ONEMercury: Towards Automatic Annotation of Environmental Science Metadata

Suppawong Tuarob
Pennsylvania State University
DataONE Summer Intern 2012

Jeff Horsburgh
Utah State University
Integration and Semantics Working Group, WG co-lead

Natasha Noy
Stanford University
Integration and Semantics Working Group

Line Pouchard
Scientific Data Group
Oak Ridge National Laboratory
Integration and Semantics Working Group

Giri Palanisamy
Oak Ridge National Laboratory
DataONE Cyber-Infrastructure

DataONE

DataONE is Foremost a Federation

Member Nodes (MNs)

- Heart of the federation
- Harness the power of local curation

Coordinating Nodes (CNs)

- Services to link Member Nodes

Investigator Toolkit (ITK)

- Tools for the whole data lifecycle

Deployed Infrastructure

Coordinating Nodes

- ORC: ORNL + UTK
- UCSB
- UNM

Investigator Toolkit

- Java and Python libs
- Command line
- ONEMercury
- (R plugin)
- (Morpho)
- (ONEDrive)

Member Nodes

Production

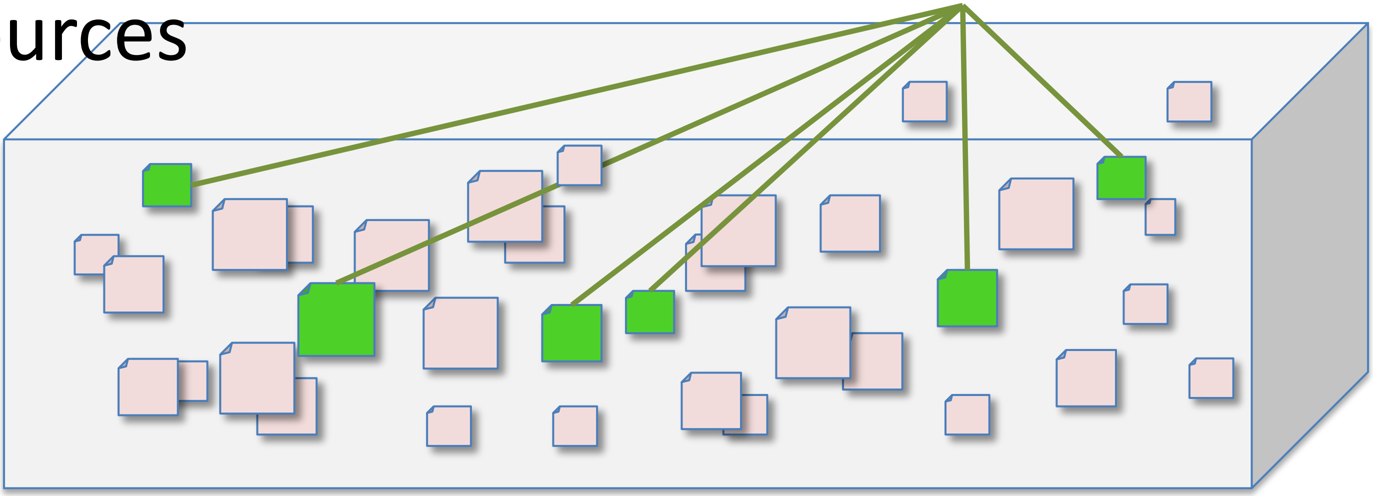
- KNB
- ESA
- SANParks
- USGS CSAS
- ORNL DAAC
- LTER
- CDL
- PISCO

Testing

- AKN
- TFRI
- Dryad
- EDAC
- KUBI

Discovery Services

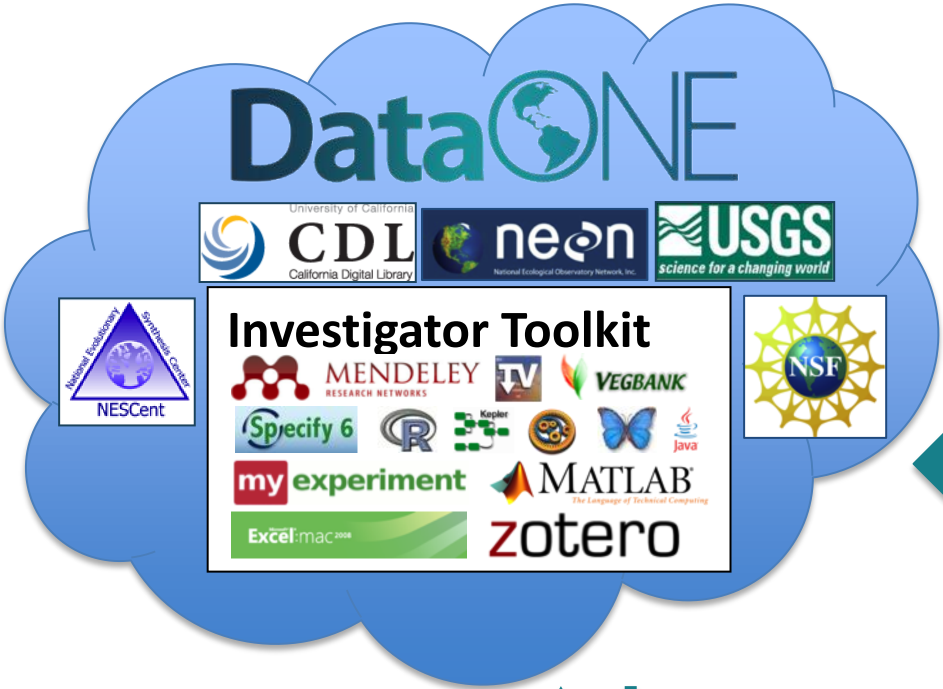
- Find the list of objects relevant to *some query*
- *Example: “Find data related to soil organic carbon concentration”*
  - Manual query: entered by user
  - Automated query: derived by navigation
  - Determined by client operation context
- Build a common index from heterogeneous sources



ORNL DAAC as a DataONE Member Node

NASA collectors

DAAC Users (UWG)

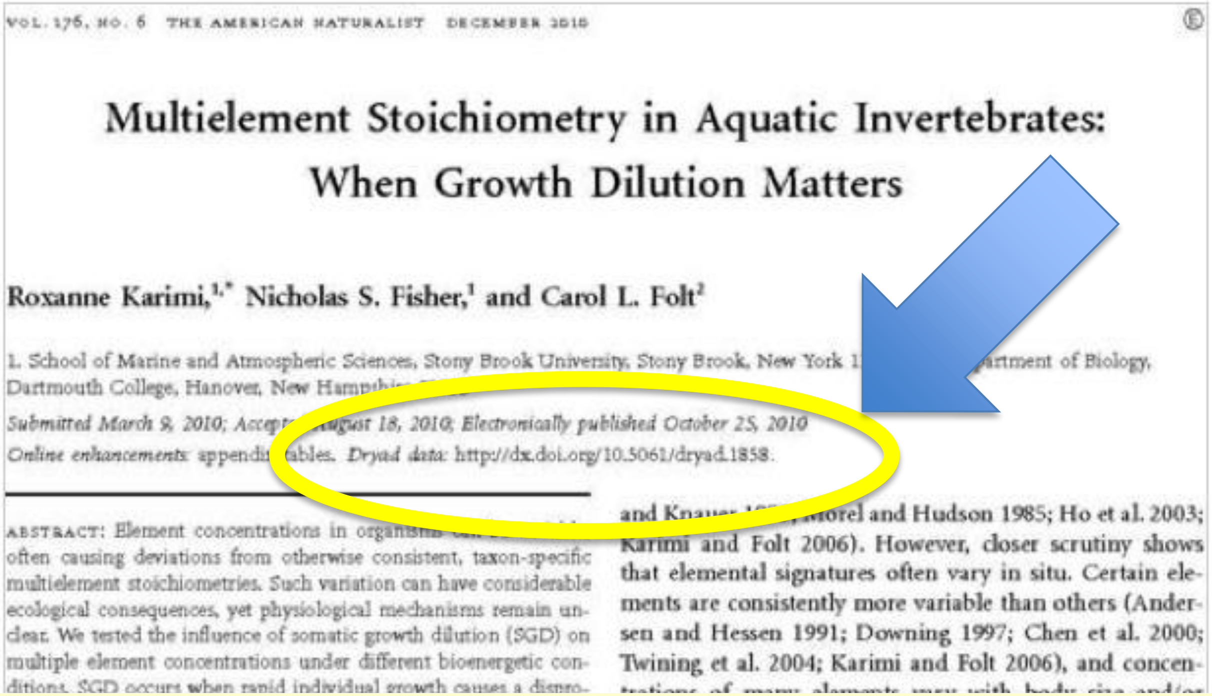


DataONE
University of California CDL California Digital Library
neon National Ecological Observatory Network, Inc.
USGS science for a changing world
National Evolutionary Synthesis Center NESCent
NSF
MENDELEY RESEARCH NETWORKS
TV
VEGBANK
Specify 6
R
Kepler
Java
my experiment
MATLAB The Language of Technical Computing
Excel:mac 2008
zotero

Investigator Toolkit

DataONE Users

Provide credit for data publication



VOL. 176, NO. 6 THE AMERICAN NATURALIST DECEMBER 2010

Multielement Stoichiometry in Aquatic Invertebrates: When Growth Dilution Matters

Roxanne Karimi,[1,*] Nicholas S. Fisher,[1] and Carol L. Folt[2]

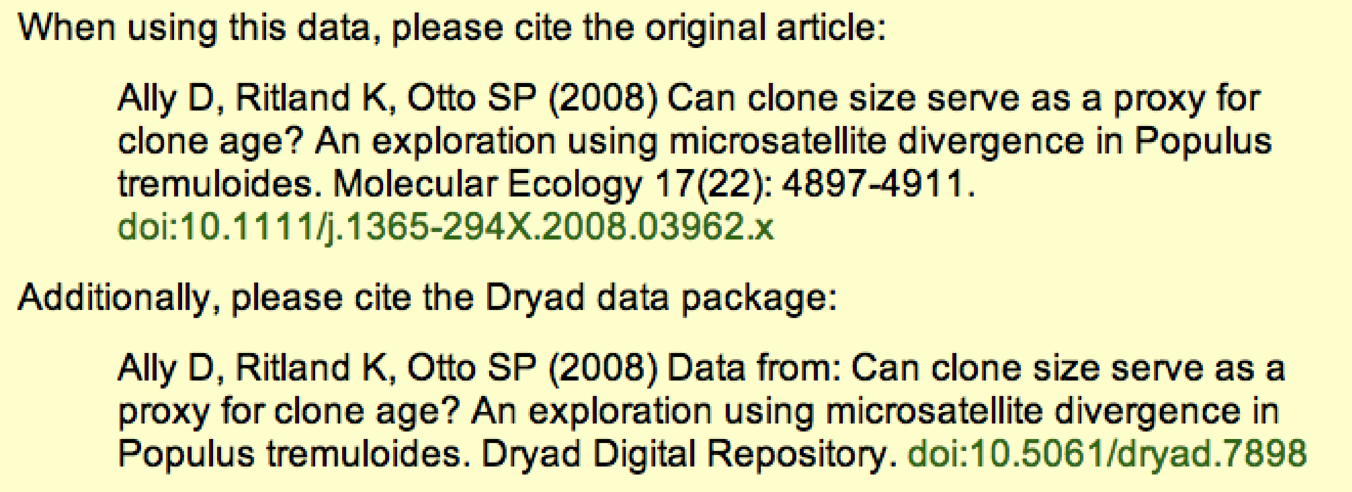
1. School of Marine and Atmospheric Sciences, Stony Brook University, Stony Brook, New York 1[illegible]artment of Biology, Dartmouth College, Hanover, New Hampshi[illegible]

Submitted March 9, 2010; Accept[illegible]ugust 18, 2010; Electronically published October 25, 2010

Online enhancements: appendi[illegible]ables. *Dryad data:* http://dx.doi.org/10.5061/dryad.1858.

ABSTRACT: Element concentrations in organisms [illegible] often causing deviations from otherwise consistent, taxon-specific multielement stoichiometries. Such variation can have considerable ecological consequences, yet physiological mechanisms remain unclear. We tested the influence of somatic growth dilution (SGD) on multiple element concentrations under different bioenergetic conditions. SGD occurs when rapid individual growth causes a dispro-

and Knauer 19[illegible]; Morel and Hudson 1985; Ho et al. 2003; Karimi and Folt 2006). However, closer scrutiny shows that elemental signatures often vary in situ. Certain elements are consistently more variable than others (Andersen and Hessen 1991; Downing 1997; Chen et al. 2000; Twining et al. 2004; Karimi and Folt 2006), and concentrations of many elements vary with body size and/or



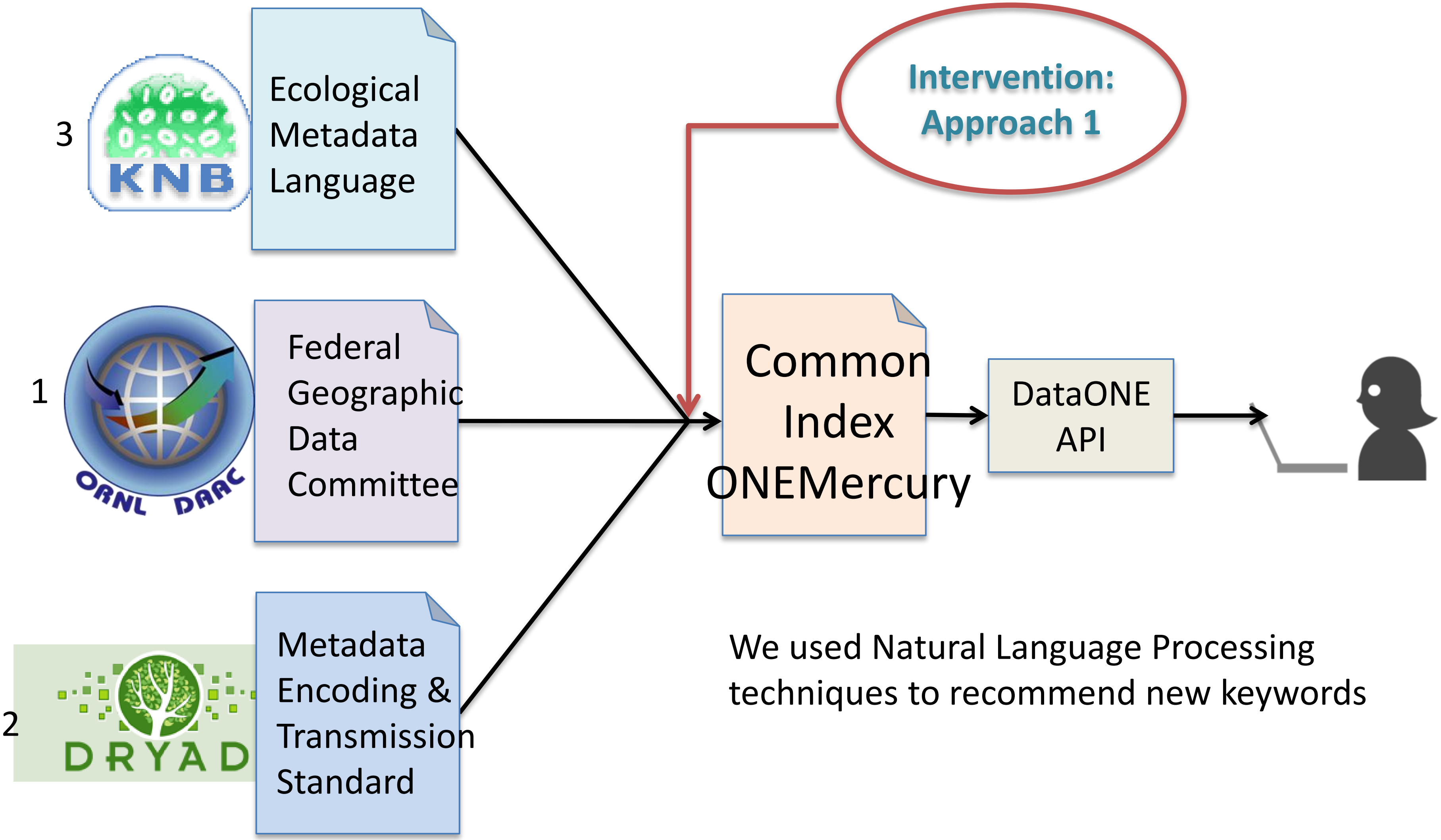
When using this data, please cite the original article:

Ally D, Ritland K, Otto SP (2008) Can clone size serve as a proxy for clone age? An exploration using microsatellite divergence in Populus tremuloides. Molecular Ecology 17(22): 4897-4911. doi:10.1111/j.1365-294X.2008.03962.x

Additionally, please cite the Dryad data package:

Ally D, Ritland K, Otto SP (2008) Data from: Can clone size serve as a proxy for clone age? An exploration using microsatellite divergence in Populus tremuloides. Dryad Digital Repository. doi:10.5061/dryad.7898

Automatic keyword recommendation



3

Ecological Metadata Language

1

Federal Geographic Data Committee

2

Metadata Encoding & Transmission Standard

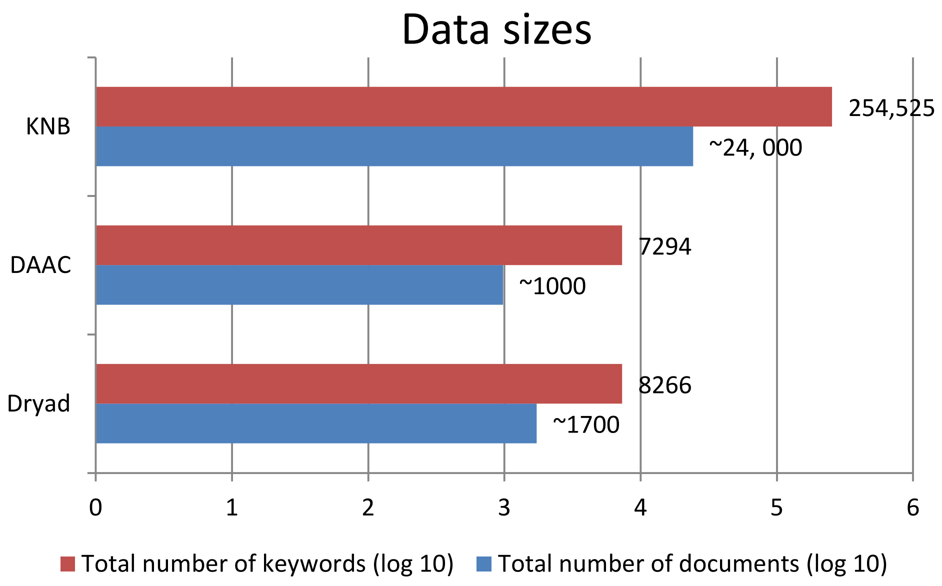
Intervention: Approach 1

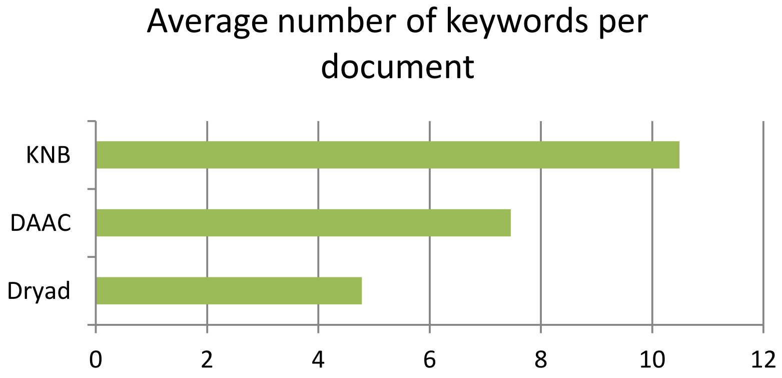
Common Index ONEMercury

DataONE API

We used Natural Language Processing techniques to recommend new keywords

Data sizes

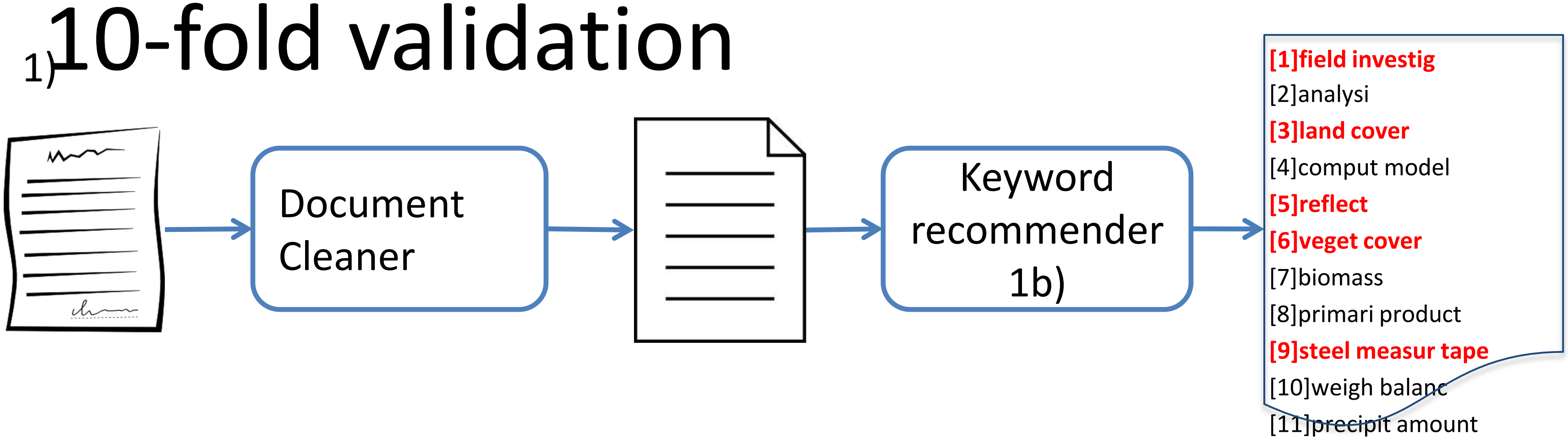




- Small, 175 MB ten-fold training set
- What matters:
  - Total number of documents
  - Total number of keywords
  - Number of topics in the training set
  - Evaluation method

Method details: Topic Model using Latent Dirichlet Allocation + 10-fold validation

1)



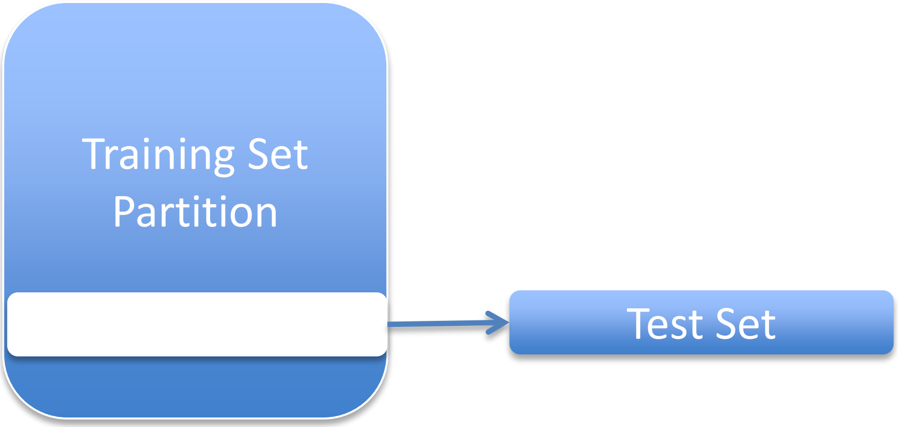
Document Cleaner

Keyword recommender 1b)

[1]field investig
[2]analysi
[3]land cover
[4]comput model
[5]reflect
[6]veget cover
[7]biomass
[8]primari product
[9]steel measur tape
[10]weigh balanc
[11]precipit amount

1b) Training set:
uses title, abstract, description, keywords

Test set:
removes keywords predict



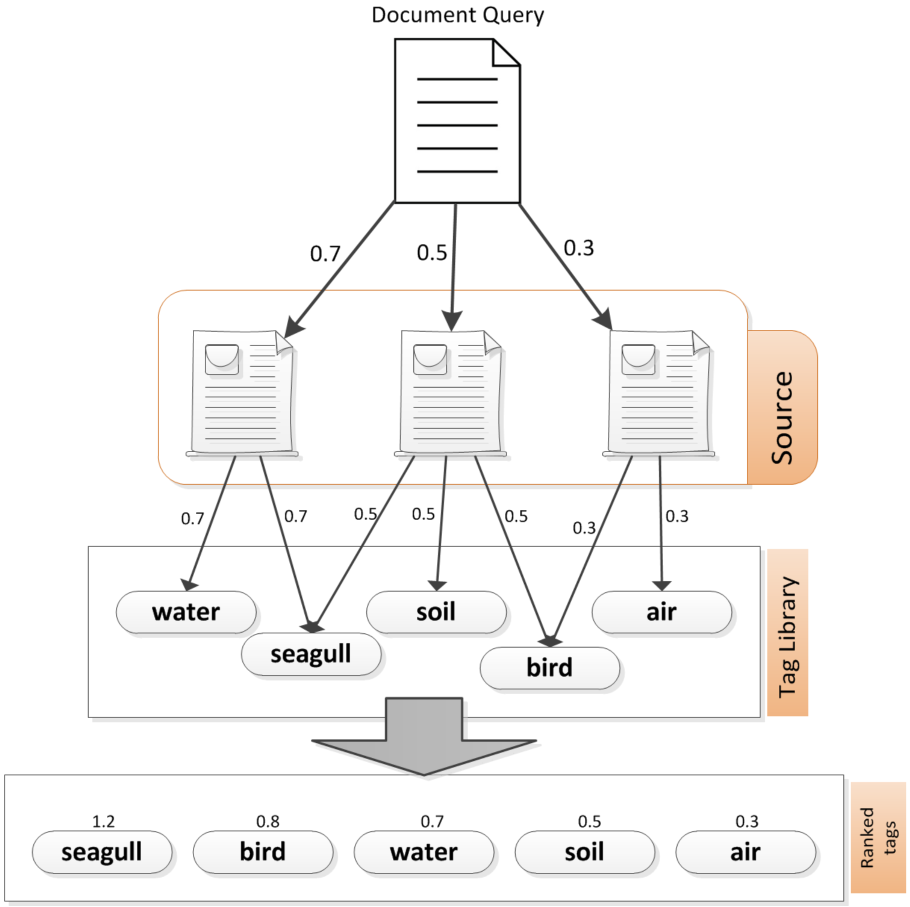
Training Set Partition

Test Set

Learn keywords

Remove keywords, and predict

System



Compute similarity scores between query and source documents.

Propagate the scores to tags.
If a tag receives multiple scores, sum the scores.

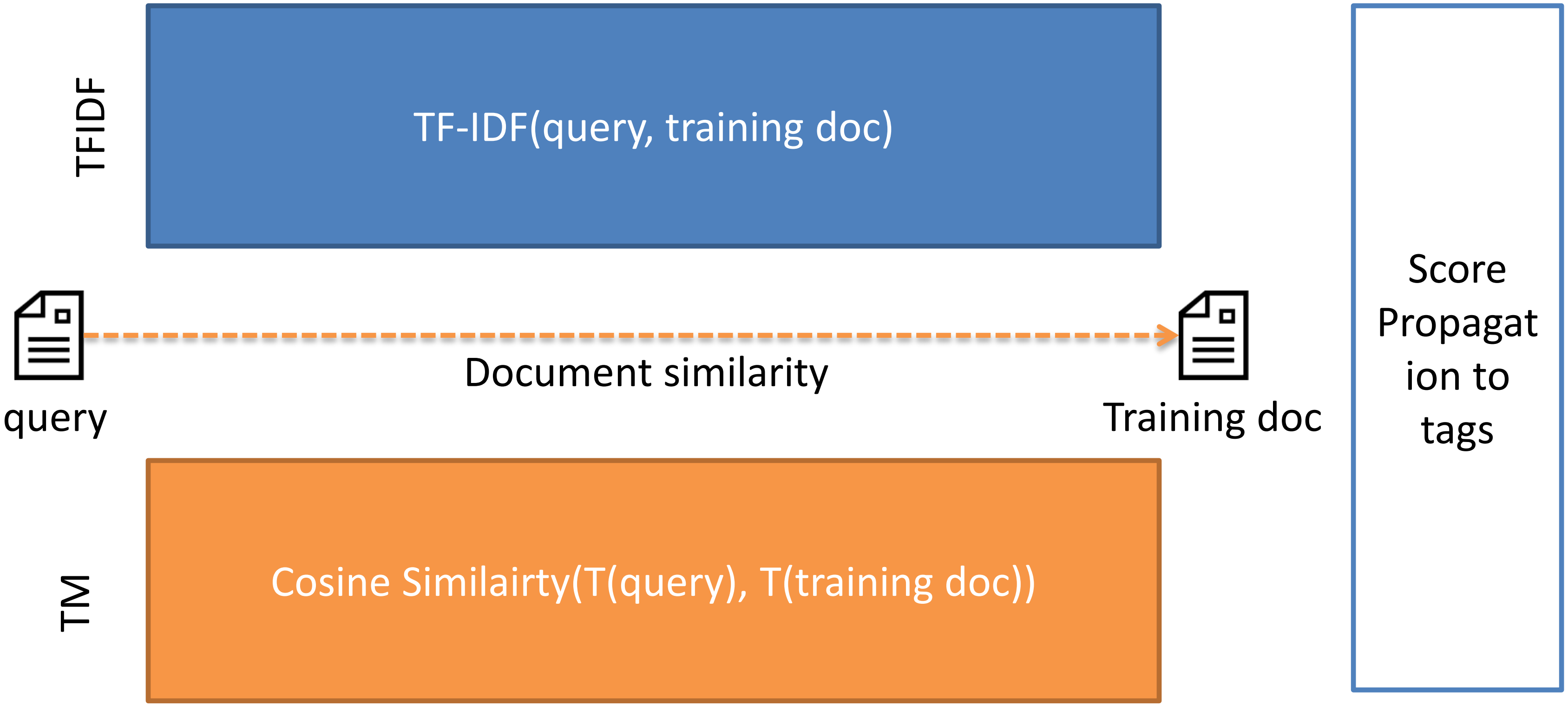
Rank and return top *K* tags.

Topic Model Based Approach (TM): Main Idea

- 1. Model topics from training corpus.
- 2. Calculate the topic similarity scores (using cosine similarity) between the query and each document in the training set.
- 3. Propagate the scores to the tags in each training document.

TFIDF vs TM

Basically, the similarity measures between the query and a training document



TFIDF

TF-IDF(query, training doc)

query

Document similarity

Training doc

Score Propagation to tags

TM

Cosine Similairty(T(query), T(training doc))

* T(d) is a topic distribution of document d, representing by a vector of probability values.

Computing T(d)

- Ex. 3 topics are modeled from the training corpus (We use *LDA* algorithm to model topics)
- T(d) = <t_1, t_2, t_3> where t_i is the probability that document *d* belongs to topic *i*.
- Note that $t_1 + t_2 + t_3 = 1.0$
- We use the inference algorithm proposed by Asuncion et al.* to find the topic distribution of a document.

* Asuncion, A., Welling, M., Smyth, P., & Teh, Y. W. (2009). On Smoothing and Inference for Topic Models. *Proceedings of the TwentyFifth Conference on Uncertainty in Artificial Intelligence*, *24*(Ml), 27–34. AUAI Press. Retrieved from http://discovery.ucl.ac.uk/150501/

Evaluation Metrics

- **Precision @ K**
- **Recall@ K**
- **Precision vs Recall**
- **F1 @ K**
- **MRR**
- **Bpref**
- **Training and Testing times**
- **Comparisons** between TFIDF and TM approaches (only self-recommendation) on each archive

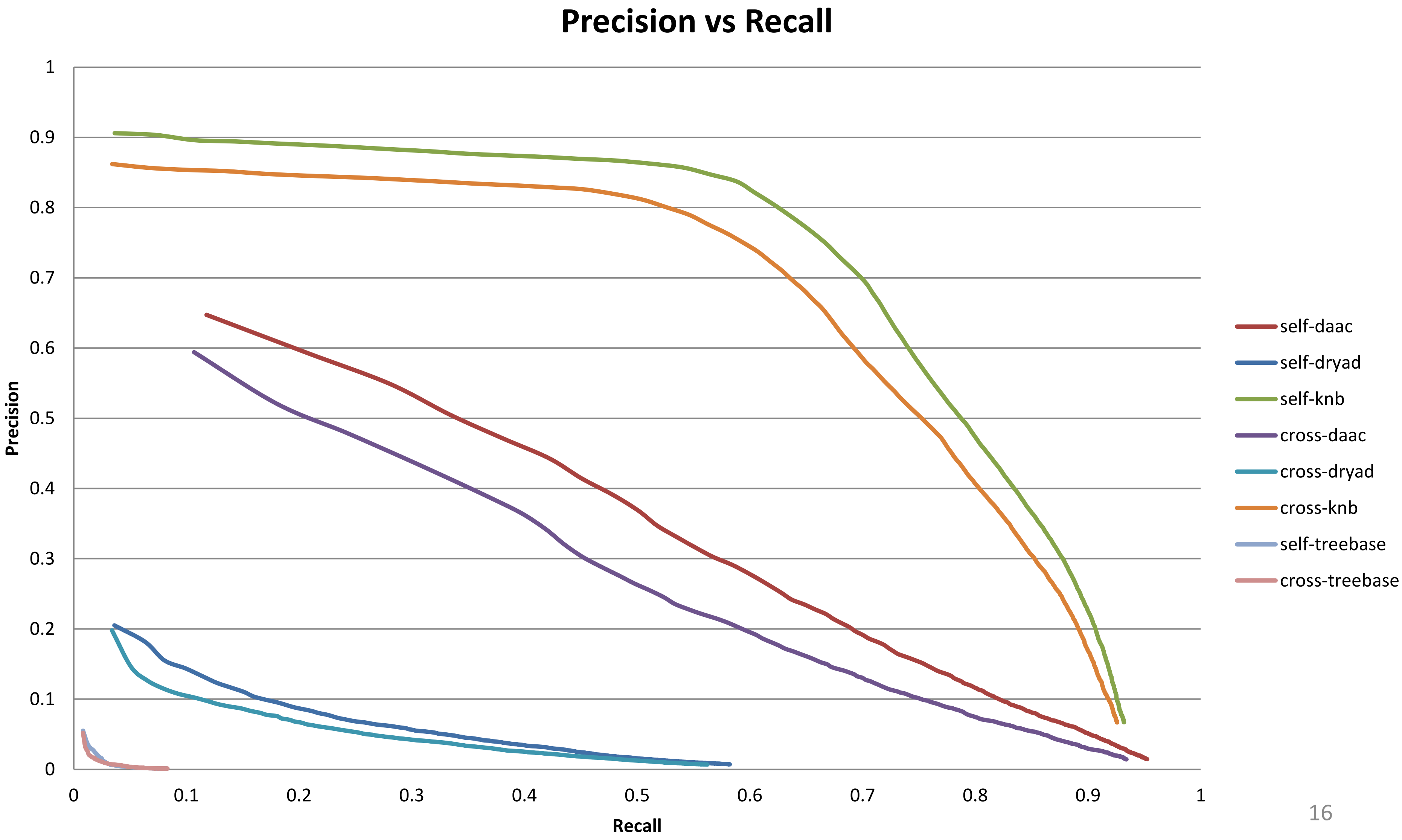
Precision and recall

	Actual Positive	Actual Negative
Predicted Positive	Tp	Fp
Predicted Negative	Fn	Tn

- Precision: $\frac{Tp}{Tp+Fp}$
- Recall: $\frac{Tp}{Tp+Fn}$

- Precision: the proportion of Actual positive in the recommended set
- Recall: the proportion of Actual that are recommended over the entire dataset

TM: Precision vs Recall



TFIDF vs TM: on Self-DAAC (record913.xml)

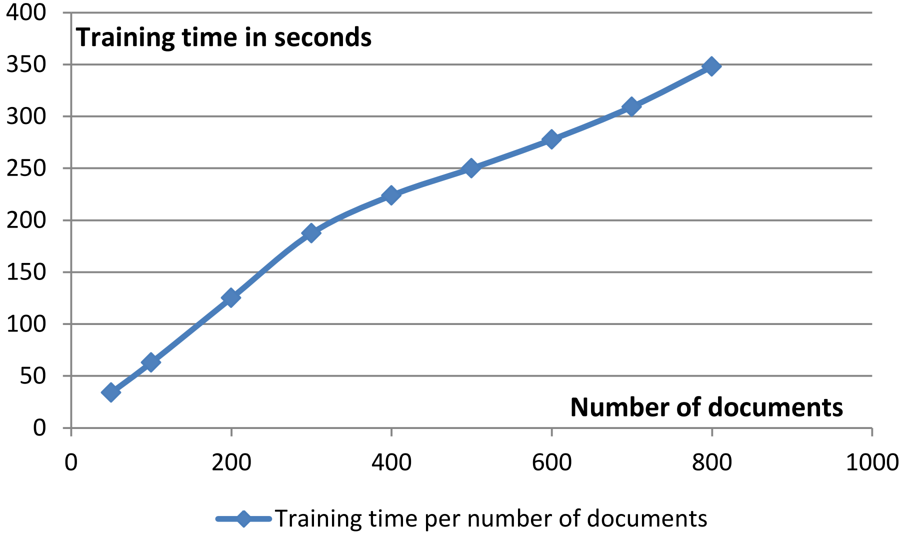
ISLSCP II IGBP DISCOVER AND SIB LAND COVER, 1992-1993

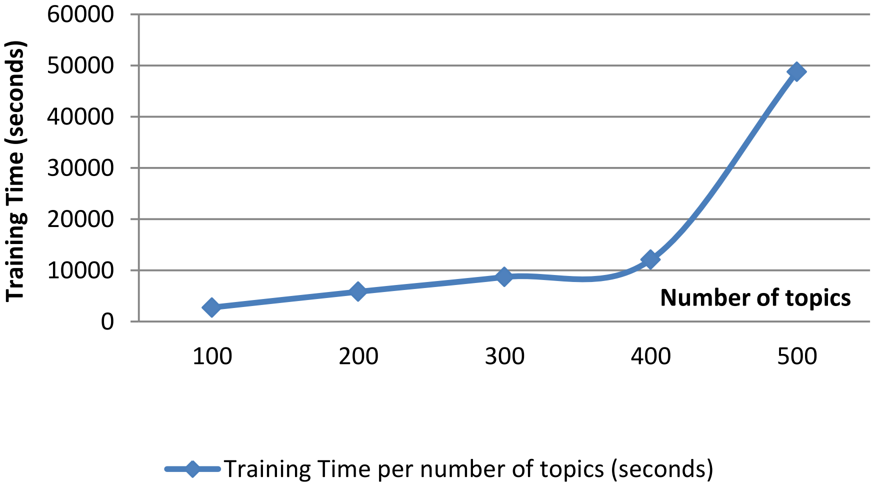
This data set describes the geographic distributions of 17 classes of land cover based on the International Geosphere-Biosphere DISCover land cover legend (Loveland and Belward 1997) and the 15 classes of the SiB model processed at the USGS EROS Data Center (EDC). Specifically, the resampled DISCover datasets were derived from the 1km DISCover data set compiled by the USGS. The 1km data sets for each classification scheme were aggregated to 1, 0.5 and 0.25 degree spatial resolutions for this ISLSCP II data collection. Each layer of the aggregated products corresponds to a single DISCover land cover category and the values represent the percentage of the coarse resolution cell (1 degree, etcÃ¢Ã¯Â¿Â½Â¦) occupied by that land cover category. The dominant class data show the land cover category that occupies the majority of the cell and is derived from the percentage files for each cover type. The objective of this study was to create a land cover map derived from 1 kilometer AVHRR data using a full year of data (April 1992-March 1993). This thematic map was resampled to 0.25, 0.5 and 1.0 degree grids for the International Satellite Land Surface Climatology Project (ISLSCP) data initiative II. During this re-processing, the original EDC land cover type and fraction maps were adjusted to match the water/land fraction of the ISLSCP II land/water mask. These maps were generated for use by global modelers and others. This data set is one of the products of the International Satellite Land-Surface Climatology Project, Initiative II (ISLSCP II) data collection which contains 50 global time series data sets for the ten-year period 1986 to 1995. Selected data sets span even longer periods. ISLSCP II is a consistent collection of data sets that were compiled from existing data sources and algorithms, and were designed to satisfy the needs of modelers and investigators of the global carbon, water and energy cycle. The data were acquired from a number of U.S. and international agencies, universities, and institutions. The global data sets were mapped at consistent spatial (1, 0.5 and 0.25 degrees) and temporal (monthly, with meteorological data at finer (e.g., 3-hour)) resolutions and reformatted into a common ASCII format. The data and documentation have undergone two peer reviews.ISLSCP is one of several projects of Global Energy and Water Cycle Experiment (GEWEX) [http://www.gewex.org/] and has the lead role in addressing land-atmosphere interactions -- process modeling, data retrieval algorithms, field experiment design and execution, and the development of global data sets.

TFIDF vs TM on Self-DAAC (record913.xml)

Actual	TFIDF	TM
[1]albedo [2]land cover [3]veget cover [4]veget index [5]leaf area meter [6]terra morn equatori cross time satellit [7]noaa nation ocean amp amp atmospher administr [8]plant characterist [9]steel measur tape [10]canopi characterist [11]modi moder resolut imag spectroradiomet [12]leaf characterist [13]avhrr advanc high resolut radiomet [14]field investig [15]reflect	[1]**field investig [2]analysi [3]**land cover [4]comput model [5]**reflect [6]**veget cover [7]biomass [8]primari product [9]**steel measur tape [10]weigh balanc [11]precipit amount [12]**canopi characterist [13]**leaf characterist [14]water vapor [15]quadrat sampl frame [16]rain gaug [17]surfac air temperatur [18]air temperatur [19]meteorolog station [20]human observ [21]**veget index [22]soil core devic [23]**plant characterist [24]surfac wind [25]**modi moder resolut imag spectroradiomet [26]**albedo [27]**terra morn equatori cross time satellit [28]**leaf area meter [29]**avhrr advanc high resolut radiomet [30]pyranomet [31]**noaa nation ocean amp amp atmospher administr [32]avhrr	[1]**land cover [2]**modi moder resolut imag spectroradiomet [3]**terra morn equatori cross time satellit [4]**field investig [5]**veget cover [6]**reflect [7]**veget index [8]leaf characterist [9]**canopi characterist [10]**plant characterist [11]**albedo [12]**steel measur tape [13]**avhrr advanc high resolut radiomet [14]**noaa nation ocean amp amp atmospher administr [15]**leaf area meter [16]analysi [17]comput model [18]noaa [19]avhrr [20]popul distribut [21]river stream [22]terrain elev [23]landsat [24]landsat tm [25]primari product [26]model analysi [27]photosynthet activ radiat [28]topograph effect [29]digit elev model [30]agricultur land [31]digit [32]biomass

Scalability issues





- Training time increases linearly with increasing number of documents and topics
- Above 400 topics, a memory issue appears

Questions?