8. Classification

INFO 202 - 24 September 2008

Bob Glushko

Plan for INFO 202 Lecture #8

Overview of Classification

Faceted Classification

What is Classification?

"Classification is a higher order thinking skill requiring the fusion of the naturalist's eye for relationships...

...with the logician's desire for structured order...

...the mathematician's compulsion to achieve consistent, predictable results...

...and the linguist's interest in explicit and tacit expressions of meaning"

-- Louise Gruenberg, "Faceted Classification, Facet Analysis and the Web"

Distinguishing Categorization and Classification

Categories are EQUIVALENCE CLASSES - sets of material and abstract things, processes, and events that we treat the same

A Classification (noun) is a SYSTEM OF CATEGORIES, ordered according to a PRE-DETERMINED SET OF PRINCIPLES and used to organize a set of instances or entities

Classification (verb) is the process of assigning instances or entities to the categories in a classification system

Functions of Classification Systems

Serve as a reference model (a semantic road map) to individual domains and the relationships among them to help people understand the concepts and relationships

Improve learning and communication

They are essential in useful work

They support information retrieval and UIs for IR systems

Classification is "Principled" -- Lawful, Systematic and Arbitrary

LAWFUL because it follows a set of principles that govern the structure of the categories and their relationships

SYSTEMATIC because it requires that those principles be followed

ARBITRARY because the criteria used to establish the categories reflect a perspective on the domain that intentionally excludes all other perspectives

What Kinds of Principles?

Classify all knowledge, or classify only the items being classified (the concept of *literary warrant*)?

Classify once and for all, or change as we learn?

Classify broadly, or classify precisely? How *enumerative* are the classifications?

Hierarchy and Uniqueness Principles

Each category is successively divided into smaller subdivisions

Every level of the hierarchy is divided according to a "feature" or "character of division"

Every item/instance is classified in only one subdivision

Guidance on the "One and Only One Place" Rule

From Introduction to Dewey Decimal Classification (http://www.oclc.org/dewey/versions/ddc22print/intro.pdf)

The title is often a clue to the subject, but should never be the only thing analyzed

A work is classed in the discipline for which it is intended, rather than the discipline from which the work derives

Works dealing with multiple subjects are classed with the subject being acted upon

Class a work multiple on subjects with the one receiving fuller treatment.

If two subjects are equal, class the work using the one that comes first in the DDC

Library of Congress Classification

Used by most university and research libraries in US and elsewhere

```
A -- GENERAL WORKS
B -- PHILOSOPHY. PSYCHOLOGY. RELIGION
C -- AUXILIARY SCIENCES OF HISTORY
D -- HISTORY (GENERAL) AND HISTORY OF EUROPE
E -- HISTORY: AMERICA
F -- HISTORY: AMERICA
G -- GEOGRAPHY. ANTHROPOLOGY. RECREATION
H -- SOCIAL SCIENCES
J -- POLITICAL SCIENCE
K -- LAW
L -- EDUCATION
M -- MUSIC AND BOOKS ON MUSIC
N -- FINE ARTS
P -- LANGUAGE AND LITERATURE
Q -- SCIENCE
R -- MEDICINE
S -- AGRICULTURE
T -- TECHNOLOGY
U -- MILITARY SCIENCE
V -- NAVAL SCIENCE
Z -- BIBLIOGRAPHY. LIBRARY SCIENCE. INFORMATION RESOURCES (GENERAL)
```

Where's Computer Science?

```
Q  Science (General)
        QA      Mathematics
        QB      Astronomy
        QC      Physics
        QD      Chemistry
        QE      Geology
        QH      Natural history - Biology
        QK      Botany
        QL      Zoology
        QM      Human anatomy
        QP      Physiology
        QR      Microbiology
```

Dewey Decimal Classification

Started in 1876, the DDC is the most widely used classification system in the world, especially in public libraries

```
000 Computers, information & general reference
100 Philosophy & psychology
200 Religion
300 Social sciences
400 Language
500 Science
600 Technology
700 Arts & recreation
800 Literature
900 History & geography

600 Technology (Applied sciences)
        630 Agriculture and related technologies
                636 Animal husbandry
                        636.7 Dogs
                        636.8 Cats
```

DDC on Religion

```
200 Religion
        210 Natural theology
        220 Bible
        230 Christian theology
        240 Christian moral & devotional theology
        250 Christian orders & local church
        260 Christian social theology
        270 Christian church history
        280 Christian sects & denominations
        290 Other religions
```

Every Classification is "Biased"

Every classification system takes a point of view

Every classification system implicitly or explicitly distinguishes between "good" or "standard" and "bad" or "nonstandard" ways of understanding things

Biased Classifications in New York City

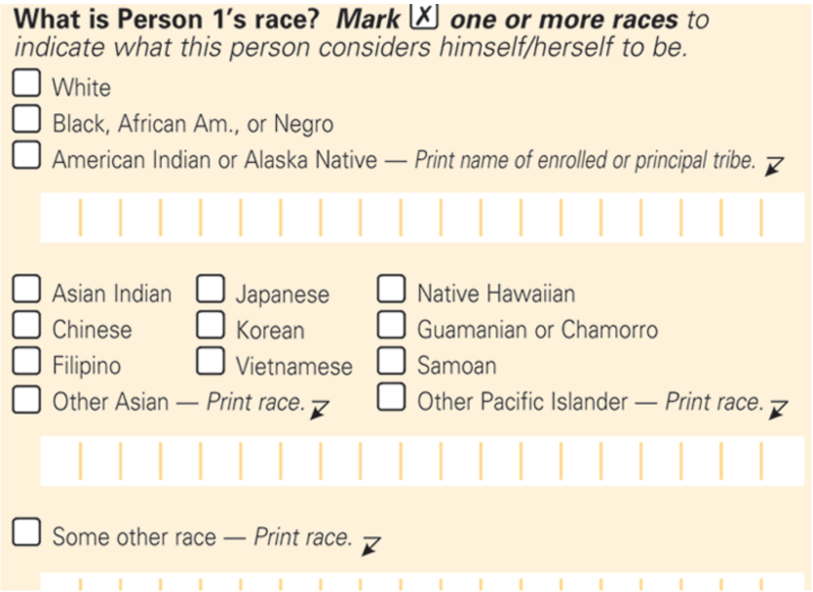
NYC Transit trains are "on time" if they arrive within three minutes of their scheduled time

"Response time" to a fire is the time it takes for the first emergency vehicle to arrive

New York Knicks treat "no shows" of season ticketholders in "paid attendance" category

Consolidated Edison talks about "customers" who were without power during a recent blackout

Non-Bibliographic Classification: "Race" and "Ethnicity" in 2000 US Census



What is Person 1's race? ***Mark*** ☒ ***one or more races*** *to indicate what this person considers himself/herself to be.*

□ White
□ Black, African Am., or Negro
□ American Indian or Alaska Native — *Print name of enrolled or principal tribe.*

□ Asian Indian
□ Chinese
□ Filipino
□ Other Asian — *Print race.*

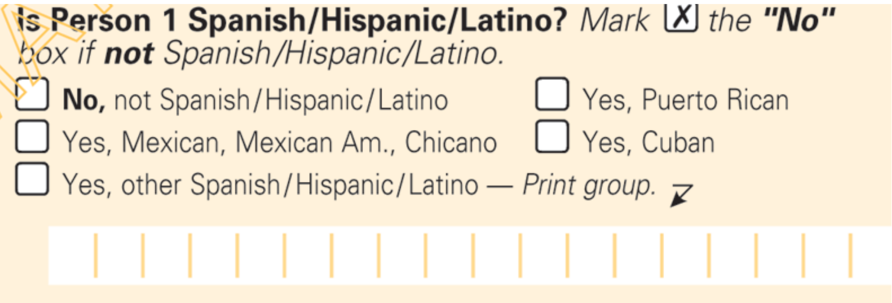
□ Japanese
□ Korean
□ Vietnamese

□ Native Hawaiian
□ Guamanian or Chamorro
□ Samoan
□ Other Pacific Islander — *Print race.*

□ Some other race — *Print race.*

"Spanish Culture or Origin" -- Oops?

Persons "of Spanish culture or origin" are collectively classed as an "ethnic group," regardless of race so they must choose one of the other categories



Is Person 1 Spanish/Hispanic/Latino? *Mark* ☒ *the* ***"No"*** *box if* ***not*** *Spanish/Hispanic/Latino.*

□ **No,** not Spanish/Hispanic/Latino □ Yes, Puerto Rican

□ Yes, Mexican, Mexican Am., Chicano □ Yes, Cuban

□ Yes, other Spanish/Hispanic/Latino — *Print group.*

Changes in US Census Race Classification

1790-1850: White, Black (free), Black (slave)

1850: White, Black (free), Black (slave), Mulatto

1860: add American Indian, Chinese (in California only)

1870: add Japanese

1890 -1920 only: add Quadroon, Octoroon

1910: add Filipino, Hindu, and Korean

1930 only: Mexican

from 1950 on: OTHER classification

Another Changing Classification Scheme

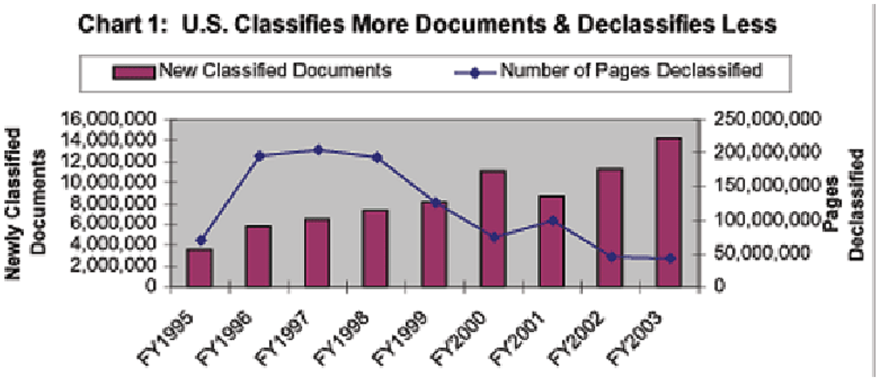


Chart 1: U.S. Classifies More Documents & Declassifies Less

US Government Classification of Documents

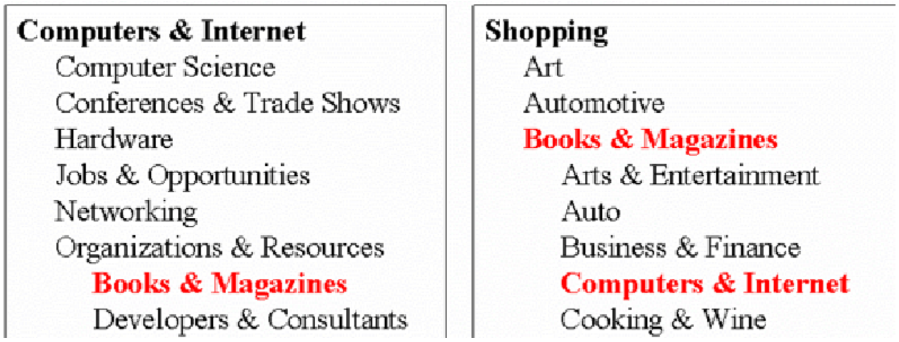
September 22, 2003 Directive from National Archives "Information Security Oversight Office"
that "prescribes a uniform system for classifying, safeguarding, and declassifying national security information
(http://www.archives.gov/isoo/policy-documents/eo-12958-implementing-dir

This Directive sets forth guidance to agencies on original and derivative classification, downgrading, declassification, and safeguarding of classified national security information.

Proliferation and expansion of categories for:

- Sensitive Security Information
- Sensitive Homeland Security Information

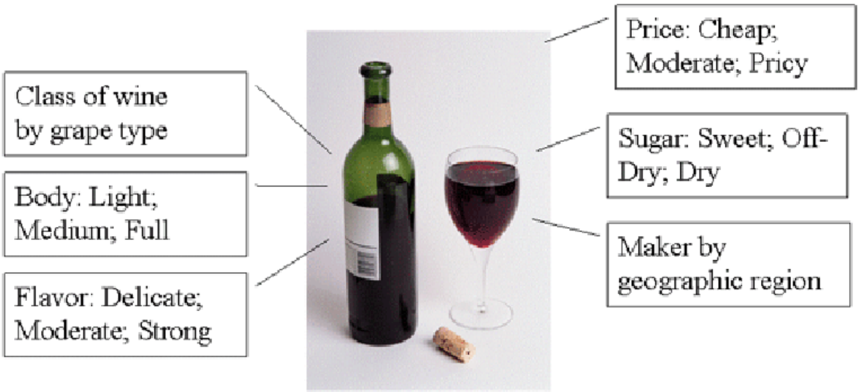
The Need For Multiple Classifications



Computers & Internet
- Computer Science
- Conferences & Trade Shows
- Hardware
- Jobs & Opportunities
- Networking
- Organizations & Resources
  - **Books & Magazines**
  - Developers & Consultants

Shopping
- Art
- Automotive
- **Books & Magazines**
  - Arts & Entertainment
  - Auto
  - Business & Finance
  - **Computers & Internet**
  - Cooking & Wine

Wine Classifications



Class of wine by grape type

Body: Light; Medium; Full

Flavor: Delicate; Moderate; Strong

Price: Cheap; Moderate; Pricy

Sugar: Sweet; Off-Dry; Dry

Maker by geographic region

Wine.com

Faceted Classification

FACETS are an alternative to hierarchical classification that overcome many of its limitations

Instead of defining categories by hierarchy, multi-dimensional categories are (defined or generated) through grammatical (composition or combination) from the (characteristics or dimensions or relations) in the domain

Facets divide a domain or subject into "homogeneous" or "semantically cohesive" categories of manageable size because the terms in a facet have similar referents

Instead of enumerating the possible instances that fit into the classification scheme in a bottom-up way, create a descriptive vocabulary and a grammar that generates all the instances

The categories are distinct and non-overlapping (mutually exclusive)

Facets as a Controlled Vocabulary

The relationships between the facets enable a small CONTROLLED VOCABULARY to generate:

- Many structured descriptions
- That are complex, but formally structured
- That enable us to describe things we don't have words for

Some Examples

Finding a Recipe at Epicurious.com

Product Finders at CDW.com

Facets and User Interfaces

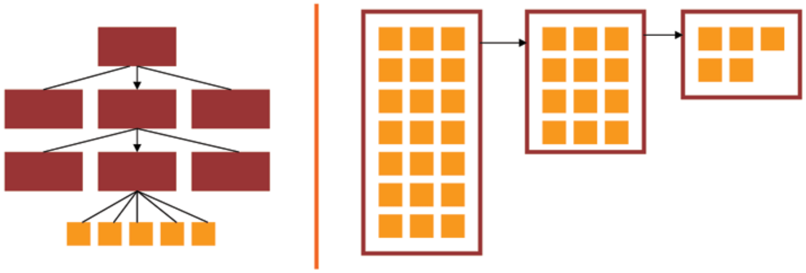
Faceted classification enables very effective user interfaces for browsing and searching in structured collections

It yields an iterative interaction style in which the searcher filters the collection or query results by progressively selecting from only valid values

Results appear at all the branches of a dynamically-generated tree, not just at the leaves of a static one -- and you never have 0 results

Facets and User Interfaces

- Old days: Click-click-click on categories and finally get to the goodies *at the leaf nodes*
- Trend: One click, get a sample of results ("1-10 of 149"), select a category to get fewer results



Types of Facets

Enumerative -- a set of mutually exclusive possible values

Boolean -- yes or no on some dimension

Hierarchical or taxonomic -- organize the instances by logical containment

Spectrum -- numerical attributes on some range, with min and max

Designing a Faceted Classification

1. Collect examples that need to be classified
2. Identify candidates for facets and subfacets by analyzing the examples
3. Order foci within facets
4. Determine grammar for ordering and combining facets and subfacets
5. Create new facets and subfacets where needed
6. Test classification scheme on new examples
7. Iterate and refine throughout

Criteria for Choosing Facets

ORTHOGONALITY - facets are independent dimensions

SEMANTIC BALANCE - top level facets are the most important semantic dimensions of the domain; values within facets are at equal semantic level

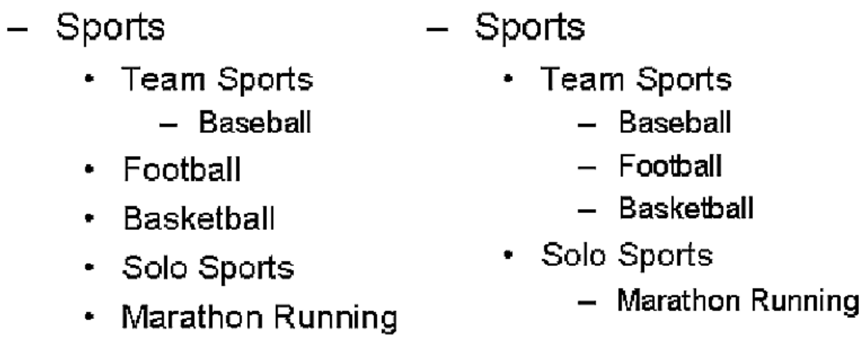
COVERAGE - all current instances can be classified

SCALABILITY - future instances can be classified

OBJECTIVITY - instances can be objectively classified; might also be called CONCRETENESS

NOT IDIOSYNCRATIC - facet semantics should be "mainstream" or "normative" and not rely on clever, fanciful or metaphoric interpretation

Maintain Hierarchical Organization of Values within a Facet



- Sports
  - Team Sports
    - Baseball
  - Football
  - Basketball
  - Solo Sports
  - Marathon Running

- Sports
  - Team Sports
    - Baseball
    - Football
    - Basketball
  - Solo Sports
    - Marathon Running

Principles for Ordering of Facet Values [1]

SIMPLE TO COMPLEX: (Locomotions: walk, run, jump, skip, hurdle, cartwheel)

FREQUENT/POPULAR TO INFREQUENT/UNPOPULAR: (Vegetarian Pizza Toppings: mushroom, onion, olive, artichoke, pineapple, pine nuts)

SPATIAL, GEOGRAPHICAL, OR GEOMETRIC: (Southwestern States: California, Nevada, Arizona, New Mexico)

CHRONOLOGICAL OR HISTORICAL: (Dinosaur Eras: Triassic, Jurassic, Cretaceous)

Principles for Ordering of Facet Values [2]

ALPHABETICAL: (Boy's Names: Al, Bob, Chuck, David, Ed, Frank, George, Harry)

SIZE: (T-Shirts: Small, Medium, Large, XL, XXL)

CANONICAL (even if arbitrary): (Playground Counting: Eenie, Meenie, Mynee, Mo)

FacetMap

FacetMap lets you try out a faceted classification for wine

It also lets you create your own facet map and see different user interface representations of your classification model

You'll do this in your next assignment, due on Monday 6 October

Origins of Faceted Classification

Condorcet, an 18th century French mathematician, proposes a "technical method for discovering the general relationship between the facts from any point of view" that has 5 categories of 10 terms each -> 10^5 = 100,000 combinations

S. R. Ranganathan, a Hindu mathematician working as a librarian, introduced facets to information science in the early 20th century

Ranganathan felt it was his desire and dharma to describe the entire universe of ideas using a single system of classification and notation that:

- Systematically describes, in detail, the contents of complex documents discussing compound subjects
- Codifies those descriptions into a sequenced numerical form

Ranganathan's Colon Classification

5 universal facets applied in fixed order to all things (PMEST):

P (ersonality) - the type of thing

M (atter) - the constituent material of the thing

E (nergy) - the action or activity of the thing

S (pace) - where the thing occurs

T (ime) - when things occur

Colon Classification - Example

CC expresses a classification using these 5 grammatical categories separated by punctuation marks

The values for each category can come from tables or schedules, so the complete expression is very compact

[P]:[E].[S]'[T] -- X62:8:44'N5 -- Management of Banks in India up to 1950

- X = economics
- 62 = banks
- 8 = management
- 55 = India
- N5 = 1950

Facets in the Library of Congress Subject Headings

CC not used much anymore, because people rejected the idea that every domain could be described using the same facets, but it strongly influenced contemporary classification schemes, most notably the LCSH

LCSH uses facets for Topic, Place, Time, and Form (but they can be ordered in a variety of ways, not as rigidly as PMEST

(Topic Main Heading - Place - Topic - Time - Form) Art criticism - France - Paris - History - Nineteenth Century - Bibliography

(Topic Main Heading - Topic - Place - Time - Form)Art - Censorship - Europe - Twentieth Century - Exhibitions

Readings for IO & IR Lecture #9

Steve Pepper, "The TAO of Topic Maps: Finding the Way in the Age of Infoglut"

W3C Recommendation, RDF Primer, through section 2.2

W3C Recommendation, OWL Web Ontology Language Use Cases and Requirements, Sections 1 and 2