

Recycled Knowledge

by John Cowan <cowan@ccil.org>

2007-08-13

Extreme Markup 2007: Thursday

This is a report on Extreme Markup 2007 for Thursday.

My first talk of the day was on a fast XSLT processor written in C++ and designed for large documents (more than 2 gigs) and high performance. Documents are loaded into memory using event records, which are essentially just reifications of SAX events, with parent and next-sibling links. Because internal links are record numbers rather than raw pointers, the internal limit is 2 GB of events rather than bytes (I think that's the reason). The authors did an interesting experiment with compiling multiple XPath's into a single DFA and executing them in parallel, but found that overall performance was not really improved by this additional complexity.

The next presentation was by Eric Freese, about chatbots and RDF. The chatbot here uses AIML, a language for representing chat scripts, and enhances it by allowing RDF triples (notably Dublin Core and FOAF information) to be stuffed into the bot so it knows a lot more about a lot more. The bot, called ALICE, follows the classical Eliza/Doctor tradition; it applies rewrite rules to the input to produce its output. (Oddly, Eric had believed that Eliza was hardcoded rather than script-driven until I pointed out otherwise during the comment period; to watch the Doctor script in action, fire up emacs and type "Alt-x doctor", or type "Alt-x psychoanalyze-pinhead" and hit Ctrl-G after a few seconds.)

James David Mason told us all about the application of topic maps to very complex systems, in this case pipe organs. Organs are even more complicated than they appear to be from the outside. (The reviewers, we were told, tended to say things like "I wouldn't attend this talk because I'm not interested in organs"; I think they missed the point. There were, however, wonderful multimedia interruptions of the talk in the form of excerpts from music being played on the organs under discussion. Usually James tells us censored versions of stories from Y-12, the secret national laboratory at Oak Ridge, Tennessee. (They're starting to make things with FOGBANK again after many years of not using it, and it's a big concern.) This time he

About Me

John Cowan

My friends say that I know at least something about practically everything; my enemies, that I know far too much about far too much. Here's the raw material for believing in both views. This is a personal blog. What I say here doesn't represent the views of my employer, Santa Claus, or anyone else. You can email me at cowan@ccil.org. Spam will be aggressively filtered.

[VIEW MY COMPLETE PROFILE](#)

Recent Comments

On Jan 23 John Cowan commented on blood groups and true parents: *"Anonymous #31 said: I was wondering... if I am a positive and my wife is A positive, could u have a..."*

On Jan 17 Andrea commented on blood groups and true parents: *"Hiya please help me i need some answers my Mum is B+ and my father is O+ is my dad step or my real..."*

On Jan 14 Anonymous commented on blood groups and true parents: *have A positive blood, my 3 younger siblings have A negative blood. My mother was pregnant when..."*

talked about his non-work instead of non-talking about his work; it was great.

Next I heard some folks from Portugal talking about digital preservation of the Portuguese national and local archives. They use PDF/A to preserve documents of all sorts, but they also need to preserve the contents of relational databases, with tables, views, triggers, stored procedures, and all. Dumping them as SQL has serious portability problems, so they decided to carefully define a database markup language to handle all the metadata as well as the data itself. That way, information can be authentically transferred from one database engine to another without loss.

The last (and plenary) presentation of the day had not a speaker but "listeners" from W3C, OASIS, and ISO, asking what things should be standardized in addition to what we already have. Audience members concentrated on processes for stabilizing and withdrawing standards (ISO has this, the other two don't), the desire for more and earlier feedback processes and fewer *faits accomplis*, and other meta-issues. There were a few requests for additional concrete standards; unfortunately, I don't have notes. The session ended without too many real fights.

On Wednesday I talked about the nocturne on naming, but it was actually held on Thursday. Much of it was spent discussing the FRBR model of works, expressions, manifestations, and items. For computer-document purposes, a work is the abstract work, e.g. *Hamlet*; an expression is a particular kind realization, like a particular edition of the text or recording of a performance; a manifestation is the expression in a particular format such as .txt, .html, .doc, or .avi; and an item is a particular copy residing on a particular server, disk, or tape. Ideally, there should be separate URIs for each of these things.

A commenter asked what about topic maps I was defending: primarily the clear topic-map distinction between a URI used to name a document (a *subject locator* in topic maps jargon) and the same URI used to refer to the subject matter of the document (a *subject indicator*). Theoretically, RDF uses different URIs for these separate purposes, but that leads to paradoxes -- the document you get when you GET a subject indicator, what's the URI that identifies that? Furthermore, how do you tell which is which? In the topic-maps data model, a topic simply has some URIs which are locators and others which are indicators.

Posted by John Cowan at 15:03 

On Jan 09 John Cowan comments on essentialist explanations 12th edition: "*Pilgrim: Actually no, I haven't written more than that. I sometimes put my own name on things that...*"

On Jan 08 びつ < 1) commented c tagsoup 12 released at long last: "*I was surprised that you identified your employer, Santa Claus, in your about paragraph. I thought.*"

Add this widget to your blog

Blog Archive

- ▼ 2008 (3)
 - ▼ January (3)
 - Taggle, a TagSoup in C++, available now
 - Revised home page
 - TagSoup 1.2 released at long last
- 2007 (26)
- 2006 (36)
- 2005 (159)