7. Controlled Names and Controlled Vocabularies

INFO 202 - 22 September 2008

Bob Glushko

Plan for INFO Lecture #7

What is a Name?

Uncontrolled and Controlled Vocabularies

Authority Control

Name Matching

What Is A Name?

A NAME is a label for some thing or some category that is used to distinguish one from another

A thing or category can often have multiple names; these are SYNONYMS or ALIASES

Different things can sometimes have the same names -- these are HOMYNYMS or POLYSEMES

If a name is used to refer to some thing and is unique in some context it is an IDENTIFIER

The Need for Controlled Vocabularies

The words people use to describe things or concepts are "embodied" in their context and experiences... so they are often different or even "bad" with respect to the words used by others

These naturally-occurring words are an "uncontrolled vocabulary" and are what people use with the limited capability of the "search box"

Searches made using an uncontrolled vocabulary will not have high recall because they will fail to match documents indexed using "good" terms

Reaching agreement on the choice of words to improve recall means that we must use a subset of the words we would each otherwise use

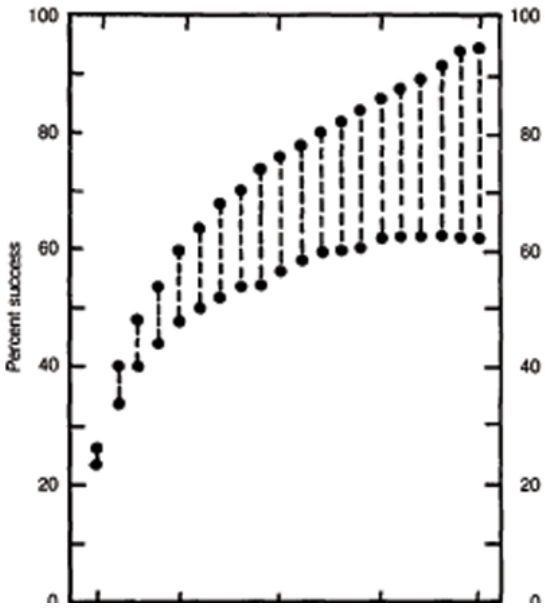
The Vocabulary Problem - Different Names for the Same Thing

Armchair" Method - Probability of two people coming up with the same term: 7, 8, 11, 12, 14, 18%

"Naming by Voting" -- Probability of someone using most popular word: 15, 21, 22, 28, 34, 36%

"Select What I Intend" -- Probability that two people using the same term intend the same referent: 13, 15, 41, 52, 62, 73%

Does Aliasing Solve the Vocabulary Problem?



Same Name for Different Things -- "Shipping Container"

"The expense of resolving ambiguous business terms over and over on a daily basis pales in comparison with the expense of NOT realizing there is an ambiguity in the term" (Farish)

The Wisdom of Svenonius...

Information is organized by describing it using a special-purpose language (p.1)

The essential and defining objective of a system for organizing information is to bring essentially the like information together and to differentiate what is not exactly alike (p. 11)

Vocabulary control is the sine qua non of information organization (p. 89)

What is 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 information
- to help searchers find information (with high recall and precision) and to identify the entity found

A CV is a content standard for use in (or as) metadata elements

The goal of a CV is to "impose some order to facilitate agreement between the concepts within the site and the vocabulary of the person [natural language] using it."

A CV can be thought of as a fixed or closed dictionary in which everything must be defined using the same set of terms

Types of Controlled Vocabularies

Dictionaries

Names and name authorities

Authority control for places and time periods

Identifiers

Code lists

Subject heading lists

Synonym rings

Thesauri

Classification schemes

Imposing Vocabulary Control (Svenonius, p. 89)

The imposition of a VC creates an artificial language out of a natural one:

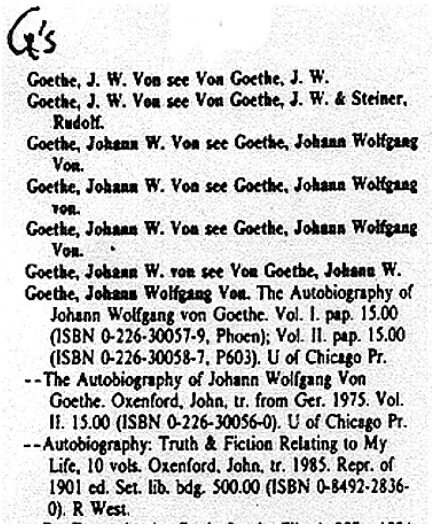
1. Choose the authoritative form of name
2. Ensure that the name is distinctive
3. Map all the variant name forms to the authoritative one

Variant Forms with Names

Proliferation of the forms of names

- Same person (or entity) uses or is given different names
- Different people (or entities) use the same name

Books in Print - Goethe



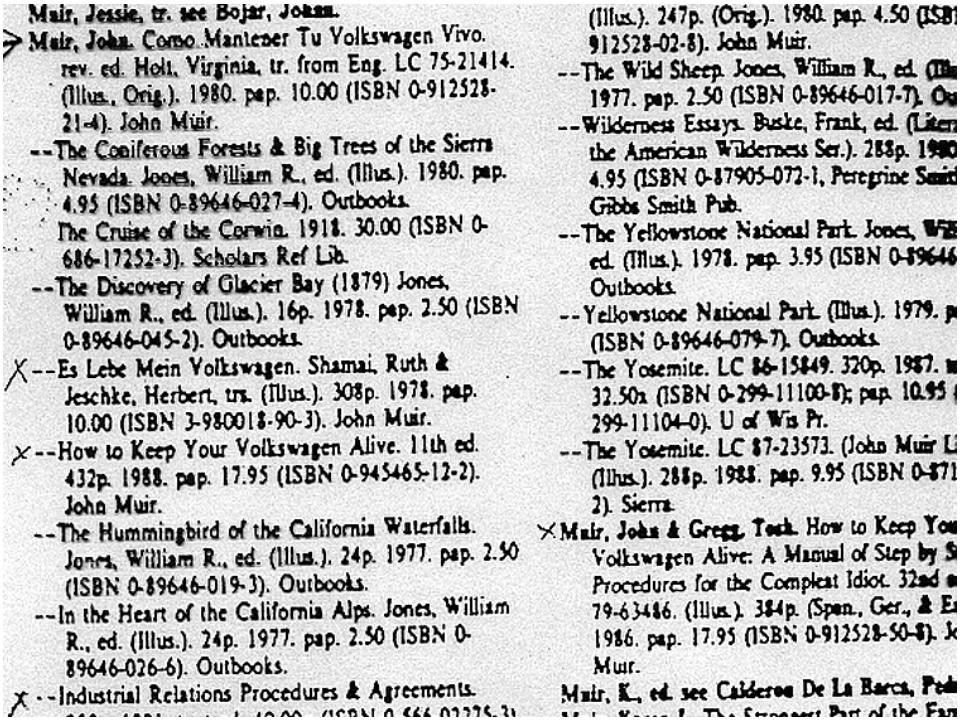
G's

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 Gnielinski, 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

Books in Print - John Muir



Muir, Jessie, tr. see Bojar, Johan.
→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.
X--Es Lebe Mein Volkswagen. Shamai, Ruth & Jeschke, Herbert, trs. (Illus.). 308p. 1978. pap. 10.00 (ISBN 3-980018-90-3). John Muir.
X--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.
X--Industrial Relations Procedures & Agreements.
(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.
X Muir, John & Gregg, Tosh. 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, Pedr

Problems With Names

How many names should be associated with a document or information object?

Which of these should be the "main entry?"

What form should each of the names take?

What references should be made from other possible forms of names that haven't been used?

Rules for Description

The AACR II (2002) and other sets of descriptive cataloging rules provide guidelines for:

Determining the number of name entries

Choosing a main entry

Deciding on the form of name to be used

Deciding when to make references to name variations

Authority Control

Authority control is concerned with creation and maintenance of a set of terms that have been chosen as the standard representatives (also known as established) 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 (http://www.loc.gov/marc/uma/)

AACR II says that the predominant form of the name used in a particular author's writings should be chosen as the form of name

References should be made from the other forms of the name

Normative Name Forms

When names appear in multiple forms, one form needs to be chosen using criteria that include:

Fullness (e.g., full names vs. initials only)

Language of the name

Spelling (choose predominant form)

Entry element

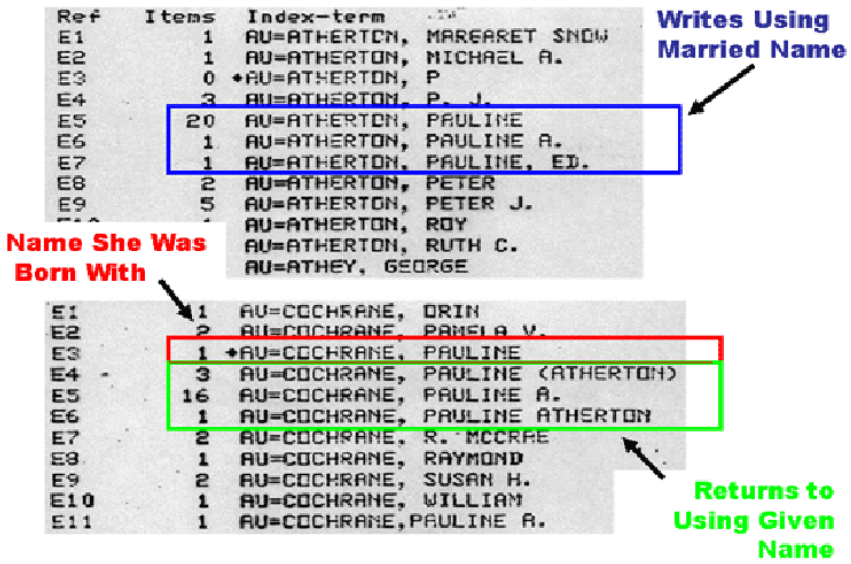
- "Smith, John" not "John Smith"
- "Mao Zedong" or "Zedong, Mao" or "Mao Tse Tung" or ?

My Authoritative Name?

LC Control Number: n 2005027744
HEADING: Glushko, Robert J.
000 00336nz a2200121n 450
001 6515030
005 20050415124620.0
008 050415n| acannaabn |n aaa
010 __ |**a** n 2005027744
040 __ |**a** DLC |**b** eng |**c** DLC
100 1_ |**a** Glushko, Robert J.
670 __ |**a** Glushko, Robert J. Document engineering, 2005: |**b** CIP t.p. (Robert J. Glushko)
953 __ |**a** jf05

LC Control Number: n 89666774
HEADING: Glushko, Robert John, 1953-
000 00431nz a2200133n 450
001 1180466
005 19891122172408.0
008 891122n| acannaab |n aad
010 __ |**a** n 89666774
035 __ |**a** (DLC)n 89666774
040 __ |**a** DLC |**c** DLC
100 10 |**a** Glushko, Robert John, |**d** 1953-
670 __ |**a** nuc89-101599: His The psychology of phonography, 1979 |**b** (hdg. on CU-S rept.:
953 __ |**a** np17

Advice to Professional Women Who Publish



Pseudonyms and Name Authority Files [1]

```
ID:NAFL8057230     ST:p    EL:n   STH:a    MS:c   UIP:a TD:19910821174242
KRC:a    NMU:a    CRC:c   UPN:a   SBU:a   SBC:a   DID:n DF:05-14-80
RFE:a    CSC:     SRU:b   SRT:n   SRN:n   TSS:    TGA:?   ROM:? MOD:
VST:d 08-21-91                      Other Versions: earlier
040    DLC$cDLC$dDLC$dOCoLC
053    PR6005.R517
100 10 Creasey, John
400 10 Cooke, M. E.
400 10 Cooke, Margaret,$d1908-1973
400 10 Cooper, Henry St. John,$d1908-1973
400 00 Credo,$d1908-1973
400 10 Fecamps, Elise
400 10 Gill, Patrick,$d1908-1973
400 10 Hope, Brian,$d1908-1973
400 10 Hughes, Colin,$d1908-1973
400 10 Marsden, James
400 10 Matheson, Rodney
400 10 Ranger, Ken
400 20 St. John, Henry,$d1908-1973
500 10 $wnnnc$aAshe, Gordon,$d1908-1973
```

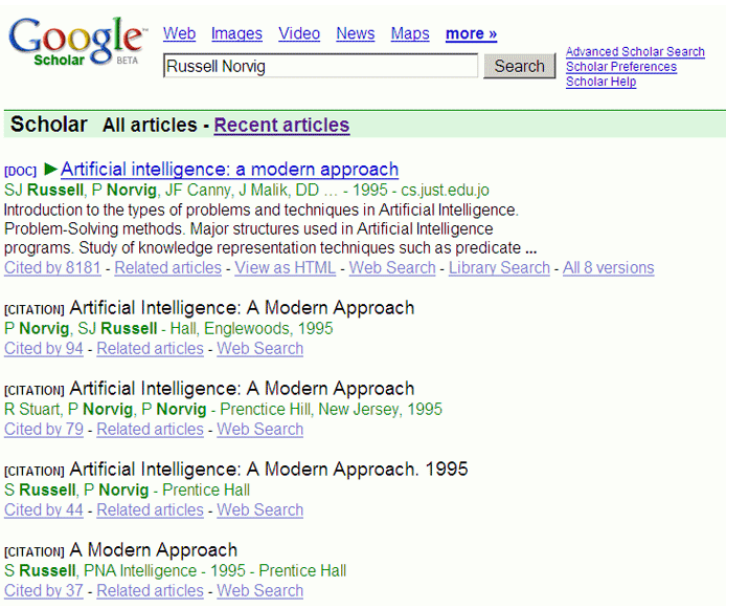
Pseudonyms and Name Authority Files [2]

```
ID:NAFO9114111     ST:p    EL:n   STH:a    MS:n   UIP:a TD:19910817053048
   KRC:a    NMU:a    CRC:c   UPN:a   SBU:a   SBC:a   DID:n DF:06-03-91
   RFE:a    CSC:c    SRU:b   SRT:n   SRN:n   TSS:    TGA:?   ROM:? MOD:
   VST:d 08-19-91
   040    OCoLC$cOCoLC
   100 10 Marric, J. J.,$d1908-1973
   500 10 $wnnnc$aCreasey, John
   663    Works by this author are entered under the name used in the item. For
        a listing of other names used by this author, search also under$bCrease
       y, John
   670    OCLC 13441825: His Gideon's day, 1955$b(hdg.: Creasey, John; usage: J
       .J. Marric)
   670    LC data base, 6/10/91$b(hdg.: Creasey, John; usage: J.J. Marric)
   670    Pseuds. and nicknames dict., c1987$b(Creasey, John, 1908-1973; Britis
       h author; pseud.: Marric, J. J.)
```

Pseudonyms and Name Authority Files [3]

```
ID:NAFL8166762     ST:p    EL:n   STH:a    MS:c   UIP:a TD:19910604053124
   KRC:a    NMU:a    CRC:c   UPN:a   SBU:a   SBC:a   DID:n DF:08-20-81
   RFE:a    CSC:     SRU:b   SRT:n   SRN:n   TSS:    TGA:?   ROM:? MOD:
   VST:d 06-06-91                     Other Versions: earlier
   040    DLC$cDLC$dDLC$dOCoLC
   100 10 Butler, William Vivian,$d1927-
   400 10 Butler, W. V.$q(William Vivian),$d1927-
   400 10 Marric, J. J.,$d1927-
   670    His The durable desperadoes, 1973.
   670    His The young detective's handbook, c1981:$bt.p. (W.V. Butler)
   670    His Gideon's way, 1986:$bCIP t.p. (William Vivian Butler writing as J
       .J. Marric)
```

It's Not Always the Author's Fault



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 also goes by many other names:

- Co-reference resolution
- Duplicate detection
- Record linkage
- Merge-purge

It arises in database maintenance, text retrieval, data mining, signal processing, and image compression tasks in numerous application areas

Why Name Matching is Hard

Misspellings (original, transcription, or OCR)

Phonological -> orthographic alternatives

Language transliteration ("Peking" or "Beijing"

Nicknames and variant forms

Name changes (for people; for countries)

Name permutations and omissions (complex language/culture rules apply)

Definite descriptions and metonymy

Technical Approches to Solving the Name Matching Problem

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 pronunication similarity

- "Hash" the names into phonetic encodings with fewer characters than original
- "Soundex" function is very commonly used

Duplicate Detection

If names are found in context of other information (e.g., address, SSN, phone, DOB), align and validate these other fields before attempting name matching and duplicate detection

	Name	Address	City	St	ZIP®	Phone
1	SMITH	2 E 13TH ST	CHICAGO	IL	60601-2407	(312) 458 9992
2		2 13 ST EAST	CHICAGO	IL	60605	SMITH
3	SMTH	2 EAST 13TH	CHICAGO LAWN	IL		312-458-9992
4	SMITH	2 E THIRTEENTH ST	CHICAGO	IL	60605	458-9992
5	SMITH	TWO EAST 13TH ST	CHICAGO IL		60602	312-458-9991

	Name	Address	City	St	ZIP	Phone
1	SMITH	2 E 13TH ST	CHICAGO	IL	60601-2407	312-458-9992
2	SMITH	2 E 13TH ST	CHICAGO	IL	60601-2407	
3	SMTH	2 E 13TH ST	CHICAGO	IL	60601-2407	312-458-9992
4	SMITH	2 E 13TH ST	CHICAGO	IL	60601-2407	XXX-458-9992
5	SMITH	2 E 13TH ST	CHICAGO	IL	60601-2407	312-458-9991

Conditions of Authorship are Complex

Single person or single corporate entity

Unknown or anonymous authors

Fictitiously ascribed works

Shared responsibility (synchronous or asynchronous?)

Collections or editorially assembled works

Works of mixed responsibility (e.g., translations)

Added Names Beyond "Author" in Bibliographic Descriptions

Personal names

Collaborators

Editors, compilers, writers

Translators (in some cases)

Illustrators (in some cases)

Other persons associated with the work (such as the honoree in a festschrift)

Corporate names - Any prominently named corporate body that has involvement in the work beyond publication, distribution

Authority Control 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

Gazetteers

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

Authority Control for Time Periods

Places and place names have temporal aspects and time period names resemble place names

Some periods have objective calendar dates as well as name

But not always:

- Imprecise: Elizabethan, Neolithic
- Ambiguous: Civil war, Renaissance, . . . which?
- Unstable: The European War, The Great War, World War I

Identifiers

Identifiers are UNIQUE if they refer to one and only thing within some defined context or scope

But the same thing can have more than one identifier

The same identifier can mean different things in different contexts or at different times

Identifiers are PERSISTENT if they resolve to the same referent indefinitely, or as long as needed; but "persistence is a functions of organizations, not technology"

Identifiers are UNSTRUCTURED or DUMB (as opposed to STRUCTURED or INTELLIGENT) if they have no inherent meaning based on their values

Many identifier schemes are designed to be STRUCTURED or INTELLIGENT (like the ISBN), but over time they often become less so

ISBN on Bar Code

Code Lists

Codes are constrained sets of values

Codes establish their meaning by reference to those values, often by abbreviations

Using codes in vocabularies and metadata promotes consistency and makes meaning unambiguous

External and Internal Codes

External codes are those maintained by some entity or organization outside of your control (ISO, ANSI, etc.)

- ISO 639 - language codes
- ISO 3166 - country codes
- ISO 4217 - currency codes
- IATA port codes
- Internet Top Level Domains

Internal codes are code sets that you can define and control

Key Points in Today's Lecture

Naming is a challenging and often contentious activity

The "uncontrolled" words that people naturally use to describe things or concepts are "embodied" in their context and experiences ... so they are often different or even "bad" with respect to the words used by others

A "controlled" vocabulary creates an artificial language to be used in place of the "natural" ones to facilitate agreement in some context and for some purposes

Readings for INFO 202 Lecture #8

Svenonius Chapter 8 (139-146), Chapter 9 (159-171), Chapter 10

William Denton, "How to make a faceted classification and put it on the web"