14. Social / Distributed Categorization

INFO 202 - 15 October 2008

Bob Glushko

Plan for Today's Lecture

"Social/distributed categorization"

the defining examples: flickr and del.icio.us

enterprise applications: fringe and dogear

from Web 2.0 to Web 3.0

Varieties of Categorization Systems - A Reminder

Cultural Categorization Systems (Language and Lakoff)

Individual Categorization ("Tagging")

Institutional Categorization ("Business Semantics")

Individual Categorization Systems

A system developed by an individual for organizing a personal domain to aid memory, retrieval, or usage

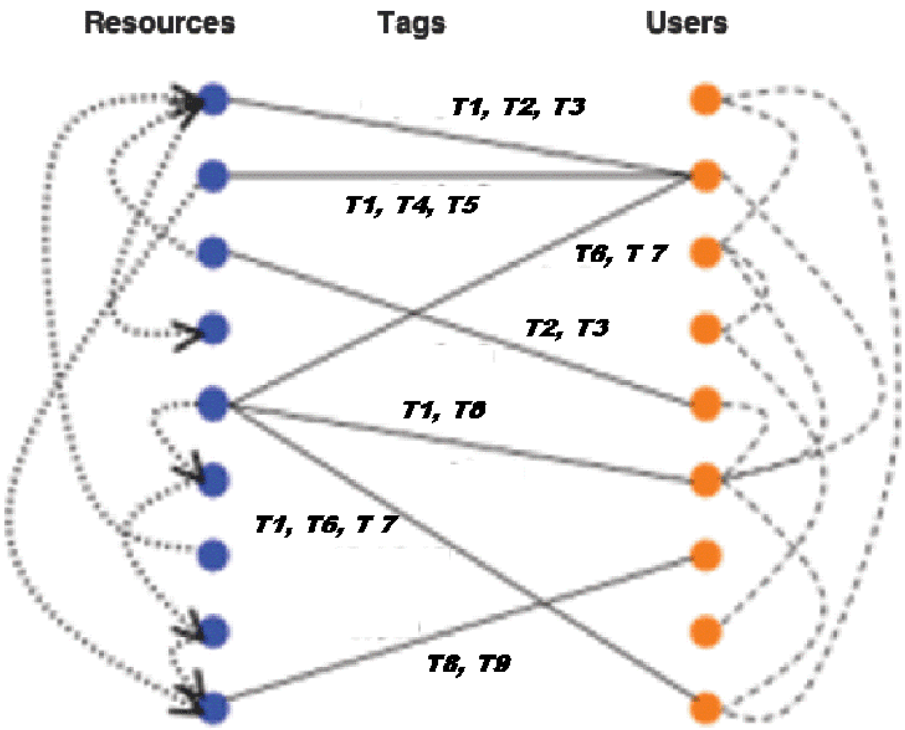
Have exploded with the advent of cyberspace, especially in applications that emphasize "tagging" / "bookmarking" / "annotation"

Why This is "Social"

Even though the tags are assigned by individuals, they can serve social goals to convey information, develop a community, manage reputation

And the outcomes have been described as: collaborative, cooperative, distributed, dynamic, community-based, folksonomic, wikified, democratic, user-assigned, or user-generated

A Conceptual Model of Tagging "Systems"



Some Call it "Classification," But It's Not

Wikipedia's article on "Folksonomy" is typically imprecise:

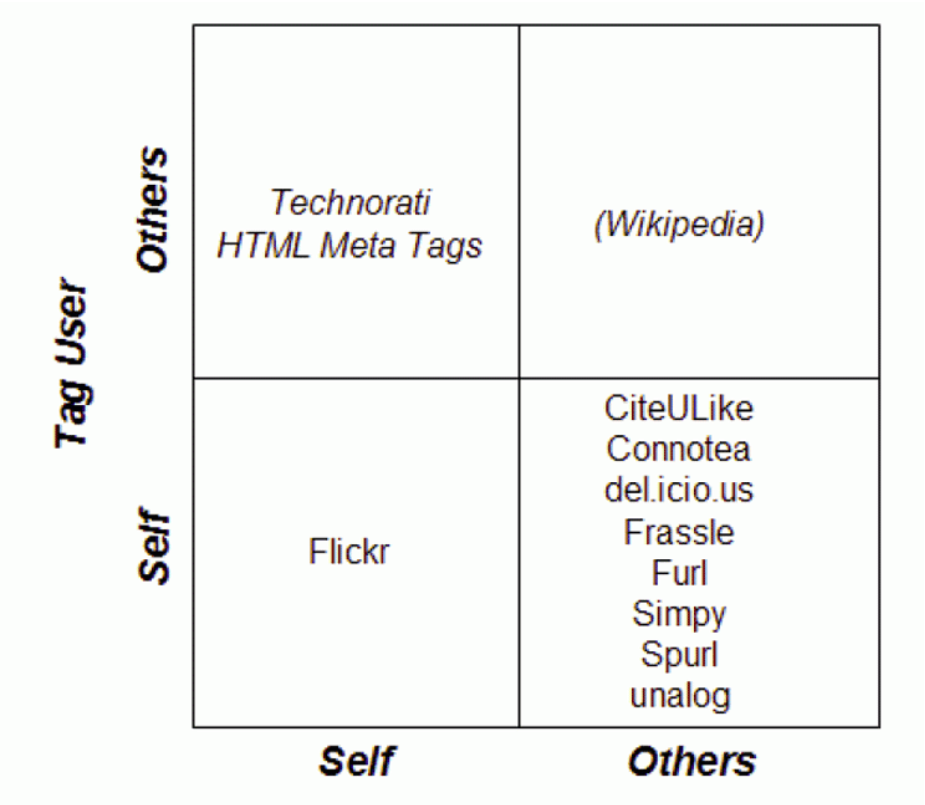
- Folksonomy (also known as collaborative tagging, social CLASSIFICATION, social indexing, and social tagging) is the practice and method of collaboratively creating and managing tags to annotate and CATEGORIZE content. Folksonomy describes the bottom-up CLASSIFICATION systems that emerge from social tagging.

CATEGORIZATION (from September 15) - Categories are equivalence classes - sets of material and abstract things, processes, and events that we treat the same

CLASSIFICATION (from September 24) - 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

Most "end user tagging" systems don't impose any pre-existing system of categories -- indeed, that's the point!

Coarse Classification of Tagging Systems



Design Dimensions for Tagging Systems

What can be tagged? (Anything, photos, web resources, bibliographic entities...)

Source of tag referents? (Global, system, user contributed)

Who can tag? (Self, permissions, anyone)

Tagging support? (None, suggested, previous tags viewable)

Aggregation model? (None, bag, labeled set)

Are tag referents linked?

Are the taggers linked?

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.
- Although it is generally preferable to use named elements which have well-defined semantics for each type of metainformation (e.g. TITLE), this element is provided for situations where strict SGML parsing is necessary and the local DTD is not extensible.
- (http://lists.w3.org/Archives/Public/www-html/1994Jun/0041.html)

The META Tag Specification: HTML 4.01 (12/99)

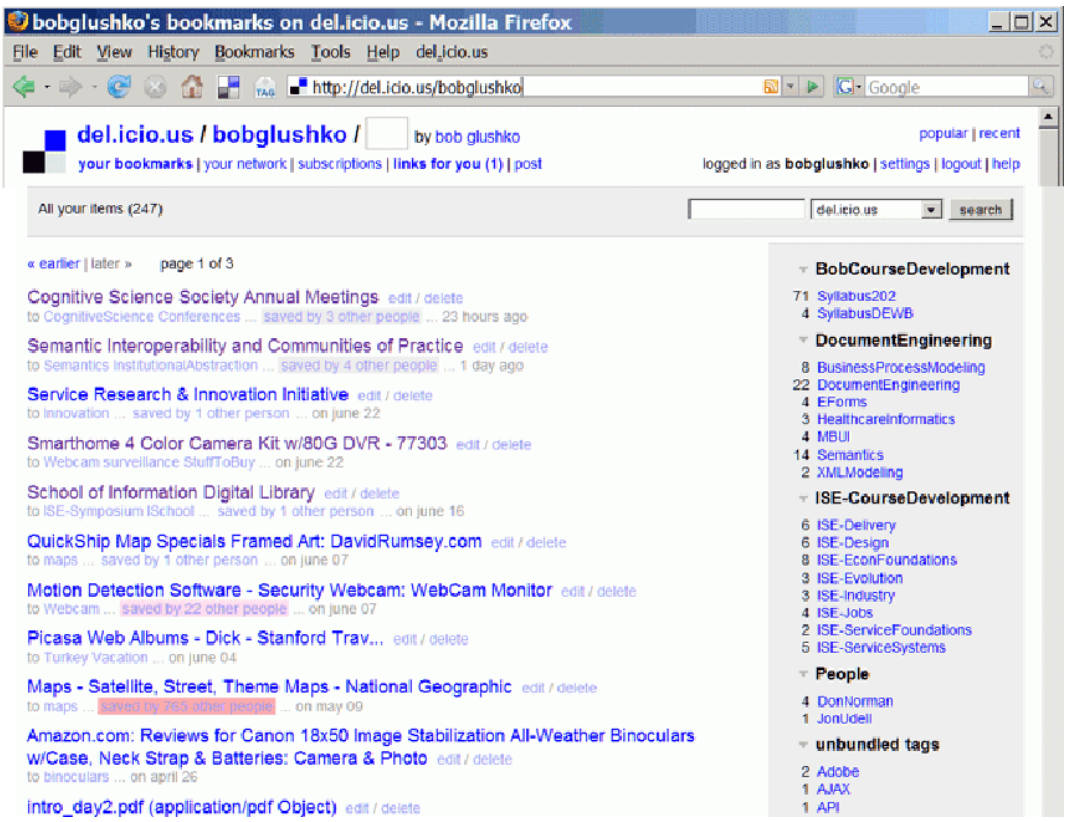
```
<!ELEMENT META - O EMPTY  -- generic metainformation -->
<!ATTLIST META
  %i18n;      -- lang, dir, for use with content --
  http-equiv  NAME  #IMPLIED  -- HTTP response header name  --
  name        NAME  #IMPLIED  -- metainformation name --
  content     CDATA  #REQUIRED -- associated information --
  scheme      CDATA #IMPLIED  -- select form of content --
  >
```

What the W3C imagined:

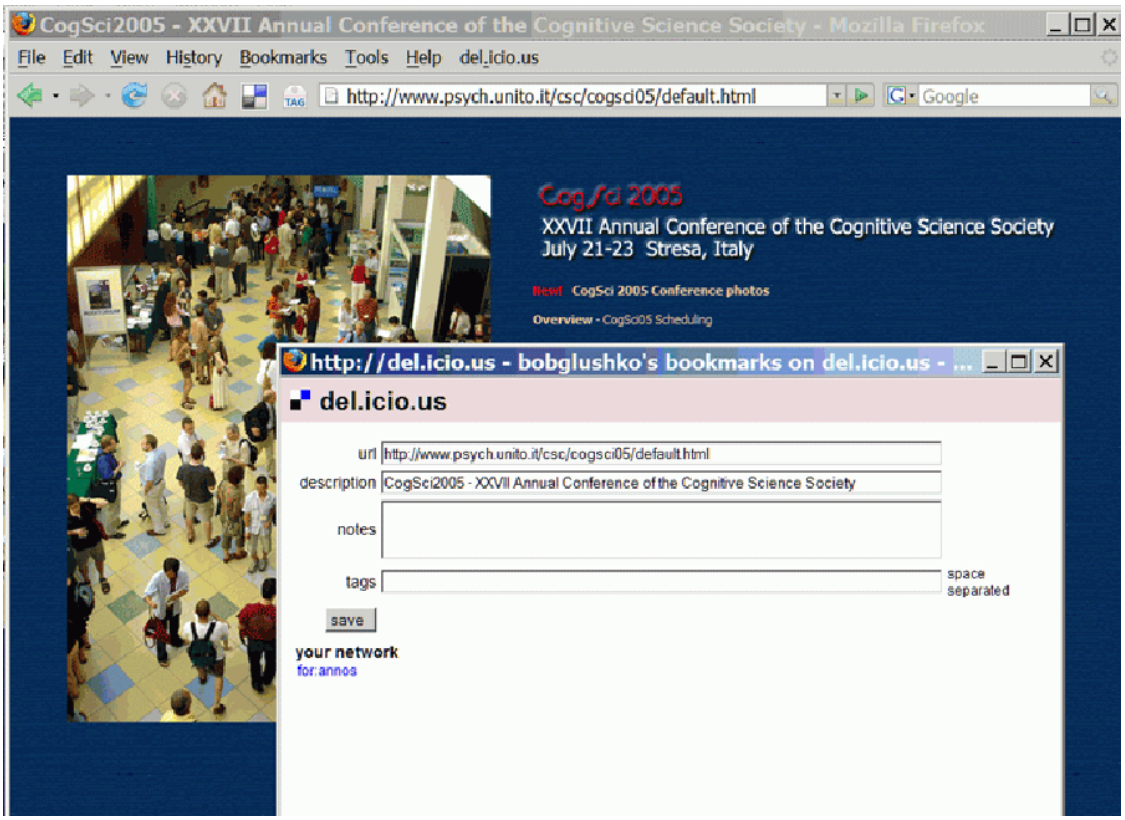
```
<META NAME="DESCRIPTION" CONTENT="accurate prose description">
<META NAME="KEYWORDS" CONTENT="useful comma-separated keywords">
```

And some sites do that... but most don't, and so META is ignored by all search engines

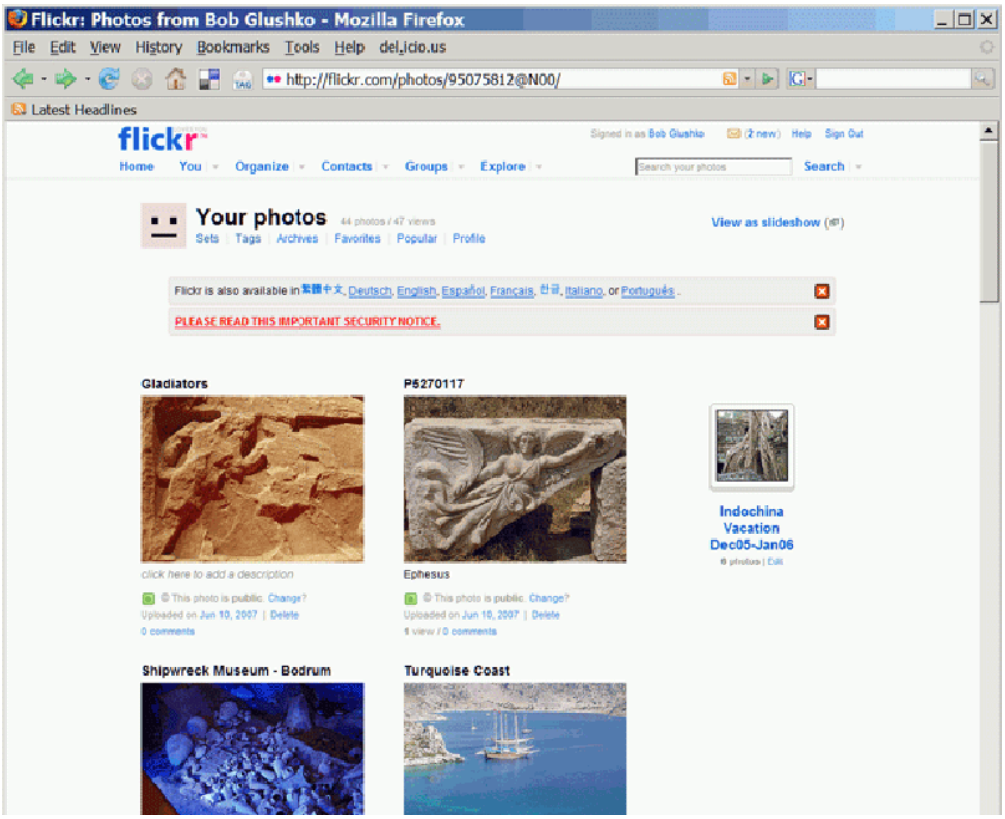
del.icio.us -- Shared Bookmarks



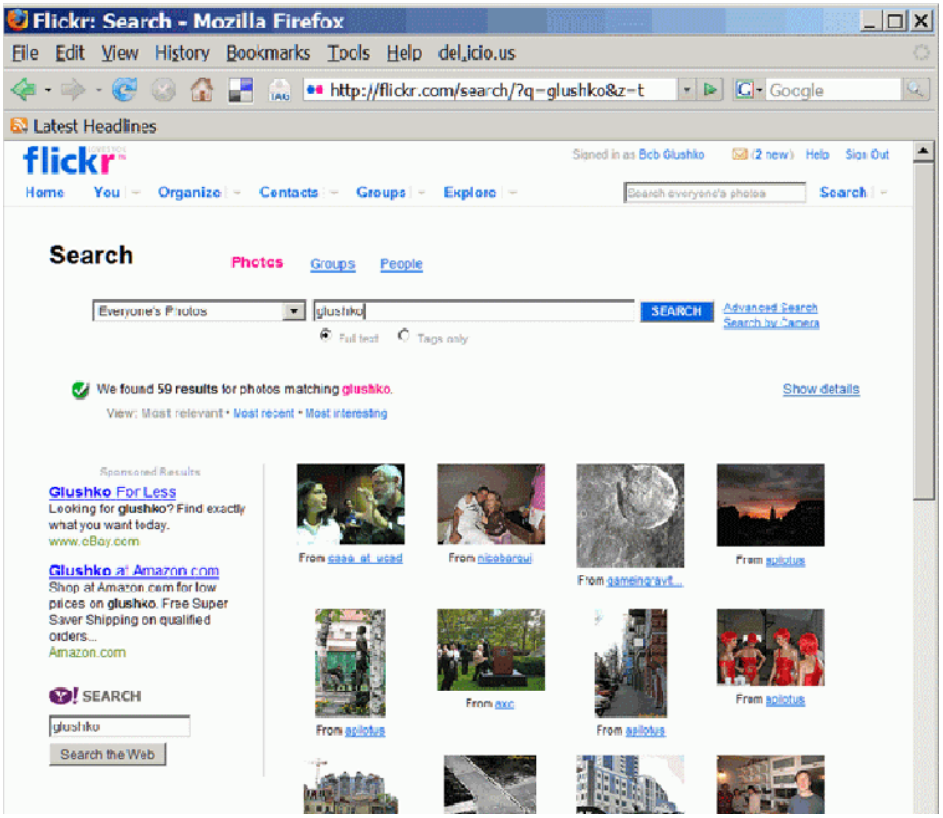
del.icio.us -- User Interface for Tagging



flickr - Photo Collections



flickr - Search for Tag "Glushko"



Tagging Is...

Creative and dynamic

Unconstrained, open-ended

Interpretive

Statistical

Why Tag?

To organize for your own future use

- Content-based organization
- Task-based organization

To enable sharing and communication to known audiences

To express opinions or to entertain

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)

Tagging Functionality / User Interfaces

Context is recorded automatically (tagger, time, date, resource name)

Share/Don't Share (or Private/Public): enable both personal organization and group organization (default is {"public")

Tag suggestion (tagging precedents) -- might be before or after your own tags are applied

Tag organization into groups or categories

Batch uploading and tagging

Tag Visualization ("tag clouds")

del.icio.us "Tag Cloud" for all Tags

del.icio.us "Tag Cloud" for "Doc Or Die" Blog

Tag Quality / Correctness?

The del.icio.us instructions say:

Tagging is intuitive

A tag can be anything you want

There are no wrong tags

Tag Me "Stanford Football" and "BarryBonds"

"Tag Soup"

Users are free to assign any number of labels or tags they choose

No vocabulary control

Responses to Tag Soup

Some people consider the unstructured, uncontrolled nature of "tag soup" to be its great strength, just as faceted classification overcomes some of the limitations of strict hierarchies

Others adopt personal conventions to encode hierarchical and derivational relationships (e.g. using CamelCase; basic and specific level categories)

Using multiple accounts for the same application is another approach for organizing tags and the resources they describe

Some systems are introducing "tag bundles" to enable more hierarchy; it might also be possible to infer the hierarchy using dictionaries or thesauri

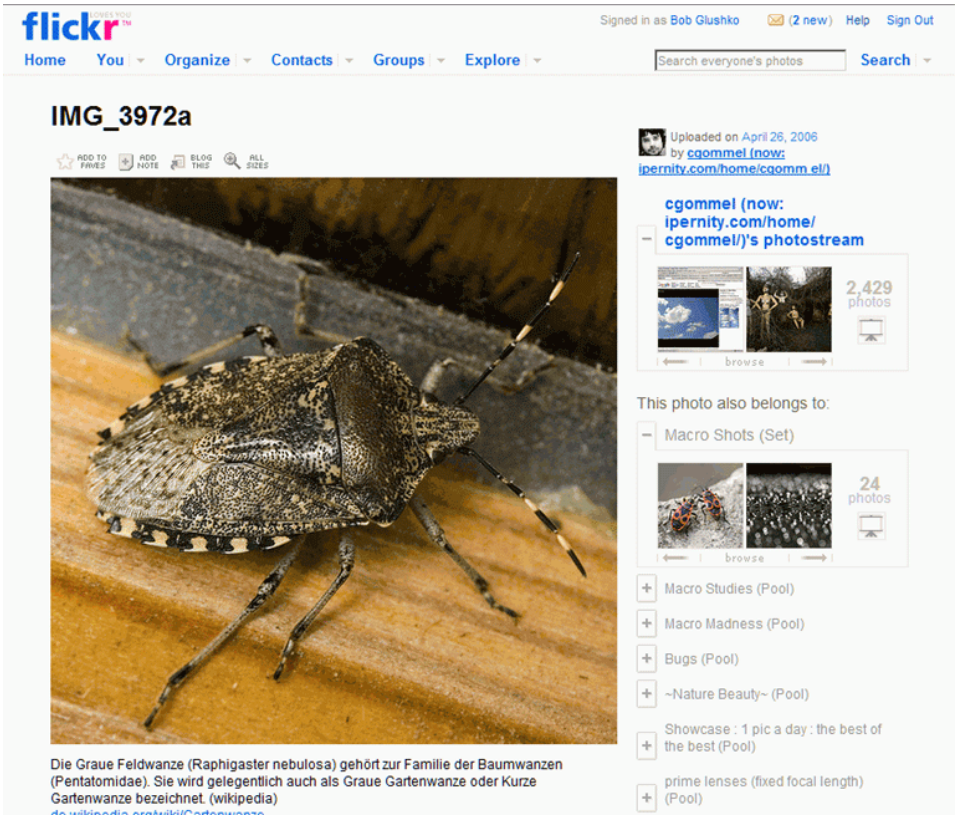
Geotagging and Taxonomic Tagging

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?

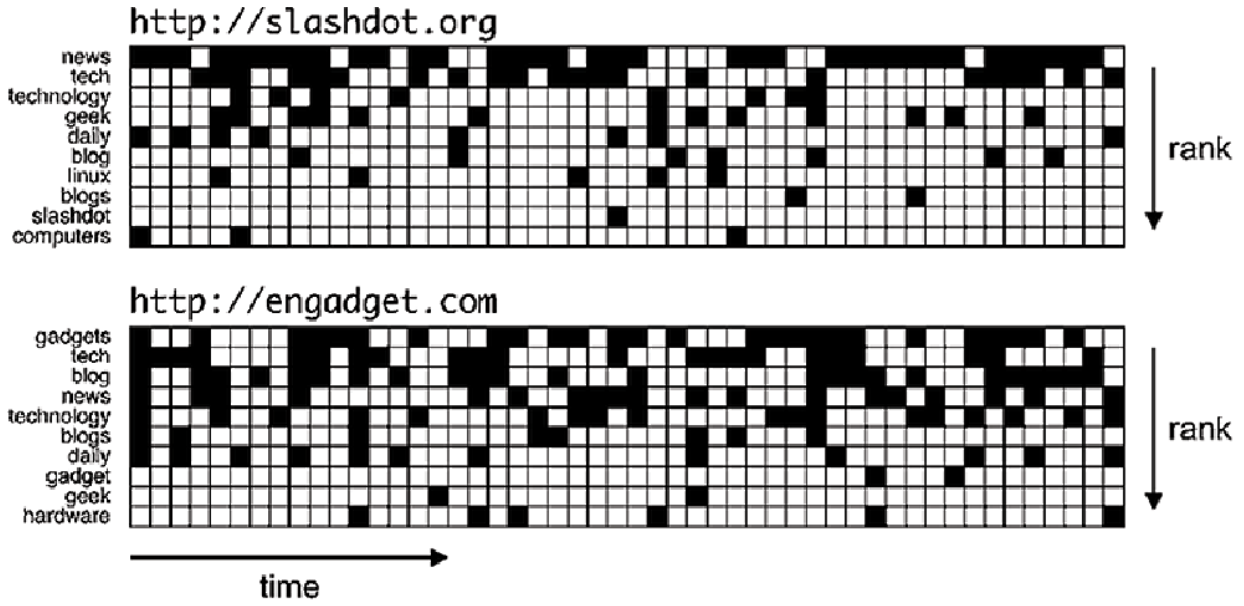
Some systems (like del.icio.us) don't allow users to see the tags assigned by other users when they are tagging a resource

But once a user tags a resource, most systems reveal the tags applied by other users

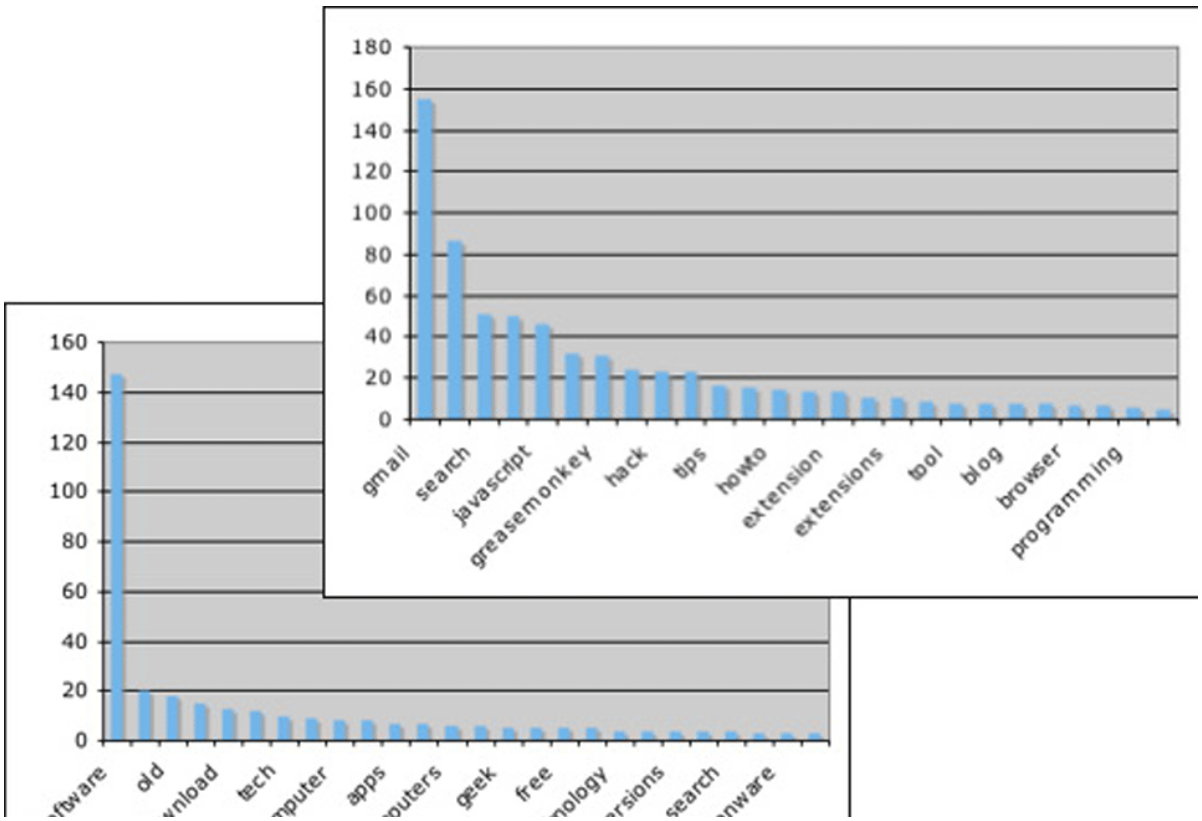
If your tag(s) don't match what others are using, do you?

- Change your tag to adapt to the group norm (maybe you'd look at the other resources with that tag to compare "senses")
- Keep your tag to influence the group norm
- Add the group tag but keep yours as well

Semiotic Dynamics, or Tagging Over Time



The Long Tail



Golder and Huberman Study

"The Structure of Collaborative Tagging Systems" studies tagging patterns for individuals and the most popular resources tagged on del.icio.us

They observe "tension between tags that may be useful to the Delicious community at large and those useful only to oneself"

The diversity of tags for many resources and tags whose meaning is intrinsic to the tagger demonstrates that a significant amount of tagging, if not all, is done for personal use rather than public benefit

Nonetheless...

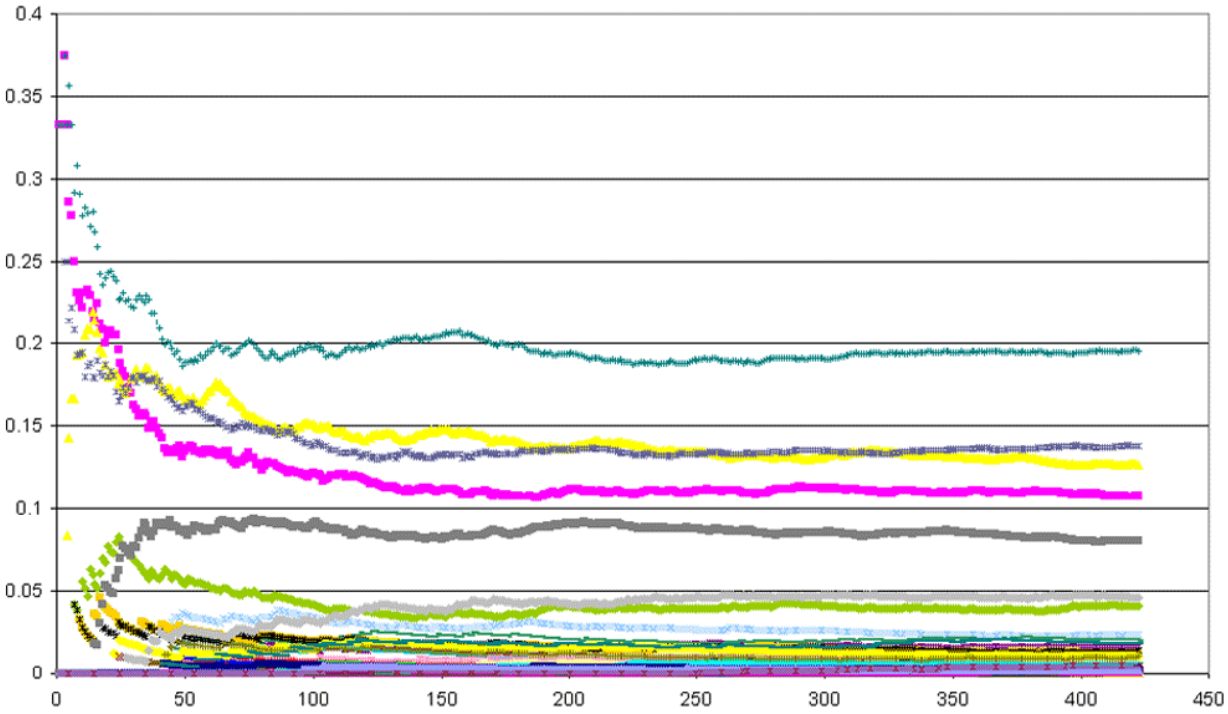
Divergence, Stabilization, or Convergence?

Will individuals' varying tag collections and personal preferences, compounded by an ever-increasing number of users, yield a chaotic pattern of tags?

Or will the combined tags of many users converge?

Or will a stable pattern emerge in which the proportions of each tag are nearly fixed?

Tag Stabilization in Golder and Huberman Study



Social Categorization in the Enterprise

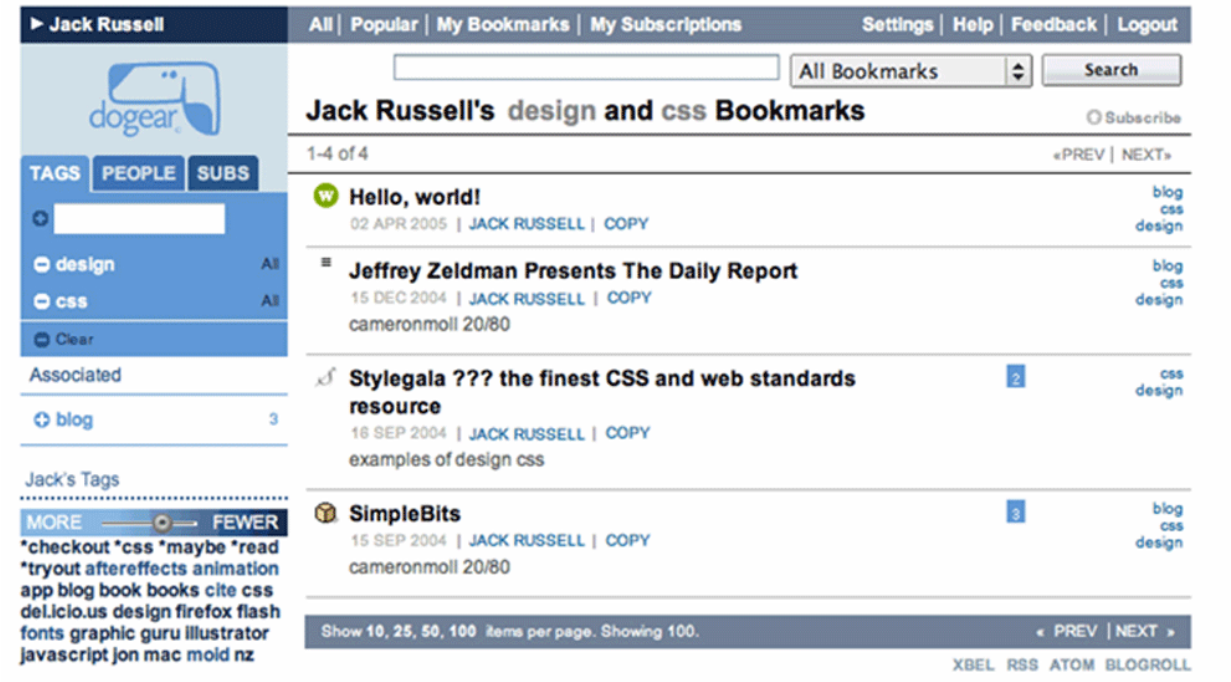
Tagging and bookmarking are being adapted for use in business organizations and large enterprises

Some significant differences with "open web" categorization

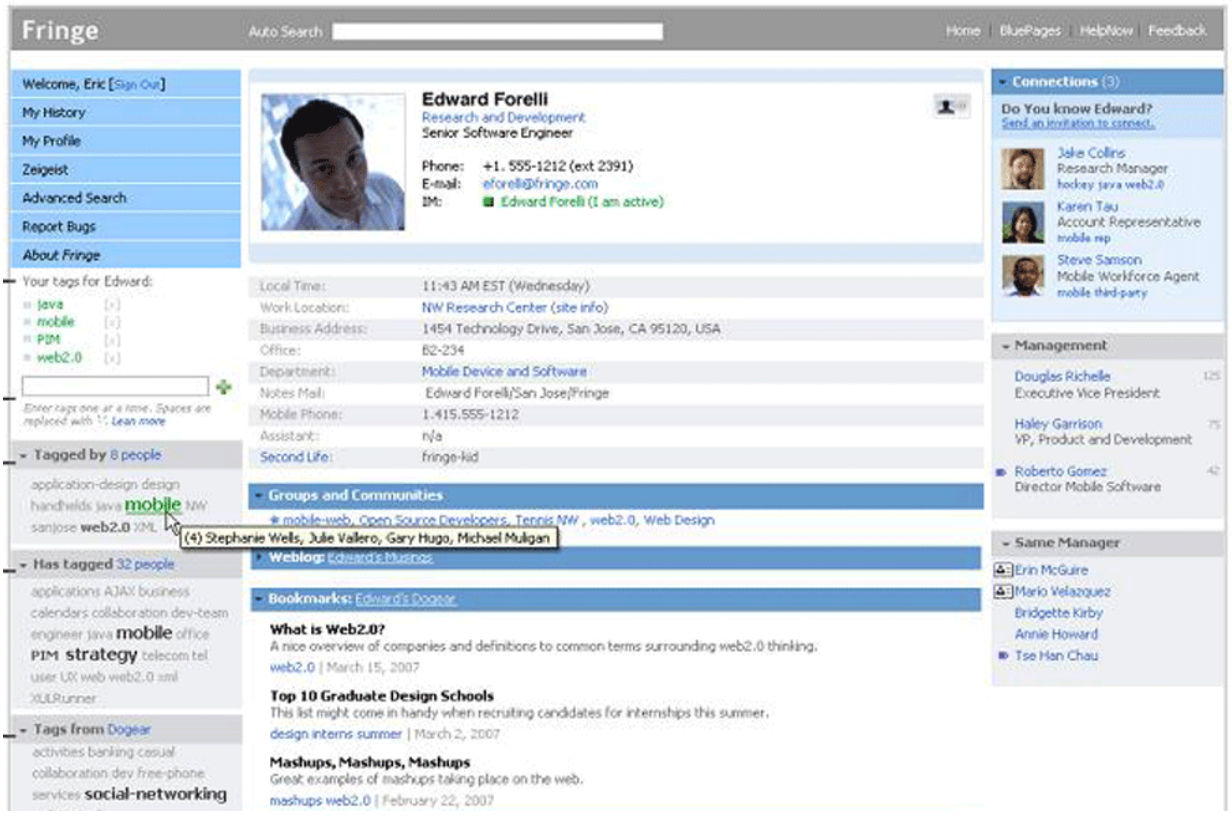
- Every user is authenticated to a "real" identity
- Organizational norms and incentives restrict/shape the purposes and nature of the categorization

These applications can capture expertise and interests implicitly and at lower cost than traditional knowledge management applications

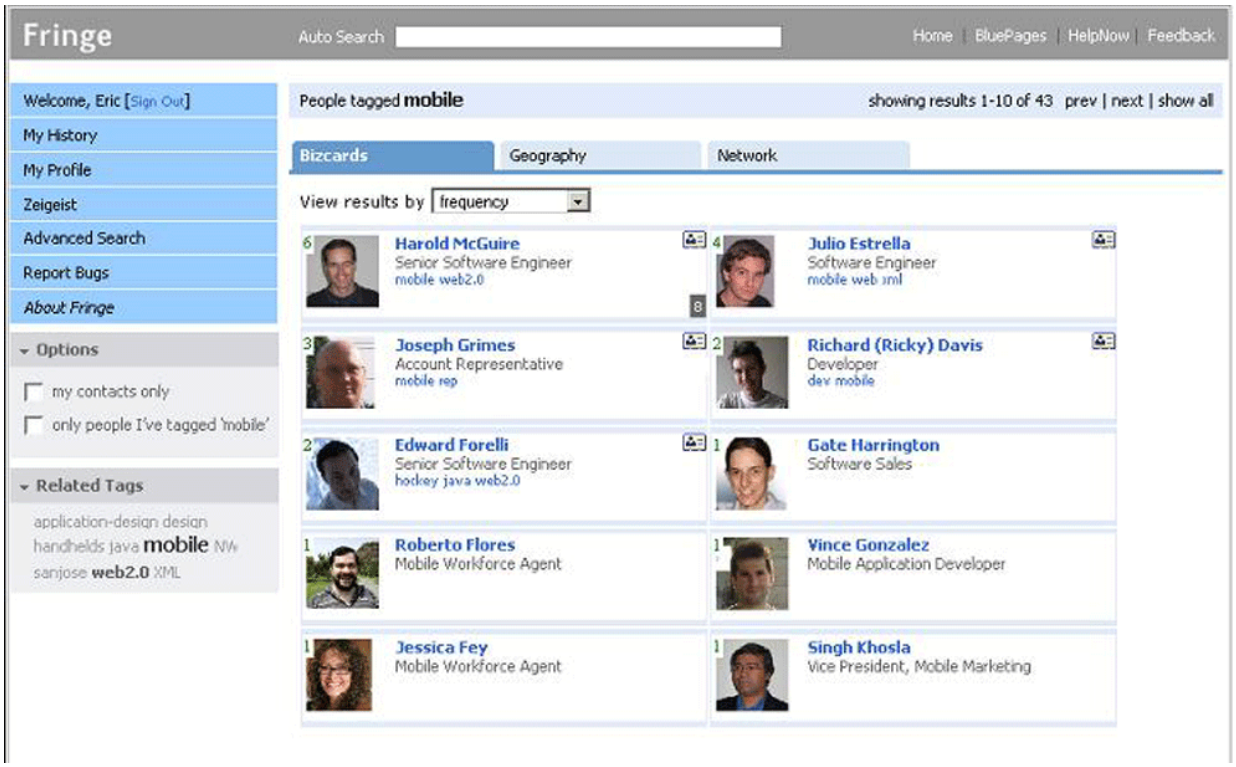
DogEar



Fringe [1]



Fringe [2]



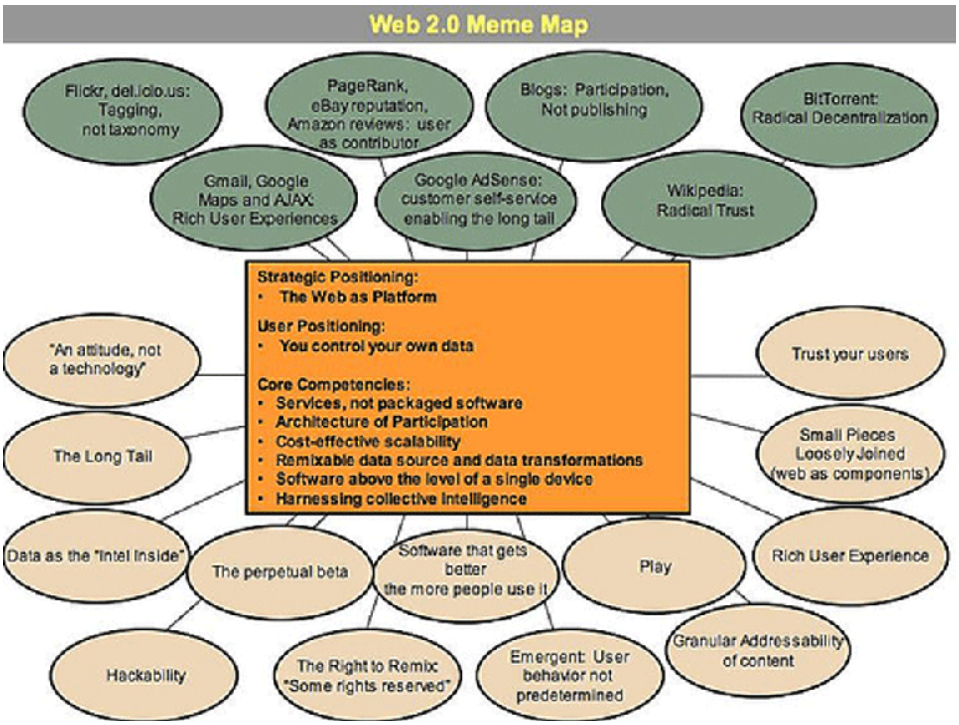
The Underlying Philosophies / Assumptions -- Why Social / Distributed Efforts (are supposed to) Work

"Architectures of Participation"

"Given Enough Eyeballs, All Bugs Are Shallow"

"Harnessing Collective Intelligence"

"Web 2.0"



From Web 2.0 to Web 3.0

The possibility of combining the "generosity" and "curation" principles embodied in Web 2.0 with the "intelligence" of the semantic web has inspired talk of a "social-semantic web" or "Web 3.0"

But as Gruber points out:

- "Collected" intelligence isn't the same as "Collective" intelligence; "Mass authoring" is not the same as "mass authority"
- An "intelligent" system must be at least as intelligent as the individuals that comprise it

The challenge is to devise mechanisms and systems that use people and computers in symbiotic or synergistic ways to harvest and exploit human-generated knowledge

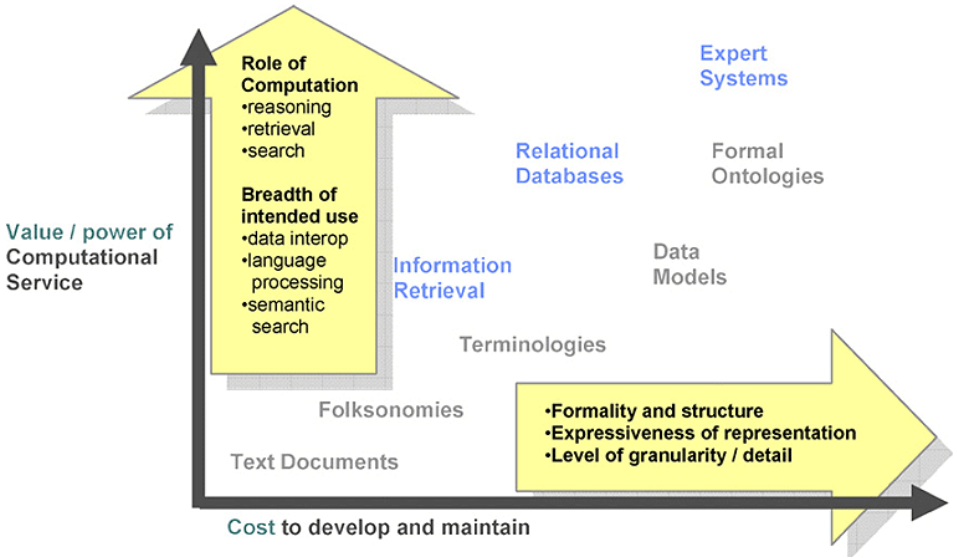
Getting There From Here

"Augmenting user-contributed data with structured data" is possible, with the caveat that "users in the social web are not there to create databases; they are there to have fun, connect with other people, promote their ideas, and share their experiences"

"A little semantics goes a long way" -- so collect data on basic dimensions of who/where/when/why to facilitate integration and inference

Can we overcome the conventional correlation between computational power of a knowledge representation and the cost of creating it?

Can We "Bootstrap" Collective Intelligence?



Readings for INFO Lecture #15

David Kirsh, "A Few Thoughts on Cognitive Overload"

Catherine Marshall, "Rethinking Personal Digital Archiving, Part 1" D-Lib Magazine