RE_ACTION:

THE DIGITAL ARCHIVE EXPERIENCE

RE_ACTION:
THE DIGITAL ARCHIVE
EXPERIENCE

RENEGOTIATING THE COMPETENCES OF THE ARCHIVE
AND THE (ART) MUSEUM IN THE 21ST CENTURY

Editor-in-Chief: Morten Søndergaard
Editors: Mogens Jacobsen and Morten Søndergaard

ary Art, Roskilde,

# ONTOLOGY IS OVERRATED: CATEGORIES, LINKS AND TAGS

Clay Shirky

This essay is based on two talks I gave in the spring of 2005, one at the O'Reilly ETech conference in March, entitled "Ontology Is Overrated" and the other at the IMCExpo in April entitled "Folksonomies & Tags: The Rise of User-developed Classification." The piece you are now reading is a heavily edited concatenation of those two talks.

Here, I write about categorization, and I want to convince you that a lot of what we think we know about categorization is wrong. In particular, I want to convince you that many of the ways we're attempting to apply categorization to the electronic world are actually a bad fit, because we've adopted habits of mind that are left over from earlier strategies.

I also want to convince you that what we're seeing when we see the world wide web is actually a radical break with previous categorization strategies, rather than an extension of them. The second part of this essay is more speculative, because it is often the case that old systems get broken before people know what's going to take their place. (Anyone observing the music industry today can see this at work.) That's what I think is happening with categorization. What I think is coming instead are much more organic ways of organizing information than our current categorization schemes allow, based on two units – the link, which can point to anything, and the tag, which is a way of attaching labels to links. The strategy of tagging – free-form labeling, without regard to categorical constraints – seems like a recipe for disaster, but as the web has shown us, we can extract a surprising amount of value from big messy data sets.

PART I: CLASSIFICATION AND ITS DISCONTENTS

More about Clay Shirky:
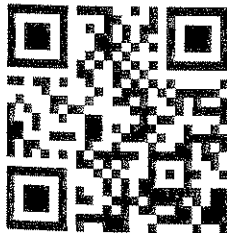
Q: What is Ontology?
A: It Depends on What the Meaning of "Is" Is.
I need to provide some quick definitions, starting with ontology. It is a rich irony that the word "ontology," which has to do with making clear and explicit statements about entities in a particular domain, has so many conflicting definitions. I'll offer two general ones.

The main thrust of ontology in the philosophical sense is the study of entities and their relations. The question ontology asks is: What kinds of things exist or can exist in the world, and what manner of relations can those things have to each other? Ontology is less concerned with what is than with what is possible.

The knowledge management and artificial intelligence (AI) communities use a related definition; they've taken the word "ontology" and applied it more directly to their problem. The sense of ontology in those communities is akin to "an explicit specification of a conceptualization."

The common thread between the two definitions is essence, "Isness." In a particular domain, what kinds of things can we say exist in that domain, and how can we say those things relate to each other?

The other pair of terms I need to define are categorization and classification. These actions organize a collection of entities, whether things or concepts, into related groups. Though there are some field-by-field distinctions, the terms are in the main used interchangeably.

And then there is ontological classification or categorization, which is organizing a set of entities into groups, based on their essences and possible relations. A library catalogue, for example, assumes that for any new book, its logical place already existed within the system, even before the book was published. That strategy of designing categories to cover possible cases in advance is what I'm primarily concerned with, because it is both widely used and badly overrated in terms of its value in the digital world. Now, anyone who deals with categorization for a living will tell you they can never get a perfect system. In working classification systems, success is not "Did we get the ideal arrangement?" but rather "How close did we come, and on what measures?" The idea of a perfect scheme is simply a Platonic ideal. However, I want to argue that even the ontological *ideal* is a mistake. Even using theoretical perfection as a measure of practical success leads to misapplication of resources.

Now, to the problems of classification.

The periodic table of the elements is my vote for "Best. Clas-

Cleaving Nature at the

IA	2A	3A	4A	5A	6A	7A		8		1B	2B	3B	4B	5B	6B	7B	0
H																	He
Li	Be											B	C	N	O	F	Ne
Na	Mg											Al	Si	P	S	Cl	Ar
K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe
Cs	Ba	L	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn
Fr	Ra	A															

	L	La	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu
	A	Ac	Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lr

Cleaving Nature at the Joints

sification. Ever." It turns out that by organizing elements by the number of protons in the nucleus, you get all of this fantastic value, both descriptive and predictive value. And because what you're doing is organizing *things*, the periodic table comes as close to making assertions about essence as it is physically possible to get. This is a really powerful scheme, almost perfect. Almost.

All the way over in the right-hand column, the pink column, are the noble gases. "Noble gas" is an odd category, because helium is no more a gas than mercury is a liquid. Helium is not fundamentally a gas, it's just a gas at most temperatures, but the people studying it at the time didn't know that because they weren't able to make it cold enough to see that helium, like everything else, has different states of matter. Lacking the right measurements, they assumed that gaseousness was an essential aspect – literally, part of the essence – of those elements.

Even in a nearly perfect categorization scheme, there are these kinds of context errors, where people place something that is true only at room temperature, absolutely unrelated to essence, right in the center of the categorization. And the category "Noble Gas" has stayed there from the day they added it, because we've all gotten used to that anomaly as a frozen accident.

If it's impossible to create a completely coherent categorization scheme, even when you're doing something as physically related to essence as chemistry, imagine the problems faced by anyone who's dealing with a domain where essence is even less obvious.

Which brings me to the subject of libraries.

Of Cards and Catalogues

The periodic table gets my vote for the best categorization scheme ever, but libraries have the best-known categorization schemes. The experience of the library catalogue is probably what people know best as a high-order categorized view of the world, and those cataloguing systems contain all kinds of odd mappings between the categories and the world they describe.

Here's the first top-level category in the Soviet library system:

A: Marxism-Leninism

A1: Classic works of Marxism-Leninism
A3: Life and work of K. Marx, F. Engels, V.I. Lenin
A5: Marxism-Leninism Philosophy
A6: Marxist-Leninist Political Economics
A7/8: Scientific Communism

Some of those categories are starting to look a little bit dated.

Or, my favorite – this is the Dewey Decimal System's categorization for religions of the world, which is the 200 category.

Dewey, 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

How much is this not the categorization you want in the 21st century?

This kind of bias is rife in categorization systems. Here's the Library of Congress' categorization of History. These are all the top-level categories – all of these things are presented as being equal.

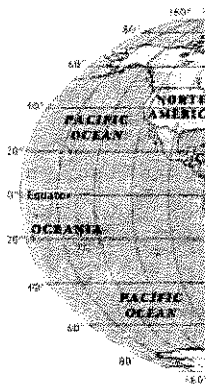D: History (general)
DA: Great Britain
DB: Austria
DC: France
DD: Germany
DE: Mediterranean
DF: Greece
DG: Italy

Spot the difference.

gorization scheme

ation schemes.

ıly what people
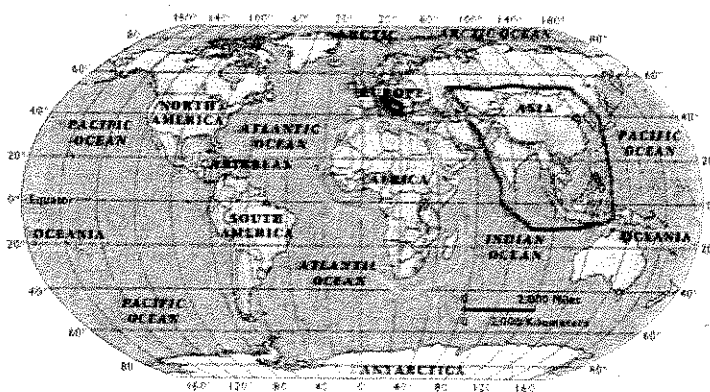
e world, and those

ıpings between the

ıet library system:

nin

ttle bit dated.

l System's catego-

00 category.

ıt in the 21st cen-

ı. Here's the Li-

ese are all the top-

d as being equal.

DH: Low Countries
DJ: Netherlands
DK: Former Soviet Union
DL: Scandinavia
DP: Iberian Peninsula
DQ: Switzerland
DR: Balkan Peninsula
DS: Asia
DT: Africa
DU: Oceania
DX: Gypsies

I'd like to call your attention to the ones in bold: The Balkan Peninsula. Asia. Africa.

And just to review the geography:

Spot the difference.

Yet, for all the oddity of placing the Balkan Peninsula and Asia in the same level, this is harder to laugh off than the Dewey Decimal System example, because it's so puzzling. The Library of Congress – founded by Thomas Jefferson and no slouches in the thinking department – has a staff of people who do nothing but think about categorization all day long. So what's being optimized here? It's not geography. It's not population. It's not regional GDP.

What's being optimized is number of books on the shelf. That's what the categorization scheme is categorizing. It's tempting to think that the classification schemes that libraries have optimized in the past can be extended in an uncomplicated way into the digital world. This badly underestimates the degree to which what libraries have historically been managing is an entirely different problem.

The musculature of the Library of Congress categorization scheme looks as though it focuses on concepts. It is organized into non-overlapping categories that get more detailed at lower levels – any concept is supposed to fit in one category and in no others. But every now and again, the skeleton pokes through, and it, the supporting structure around which the system is really built, is designed to minimize seek time on shelves.

The essence of a book isn't the ideas it contains. The essence of a book is "book." Thinking that library catalogues exist to organize concepts confuses the container of the thing contained.

The categorization scheme is a response to physical constraints on storage and to people's inability to keep the location of more than a few hundred things in their mind at once. Once you own more than a few hundred books, you have to organize them somehow. (My mother, who was a reference librarian, said she wanted to reshelve the entire university library by color, because students would come in and say "I'm looking for a sociology book. It's green…") But however you do it, the frailty of human memory and the physical fact of books make some sort of organizational scheme a requirement, and hierarchy is a good way to manage physical objects.

The "Balkans/Asia" kind of imbalance is simply a byproduct of physical constraints. It isn't the ideas in a book that have to be in one place; a book can be about several things at once. It is the book itself, the physical fact of the bound object, that has to be one place, and if it's in one place, it can't also be in another. This, in turn, means that a book has to be declared to be *about* some main thing. A book that is equally about two things breaks the "be in one place" requirement, so each book needs to be declared to be about one thing more than others, regardless of its contents. People have been getting excited about the virtuality of data for decades, and you'd think we'd have internalized the obvious truth: there is no shelf. In the digital world, there is no physical constraint that is forcing this kind of organization on us any longer. We can do without it, and you'd think we'd have learned that lesson by now.

And yet …

Yahoo's Entertainment Category

gorization

is organized into

l at lower levels

nd in no others.

ugh, and it, the

really built, is

The essence of a

exist to organize

ained.

al constraints on

on of more than

you own more

hem somehow.

ne wanted to

ause students

gy book. It's

man memory and

nizational scheme

nage physical

byproduct of

at have to be

once. It is the

that has to be one

other. This, in

bout some main

aks the "be in

declared to be

s contents.

lity of data for

ne obvious truth:

 physical con-

us any longer.

earned that les-

INSIDE YAHOO!

Entertainment: Movies - Music - TV - ET Online
CATEGORIES

Top Categories

* Music (77338) NEW!
* Actors and Actresses (17656) NEW!
* Movies and Film (31630) NEW!

* Television Shows (13577) NEW!
* Humor (4245) NEW!
* Comics and Animation (5522) NEW!

Additional Categories

* Amusement and Theme Parks (454)
* Awards (21) NEW!
* Books and Literature@
* Chats and Forums (56)
* Comedy (1389) NEW!
* Consumer Electronics (1290)

* History (15)
* Magic (303) NEW!
* News and Media (340)
* Organizations (35)
* Performing Arts@
* Radio@

Yahoo's Entertainment
Category

The Parable of the Ontologist, or, "There Is No Shelf"

A little over ten years ago, a couple of developers in Stanford launched a service called Yahoo that offered a list of things available on the web. It was the first really significant attempt to bring order to the web. As the web expanded, the Yahoo list grew into a hierarchy with categories. Then the developers realized that, to maintain the value in the directory, they were going to have to systematize, so they hired a professional ontologist. They developed their now-familiar, top-level categories, which lead to subcategories; each subcategory contains links to still other subcategories, and so on. Now we have an ontologically managed list of what's out there.

Here we are in one of Yahoo's top-level categories, Entertainment.

You can see what the subcategories of Entertainment are, whether or not there are new additions, and how many links roll up under those subcategories. However, the subcategory of Books and Literature doesn't tell viewers how many links roll up under it. Books and Literature doesn't end with a number of links, but with an "@" sign. That "@" is telling us that the category of Books and Literature isn't "really" in the category Entertainment. Yahoo is saying "We've put this link here for your convenience, but that's only to take you to where Books and Literature 'really' are." To which one can only respond, "What's real?"

Yahoo is saying "We understand better than you how the world is organized, because we are trained professionals. So if you mistakenly think that Books and Literature are entertainment, we'll put

125

a little flag up so we can set you right, but to see those links, you have to 'go' to where they 'are'." (My fingers are going to fall off from all the air quotes.) When you go to Literature – which is part of Humanities, not Entertainment – you are told, similarly, that booksellers are not "really" there. Because they provide a commercial service, booksellers are "really" in Business.

Humanities > Literature

Directory > Arts > Humanities > Literature

INSIDE YAHOO!
Shop for Books: Novels on Yahoo! Shopping
CATEGORIES

* Authors (14155) NEW!
* Awards (41) NEW!
* Banned Books (22)
* Bestseller Lists (11)
* Book Arts@
* Booksellers@
* Chats and Forums (44)

* Libraries@
* Literary Libraries (7)
* Literature Weblogs@
* Museums (49)
* News and Media (425)
* Organizations (167)
* Periods and Movements (366)

Literature on Yahoo

Look what's happened here. Faced with the possibility that they could organize things with no physical constraints, Yahoo *added back the shelf*. They couldn't imagine organization without the constraints of the shelf, so they replaced it. It is perfectly possible for any number of links to be in any number of places in a hierarchy, or in many hierarchies, or in no hierarchy at all. But Yahoo decided to privilege one way of organizing links over all others, because they wanted to make assertions about what is "real."

The charitable explanation for this is that Yahoo thought that this kind of *a priori* organization was their job, and they thought their users would value it. The less charitable explanation is that they thought there was business value in determining the view the user would need to use the system. Both of these explanations might have been true at different times and in different measures, but the effect was to override the users' sense of where things ought to be, and to insist on the Yahoo view instead.

File Systems and Hierarchy
It's easy to see how the Yahoo hierarchy maps to technological constraints as well as physical ones. The constraints in the Yahoo directory describe both a library categorization scheme and, obviously, a file system. It is both a powerful tool and a powerful metaphor. We're all so used to it – it seems natural.
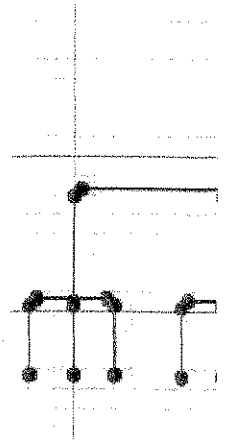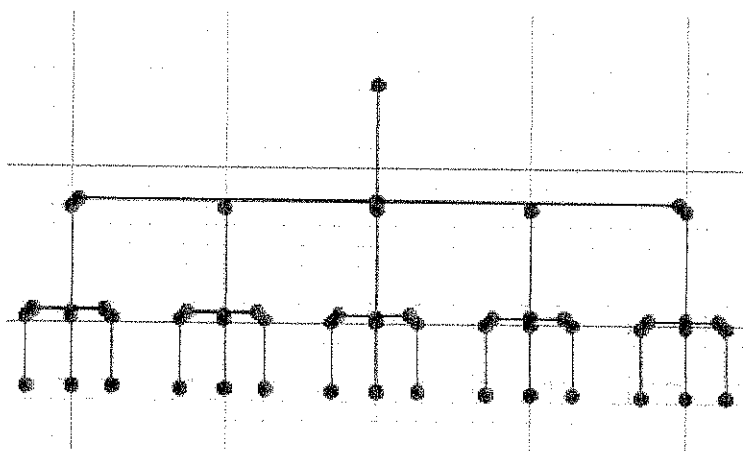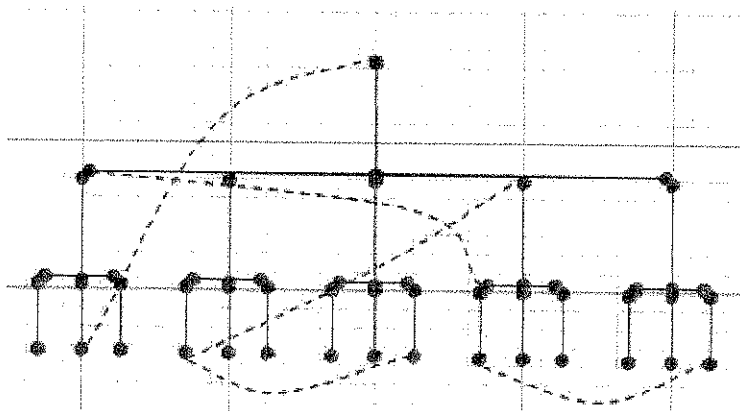

Illus Hierachy
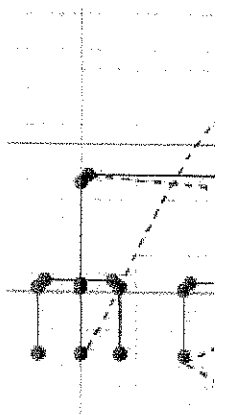
Plus Link

hose links, you
going to fall off
re – which is part
 similarly, that
rovide a commer-

Illus Hierachy

Here we see a top level, with subdirectories that roll up under it.
Subdirectories contain files or further subdirectories and so on, all
the way down to the lowest category. Both librarians and computer
scientists hit on the same subsequent idea: "You know, it wouldn't
hurt to add a few secondary links in here" – symbolic links, aliases,
shortcuts, whatever you want to call them.

bility that they
s, Yahoo *added*
 without the con-
ctly possible for
s in a hierarchy,
But Yahoo decided
thers, because
ıl."

thought that this
ey thought their
on is that they
the view the user
ınations might
measures, but the
ings ought to be,

Plus Link

The Library of Congress shows something similar in its second-
order categorization – "This book is mainly about the Balkans, but
it's also about art, or it's mainly about art, but it's also about the
Balkans." Most hierarchical attempts to subdivide the world use a
system like this.

Then, in the early 1990s, one of the things that Berners-Lee
showed us is that one could have a lot of links. A system doesn't
need to have just a few links; it could have a whole lot of links.

technological
ıts in the Ya-
on scheme and,
 and a powerful
al.

127
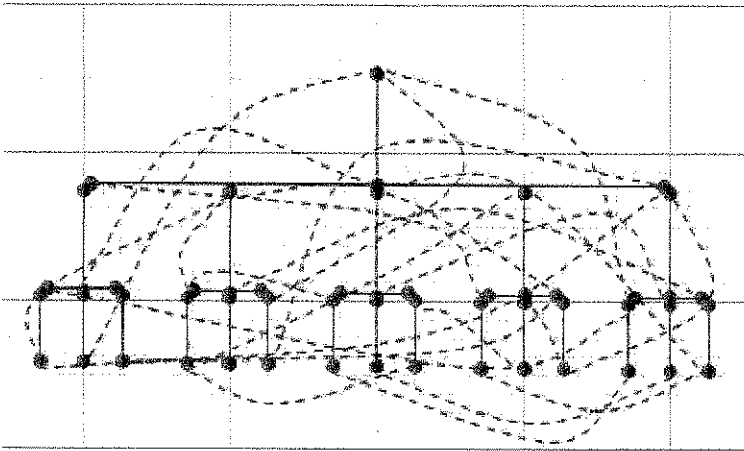
Plus Lots of Links

This is where Yahoo got off the boat. They said, "Get out of here with that crazy talk. A URL can only appear in three places. That's the Yahoo rule." They did that in part because they didn't want to get spammed, since they were doing a commercial direc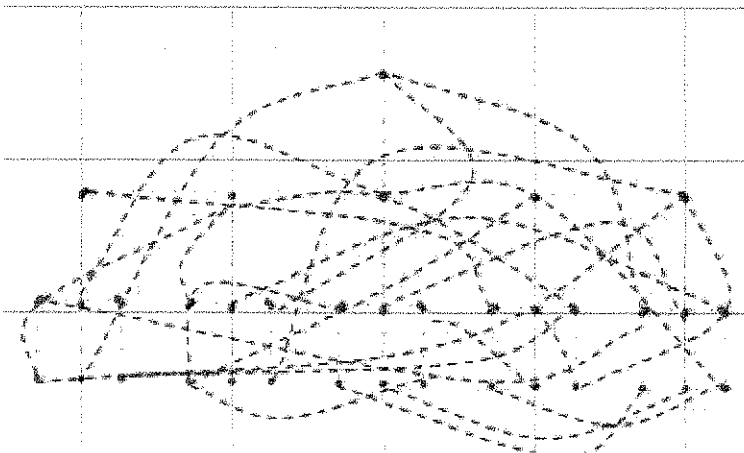tory. They put an upper limit on the number of symbolic links that could fit into their view of the world. But they missed the end of this progression, which is that, if you've got enough links, you don't need the hierarchy anymore. There is no shelf. There is no file system.

Just like (There Is NO File-system)

The links alone are enough.

One reason Google was adopted so quickly is that it understood there is no shelf, and that there is no file system. Google can decide what goes with what *after* hearing from the user, rather than trying to predict in advance what it is the user needs to know.

. "Get out of here
hree places. That's
hey didn't want to
ial directory. They
nks that could fit
end of this pro-
s, you don't need
is no file system.

at it understood
Google can
ne user, rather than
eeds to know.

Let's say I need to find every web page with the words "obstreperous" and "Minnesota." You can't ask a cataloguer in advance to say, "Well, that's going to be a useful category; we should encode that in advance." Instead, what the cataloguer is going to say is, "Obstreperous plus Minnesota! Forget it, we're not going to optimize for one-offs like that." Google, on the other hand, says, "Who cares? We're not going to tell the user what to do, because the link structure is more complex than we can read, except in response to a user query."

Browse versus search is a radical increase in the amount of trust we put in link infrastructure, and in the degree of power derived from that link structure. "Browse" says the people making the ontology, the people doing the categorization, have the responsibility of organizing the world in advance. Given this requirement, the views of the cataloguers necessarily override the user's needs and the user's view of the world. If someone wants something that hasn't been categorized in the way you think about it, they are out of luck.

The search paradigm says the reverse. It says nobody gets to tell the user in advance what he or she needs. "Search" says that, at the moment that the search is initiated, we will do our best to service it based on this link structure. We believe we can build a world where we don't need the hierarchy to coexist with the link structure.

A lot of the current conversation about categorization starts at a second step: "Since categorization is a good way to organize the world, we should…" But the first step is to ask the critical question: Is categorization a good idea? We can see, from the example of Yahoo vs. Google, that there are a number of cases where there is a significant value in *not* categorizing. Even Google adopted DMOZ, the open source version of the Yahoo directory, and later they downgraded its presence on the site because almost no one was using it. When people were offered search and categorization side-by-side, fewer and fewer people were using categorization to find things.

When Does Ontological Classification Work Well?
Ontological classification works well in some places, of course. You need a card catalogue if you are managing a physical library. You need a hierarchy to manage a file system. So what you want to

129

know, when thinking about how to organize anything, is whether that kind of classification is a good strategy.

Here is a partial list of characteristics that help make a classification system work:

Domain to be Organized
→ Small corpus
→ Formal categories
→ Stable entities
→ Restricted entities
→ Clear edges

These are all the domain-specific criteria that you would like to be true if you're trying to classify cleanly. The Periodic Table has all of these things – there are just over one hundred elements; the categories are simple and derivable; protons don't change because of political circumstances; only elements can be classified, not molecules; there are no blended elements; and so on. The more of those characteristics that are true, the better fit ontology is likely to be.

The other key question to consider along with the characteristics of the domain is "What are the participants like?" Here are some things that, if true, help make ontology a workable classification strategy:

Participants
→ Expert cataloguers
→ Authoritative source of judgment
→ Coordinated users
→ Expert users

DSM-IV, the fourth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (*DSM*) that is published by the American Psychiatric Association (APA), is a classic example of an classification scheme that works because of the characteristics listed above. *DSM-IV* (1994) allows psychiatrists all over the United States, in theory, to make the same diagnosis of a mental illness when presented with the same list of symptoms. Furthermore, the source for *DSM-IV*, the APA, is authoritative. The association has the authority to say, for example, which symptoms are manifestations of psychosis. The APA has both expert cataloguers and expert

rthing, is whether

nake a classifica-

ɔu would like to
'eriodic Table has
lred elements; the
ı't change because
classified, not
o on. The more of
ontology is likely

he characteristics
?" Here are some
ble classification

d Statistical Manual
ɪe American
ɪe of an classifica-
istics listed above.
United States,
ıl illness when
ɪermore, the source
ɪciation has the
ɪre manifestations
ɪrs and expert

users. The number of "people infrastructure" that is hidden in a working system like *DSM-IV* is a big part of what makes this sort of categorization work.

This "people infrastructure" is very expensive, though. One of the major problems we encounter is that when we do head-to-head tests – we describe something and then we ask users to guess how we described it – we get a very poor match. Users have a terrifically hard time guessing how something they want will have been categorized in advance, unless they have been educated about those categories in advance as well. The bigger the user base, the more work that education is.

The Domain list can also be viewed another way, saying, "Here are some characteristics for which ontological classification *doesn't* work well":

Domain
→ Large corpus
→ No formal categories
→ Unstable entities
→ Unrestricted entities
→ No clear edges

Participants
→ Uncoordinated users
→ Amateur users
→ Naive cataloguers
→ No authority

In other words, ontology will be a bad strategy under these conditions: a large, ill-defined corpus, naive users, non-expert cataloguers, and no authority at hand to supervise the process.

The list of factors making ontology a bad fit is, also, an almost perfect description of the web – largest corpus, most naive users, no global authority, and so on. The greater the push in the direction of scale, spread, fluidity, flexibility, the harder it becomes to handle the expense of starting a cataloguing system and the hassle of maintaining it, to say nothing of the amount of force that must be exerted over users to convince them to drop their world view in favor of yours.

We know SUVs are light trucks instead of cars because the government says so. This is voodoo categorization, where acting on the model changes the world: when the government says an SUV is a truck, it *is* a truck, by definition. Much of the appeal of categorization comes from this sort of voodoo, where the people doing the categorizing believe, even if only unconsciously, that naming the world changes it. Unfortunately, most of the world is not amenable to voodoo categorization.

The reason we don't know whether or not *Buffy, The Vampire Slayer* is science fiction, for example, is because no one can say definitively yes or no. In environments where there is no authority and no force that can be applied to the user, it's very difficult to support the voodoo style of organization. Merely naming the world creates no actual change, either in the world or in the minds of potential users who don't understand the system.

Mind reading

One of the biggest problems with categorizing things in advance is that it forces the categorizers to take on two historically difficult jobs: mind reading and fortune telling. Categorizers must guess what their users are thinking and make predictions about the future.

The mind-reading aspect shows up in conversations about controlled vocabularies. Whenever users are allowed to label or tag things, someone always says, "Hey, I know! Let's make a thesaurus, so that if you tag something 'Mac' and I tag it 'Apple' and somebody else tags it 'OSX,' we all end up looking at the same thing!" They point to the signal loss from the fact that users, although they use these three different labels, are talking about the same thing.

The assumption is that we both can and should read the minds of users, that we can understand what they meant when they used a particular label, and, understanding that, we can start to restrict those labels or at least map them easily onto one another.

This looks relatively simple with the Apple/Mac/OSX example, but when we start to expand to other groups of related words like movies, film, and cinema, the case for the thesaurus becomes much less clear. I learned this from Brad Fitzpatrick's design for Live-Journal, which allows users to list their own interests. LiveJournal makes absolutely no attempt to enforce solidarity, a thesaurus,

because the govern-
ere acting on the
: says an SUV is a
peal of categoriza-
people doing the
 that naming the
rld is not amenable

, *The Vampire*
no one can say
.ere is no author-
it's very difficult
rely naming the
rld or in the minds
em.

.hings in advance
.istorically dif-
tegorizers must
edictions about the

ions about control-
 label or tag things,
a thesaurus, so that
nd somebody else
thing!" They point
gh they use these
: thing.

read the minds of
vhen they used a
: start to restrict
: another.

:/OSX example,
related words like
irus becomes much
lesign for Live-
:rests. LiveJournal
:y, a thesaurus,

---

or a minimal set of terms: there are no check-box or drop-box, just free-text typing. Some people say they're interested in movies. Some people say they're interested in film. Others say they're interested in cinema.

The cataloguers' first reaction to that is, "Oh my God, that means you won't be introducing the movies people to the cinema people!" To which the obvious answer is "Good. The movie people don't *want* to hang out with the cinema people." Those terms actually encode different things, and the assertion that restricting vocabularies improves signal assumes that that there is no signal in the difference itself and no value in protecting the user from too many matches.

When we reach the point of hotly contested terms like queer/gay/homosexual, all the signal loss is in the collapse, not in the expansion. "Oh, the people talking about 'queer politics' and the people talking about 'the homosexual agenda,' they're really talking about the same thing." Oh, no, they're not! If you think the movies and cinema people were going to have a fight, wait until you get the queer politics and homosexual agenda people in the same room. You can't do it. You can't collapse these categorizations without some signal loss. The problem is, because the cataloguers assume their classification should have force on the world, they underestimate the difficulty of understanding what users are thinking, and they overestimate the amount to which users will agree, either with one another or with the cataloguers, about the best way to categorize. They also underestimate the loss sustained by erasing difference of expression, and they overestimate loss from the lack of a thesaurus.

Fortune telling

Predicting the future turns out to be hard. Yet any classification system meant to be stable over time puts the categorizer in the position of fortune teller.

Alert readers will be able to spot the difference between Sentence A and Sentence B.
> A: "I love you."
> B: "I will always love you."

Woe betide the person who utters Sentence B when what they

mean is Sentence A. Sentence A is a statement. Sentence B is a prediction.

But this is the ontological dilemma. Consider the following statements:

A: "This is a book about Dresden."

B: "This is a book about Dresden, and it goes in the category 'East Germany.'"

The second sentence seems so obvious, but East Germany actually turned out to be an unstable category. Cities are real. They are real, physical facts. Countries are social fictions. It is much easier for a country to disappear than for a city to disappear, so when you're saying that the small thing is contained by the large thing, you're mixing radically different kinds of entities. We pretend that "country" refers to a physical area the same way "city" does, but it's not true, as we know from places like the former Yugoslavia.

There is a top-level category – you might have seen it earlier in the Library of Congress scheme – called Former Soviet Union. The best they were able to do was tack "former" onto that entire zone that they'd previously categorized as the Soviet Union. The Library of Congress did this not because that's what they thought was true about the world, but because they don't have the staff to reshelve all the books. That's the constraint.
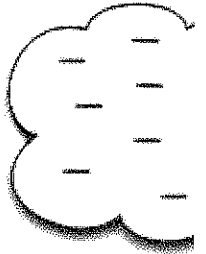
PART II:
THE ONLY GROUP THAT CAN CATEGORIZE EVERYTHING IS EVERYBODY

"My God! It's full of links!"
When we reexamine categorization without assuming the physical constraint either of hierarchy on disk or hierarchy in the physical world, we get very different answers. Let's say you wanted to merge two libraries – mine and the Library of Congress. (You can tell it's the Library of Congress on the right, because they have a few more books than I do.)

So, how do we do this? Do I have to sit down with the Librarian of Congress and say, "Well, in my world, *Python In A Nutshell* is a reference work, and I keep together all of my books on creativity." Do we have to hash out the difference between my categorization scheme and theirs before the Library of Congress is able to take my books?

Two Categorized Colle[c]
of Books

Merge ISBNs

Sentence B is a

te following state-

s in the category

t Germany actu-
s are real. They are
It is much easier
ppear, so when
y the large thing,
s. We pretend that
"city" does, but it's
r Yugoslavia.

seen it earlier in
r Soviet Union.
" onto that entire
oviet Union. The
what they thought
t have the staff to

HING IS EVERYBODY

aming the physical
hy in the physi-
y you wanted to
Congress. (You can
cause they have a

ith the Librarian
In A Nutshell is a
ooks on creativity."
my categorization
ss is able to take
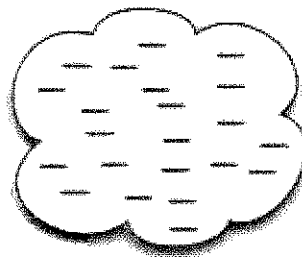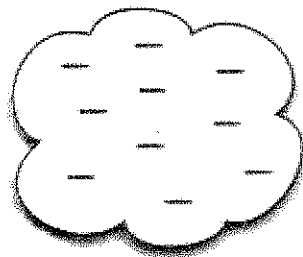
Two Categorized Collections
of Books

No, of course not. They're able to accept my books while ignor-
ing my categories, because all my books have ISBNs, International
Standard Book Numbers. They're not merging at the category
level. They're merging at the globally unique item level. My
entities, my uniquely labeled books, go into Library of Congress
scheme trivially. The presence of unique labels means that merging
libraries doesn't require merging categorization schemes.
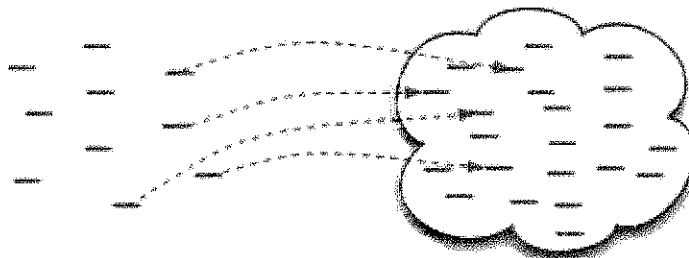
Merge ISBNs

Now imagine a world where *every thing* can have a unique identi-
fier. This should be easy, since that's the world in which we cur-
rently live. The URL gives us a way to create a globally unique ID
for anything to which we need to point. Sometimes the pointers
are direct, as when a URL points to the contents of a web page.
Sometimes they are indirect, as when you use an Amazon link to
point to a book. Sometimes there are layers of indirection, as when
you use a URI, a uniform resource identifier, to name something
whose location is indeterminate. But the basic scheme gives us
ways to create a globally unique identifier for anything.

And once you can do that, anyone can label those pointers, can
tag those URLs, in ways that make them more valuable, and all
without requiring top-down organization schemes. And this – an
explosion in free-form labeling of links, followed by all sorts of

ways of grabbing value from those labels – is what I think is happening now.

Great Minds Don't Think Alike

Here is del.icio.us, Joshua Shachter's social bookmarking service. It's for people who are keeping track of their URLs for themselves, but who are willing to share globally a view of what they're doing, creating an aggregate view of all users' bookmarks, as well as a personal view for each user.

Front Page of del.icio.us

As you can see here, the characteristics of a del.icio.us entry are a link, an optional extended description, and a set of tags, which are words or phrases users attach to a link. Each user who adds a link to the system can give it a set of tags – some do, some don't. Attached to each link on the home page are the tags, the username of the person who added it, the number of other people who have added that same link, and the time.

Tags are simply labels for URLs, selected to help the user retrieve those URLs later. Tags also group together related URLs. There is no fixed set of categories or officially approved choices. You can use words, acronyms, numbers, whatever makes sense to you, without regard for anyone else's needs, interests, or requirements.

The addition of a few simple labels hardly seems so momentous, but the surprise here, as so often with the web, is the surprise of simplicity. Tags are important mainly for what they leave out. By foregoing formal classification, tags enable a huge amount of user-produced organizational value at small cost.

at I think is hap-

cmarking service.
RLs for themselves,
what they're doing,
·ks, as well as a

cio.us entry are
et of tags, which
 user who adds a
: do, some don't.
tags, the username
· people who have

ɔ the user retrieve
:ed URLs. There
d choices. You can
 sense to you, with-
·equirements.

; so momentous,
; the surprise of
ney leave out. By
ge amount of user-

There's a useful comparison here between gopher and the web, where gopher was better organized, better mapped to existing institutional practices, and utterly unfit to work at Internet scale. The web, by contrast, was and is a complete mess, with only one brand of pointer, the URL, and no mechanism for global organization or resources. The web is mainly notable for two things: the way it ignored most of the theories of hypertext and rich metadata, and how much better it works than any of the proposed alternatives. (The Yahoo/Google strategies I mentioned earlier also split along those lines.)

With changes afoot, here are some of the things that I think are coming, with regard to the advantages of tagging systems:

→ **Market Logic**: As we get used to the lack of physical constraints, as we internalize the fact that there is no shelf and there is no disk, we're moving towards market logic, where you deal not with individual motivation, but group value.

As Schachter says of del.icio.us, "Each individual categorization scheme is worth less than a professional categorization scheme. But there are many, many more of them." If you find a way to make it valuable to individuals to tag their information, you'll generate a lot more data about any given object than if you pay a professional to tag it once and only once. And if you can find any way to create value from combining myriad amateur classifications over time, those classifications will become more valuable than professional categorization schemes, particularly with regard to robustness and cost of creation.

The other essential value of market logic is that individual differences don't have to be homogenized. Look for the word "queer" in almost any top-level categorization. You will not find it, even though, as an organizing principle for a large group of people, that word matters enormously. Users don't get to participate in those kinds of discussions around traditional categorization schemes, but with tagging, anyone is free to use the words he or she thinks are appropriate without having to agree with anyone else about how something "should" be tagged. Market logic allows many distinct points of view to coexist, because it allows individuals to preserve their point of view, even in the face of general disagreement.

→ **User and Time Are Core Attributes**: This is absolutely essential. The attitude of the Yahoo ontologist and her staff was – "We are Yahoo. We do not have biases. This is just how the world is. The world is organized into a dozen categories." You don't know who those people were, where they came from, what their background was, what their political biases might be.

Here, because you can derive "this is who this link was tagged by" and "this is when it was tagged," you can start to make inclusion and exclusion around people and time, not just tags. You can start to do grouping. You can start to do decay. "Roll up tags from just this group of users. I'd like to see what they are talking about," or "Give me all tags with this signature, but anything that's more than a week old or a year old."

This is group tagging – not the entire population, and not just me. It's like Unix permissions; right now we've got tags for user and world, and this is the basis on which we will be inventing group tags. We're going to start to be able to subset our categorization schemes. Instead of having massive categorizations and then specialty categorization, we're going to have a spectrum between them, based on the size and makeup of various tagging groups.

→ **Signal Loss from Expression:** The signal loss in traditional categorization schemes comes from compressing things into a restricted number of categories. With tagging, when there is signal loss, it comes from people not having any commonality when talking about things. The loss is from the multiplicity of points of view, rather than from compression around a single point of view. But in a world where enough points of view are likely to provide some commonality, the aggregate signal loss falls with scale in tagging systems, while it grows with scale in systems with single points of view.

The solution to this sort of signal loss is growth. Well-managed, well-groomed organizational schemes get worse with scale, both because the costs of supporting such schemes at large volumes are prohibitive, and, as I noted earlier, scaling over time is also a serious problem. Tagging, by contrast, gets better with scale. With a multiplicity of points of view the question isn't "Is everyone tagging any given link 'correctly,'" but, rather, "Is anyone tagging it the way I do?" As long as at least one other person tags something the way you would, you'll find it. Using a thesaurus to force every-

is absolutely essen-
:r staff was – "We
ow the world is.
' You don't know
what their back-

nk was tagged by"
) make inclusion
:ags. You can start
up tags from just
talking about," or
ng that's more than

n, and not just me.
ags for user and
nventing group
ır categoriza-
izations and then
pectrum between
agging groups.

ss in traditional
g things into a
when there is signal
ıonality when
plicity of points of
ıgle point of view.
likely to provide
Is with scale in
tems with single

.. Well-managed,
: with scale, both
large volumes
ver time is also a
er with scale. With
t "Is everyone tag-
ınyone tagging it
on tags something
ırus to force every-

one's tags into tighter synchrony would actually worsen the noise
you'll get with your signal. If there is no shelf, then even *imagining*
that there is a single right way to organize things is an error.

The Filtering is Done Post-Hoc: There's an analogy here with
every journalist who has ever looked at the web and said "Well,
it needs an editor." The web has an editor: everybody. In a world
where publishing is expensive, the act of publishing is also a
statement of quality. The filter comes before the publication. In a
world where publishing is cheap, putting something out there says
nothing about its quality. It's what happens after it gets published
that matters. If people don't point to it, others won't read it. But
the idea that the filtering comes *after* the publishing is incredibly
foreign to journalists.

Similarly, the idea that the categorization is done after things are
tagged is just as foreign to cataloguers. Much of the expense of
existing cataloguing systems comes in trying to prevent one-off
categories. With tagging, what you say is, "As long as a lot of peo-
ple are tagging any given link, the rare tags can be used or ignored,
as the user likes. We won't even have to expend the cost to prevent
people from using them. We'll just help other users ignore them if
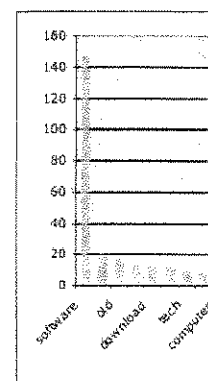they want to."

Again, economy of scale comes to the rescue of the system in a way
that would break traditional cataloguing schemes. The existence
of an odd or unusual tag is a problem if it's the only way a given
link has been tagged, or if there is no way for a user to avoid that
tag. Once a link has been tagged more than once, though, users
can view or ignore the odd tags as it suits them, and the decision
about which tags to use comes after the links have been tagged, not
before.

→ **Merged from URLs, Not Categories:** You don't merge tag-
ging schemes at the category level and then see what the contents
are. As with the "merging ISBNs" idea, we can merge individual
contents because we now have URLs to use as unique handles. You
merge from the URLs, and then try to derive something about the
categorization from there. This allows for partial, incomplete, or
probabilistic merges that are better fits for uncertain environments
– such as the real world – than rigid classification schemes.

→ **Merges are Probabilistic, Not Binary.** Merges create partial overlap between tags, rather than defining tags as synonyms. Instead of saying that any given tag "is" or "is not" the same as another tag, del.icio.us is able to recommend related tags by saying, "A lot of people who tagged this 'Mac' also tagged it 'OSX.'" We move from a binary choice between saying two tags are the same or different to the Venn diagram option of "kind of is/somewhat is/sort of is/overlaps to this degree." That is a really profound change.

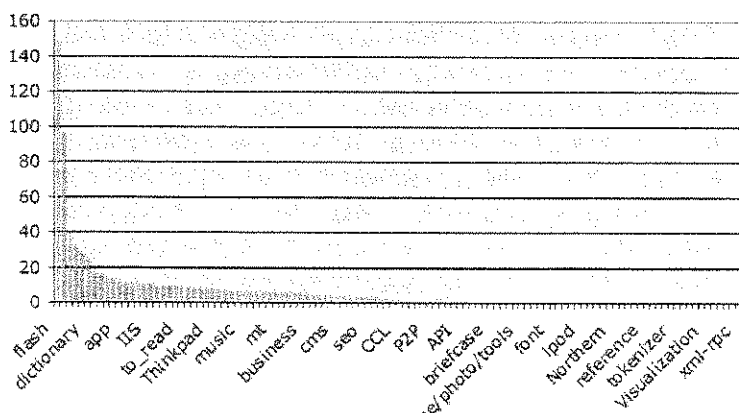**Tag Distributions on del.icio.us**
Here is what I mean about the breakdown of binary categorization. This is a chart based on a small sample of links from the del.icio.us front page, taken during a two-hour window. On the x-axis are the 64 users who posted links during that period. On the y-axis is the total number of discrete kinds of tags that those subjects have ever used in their history on del.icio.us.

The chart shows a great variability in tagging strategies among the users. The user farthest to the left has an enormous number of unique tags, almost 600 of them. Moving to the right, we see charted a group of people who are not quite power taggers but who tag quite a bit. Father to the right, we see the characteristic long line of people who use many fewer tags than the power taggers. (Because this is a two-hour snapshot, it has a natural bias towards frequent del.icio.us users. I'm trying to get a larger data set. My guess is the tail extends much further.) But this is what organization looks like when you turn it over to the users – many different strategies, each of which works in its own context, but which can also be merged.

Different tag "signature different URL's

A single user's tags

ges create par-
gs as synonyms.
ot" the same as
ated tags by say-
tagged it 'OSX.'"
wo tags are the
"kind of is/some-
is a really profound

ary categorization.
rom the del.icio.us
n the x-axis are the
n the y-axis is the
subjects have ever

rategies among
ormous number
he right, we see
ver taggers but
he characteris-
than the power
has a natural bias
get a larger data
But this is what
the users – many
wn context, but

Here are a single user's tags. You can readily tell something about this person: he or she is a Flash programmer; the commonest tag
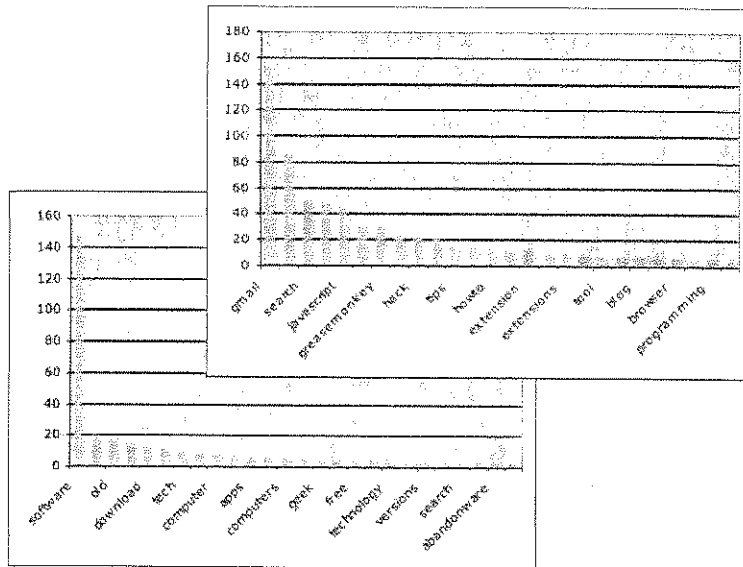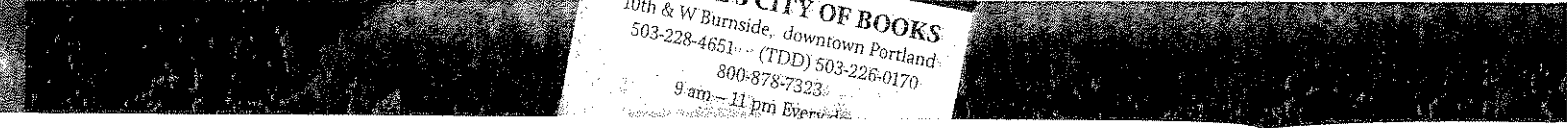

Different tag "signatures" for different URL's

here is Flash, followed by a number of other frequently used tags related mainly to programming. Like the front page, this distribution has an organic signature. Experts don't catalogue this way; experts who learn how to catalogue produce much more consistent labeling. Here, the user has tagged with whatever he or she thought would help to remember the link later.

See the tag "to_read." A professional cataloguer would look at this tag in horror – "This is context-dependent and temporary." Well, so was the category "East Germany." Once you expand your time scale to include the actual life of the categorization scheme itself, you recognize that the distinction between temporary and permanent is awfully vague. There isn't, in fact, a binary condition of a tag that can or cannot survive any kind of long-term examination.

Consider this set of graphs. In my opinion, this is the most interesting and least well understood part of the del.icio.us right now. Pictured here are two different URLs and the tags that a whole group of users applied to them. The graph at the bottom left refers to a site for downloading old versions of programs that are no longer supported. You can see that there is broad communal consensus. One hundred forty people tagged this Software. Then, the next commonest tag, with only twenty occurrences, is

141

Windows, then Old, then Download, and so forth. For this URL, there's a core consensus – this link is about software – and after that one strand of commonality, the number of tags falls off clearly and sharply.

The graph at the upper right, by contrast, shows the tags for a page detailing how to embed standing searches in Gmail. You can see the tags – Gmail, Firefox, Search, Javascript, GreaseMonkey. This is a much messier distribution, with a much less sharp fall-off. The consensus view is that this link is about more kinds of things than the software download link, or, rather, it occupies more contexts for del.icio.us users than the software download link.

Looking at this sort of data, we can start to say of particular URLs that the users tagging this URL either did or did not focus on certain core tags, with this degree of certainty. Thanks to the time stamps, we can even start to understand how the distribution of a URL's tags changes over time. Five years passed between the spread of the link and Google's figuring out how to use whole collections of links to create additional value. We're early in the use of tags, so we don't yet have large, long-lived data sets to look at, but they are being built up quickly. We are just starting to figure out how to extract novel value from entire collections of tags.

Organization Goes Organic

We are moving away from binary categorization – books either are or are not entertainment – and into this probabilistic world, where N% of users think books are entertainment. It may well be that, within Yahoo, there was a big debate about whether or not books are entertainment. But they either had no way of reflecting that debate or they decided not to reveal it to users. What happened instead was it became an all-or-nothing categorization, "This is entertainment, this is not entertainment." We're moving away from that sort of absolute declaration and towards being able to roll up this kind of value by observing how users handle it in practice. It comes down to a question of philosophy. Does the world make sense or do we make sense of the world? If you believe the former, then anyone who tries to make sense of the world differently than you is presenting you with a situation that needs to be reconciled formally; if you get it wrong, you're getting it wrong about the real world.

If, on the other hand, you believe that we make sense of the world,

rth. For this URL,
ware – and after
tags falls off clearly

's the tags for a
in Gmail. You can
, GreaseMonkey.
h less sharp fall-off.
re kinds of things
ccupies more con-
nload link.

of particular URLs
d not focus on
Thanks to the time
e distribution of
ed between the
v to use whole col-
e early in the use of
sets to look at, but
ting to figure out
ns of tags.

n – books either are
ilistic world, where
may well be that,
ther or not books
of reflecting that
What happened
ization, "This is en-
moving away from
ing able to roll up
e it in practice.
s the world make
believe the former,
ld differently than
s to be reconciled
rong about the real

sense of the world,

if we are, from a bunch of different points of view, applying some kind of sense to the world, then you don't privilege one top level of sense making over the other. What you do instead is to try to find ways that the individual sense making can roll up to something that is of value in the aggregate, but you do it without an ontological goal. You do it without the goal of explicitly getting to or even closely matching a theoretically perfect view of the world. Critically, the semantics here rest with the users, not the system. This is not a way to get computers to understand things. When del.icio.us is recommending tags to me, the system is not saying, "I know that OSX is an operating system. Therefore, I can use predicate logic to come up with recommendations – users run software, software runs on operating systems, OSX is a type of operating system – and then say 'Here, Mr. User, you may like these links.'"

What is happening is a lot simpler: "A lot of users tagging things foobar are also tagging them frobnitz. I'll tell the user foobar and frobnitz are related." It's up to the user to decide whether that recommendation is useful – del.icio.us has no idea what the tags *mean*. The tag overlap is in the system, but the tag semantics are in the users. This is not a way to inject linguistic meaning into the machine.

It's all dependent on human context. This is what we're starting to see with del.icio.us, with Flickr, with systems that allow for and are aggregating tags. The signal benefit of these systems is that they don't recreate the structured, hierarchical categorization so often forced upon us by our physical systems. Instead, we're dealing with a significant break: By letting users tag URLs and then aggregating those tags, we're going to be able to build alternate organizational systems, systems that, like the web itself, do a better job of letting individuals create value for each other, often without realizing it.