Finding Information

- All about how we now search for and (hopefully) find information and determine (or not!) its veracity. Theories of classification and the revolt against old hierarchical systems. In passing, issues of how information finding systems morph into knowledge navigation and knowledge architecture frameworks. Metadata as the new magic bullet. Tagging as a solution, but a laborious one. The advantages and shortcomings of Web search engines. What we [think] we know about how people try to find information.

Ordering and Finding Information

- Traditions of cataloguing and classif cation
  - Collections, relics, and cabinets of curiousities
  - Diderot and the Tree of Knowledge; Linneas
    - Hierarchies of knowledge
    - Social & cultural hierarchies
  - Melvil Dewey and his system: a single subject… +
    - How many places right of the dot?
    - Problems of scalability
  - LC headings: Dewey made scalable
    - Dilemmas of LC headings: reproducing the cataloguer's thinking
  - Now: f at structures, relations, parsing, semantics
  - Records and the rule of provenance

The Key Problem in "Information Retrieval": Increasing the Signal-to-Noise Ratio

- A consequence of "information overload"
  - Most searches yield oceans of spurious "hits"
- Strategies to attain greater accuracy, better results:
  - Rethink the use of hierarchies
  - Be aware of the Bradford distribution (the 80/20 rule)
  - Shift toward faceted systems, relational aspects
  - Structure with rich metadata structures/fields, then "populate it"
  - Problems in tagging information objects
    - —but who will do the tagging? Our "objects dB" project…
    - is automated tagging a credible option?

Problems of Hierarchies

- Taxonomies, nomenclatures, and controlled vocabularies—who defines these??
  - Buffon, Linneas, and the rage for order: phenetics vs. cladistics
  - Standardizing language: nomenclatures, thesauri, etc.
  - Controlled vocabularies and authority lists
    - LC headings
    - Authority lists as defining proper nouns
    - Controlled vocabularies are broader
- *Nota bene:* Taxonomies do *not* yield new knowledge—they can become conceptual prisons as they order things
- New issues: loadable lists & DTDs via XML + friends

From Hierarchies to Flatness

- The birth of relational databases
- Hierarchic and relational databases compared
  - Parallels to org charts and peer-to-peer
  - Consider EndNote: can we conf dently share our f les?
- Computation power and the ease of the f at search
  - Illusory “hits”
  - Second-guessing non-standard vocabularies and keywords

From Simple to Smart Searches

- Booleans for simple filtering
- Standardizing records
  - Standardized vocabularies
  - OPACs and MARC format; Z39.50
  - EAD and quasi-searchable archival finding aids; CIDOC-CRM
- Data descriptors defined: form vs. content
  - Page languages, SGML, Postscript
  - Metadata
  - From HTML to XML; semantics and auto-tagging
  - Endless need to “metatag” digital objects
  - Is there a politics of tagging? (we know there’s a politics of classifying)

The Compromise: Flat/Relational Systems vs. Hierarchies

- Hierarchies yield better-focused results, and less processing, but require knowledge of the classification system

- Primitive flat systems vastly ease access—no need to "second guess" how hierarchy is ordered, but often yield too many [useless] results

- Faceted searching (Bates): use tagging (read: XML) and metadata to define several "faces" of information objects, then access them with flat searching

Searching for Meanings and Ideas, Not Just Words

- First step: “do you mean…?” —Google
- Web caching: meanings defined by frequencies & contexts of queries, and by analysis of hyperlinks
- Semantics
  - Defined
  - From word-search algorithms to deriving meanings
  - The role of statistics

How Well Do "Finding Aids" Filter Information?

- First, how Google works…
  - Step 1: "spidering," to harvest what's publicly available on the Web—a constant process (and sites can implement "do not spider" rules)
  - Step 2: indexing with word searches, then building rankings based in large part on frequency of links to specif c pages
    - Consequences: a. no validity tests, so "mind-share" equates to "truth;" b. massive parts of the Web don't appear in Google; c. system is more a gathering and ranking than a f lter
- Contrast this to for-pay and specialized databases (OCLC, RLIN-Eureka, Ovid, ProQuest, Dialog, MedLine…)
  - Minimal ranking, as the ordering algorithm is the traditional author/title/subject/[keyword] framework
  - Better general veracity, as content has already been vetted
  - At the same time, access rules and interfaces stink

More on Databases and Search Engines…

- Keep in mind that Google and Yahoo do also maintain canonical (or near-canonical) subject heading indexes…
- They also work closely with the older, more "mainstream" info aggregators such as OCLC and LC's National Union Catalog effort
- Many database services (ProQuest, MedLine, Nexis/Lexis, Gale, etc.) are actually "data aggregators;" they thus inherit the foibles and oddnesses of the originals
- Both Web search engines and higher-end databases redefine "browsing" in ways that are probably not as fruitful as old-fashioned "stack browsing"
- Note well the emergence of "grass-roots" indexing with Wikipedia, del.icio.us: do folksonomies consolidate or diffuse findability frameworks? Do they undermine the utility of thesauri and controlled vocabularies?

An Important Insight…
with startling implications.

- In terms of the development of personally-held knowledge (what and how people learn), research has often indicated that discovery by unintentional means, from browsing to thinking by metaphor, is often quite effective and profound. Consider the implications:
  - Search-engine results may often be too narrow, precluding serendipity
  - The best search-behavior model is recursive: people learn and adapt in an extended, often less-directed fashion
  - “Stack browsing” is often more useful than we would otherwise believe
  - Maybe the best interaction with a reference librarian is “Socratic”…
  - [This is what “Information Searching Behavior” is all about]—current preferred model is “berry-picking”