YAMZ—Yet Another Metadata Zoo, HIVE (Helping Interdisciplinary Vocabulary Engineering) , and Semantic Interoperability

Semantic Registry workshop
9 Dec. 2013, Utrecht University
(http://www.isocat.org/2013-SR/)

Jane Greenberg
Metadata Research Center/SILS UNC-CH
janeg@email.unc.edu

DFC DataNet FEDERATION CONSORTIUM

<MRC>

DataONE

Overview

1. Assumptions, motivation
2. Introduce YAMZ (formerly ‘SeaIce’)
3. Address questions
4. “briefly” share about HIVE (Helping Interdisciplinary Vocabulary Engineering)
   - Dynamic, on-the-fly registration
5. Address questions
6. Conclusions, Q&A

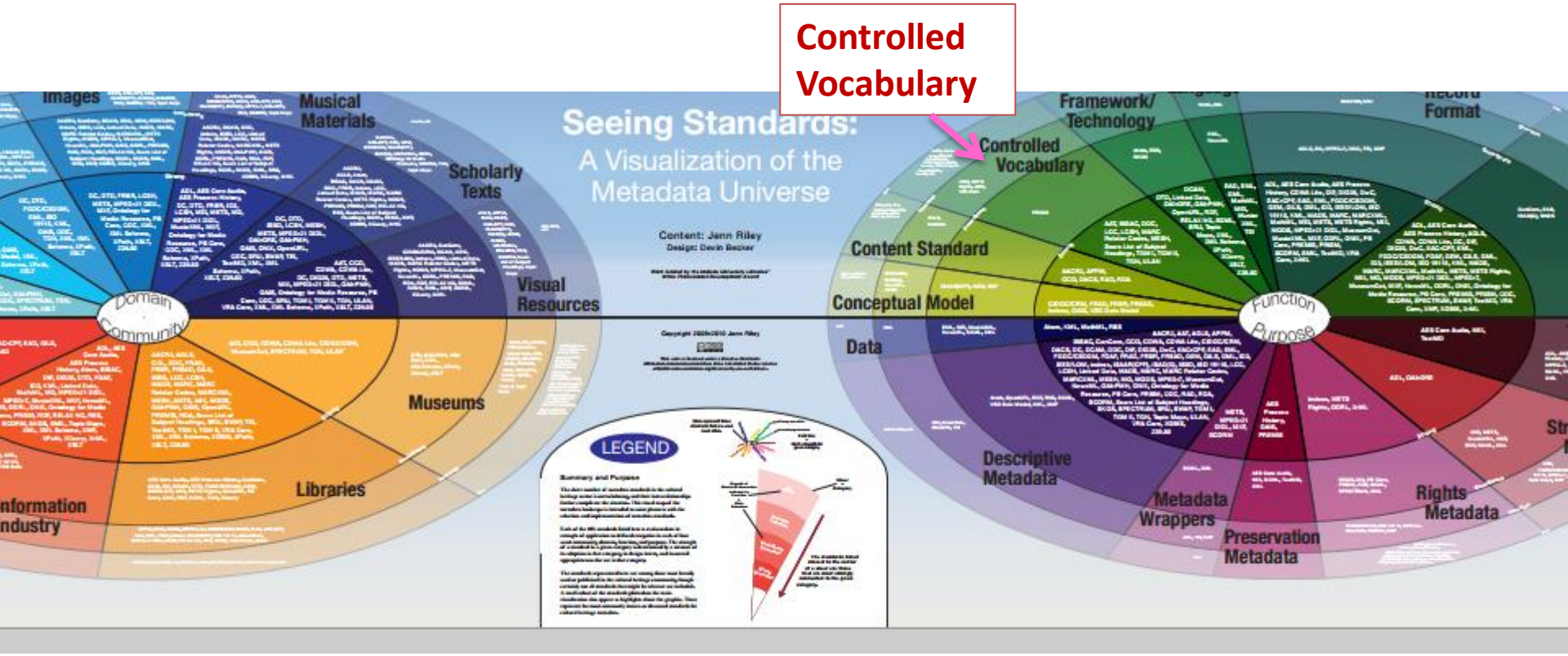
Assumptions and motivation

- Prevailing methods of semantic registration + / -
- More than one way to skin a cat
  - Complementary, alternative approaches; DataONE: social technologies; HIVE: LOD/LOV

Toothbrush

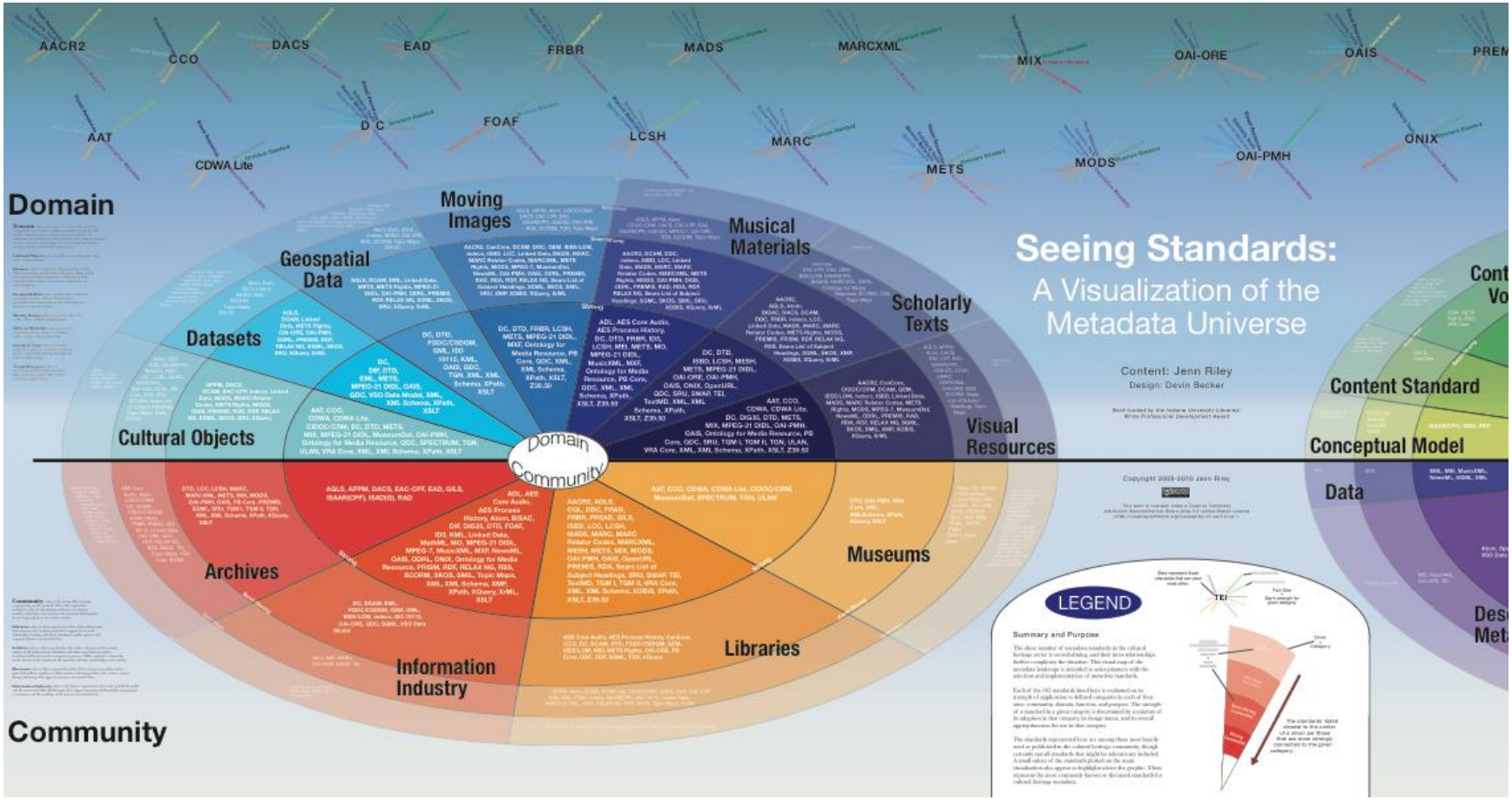
Long Tail

The Metadata Universe



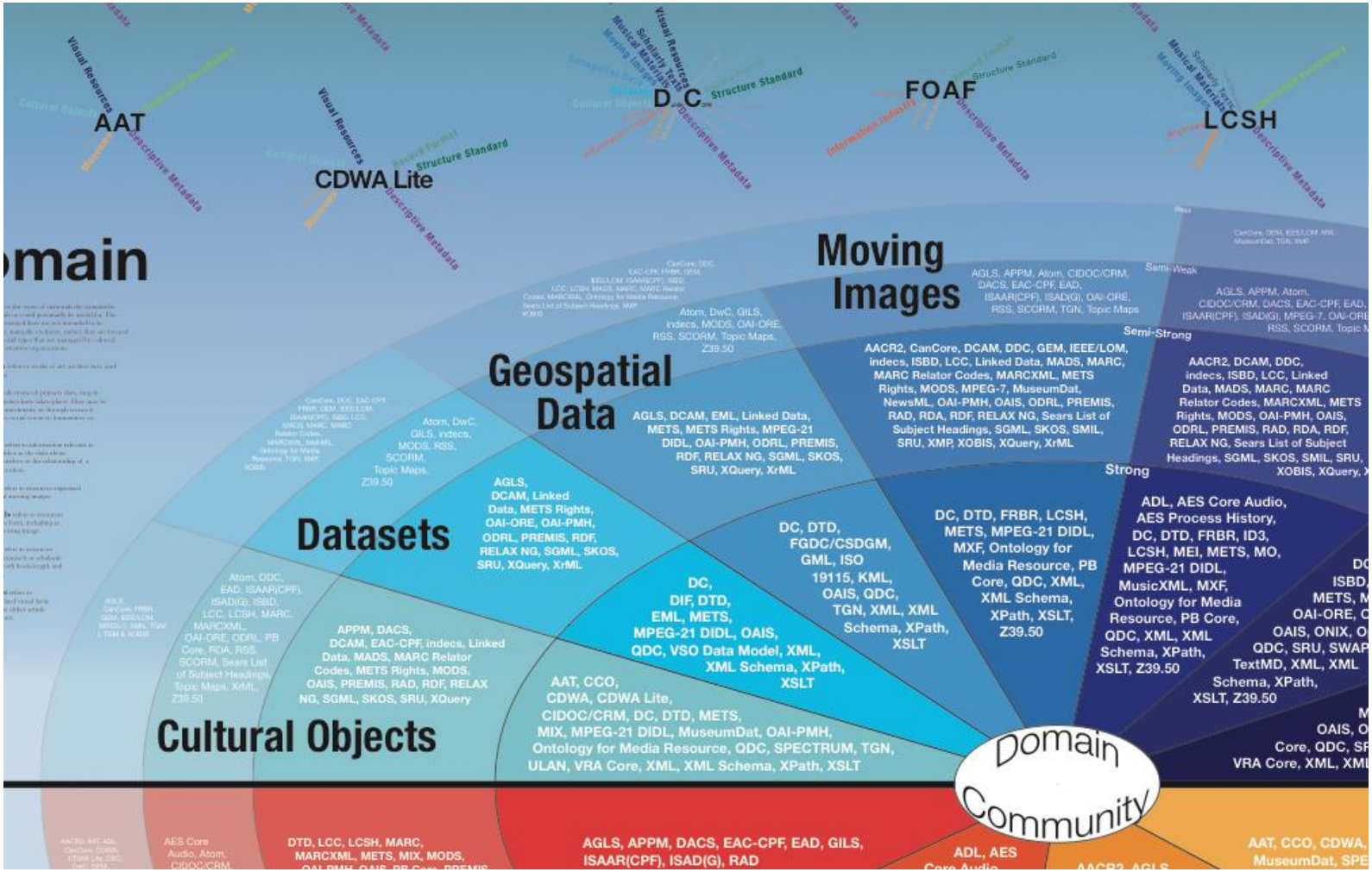
 Jenn Riley, IU , 2009-2010

The Metadata Universe



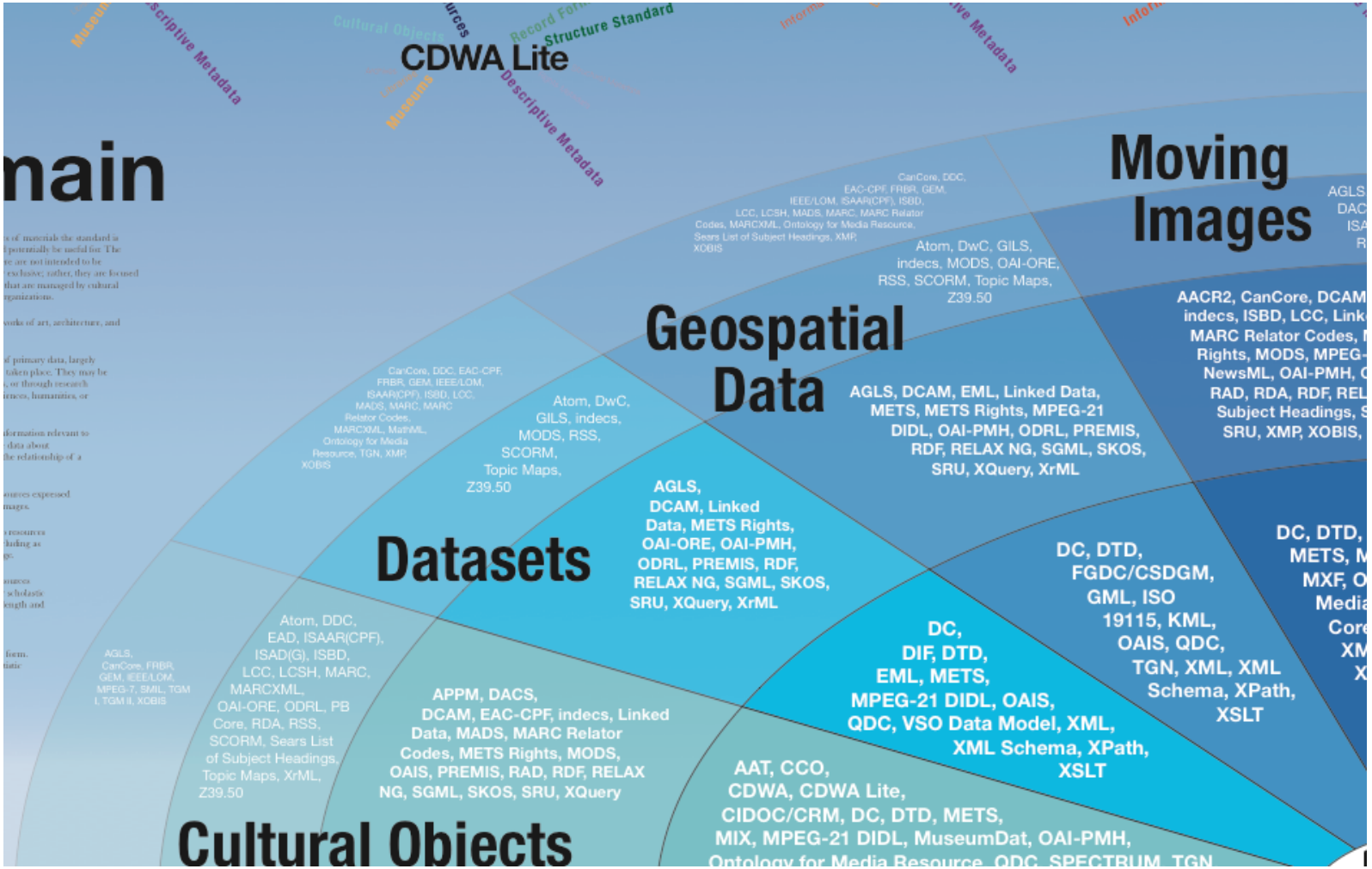
 Jenn Riley, IU , 2009-2010

The Metadata Universe



Jenn Riley, IU , 2009-2010

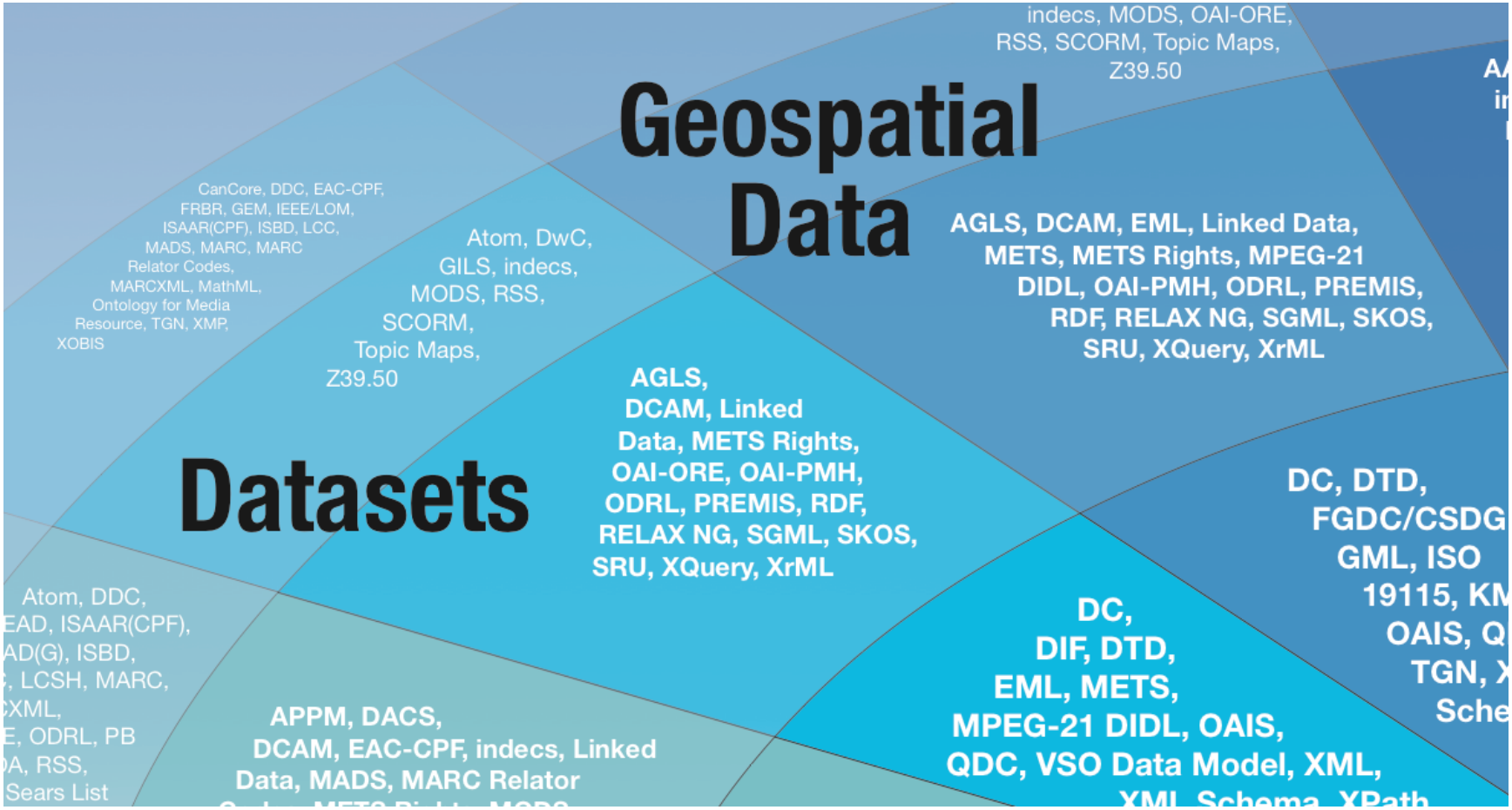
The Metadata Universe



 Jenn Riley, IU , 2009-2010

The Metadata Universe

How can stakeholders including machines navigate this space more efficiently and effectively?

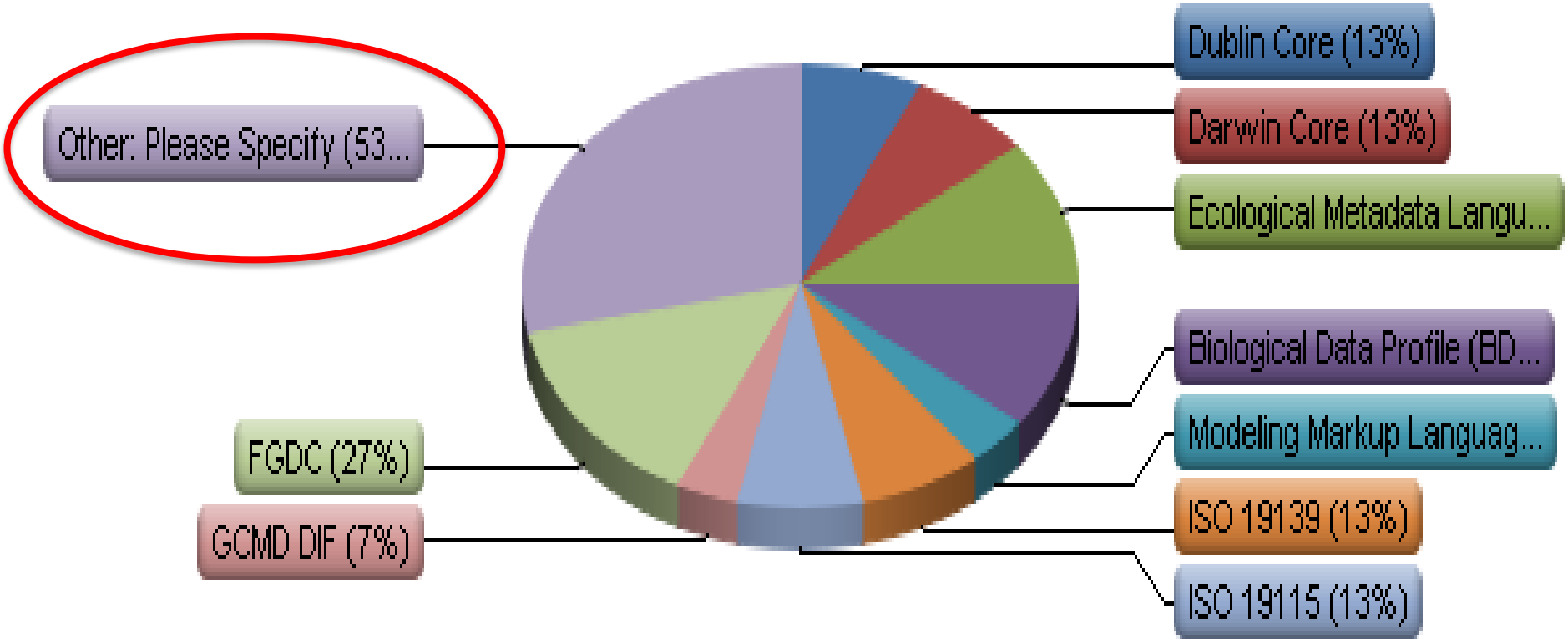


EVIDENCE: Barriers to access, not using standard semantics

Components of Successful Metadata Registry Frameworks (**A. Murillo, 2012)**

14 standardized schemes used, lots of in-house

n = ~ 100 (biology, earth science, computer science, etc.)



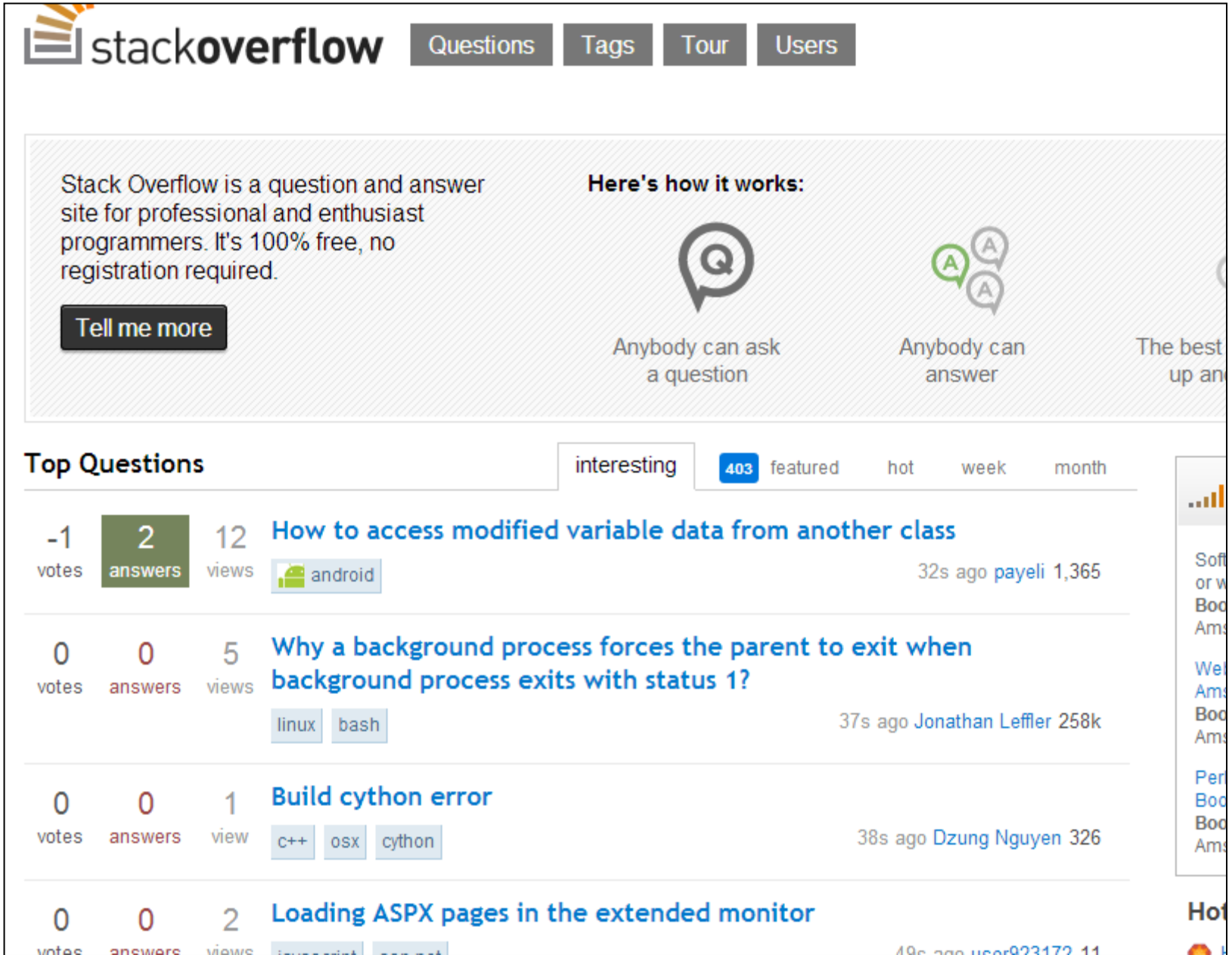
YAMZ – yet another metadata zoo

Metadata *Vision* for YAMZ (yamz.net)

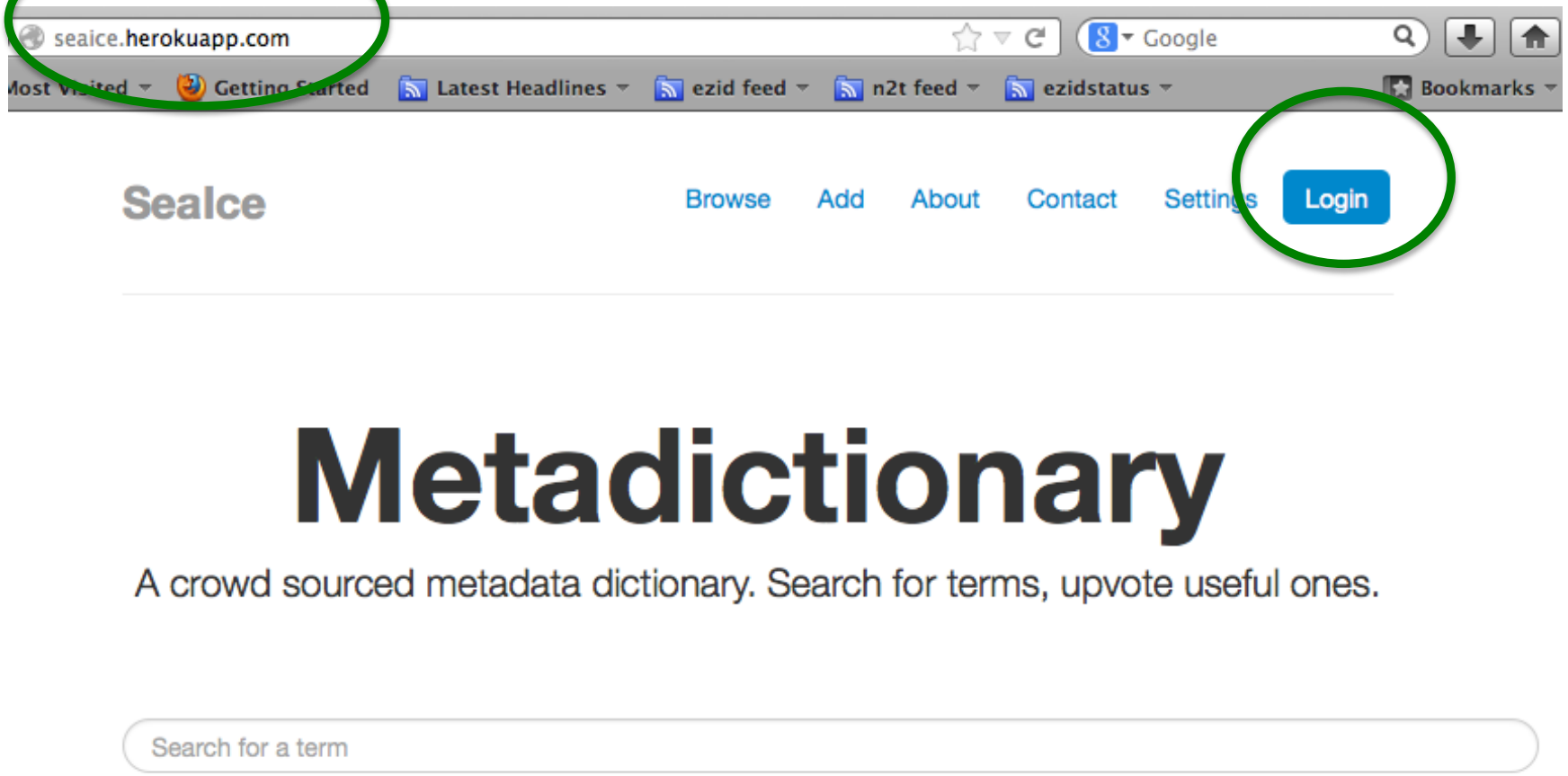
- *One dictionary, one namespace*
- Crowd sourced plus lightly supervised canon
- Anyone can look up terms
- Any domain, any part of “metadata speech”
  - Names, values, units, relationships, …
- Anyone can propose and refine their terms
- Strong terms rise, weak terms decline

What can we glean from Wikipedia, internet RFCs, and American Heritage Dictionary?

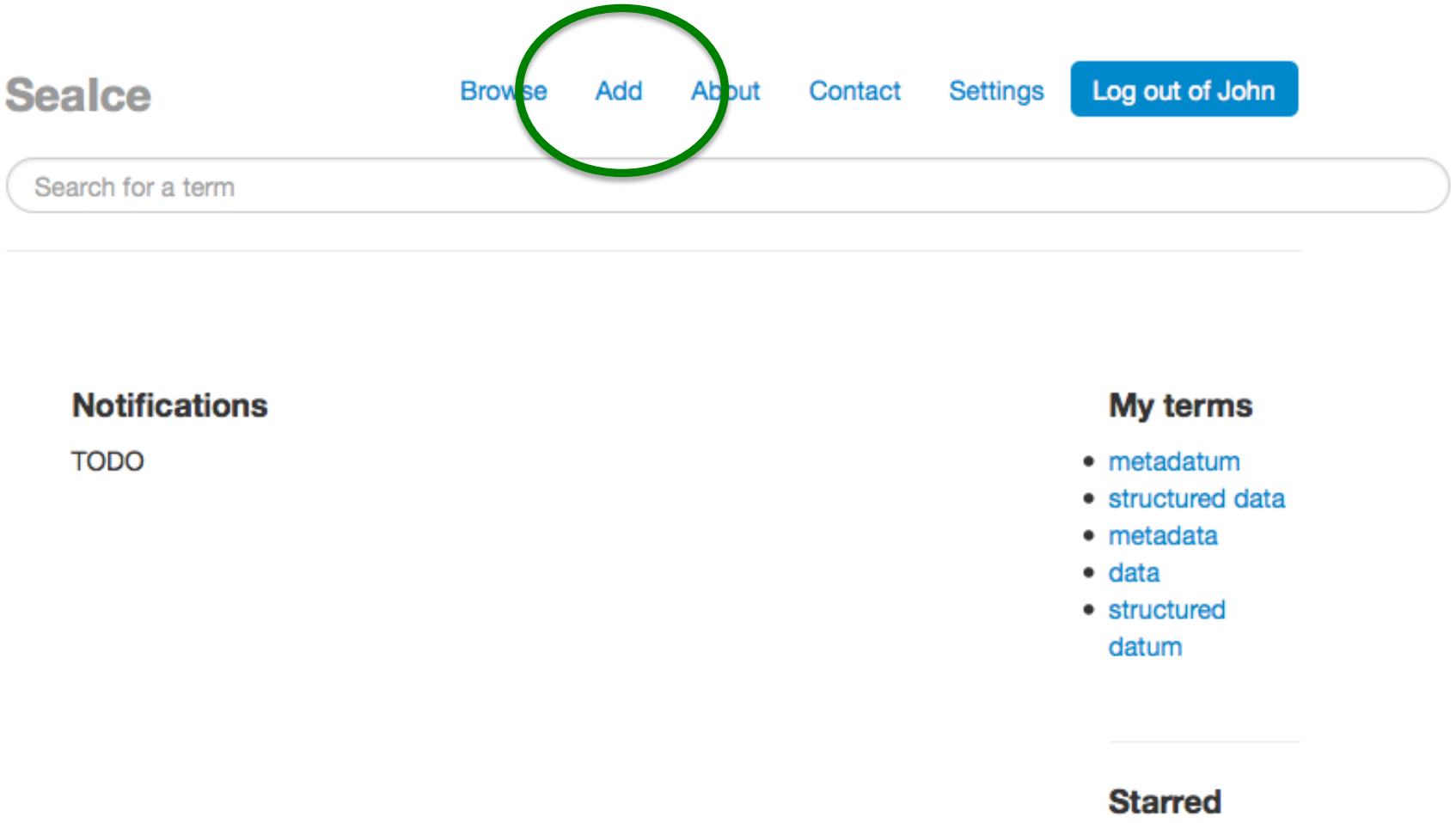
Stackoverlow

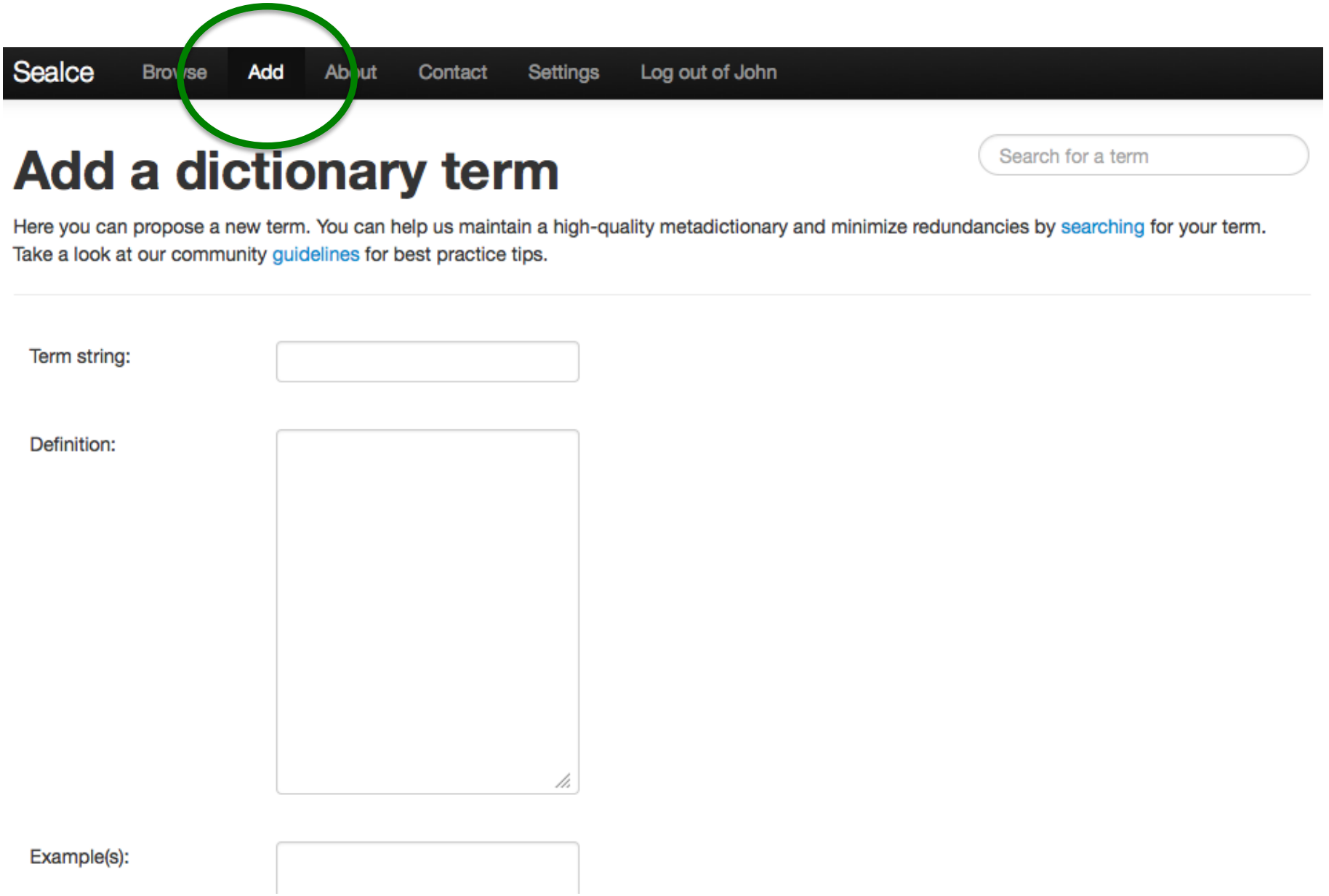


http://yamz.net/ Formerly Sealce



Tech stuff: Python + off-the shelf tools; freely hosted on heroku with the evolving code on github; Project code name is Sealce.


Sealce
Browse
Add
About
Contact
Settings
Log out of John
Search for a term
Notifications
TODO
My terms
metadatum
structured data
metadata
data
structured datum
Starred

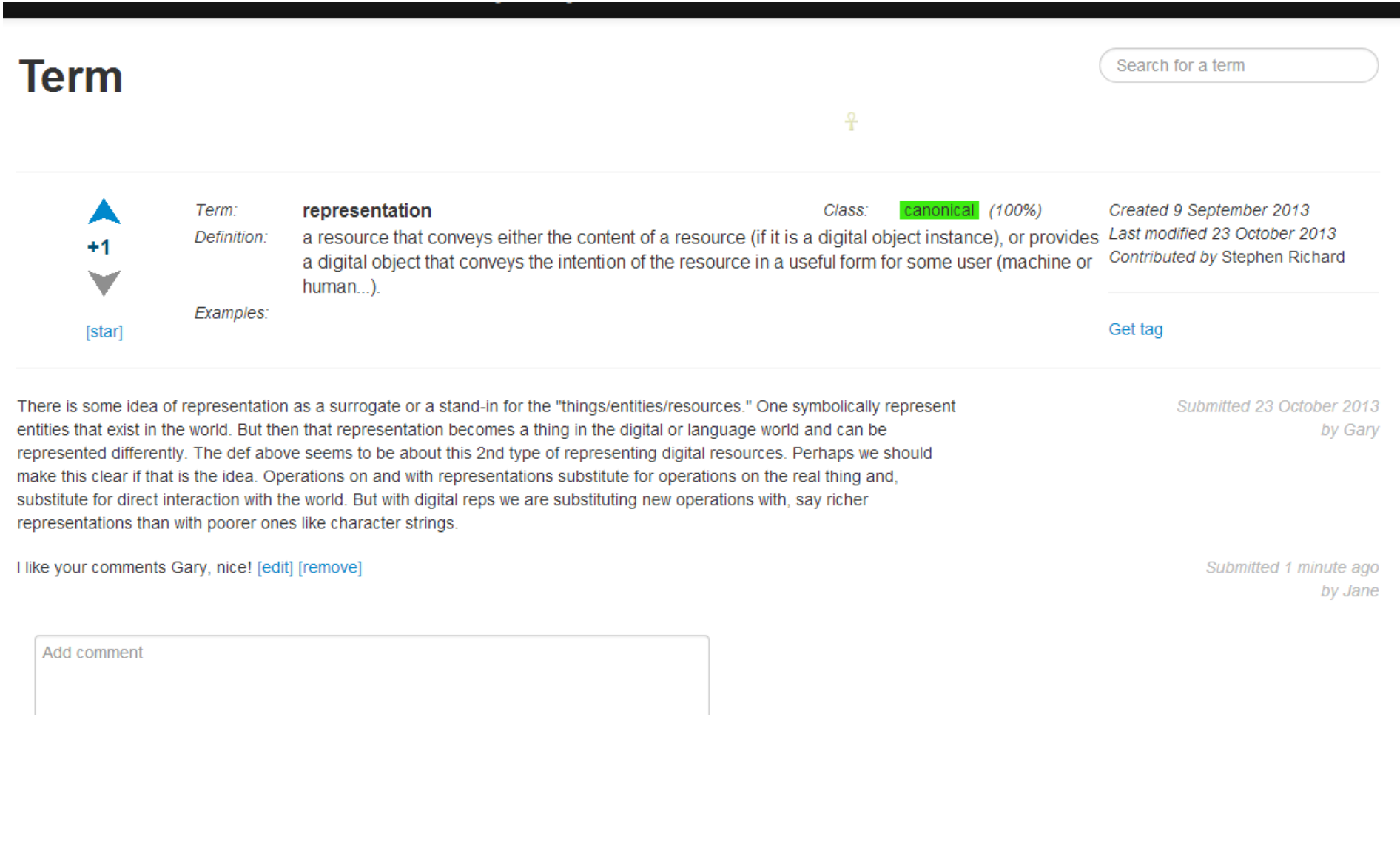

Sealce
Browse
Add
About
Contact
Settings
Log out of John
Add a dictionary term
Search for a term
Here you can propose a new term. You can help us maintain a high-quality metadictionary and minimize redundancies by searching for your term.
Take a look at our community guidelines for best practice tips.
Term string:
Definition:
Example(s):

Browse dictionary

high score | **recent** | **volatile** | **stable** | **alphabetical**

Term	Score	Consensus	Class	Contributed by	Last modified
data	2	100%	canonical	John Kunze	1 day ago
publisher	2	100%	canonical	John Kunze	5 days ago
creator	1	100%	canonical	John Kunze	6 days ago
datum	1	100%	canonical	John Kunze	1 day ago
description	1	100%	canonical	John Kunze	6 days ago
identifier	1	100%	canonical	John Kunze	1 day ago
metadata	1	100%	canonical	John Kunze	6 days ago
resource	1	100%	canonical	Chris Patton	2 days ago
identifier	1	66%	vernacular	John Kunze	5 days ago
datum	0	50%	vernacular	Chris Patton	1 day ago
hydraulic gradient	0	0%	vernacular	Angela Murillo	14 August 2013
structured data	0	0%	deprecated	John Kunze	9 August 2013
structured datum	0	0%	deprecated	John Kunze	5 days ago
talus slope	0	0%	deprecated	Angela Murillo	12 August 2013
great	-1	12%	deprecated	Nassib Nassar	6 days ago
CHL	-1	0%	vernacular	Greg Janée	6 days ago
metadatum	-1	0%	deprecated	John Kunze	12 August 2013
token	-1	0%	deprecated	John Kunze	1 day ago
talus	-2	0%	deprecated	Angela Murillo	6 days ago

Example of dialog



Term Classes and Voting Impact

Vernacular → canonical -- term is stable after two days and consensus is above 75%.
Vernacular → deprecated -- term is stable after two days and consensus is below 25%.
Canonical → vernacular -- term has been updated, restabilized, and consensus has dropped below 75%.
Deprecated → vernacular -- term has been updated, restabilized, and consensus has risen above 25%.

* Nothing firmed about percentages, *just an illustrative consideration*

Questions

1. In your community, how is the support for a semantic registry?
2. How open is the registry, who can add or modify things, are the semantics stable?
3. Does the data model of the registry match your needs (information entry, relationships, granularity, possible inconsistency, …)
4. How is the sustainability of the registry: financially, organizational, … ?

1. How is the support for a semantic registry?

- DataONE: Lots of exploration taking place with the Semantic WG, automatic ontology applications
- YAMZ:
  - Nascent state, exploratory, pre-beta
    - PPSR (Public Participants in Scientific Research)
    - Provenance WG
    - RDA Data Foundations and Terminology WG
  - Has tremendous potential
  - Concerns

    Not ready for prime time

    Identifier matters (Kunze, et al, 2013, CAMP-4-DATA), ARKS via EZID, LOD compliant

2. How open is the registry, who can add or modify things, are the semantics stable?

- Open to anyone
- Modifications:
  - Anyone can add
  - Anyone can suggest modifications/vote
  - Only creator can modify his/her term
  - Different view – suggest your own term/definition, etc.
- Semantics will be stable via identifiers

3. Does the data model of the registry match your needs (information entry, relationships, granularity, possible inconsistency, …)

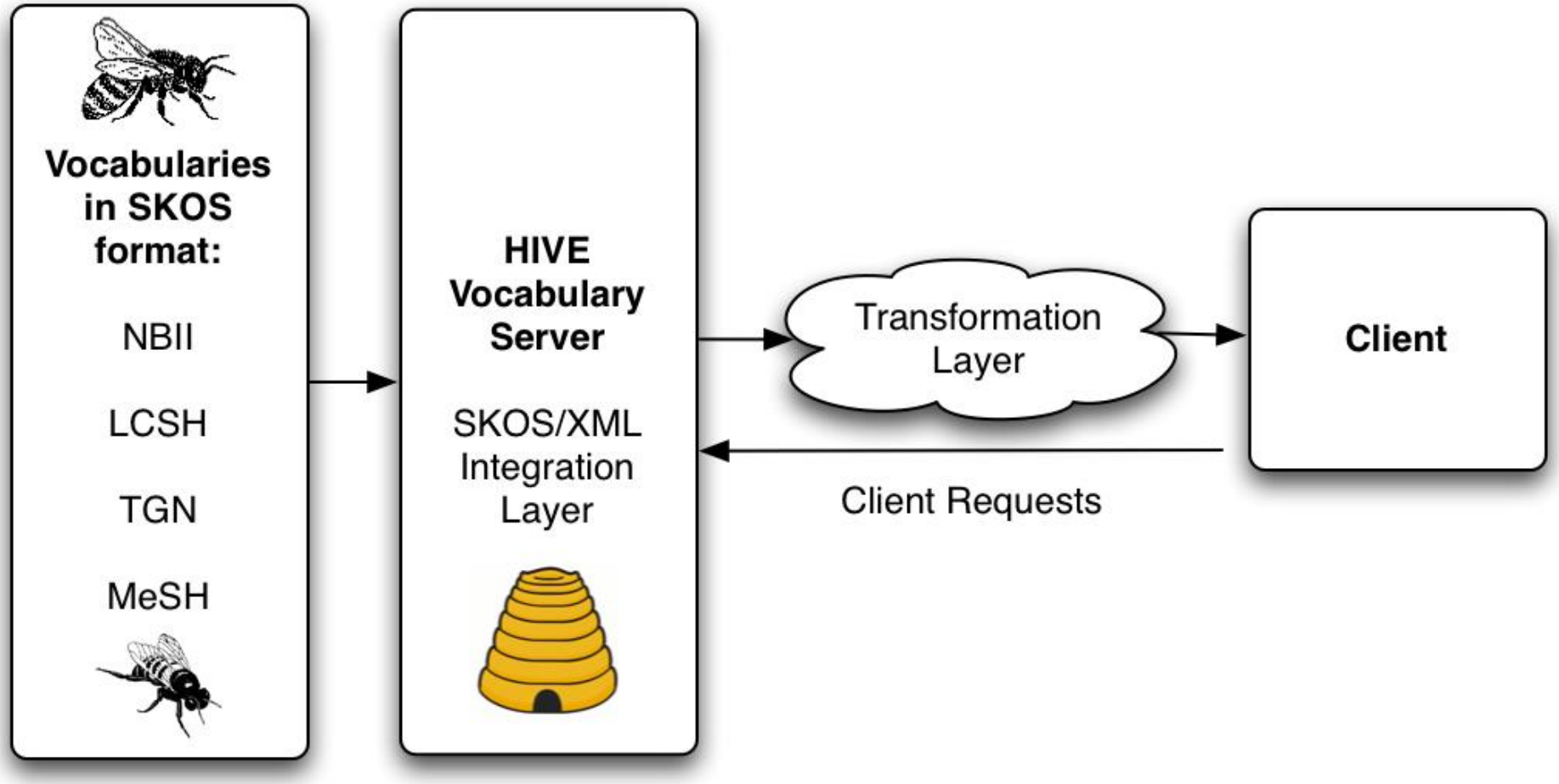
- *Perhaps…*too early to tell, experimental
- Signs of “yes” given update/interest within DataONE and RDA

4. How is the sustainability of the registry: financially, organizational, … ?

- Initial support via DataONE summer intern
- BIG question
  - Communication underway with other organizations
- Crucial question, hopes/reality for community buy-in

Dynamic, on-the-fly registration, taking advantage of LOD…

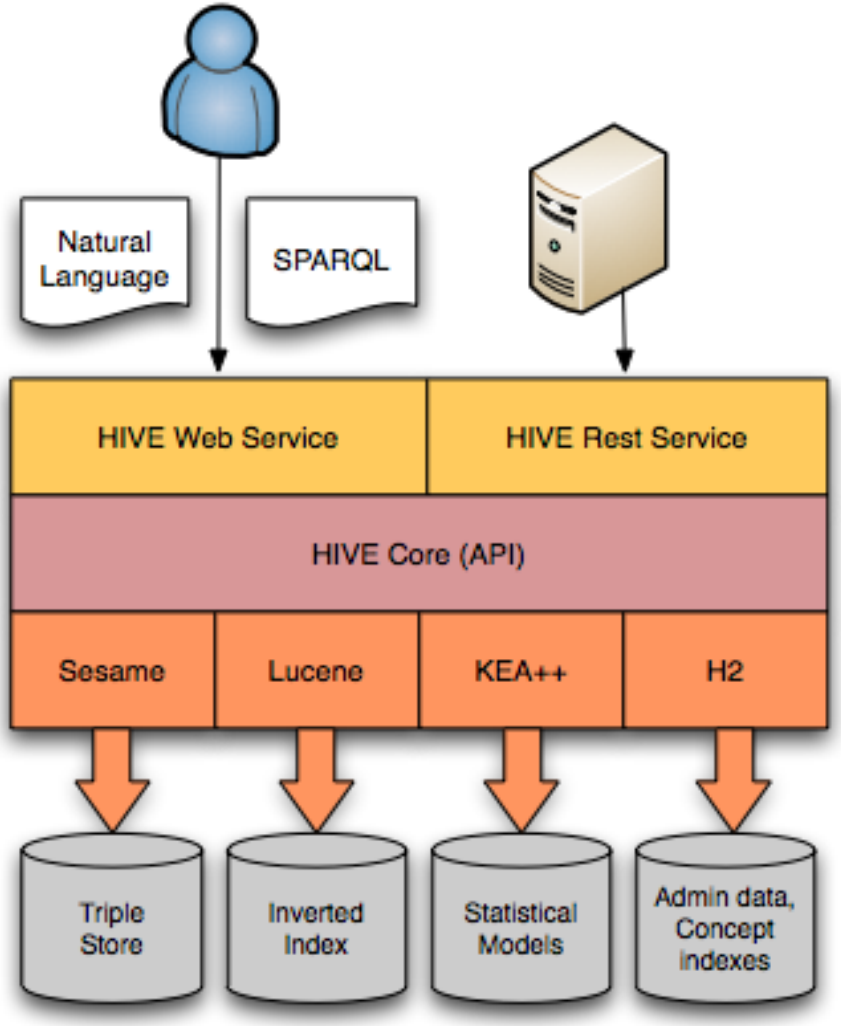
Helping Interdisciplinary Vocabulary Engineering (HIVE)

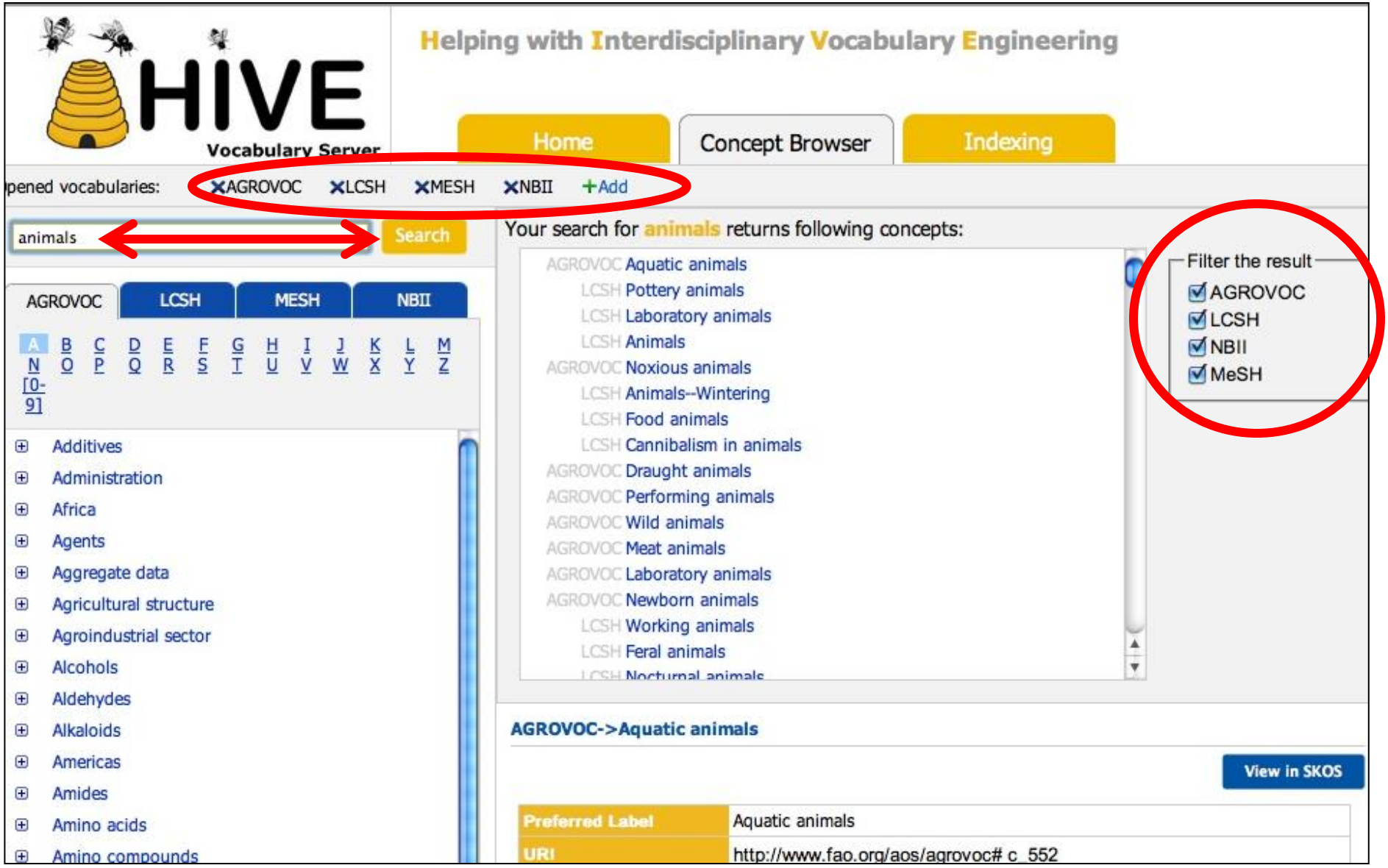


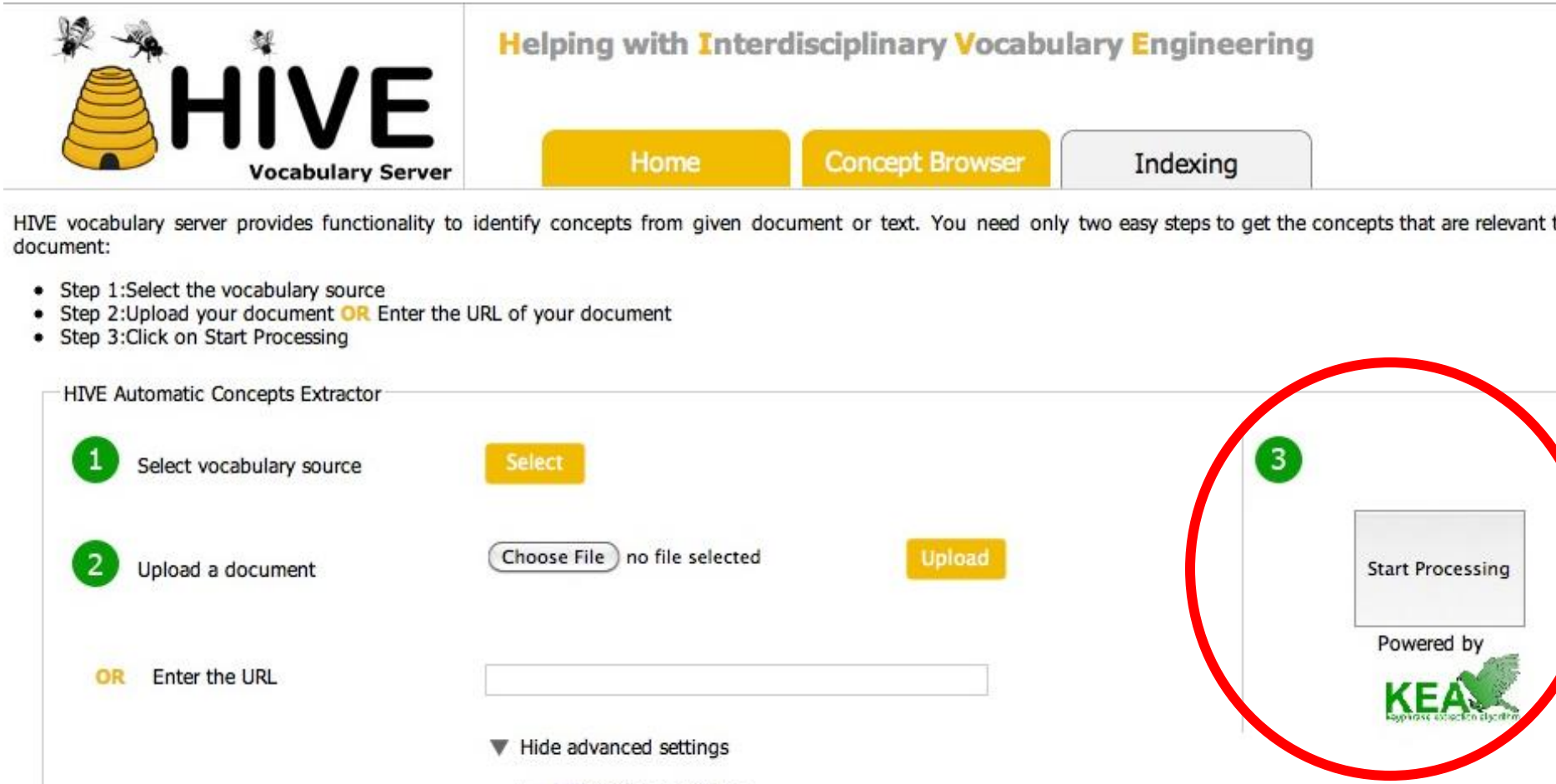
- **Linked Open Vocabulary** initiative, to support inter/transdisciplinary….
- SKOS (a little dumb)
- AMG + machine learning approach for integrating discipline terminologies
- 2 NSF DataNet projects: DataONE and DFC, prototype in Dryad, other uses

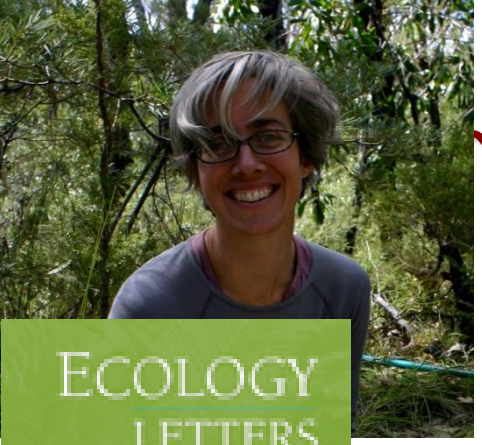
Technical overview and architecture

- HIVE combines several open-source technologies to provide a framework for vocabulary services.
- Java-based web services can run in any Java application server
- Demonstration website @ RENCI and NESCent
- Open-source Google Code project, in process of moving to Github (http://code.google.com/p/hive-mrc/)



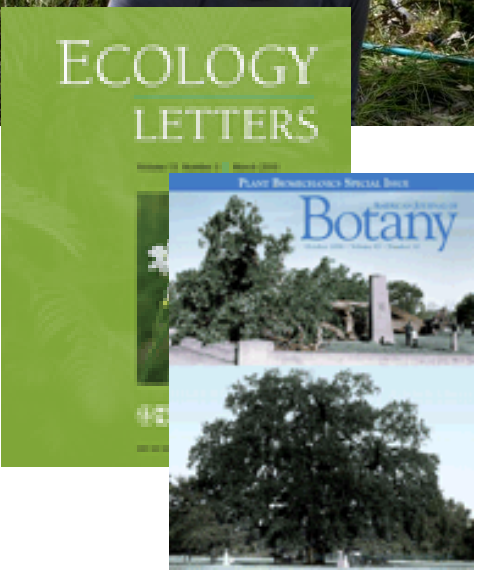

HIVE
Vocabulary Server
Helping with Interdisciplinary Vocabulary Engineering
Home
Concept Browser
Indexing
pened vocabularies:
AGROVOC
LCSH
MESH
NBII
+Add
animals
Search
Your search for animals returns following concepts:
AGROVOC Aquatic animals
LCSH Pottery animals
LCSH Laboratory animals
LCSH Animals
AGROVOC Noxious animals
LCSH Animals--Wintering
LCSH Food animals
LCSH Cannibalism in animals
AGROVOC Draught animals
AGROVOC Performing animals
AGROVOC Wild animals
AGROVOC Meat animals
AGROVOC Laboratory animals
AGROVOC Newborn animals
LCSH Working animals
LCSH Feral animals
Filter the result
AGROVOC
LCSH
NBII
MeSH
AGROVOC
LCSH
MESH
NBII
A B C D E F G H I J K L M
N O P Q R S T U V W X Y Z
[0-9]
Additives
Administration
Africa
Agents
Aggregate data
Agricultural structure
Agroindustrial sector
Alcohols
Aldehydes
Alkaloids
Americas
Amides
Amino acids
Amino compounds
AGROVOC->Aquatic animals
View in SKOS
Preferred Label
Aquatic animals
URI
http://www.fao.org/aos/agrovoc# c_552


HIVE
Vocabulary Server
Helping with Interdisciplinary Vocabulary Engineering
Home
Concept Browser
Indexing
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
Select
2
Upload a document
Choose File
no file selected
Upload
OR
Enter the URL
Hide advanced settings
0
Number of hops
10
Maximum number of terms
3
Start Processing
Powered by
KEA



~Amy

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



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			16.0	7.08822E-05
Fabaceae	Acacia galp			8.0	0.001509535
Fabaceae	Acacia ger			7.5	0.001543255
Fabaceae	Acacia gra			7.0	0.000929126
Fabaceae	Acacia hae			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

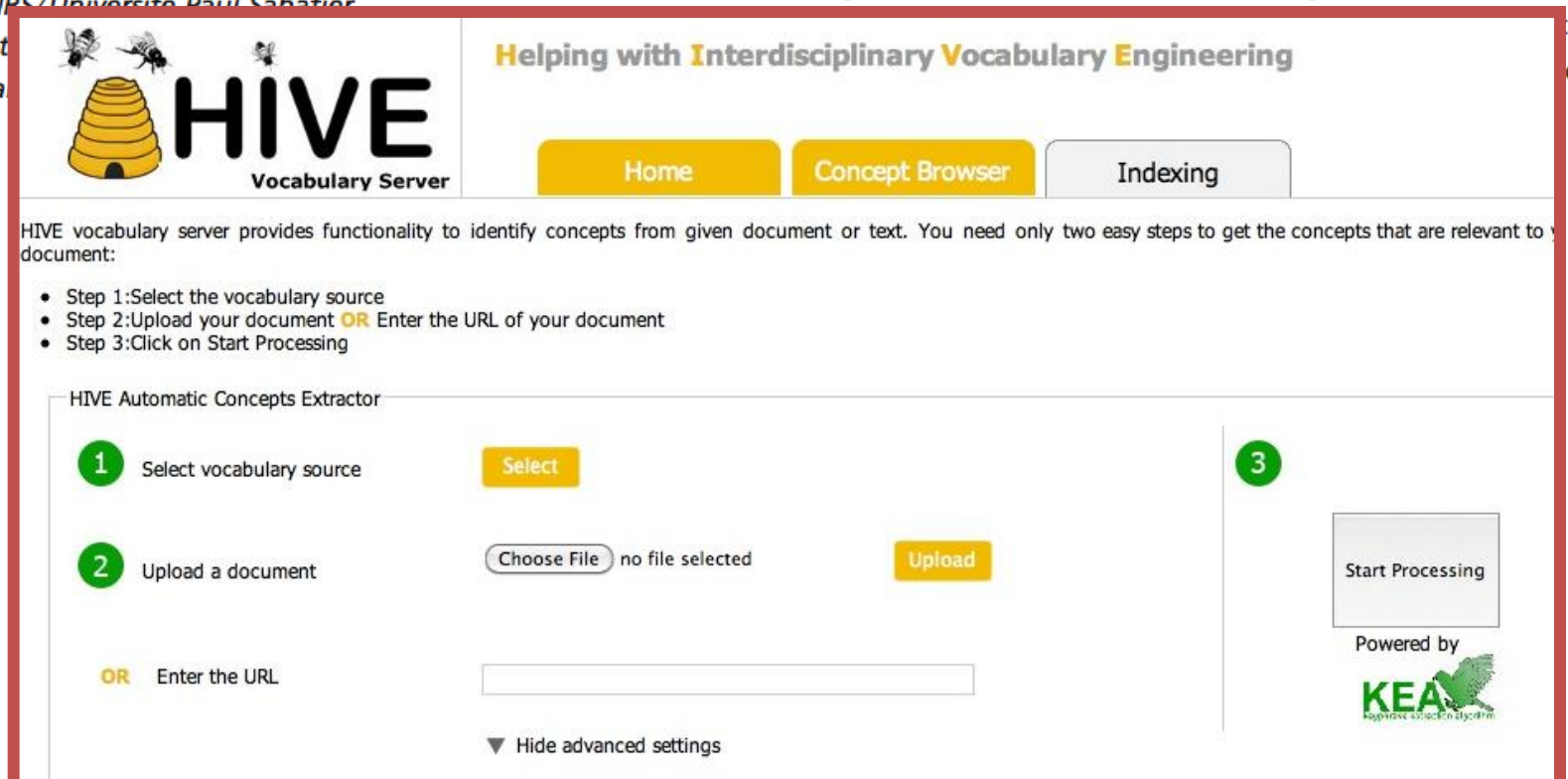
Towards a worldwide wood economics spectrum

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 Bât Fra*

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 wood omical



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

Pyenson N, Goldbogen J, Shadwick R (2012) Data from: Mandible allometry in extant and fossil Balaenopteridae (Cetacea: Mammalia): the largest vertebrate skeletal element and its role in rorqual lunge-feeding. Dryad Digital Repository. doi:10.5061/dryad.bt739.

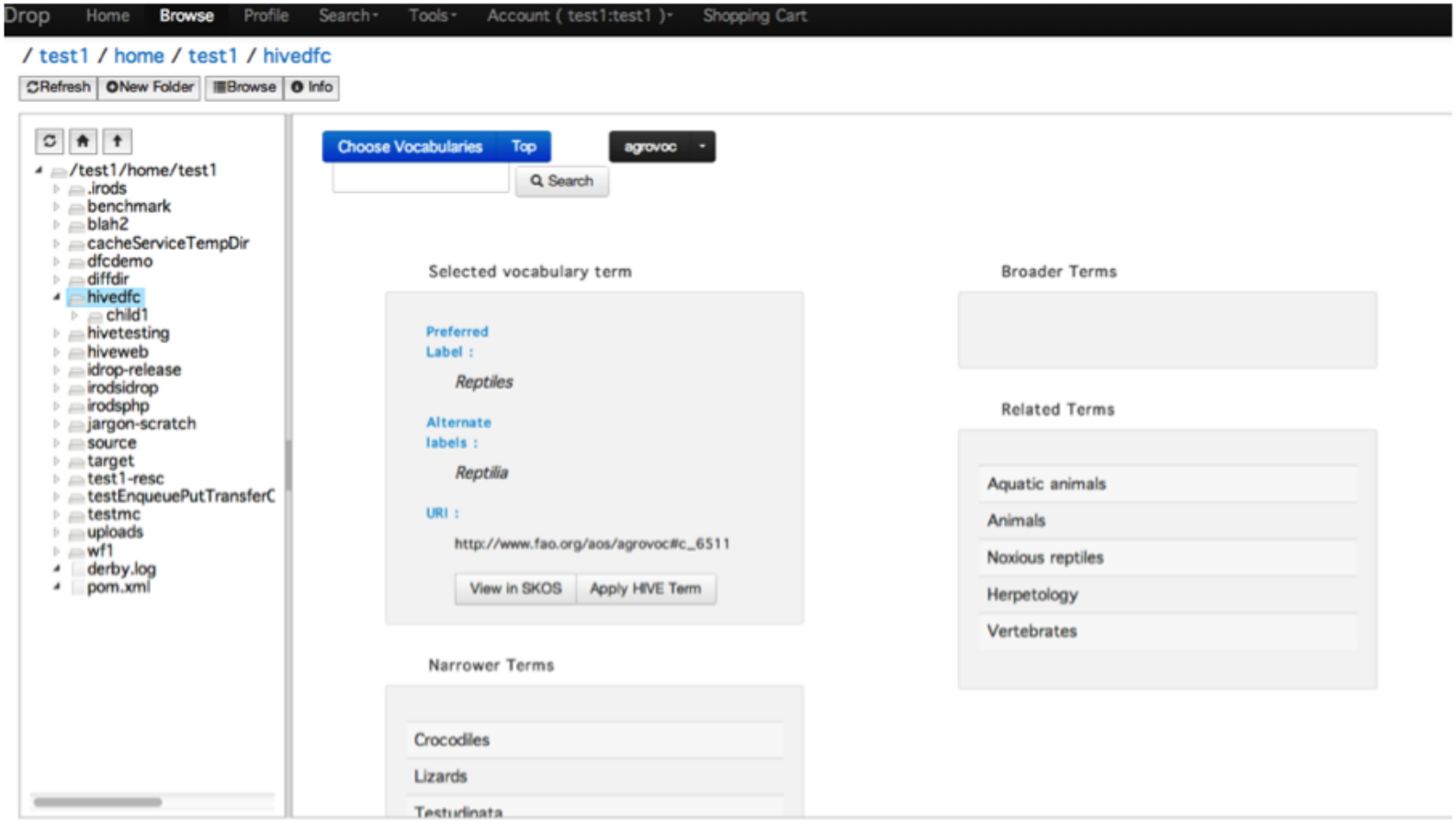
View in multiple formats

DFC DataNet FEDERATION CONSORTIUM

HIVE in iRODS for DFC

- Searchable
- Navigable
- Easy to integrate

```
</skos:ConceptScheme>
 <skos:Concept
rdf:about="http://www.fao.org/aos/agrovoc#c_3">
   <skos:prefLabel xml:lang="en">ABA</skos:prefLabel>
   <skos:altLabel xml:lang="en">Abscisic acid</skos:altLabel>
   <skos:broader
rdf:resource="http://www.fao.org/aos/agrovoc#c_3397"/>
   <skos:broader
rdf:resource="http://www.fao.org/aos/agrovoc#c_32543"/>
 </skos:Concept>
```



Questions

1. In your community, how is the support for a semantic registry?
   - Support depends on agencies to publish terminologies in LOD
   - Support/registry construction can be dynamic, on-the-fly, depending on how often a HIVE server is updated
2. How open is the registry, who can add or modify things, are the semantics stable?
   - N/A
3. Does the data model of the registry match your needs (information entry, relationships, granularity, possible inconsistency, …)
   - Matching needs in DFC, potentially DataONE, used in LTER, Dryad prototype, others…
4. How is the sustainability of the registry: financially, organizational, … ?
   - IMLS, NSF supplements, dependency on agencies

Conclusions, Feedback welcome!

- Complementary and alternative approaches
  - *More than one way to skin a cat…*
- YAMZ - Next steps…populate, test, engage
  - *User profiles*
  - YAMZ terms to EZID identifiers (ARKs)
- HIVE: https://code.google.com/p/hive-mrc/ moving to github, RENCI demo., testing.

Acknowledgments

- DataONE PAMWG: **John Kunze, Angela Murillo, **Christopher Patton, Sarah Callaghan, Rob Guralnick, Vivian Hutchinson, Greg Janee, Nassib Nassar, Karthik Ram, Tim Robertson
- DataONE HIVE: Bo Yao, Luke Williamson
- DataNet Federation Consortium: (Reagan Moore, **Mike Conway, Le Zhan, Mary Whitton)
- SILS-UNC-CH <Metadata Research Center>
- NESCent/Dryad: Ryan Scherle

DataONE (Grant #OCI-0830944)

DFC DataNet FEDERATION CONSORTIUM (Grant #OCI-0940841).

HIVE Helping Interdisciplinary Vocabulary Engineering

INSTITUTE of Museum and Library SERVICES

NSF

UNC SCHOOL OF INFORMATION AND LIBRARY SCIENCE
Metadata Research Center <MRC>

DRYAD

National Evolutionary Synthesis Center NESCent