MetaDataCAPT'L

# Metadata Quality and Capital

## EUDAT/B2FIND & Quality of metadata

25, Sept. 2014, Amsterdam

ncds | THE NATIONAL CONSORTIUM for DATA SCIENCE

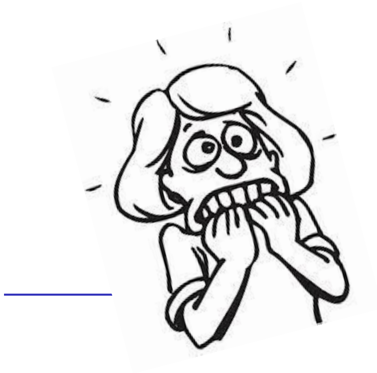
Jane Greenberg, CCI/Drexel University
Director, SILS Metadata Research Center

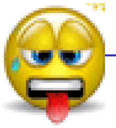
DREXEL UNIVERSITY
Metadata Research Center
College of Computing & Informatics

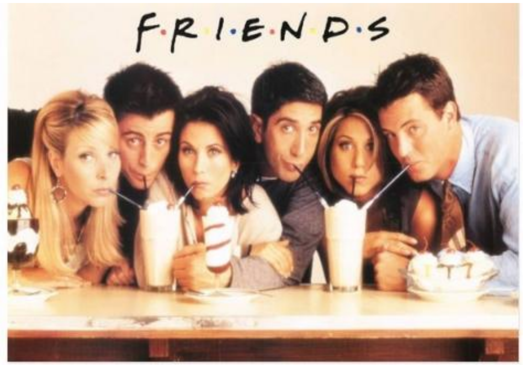
# Overview

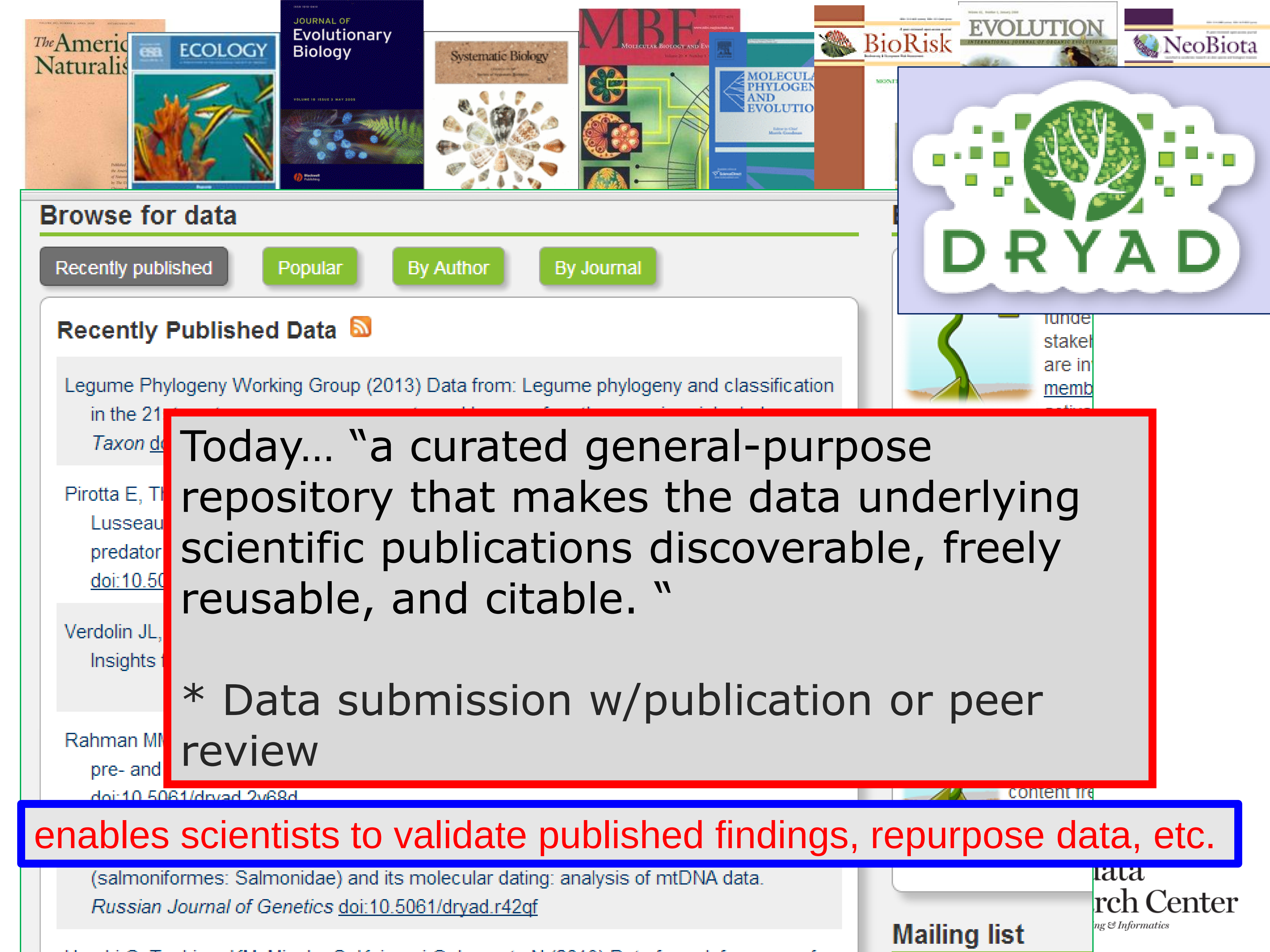
1. Motivation
   - Dryad / Metadata capital, some early data
2. HIVE – Helping Interdisciplinary Vocabulary Engineering (linked data)
3. MetaDataCAPT’L
4. Quality and capital (observations)
5. Conclusions, discussion, criticism…

conceptual








The American Naturalist
ECOLOGY
Journal of Evolutionary Biology
Systematic Biology
MBE
Molecular Biology and Evolution
Molecular Phylogenetics and Evolution
BioRisk
EVOLUTION
NeoBiota
DRYAD

Browse for data

Recently published | Popular | By Author | By Journal

Recently Published Data

Legume Phylogeny Working Group (2013) Data from: Legume phylogeny and classification in the 21

*Taxon* d

Pirotta E, Th

Lusseau

predator

doi:10.50

Verdolin JL,

Insights f

Rahman MM

pre- and

doi:10.5061/dryad.2v68d

(salmoniformes: Salmonidae) and its molecular dating: analysis of mtDNA data. *Russian Journal of Genetics* doi:10.5061/dryad.r42qf

funde
stake
are in
memb
content fre

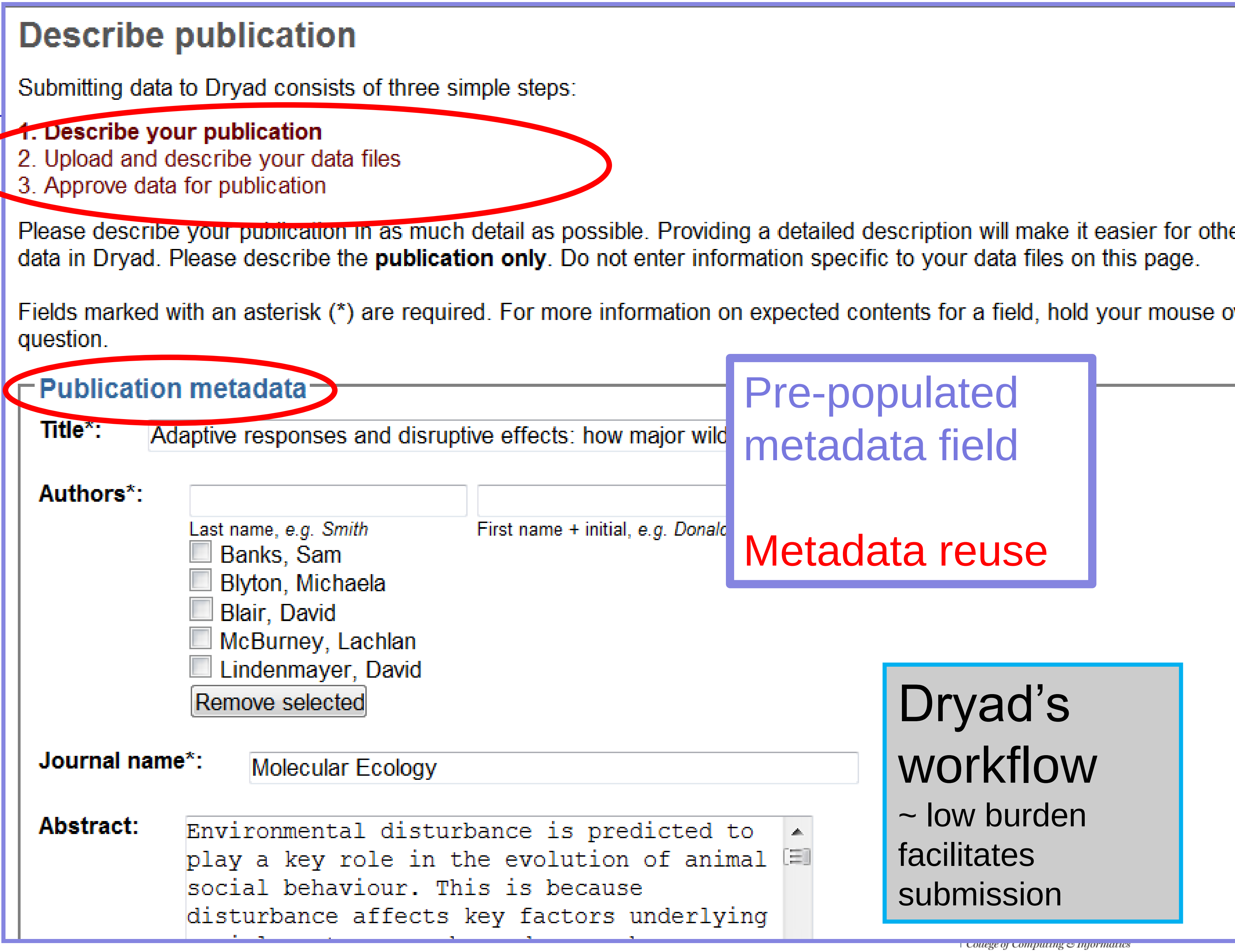
data
rch Center
ng & Informatics

Mailing list

Today… "a curated general-purpose repository that makes the data underlying scientific publications discoverable, freely reusable, and citable. "

* Data submission w/publication or peer review

enables scientists to validate published findings, repurpose data, etc.



## 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 wild

Authors*:

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

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

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

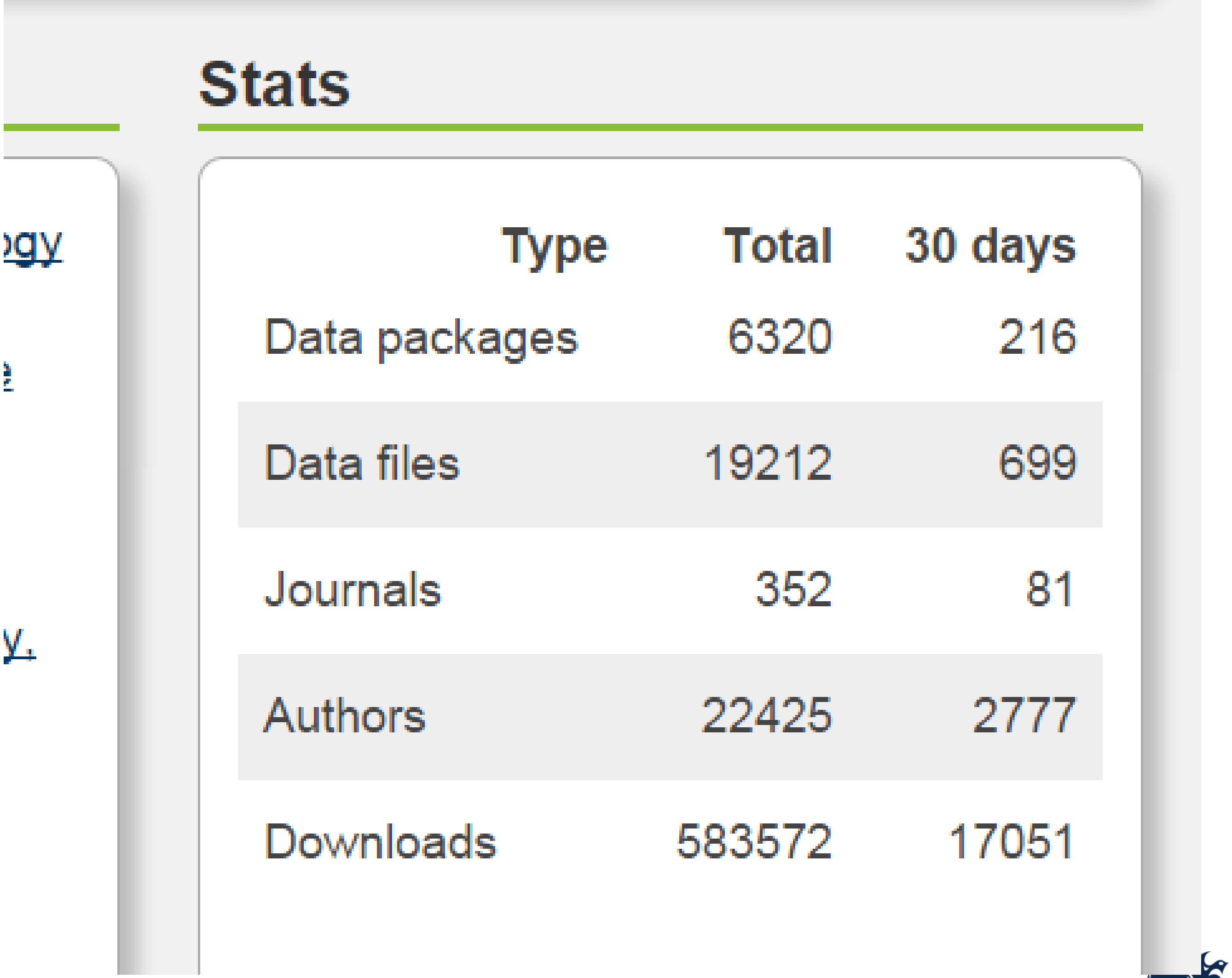
Pre-populated metadata field

Metadata reuse

Dryad’s workflow

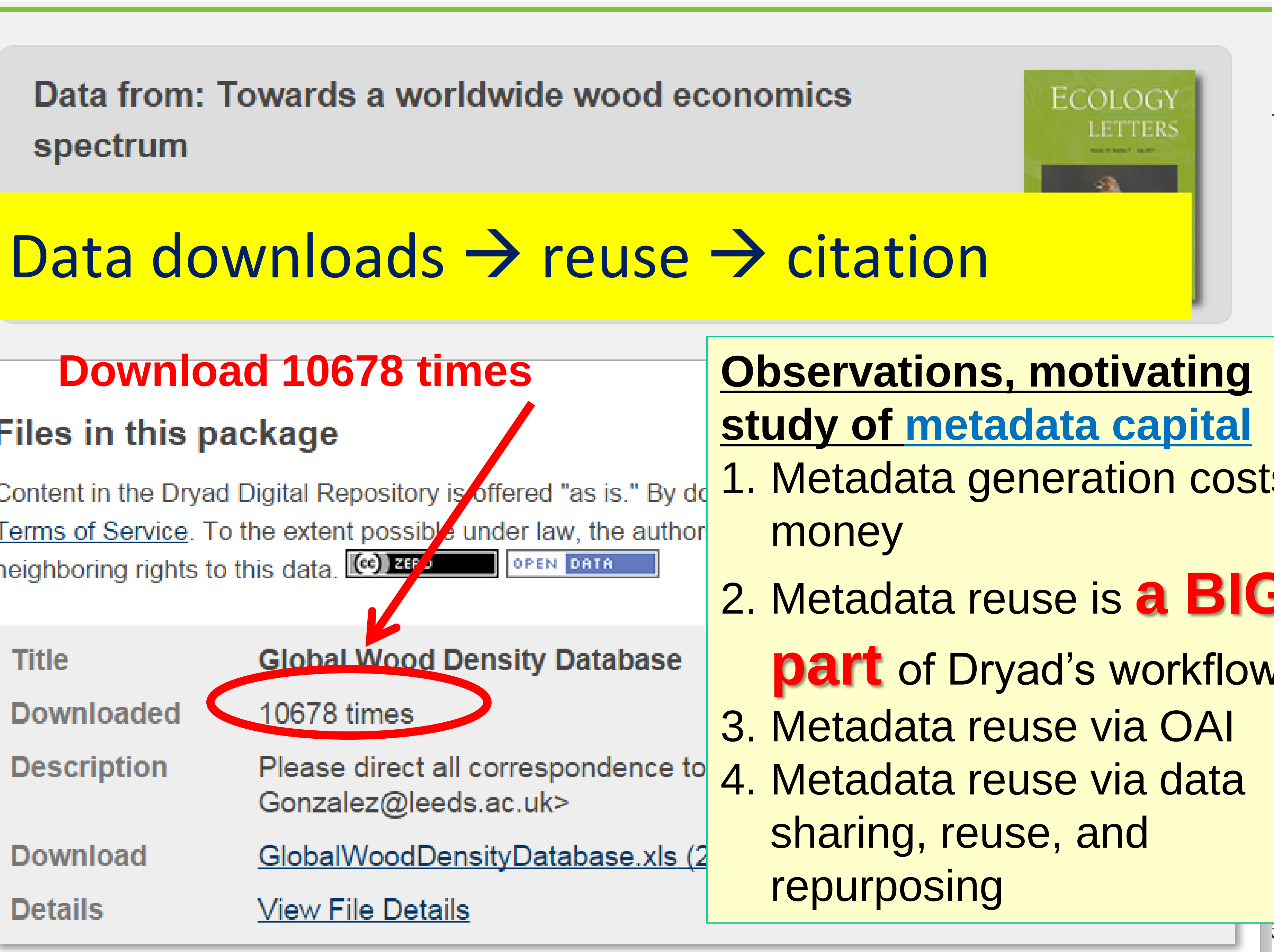
~ low burden facilitates submission

# Dryad statistics from Monday AM this week



## Stats

| Type | Total | 30 days |
|---|---|---|
| Data packages | 6320 | 216 |
| Data files | 19212 | 699 |
| Journals | 352 | 81 |
| Authors | 22425 | 2777 |
| Downloads | 583572 | 17051 |



Data from: Towards a worldwide wood economics spectrum

Files in this package

Content in the Dryad Digital Repository is offered "as is." By do
Terms of Service. To the extent possible under law, the author
neighboring rights to this data. (cc) ZERO OPEN DATA

| | |
|---|---|
| Title | Global Wood Density Database |
| Downloaded | 10678 times |
| Description | Please direct all correspondence to<br>Gonzalez@leeds.ac.uk> |
| Download | GlobalWoodDensityDatabase.xls (2 |
| Details | View File Details |

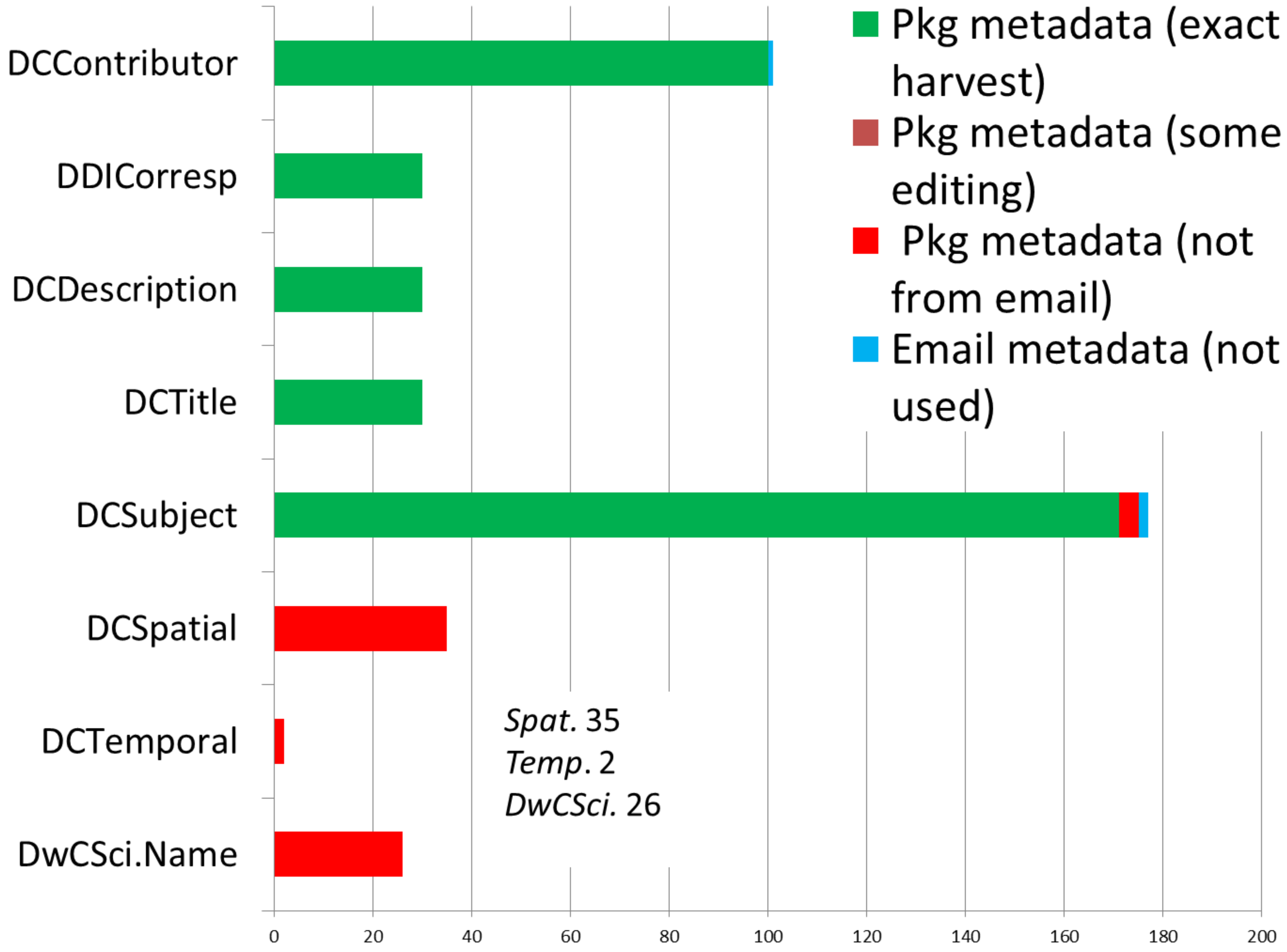
# Data downloads → reuse → citation

**Download 10678 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

# Package metadata harvested from email



# The leap - capital to metadata capital

- An economic concept (Weber, 1905; Smith's, 1776)
  - Business and operations (net gains or losses)
  - Finances, goods and services, and public needs
  - Intellectual capital, social capital
  - ***a tangible result, value increase***
- **Metadata as an asset, a product**
  - Reuse of good *quality metadata* increase value of initial investment
    - Poor quality may reduce metadata capital ?
  - Metadata reuse prevalence
    - Cooperative cataloging , CIP, ISBD, MARC, FRBR, LCC, VIAF, OAI-PMH, CrossRef, PubMed, Zotero, BibTex, DataCite. Linked data/Semantic Web, PIDs, etc.

# Modified Capital-sigma notation

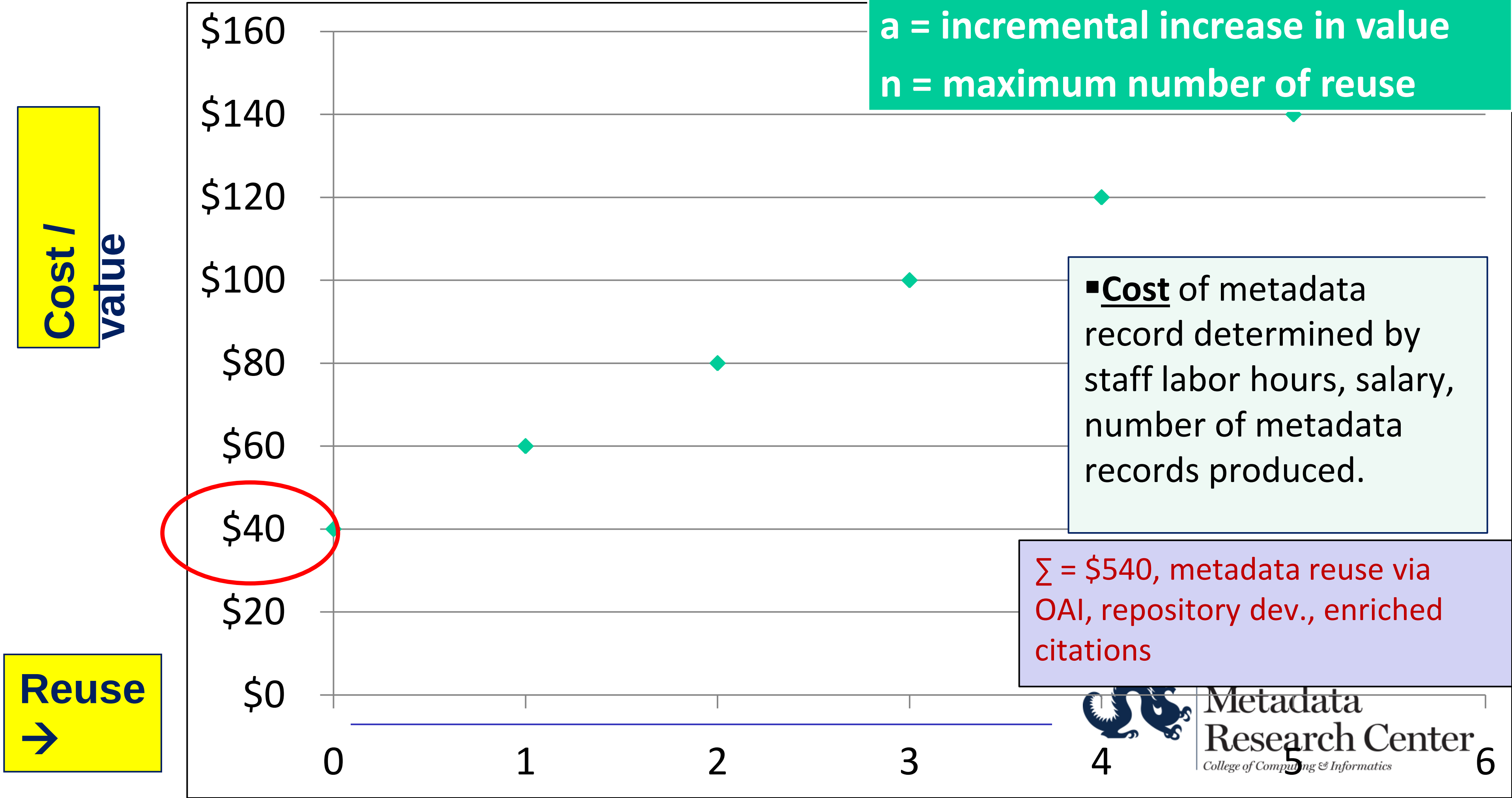
$$R + \sum_{i=1}^{n} a_i = R + a_1 + a_2 + a_3 + \ldots a_n$$

R = value of the metadata record

i= number of usages

a = incremental increase in value

n = maximum number of reuse

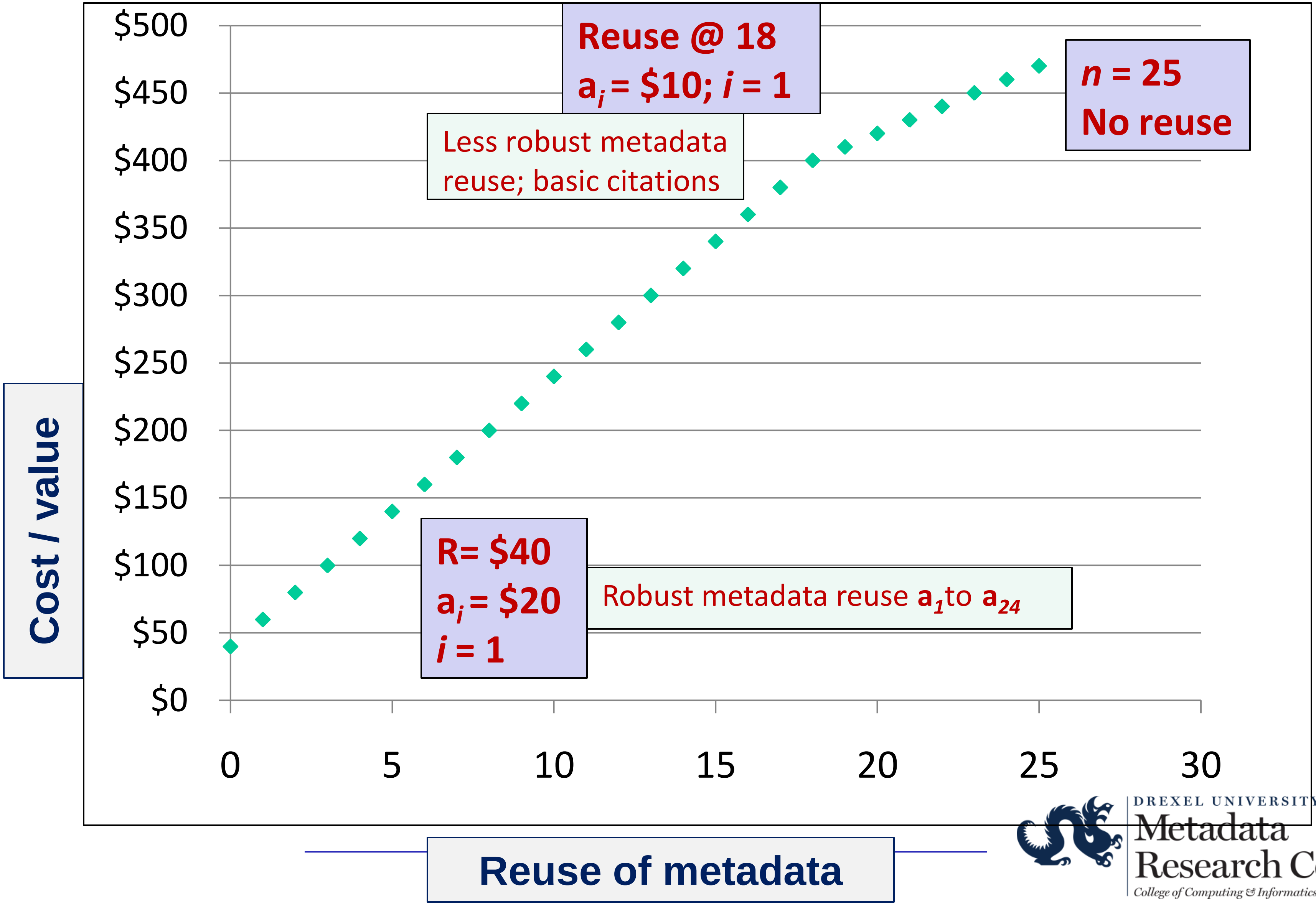


- **Cost** of metadata record determined by staff labor hours, salary, number of metadata records produced.

∑ = $540, metadata reuse via OAI, repository dev., enriched citations

Metadata Research Center
College of Computing & Informatics

# Modified Capital-sigma notation



**Reuse @ 18**
**$a_i$ = $10; *i* = 1**

Less robust metadata reuse; basic citations

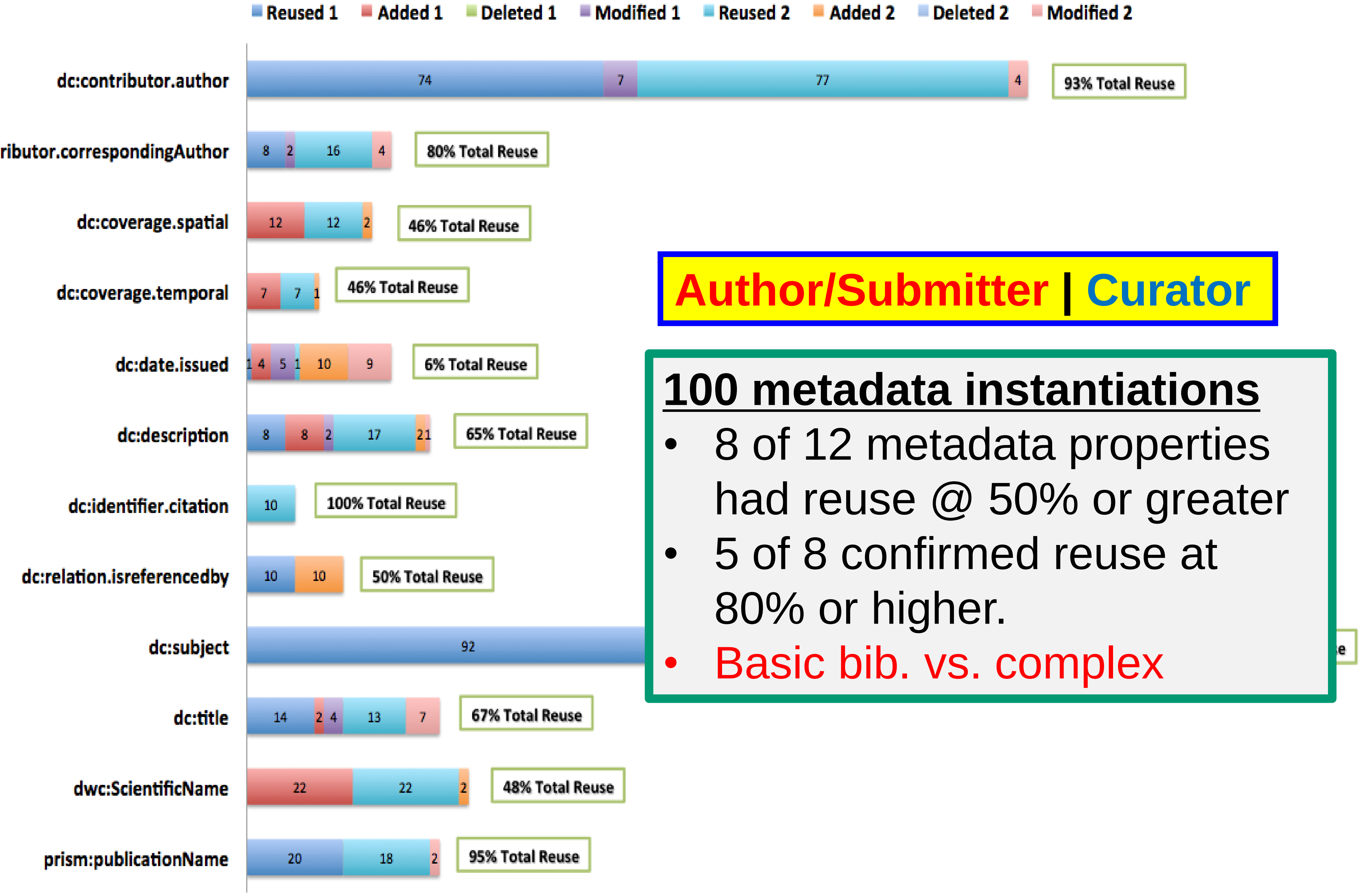
***n* = 25**
**No reuse**

**R= $40**
**$a_i$ = $20**
***i* = 1**

Robust metadata reuse $\mathbf{a_1}$ to $\mathbf{a_{24}}$

DREXEL UNIVERSITY
Metadata Research Center
College of Computing & Informatics

# FIGURE 4: TOTAL METADATA WORKFLOW PHASES 1 & 2 - CASES A & B

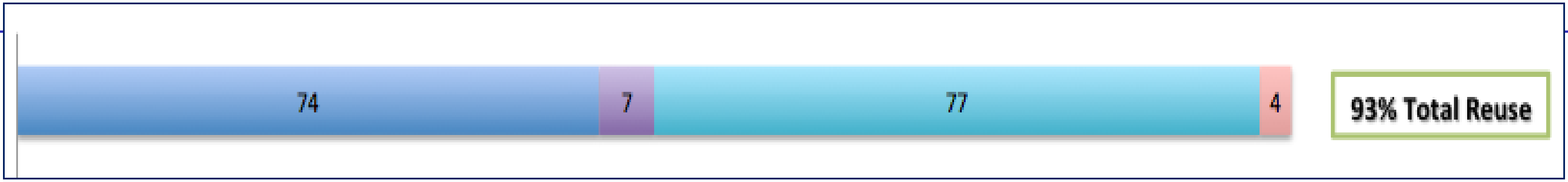


**Author/Submitter | Curator**

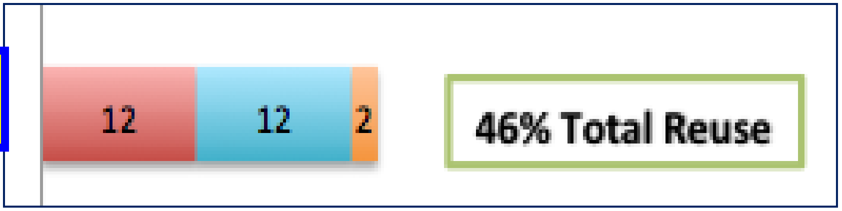
**100 metadata instantiations**

- 8 of 12 metadata properties had reuse @ 50% or greater
- 5 of 8 confirmed reuse at 80% or higher.
- Basic bib. vs. complex

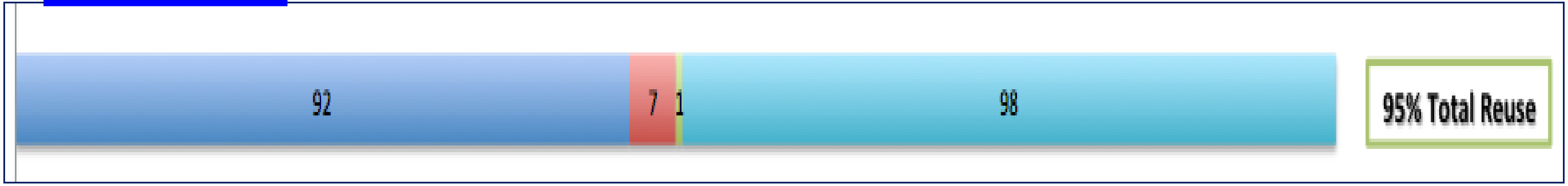
Author

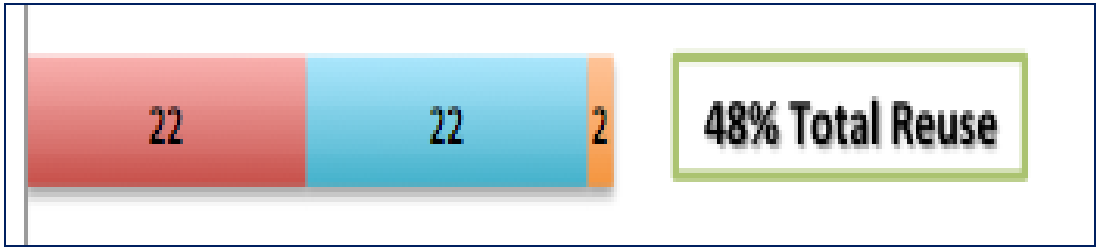


Dcterms.spatial→



Subject





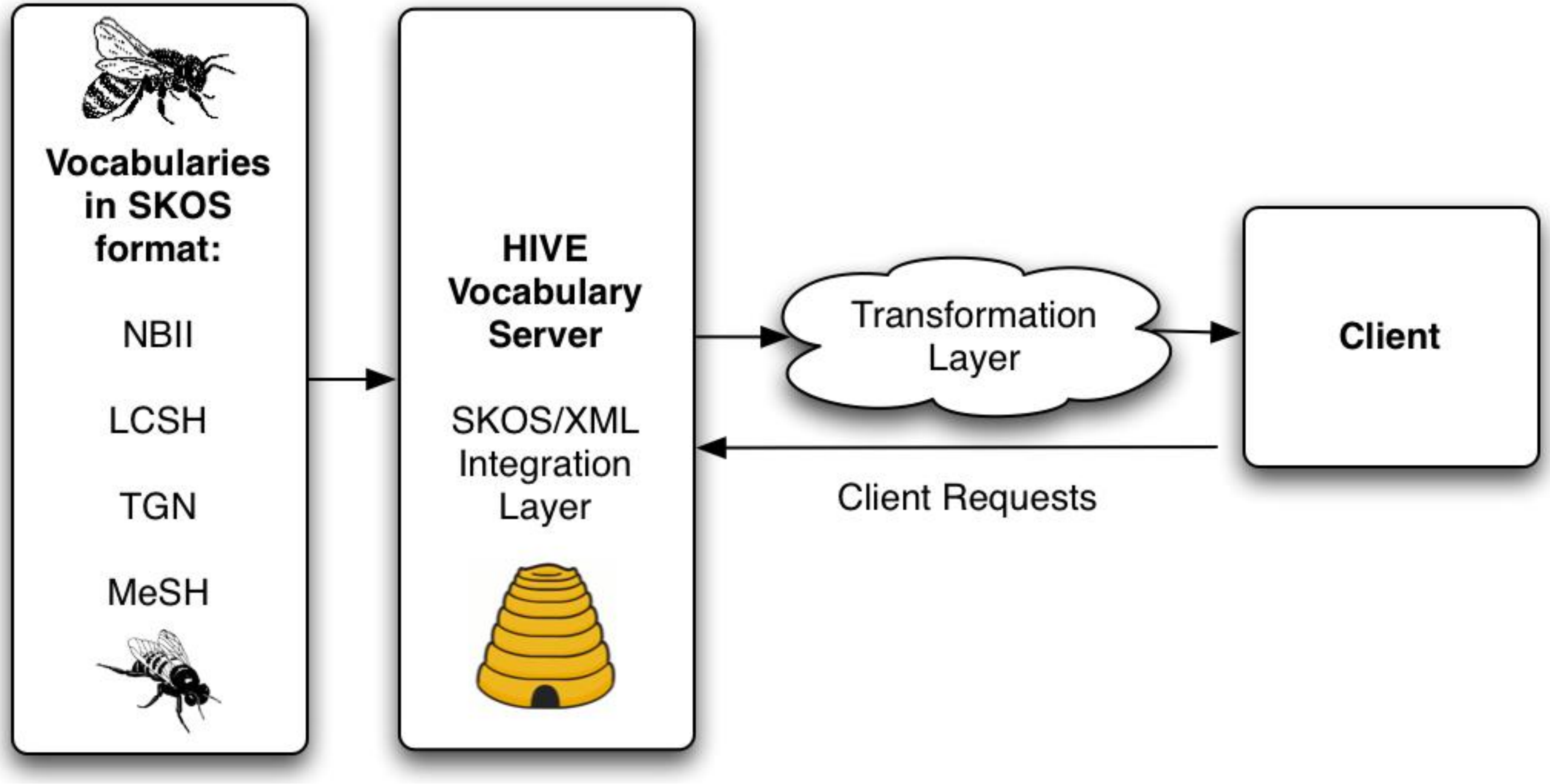
← DwC.ScientificName

# KO – knowledge organization

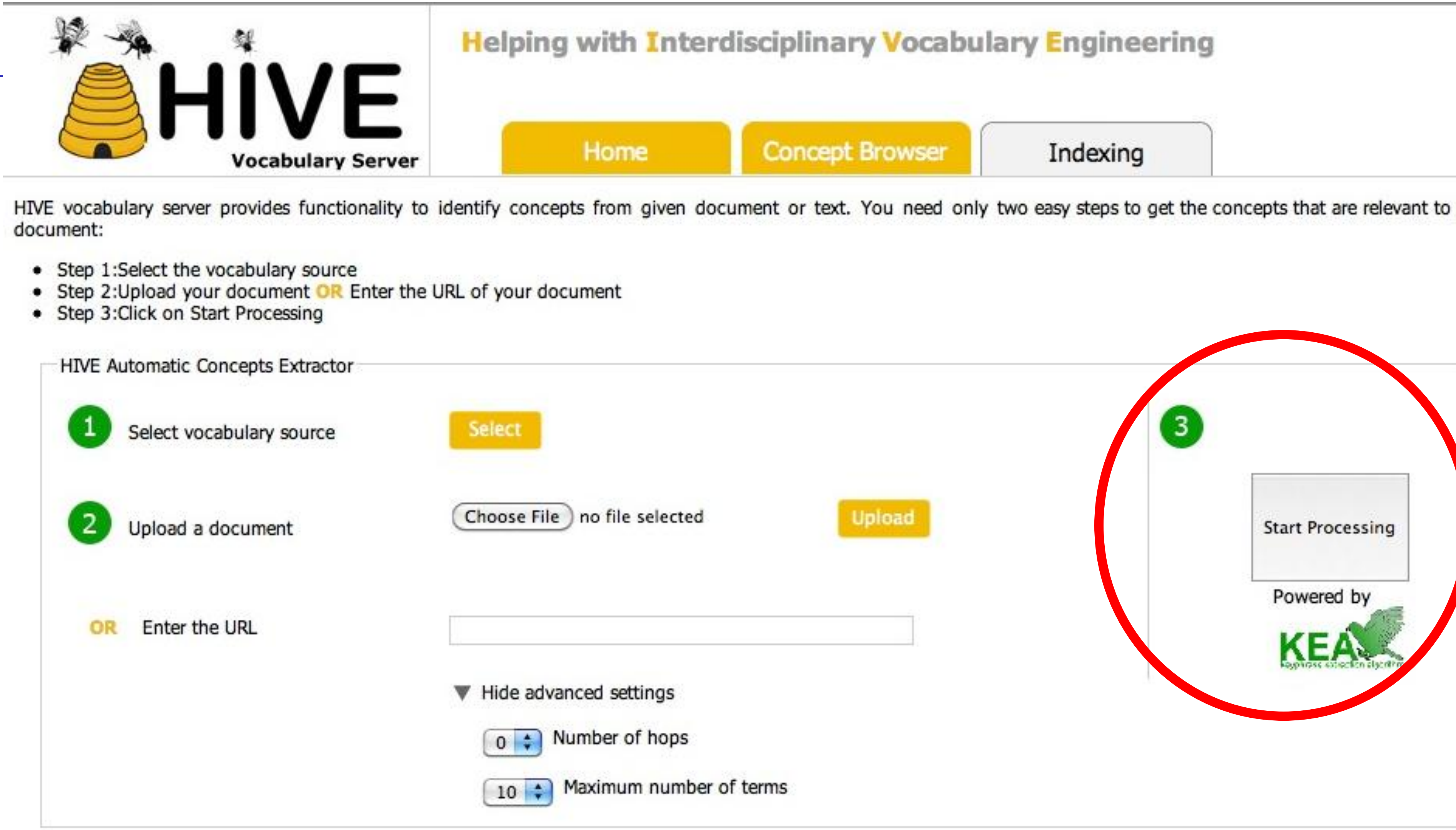
*If we [can] think about reuse as capital…?*

*Does this fit w/Robert Stevens notion of "active ontologies"?*

# Helping Interdisciplinary Vocabulary Engineering (HIVE)



- **Linked Open Vocabulary** initiative, to support inter/transdisciplinary….
- SKOS (a little dumb)
- AMG + machine learning approach for integrating discipline terminologies
- **Capital:** Productivity with metadata generation…



Helping with Interdisciplinary Vocabulary Engineering

HIVE vocabulary server provides functionality to identify concepts from given document or text. You need only two easy steps to get the concepts that are relevant to document:

- Step 1:Select the vocabulary source
- Step 2:Upload your document OR Enter the URL of your document
- Step 3:Click on Start Processing

HIVE Automatic Concepts Extractor

1 Select vocabulary source

2 Upload a document

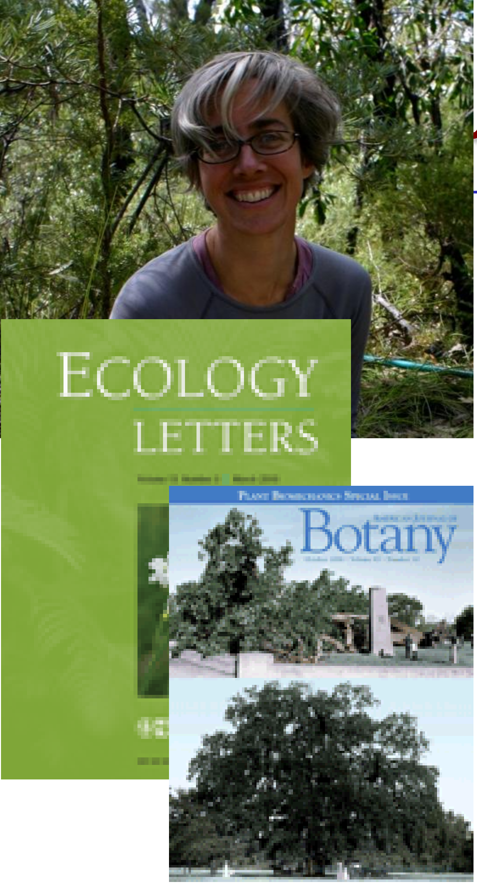
OR Enter the URL

▼ Hide advanced settings

0 Number of hops

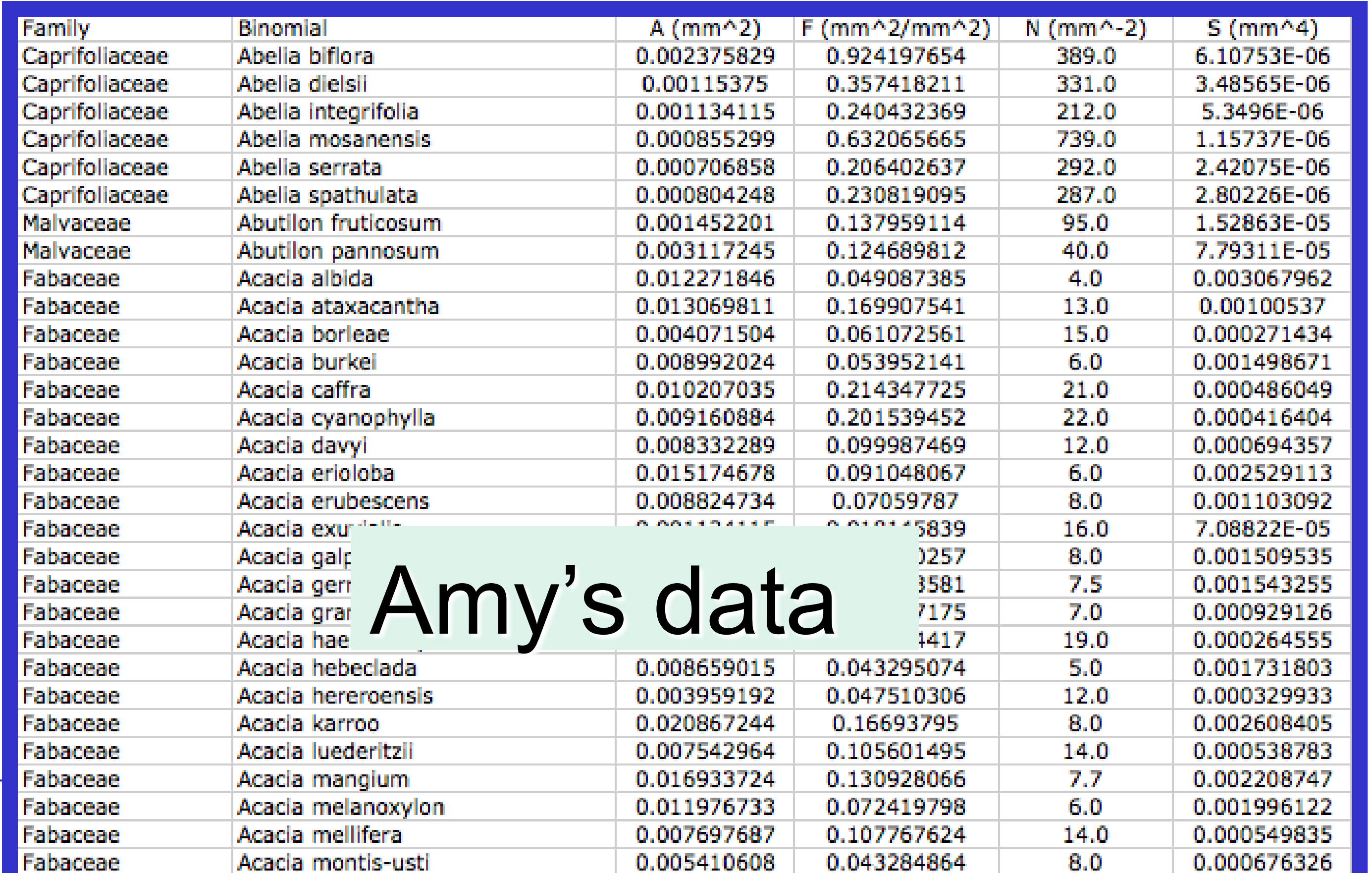
10 Maximum number of terms

3



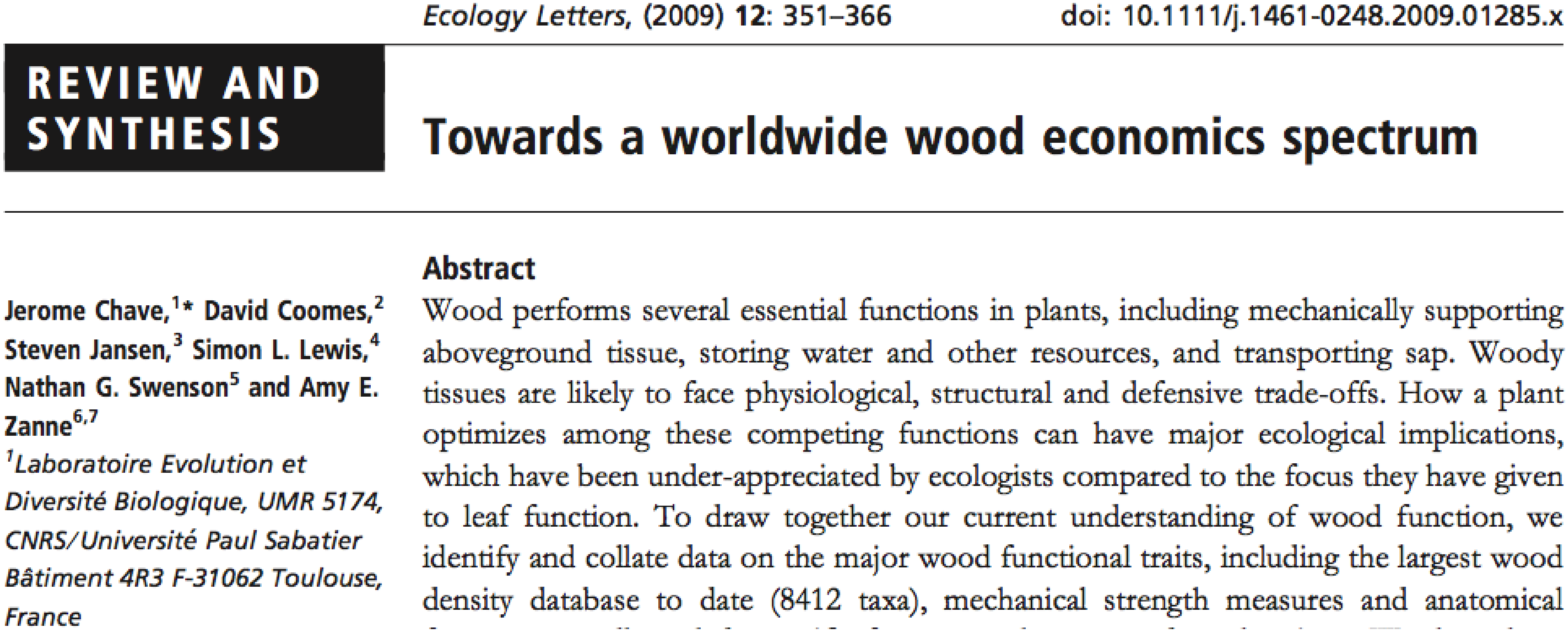
# ’Amy

- Meet Amy Zanne. She is a botanist.
- Like every good scientist, she publishes, and she deposits data in Dryad.



| Family | Binomial | A (mm^2) | F (mm^2/mm^2) | N (mm^-2) | S (mm^4) |
|---|---|---|---|---|---|
| Caprifoliaceae | Abelia biflora | 0.002375829 | 0.924197654 | 389.0 | 6.10753E-06 |
| Caprifoliaceae | Abelia dielsii | 0.00115375 | 0.357418211 | 331.0 | 3.48565E-06 |
| Caprifoliaceae | Abelia integrifolia | 0.001134115 | 0.240432369 | 212.0 | 5.3496E-06 |
| Caprifoliaceae | Abelia mosanensis | 0.000855299 | 0.632065665 | 739.0 | 1.15737E-06 |
| Caprifoliaceae | Abelia serrata | 0.000706858 | 0.206402637 | 292.0 | 2.42075E-06 |
| Caprifoliaceae | Abelia spathulata | 0.000804248 | 0.230819095 | 287.0 | 2.80226E-06 |
| Malvaceae | Abutilon fruticosum | 0.001452201 | 0.137959114 | 95.0 | 1.52863E-05 |
| Malvaceae | Abutilon pannosum | 0.003117245 | 0.124689812 | 40.0 | 7.79311E-05 |
| Fabaceae | Acacia albida | 0.012271846 | 0.049087385 | 4.0 | 0.003067962 |
| Fabaceae | Acacia ataxacantha | 0.013069811 | 0.169907541 | 13.0 | 0.00100537 |
| Fabaceae | Acacia borleae | 0.004071504 | 0.061072561 | 15.0 | 0.000271434 |
| Fabaceae | Acacia burkei | 0.008992024 | 0.053952141 | 6.0 | 0.001498671 |
| Fabaceae | Acacia caffra | 0.010207035 | 0.214347725 | 21.0 | 0.000486049 |
| Fabaceae | Acacia cyanophylla | 0.009160884 | 0.201539452 | 22.0 | 0.000416404 |
| Fabaceae | Acacia davyi | 0.008332289 | 0.099987469 | 12.0 | 0.000694357 |
| Fabaceae | Acacia erioloba | 0.015174678 | 0.091048067 | 6.0 | 0.002529113 |
| Fabaceae | Acacia erubescens | 0.008824734 | 0.07059787 | 8.0 | 0.001103092 |
| Fabaceae | Acacia exu[illegible] | [illegible] | [illegible]5839 | 16.0 | 7.08822E-05 |
| Fabaceae | Acacia galp[illegible] | [illegible] | [illegible]0257 | 8.0 | 0.001509535 |
| Fabaceae | Acacia gerr[illegible] | [illegible] | [illegible]3581 | 7.5 | 0.001543255 |
| Fabaceae | Acacia grar[illegible] | [illegible] | [illegible]7175 | 7.0 | 0.000929126 |
| Fabaceae | Acacia hae[illegible] | [illegible] | [illegible]4417 | 19.0 | 0.000264555 |
| Fabaceae | Acacia hebeclada | 0.008659015 | 0.043295074 | 5.0 | 0.001731803 |
| Fabaceae | Acacia hereroensis | 0.003959192 | 0.047510306 | 12.0 | 0.000329933 |
| Fabaceae | Acacia karroo | 0.020867244 | 0.16693795 | 8.0 | 0.002608405 |
| Fabaceae | Acacia luederitzii | 0.007542964 | 0.105601495 | 14.0 | 0.000538783 |
| Fabaceae | Acacia mangium | 0.016933724 | 0.130928066 | 7.7 | 0.002208747 |
| Fabaceae | Acacia melanoxylon | 0.011976733 | 0.072419798 | 6.0 | 0.001996122 |
| Fabaceae | Acacia mellifera | 0.007697687 | 0.107767624 | 14.0 | 0.000549835 |
| Fabaceae | Acacia montis-usti | 0.005410608 | 0.043284864 | 8.0 | 0.000676326 |

Amy’s data



*Ecology Letters*, (2009) **12:** 351–366 doi: 10.1111/j.1461-0248.2009.01285.x

**REVIEW AND SYNTHESIS**

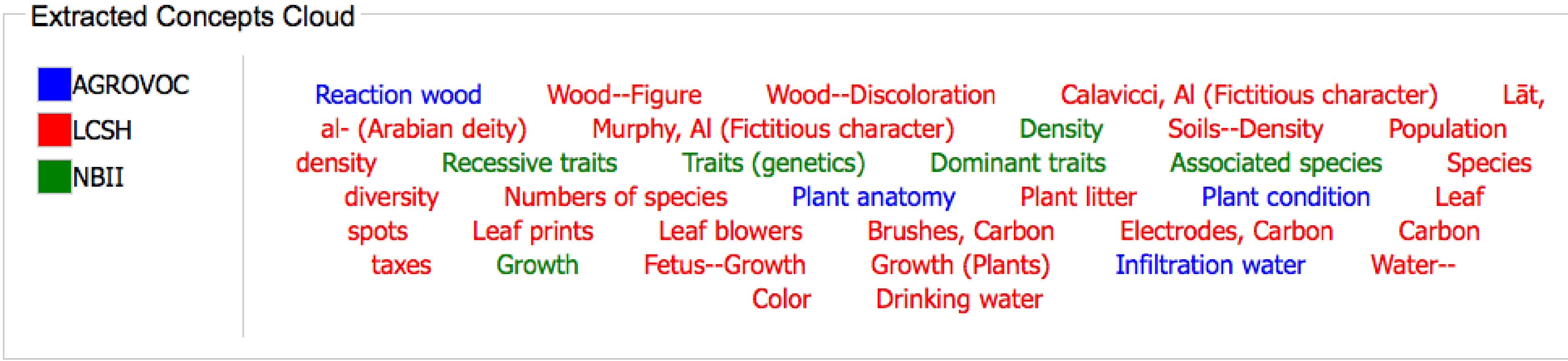
# Towards a worldwide wood economics spectrum

**Jerome Chave,[1]* David Coomes,[2] Steven Jansen,[3] Simon L. Lewis,[4] Nathan G. Swenson[5] and Amy E. Zanne[6,7]**

*[1]Laboratoire Evolution et Diversité Biologique, UMR 5174, CNRS/Université Paul Sabatier Bâtiment 4R3 F-31062 Toulouse, France*

## Abstract

Wood performs several essential functions in plants, including mechanically supporting aboveground tissue, storing water and other resources, and transporting sap. Woody tissues are likely to face physiological, structural and defensive trade-offs. How a plant optimizes among these competing functions can have major ecological implications, which have been under-appreciated by ecologists compared to the focus they have given to leaf function. To draw together our current understanding of wood function, we identify and collate data on the major wood functional traits, including the largest wood density database to date (8412 taxa), mechanical strength measures and anatomical



Extracted Concepts Cloud

🟦 AGROVOC
🟥 LCSH
🟩 NBII

Reaction wood Wood--Figure Wood--Discoloration Calavicci, Al (Fictitious character) Lāt, al- (Arabian deity) Murphy, Al (Fictitious character) Density Soils--Density Population density Recessive traits Traits (genetics) Dominant traits Associated species Species diversity Numbers of species Plant anatomy Plant litter Plant condition Leaf spots Leaf prints Leaf blowers Brushes, Carbon Electrodes, Carbon Carbon taxes Growth Fetus--Growth Growth (Plants) Infiltration water Water--Color Drinking water

# Modified Capital-sigma notation

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

P = value of the metadata property (w/HIVE the linked data concept)

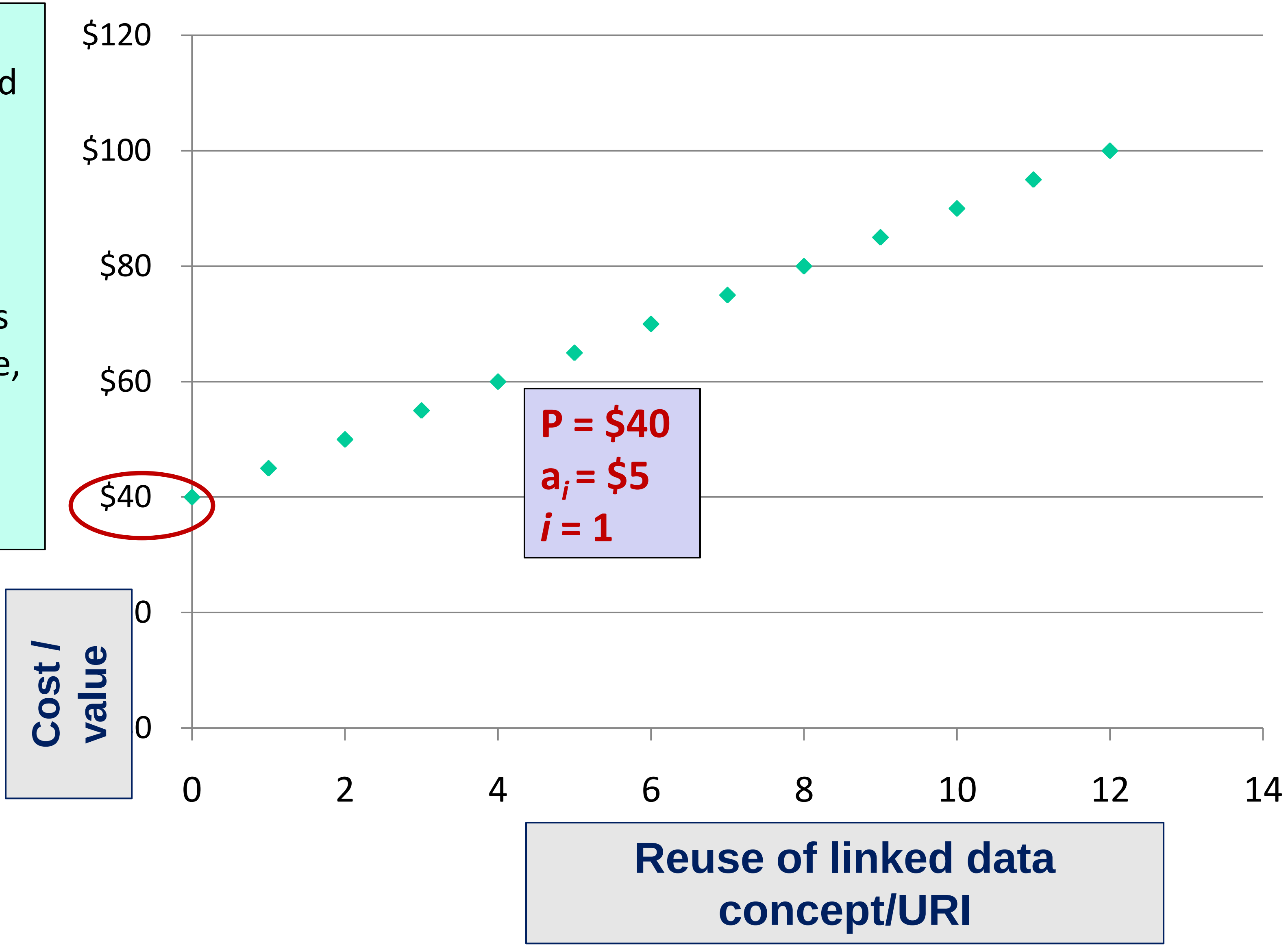
i = number of usages (reuse)

a = incremental increase in value

n = maximum number of reuse ?

# Modified Capital-sigma notation for linked data

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

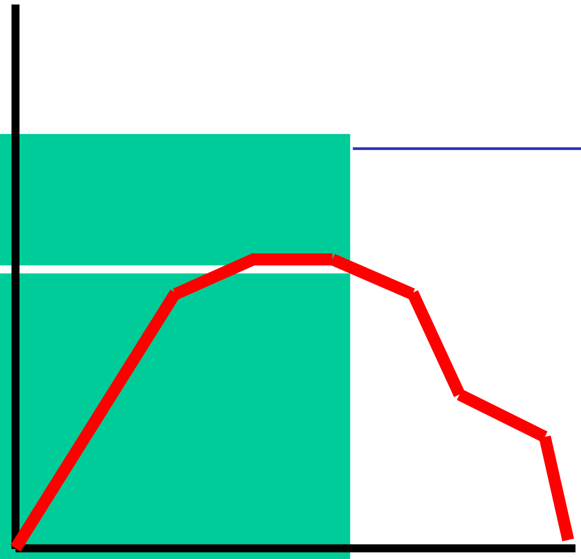


## Successive growth rates

N

$\sum i^c = \Theta (n^c +1)$

i=1



**Cycles…**

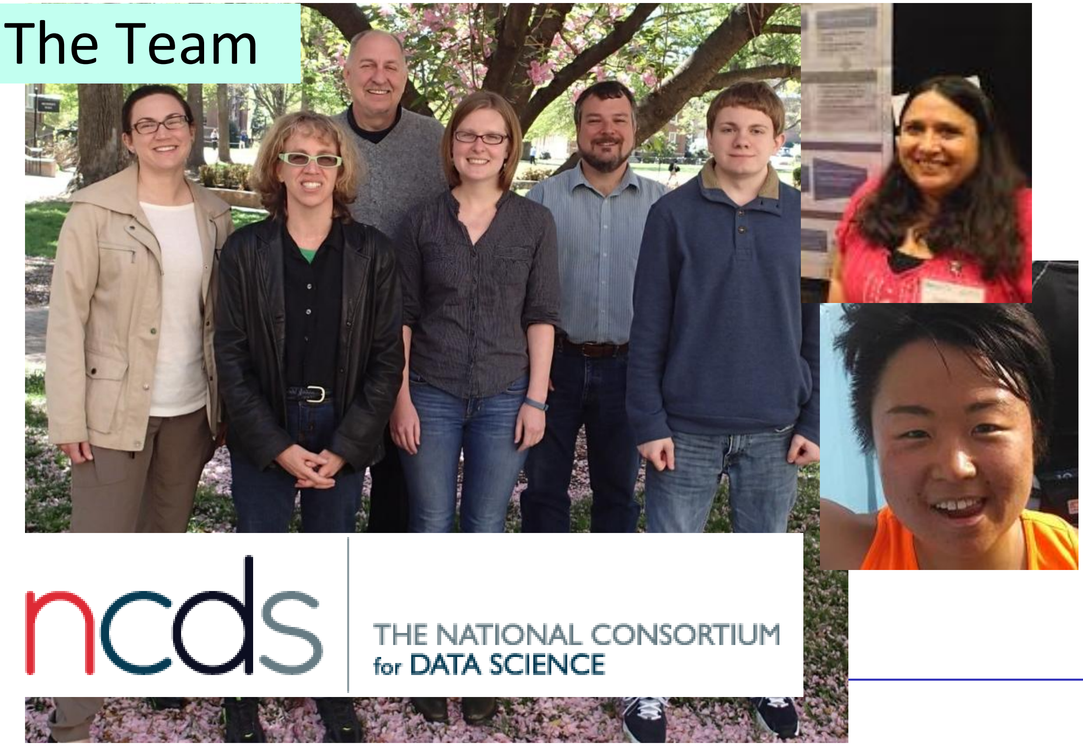
What about successive growth rate tied to a concept?

A concept can be

- in ~ vernacular to canonical
- fall by the wayside, less popular
- out (deprecated)

# Discover and advance the application of methods for quantifying the cost and value of metadata over time; raise dialog

1. Advance nascent work on **“metadata capital”** for data science
2. Actively engage with the NCDS community



The Team

3. Connect NCDS metadata efforts w/the **Research Data Alliance**

DREXEL UNIVERSITY
Metadata Research Center
College of Computing & Informatics

# Research environments

1. **Self-generated health information (SGHI)** monitoring daily activity in collaboration with the Research Triangle Institute (RTI) (**Tom Caruso**, Health Information Liaison Research Associate, UNC-SILS/RTI)
   - Fitbit; mobile health
   - Consumer/patient awareness
   - Metadata/data ownership; cost generating, capital via use/re-use

SGHIx CONSORTIUM

RTI INTERNATIONAL

2. **Data management/ontology development** in collaboration with the National Institute for Environmental Health Sciences (NIEHS). **Rebecca Boyles**, Data Scientist, NIEHS

- Viral vector core
- Prevent re-running experiments

Accounting factor/cost analysis

NIEHS

i·R·O·D·S

Integrated Rule-Oriented Data System

DREXEL UNIVERSITY
Metadata
Research Center
*College of Computing & Informatics*

# Capital: *does this work have merit for <m> quality?*

Possibly – to get R&D support
Hard work

Quality

- Support metadata functions – discovery, provenance tracking, authenticity
- Standard of taste
- We like it
- Impact

Capital

- Financial
- Social
- Intellectual
- Objects
  - Asset, product, service, good, public good

(Greenberg, ASIST Bulletin, 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

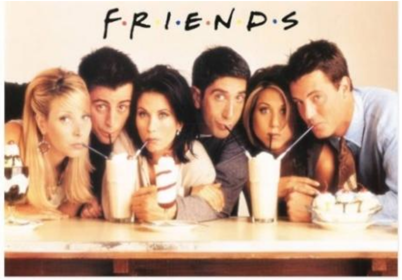
# Conclusion…other Valuation Approaches

- **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.

# Concluding remarks

- Interest….traction
- Limitations: bad data, cost/value, more metadata
- We should care about cost
- Metadata capital can contextualize the discussion, provide a foundation
- Generic formula for further research
  - Proof

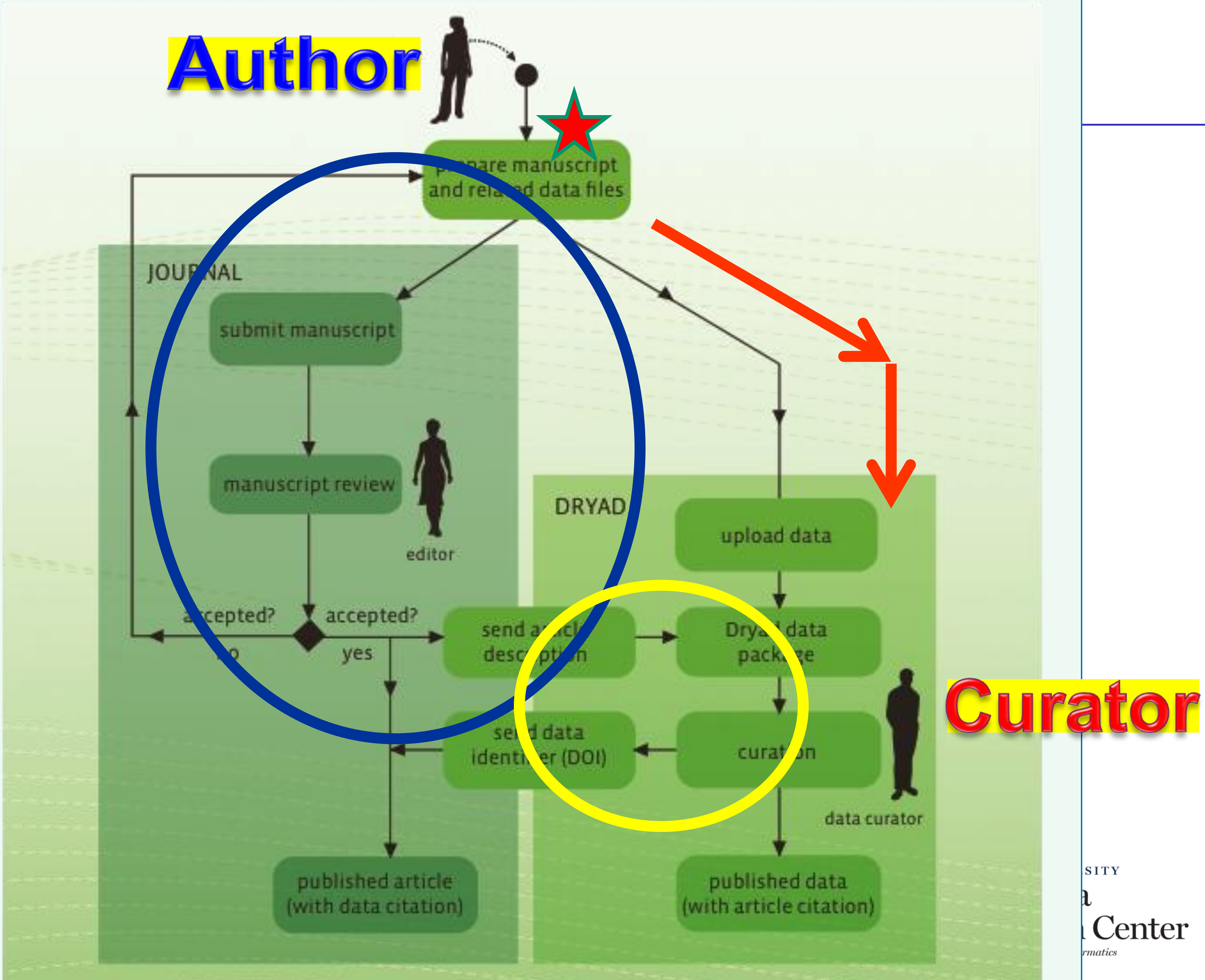


# Acknowledgments and more information

- National Consortium for Data Science (NCDS)
- CCI/Drexel <Metadata Research Center>.
- NESCent
- DataNet Federation Consortium (Reagan Moore, Mike Conway, Le Zhan)
- Dryad + HIVE (many many people)

------------------------------------------------------------

- NCDS Fellowship program
- U.S. National Science Foundation Award number: OCI 0940841, and award number: 1147166
- [http.ccc.drexel.edu/mrc](http.ccc.drexel.edu/mrc)



SITY
a
Center
rmatics