B Sides
http://ir.uiowa.edu/bsides

U of I SLIS Journal

01|12|2011

Why I Can’t Love the Homemade Semantic Web

Jason A. Bengtson

Abstract:

Almost all information professionals agree that the web needs to move to a semantic structure. While work is proceeding in this area, movements to get individual web authors to use semantic markup tools have also been on the rise. This author argues that such efforts are ill conceived and he proposes an automated alternative.

Keywords:
semantic | web | internet | search | data | mining

B Sides Spring 2011
http://ir.uiowa.edu/bsides/20
U of I SLIS Journal

Why I Can't Love the Homemade Semantic Web

By Jason Bengtson, MLIS

What is the homemade semantic web? The homemade semantic web is a dream by web designers and librarians; a utopian ideal that extends the tantalizing promise of a World Wide Web with machine readable architecture applied from the ground up. Put simply, it is the idea that web authors will be successfully compelled to mark up their own creations with some standardized form of metadata (information about information) so that the resulting web documents can be more easily read and manipulated by computer applications. It is an optimistic notion, and it is one that is doomed to fail.

One of the old movies that I love to dust off every once in a while is the sci-fi classic *Runaway*. For those of you who don't remember or who haven't seen it, Tom Selleck, Kirstie Alley, and Gene Simmons, deal with robots run amok. In the film, the protagonist is a police officer on the "Runaway squad"; a part of the force tasked with dealing with rogue, malfunctioning devices in a near future world full of robots. Gene Simmons is the delightfully over the top villain, who uses a combination of heat-seeking bullets and toaster shaped spider robots to strike at his enemies, while at the same time orchestrating his plan to insert special chips in robots everywhere that will make them do what he wants[1].

For me, one of the most enjoyable aspects of this movie is the presence of the many anachronisms. As such, the most noticeable incongruity is the robots themselves. These aren't the increasingly biological creations that we're seeing now, but the old eighties ideal of a robot, from an era in technology when designers were increasingly trying to convince the public that robots *shouldn't* look biological. Instead, those designers reasoned, robots should look like the tools they are. We should expect to see little tanks with a claw arm, or a metal cylinder with a crude voice synthesizer. Robots were things that did a job, and anthropomorphic and biomorphic considerations were things that the public needed to relegate to the realm of fantasy. So why don't we see more of those kinds of robots being developed? There are a few, certainly, especially as used by industry and the military. But increasingly, robots are looking, if not more human, more biological.

The answer is simple: biology works for the real world. This is a world that the biological has adapted itself to function in, and function well. In our own segment of this world, grown out of the complexities of human intellect and needs, it is the human form that works. After all, our technologies, toys and various other trinkets were designed to be used by the human body and the human mind. In *Runaway* there is a particularly hilarious scene where a household robot is terrorizing a home with a baby in it after appropriating the owner's revolver out of a dresser drawer. If anything approaches the area, it shoots. The robot in question is essentially a low-sitting box on tracks with a robot arm. There is never any explanation given about: A. How it managed to see over the edge of the drawer, much less successfully fish around in it with its claw to get the gun or B. How it plans to, with only one arm, reload the six shot pistol cylinder after all the rounds are discharged. As the brave hero dons body armor and enters the home to take action, we are left to wonder why someone doesn't just throw rocks at the house until the robot runs out of bullets.

Biology works on the physical level, and we would do well to emulate its success. But all too often, in our digital age, I feel as if some in the Information Technology and Information Science fields are trying to get us to take a step back to eighties thinking when considering informational questions. The homemade semantic web is a perfect example. The idea with a semantic web is to design websites and applications with the use of metadata that will make them more readable to machines. On its face this is a way to revolutionize information transfer,

[1] Crichton, *Runaway*.

web publishing, and unnumbered other applications. It sounds good, and there's a lot of potential within this idea. The research done on social networking and the use of semantic mechanisms for web publishing or other applications is impressive. But there has also been a movement going to get individual web authors to mark their work up semantically, and in this area I can't help but feel like we're making a mistake.[2] [3] Shouldn't our tools be designed to work in our world, to see the world as we see it? Shouldn't we be designing tools that function in the human world, instead of trying to get humans to adapt themselves to those tools?

Of course, since we have created this new level of complexity by creating a mechanical world, we have to understand that within the world of the machine, things need to be done in a mechanical way. It doesn't make sense to have a factory robot that's built to look like an auto plant worker. A factory isn't, by nature, the human world. It's only housed humans by necessity, due to the limitations of our past technology. Robots in a factory should be built to function, because there is no inherent need for them to have human capabilities. The only places where there must be human friendly interfaces are those parts of the facility where the mechanical world touches the human world again. The interface to program the robots needs to be one that humans can use effectively. The cars that roll off of the assembly line need to have their production information formatted in a way that humans can make sense of. *When they interact with us, machines need to be built around our needs. When they don't, they need to be allowed to do what they do in a way that works for the machine*. That's one distinction that is fundamental to a harmonious set of interactions between people and their tools. Here's another: *Machines need to make our information into something they can use and they need to make their information into something we can use because they're the tools. Making things easier for humans is their job, making things easier for machines shouldn't be ours.*

As a librarian, I'm well acquainted with metadata. Metadata, or information about information, is vital to cataloging and information recovery because it allows us to find a particular item and put it to practical use in much the same way that labels on boxes in a supermarket help us determine the difference between a box of laundry detergent and a container of breakfast cereal. Metadata is normally field driven. That is to say, a nominal metadata record will contain fields for things like title, author, subject headings, etc. Human beings fill out these fields in a record and then assign the record as a sort of digital proxy for the item itself.

This is how things are done in libraries, and many would like to see things done this way on the web, with web documents assigned some sort of common metadata schema so that they could be more easily found and put to use. Currently much of the web still operates through keyword searching of barely formatted documents, which, as any librarian will tell you, is often less effective than a search based on a more organized set of metadata. Even more complex web searches, using the tags in HTML as a kind of metadata, are often problematic because people don't always apply those tags consistently, and tags can often mean more than one thing. Hence the creation of XML, Tag Clouds, Dublin Core, and other schemes designed to make the information in a web document (and on other non-HTML documents posted to the web) more searchable via machine.

More important than improving the searchability of the web, however, is the task of making the web more generally machine readable. It can currently be difficult to design applications in which software interfaces in a meaningful way with the web, because too many of the web documents to be found are designed to be read only by humans and lack the metadata tags that are essential for allowing a software application to understand exactly what

[2] Karandikar, Nitin, *Semantic Web: Where Are The Meaning-Enabled Authoring Tools?*

[3] Syn, Sue Yeon and Michael B. Spring, *Can a System make Novice Users Experts?; Analysis of Metadata Created by Novices and Experts with Varying Levels of Assistance*.

it's reading. Metadata, or *semantic markup* would make the documents more readable to machine applications, by making available explicit information about the content and intent of the document. Many designers and web experts seem to regard this semantic web as an essential next step in making the web more useful.[4] [5]

And let's be clear; pretty much everybody agrees that some kind of semantic web is a good idea. If we want the web to be more useful and if we ever want the web to behave more like a database than the bulletin board at the laundromat down the street, we're going to need to add a semantic dimension to it. New tools like OWL, Dublin Core, and even far more user friendly (if less semantically useful) applications like Microformats and tag clouds, have been developed to expedite a new semantic web, but who enforces the use of these tools? Who ensures their consistency and accuracy? How do we establish a common format and keep it up to date? If we simply make the tools available, without seeing to these other aspects of implementation, the semantic web may have a grim future.

Humans, unlike machines, don't normally read or reason by use of explicit metadata. We can determine, based on known conventions and the construction of the document, what part is the title, what part is the name of the author, who the publisher is, and so on. In fact, *the need to determine those things is the reason that those conventions were created in the first place*. Because we, as human beings, attach meaning to the title of a piece, or its creator, we have created a format so we know within the document where that information can be found. It seems to me to be a more useful approach to make machines that can use our conventions in ways that have meaning to them, not the other way around. Computers need to learn from people, and operate in ways that are similar to the way people operate, when using information from the human world. I'm of course not suggesting that every application out there needs to have a natural language comprehension component. Instead I think we need to automate the production of metadata.

The idea of making the world around machines and of making men conform to machines, is nothing new. Our culture has been deeply concerned with the mechanizing effects of the modern world on human beings ever since the industrial revolution. Examples of angst in this area abound, as seen so explicitly in the curious "man factory" that Twain's protagonist creates in "A Connecticut Yankee in King Arthur's Court". We see within it an allusion to some mythical standard formula by which a man of worth can be created from "common stock", shaped and molded by factory conditions[6]. But humans are not made from interchangeable parts. Only their works can be (often imperfectly) reproduced in such a way. Because we are imperfect and imprecise, our modes of communication often are as well. A fact that, ironically, lends to them a depth and meaning that is probably impossible to achieve with the sterile language of the machine. While we can probably never teach a machine the tragic nature of Macbeth, or the epic strivings of the Odyssey, we can teach machines to see the structural elements in a work in something like the way we do. Such a technique seems more likely to bear fruit than the alternative, and it is an approach that has already seen extensive development.

'Traditionally', metadata is produced in one of two ways; from a primary source, or from a cataloger. The primary source, such as the author or publisher, can directly transmit the "official information" in a top down fashion, populating the metadata fields themselves. These entities should, after all, have at least some idea of how to assign subject headings; answering that most basic question of "what is this about?". The alternative is to have a cataloger decide how to populate the metadata fields, and to do this the cataloger must rely on those old human

[4] Al-Feel, Haytham, M. A. Koutb, and Hoda Suoror, *Toward an Agreement on Semantic Web Architecture*.

[5] Yee, Martha M, *Can Bibliographic Data be Put Directly Onto the Semantic Web?*

[6] Twain and Twain, *A Connecticut Yankee in King Arthur's Court*.

conventions. This is the title and this is not, because of position and font size in relation to the rest of the text. The book is so large and so long. It is about this or that, based on the information from the cover, or information from others about the book, or based on the cataloger's perusal of the material. In any case something human is made into something mechanical, like the freakish creation of a science fiction story, where a human explorer is turned into a lurching, parody of the biological by some arcane, alien process. And what of our example of the author making his own metadata? Humans create constantly, but is it realistic to expect them to sit still and document each creation in minute detail? The idea that a legion of web authors will pause after each frenetic act of creation to carefully document their work according to Dublin Core, or any other standard seems unlikely, while the idea that such documentation would be reliably carried out and consistently produce usable information seems absurd. One example of what we are seeing to this point can be found in the article, *Can We Trust Webpage Metadata*, by Anders Ardo, where webpage metadata was found to be highly questionable.[7]

Alternatives are possible, and the advances in the field of natural language translation have been impressive. Data mining and related tools are getting better every day at translating the confusing language of humans into something that makes sense in the digital realm.[8] New applications are generating metadata from fields found within documents or search boxes by using the same conventions relied upon by humans.[9] [10] Applications like Zotero and Google search can then ferret out the metadata in markup tags, snapping up this information, be it human or machine generated, with alacrity.[11] The World Wide Web acts almost as another player in this effort, providing the brute force needed, through the analysis of masses of human-generated material, to make such translational efforts possible for the machine. This is language learning through total immersion, the application of an environment never available and possibly never imagined by the mechanistic champions of eighties robotics. Instead of making human speech into the speech of machines so that we can eventually convert it back again, we will teach the machine to speak and interact with us. The machines then, from that understanding, can create metadata, if we wish. Like visitors from another world they will take our speech, discuss it among themselves in their own, and create a new level of complexity that they can use to describe it, before returning it us in ways that allow us to better use it.

Of course, as I use terms like "understanding" I don't mean them in the sense of true sapience. I'm anthropomorphizing a bit. We can (and do) create software that recognizes the same conventions we use as humans, before translating the data from that context into machine readable metadata. To put it another way, instead of having humans compile metadata, we would have machines do it.

After all, all modern searching relies on some form of metadata. When we search a database, or the web for that matter, we don't really search either. We don't have the processing technology or bandwidth to hope to do that. Instead we're searching indexes generated in both. In academic databases it's largely the article metadata that's indexed, whereas with the web keyword indexes are created from web pages grabbed off of the web by "bots" or "crawlers". The index itself forms a variety of metadata, but there's no reason why the

[7] Ardo, Anders, *Can we Trust Web Page Metadata?*

[8] Zaremba, Sam, Mila Ramos-Santacruz, Thomas Hampton, Panna Shetty, Joel Fedorko, Jon Whitmore, John M. Greene, et al., *Text-Mining of PubMed Abstracts by Natural Language Processing to Create a Public Knowledge Base on Molecular Mechanisms of Bacterial Enteropathogens*.

[9] Karande, N. D. and G. A. Patil, *Natural Language Database Interface for Selection of Data using Grammar and Parsing.*

[10] Han, Lixin, Guihai Chen, and Li Xie, *AASA: A Method of Automatically Acquiring Semantic Annotations*.

[11] Beall, Jeffrey, *How Google Uses Metadata to Improve Search Results*.

software used by search engine providers couldn't break pages into more structured metadata formats before generating indexes. Instead of relying on the HTML tags (which could still contribute on a limited level), or other pre-structured data already present in the document, systems employing more sophisticated data mining and natural language translation would be used. It would require even more storage and server space than the gargantuan facilities currently in use, but it could be done. We could even have bots crawl the web and create metadata, attaching it to the many web pages that need it, as seen in the paper by Yang and Lee.[12] And unlike some utopian vision of a standardized web, this is a version of the homemade semantic web that could happen.

This solution wouldn't be perfect, but it would be a much more realistic solution than trying to turn web authors into librarians and humans into translators for machines. It also has the advantage of being more easily upgraded. If the process is automatic, and largely invisible, then as the technology for better interpretation and metadata solution presents itself it can be implemented automatically, with our ubiquitous web crawlers being programmed to upgrade the previously generated, lower quality files that they detect. In this way we let machines generate the data that they need to make the web work.

Later on in his life, Mark Twain, originator of the chilling concept of the "man factory", threw a fortune away in the pursuit of the Paige typesetting machine. The mechanism was an ultimately fruitless attempt to set type by replicating the movements of a human typesetter[13]. Years later, the Linotype machine made automatic typesetting possible by processing the type in a whole new way, a way that was purely mechanical rather than an attempt to emulate the human typesetter. The Linotype machine used a keyboard to translate human data into a mechanical format, and after the keys were punched the machine built a plate using mechanical processes[14]. Paige and Twain had overlooked something fundamental: the human typesetter had adapted to do something mechanical. As such, having a machine copy a human typesetter was a fundamentally flawed solution. The machine needed to process that information and that task in a way that made sense mechanically; a way that transmitted the information effectively to other mechanical components. It wasn't until that information passed back into the human realm that it needed to exist, once more, in a human friendly format. We need to take a lesson from this. Machines do a better job of interacting with machines than people do. We should let them do it. We should let machines generate the information that machines need. And we should create smarter machines that are compatible with us instead of focusing on making ourselves more compatible with our creations.

[12] Hsin-Chang Yang, and Chung-Hong Lee, *Automatic Metadata Generation forWeb Pages Using a Text Mining Approach*.

[13] "The Paige Typesetting Machine."

[14] "The Linotype."

Crichton, Michael. Runaway. Sony Pictures, 2000.

Karandikar, Nitin. “Semantic Web: Where Are The Meaning-Enabled Authoring Tools?,” n.d. http://www.readwriteweb.com/archives/semantic_web_authoring_tools.php.

Syn, Sue Yeon and Michael B. Spring. "Can a System make Novice Users Experts?; Analysis of Metadata Created by Novices and Experts with Varying Levels of Assistance." Int.J.Metadata Semant.Ontologies 3, no. 2 (December, 2008): 122-131.

Al-Feel, Haytham, M. A. Koutb, and Hoda Suoror. "Toward an Agreement on Semantic Web Architecture." Proceedings of World Academy of Science: Engineering & Technology 49, (02, 2009): 806-810.

Yee, Martha M. "Can Bibliographic Data be Put Directly Onto the Semantic Web?" Information Technology & Libraries 28, no. 2 (06, 2009): 55-80.

Twain, Mark, and Mark Twain. A Connecticut Yankee in King Arthur's Court. Berkeley: Published in cooperation with the University of Iowa [by] University of California Press, 1983.

Ardo, Anders. "Can we Trust Web Page Metadata?" Journal of Library Metadata 10, no. 1 (Jan, 2010): 58-74.

Zaremba, Sam, Mila Ramos-Santacruz, Thomas Hampton, Panna Shetty, Joel Fedorko, Jon Whitmore, John M. Greene, et al. . Text-Mining of PubMed Abstracts by Natural Language Processing to Create a Public Knowledge Base on Molecular Mechanisms of Bacterial Enteropathogens.(Database)(Report). Vol. 10, 2009RP: Not in File; undefined.

Karande, N. D. and G. A. Patil. "Natural Language Database Interface for Selection of Data using Grammar and Parsing." Proceedings of World Academy of Science: Engineering & Technology 59, (11, 2009): 484-487.

Han, Lixin, Guihai Chen, and Li Xie. "AASA: A Method of Automatically Acquiring Semantic Annotations." Journal of Information Science 33, no. 4 (08, 2007): 435-450.

Beall, Jeffrey. "How Google Uses Metadata to Improve Search Results." Serials Librarian 59, no. 1 (07, 2010): 40-53.

Hsin-Chang Yang, and Chung-Hong Lee. “Automatic Metadata Generation forWeb Pages Using a Text Mining Approach.” In Web Information Retrieval and Integration, 2005. WIRI '05. Proceedings. International Workshop on Challenges in, 186-194, 2005.

“The Paige Typesetting Machine.” http://etext.virginia.edu/railton/yankee/cymach6.html.

“The Linotype.” http://www.woodsidepress.com/LINOTYPE.HTML.