Plan for Today's Lecture(s)

- MITRE Reference Model for Understanding and Comparing Standards (holdover from 10/22)
- Bibliographic Classification
- Faceted Classification
- Taskonomic Classification

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
“Information Organization & Retrieval”
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

22 October 2013
Lecture 16.4 – MITRE Reference Model for Understanding and Comparing Standards

MITRE Reference Model for Understanding and Comparing Standards

- "We need to proactively produce new areas of useful semantic agreement, and not simply document correspondences among existing systems"
- "An approach to semantics management must tolerate organizational realities that are often ignored“
- Rosenthal, A., Seligman, L., and Renner, S. 2004. From semantic integration to semantics management: case studies and a way forward. SIGMOD Rec. 33, 4 (Dec. 2004), 44-50

Lessons From Standards Making: The "Person-Concept" Tradeoff

- Semantic agreement comes at a cost driven by the number of people who require a shared understanding, and by the number of concepts they must all understand

- So a small set of people can agree on a complex standard, or a large set of people can agree on a simple one

- …especially when participants are just trying to agree on how to describe pre-existing shared concepts rather than having to define them first

Lessons from Standards Making: Incentives and Disincentives

- Approaches that require perfect coordination and altruism are of no practical interest
- Disincentives to agree on semantics arise if agreement means that someone has to change an implementation and pay the cost of doing so
- Model-driven development tools that generate needed software artifacts (e.g., system and user interfaces) from specifications can encourage standards adoption by reducing the transition costs

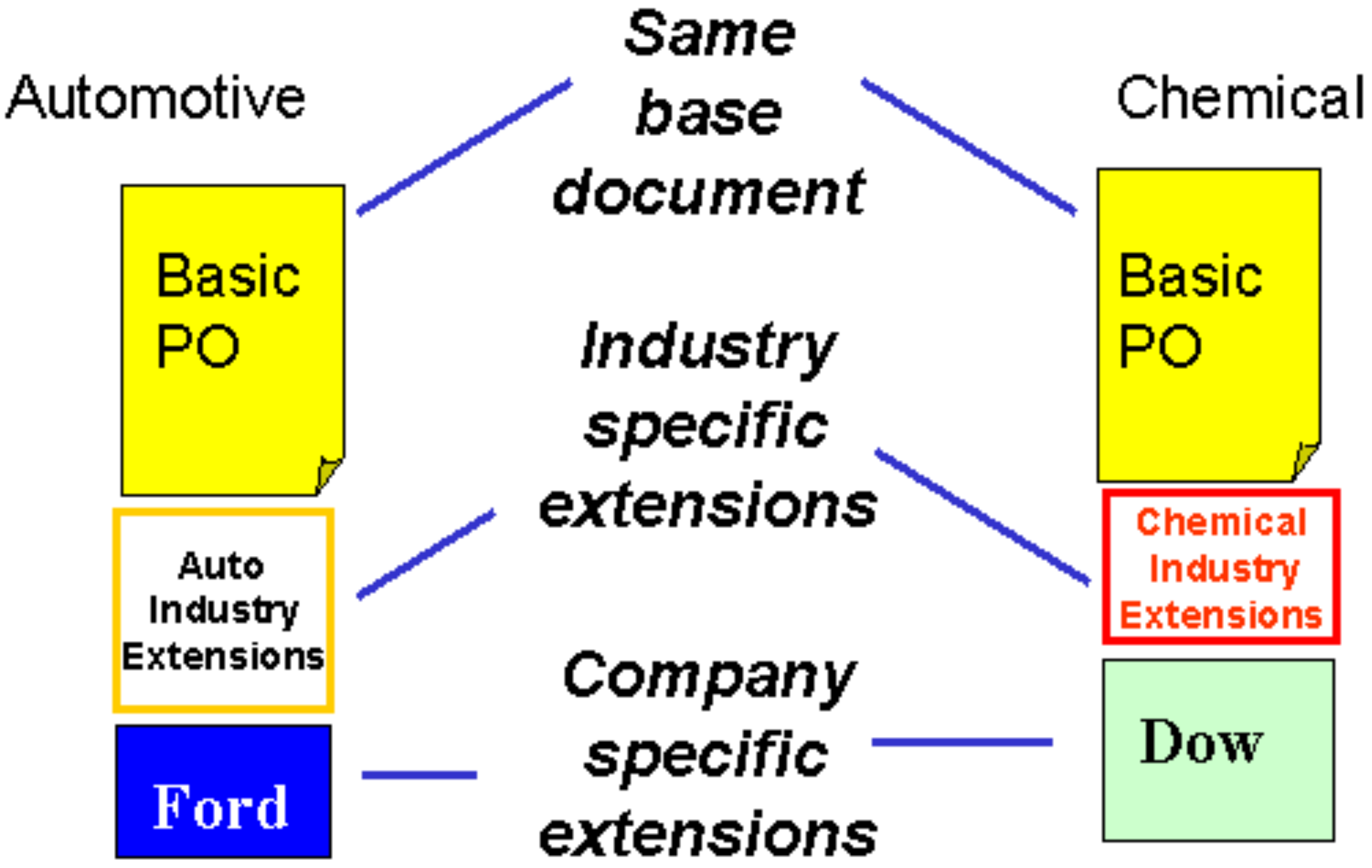
MITRE Reference Model for Comparing Standards: Data Objects

- What is the semantic granularity of the concepts being standardized?
- Is there also a standard for how the concepts are encoding in some syntactic or physical representation?
- Does the standard also specify "instance sets" or possible values for each data element concept?

MITRE Reference Model for Comparing Standards: Structures

- In addition to standards for data objects, are there standards for schemas or "document architectures" that structurally organize them?
- Are there standards for the instances or interchange formats used by publishers/producers or expected by subscribers/consumers?

Horizontal and Vertical Document Components



Automotive

Chemical

Same base document

Industry specific extensions

Company specific extensions

Basic PO

Basic PO

Auto Industry Extensions

Chemical Industry Extensions

Ford

Dow

The Universal Business Language (UBL)

- Over 100 document types built using core components and a common architecture/metamodel needed for supply chains (European) and International Trade (Asia and US), collaborative planning, forecasting, and replenishment; vendor managed inventory; intermodal freight management; and utility billing
- These document types use a library of XML schemas for reusable aggregate data components such as "Address," "Item," and "Payment“

MITRE Reference Model for Comparing Standards: Community Characteristics

- Is there a primary stakeholder with decision making authority, or is authority distributed?
- What are participants' obligations to support the standard?
- Do the participants already share an understand of the domain to be standardized?

Lessons from Standards Making: Enterprise Data Standards

- Standards-making can be successful when a "single authority exercises effective control over the system requirements, funding, the developers, and the users"
- But a very large enterprise cannot hope to construct a single data model (or even a single-set of universally understood concept definitions) for all the data it requires

Lessons from Standards Making: "Communities of Interest"

- Standards making is best organized around naturally formed communities of interest rather than "org chart" organizations
- Different types of communities might be needed to develop, deploy, and maintain a standard

UNIVERSITY OF CALIFORNIA, BERKELEY
SCHOOL OF INFORMATION

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

24 October 2013
Lecture 17.1 – Bibliographic Classification

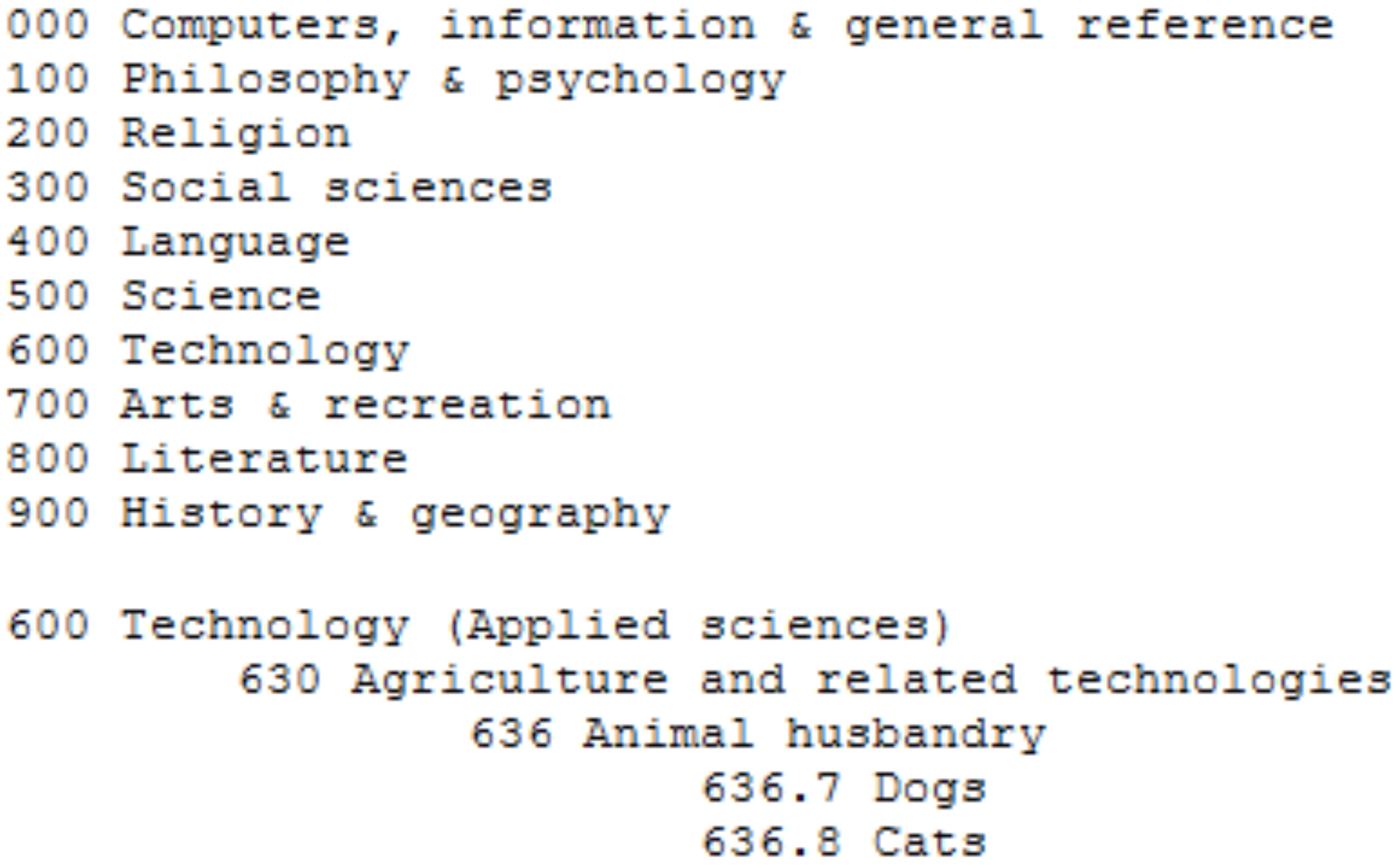
Bibliographic Classification

- Much of our thinking about classification systems comes from the bibliographic domain, which is distinctive because of:
  - Scale, complexity, and degree of standardization
  - Standards and rules for classification
  - Legacy of physical arrangement, user access, circulation

Dewey Decimal Classification

- Started in 1876, the DDC is the most widely used classification system in the world, especially in public libraries
- Easy to use to locate resources in libraries because of its numerical notation
- The DDC is proprietary and must be licensed from OCLC in Dublin OH

Dewey Decimal Classification



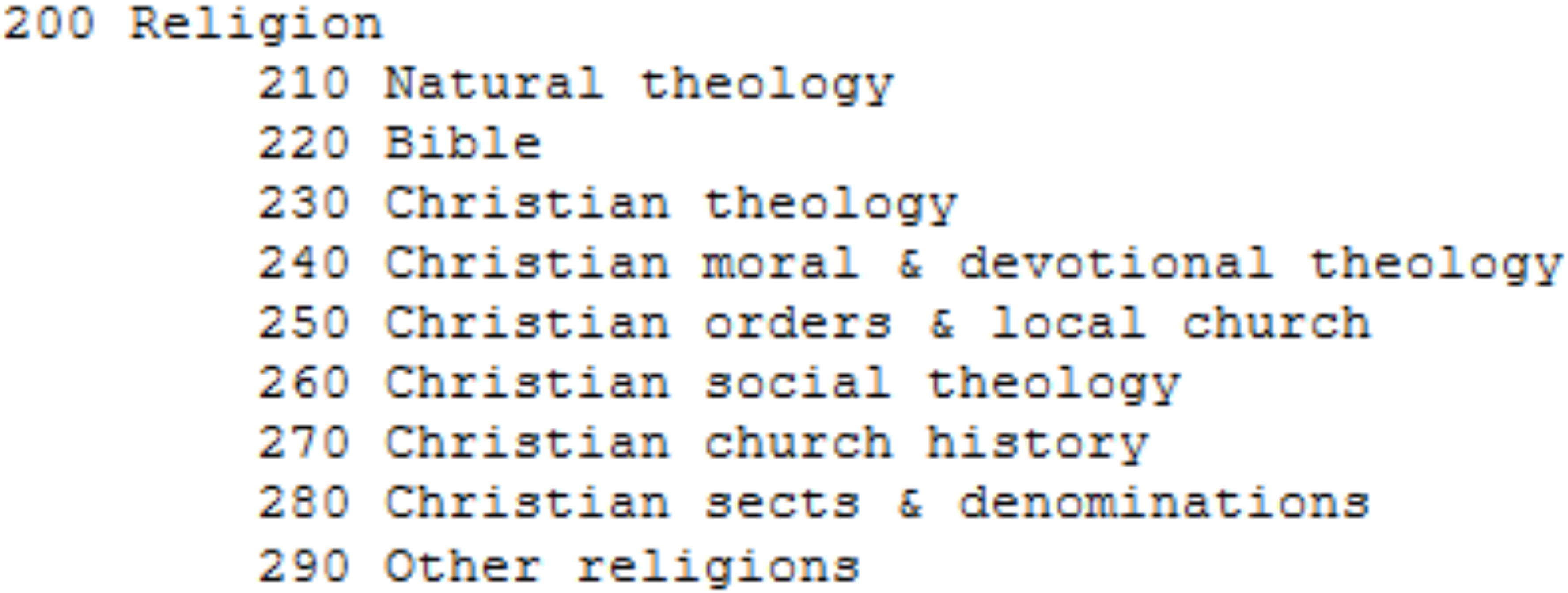
```
000 Computers, information & general reference
100 Philosophy & psychology
200 Religion
300 Social sciences
400 Language
500 Science
600 Technology
700 Arts & recreation
800 Literature
900 History & geography

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

Dewey Decimal Classification

- The DDC is divided into ten main classes, which together cover the entire world of knowledge. Each main class is further divided into ten divisions, and each division into ten sections
- Melvil Dewey wanted to develop a universal classification, but it was done in the context of the library of Amherst College, which introduced significant bias because of the contents of its collection and school's religious orientation at the time

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
```

Dewey Religion Browser hides the Christian bias

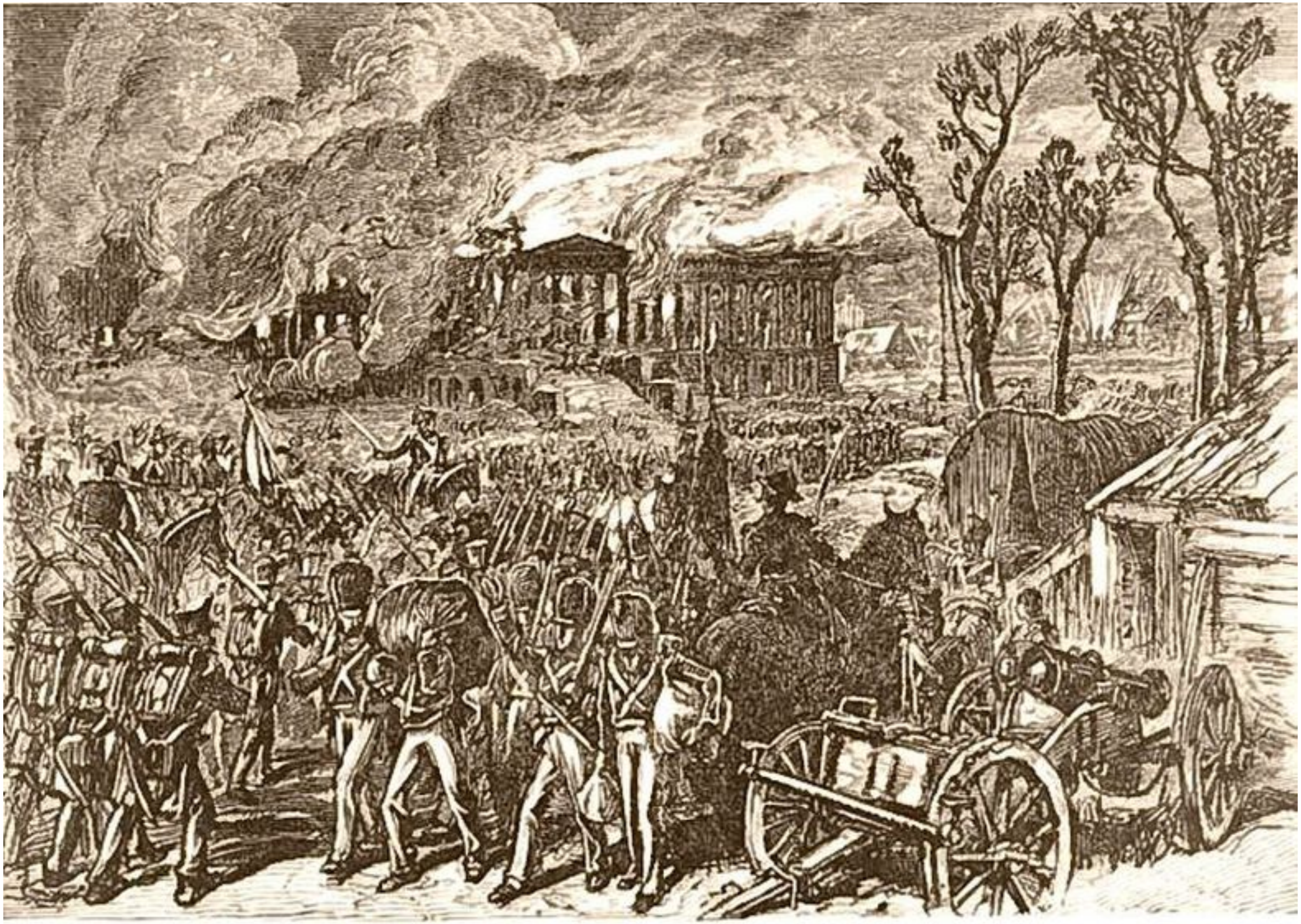
DDC and the "One and Only One Place" Rule

- 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

- The US Library of Congress was established in 1800 with a very practical focus, with no intent to devise an organizing scheme that could be used elsewhere
- It got off to a bad start when the British burned it down (along with the White House) during the "War of 1812" in 1814
- The library was restarted with Thomas Jefferson's personal library, which broadened its scope, influenced the subsequent LCC scheme

British Burn the White House (1814)



CAPTURE AND BURNING OF WASHINGTON BY THE BRITISH, IN 1814.

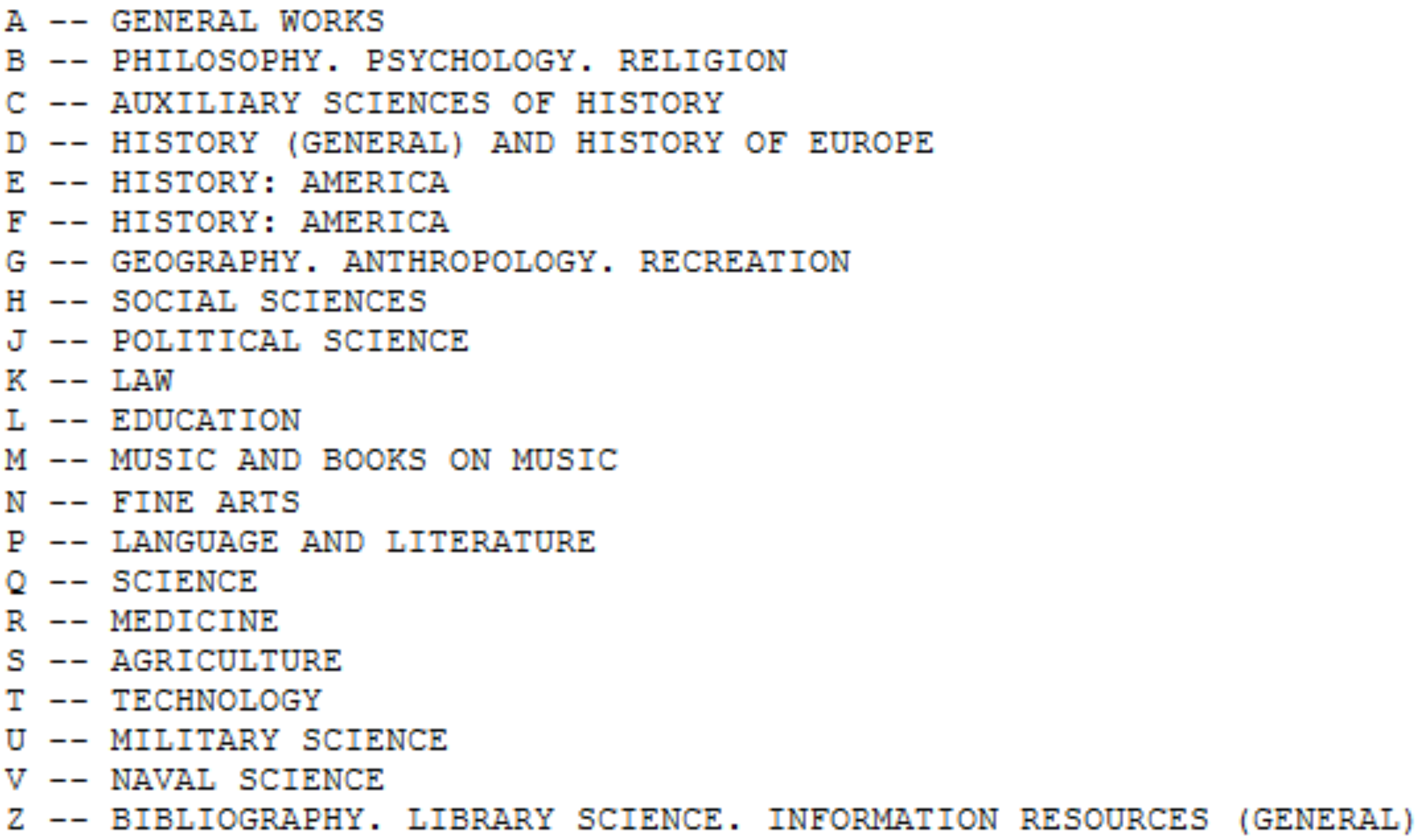
An Important Distinction

- The LCC is not the same as the LCSH, the controlled vocabulary for subject description
  - The LCSH is used to describe individual resources
  - The LCC positions them in a collection
  - These two activities are distinguished in LIS education and practice
- In non-library organizing contexts, describing resources and organizing them are much more intertwined activities

Library of Congress Classification

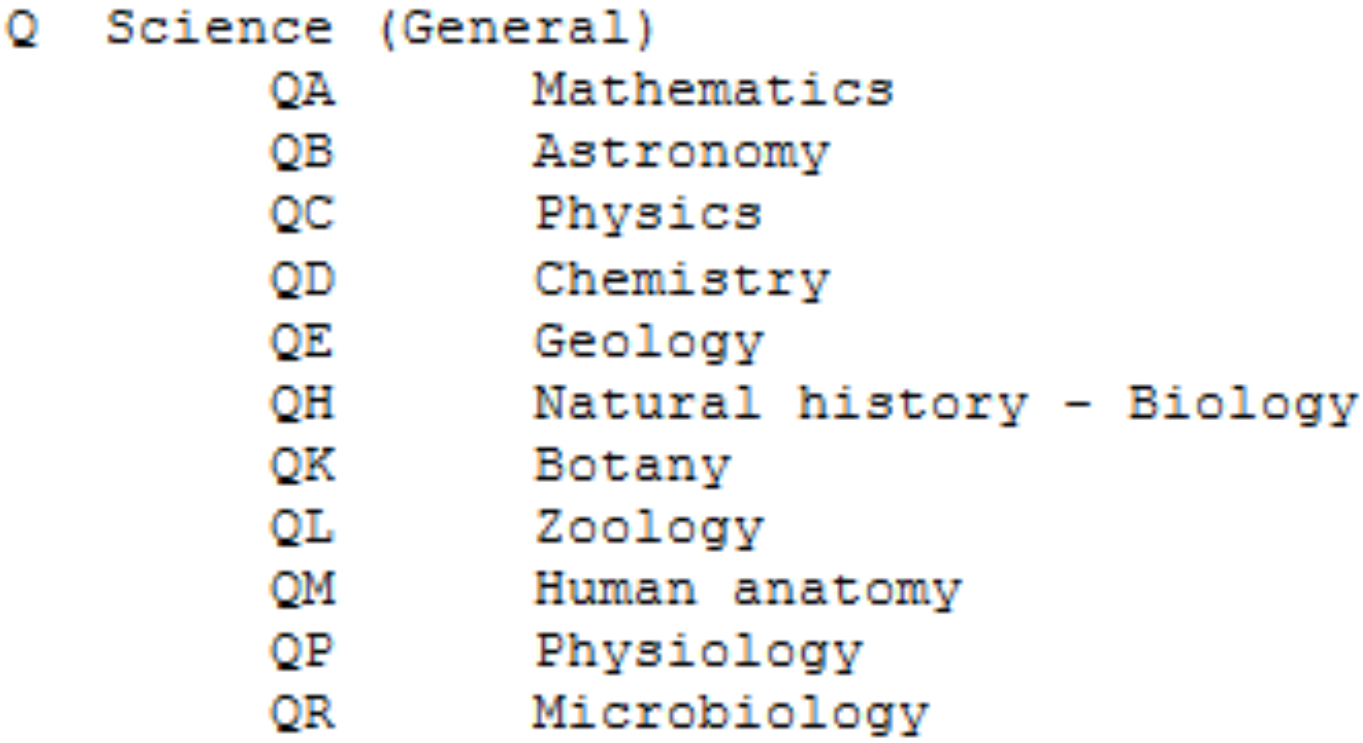
- The LCC has 21 top level categories, identified by (arbitrary?) letters instead of numbers like the DCC
- Each top level category is further divided, and then once again… a very deep system that makes it practical but not theoretically grounded
- The LCC is biased toward the US and toward the needs of a national government, and definitely shows its age

Library of Congress Classification



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?



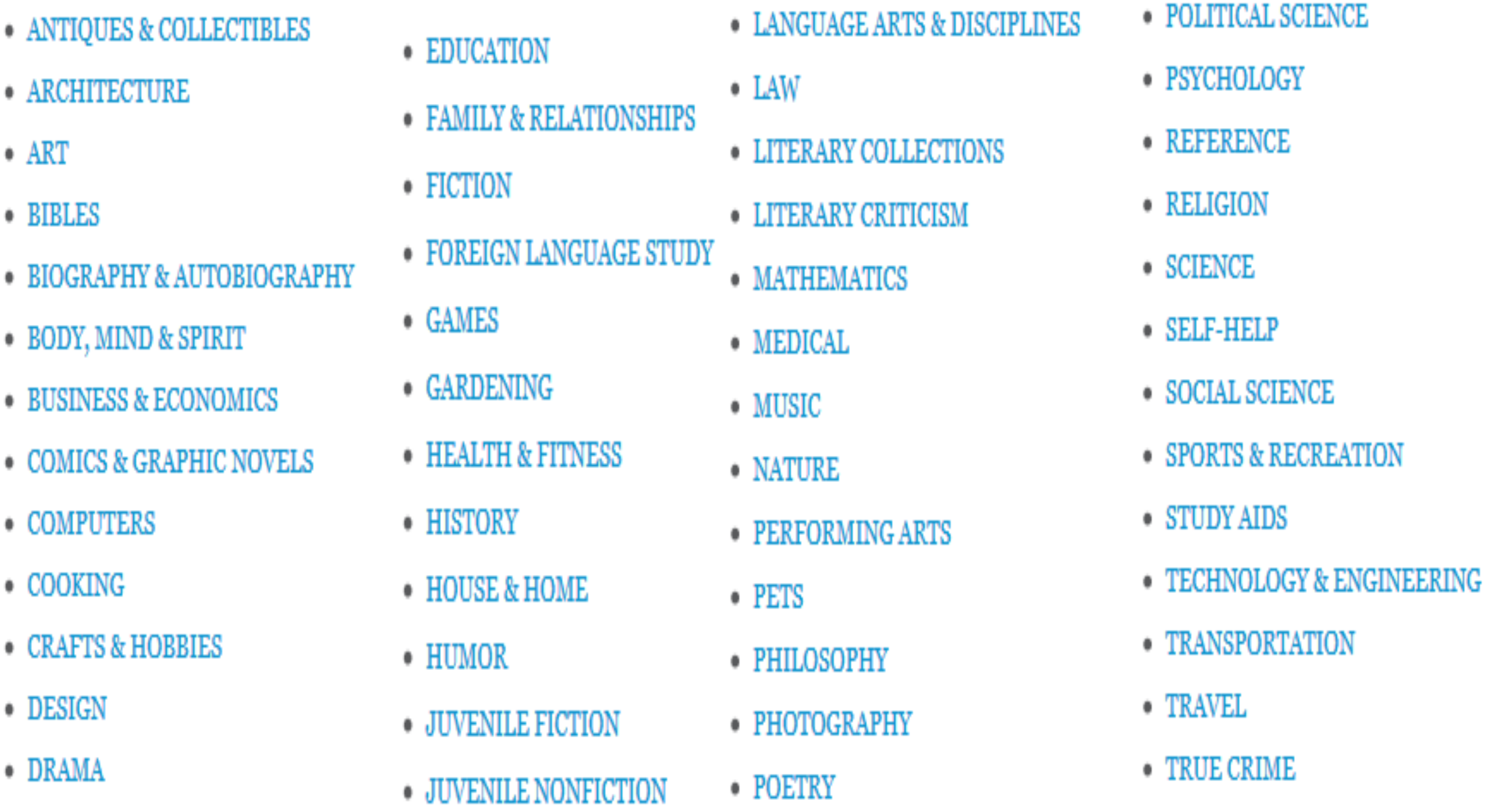
Q Science (General)

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

BISAC

- The Book Industry Standards Advisory Committee Classification (BISAC) differs substantially in intent and design from the DDC and LCC schemes
- BISAC is developed by the Book Industry Study Group, an organization with a business efficiency focus (logistics, marketing)
- BISAC is used by publishers to suggest how a book should be classified in physical and online bookstores, so its categories are biased toward common language usage and popular culture

Top Level BISAC Categories



- ANTIQUES & COLLECTIBLES
- ARCHITECTURE
- ART
- BIBLES
- BIOGRAPHY & AUTOBIOGRAPHY
- BODY, MIND & SPIRIT
- BUSINESS & ECONOMICS
- COMICS & GRAPHIC NOVELS
- COMPUTERS
- COOKING
- CRAFTS & HOBBIES
- DESIGN
- DRAMA
- EDUCATION
- FAMILY & RELATIONSHIPS
- FICTION
- FOREIGN LANGUAGE STUDY
- GAMES
- GARDENING
- HEALTH & FITNESS
- HISTORY
- HOUSE & HOME
- HUMOR
- JUVENILE FICTION
- JUVENILE NONFICTION
- LANGUAGE ARTS & DISCIPLINES
- LAW
- LITERARY COLLECTIONS
- LITERARY CRITICISM
- MATHEMATICS
- MEDICAL
- MUSIC
- NATURE
- PERFORMING ARTS
- PETS
- PHILOSOPHY
- PHOTOGRAPHY
- POETRY
- POLITICAL SCIENCE
- PSYCHOLOGY
- REFERENCE
- RELIGION
- SCIENCE
- SELF-HELP
- SOCIAL SCIENCE
- SPORTS & RECREATION
- STUDY AIDS
- TECHNOLOGY & ENGINEERING
- TRANSPORTATION
- TRAVEL
- TRUE CRIME

The “Dewey Dilemma”

- Some new public libraries are adopting BISAC rather than DDC; their patrons like this (“Use Warrant”) but traditional librarians see this as heresy and selling out to commercial interests

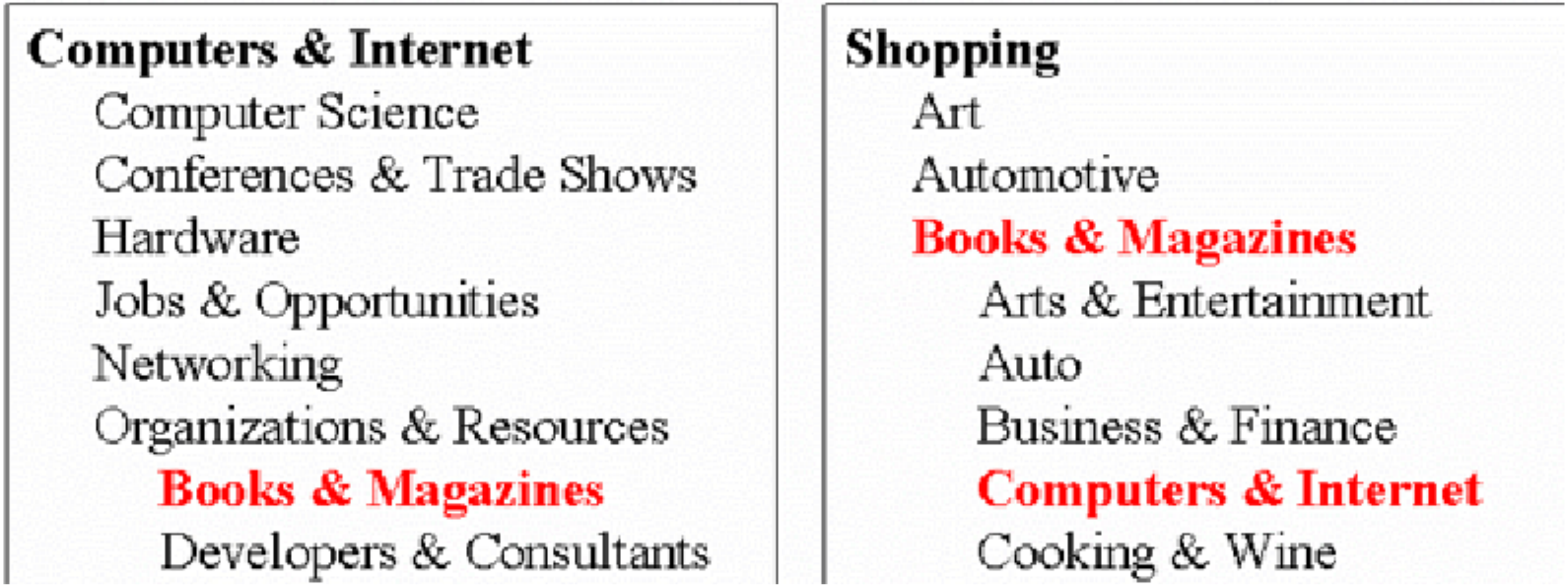
> “The books everywhere, but especially in the children’s room, have been shelved, labeled, and organized in a way that makes me feel less like a moron and more empowered to find what I’m looking for on my own.”

INFO 202
"Information Organization & Retrieval"
Fall 2013

Robert J. Glushko
glushko@berkeley.edu
@rjglushko

24 October 2013
Lecture 17.2 – Faceted Classification

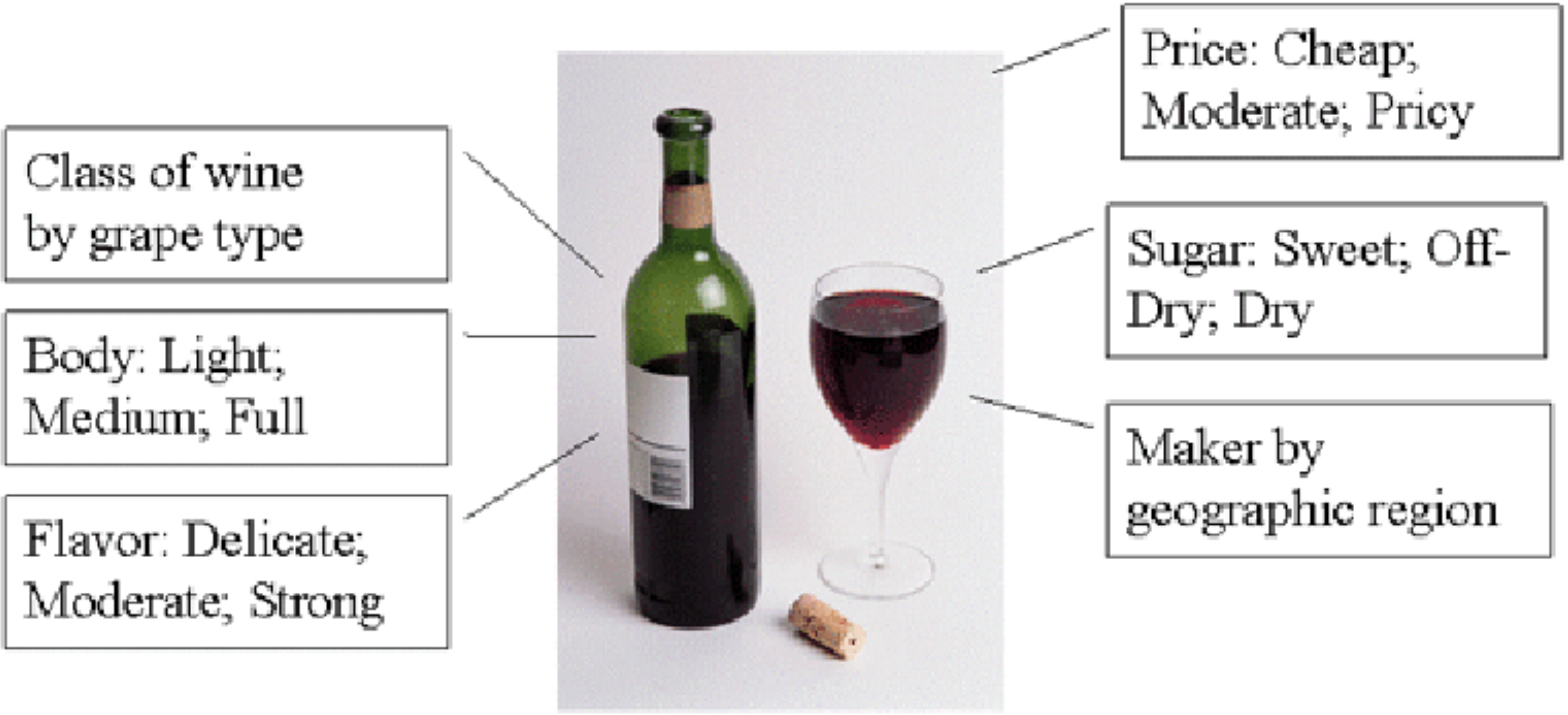
The Need For Multiple Classifications



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

Faceted Classification

- FACETS are an alternative to hierarchical classification that overcome many of its limitations
- Instead of creating a deep category hierarchy, facets create multi-dimensional categories that 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

Facets as a Controlled Vocabulary

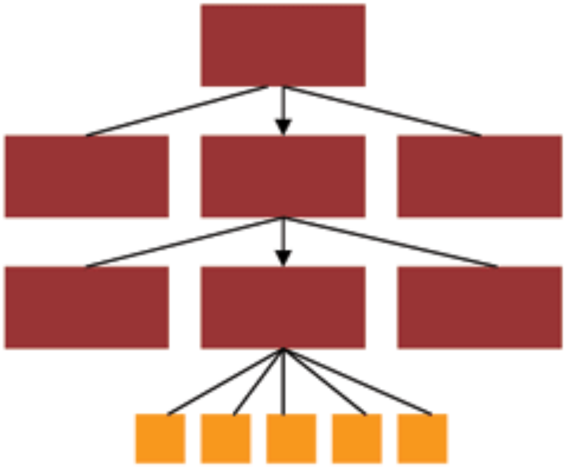
- The relationships between the facets enable a small CONTROLLED VOCABULARY to generate:
  - Many structured descriptions
  - That are complex, but formally structured
  - That enable us to describe things we don't have words for

Marti Hearst and Faceted UIs

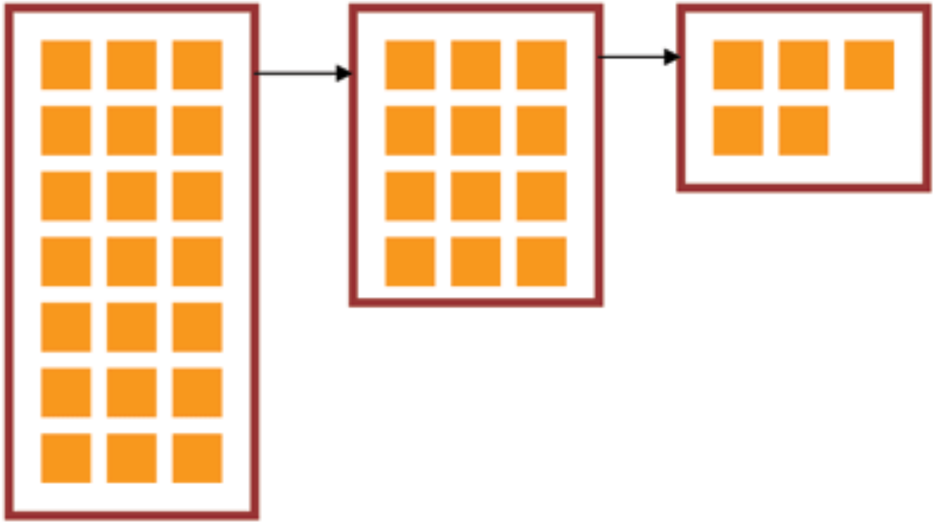
- Professor Marti Hearst did much of the fundamental research on faceted user interfaces for search and navigation in large resource collections
- Section 8.6 of "Search User Interfaces"
- "
  Faceted metadata for image search and browsing." In Proceedings of the SIGCHI conference on Human factors in computing systems, pp. 401-408. ACM, 2003.

Facets and User Interfaces

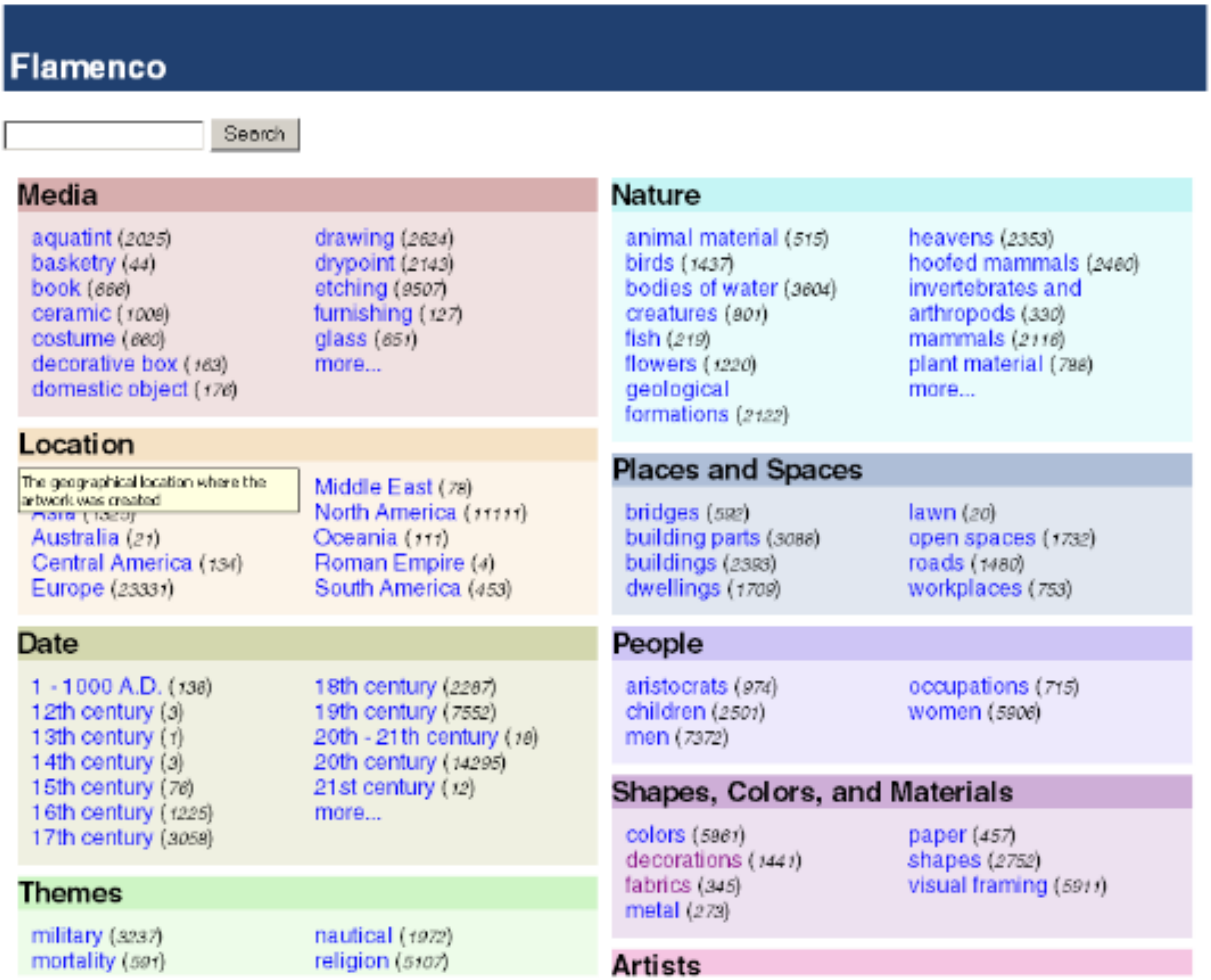
Sequential category selection, results only at leaf nodes



Each selection yields results, subsequent selections refine the results set



Facets in “Flamenco”



Media

aquatint (2025)
basketry (44)
book (666)
ceramic (1008)
costume (660)
decorative box (163)
domestic object (176)
drawing (2624)
drypoint (2143)
etching (9507)
furnishing (127)
glass (651)
more...

Location

The geographical location where the artwork was created

Asia ([illegible])
Australia (21)
Central America (134)
Europe (23331)
Middle East (78)
North America (11111)
Oceania (111)
Roman Empire (4)
South America (453)

Date

1 - 1000 A.D. (136)
12th century (3)
13th century (1)
14th century (3)
15th century (76)
16th century (1225)
17th century (3058)
18th century (2287)
19th century (7552)
20th - 21th century (18)
20th century (14295)
21st century (12)
more...

Themes

military (3237)
mortality (591)
nautical (1972)
religion (5107)

Nature

animal material (515)
birds (1437)
bodies of water (3604)
creatures (801)
fish (219)
flowers (1220)
geological formations (2122)
heavens (2353)
hoofed mammals (2460)
invertebrates and arthropods (330)
mammals (2116)
plant material (788)
more...

Places and Spaces

bridges (592)
building parts (3088)
buildings (2383)
dwellings (1709)
lawn (20)
open spaces (1732)
roads (1480)
workplaces (753)

People

aristocrats (974)
children (2501)
men (7372)
occupations (715)
women (5906)

Shapes, Colors, and Materials

colors (5861)
decorations (1441)
fabrics (345)
metal (273)
paper (457)
shapes (2752)
visual framing (5911)

Artists

Figure 1 of “Faceted Metadata for Image Search and Browsing”

Some Live Examples

- Epicurious
- CDW
- Hearst Museum

Faceted Classification - Condorcet

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

(He's on the Left Bank – Quai Malaquais - looking at the Louvre on the Right Bank of the Seine)

Faceted Classification - Ranganathan

- 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 (PMEST)

- 5 universal facets applied in fixed order to all things:
  - 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

L,45;421:6;253:f.44'N5

Medicine,Lungs;Tuberculosis:Treatment;Xray
:Research.India'1950

Facets in the Library of Congress Subject Headings

- LCSH uses facets for Topic, Place, Time, and Form (but they can be ordered in a variety of ways)
- (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

Designing a Faceted Classification (TDO 7.4.4)

- Collect examples that need to be classified
- Identify candidates for facets and subfacets by analyzing the examples
- Order foci within facets
- Determine grammar for ordering and combining facets and subfacets
- Create new facets and subfacets where needed
- Test classification scheme on new examples
- Iterate and refine throughout

Types of Facets

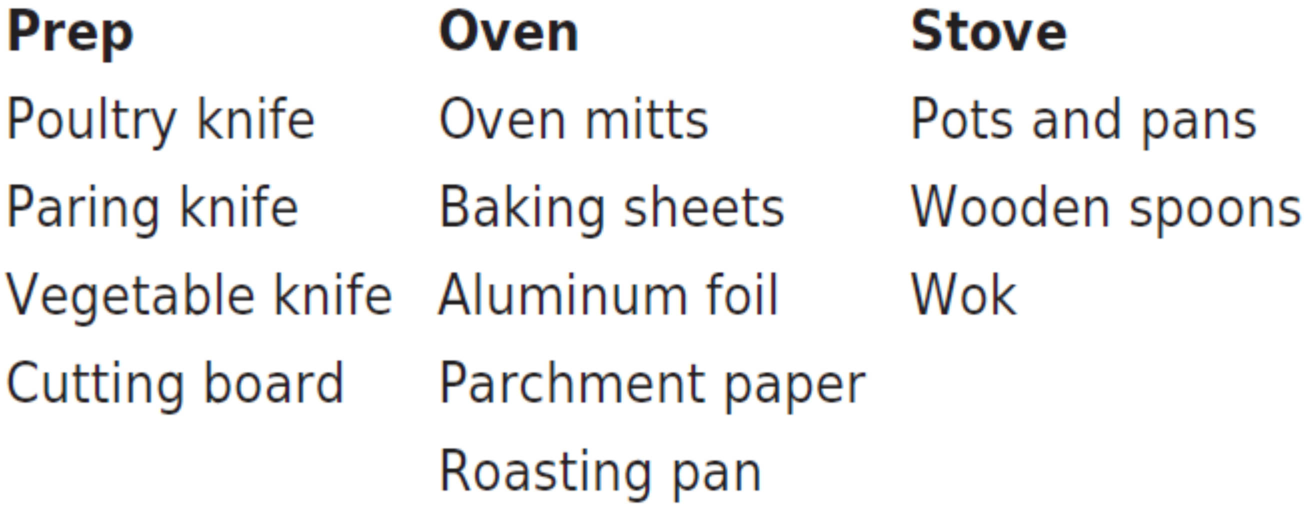
- Enumerative -- a set of mutually exclusive possible values
- Boolean -- yes or no on some dimension
- Hierarchical or taxonomic -- organize the instances by logical containment
- Spectrum -- numerical attributes on some range, with min and max

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

Taskonomy

- In contrast to "taxonomy" - an approach to organization based on similarity of content or characteristics - a "taskonomy" organizes on the basis of purpose or "activity structure“
- A COOK’S TASKONOMY (Figure 7.1 in TDO)



Prep	Oven	Stove
Poultry knife	Oven mitts	Pots and pans
Paring knife	Baking sheets	Wooden spoons
Vegetable knife	Aluminum foil	Wok
Cutting board	Parchment paper	
	Roasting pan	

“Distributed Cognition”

- "Distributed cognition" is a "dialect" of cognitive science that views cognitive activity as not just something that takes place within the mind of a single person a some moment in time; it is:
  - Distributed across the members of a social or goal-oriented group
  - Distributed and coordinated between people and the technology and artifacts they create and use

Hollan, Hutchins, & Kirsh, "Distributed Cognition: Toward a New Foundation for Human-Computer Interaction Research," ACM TOCHI, 2000

“The Intelligent Use of Space”

- "Whether we are aware of it or not, we are constantly organizing and reorganizing our workplace to enhance performance"
- "The physics of the world is such that at times the histories of use are perceptually available to us in ways that support the tasks we are doing"
- "When space is used well, it reduces the time and memory demands of our tasks... to simplify choice, to simplify perception, and simplify internal computation”

Kirsh, David. "The intelligent use of space."
Artificial intelligence 73, no. 1 (1995): 31-68.

Readings for Next Lecture

- TDO 9 through 9.3
- Trant, Jennifer. “Emerging convergence? Thoughts on museums, archives, libraries, and professional training.” Museum Management and Curatorship 24, no. 4 (2009): 369-387
- Williams, Ashley. “User-centered design, activity-centered design, and goal-directed design: a review of three methods for designing web applications.” In Proceedings of the 27th ACM international conference on Design of communication, pp. 1-8. ACM, 2009.