Plan for Today's Lecture(s)

- The Dublin Core
- Metacrap
- Tagging
- Authority Control / Duplicate Detection

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

26 September 2013
Lecture 9.1 – The Dublin Core

Dublin Core: Motivation

- When the web was just a few years old, professional "metadata-makers" were already concerned that web resources would be difficult to discover and identify because they lacked good descriptions
- An international, cross-disciplinary group of professionals from librarianship, computer science, text encoding, the museum community, and other related fields convened in 1995 to propose a standard set of metadata elements for web pages

Dublin Core: Goals

- **Simplicity**: Must be usable by web page authors, because web scale defies description by professional metadata-makers
- **Broad semantics**: Description elements must not require precise semantic distinctions that not everyone could or would want to make
- **Extensibility**: Must be easy to extend when more precise semantic descriptions are needed
- **International**: Element names need equivalents in all the languages used in web pages

Dublin Core

- Named after the place where it was created (but not where you think)
- Widely used as the basis for more specialized models (because it is the "core" of all of them)
  - See http://dublincore.org/documents/profile-guidelines/
  - Example: PBCORE (used by PopUpArchive)
- There are specifications of how to use it in numerous syntaxes (especially XML and RDF) and languages

The Dublin Core

Dublin Core® Metadata Initiative

Making it easier to find information.

Home | Metadata Basics | DCMI Specifications | Community and Events | Join/Support | About Us

Enter keyword Search

Dublin Core Metadata Element Set, Version 1.1

Identifier: http://dublincore.org/documents/2012/06/14/dces/

Replaces: http://dublincore.org/documents/2010/10/11/dces/

Latest version: http://dublincore.org/documents/dces/

Date Issued: 2012-06-14

Status of document: This is a DCMI Recommendation.

Description of document: This document provides ready reference for the Dublin Core Metadata Element Set, Version 1.1. For more detailed documentation and links to historical versioning information, see the document "DCMI Metadata Terms".

Introduction

The Dublin Core Metadata Element Set is a vocabulary of fifteen properties for use in resource description. The name "Dublin" is due to its origin at a 1995 invitational workshop in Dublin, Ohio; "core" because its elements are broad and generic, usable for describing a wide range of resources.

The fifteen element "Dublin Core" described in this standard is part of a larger set of metadata vocabularies and technical specifications maintained by the Dublin Core Metadata Initiative (DCMI). The full set of vocabularies, DCMI Metadata Terms [DCMI-TERMS], also includes sets of resource classes (including the DCMI Type Vocabulary [DCMI-TYPE]), vocabulary encoding schemes, and syntax encoding schemes. The terms in DCMI vocabularies are intended to be used in combination with terms from other, compatible vocabularies in the context of application profiles and on the basis of the DCMI Abstract Model [DCAM].

More than a Controlled Vocabulary

- A controlled vocabulary is a standardized set of terms (such as subject headings, names, classifications, etc.) assigned by organizers / cataloguers / indexers of resources
- A metadata schema like the Dublin Core controls the kinds of assertions about resources that you can make in the first place
- Controlled vocabularies can be very useful requirements or recommendations about the values that are contained in the assertions (the information content of the assertion)

Dublin Core: The Elements

- TITLE -- a name given to the resource
- IDENTIFIER -- an unambiguous reference to the resource within a given context
- SUBJECT -- the topic of the resource; key words or classification phrases
- CREATOR -- an entity primarily responsible for making the resource

Dublin Core: The Elements

- CONTRIBUTOR -- An entity responsible for making contributions to the resource
- PUBLISHER -- the entity primarily responsible for making the resource available
- DATE -- a point or period of time associated with an event in the life cycle of the resource
- FORMAT -- the file format, physical medium, or dimensions of the resource.

Dublin Core: The Elements

- DESCRIPTION -- an account of the resource; abstract, TOC, etc.
- LANGUAGE -- a language of the resource
- TYPE -- the nature or genre of the content of the resource
- RIGHTS -- information about rights held in and over the resource

Dublin Core: The Elements

- SOURCE – a related resource from which the described resource is derived
- RELATION -- a related resource
- COVERAGE -- the spatial or temporal topic of the resource, relevant jurisdiction

Dublin Core "Content Model" is Unrestrictive

- All elements are optional
- All elements are repeatable
- Elements can be in any order

The model for each element has gotten more general over time… e.g., used to say "content of a resource"… where it now says "resource"

Hypothetical Metadata Description for 202 Course Home Page

<meta name="author" content="Robert J. Glushko">

<meta name="title" content="I202 Information Organization and Retrieval">

<meta name="webmaster" content="Fred Chasen">

<meta name="publisher" content="UC Berkeley School of Information">

<meta name="date" content="Fall 2013">

“DublinCore-ized” Hypothetical Example for 202 Course Home Page

<meta name=“DC.creator” content=“Robert J. Glushko”>

<meta name=“DC.title” content=“I202 Information Organization and Retrieval”>

<meta name=“DC.contributor” content=“Fred Chasen”>

<meta name=“DC.publisher” content=“UC Berkeley School of Information”>

<meta name=“DC.date” content=“Fall 2013”>

Dublin Core in XML

- We can define the DC as an XML schema
- This allows DC statements to be embedded in any XML document using a DC namespace

Dublin Core in XML as Separate Record

<?xml version="1.0"?>
<metadata xmlns:dc=http://purl.org/dc/elements/1.1/>

<dc:creator>Robert J. Glushko</dc:creator>
<dc:title>I202 Information Organization and Retrieval </dc:title>
<dc:contributor>Fred Chasen</dc:contributor>
<dc:publisher>UC Berkeley School of Information</dc:publisher>
<dc:date>Fall 2013</dc:date>

</metadata>

Using the Dublin Core – Pragmatics and Problems

- Some information may appear to belong in more than one metadata element“
- There is potential semantic overlap between some elements
- There will occasionally be some judgment required from the person assigning the metadata

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

26 September 2013
Lecture 9.2– Metacrap

Cory Doctorow

- Cory Efram Doctorow (born July 17, 1971) is a Canadian-British blogger, journalist, and science fiction author who serves as co-editor of the weblog Boing Boing.
- He is an activist in favour of liberalising copyright laws and a proponent of the Creative Commons organization, using some of their licenses for his books.

(from Wikipedia)

Metacrap

- "A world of exhaustive, reliable metadata would be a utopia. It's also a pipe-dream, founded on self-delusion, nerd hubris and hysterically inflated market opportunities”
- People lie, are lazy, stupid, deluded, biased…
- The vocabulary problem…
- What is Doctorow's recommendation?

The HTML META Tag

- In 1994 (very early in Web history) a Computer Science graduate student proposed that HTML be revised to include a META tag
- "The META element can be used within the HEAD element to embed document metainformation not defined by other HTML elements. Such information can be extracted by servers/clients for use in identifying, indexing, and cataloging specialized document metainformation”
- (http://lists.w3.org/Archives/Public/www-html/1994Jun/0041.html)

What the W3C Imagined

<META NAME="DESCRIPTION" CONTENT="accurate prose description">

<META NAME="KEYWORDS" CONTENT="useful comma-separated keywords">

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

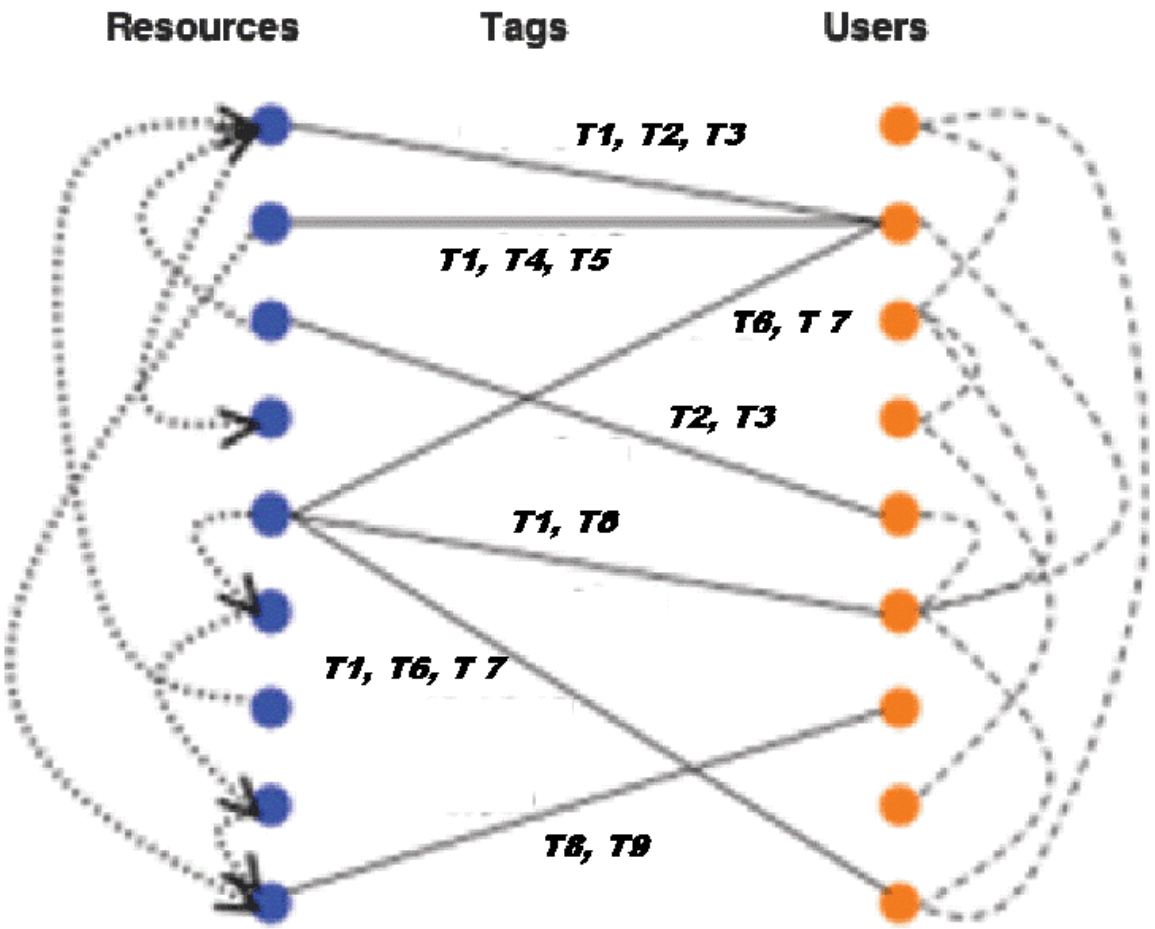
Robert J. Glushko
glushko@berkeley.edu
@rjglushko

26 September 2013
Lecture 9.3 – Resource Tagging

Social Tagging - Overview

- Trant's definition: "Publicly labeling or categorizing resources in a shared online environment"
- The aggregated results of individual tags have been described as: collaborative, cooperative, distributed, dynamic, community-based, folksonomic, wikified, ethnoclassification, democratic, user-assigned, or user-generated
- Tag sharing (or visibility) creates additional value through network effects

The Conceptual Model of Social Tagging



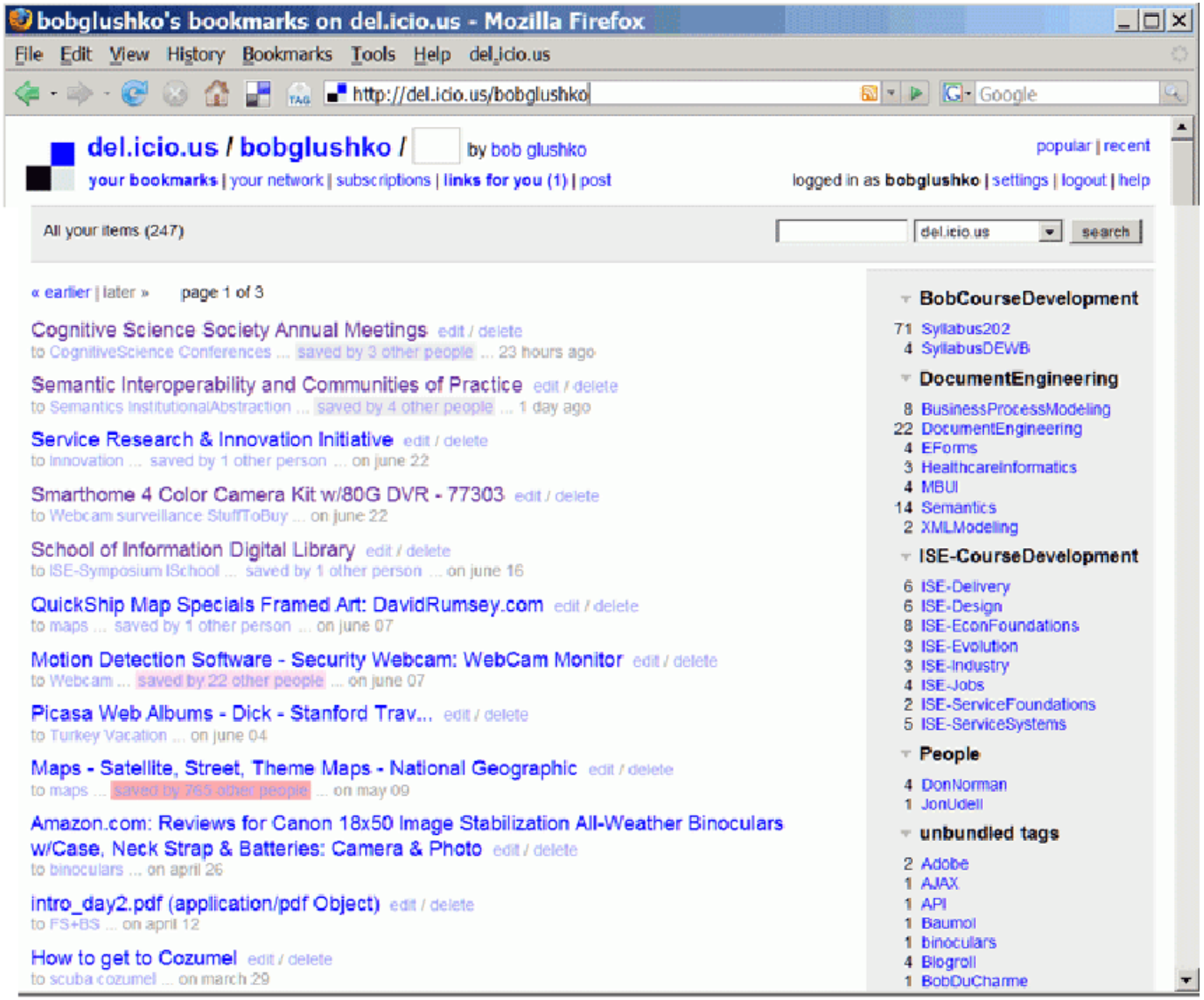
Design Dimensions for Tagging Systems

- What can be tagged? (anything, photos, web resources, bibliographic resources…)
- Tagging lexicography? (whitespace delimited string, phrases, normalization)
- Who can tag? (self, permitted people, anyone)
- Tagging support? (none, suggested, previous tags visible)
- Aggregation model? (none, “bag”, labeled set)

Types of Tags

- Subject /Taxonomic or Keyword Tags (most common, but rarely from a controlled vocabulary)
- Property or Attribute Tags ("red," "expensive“)
- "Purpose" Tags (e.g, "toread" or "buythis" or "tagthis“)
- Evaluative Tags ("interesting," "good)

Delicious (2005)



Tagging Functionality / UIs for Tagging

- Context is recorded automatically
- Share/Don't Share (or Private/Public): enable both personal organization and group organization
- Tag suggestion (tagging precedents)
- Tag organization into groups or categories
- Batch uploading and tagging
- Tag Visualization ("tag clouds")

Tag Quality / Correctness?

- When I first signed up for the popular bookmarking/ tagging site "delicious" the instructions said:
  - *Tagging is intuitive*
  - *A tag can be anything you want*
  - *There are no wrong tags*

Tag Me “Stanford” “Hillary” “Obama”

"Tag Soup"

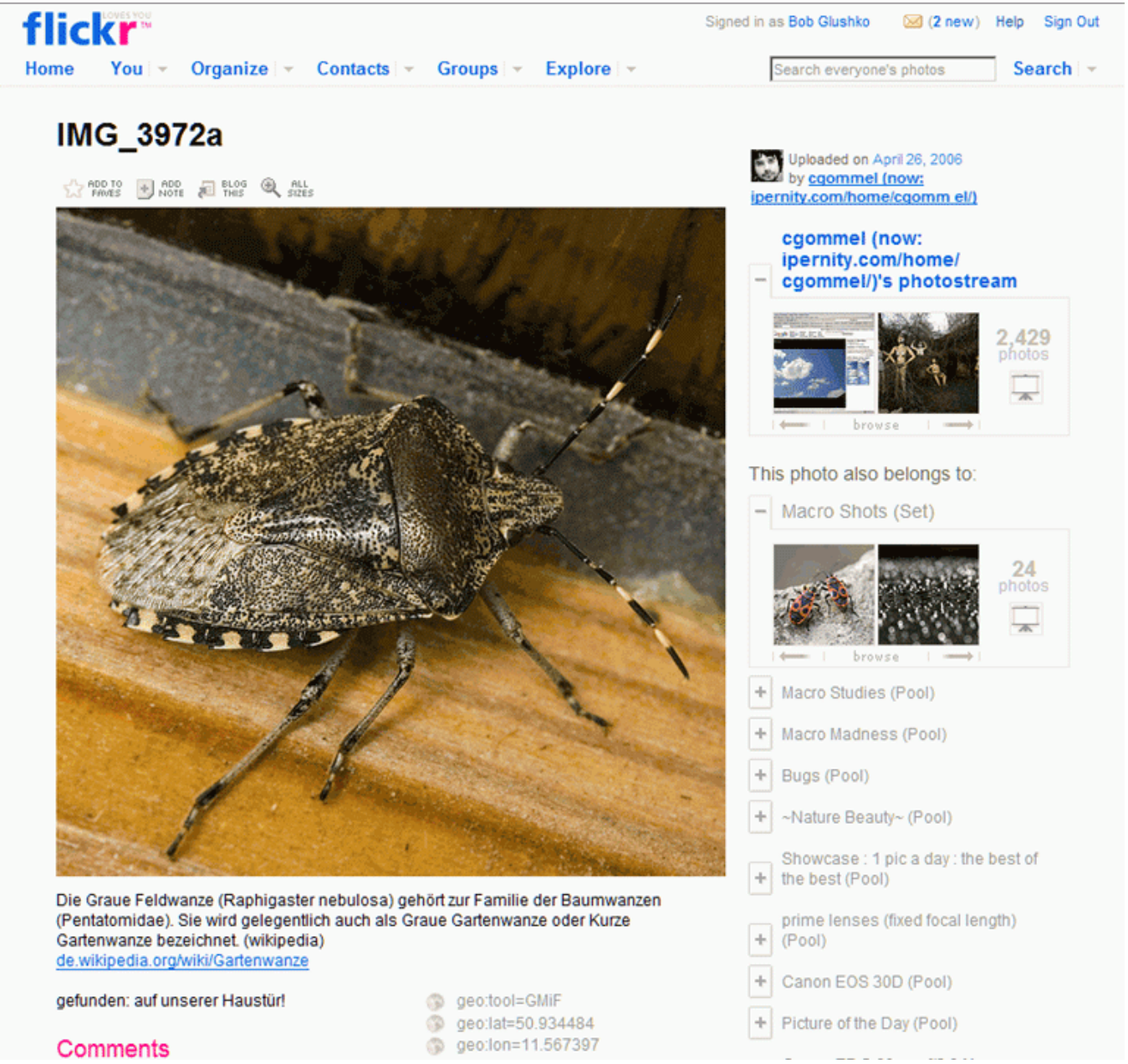
- Users are free to assign any number of labels or tags they choose
- No vocabulary control
- No text processing to handle derivational and morphological variants

Responses to Tag Soup

- Some people consider the unstructured, uncontrolled nature of "tag soup" to be its great strength
- Others adopt personal conventions to encode hierarchical and derivational relationships
- Using multiple accounts for the same application for different purposes
- “Tag bundles" to enable more hierarchy

Geotagging and Taxonomic Tagging

- Most tags don't come from controlled vocabularies, but geotagging and biological tagging are the exceptions that prove the rule
- Map interfaces in flickr and google earth can be used for geotagging but any GPS will do - by convention 3 tags are used:
  - geotagged
  - geo:lat=latitude e.g. geo:lat=51.4989
  - geo:lon=longitude e.g. geo:lon=-0.1786



Combined Geo and Bio Tagging

Tag Convergence

- Some systems (like del.icio.us) don't allow users to see the tags assigned by other users when they are tagging a resource
- But once a user tags a resource, most systems reveal the tags applied by other users
- If your tag(s) don't match, do you?
  - Change your tag to adapt to the group norm
  - Keep your tag to influence the group norm
  - Add the group tag but keep yours as well

Tag Clouds


Savignon
blackberry
Audi
Austin
vegetables
Glushko
grapefruit
boxer
yuck
anise
fresh
finish
grass
straw
sonoma
wicked
fruit
strong
grapey
pinot
nosy
noir
pear
bitter
Portugal
chardonnay
bright
malbec
Alcohol
Taylor
merlot
citrus
red
tree
astringent
tea
floral
lemon
lemony
metallic
Hills
bell
green
artificial
not
buzzed
taste
vino
sparkly
vanilla
black
rose
apricot
dry
light
Pino
carbonated
carbernet
prune
carmenere
windex
tart
fig
sour
berry
white
spicy
verde
acerip
Lotus
warm
spice
vinoverde
artichoke
No
juice
currant
crisp
garden
nape
whitewinesucks
raspberry
Wales
Chile
clove
CSNY
grassy
Yanni
syrah
sweet
vegetable
earthy
drunk
Embarcadero
Healy
tropical
beans
young
zinfandel
Sunshine
apple
sweetness
delicious
Mercedes
pepper
Alfa
grape
undry
Boston
peach
Romeo
good
James
grain
Chicago
plum
banana
rasin
clean
tipsy
blanc
Coye
peppery
fruity
tomato
raisin
sparkling

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

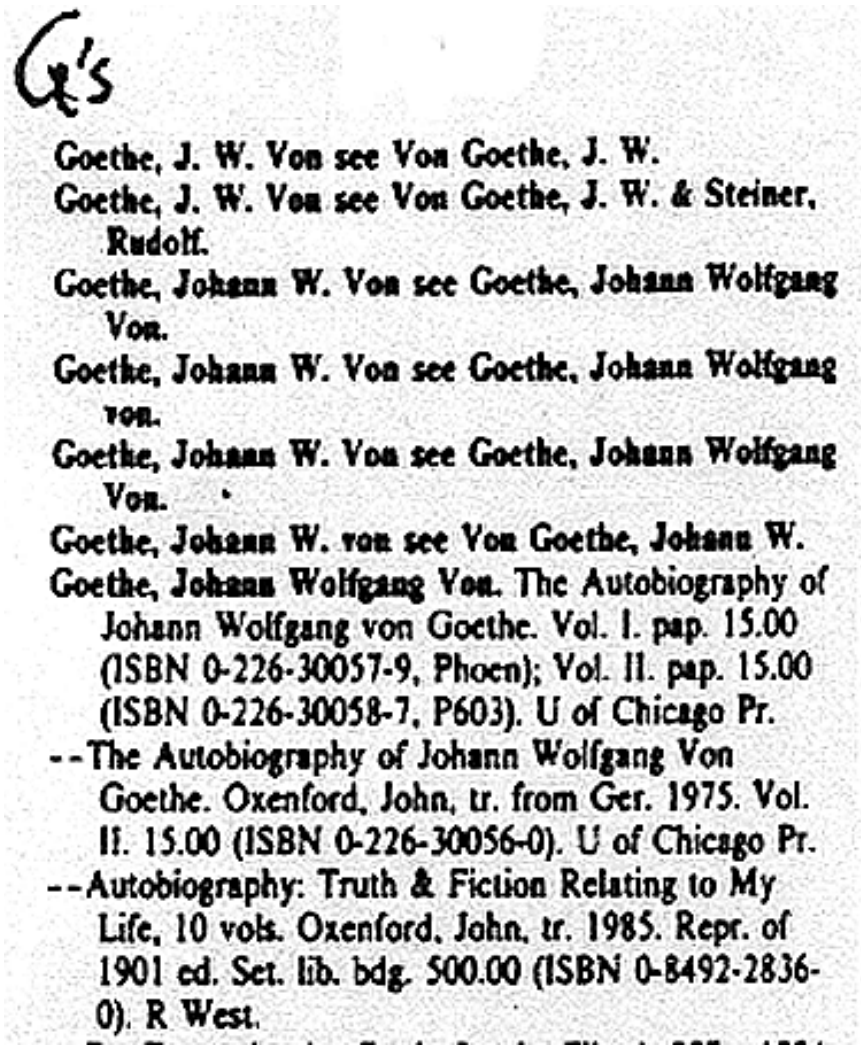
Robert J. Glushko
glushko@berkeley.edu
@rjglushko

26 September 2013
Lecture 9.4 – Authority Control
and Duplicate Detection

Variant Forms of Names

- A generic problem in resource description is dealing with variant name forms
  - Same person (or resource) uses or is given different names
  - Different people (or resources) use the same name

Different Forms of "Same" Author Name



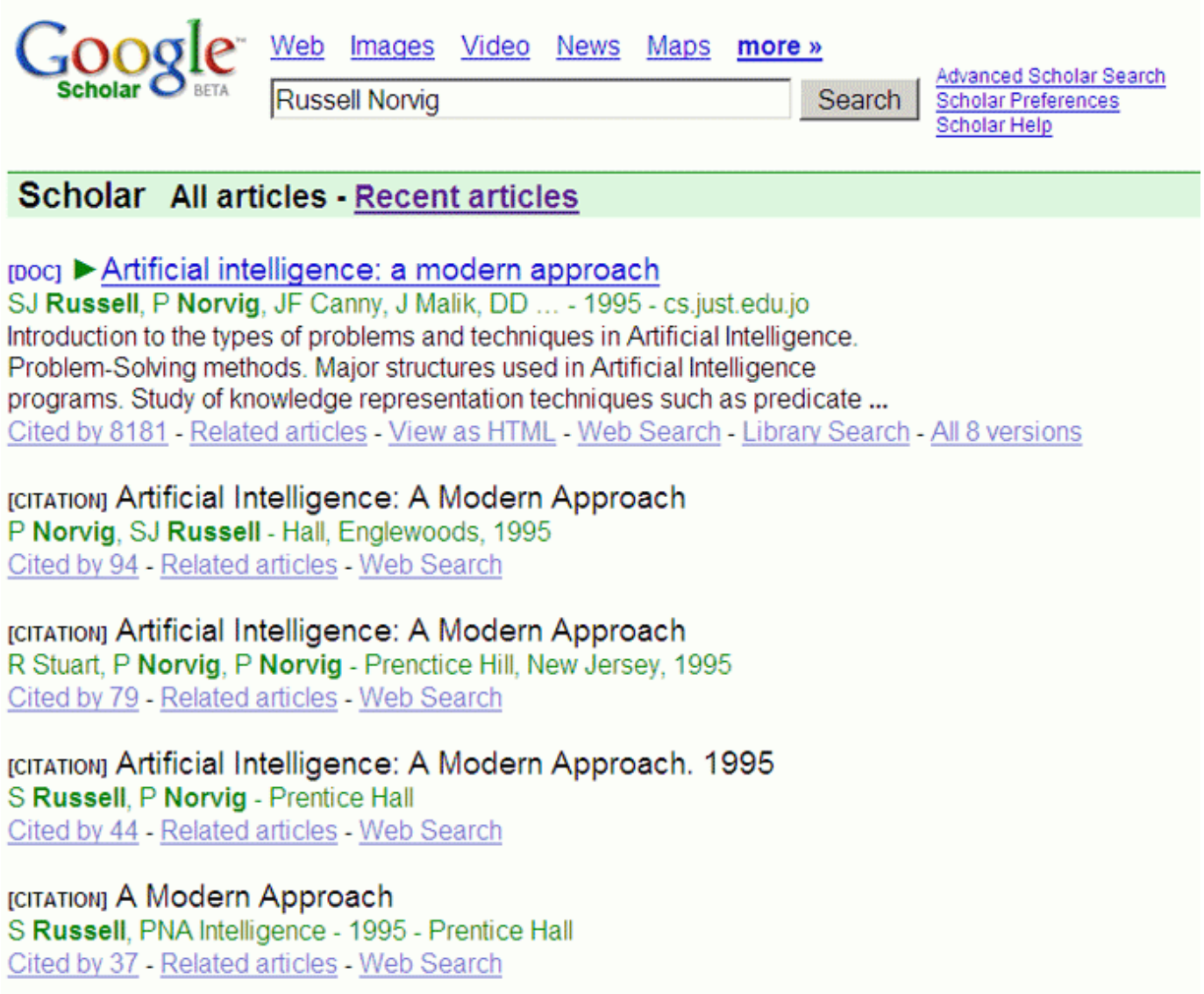
G's

Goethe, J. W. Von see Von Goethe, J. W.
Goethe, J. W. Von see Von Goethe, J. W. & Steiner, Rudolf.
Goethe, Johann W. Von see Goethe, Johann Wolfgang Von.
Goethe, Johann W. Von see Goethe, Johann Wolfgang von.
Goethe, Johann W. Von see Goethe, Johann Wolfgang Von.
Goethe, Johann W. von see Von Goethe, Johann W.
Goethe, Johann Wolfgang Von. The Autobiography of Johann Wolfgang von Goethe. Vol. I. pap. 15.00 (ISBN 0-226-30057-9, Phoen); Vol. II. pap. 15.00 (ISBN 0-226-30058-7, P603). U of Chicago Pr.
--The Autobiography of Johann Wolfgang Von Goethe. Oxenford, John, tr. from Ger. 1975. Vol. II. 15.00 (ISBN 0-226-30056-0). U of Chicago Pr.
--Autobiography: Truth & Fiction Relating to My Life, 10 vols. Oxenford, John, tr. 1985. Repr. of 1901 ed. Set. lib. bdg. 500.00 (ISBN 0-8492-2836-0). R West.

V's

Ballinger Pub.
Von Gloeden, Wilhelm, photos by. Taormina. (Illus.). 112p. 1986. 50.00 (ISBN 0-942642-22-8). Twelvetrees Pr.
Von Gaielinski, Stefan, ed. Liberia in Maps. LC 72-80411. (Graphic Perspectives of Developing Countries Ser.). (Illus.). 111p. 1973. 35.00 (ISBN 0-8419-0126-0, Africana). Holmes & Meier.
→Von Goethe, J. W. Conversations with Eckermann. Oxenford, John, tr. from Ger. 384p. (Orig.). 1984. pap. 16.50 (ISBN 0-86547-148-7). N Point Pr.
→Von Goethe, J. W. & Steiner, Rudolf. The Fairy Tale of the Green Snake & the Beautiful Lily. 2nd ed. LC 78-73644. 72p. (Orig.). 1981. pap. 3.50 (ISBN 0-89345-203-3, Steinerbks). Garber Comm.
→Von Goethe, J. W. see Goethe, Johann Wolfgang Von.
→Von Goethe, Johann see Goethe, Johann Wolfgang Von.
Von Goethe, Johann W. Goethe, Johann Wolfgang von, Italian Journey. Saine, Thomas P. & Sammons, Henry, eds. Heitner, Robert P., tr. from

“Dirty” Citations for the Same Book



Google Scholar BETA

Web Images Video News Maps more »

Russell Norvig Search

Advanced Scholar Search
Scholar Preferences
Scholar Help

Scholar All articles - Recent articles

[DOC] ► Artificial intelligence: a modern approach
SJ **Russell**, P **Norvig**, JF Canny, J Malik, DD … - 1995 - cs.just.edu.jo
Introduction to the types of problems and techniques in Artificial Intelligence. Problem-Solving methods. Major structures used in Artificial Intelligence programs. Study of knowledge representation techniques such as predicate ...
Cited by 8181 - Related articles - View as HTML - Web Search - Library Search - All 8 versions

[CITATION] Artificial Intelligence: A Modern Approach
P **Norvig**, SJ **Russell** - Hall, Englewoods, 1995
Cited by 94 - Related articles - Web Search

[CITATION] Artificial Intelligence: A Modern Approach
R Stuart, P **Norvig**, P **Norvig** - Prenctice Hill, New Jersey, 1995
Cited by 79 - Related articles - Web Search

[CITATION] Artificial Intelligence: A Modern Approach. 1995
S **Russell**, P **Norvig** - Prentice Hall
Cited by 44 - Related articles - Web Search

[CITATION] A Modern Approach
S **Russell**, PNA Intelligence - 1995 - Prentice Hall
Cited by 37 - Related articles - Web Search

Different Forms of Institutional Name

Same Name for Different Authors

Muir, Jessie, tr. see Bojar, Joann.

Muir, John. Como Mantener Tu Volkswagen Vivo. rev. ed. Holt, Virginia, tr. from Eng. LC 75-21414. (Illus., Orig.). 1980. pap. 10.00 (ISBN 0-912528-21-4). John Muir.

--The Coniferous Forests & Big Trees of the Sierra Nevada. Jones, William R., ed. (Illus.). 1980. pap. 4.95 (ISBN 0-89646-027-4). Outbooks.

The Cruise of the Corwin. 1918. 30.00 (ISBN 0-686-17252-3). Scholars Ref Lib.

--The Discovery of Glacier Bay (1879) Jones, William R., ed. (Illus.). 16p. 1978. pap. 2.50 (ISBN 0-89646-045-2). Outbooks.

--Es Lebe Mein Volkswagen. Shamai, Ruth & Jeschke, Herbert, trs. (Illus.). 308p. 1978. pap. 10.00 (ISBN 3-980018-90-3). John Muir.

--How to Keep Your Volkswagen Alive. 11th ed. 432p. 1988. pap. 17.95 (ISBN 0-945465-12-2). John Muir.

--The Hummingbird of the California Waterfalls. Jones, William R., ed. (Illus.). 24p. 1977. pap. 2.50 (ISBN 0-89646-019-3). Outbooks.

--In the Heart of the California Alps. Jones, William R., ed. (Illus.). 24p. 1977. pap. 2.50 (ISBN 0-89646-026-6). Outbooks.

--Industrial Relations Procedures & Agreements. 200p. 1981. text ed. 40.00x (ISBN 0-566-02275-3).

(Illus.). 247p. (Orig.). 1980. pap. 4.50 (ISBN 912528-02-8). John Muir.

--The Wild Sheep. Jones, William R., ed. (Illus. 1977. pap. 2.50 (ISBN 0-89646-017-7). Ou

--Wilderness Essays. Buske, Frank, ed. (Liter the American Wilderness Ser.). 288p. 1980 4.95 (ISBN 0-87905-072-1, Peregrine Smith Gibbs Smith Pub.

--The Yellowstone National Park. Jones, Wil ed. (Illus.). 1978. pap. 3.95 (ISBN 0-89646 Outbooks.

--Yellowstone National Park. (Illus.). 1979. p (ISBN 0-89646-079-7). Outbooks.

--The Yosemite. LC 86-15849. 320p. 1987. 32.50x (ISBN 0-299-11100-8); pap. 10.95 299-11104-0). U of Wis Pr.

--The Yosemite. LC 87-23573. (John Muir L (Illus.). 288p. 1988. pap. 9.95 (ISBN 0-871 2). Sierra.

Muir, John & Gregg, Tech. How to Keep You Volkswagen Alive: A Manual of Step by S Procedures for the Compleat Idiot. 32nd 79-63486. (Illus.). 384p. (Span., Ger., & E 1986. pap. 17.95 (ISBN 0-912528-50-8). J Muir.

Muir, K., ed. see Calderon De La Barca, Pe

Muir, Karen L. The Strongest Part of the Fa

Questions about Names

- How many names should be associated with a information object or resource?
- Should one be designated as the preferred or authoritative form?
- What references should be made from other possible forms of names that haven't been used?

Authority Control

- Authority control is concerned with creation and maintenance of a set of terms that have been chosen as the standard representatives for some resource based on some set of rules
- The Library of Congress maintains an "authority file" (in MARC format) for the names of persons, corporate entities, geographic names of political entities, and titles of works
- See http://www.loc.gov/marc/uma/ and http://authorities.loc.gov/

Normative Forms of Names

- When names appear in multiple forms, one form needs to be chosen; criteria include:
  - Fullness (e.g., full names vs. initials only)
  - Language of the name
  - Spelling (choose predominant form)
  - Entry element
    - "Smith, John" not "John Smith"
    - "Mao Zedong" or "Zedong, Mao" or "Mao Tse Tung" or ?

Naming Problems for Places

- Variant forms: St. Petersburg, Санкт Пербургскйй, Saint-Pétersbourg
- Multiple names: Cluj, in Romania / Roumania / Rumania, is also called Klausenburg and Kolozsvar
- Name changes: Bombay -> Mumbai.
- Homographs:Vienna, VA, and Vienna, Austria; 50 Springfields
- Anachronisms: No Germany before 1870
- Vague, e.g. Midwest, Silicon Valley
- Unstable boundaries: 19th century Poland; Balkans; USSR

Authoritative Place Names

- Places have latitude and longitude coordinates, so we can link places and spaces with a GAZETTEER
- A gazetteer is a place name authority file that:
  - Indicates what kinds of place: "Feature type"
  - Objectively specifies latitude and longitude
  - Disambiguates similar place names
  - Brings variant names together
  - Allows places to be displayed on maps

Code Sets

- Code sets are constrained sets of values that are often completely arbitrary but they are unambiguous and authoritative names
- The ISO code sets for countries (3166), currencies (4217), quantities and units of measure (31) are very commonly used
- Most organizations have internal code sets or business rules that implicitly define them

The Name Matching Problem

- Name matching" is the task of determining when two different strings denote the same person, object, or other named entity
- It is ironic that this problem has many other names:
  - Co-reference resolution
  - Duplicate detection
  - Record linkage
  - Object consolidation
  - Merge-purge
  - Householding
  - Fuzzy/approximate matching

Duplicate Detection vs. Exact Copies

Table 1.1: Exact duplicate r_1 and r_2 and fuzzy duplicate r_1 and r_3

	FN	LN	Phone	email
r_1	John	Doe	(407) 356 8888	john@doe.com
r_2	John	Doe	(407) 356 8888	john@doe.com
r_3	Jon	Doe	(407) 356 8887	john@doe.com

Domains For The Name Matching Problem

- Customer relationship management
- Direct mail
- Law enforcement / counter-terrrorism
- Demography / population statistics (e.g. census)
- Computational biology
- Data mining, signal processing, and image compression tasks in numerous application areas

Causes of Duplicate Names / Why Name Matching is Hard

- Different data capture formats
- Poor or mis-spellings (original, transcription, OCR, “autocorrection”)
- Phonological -> orthographic alternatives (multiple ways to spell the same sound… which might not have been heard correctly)
- Language transliteration ("Peking" or "Beijing")
- Nicknames and variant forms

Naming the Same Song?

Figure 1: Top 25 Representations of "Knockin' On Heaven's Door" [35]

```
Guns N' Roses  -  Knockin' On Heaven's Door
Guns N' Roses  -  Knocking On Heavens Door
Guns 'N' Roses  -  Knockin' On Heaven's Door
Guns N' Roses  -  Knockin On Heavens Door
Guns N' Roses  -  Knockin' On Heavens Door
Guns N'roses  -  knockin on heavens door
Guns N' Roses  -  Knocking On Heaven's Door
Guns N Roses  -  Knockin' On Heaven's Door
Guns N Roses  -  Knockin On Heavens Door
Guns And Roses  -  Knocking On Heavens Door
Guns Nroses  -  Knockin On Heavens Door
Guns'n' Roses  -  Knockin' On Heaven's Door
Guns N Roses  -  Knocking On Heavens Door
Guns'n'Roses  -  Knockin' On Heaven's Door
Guns 'N Roses  -  Knockin' On Heaven's Door
Guns & Roses  -  Knockin' on Heaven's Door
Guns N'roses  -  Knockin' On Heaven's Door
Guns and Roses  -  Knockin' On Heaven's Door
Guns'n Roses  -  Knocking On Heavens Door
Guns 'n' Roses  -  Knockin' On Heavens Door
Aerosmith  -  Knocking On Heaven's Door
Guns 'n Roses  -  Knocking On Heaven's Door
Guns 'n' Roses  -  Knocking On Heavens Door
Guns N Roses  -  Knocking On Heaven's Door
Guns N' Roses  -  Knockin On Heaven's Door
```

Causes of Duplicate Names / Why Name Matching is Hard

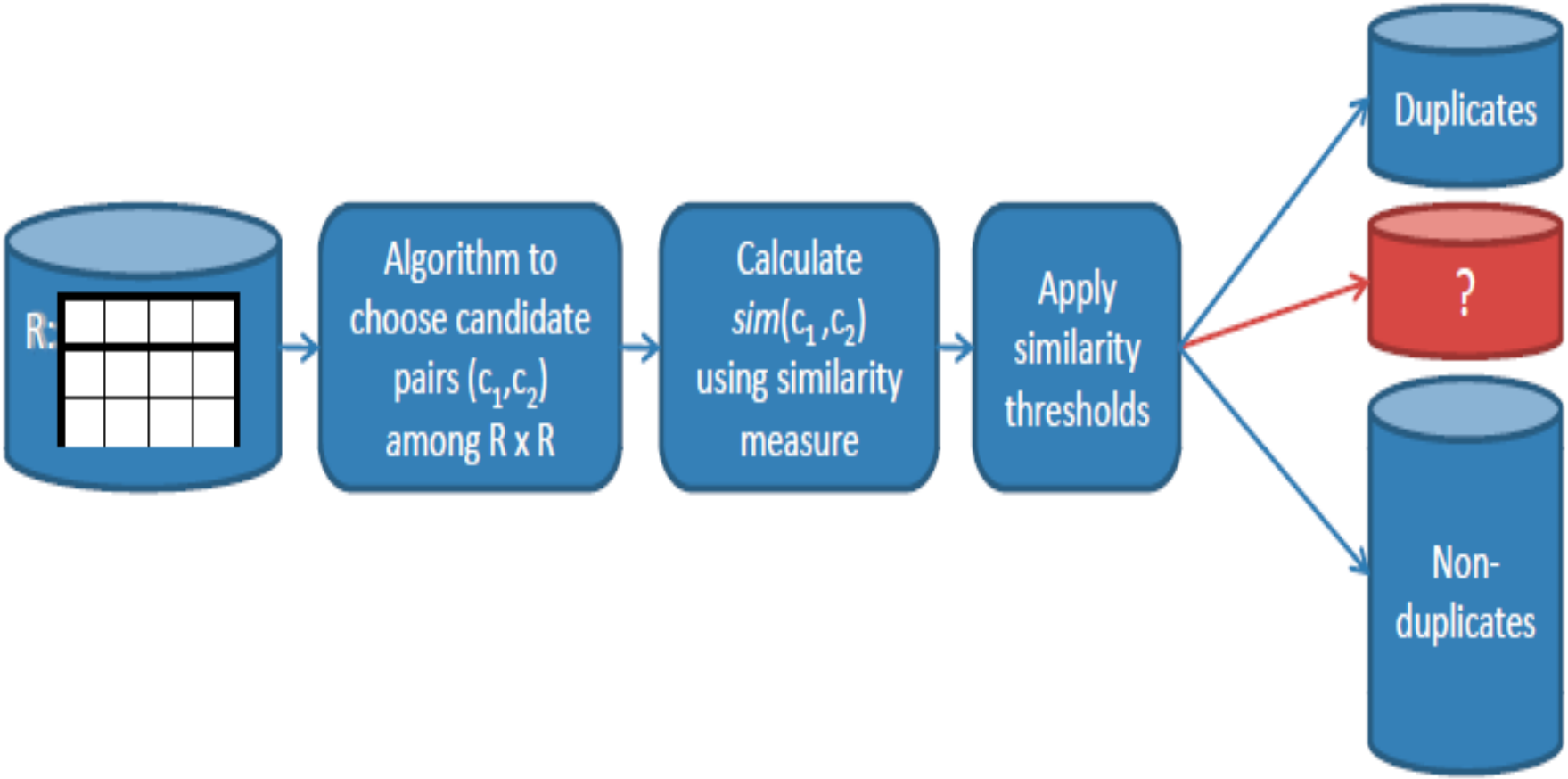
- Name changes for people (http://www.famousnamechanges.net/)
- Name changes for countries (http://www.nationsonline.org/oneworld/hist_country_names.htm)

(effectivity of names)

Causes of Duplicate Names / Why Name Matching is Hard

- Name permutations and omissions (complex rules for language and culture apply http://en.wikipedia.org/wiki/Family_name
- Definite descriptions and metonymy

Generic Approach to Name Matching



Computing Similarity between Names

- Measures of orthographic similarity
  - "Edit distance" - how many insertions, deletions, or substitutions to turn one string into another
  - Can be weighted using likelihood or word order considerations
- Measuring pronunciation similarity
  - "Hash" the names into phonetic encodings with fewer characters than original
  - "Soundex" function is very commonly used http://www.searchforancestors.com/soundex.html

Operation Clean Data

- What were the symptoms or implications of "dirty" data in the British army's supply chains?
- What were the primary causes of this "dirty" data?
- Which data items were the focus of the data cleanup effort? Why?
- What technologies or tools were used in the data cleanup effort?

Data Quality

- Data quality problems can have any combination of technological causes, process causes, or managerial causes
- Does data have to be perfectly clean? Can it ever be?
- How can your own actions contribute to data quality problems or to their resolution?

Principles and Processes for Quality Information

- Prioritize the data items
- Involve the data owners
- Find the data owners and the "headwaters“
- Validate at the time of capture or creation
- Set realistic goals for data quality

Readings for Next Lecture

- TDO 4.4
- Harpring, Patricia. “The language of images: enhancing access to images by applying metadata schemas and structured vocabularies.”
- Bailer, Werner et al., Multimedia Semantics: Use Case Scenarios.
- Christel, Michael. Automated Metadata in Multimedia Information Systems, Chapter 2