Clay Shirky's Writings About the Internet
Economics & Culture, Media & Community

clay@shirky.com

Ontology is Overrated: Categories, Links, and Tags

This piece 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 one at the IMCExpo in April entitled "Folksonomies & Tags: The rise of user-developed classification." The written version is a heavily edited concatenation of those two talks.

Today I want to talk 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 Web is actually a radical break with previous categorization strategies, rather than an extension of them. The second part of the talk 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 watching the music industry can see this at work today.) 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, you can extract a surprising amount of value from big messy data sets.

PART I: Classification and Its Discontents

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 thread 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 AI communities have a related definition -- they've taken the word "ontology" and applied it more directly to their problem. The sense of ontology there is something like "an explicit specification of a conceptualization."

The common thread between the two definitions is essence, "Is-ness." 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 are the act of organizing a collection of entities, whether things or concepts, into related

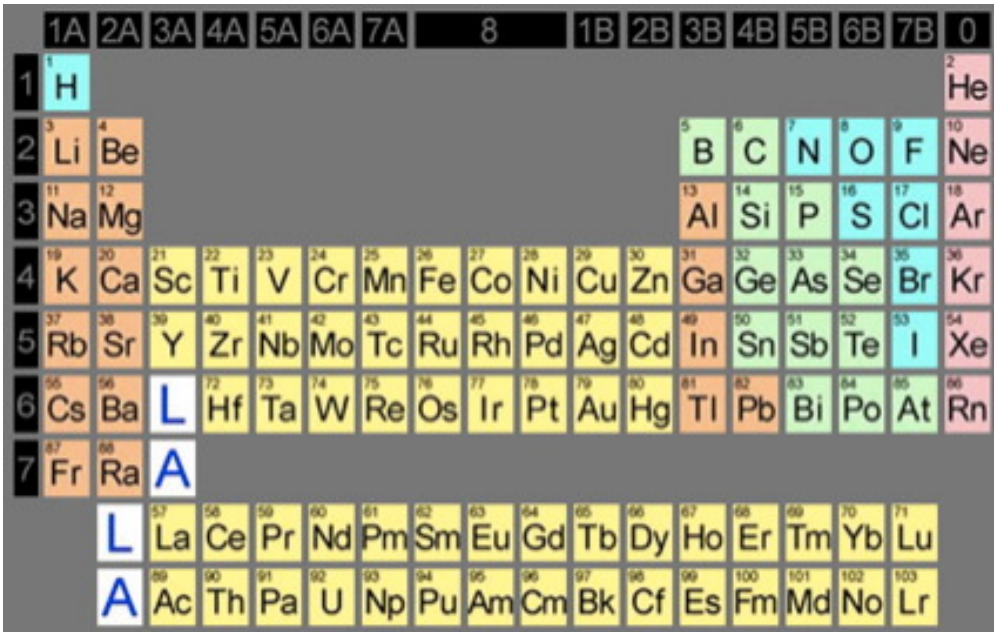
groups. Though there are some field-by-field distinctions, the terms are in the main used interchangeably.

And then there's ontological classification or categorization, which is organizing a set of entities into groups, based on their essences and possible relations. A library catalog, for example, assumes that for any new book, its logical place already exists 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.

Cleaving Nature at the Joints #



[The Periodic Table of the Elements]

The periodic table of the elements is my vote for "Best. Classification. Evar." 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 is 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 noble gases. Now 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 are placing something that is merely true at room temperature, and is 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 just gotten used to that anomaly as a frozen accident.

If it's impossible to create a completely coherent categorization, 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 Catalogs #

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 catalog is probably what people know best as a high-order categorized view of the world, and those cataloging 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 C.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 co-equal.

D: History (general)

DA: Great Britain	DK: Former Soviet Union
DB: Austria	DL: Scandinavia
DC: France	DP: Iberian Peninsula
DD: Germany	DQ: Switzerland
DE: Mediterranean	**DR: Balkan Peninsula**
DF: Greece	**DS: Asia**
DG: Italy	**DT: Africa**

```
      DH: Low Countries  DU: Oceania

      DJ: Netherlands    DX: Gypsies
```

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

And just, you know, 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 example, because it's so puzzling. The Library of Congress -- no slouches in the thinking department, founded by Thomas Jefferson -- 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 for in the past can be extended in an uncomplicated way into the digital world. This badly underestimates, in my view, 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 like it's about concepts. It is organized into non-overlapping categories that get more detailed at lower and lower levels -- any concept is supposed to fit in one category and in no other categories. But every now and again, the skeleton pokes through, and the skeleton, 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 catalogs exist to organize concepts confuses the container for 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 one place, it can't also be in another place. And this in turn means that a book has to be declared to be *about* some main thing. A book which is equally

about two things breaks the 'be in one place' requirement, so each book needs to be declared to about one thing more than others, regardless of its actual contents.

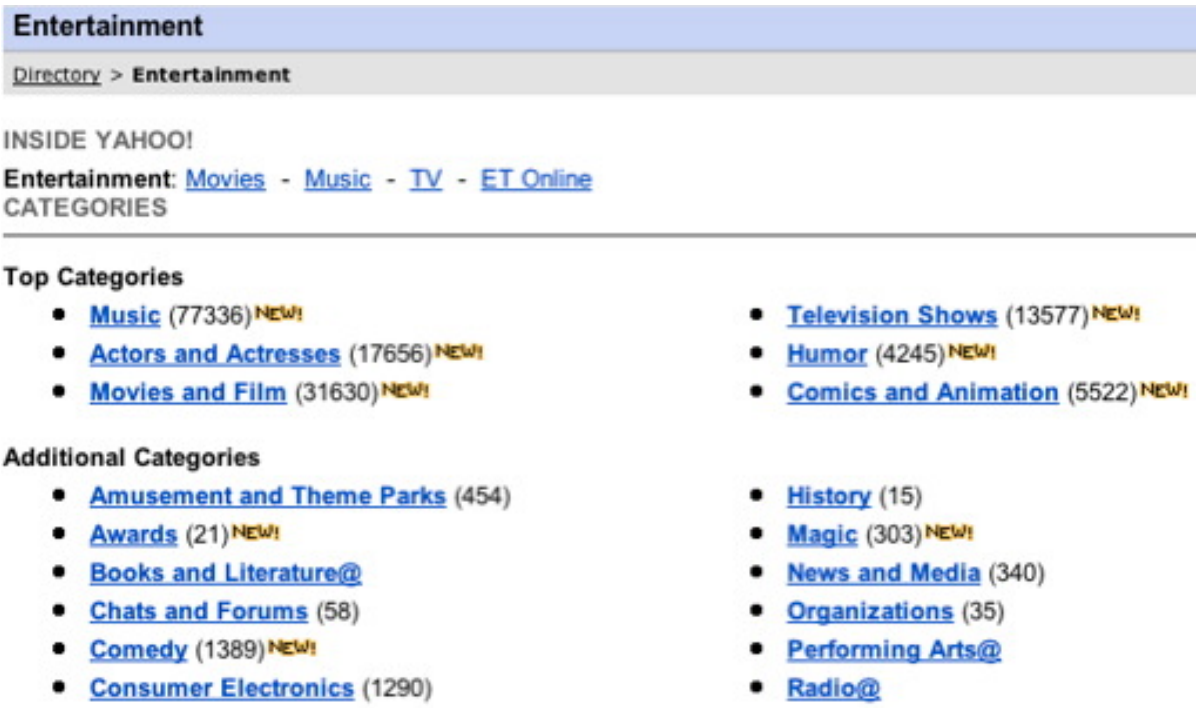
People have been freaking out about the virtuality of data for decades, and you'd think we'd have internalized the obvious truth: there is no shelf. In the digital world, there is no physical constraint that's forcing this kind of organization on us any longer. We can do without it, and you'd think we'd have learned that lesson by now.

And yet.

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

A little over ten years ago, a couple of guys out of 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. As the Web expanded more they realized that, to maintain the value in the directory, they were going to have to systematize, so they hired a professional ontologist, and they developed their now-familiar top-level categories, which go to subcategories, each subcategory contains links to still other subcategories, and so on. Now we have this ontologically managed list of what's out there.

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



Entertainment

Directory > **Entertainment**

INSIDE YAHOO!

Entertainment: Movies - Music - TV - ET Online

CATEGORIES

Top Categories

- Music (77336) 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 (58)
- Comedy (1389) NEW!
- Consumer Electronics (1290)
- History (15)
- Magic (303) NEW!
- News and Media (340)
- Organizations (35)
- Performing Arts@
- Radio@

[Yahoo's Entertainment Category]

You can see what the sub-categories of Entertainment are, whether or not there are new additions, and how many links roll up under those sub-categories. Except, in the case of Books and Literature, that sub-category doesn't tell you how many links roll up under it. Books and Literature doesn't end with a number of links, but with an "@" sign. That "@" sign is telling you 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 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 are a

commercial service, booksellers are 'really' in Business.

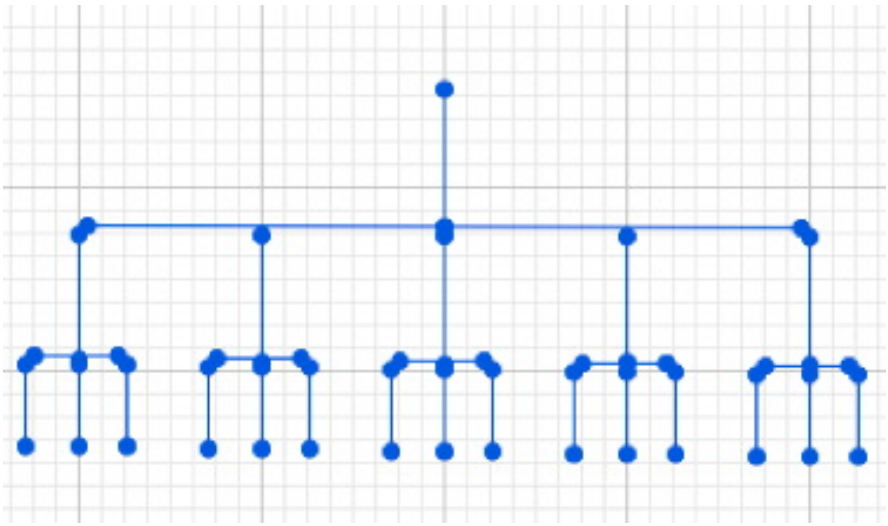
['Literature' on Yahoo]

Look what's happened here. Yahoo, faced with the possibility that they could organize things with no physical constraints, *added the shelf back*. They couldn't imagine organization without the constraints of the shelf, so they added it back. 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 they thought of this kind of a priori organization as their job, and as something their users would value. The uncharitable explanation is that they thought there was business value in determining the view the user would have to adopt to use the system. Both of those explanations may 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.

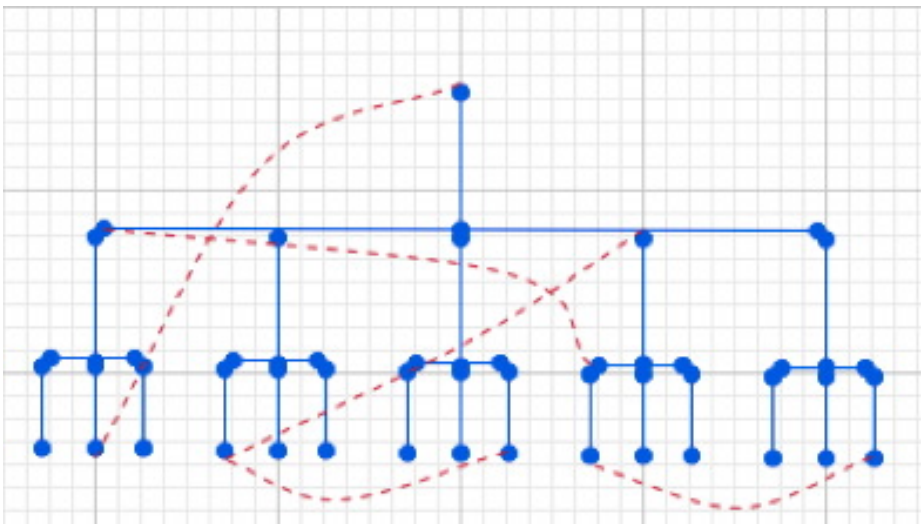
File Systems and Hierarchy #

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



[Hierarchy]

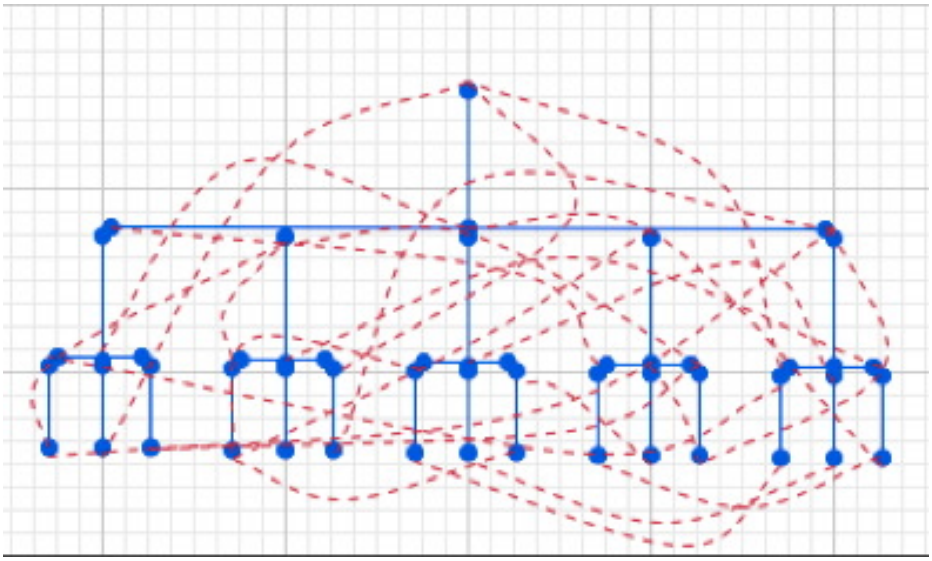
There's a top level, and subdirectories roll up under that. Subdirectories contain files or further subdirectories and so on, all the way down. Both librarians and computer scientists hit the same next idea, which is "You know, it wouldn't hurt to add a few secondary links in here" -- symbolic links, aliases, shortcuts, whatever you want to call them.



[Plus Links]

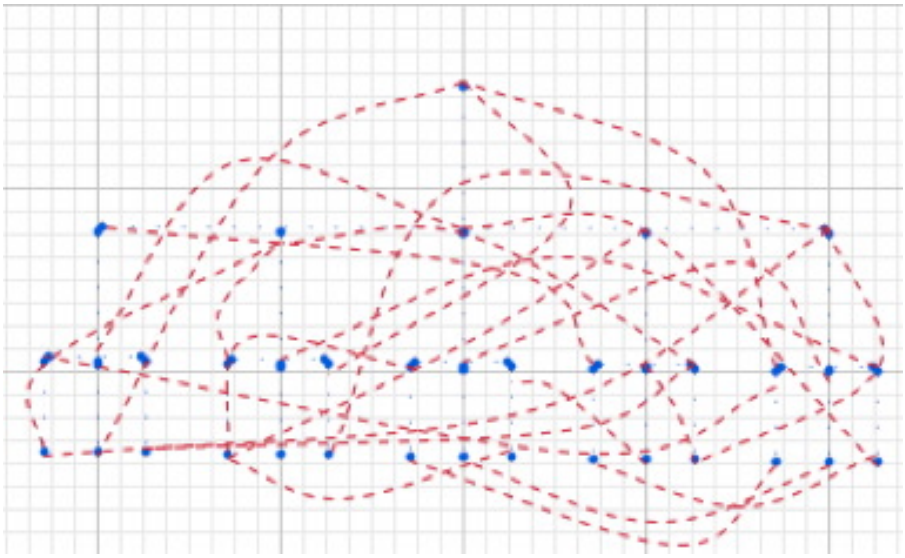
The Library of Congress has something similar in its second-order categorization -- "This book is mainly about the Balkans, but it's also about art, or it's mainly about art, but it's also about the Balkans." Most hierarchical attempts to subdivide the world use some system like this.

Then, in the early 90s, one of the things that Berners-Lee showed us is that you could have a lot of links. You don't have to have just a few links, you could have a whole lot of links.



[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 directory, so they put an upper limit on the number of symbolic links that could go into their view of the world. 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. The links alone are enough.



[Just Links (There Is No Filesystem)]

One reason Google was adopted so quickly when it came along is that Google 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 you need to know.

Let's say I need every Web page with the word "obstreperous" and "Minnesota" in it. 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 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 to organize the world in advance. Given this requirement, the views of the catalogers necessarily override the user's needs and the user's view of the world. If you want something that

hasn't been categorized in the way you think about it, you're out of luck.

The search paradigm says the reverse. It says nobody gets to tell you in advance what it is you need. Search says that, at the moment that you are looking for it, we will do our best to service it based on this link structure, because we believe we can build a world where we don't need the hierarchy to coexist with the link structure.

A lot of the conversation that's going on now 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 Yahoo versus Google example, that there are a number of cases where you get significant value out of *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 catalog if you are managing a physical library. You need a hierarchy to manage a file system. So what you want to 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 it work:

Domain to be Organized

- Small corpus

- Formal categories
- Stable entities
- Restricted entities
- Clear edges

This is all the domain-specific stuff that you would like to be true if you're trying to classify cleanly. The periodic table of the elements has all of these things -- there are only a hundred or so 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 a fit ontology is likely to be.

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

Participants

- Expert catalogers
- Authoritative source of judgment
- Coordinated users
- Expert users

DSM-IV, the 4th version of the psychiatrists' Diagnostic and Statistical Manual, is a classic example of an classification scheme that works because of these characteristics. DSM IV allows psychiatrists all over the US, in theory, to make the same judgment about a mental illness, when presented with the same list of symptoms. There is an authoritative source for DSM-IV, the American Psychiatric Association. The APA gets to

say what symptoms add up to psychosis. They have both expert cataloguers and expert users. The amount of 'people infrastructure' that's 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 problem users have with categories is that when we do head-to-head tests -- we describe something and then we ask users to guess how we described it -- there's 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, and the bigger the user base, the more work that user education is.

You can also turn that list around. You can say "Here are some characteristics where 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 catalogers
- No Authority

If you've got a large, ill-defined corpus, if you've got naive users, if your cataloguers aren't expert, if there's no one to say authoritatively what's going on, then ontology is going to be a bad strategy.

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 more you 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 you have to get to exert over users to get them to drop their own world view in favor of yours.

The reason we know SUVs are a light truck instead of a car is that the Government says they're a light truck. 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 unconciously, that naming the world changes it. Unfortunately, most of the world is not actually amenable to voodoo categorization.

The reason we don't know whether or not *Buffy, The Vampire Slayer* is science fiction, for example, is because there's no one who can say definitively yes or no. In environments where there's 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 jobs that have historically been quite hard: mind reading, and fortune telling. It forces categorizers to guess what their users are thinking, and to 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 people's minds, 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 LiveJournal, which allows user to list their own interests. LiveJournal makes absolutely no attempt to enforce solidarity or a thesaurus or a minimal set of terms, no check-box, no drop-box, just free-text typing. Some people say they're interested in movies. Some people say they're interested in film. Some people 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's no signal in the difference itself, and no value in protecting the user from too many matches.

When we get to really contested terms like queer/gay/homosexual, by this point, 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 til 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 catalogers, about the best way to categorize. They also underestimate the loss from erasing difference of expression, and they overestimate loss from the lack of a thesaurus.

Fortune Telling #

The other big problem is that predicting the future turns out to be hard, and 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'."
```

That 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 actually 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 may have seen it earlier in the Library of Congress scheme, called Former Soviet Union. The best they were able to do was just tack "former" onto that entire zone that they'd previously categorized as the Soviet Union. 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

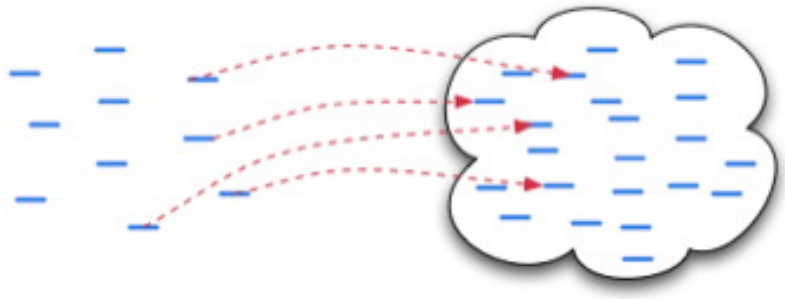
"My God. It's full of links!"

When we reexamine categorization without assuming the physical constraint either of hierarchy on disk or of 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's. (You can tell it's the Library of Congress on the right, because they have a few more books than I do.)

[Two Categorized Collections of Books]

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 all of my books on creativity together." 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?

No, of course we don't have to do anything of the sort. They're able to take my books in while ignoring my categories, because all my books have ISBN numbers, 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.



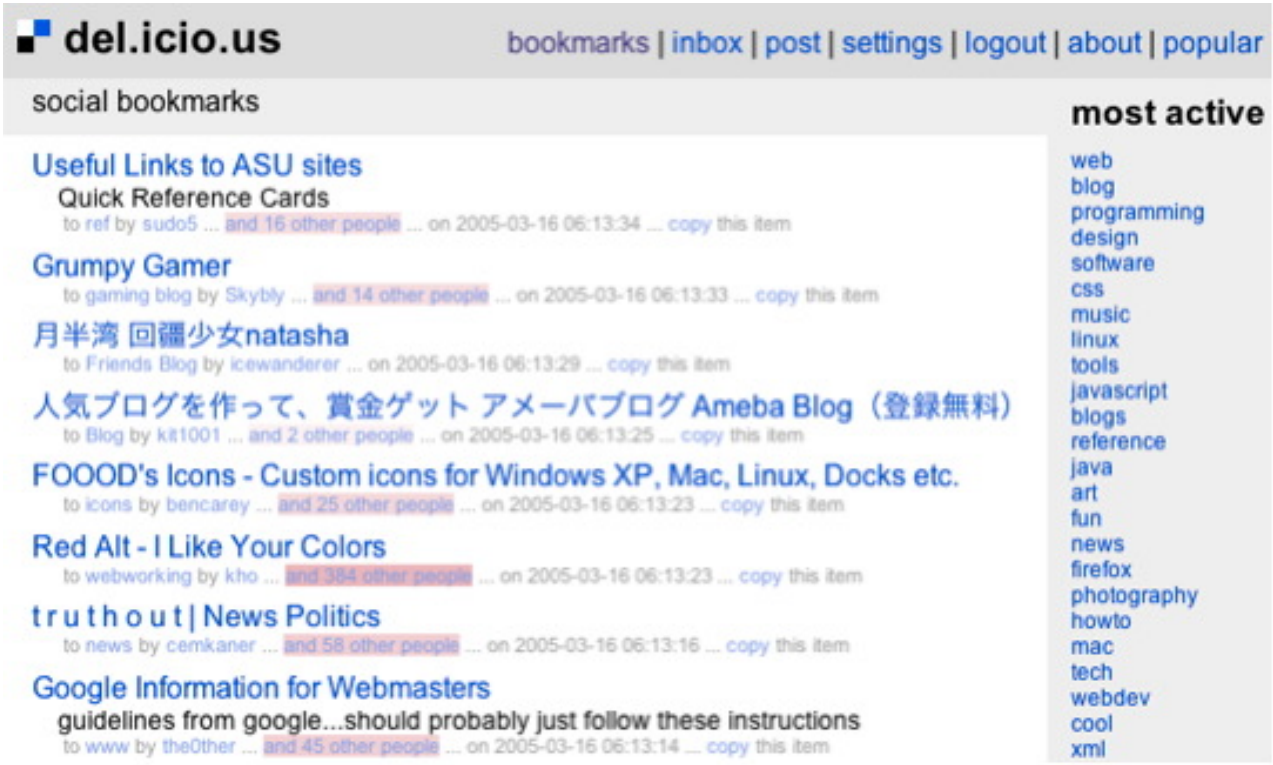
[Merge ISBNs]

Now imagine a world where *everything* can have a unique identifier. This should be easy, since that's the world we currently live in -- the URL gives us a way to create a globally unique ID for anything we need to point to. 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 in later retrieval of those URLs. Tags have the additional effect of grouping related URLs together. 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 forgoing formal classification, tags enable a huge amount of

user-produced organizational value, at vanishingly small cost.

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 those changes afoot, here are some of the things that I think are coming, as 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 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 stuff, 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, they will come to be more valuable than professional categorization schemes, particularly with regards 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 those kind 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 co-exist, 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 is was tagged by' and 'this is when it was tagged, you can start to do 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 base 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 make-up 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 in 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 they way you would, you'll find it -- using a thesaurus to force everyone'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 one 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, it's 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, other people won't read it. But the idea that the filtering is *after* the publishing is incredibly foreign to journalists.

Similarly, the idea that the categorization is done after things are tagged is incredibly foreign to cataloguers. Much of the expense of existing catalogue systems is in trying to prevent one-off categories. With tagging, what you say is "As long as a lot of people 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."

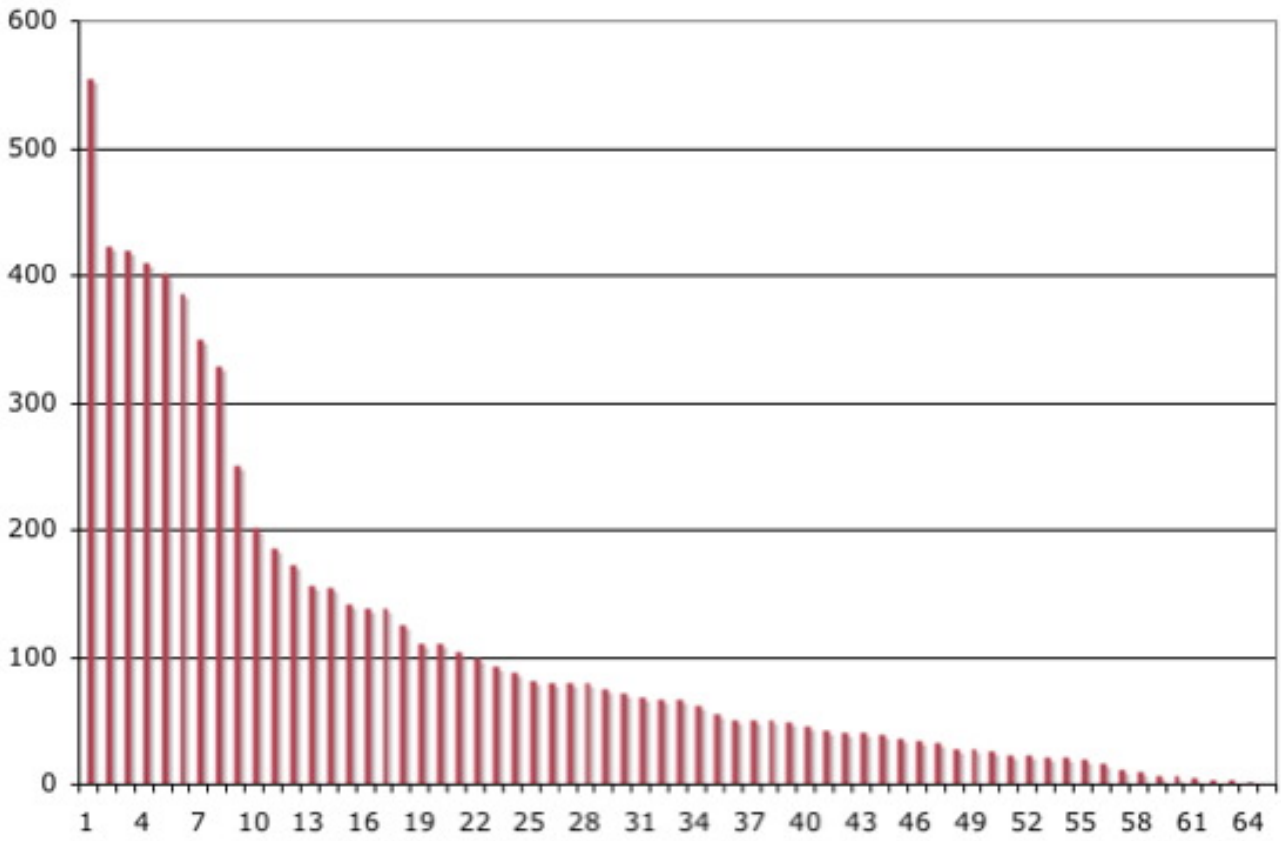
Again, scale comes to the rescue of the system in a way that would simply break traditional cataloging schemes. The existence of an odd or unusual tag is a problem if it's the only way a given link has been tagged, or if there is no way for a user to avoid that tag. Once a link has been tagged more than 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 tagging schemes at the category level and then see what the contents are. As with the 'merging ISBNs' idea, you merge individual contents, because we now have URLs as unique handles. You merge from the URLs, and then try and derive something about the categorization from there. This allows for partial, incomplete, or probabilistic merges that are better fits to 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's something showing what I mean about the breakdown of binary categorization.

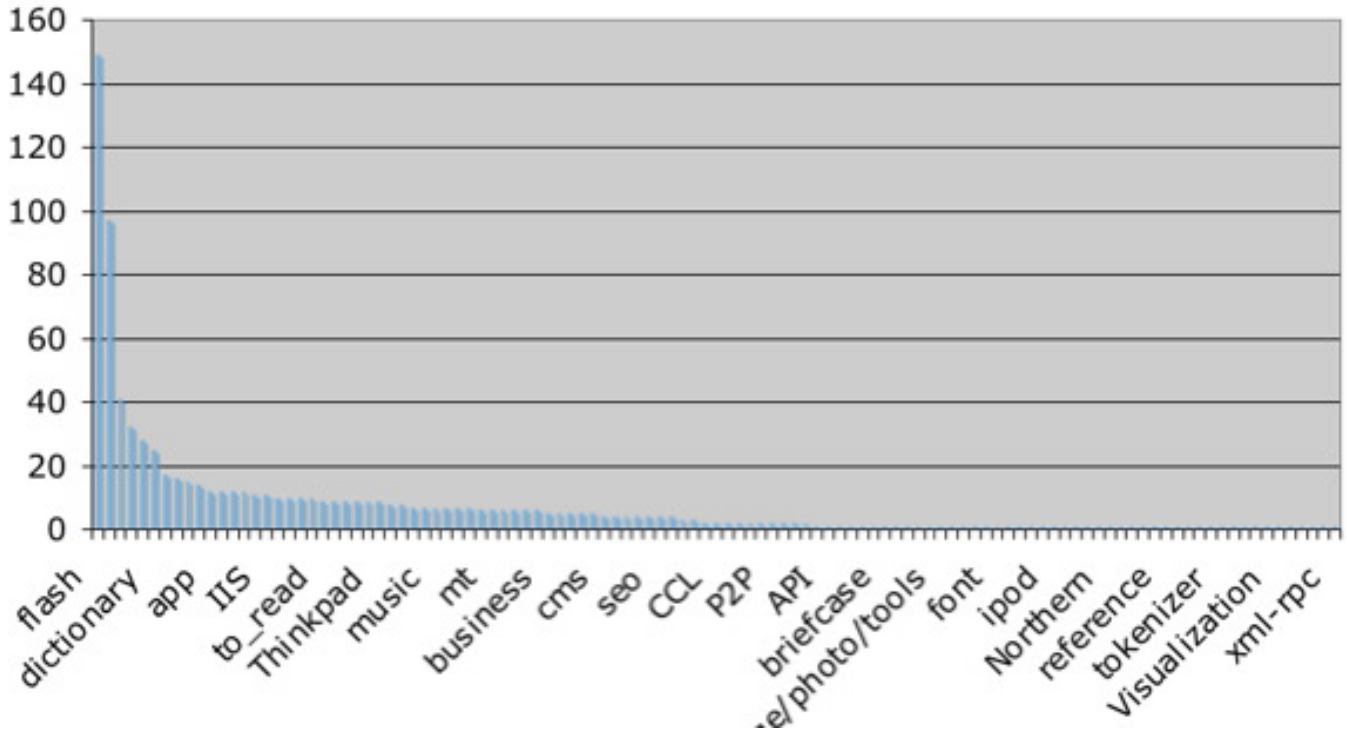


[Tags per user]

This is a chart based on a small sample of links from the del.icio.us front page, taken during a 2-hour window. The X axis is the 64 users who posted links during that period.

The Y axis is the total number of discrete kinds of tags that those users have ever used in their history on del.icio.us.

The chart shows a great variability in tagging strategies among the various users. The user all the way to the left has an enormous number of unique tags, almost 600 of them. Then there's this group of people who are not quite power taggers but who tag quite a bit, and of course to the right of them there's the characteristic long tail 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 goes out quite a bit further than this.) 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.

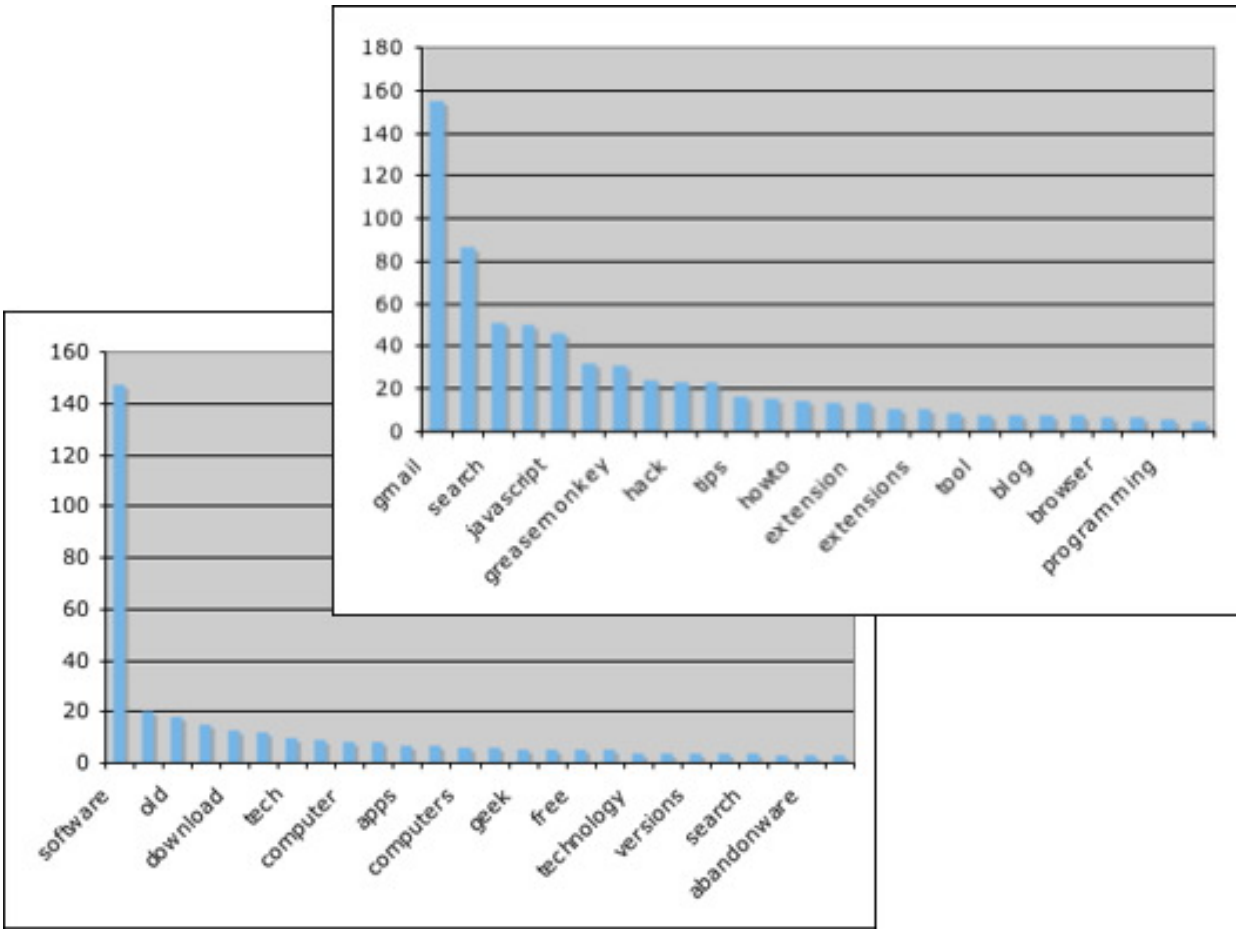


[A single user's tags]

This is a single user's tags. From here, you can tell something about this person -- he or she is obviously a Flash programmer -- the commonest tag here is Flash, followed by a number of other frequently used tags mainly related to programming. Like the front

page, this distribution has the organic signature. Experts don't catalog this way; experts who learn how to catalogue produce much more consistent labeling. Here, it's whatever the user thought would help them remember the link later.

You can see there's a 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.



[Different tag 'signatures' for different URLs]

Then there's this set of graphs. This is to me in a way the most interesting and least well understood part of the del.icio.us right now -- these 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 here that there is broad communal consensus. 140 people tagged this Software. Then, the next commonest tag, with only 20 occurrences, is 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 bit of commonality, there is a really sharp, clear fall off in tags.

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 smearier 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 is, or, rather, occupies more contexts for del.icio.us users than the software download link does.

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 center around a certain core tags, with this degree of certainty, and, thanks to the time stamps, we can even start to understand how the distribution of a URLs tags changes over time. It was 5 years 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, and we're just figuring out how to extract novel value from whole 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 expose it to the users. What instead happened 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 people handle it in practice.

It comes down ultimately to a question of philosophy. Does the world make sense or do we make sense of the world? If you believe the world makes sense, 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, because 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, 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 you try to find ways that the individual sense-making can roll up to something which is of value in aggregate, but you do it without an ontological goal. You do it without a goal of explicitly getting to or even closely matching some theoretically perfect view of the world.

Critically, the semantics here are in the users, not in 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 it's doing instead 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 or not 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 are allowing for and aggregating tags. The signal benefit of these systems is that they don't recreate the structured, hierarchical categorization so often forced onto 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 one another, often without realizing it.

Thank you very much.

Thanks to Alicia Cervini for invaluable editorial help.

clay@shirky.com