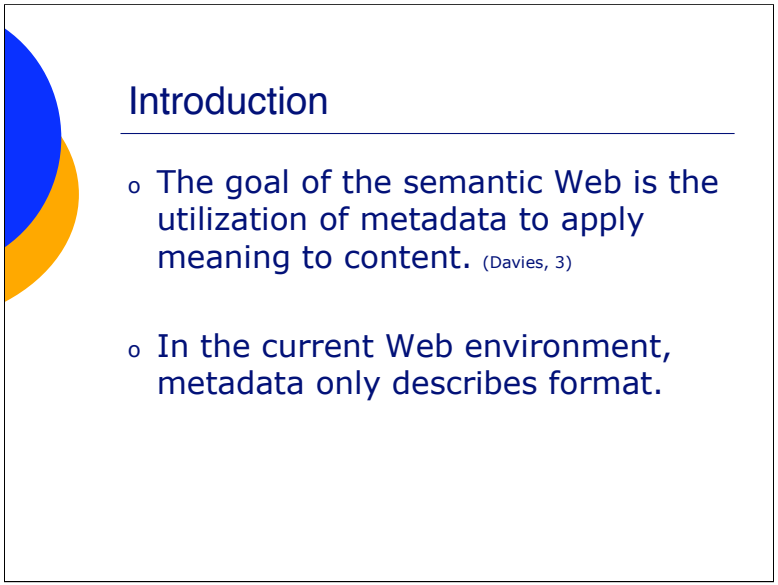
twine

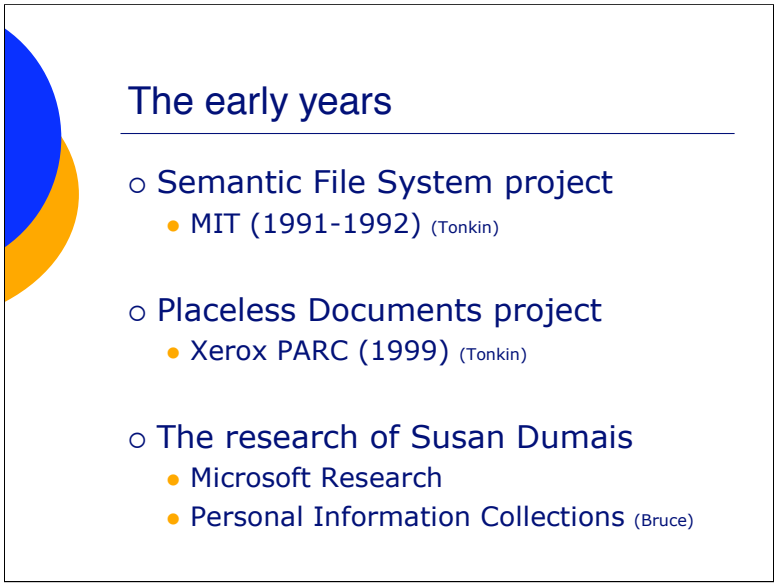
Accessible Semantic Tagging

Sarah Miller
Visiting Electronic Resources Librarian
Illinois Wesleyan University

Coalition for Networked Information
Spring Task Force Meeting
Minneapolis, Minnesota
April 8, 2008



I started out looking for a definition, but I think goal is a better term for it. From my perspective, computer scientists and companies are converging on the same end result using a variety of techniques.



In the process of my research, I found several papers that reflected on the origins of this area of research. By no means is the list exhaustive, nor do I claim to tell you who the person to dream up the Semantic Web is. I just want to contextualize that people have been working on this idea for years.

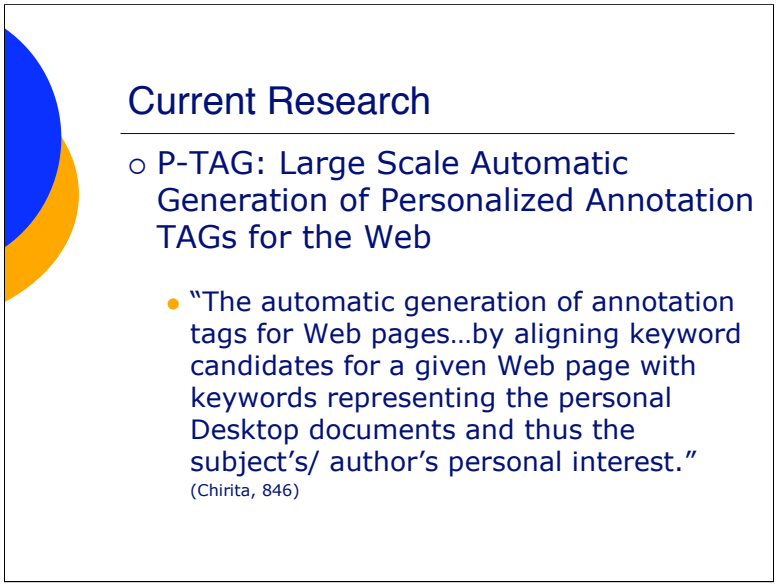
In her research of plain-text tagging, Emma Tonkin of UKOLN outlines the first two projects.

At MIT, the Semantic File System project worked to extract attributes from files and provide access via query based virtual directories.

The Placeless Documents project of Xerox PARC

"Document attributes capture the multiple different roles that a single document might play, and allow users to rapidly reorganise their document space for the task at hand." As Emma says, the intention of this project was to deal with the problem of "single inheritance" – the idea that a file can only reside in one place in a file system.

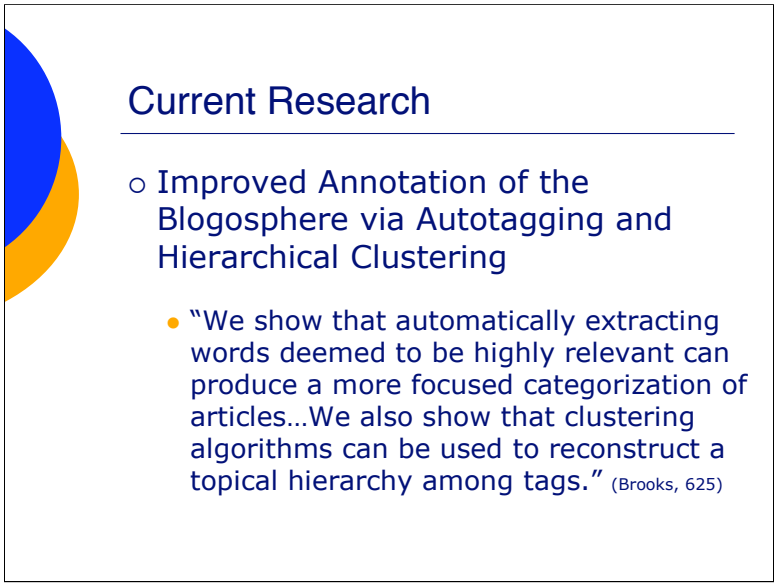
Susan Dumais- If I was going to drop everything and start over again, I want to study what she does. Originally came across her in the paper, "Keeping Found Things Found." She has studied user reactions to semantically categorized search results and what people do with information after they've located it. Written in conjunction with Hao Chen, the paper, "Bringing Order to the Web: Automatically Categorizing Search Results," highlights the use of metadata in search. This paper exemplifies to me the transition from the semantic organization of information on the desktop to that which users find on the web.



L3S Research Center at the University of Hannover and the National University of Ireland (IDA business park)

In this study, researchers indexed all emails, Web Cache documents, and files of the participating users. Simply put, the authors extracted a generic set of keywords from the Web page being annotated. To personalize this annotation, these keywords are then aligned with related terms from the personal desktop. At this point I feel that it is important to note that, according to the authors, the algorithms in this paper generate keywords that do not necessarily appear on the Web page, but are in its context.

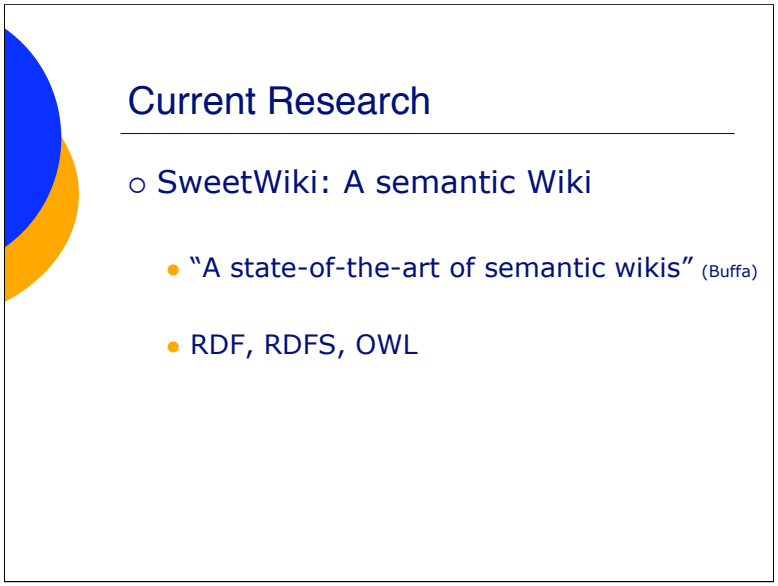
Results: User satisfaction was generally found to be above 70%. The authors would like to see their research applied to personalized Web searches, Web recommendations for desktop tasks, and ontology learning.



Computer Science Department, University of San Francisco. Brooks and Montanez.

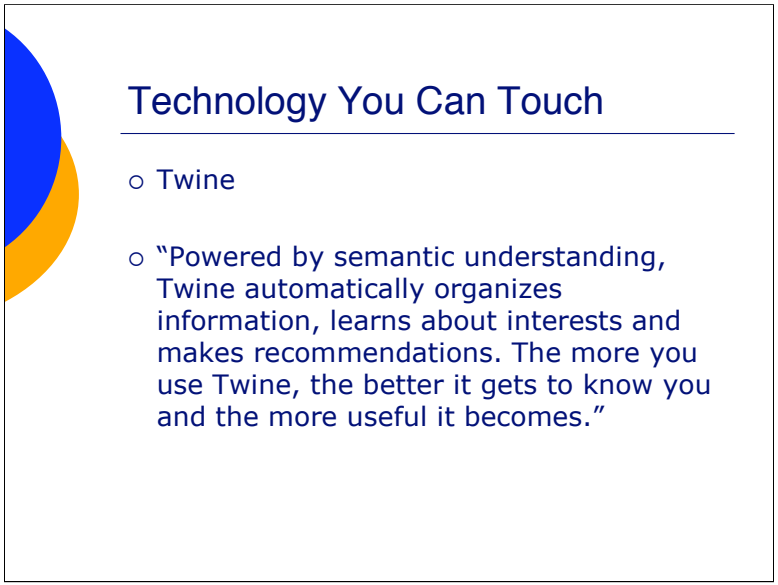
This papers focuses on analyzing blog entries indexed by Technorati. They explore the similarity of articles that share tags to determine whether articles that have similar tages actually contain similar content. They also used agglomerative clustering to record the order in which tag clusters are grouped into progressively more abstract clusters (every time I read the article I understand what they are saying a little better, but please recognize I'm not an expert on this hierarchical method).

Results: Keywords were extracted from articles, and then those that fell in the top 3 of the TFIDF score range were used to compare similarities. This resulted in significantly more similar articles sharing tags and smaller clusters indicating a more tightly focused topical grouping of articles.



A collaborative project between the KEWI Group at the University of Nice in France and the Edelweiss Group, also of France.

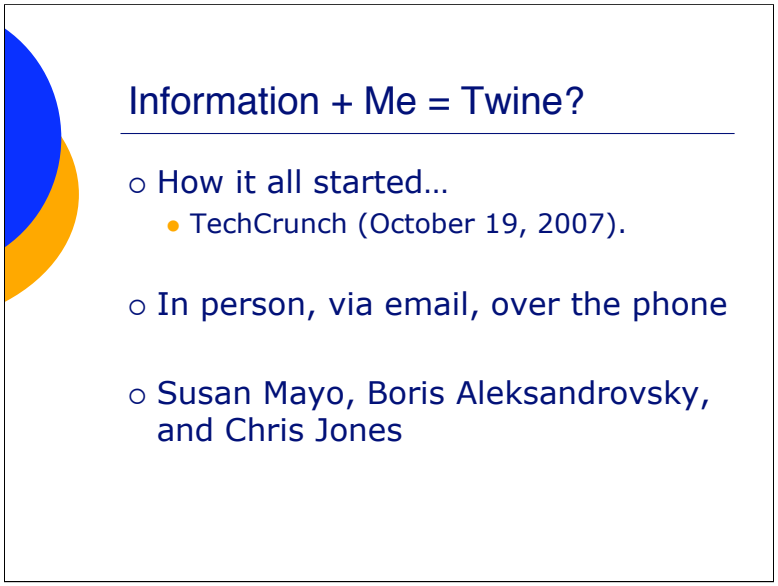
For the highly-technical audience member, I recommend this article. If you understand it, please get in touch, I'd like to have it explained to me. I included it for several reasons, one being that I think it's important to recognize that wikis are rich data sources and I'm glad that someone (not me) is trying to figure out how to weave them into the semantic Web. Two, it's full of acronyms. RDF, RDFS, and OWL are all acronyms I've heard come out of Nova Spivack's mouth, head of Radar Networks, creator of Twine. Can you see where I'm going with this? Also, it's in the Journal of Web Semantics. All semantics, all the time.



Doesn't that sound so simple? I thought in all fairness I would quote Twine's documentation in much the same way that I did the other projects.

Quick Facts

- Twine is a product of Radar Networks
- Founded by Nova Spivack
- Currently there are 40,000+ people on the waiting list



My husband and I wanted a place to store information we find on the web about teaching and learning. We dreamt of a product that does auto-tagging.

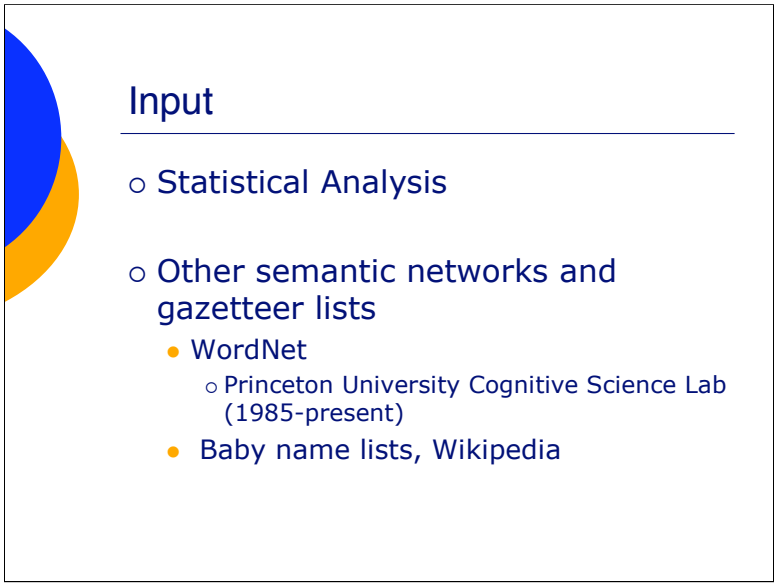
Low and behold, just a few days later I read about Twine on the blog TechCrunch.

I emailed Nova Spivack directly to get an account on the premise that I was looking for a new job (my current position is temporary) and wouldn't it be great to go into an interview and say I already had an account.

Visited Radar Networks in December while attending the American Geophysical Union Confernce. Great session for my husband and I to give them feedback. We also shared products like LibraryThing for Libraries and LibGuides with them.

In January, I was interviewed by Anne Eisenberg for the New York Times. Twine recommended me to her as one of their users.

Susan, Boris, and Chris have been my main contact points at Radar Networks.



This slide is the reason I started out with a goal to describe the semantic Web rather than a definition. There are still many different means to an end. Twine, like all of the other projects in this field, utilizes algorithms. This helps deal with CamelCase and Multi-word phrases.

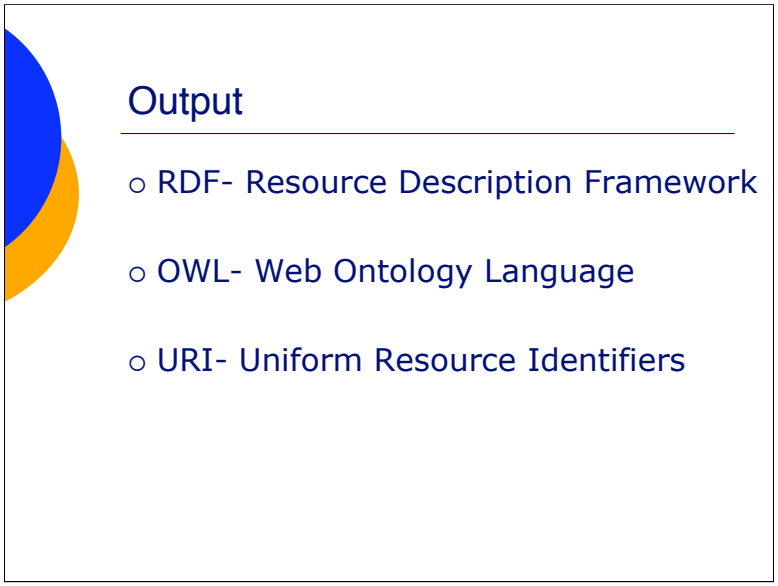
It also based on Princeton's handcrafted WordNet Project. This is taken from their website.

"Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be navigated with the browser."

Explanatory quote from Boris via email.

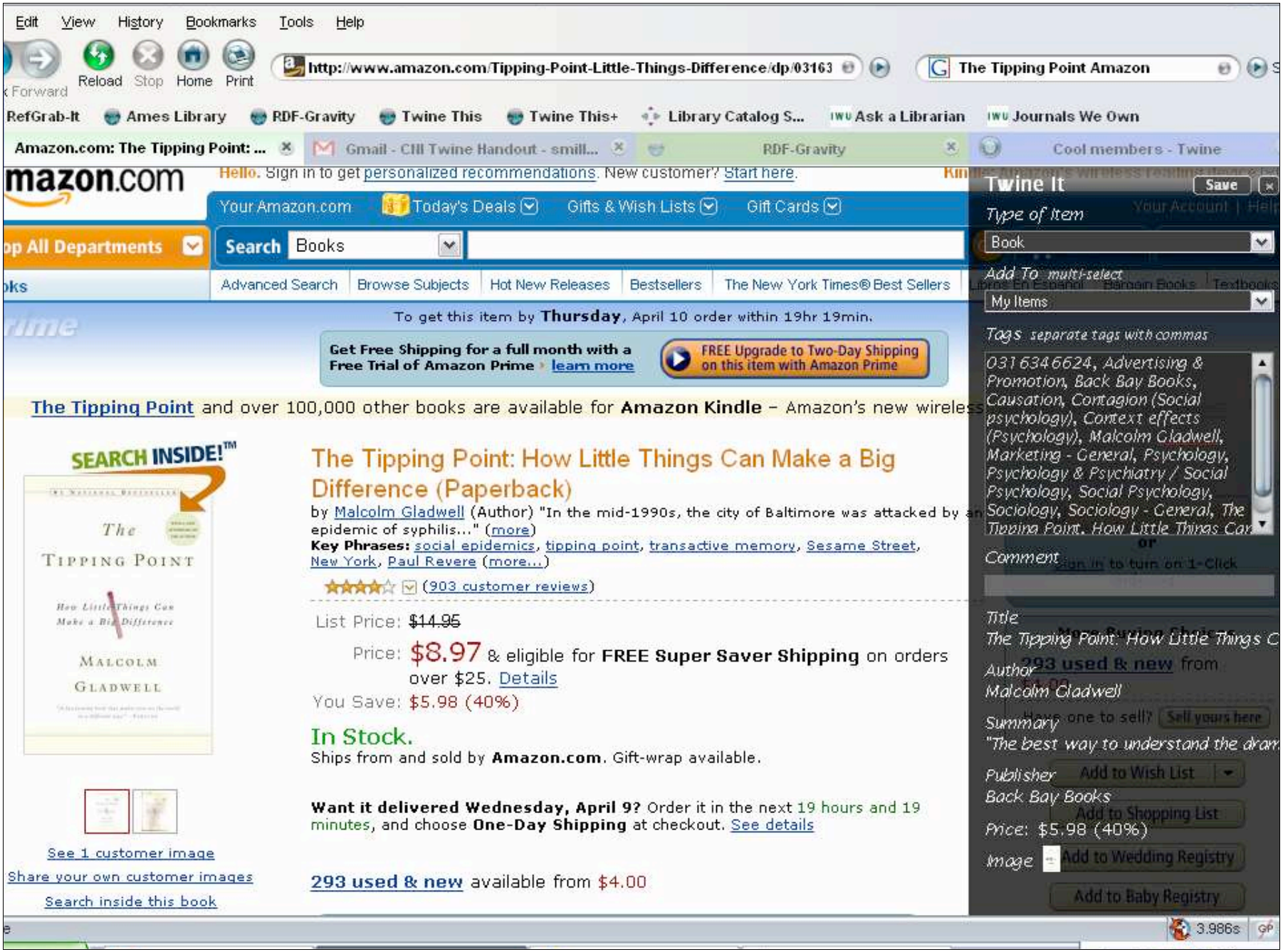
"Extractor based on the WordNet is the component which is looking for those concepts - people, organization, places and others. WordNet-like component descibes relationships between types (or classes) of things, not between particular instances of those classes. For instance it might know that "basketball player" is a kind of PERSON, but is does not know any particular basketball player. This is where the wikipedia-based gazeteers come into play - by uploading the list of basketball players into the net, it would immediately know that Michael Jordan is one. "

This helps to deal with situations of polysemy and other lexical semantic constructs. Boris acknowledges that this method can be somewhat brittle and sensitive to human error. In the future Twine will move further towards the statistical analysis of the words in the information input by its users.

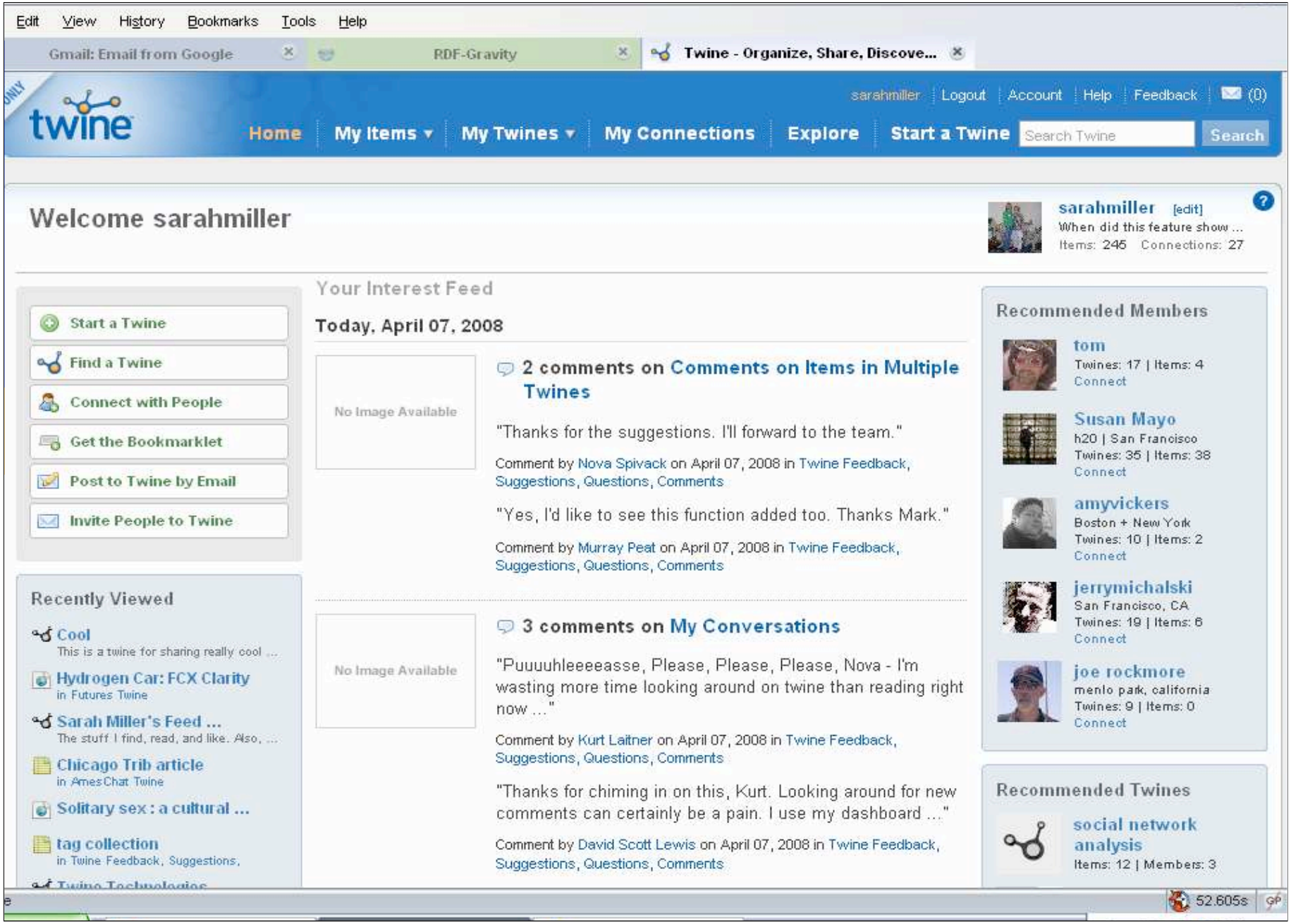


The information output of Twine is intended for situations where information needs to be processed by applications. There are two places I would encourage everyone to go for further information. The first is Twine's Website and the second is the W3C website. By utilizing these languages, the information in Twine is being formatted in a manner that will make it easy to share with other applications.

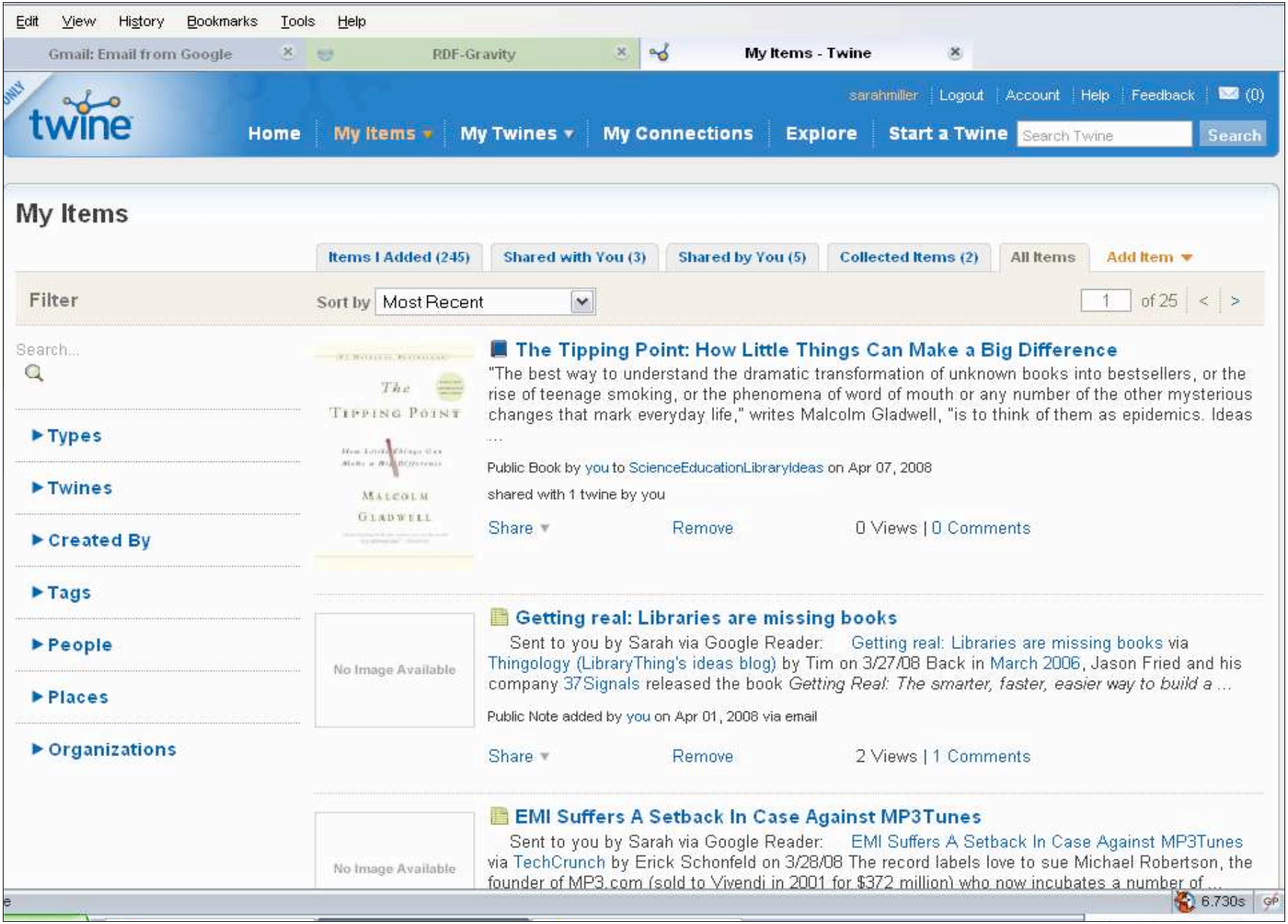
And it looks like this…



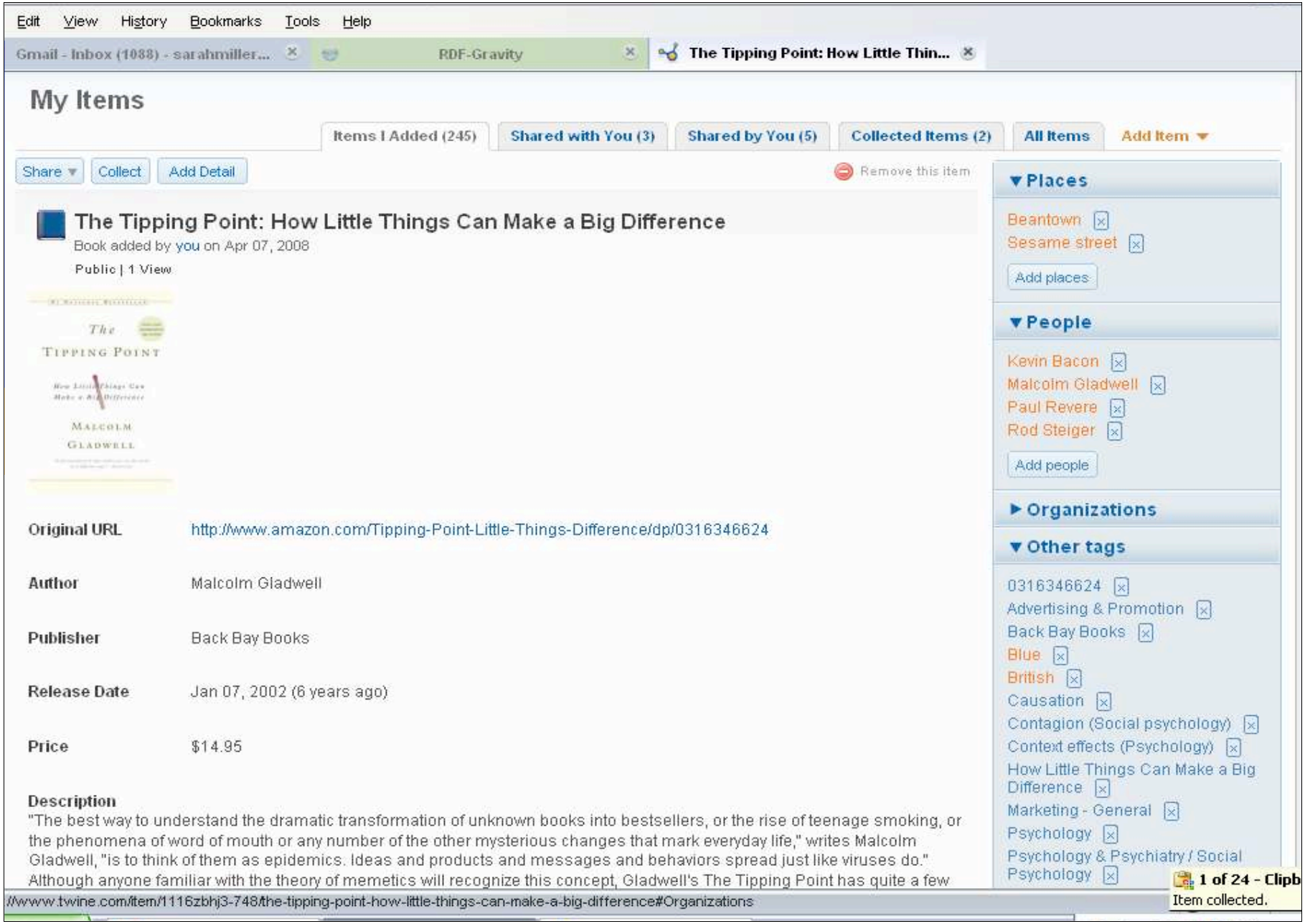
Adding an item to Twine from Amazon



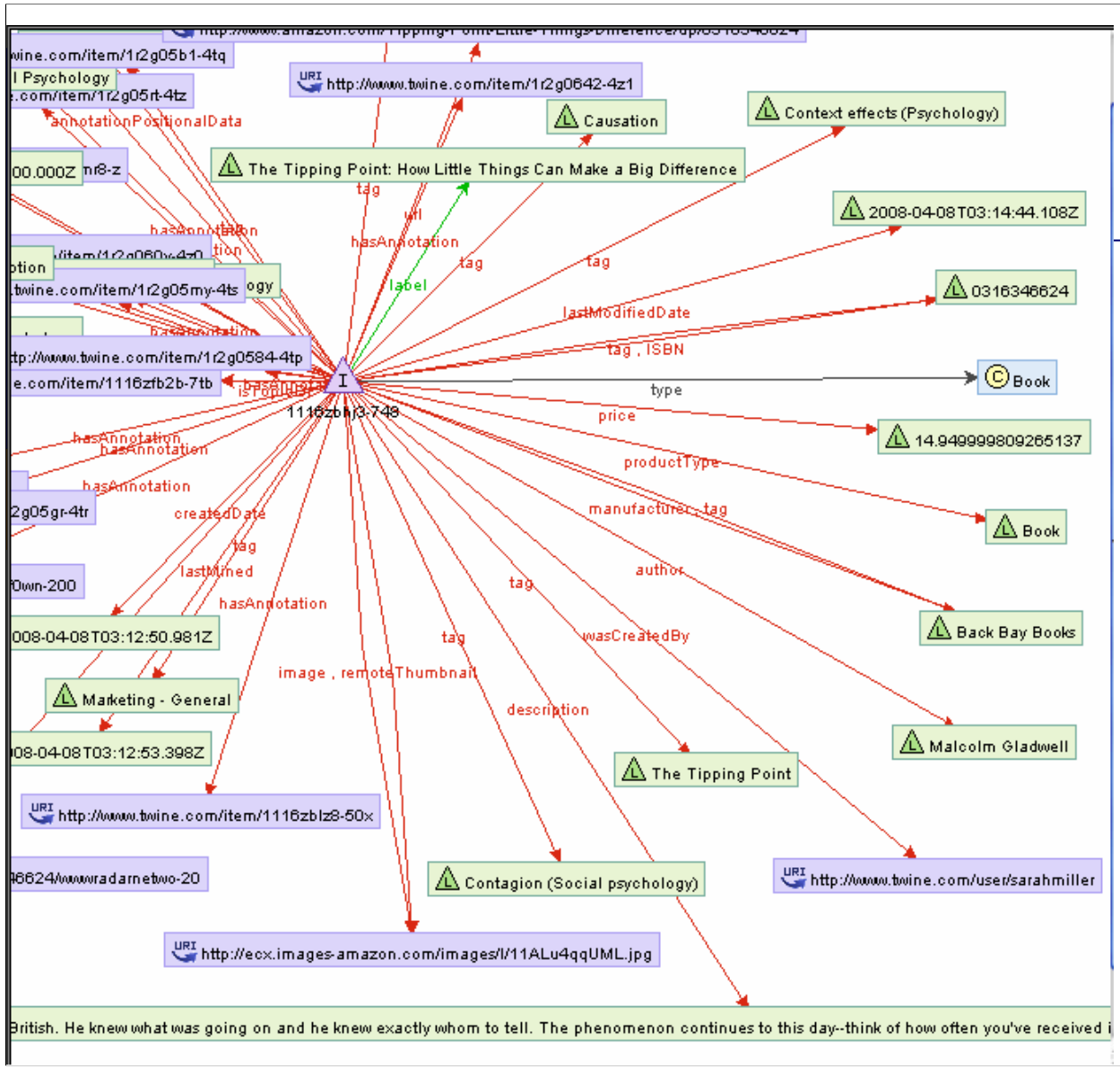
What you see when you log-in.



Items I've added to Twine



Be aware that the top of this screen is the same as the homepage.



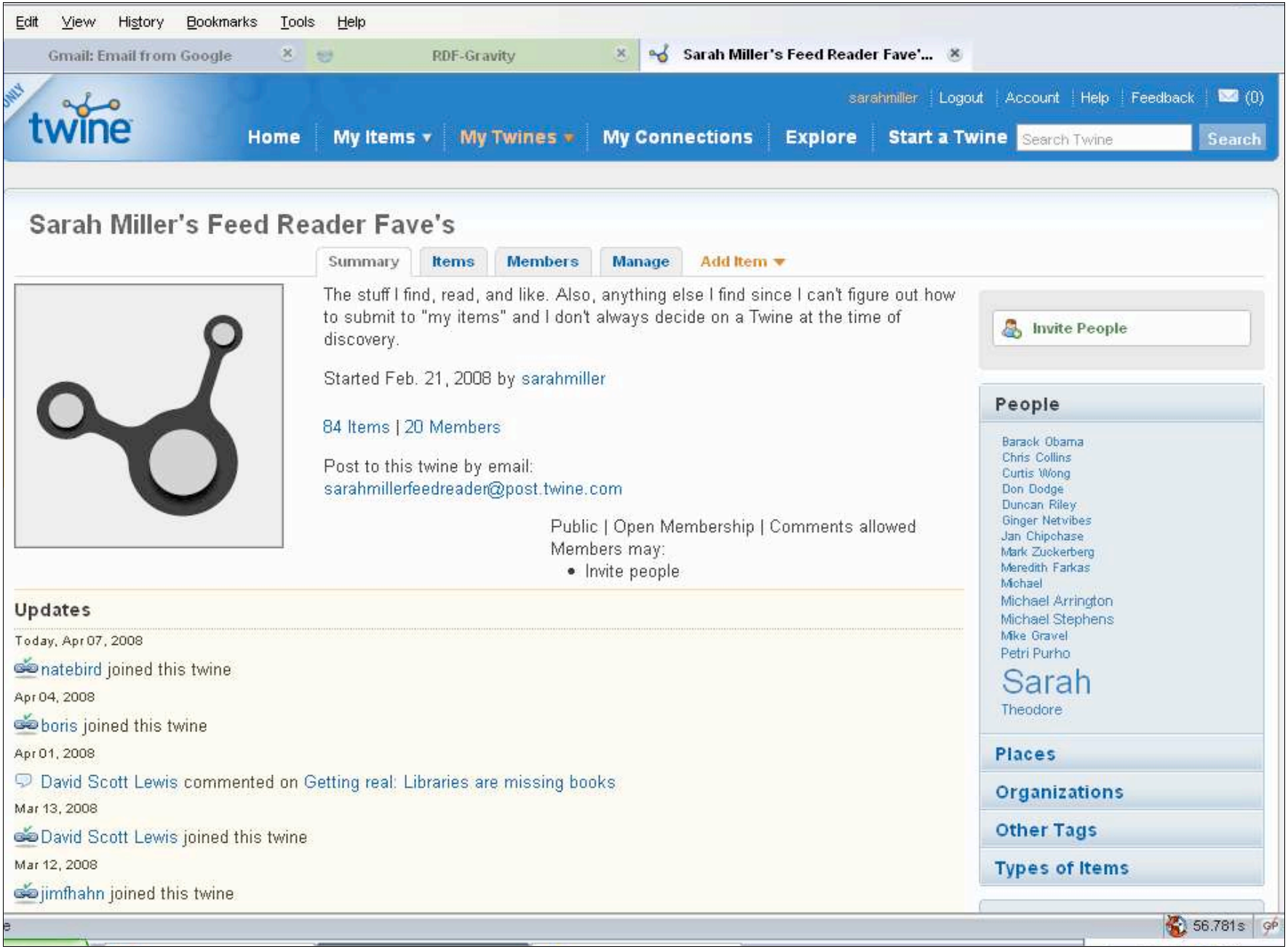
RDF Gravity: Graph Visualization Tool

Created by

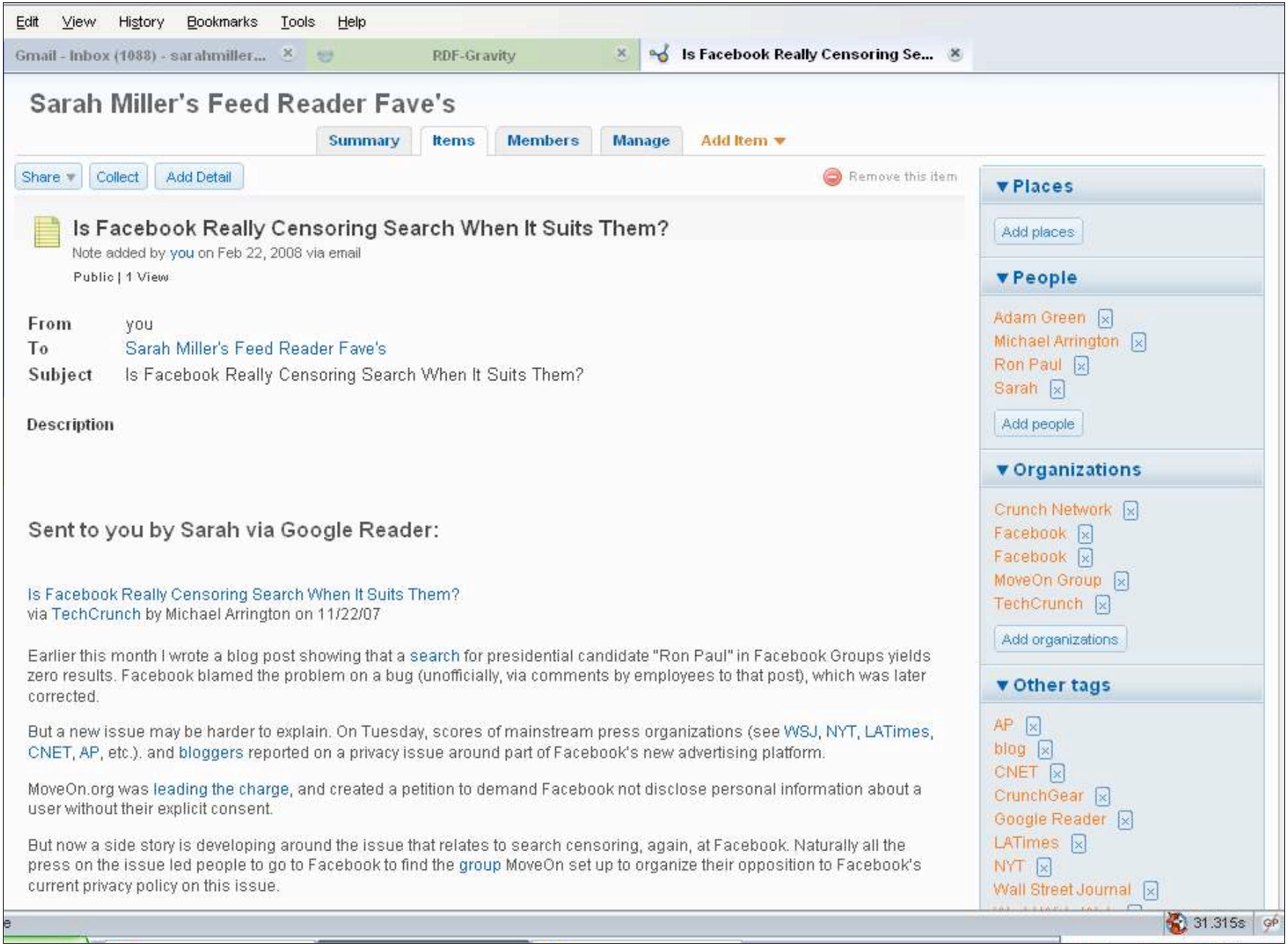
Sunil Goyal & Rupert Westenthaler

Subject-> Predicate -> Object

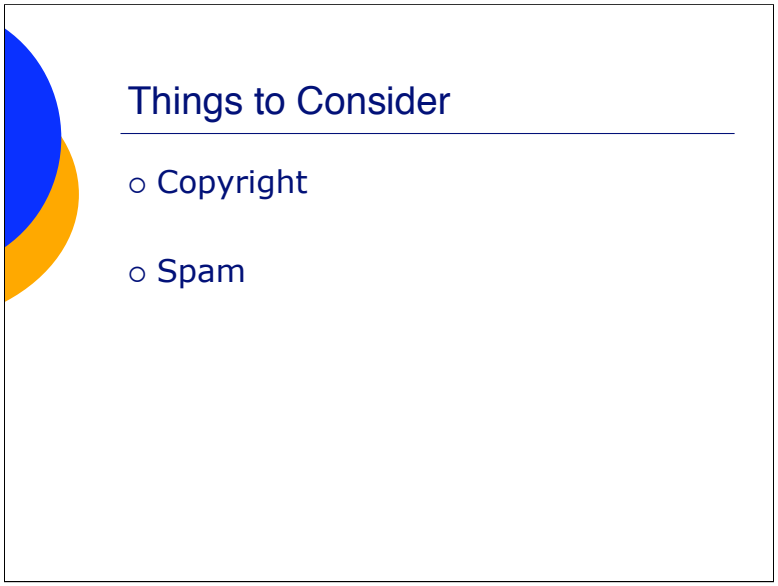
A visualization of *The Tipping Point* as one of my items in Twine.



I chose to email items from my feed reader into Twine to simulate what it might be like if my feed reader offered auto-tagging. Wouldn't that be an awesome way to track trends across feeds?

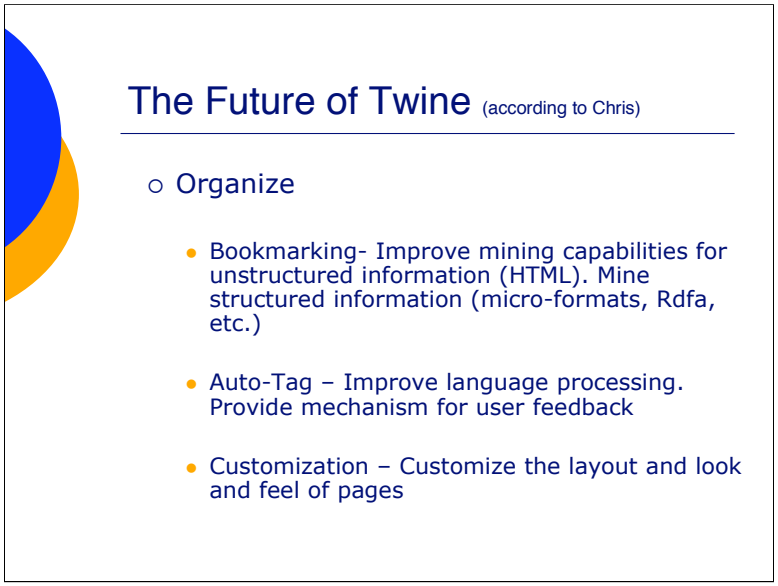


Notice all the orange auto-tags from Twine on this entry. Many aren't perfect, but it will get there.



Because users can upload content, they will be adding a question to “verify” that users own the content they are uploading. PDFs, photos, etc…

As for spam, they haven’t had a problem yet. They will be able to monitor it through individuals who will be “smoke testing” and by watching for data growth anomalies. In the future they plan to add a “report” this button. Although they have asked a user to remove less than desirable content. Not illegal, but considered the content to be offensive and in bad taste. Response to this could be to ask the user to keep the content private.



Response from Chris Jones via email on the future of Twine.

Response from Chris Jones via email on the future of Twine.

Sarah's Suggestions

- Diversify the interface
- Improved entity recognition
- Acknowledge that email is not where it's at

My humble opinion.

Bibliograpy

- Brooks, Christopher H., and Nancy Montanez. "Improved Annotation of the Blogosphere Via Autotagging and Hierarchical Clustering." WWW '06: Proceedings of the 15th International Conference on World Wide Web. Edinburgh, Scotland, <http://doi.acm.org/10.1145/1135777.1135869>.
- Bruce, Harry, William Jones, and Susan Dumais. "Information behaviour that keeps found things found." Information Research 10.1 (2004) <http://informationr.net/ir/10-1/paper207.html>.
- Buffa, Michel, et al. "SweetWiki: A Semantic Wiki." Journal of Web Semantics 6.1 (2008) <http://www.websemanticsjournal.org/papers/2007119/SweetWikiBuffaV6I1.pdf>.
- Chen, Hao, and Susan Dumais. "Bringing Order to the Web: Automatically Categorizing Search Results." CHI '00: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. The Hague, The Netherlands, <http://doi.acm.org.proxy2.library.uiuc.edu/10.1145/332040.332418>.
- Chirita, Paul Alexandru, et al. "P-TAG: Large Scale Automatic Generation of Personalized Annotation Tags for the Web." WWW '07: Proceedings of the 16th International Conference on World Wide Web. Banff, Alberta, Canada, <http://doi.acm.org/10.1145/1242572.1242686>.
- Davies, John, Rudi Studer, and Paul Warren. Semantic Web Technologies: trends and research in ontology-based systems. John Wiley & Sons, Ltd., 2006.
- Dourish, Paul, et al. "Presto: An Experimental Architecture for Fluid Interactive Document Spaces." ACM Trans.Comput.-Hum.Interact. 6.2 (1999). <http://doi.acm.org.proxy2.library.uiuc.edu/10.1145/319091.319099>.
- Falconer, S. M., D. Maslov and M. -A Storey. "Combining Web-Based Searching with Latent Semantic Analysis to Discover Similarity between Phrases." ODBASE. Montpellier, France, Oct. 31-Nov. 2, 2006.
- Gifford, David K., et al. "Semantic File Systems." SOSP '91: Proceedings of the Thirteenth ACM Symposium on Operating Systems Principles. Pacific Grove, California, United States, <http://doi.acm.org.proxy2.library.uiuc.edu/10.1145/121132.121138>.
- Lee, Sigma On Kee, and Andy Hon Wai Chun. "Automatic Tag Recommendation for the Web 2.0 Blogosphere using Collaborative Tagging and Hybrid ANN Semantic Structures." ACOS'07: Proceedings of the 6th Conference on WSEAS International Conference on Applied Computer Science. Hangzhou, China.
- Marieke, Guy, and Emma Tonkin. "Folksonomies: Tidying up Tags?" D-Lib Magazine. 12.1 (2006). 3/12/2008 <http://www.dlib.org/dlib/january06/guy/01guy.html>.
- Mikroyannidis, Alexander. "Toward a Social Semantic Web." Computer 40.11 (2007).
- "OWL Web Ontology Language Overview." 4/6/2008 <http://www.w3.org/TR/owl-features/>.
- "RDF-Gravity." 4/7/2008 <http://semweb.salzburgresearch.at/apps/rdf-gravity/index.html>.
- "RDF Primer." 4/6/2008 <http://www.w3.org/TR/rdf-primer/>.
- Sinclair, James, and Michael Cardew-Hall. "The Folksonomy Tag Cloud: When is it Useful?" Journal of Information Science 34.1 (2007).
- "Susan Dumais Homepage." 4/7/2008 <http://research.microsoft.com/~sdumais/>.
- Tonkin, Emma. "Folksonomies: The Fall and Rise of Plain-text Tagging." Ariadne. Issue 47. 3/12/2008 <http://www.ariadne.ac.uk/issue47/tonkin/>.
- "Twine." 4/7/2008 <http://www.twine.com>.
- "WordNet - Princeton University Cognitive Science Laboratory." 4/7/2008 < http://wordnet.princeton.edu/>.

Further Reading on Social Bookmarking and Tagging

- Al-Khalifa, Hend S., and Hugh C. Davis. "Towards Better Understanding of Folksonomic Patterns." Proceedings of the 18th Conference on Hypertext and Hypermedia. Manchester, UK. 3/13/2008.
- "A cognitive analysis of tagging." 3/12/2008 <http://www.rashmisinha.com/archives/05_09/tagging-cognitive.html>.
- Gordon-Murnane, Laura. "Social Bookmarking, Folksonomies, and Web 2.0 Tools." Searcher 14.6 (2006).
- Halpin, Harry, Valentin Robu and Hana Shepherd. "The Complex Dynamics of Collaborative Tagging." Proceedings of the 16th International Conference on World Wide Web. Banff, Alberta, Canada.
- "Folksonomies and Social Tagging Presentation Video." Infotangle(2007) 3/12/2008 <http://infotangle.blogsome.com/2007/07/04/folksonomies-and-social-tagging-presentation-video/>.
- Golder, Scott, and Bernardo A. Huberman "Usage Patterns of Collaborative Tagging Systems." Journal of Information Science 32.2 (2006). 3/12/2008.
- Hammond, Tony, Timo Hannay, Ben Lund, and Joanna Scott. "Social Bookmarking Tools (I): a General Review." D-Lib Magazine 11.4 (2005). 3/12/2008.
- "The Hive Mind: Folksonomies and User-Based Tagging." Infotangle. (2005) 3/12/2008 <http://infotangle.blogsome.com/2005/12/07/the-hive-mind-folksonomies-and-user-based-tagging/>.
- Jaschke, Robert, et al. "Discovering Shared Conceptualizations in Folksonomies." Journal of Web Semantics 6.1 (2008) <http://www.websemanticsjournal.org/papers/2007119/DiscoveringSharedConceptualizationsStummeV6I1.pdf>.
- Marlow, Cameron, et al. "HT06, Tagging Paper, Taxonomy, Flickr, Academic Article, to Read." HYPERTEXT '06: Proceedings of the Seventeenth Conference on Hypertext and Hypermedia. Odense, Denmark, <http://doi.acm.org/10.1145/1149941.1149949>.