director), Hilmar Lapp, Heather Piwowar, Peggy Schaeffer, Ryan Scherle, Todd Vision (PI)

**Drexel/UNC <Metadata Research Center>: Jose R. Pérez-Agüera, Sarah Carrier, Elena Feinstein, Lina Huang, Robert Losee, Hollie White, Craig Willis, Jane Smith, Shea Swuager, Liz Turner, Christine Mayo, Adrian Ogletree, Erin Clary

U British Columbia: Michael Whitlock

NCSU Digital Libraries: Kristin Antelman

HIVE: Library of Congress, USGS, and The Getty Research Institute; and workshop hosts

Yale/TreeBASE: Youjun Guo, Bill Piel

DataONE: Rebecca Koskela, Bill Michener, Dave Veiglais, and many others

British Library: Lee-Ann Coleman, Adam Farquhar, Brian Hole

Oxford University: David Shotton