Metacrap: Putting the torch to seven straw-men of the meta-utopia

Cory Doctorow
doctorow@craphound.com

Version 1.3: 26 August 2001

BACK TO TOP

---

0. ToC:

BACK TO TOP

0.1. Version History

Version 1.3, August 26 2001. Fixed typos. First published version.

Version 1.2, May 23 2001. Tweaked intro (Thanks, Fred).

Version 1.1, May 18 2001. Changed section orders for better organization. (Thanks, Raffi). Clarified "metrics" in 2.6 (Thanks, Andy).

Version 1.0, May 15 2001. First draft

BACK TO TOP

---

1. Introduction

Metadata is "data about data" -- information like keywords, page-length, title, word-count, abstract, location, SKU, ISBN, and so on. Explicit, human-generated metadata has enjoyed recent trendiness, especially in the world of XML. A typical scenario goes like this: a number of suppliers get together and agree on a metadata standard -- a Document Type Definition or scheme -- for a given subject area, say washing machines. They agree to a common vocabulary for describing washing machines: size, capacity, energy consumption, water consumption, price. They create machine-readable databases of their inventory, which are available in whole or

part to search agents and other databases, so that a consumer can enter the parameters of the washing machine he's seeking and query multiple sites simultaneously for an exhaustive list of the available washing machines that meet his criteria.

If everyone would subscribe to such a system and create good metadata for the purposes of describing their goods, services and information, it would be a trivial matter to search the Internet for highly qualified, context-sensitive results: a fan could find all the downloadable music in a given genre, a manufacturer could efficiently discover suppliers, travelers could easily choose a hotel room for an upcoming trip.

A world of exhaustive, reliable metadata would be a utopia. It's also a pipe-dream, founded on self-delusion, nerd hubris and hysterically inflated market opportunities.

BACK TO TOP

---

2. The problems

There are at least seven insurmountable obstacles between the world as we know it and meta-utopia. I'll enumerate them below:.

BACK TO TOP

2.1 People lie

Metadata exists in a competitive world. Suppliers compete to sell their goods, cranks compete to convey their crackpot theories (mea culpa), artists compete for audience. Attention-spans and wallets may not be zero-sum, but they're damned close.

That's why:

- A search for any commonly referenced term at a search-engine like Altavista will often turn up at least one porn link in the first ten results.
- Your mailbox is full of spam with subject lines like "Re: The information you requested."
- Publisher's Clearing House sends out advertisements that holler "You may already be a winner!"
- Press-releases have gargantuan lists of empty buzzwords attached to them.

Meta-utopia is a world of reliable metadata. When poisoning the well confers benefits to the poisoners, the meta-waters get awfully toxic in short order.

BACK TO TOP

2.2 People are lazy

You and me are engaged in the incredibly serious business of creating information. Here in the Info-Ivory-Tower, we understand the importance of creating and maintaining excellent metadata for our information.

But info-civilians are remarkably cavalier about their information. Your clueless aunt sends you email with no subject line, half the pages on Geocities are called "Please title this page" and your boss stores all of his files on his desktop with helpful titles like "UNTITLED.DOC."

This laziness is bottomless. No amount of ease-of-use will end it. To understand the true depths of meta-laziness, download ten random MP3 files from Napster. Chances are, at least one will have no title, artist or track

information -- this despite the fact that adding in this info merely requires clicking the "Fetch Track Info from CDDB" button on every MP3-ripping application.

Short of breaking fingers or sending out squads of vengeful info-ninjas to add metadata to the average user's files, we're never gonna get there.

BACK TO TOP

2.3 People are stupid

Even when there's a positive benefit to creating good metadata, people steadfastly refuse to exercise care and diligence in their metadata creation.

Take eBay: every seller there has a damned good reason for double-checking their listings for typos and misspellings. Try searching for "plam" on eBay. Right now, that turns up *nine* typoed listings for "Plam Pilots." Misspelled listings don't show up in correctly-spelled searches and hence garner fewer bids and lower sale-prices. You can almost always get a bargain on a Plam Pilot at eBay.

The fine (and gross) points of literacy -- spelling, punctuation, grammar -- elude the vast majority of the Internet's users. To believe that J. Random Users will suddenly and *en masse* learn to spell and punctuate -- let alone accurately categorize their information according to whatever hierarchy they're supposed to be using -- is self-delusion of the first water.

BACK TO TOP

2.4 Mission: Impossible -- know thyself

In meta-utopia, everyone engaged in the heady business of describing stuff carefully weighs the stuff in the balance and accurately divines the stuff's properties, noting those results.

Simple observation demonstrates the fallacy of this assumption. When Nielsen used log-books to gather information on the viewing habits of their sample families, the results were heavily skewed to *Masterpiece Theater* and *Sesame Street*. Replacing the journals with set-top boxes that reported what the set was actually tuned to showed what the average American family was really watching: naked midget wrestling, *America's Funniest Botched Cosmetic Surgeries* and Jerry Springer presents: "My daughter dresses like a slut!"

Ask a programmer how long it'll take to write a given module, or a contractor how long it'll take to fix your roof. Ask a laconic Southerner how far it is to the creek. Better yet, throw darts -- the answer's likely to be just as reliable.

People are lousy observers of their own behaviors. Entire religions are formed with the goal of helping people understand themselves better; therapists rake in billions working for this very end.

Why should we believe that using metadata will help J. Random User get in touch with her Buddha nature?

BACK TO TOP

2.5 Schemas aren't neutral

In meta-utopia, the lab-coated guardians of epistemology sit down and rationally map out a hierarchy of ideas, something like this:

```
Nothing:
    Black holes

Everything:
    Matter:
        Earth:
            Planets
            Washing Machines
        Wind:
            Oxygen
            Poo-gas
        Fire:
            Nuclear fission
            Nuclear fusion
            "Mean Devil Woman" Louisiana Hot-Sauce
```

In a given sub-domain, say, Washing Machines, experts agree on sub-hierarchies, with classes for reliability, energy consumption, color, size, etc.

This presumes that there is a "correct" way of categorizing ideas, and that reasonable people, given enough time and incentive, can agree on the proper means for building a hierarchy.

Nothing could be farther from the truth. Any hierarchy of ideas necessarily implies the importance of some axes over others. A manufacturer of small, environmentally conscious washing machines would draw a hierarchy that looks like this:

```
Energy consumption:
    Water consumption:
        Size:
            Capacity:
                Reliability
```

While a manufacturer of glitzy, feature-laden washing machines would want something like this:

```
Color:
    Size:
        Programmability:
            Reliability
```

The conceit that competing interests can come to easy accord on a common vocabulary totally ignores the power of organizing principles in a marketplace.

BACK TO TOP

2.6 Metrics influence results

Agreeing to a common yardstick for measuring the important stuff in any domain necessarily privileges the items that score high on that metric, regardless of those items' overall suitability. IQ tests privilege people who are good at IQ tests, Nielsen Ratings privilege 30- and 60-minute TV shows (which is why MTV doesn't show videos any more -- Nielsen couldn't generate ratings for three-minute mini-programs, and so MTV couldn't demonstrate the value of advertising on its network), raw megahertz scores privilege Intel's CISC chips over Motorola's RISC chips.

Ranking axes are mutually exclusive: software that scores high for security scores low for convenience, desserts that score high for decadence score low for healthiness. Every player in a metadata standards body wants to emphasize their high-scoring axes and de-emphasize (or, if possible, ignore altogether) their low-scoring axes.

It's wishful thinking to believe that a group of people competing to advance their agendas will be universally pleased with any hierarchy of knowledge. The best that we can hope for is a *detente* in which everyone is

equally miserable.

BACK TO TOP

2.7 There's more than one way to describe something

"No, I'm not watching cartoons! It's *cultural anthropology*."

"This isn't smut, it's *art*."

"It's not a bald spot, it's a *solar panel for a sex-machine*."

Reasonable people can disagree forever on how to describe something. Arguably, your Self is the collection of associations and descriptors you ascribe to ideas. Requiring everyone to use the same vocabulary to describe their material denudes the cognitive landscape, enforces homogeneity in ideas.

And that's just not right.

BACK TO TOP

---

3. Reliable metadata

Do we throw out metadata, then?

Of course not. Metadata can be quite useful, if taken with a sufficiently large pinch of salt. The meta-utopia will never come into being, but metadata is often a good means of making rough assumptions about the information that floats through the Internet.

Certain kinds of implicit metadata is awfully useful, in fact. Google exploits metadata about the structure of the World Wide Web: by examining the number of links pointing at a page (and the number of links pointing at each linker), Google can derive statistics about the number of Web-authors who believe that that page is important enough to link to, and hence make extremely reliable guesses about how reputable the information on that page is.

This sort of observational metadata is far more reliable than the stuff that human beings create for the purposes of having their documents found. It cuts through the marketing bullshit, the self-delusion, and the vocabulary collisions.

Taken more broadly, this kind of metadata can be thought of as a pedigree: who thinks that this document is valuable? How closely correlated have this person's value judgments been with mine in times gone by? This kind of implicit endorsement of information is a far better candidate for an information-retrieval panacea than all the world's schema combined.