— DATA UNIFICATION AND DISAMBIGUATION

LINGE BAI
11/15/2016

Clarivate Analytics | Formerly the IP & Science business of Thomson Reuters

ENTITY RESOLUTION

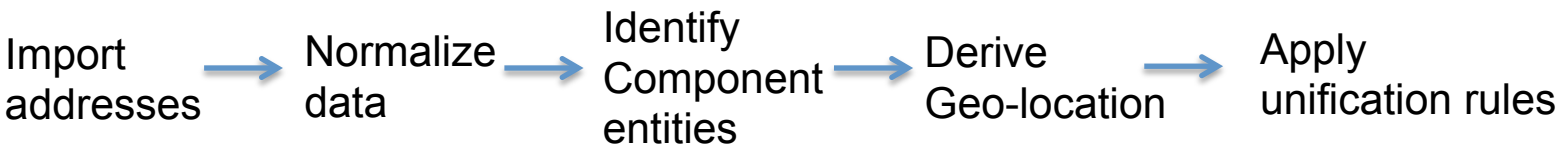
- Academic Institutions
  - Quantity: tens of thousands worldwide
  - Name and address variance
  - Hierarchy, merging and splitting

- Scholarly Authors
  - Quantity: hundreds of millions of name occurrences
  - Same name, different people
  - Same person, different names

INSTITUTION UNIFICATION

To establish accurate publications for institutions and institutional hierarchies.

- Rule based system driven by domain knowledge
- Web application to integrate computational processing and domain expert inputs



INSTITUTION UNIFICATION

- Normalize data: to parse and normalize captured addresses

Published Address:
Department of Biology, Indiana University, Bloomington, IN 47405 USA

WoS Editorial Capture:
Indiana Univ, Dept Biol, Bloomington, IN 47405 USA

WAAN:
INDIANA UNIV, DEPT BIOL, BLOOMINGTON, IN 47405 USA

- Identify component entities: to tokenize address components

INDIANA UNIV, DEPT BIOL, BLOOMINGTON, IN 47405 USA
- Component 1: INDIANA UNIV
- Component 2: DEPT BIOL
- Component 3: BLOOMINGTON
- Component 4: IN 47405 USA

- Derive Geographical Location: to utilize location levels and hierarchy

INDIANA UNIV, DEPT BIOL, BLOOMINGTON, IN 47405 USA

City: BLOOMINGTON
State: INDIANA
Country: USA

INSTITUTION UNIFICATION

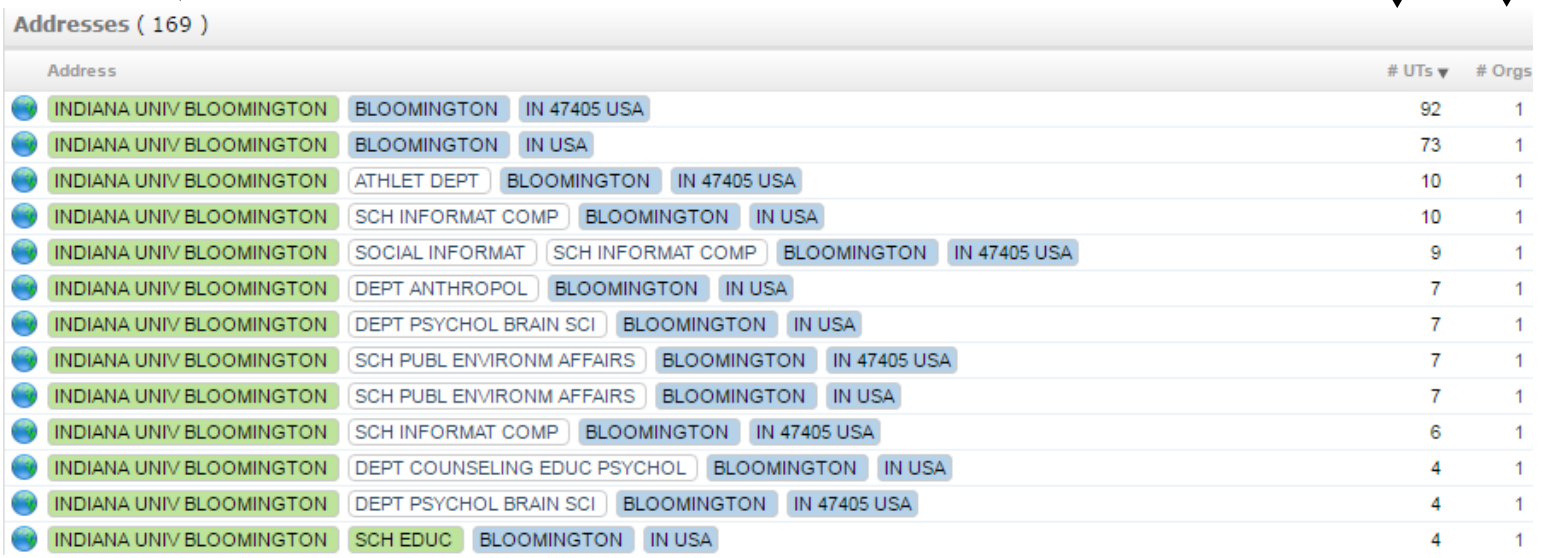
Apply unification rules: to attribute address to one or more organizations.

Organization components

Locations

Items

Orgs



Addresses (169)

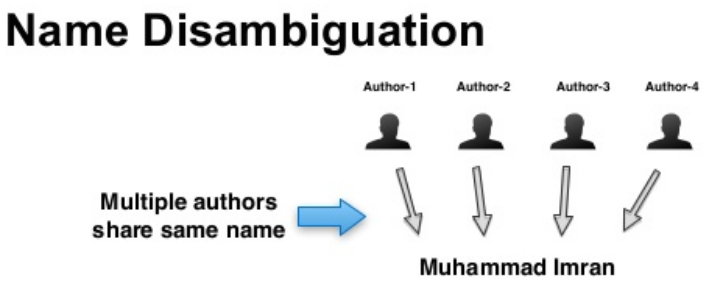
Address	# UTs	# Orgs
INDIANA UNIV BLOOMINGTON BLOOMINGTON IN 47405 USA	92	1
INDIANA UNIV BLOOMINGTON BLOOMINGTON IN USA	73	1
INDIANA UNIV BLOOMINGTON ATHLET DEPT BLOOMINGTON IN 47405 USA	10	1
INDIANA UNIV BLOOMINGTON SCH INFORMAT COMP BLOOMINGTON IN USA	10	1
INDIANA UNIV BLOOMINGTON SOCIAL INFORMAT SCH INFORMAT COMP BLOOMINGTON IN 47405 USA	9	1
INDIANA UNIV BLOOMINGTON DEPT ANTHROPOL BLOOMINGTON IN USA	7	1
INDIANA UNIV BLOOMINGTON DEPT PSYCHOL BRAIN SCI BLOOMINGTON IN USA	7	1
INDIANA UNIV BLOOMINGTON SCH PUBL ENVIRONM AFFAIRS BLOOMINGTON IN 47405 USA	7	1
INDIANA UNIV BLOOMINGTON SCH PUBL ENVIRONM AFFAIRS BLOOMINGTON IN USA	7	1
INDIANA UNIV BLOOMINGTON SCH INFORMAT COMP BLOOMINGTON IN 47405 USA	6	1
INDIANA UNIV BLOOMINGTON DEPT COUNSELING EDUC PSYCHOL BLOOMINGTON IN USA	4	1
INDIANA UNIV BLOOMINGTON DEPT PSYCHOL BRAIN SCI BLOOMINGTON IN 47405 USA	4	1
INDIANA UNIV BLOOMINGTON SCH EDUC BLOOMINGTON IN USA	4	1

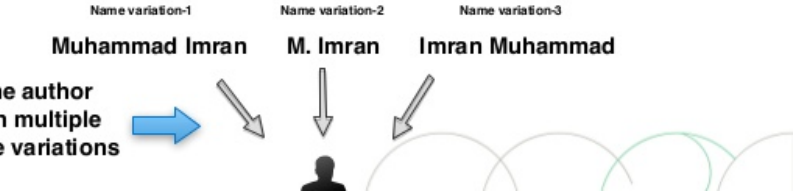
Addresses are unified to an organization

AUTHOR DISAMBIGUATION

"Of the more than 6 million authors in a major journal citations and abstracts database, **more than two-thirds of them share a last name and single initial with another author**, and an ambiguous name in the same database refers on average to eight people."

Name ambiguity is a frequently encountered problem in the scholarly community:





Also published as:

- Avram Noam Chomsky
- N. Chomsky
- نعوم تشومسكي
- נועם חומסקי

AUTHOR DISAMBIGUATION

THREE TIERED APPROACH

Machine learning

- Updating algorithms to automatically disambiguate author names with high precision and recall.*
- Improved author clusters algorithmically.

Learn from multiple data sources

- Identifying trusted sources for author data.
- Multiple data sources – Internal and External.
- Improving disambiguation by learning from external trusted data sources.

User Feedback

- Capability to accept, store and apply customer feedback.
- Improved author clusters with user feedback.

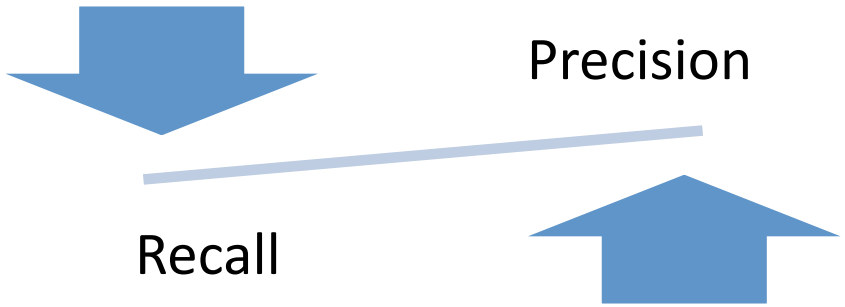
ORCID

RESEARCHERID

***Levin, M., Krawczyk, S., Bethard, S. and Jurafsky, D. (2012), Citation-based bootstrapping for large-scale author disambiguation. J Am Soc Inf Sci Tec, 63: 1030–1047. doi:10.1002/asi.22621**

BALANCING ACT

- Although customers perceive author clustering as data, in fact it is the result of programs that evaluate pairs for linking
- If the programs are tuned for precision (reduce false-positive links) then some links that should be made are not.
- But if the programs are not tuned for precision, we see false positives – “clumping”.



Comparing methods for partitioning a decade of carbon dioxide and water vapor fluxes in a temperate forest

By: Sulman, BN (Sulman, Benjamin N.)[1,2,3]; Roman, DT (Roman, D. Tyler)[1,4]; Scanlon, TM (Scanlon, Todd M.)[5]; Wang, LX (Wang, Lixin)[6]; Novick, KA (Novick, Kimberly A.)[1]

AGRICULTURAL AND FOREST METEOROLOGY
Volume: 226 **Pages:** 229-245
DOI: 10.1016/j.agrformet.2016.06.002
Published: OCT 15 2016
View Journal Information

Abstract

The eddy covariance (EC) method is routinely used to measure net ecosystem fluxes of carbon dioxide (CO2) and evapotranspiration (ET) in terrestrial ecosystems. It is ofte ... and transpiration. We ... record of ET and CO ... the FVS method has ... partitioned using non ... canopy potential eva ... parameterize a mode ... independent estimate ... with scaled leaf gas e ... estimates of evapora ... the canopy but above ... sensitive to errors in ... necessary for the commonly used nonlinear regression technique. (C) 2016 Elsevier B.V. All rights reserved.

Keywords

Author Keywords: CO2 flux; Ecohydrology; Eddy covariance; Evapotranspiration; Flux partitioning; Water use efficiency
KeyWords Plus: NET ECOSYSTEM EXCHANGE; EDDY-COVARIANCE MEASUREMENTS; NORTHERN HARDWOOD FOREST; DECIDUOUS FOREST; USE EFFICIENCY; ENERGY FLUXES; UNITED-STATES; GAS-EXCHANGE; SAP FLOW; CO2

Author Information

Reprint Address: Sulman, BN (reprint author)

+ Princeton Univ, Dept Geosci, Program Atmospher & Ocean Sci, 300 Forrestal Rd, Princeton, NJ 08544 USA.

Addresses:

− [1] Indiana Univ, Sch Publ & Environm Affairs, 702 N Walnut Grove Ave, Bloomington, IN 47405 USA
Organization-Enhanced Name(s)
Indiana University Bloomington
Indiana University System

− [2] Indiana Univ, Dept Biol, 1001 E 3rd St, Bloomington, IN 47405 USA
Organization-Enhanced Name(s)
Indiana University Bloomington
Indiana University System

+ [3] Princeton Univ, Dept Geosci, Program Atmospher & Ocean Sci, 300 Forrestal Rd, Princeton, NJ 08544 USA

+ [4] Forest Serv, USDA, Northern Res Stn, 1831 Hwy 169 E, Grand Rapids, MN 55744 USA

+ [5] Univ Virginia, Dept Environm Sci, 291 McCormick Rd, Charlottesville, VA 22904 USA

+ [6] IUPUI, Dept Earth Sci, 723 W Michigan St, Indianapolis, IN 46202 USA

E-mail Addresses: bsulman@princeton.edu

Citation Network

Clarivate
Analytics

Formerly the IP & Science
business of Thomson Reuters