What next for Semantic Blogging?♦

Steve Cayzer
Digital Media Systems Laboratory
HP Laboratories Bristol
HPL-2006-149
October 31, 2006*

Semantic Blogging is the use of rich metadata to transform blogs from simple online diaries to full participants in an information sharing ecosystem. Originally the semantic blogging vision centred around informal decentralized knowledge management, but recent developments in social network analysis, microformats, semantic desktop applications and wikis have challenged, enriched and extended this vision. In this paper I review the history of semantic blogging, present a snapshot of where we are now, including two of my current experiments. BlogAccord is an exploration of music blogging, while the Snippet Manager is an information integration portal. I conclude this paper by offering a personal take on some promising future directions.

What next for Semantic Blogging?

Steve Cayzer[1]

Abstract
Semantic Blogging is the use of rich metadata to transform blogs from simple online diaries to full participants in an information sharing ecosystem. Originally the semantic blogging vision centred around informal decentralized knowledge management, but recent developments in social network analysis, microformats, semantic desktop applications and wikis have challenged, enriched and extended this vision. In this paper I review the history of semantic blogging, present a snapshot of where we are now, including two of my current experiments. BlogAccord is an exploration of music blogging, while the Snippet Manager is an information integration portal. I conclude this paper by offering a personal take on some promising future directions.

1. Introduction

Semantic blogging has attracted quite a lot of interest since its introduction in 2003 [1]. Indeed, several groups from around the world (UK, USA and Japan) have built systems with precisely this theme. The purpose of this paper is to review this work and to offer some reflections on promising new directions. Along the way I shall refer to work which, although not directly involved with semantic blogging, is I think related enough to influence the field. First and foremost is structured blogging, a lightweight metadata approach that is very similar in spirit to semantic blogging. At a hardly greater conceptual distance are wikis, which, together with email and IRC, have an important role to play in the information ecosystem in which semantic blogs participate. These related developments have guided my two current projects in this space. The snippet manager focuses on the whole information ecosystem, in which a semantic blog is an important, but not central, component. BlogAccord is a music blog inspired by structured blogging. In the final section I reflect upon some future trends: information integration, concept mapping, social networking and folksonomy.

2. Where did semantic blogging come from?

Semantic Blogging, at least as I use the term, was introduced at the first BlogTalk conference [1]. The idea, a natural fusion of the semantic web technologies and weblogs, comprises three central modalities. First, semantic view, in which metadata is used to adjust the presentation of the blog in a semantically appropriate manner. Secondly, semantic navigation, in which metadata is used to drive browsing behaviour (for example 'more like this'). Finally, semantic query refers to an enriched search where, for example, the user can search for blog entries on papers written by X. These modalities are, I believe, still helpful, but the idea has been beneficially enriched and extended in

---

[1] Hewlett-Packard Laboratories, Filton Road, Stoke Gifford, Bristol BS34 8QZ UK

the intervening years as we shall see later. Semantic blogging formed a central component of the SWAD-E project and a demonstrator was produced around bibliography management [2]. The software has subsequently been used to power my own blog [3] and is available for download.

As I had hoped [4], a number of groups have subsequently picked up on the theme. The semblog platform [5] allows users to annotate content using their personal ontologies, syndicating this metadata over RSS. Both topic and social network [6] in-formation are thus available for information retrieval and recommendation. MIT's Haystack platform [7] has also been applied to metadata authoring. In the UK, the University of Southampton are enamoured with blogs as a way of publishing information from their mSpace platform [8]. A rather more ambitious approach is taken by the KMI group at the UK's Open University, who envision a system to build up argumentation networks [9]. Linking this to blogging [10] is largely a matter of making the annotation easier [11]. Most recently, Knud Möller from DERI at the University of Galway has proposed a semantic blogging approach called semiBlog [12], which exhibits tight integration with various desktop applications.

Thus we can see that there are a number of concrete implementations of the semantic blogging vision. However, that is not all there is to the story as we shall see in the next section.

3. Where is it right now?

Blogs syndicate their latest posts in RSS[2], a common XML format which allows software readers to automatically poll, process and present posts from any blog. RSS comes in a variety of dialects including the RDF based RSS1.0 [13]. In the original formulation [1], semantic blogs express their metadata in the semantic web format RDF [36], and hence use RSS1.0. However, this approach brings with it two problems. Firstly, not all blogs publish RSS1.0, and many aggregators favour a different syndication format (the newest option is Atom [14] although Microsoft's adoption of RSS2.0 [15] for 'web feeds' may give that format something of a renaissance). Another problem is that adding arbitrary RDF to an RSS feed at best produces feed bloat, and at worst chokes RSS parsers that are not fully RDF aware (this issue is discussed in [2]). In such cases it may be better to embed a link to the RDF in the RSS/Atom (and HTML), allowing RDF aware consumers to download the bulk of the metadata separately.

Notwithstanding this discussion, it is in fact possible to achieve much of the semantic blogging vision without using RDF. Microformats [16] are methods for embedding simple structured information in an HTML page. It is not difficult to see how this could be useful for blogs; in fact, it even has a name, *structured blogging* [17]. The structured blogging approach is a good way to achieve a context sensitive presentations (for example, blogs about music tracks are presented differently to blog entries about restaurants); this is a concrete instantiation of what I call *semantic view* [1]. In addition, structured blogging is complementary to semantic blogging, since the XML metadata contained in structured blog entries can easily be translated into RDF [18].

How can structured blogging inform semantic blogging? I recently supervised a group of Sydney University students who were tasked with applying semantic blogging to a new domain. Since the original demonstrator [2] targeted the bibliography domain, these students chose music blogging as

---

2 http://www.xml.com/pub/a/2002/12/18/dive-into-xml.html accessed 13 October 2006

a motivating example. BlogAccord [19] allows a user to blog an artist, album or track, the blog entry being automatically enriched with MusicBrainz [20] metadata. When viewing the blog, the metadata is used to enrich the display – for example with rating information being presented as stars (*fig 1*).
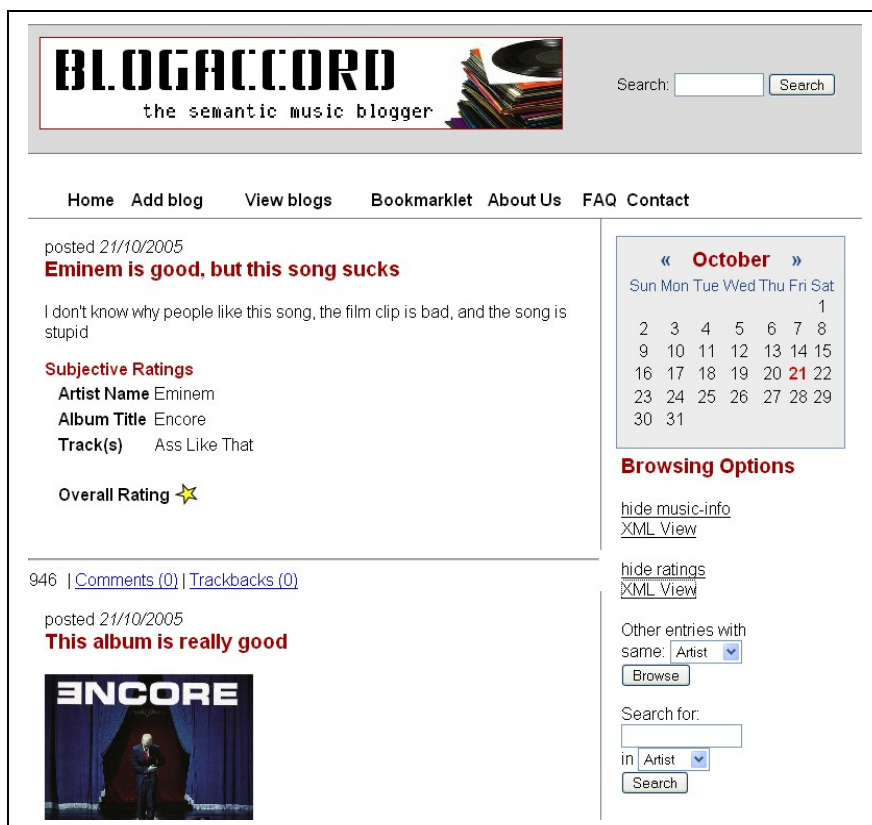

Figure 1. BlogAccord prototype screenshot. Each blog entry is annotated with music related information (album title, track) and rating (stars) which can be toggled on or off as the user desires. The metadata (except the rating) was autopopulated from MusicBrainz.

While the existing application is merely a partly functional prototype, I believe it does demonstrate a promising future direction for semantic blogging. Although on the face of it looks little different to structured blogging, the metadata is both richer and more explicitly defined (*fig 2*).

The core concept in the data model is that a blog entry is *about* a data item. Both these entities are represented as nodes in a directed graph. Metadata about the music item (artist, item, track) are attached to the music item node, whereas metadata about the blog entry (author, title, description) are attached to the blog entry node. Each unique entity (e.g. musician, album) is represented by a unique tag, called a URI³. This allows the music blog to go beyond cosmetic formatting, for example finding 'other entries about this artist'. Extending this search across blogs requires a shared vocabulary (one recent example for classical music is described in [37]). Such vocabularies allow the blog to be linked with third party metadata about the music (such as artist biography, chart position of album etc) and allow the inference of relationships (tracks with similar rhythmic or harmonic quality; bands with some overlap of membership, like The Libertines and Babyshambles). Further, one need not be limited to the music domain. One of the strengths of RDF is data integration which allows us to link into geographical data ('Upcoming local gigs from bands

---

³ Uniform Resource Identifier http://www.w3.org/Addressing/Activity accessed 13 October 2006

mentioned on your blog) and social networks ('bands your friends like'). This, incidentally, takes semantic blogging beyond the initial vision of 'semantic view, semantic navigation, semantic query' [1] and is a theme I return to later.
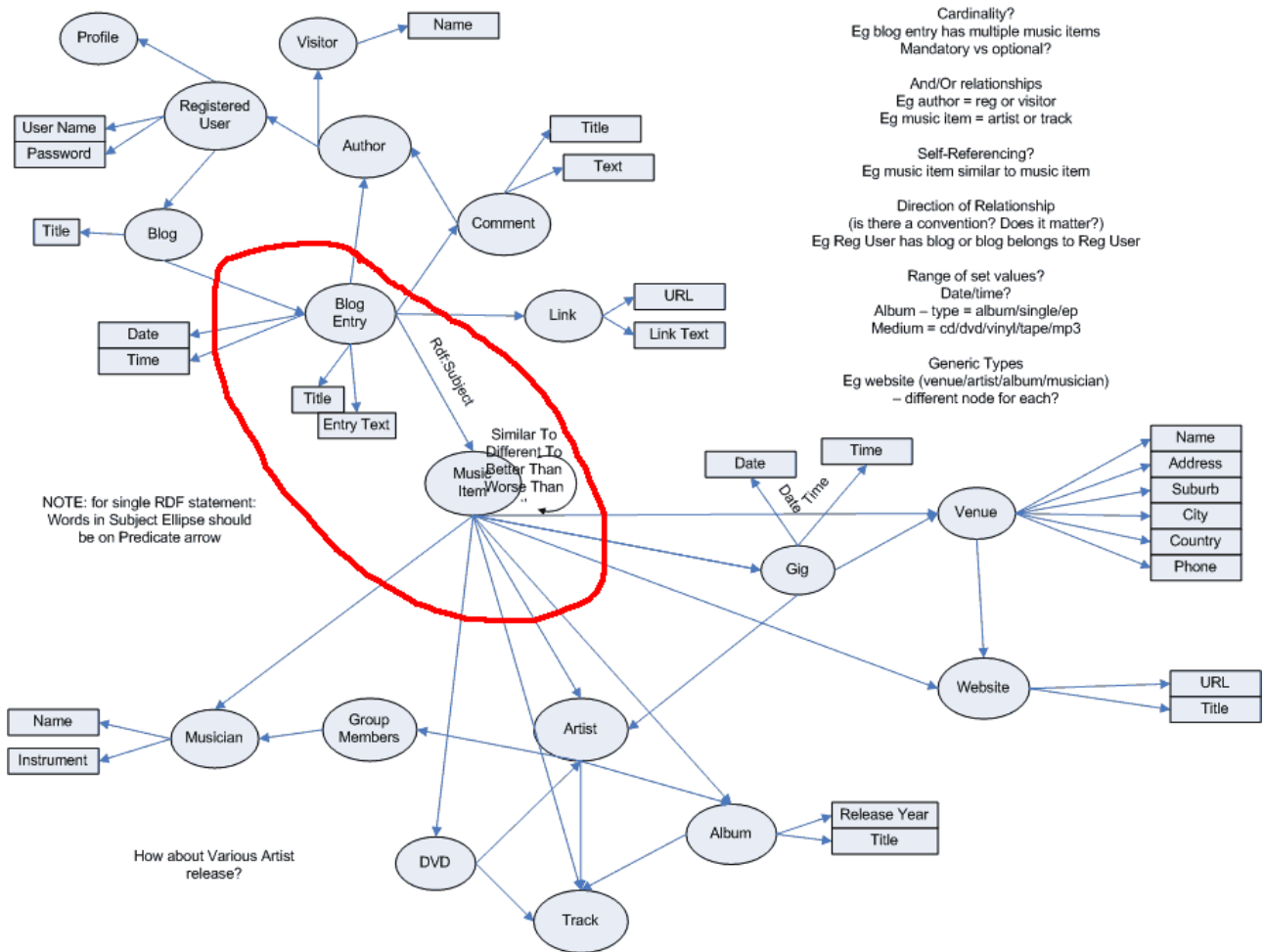

Figure 2. Data model underlying our BlogAccord prototype. The central entities (blog entry and music item) are circled. Note that, according to the semantic blogging idea, a blog entry is an annotation over an underlying music item, which is linked to a rich underlying information space via Artist, Album, Venue and other entities.

Wikis, like blogs, are lightweight data capture tools which allow collaboration and discourse around a topic. The difference is mainly one of tone: blogs tend to have a single author, with community commentary; whereas wikis are designed to be communally authored (the dividing line is not crisp and one can combine the two[4]). Recently, much has been made of semantic wikis [21], where collaborative editing of both instance data and ontologies is encouraged. A little further removed, though still strongly related, is email, widely used as a tool to disseminate 'snippets' of information. Annotating such snippets with rich metadata enables what has been termed semantic email [22].

Semantic blogs are not the same thing as blogs about the semantic web. Nevertheless, PlanetRDF [23] is worth mentioning because it is an aggregator blog, generated from RDF metadata that describes the source blogs. Aggregating semantic information touches on a wider theme that has been developing in the semantic web community. Piggybank [24] is a way of collecting information

---

[4] Blog + wiki = *bliki*: http://en.wikipedia.org/wiki/Bliki accessed 13 October 2006

from web pages in a lightweight manner, using RDF if available, or failing that using screenscraping techniques. The data collected can be stored in a semantic bank, and users can browse this information using a faceted browser. A closely related tool is Trailblazer [25], which uses simple machine learning to detect both the likely topic of a page and likely context-dependent entities in that page, thus enabling users to enter relevant annotations which can be linked and shared post-hoc.

Our snippet manager explorations [26] stem from a similar motivation. Specifically, we are interested in capturing, annotating and sharing information snippets. However, in this work we have brought in a new and important dimension, that of integration. The semantic blogging vision has become [27] information centric (rather than blog- or indeed tool-centric). In the snippet manager we have made this desire explicit, in that blogs form just one of the inputs (albeit a particularly convenient one). Email, wikis, bookmarklets, web resources and intranet databases all provide a steady stream of tagged information which it is the Snippet Manager's task to collate and present to its users (*figures 3, 4*).

This move from semantic blogging to semantic information management is one of the many interesting trends that I believe are keeping the semantic blogging meme alive and well. I examine some of these trends in the next section.
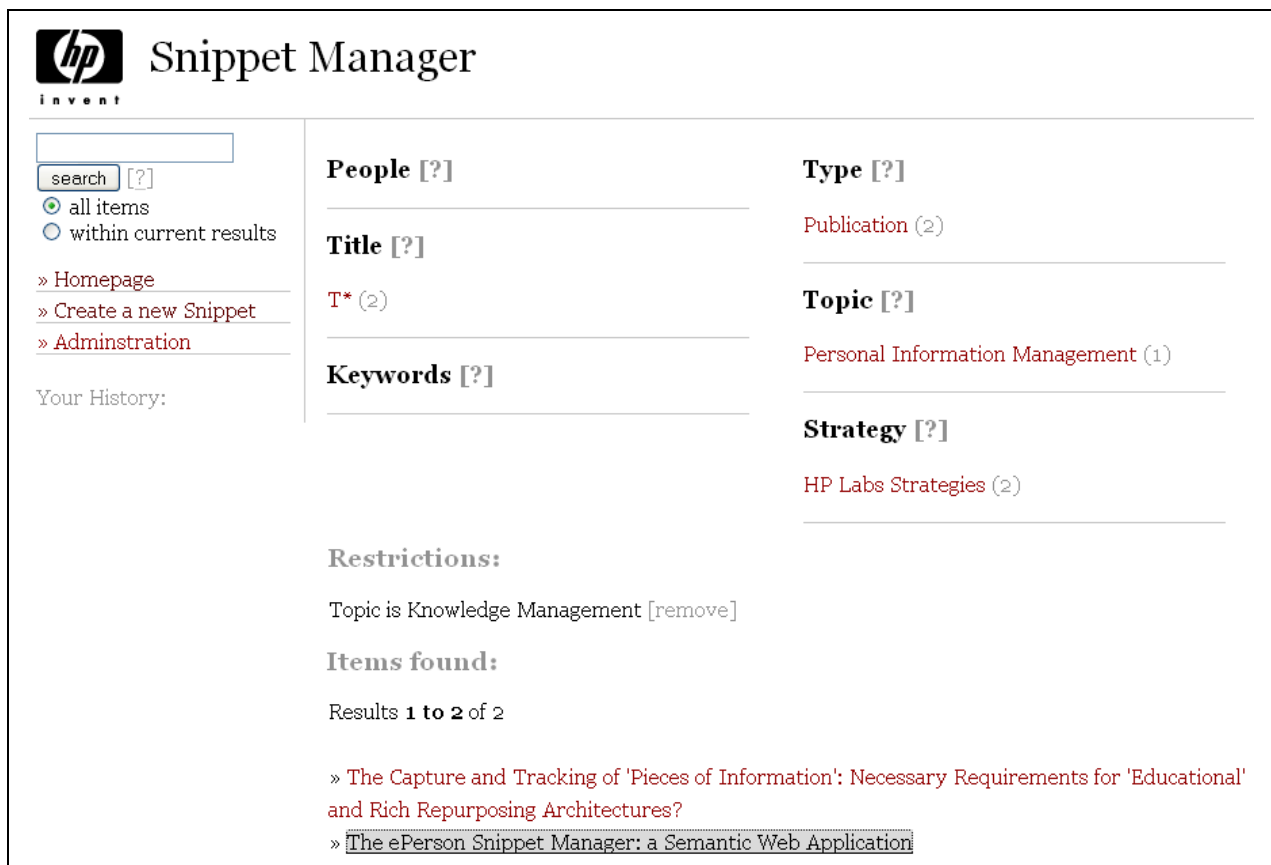

Figure 3: Sample screenshots of an early snippet manager prototype. One can browse snippets by type (publication, email, person), topic, keyword, strategy and so on.
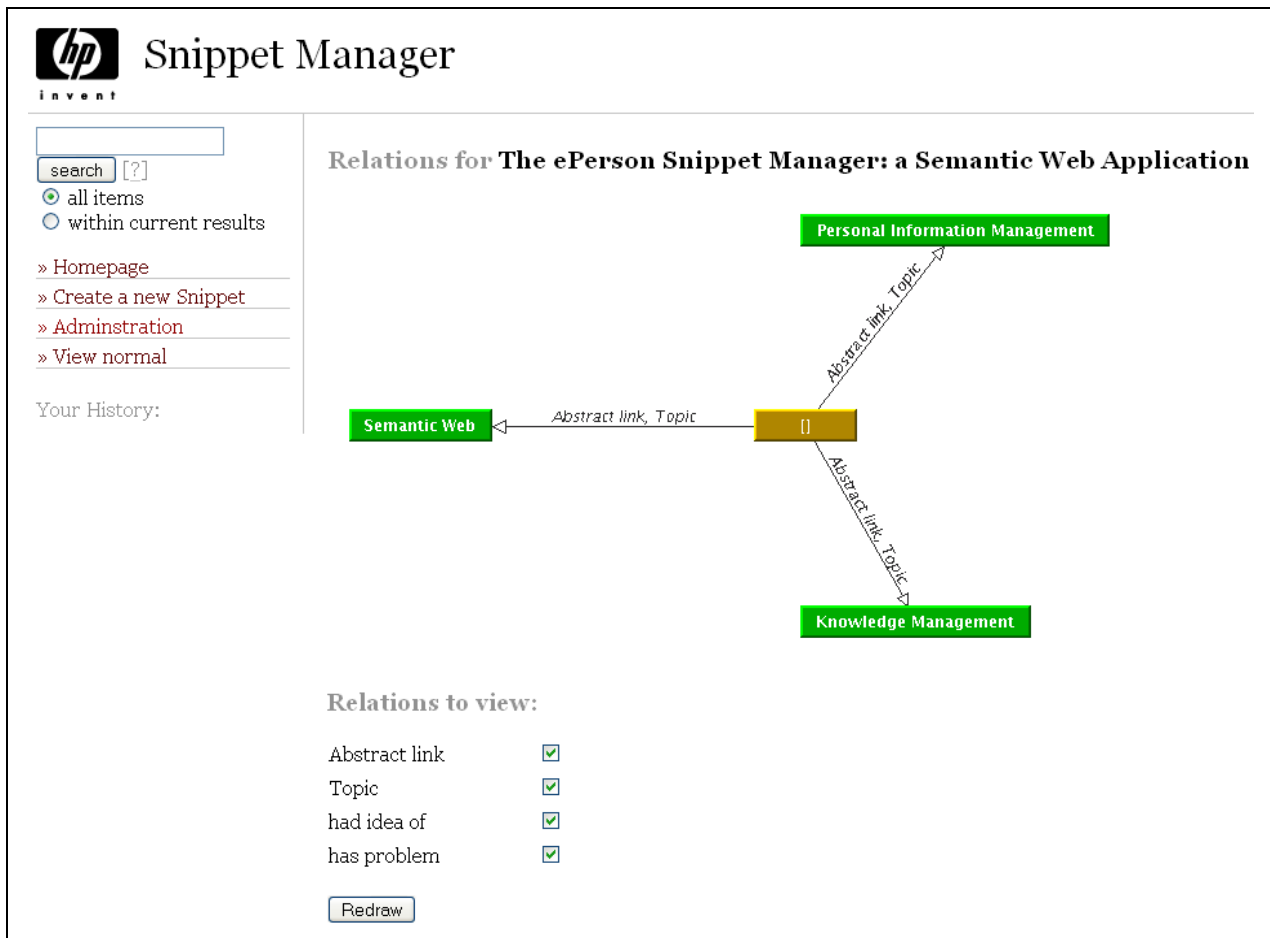
Figure 4: Sample screenshots of an early snippet manager prototype. Faceted browsing, folksonomy cloud views and graph type navigation are all supported modalities.

4 Where are we headed

Having surveyed the past and present in what I hope is a reasonably objective manner I now present a necessarily subjective view of what the future holds for semantic blogging. Although blogs are often seen to be a Web2.0[5] technology, I will avoid that term, since it does not appear to be universally understood or accepted[6]. Rather, I will take three concrete examples of semantic blogging technology and talk about their technical and commercial potential.

4.1 Semantic Blogs and Enterprise Information Management

Gartner define Enterprise Information Management (EIM) as "*A technical strategy to organize, design, catalog and safeguard structured and unstructured content to maximize its value, usefulness, accessibility and security*" [39]. Blogs clearly play a role in generating such content; nearly 70 percent of all large (>$50M) companies will have deployed corporate blogs by the end of 2006 [40]. Metadata associated with blogs are seen to be a significant benefit for enterprise search functionality. However, a major limitation of current technologies is that "*information created in one tool [is] often not visible in another ... information becomes fragmented*" [41]. The Snippet

---
[5] http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html
[6] http://www.theregister.co.uk/2005/10/21/web_two_point_nought_poll/

Manager [26] steps up to this challenge directly, merging disparate information sources in an application independent manner. SIOC⁷ [42] is a recently proposed semantic web ontology for representing (and combining) topics and conversations from blogs, forums and newsgroups. Thus, semantic blogs and associated technologies provide a channel for solving semantic inconsistencies, an ability which Gartner believes to be crucial [39] in the burgeoning EIM market.

4.2 Semantic Blogs and Concept Mapping

In my discussions of semantic blogging, I have always been careful to draw a distinction between blog entries and information items (such as papers). In effect, the entries are merely annotations over the items which are first class entities in the information space. Concept mapping takes this one stage further. Here, blog entries are associated with *ideas about an information item*. For example, one might blog about a claim made in a paper [11] and link these claims together post-hoc. Such an approach is currently being applied in the neuroscience community, for the benefit of researchers studying Alzheimer's disease [28].

4.3 Semantic Blogs, Social Networks and Folksonomy

It seems natural to augment blogs with social networks. After all, blogrolls have performed this function almost since the inception of blogging. In the RDF world, Friend of a Friend, or foaf [6] is a simple format to encode such social networks. Indeed, social network analysis techniques have been applied to the foaf world [43]. However, I believe that caution is required. Relying on social connections to provide recommendations [29], for example, risks conflating social relationships with correlated interests. A more principled approach is taken by SIMAC [30] in which foaf connections are used to find related artists ("here are some artists related to your favourite bands"). The connection with blogging, in particular our BlogAccord prototype, is obvious.

Another link is that with computational analysis. Blogs are a rich source of data to be mined for insight into social networks [31], two recent examples being information epidemics [32] and collective mood [33]. What has this to do with semantic blogging? I think there are a number of promising directions.

Firstly, metadata can be used as input to such analyses. As Downes [38] argues, metadata can be used to define the role of a social network actor (author, user, commentator). Alternatively, metadata about an information item, both internal (eg topic hierarchy) and external (eg thesauri), can be used to refine blog categorization, crosslinking and recommendation [34].

Secondly, the output of computational analysis can usefully be transported and communicated in the blog metadata. To my knowledge, no-one has yet included *inferred* knowledge such as 'authors related to me' or 'resources relevant to this blog entry' to their blog. Yet such a thing is easily possible, and can even be syndicated using an extensible format like RDF or Atom.

---

⁷ Pronounced 'shock' (Gaelic word)

Many people tag their blog posts with what are effectively free text labels, or tags. This is similar to tagging behaviour found on sites like del.icio.us[8] and flickr[9]. Such tagging behaviour is of enormous commercial relevance; both companies have been bought by Yahoo[10] and the field is attracting other players both large and small [44]. Tom Gruber has proposed a tag ontology [45]; not a controlled vocabulary for tagging; rather a way of specifying *who* said *what*, and in what *context* (tagging system and/or tagging weight; one can 'negatively' tag). Such an approach seems to offer the possibility of combining tagging information from different systems. This is particularly interesting when we mine the tags for their hidden structure; Peter Mika calls this *emergent semantics* [35]. The idea of bootstrapping an ontology as a side effect of a community's tagging behaviour is I think a powerful one whose potential has not yet been fully realized.

5. Conclusion

In summary then I think semantic blogging has a strong future – not as a standalone application, but rather as a player in a world where information is treated as a first class citizen. Rich structuring of exported data will allow blog information to be coupled with other sources like email, wikis, IM, newgroups, forums and publication databases. Emerging methods in information extraction, concept tracking and social network analysis will extend the value we can derive from blog information. Collaborative tagging approaches will encourage participation and contribute to evolving information models. The semantic blogging meme will continue to influence and enhance our semantic applications for many years to come.

Acknowledgements

References

[1] Cayzer, Steve, Shabajee, Paul. Semantic Blogging and Bibliography Management Blogtalk - the First European Conference on Weblogs (Blogtalk 2003) Vienna, Austria 23-24 May 2003

[2] Dave Reynolds, Steve Cayzer, Paul Shabajee, Damian Steer (2004) SWAD-E Demonstrators - Lessons Learnt available at: http://www.w3.org/2001/sw/Europe/reports/demo-lessons-report/ accessed 13 October 2006

[3] Semantic Blog (HP) http://www.semanticblogging.org/ accessed 13 October 2006

[4] Cayzer, Steve. Semantic Blogging : Spreading the Semantic Web Meme. In: Proceedings of XML Europe 2004, Amsterdam Netherlands April 18-21 2004

---

[8] http://del.icio.us
[9] http://www.flickr.com
[10] Yahoo bought Flickr in March 2005, and del.icio.us in December that year.

[5] I. Ohmukai and H. Takeda. Semblog: Personal Publishing Platform with RSS and FOAF. In 1st Workshop on Friend of a Friend, Social Networking and the (Semantic) Web, Proceedings, pages 217–221, Galway, September 2004.

[6] The Friend of a Friend Project http://www.foaf-project.org/ accessed 13 October 2006. Current specification http://xmlns.com/foaf/0.1/ accessed 13 October 2006

[7] Karger, David R. Quan, Dennis. What Would It Mean to Blog on the Semantic Web? Journal of Web Semantics 3 (2) 147-157 (2005)

[8] schraefel, m. c., Smith, D. A., Owens, A., Russell, A., Harris, C. and Wilson, M. L. (2005) The evolving mSpace platform: leveraging the Semantic Web on the Trail of the Memex. In *Proceedings of Hypertext, 2005*, Salzburg.

[9] G. Li, V. Uren, E. Motta, S. Buckingham-Shum, J. Domingue. ClaiMaker: Weaving a Semantic Web of Research Papers In: Proceedings of the 1st International Semantic Web Conference, Sardinia, June 2002.

[10] Semantic Blogging (KMI Project) http://kmi.open.ac.uk/projects/semanticblog/ accessed 13 October 2006

[11] Bertrand Sereno, Simon Buckingham Shum and Enrico Motta. ClaimSpotter: an Environment to Support Sensemaking with Knowledge Triples. International conference on Intelligent User Interfaces (IUI 2005), San Diego, CA, USA. 9-12 January 2005

[12] K. Möller, U. Bojars and J. G. Breslin. Using Semantics to Enhance the Blogging Experience. In 3rd European Semantic Web Conference (ESWC2006), Budva, Montenegro, June 2006.

[13] RDF Site Summary (RSS) 1.0 http://web.resource.org/rss/1.0/ accessed 13 October 2006

[14] Atom publishing protocol IETF RFC 4287 (December 2005) http://tools.ietf.org/html/rfc4287 accessed 13 October 2006

[15] RSS 2.0 Specification http://blogs.law.harvard.edu/tech/rss accessed 13 October 2006

[16] http://microformats.org/ accessed 13 October 2006

[17] http://structuredblogging.org/ accessed 13 October 2006

[18] http://esw.w3.org/topic/MicroModels accessed 13 October 2006

[19] BlogAccord demo: http://www.ug.it.usyd.edu.au/~vhou4029/soft3200/demo/index.html accessed 13 October 2006

[20] MusicBrainz http://musicbrainz.org/ accessed 13 October 2006

[21] http://wiki.ontoworld.org/index.php/Semantic_Wiki_Interest_Group accessed 13 October 2006 and http://ontoworld.org/index.php/Semantic_MediaWiki accessed 13 October 2006

[22] McDowell, Luke; Etzioni, Oren; Halevy, Alon. Semantic Email: Theory and Applications Journal of Web Semantics 2 (2): 153-183 (2004)

[23] http://planetrdf.com/ accessed 13 October 2006 (for a description of how it is generated, see http://www-128.ibm.com/developerworks/xml/library/x-pblog/ accessed 13 October 2006)

[24] David Huynh, Stefano Mazzocchi, and David Karger. Piggy Bank: Experience the Semantic Web Inside Your Web Browser. In: Proceedings of the 4th International Semantic Web Conference, Galway, Ireland 6-10 November 2005.

[25] Pat Croke, Ann Johnston and Kim Tighe. Using Named Entities as a basis to share associative trails between Semantic Desktops In: Proceedings of the 1st Workshop on The Semantic Desktop. 4th International Semantic Web Conference (Galway, Ireland) eds Stefan Decker, Jack Park, Dennis Quan, Leo Sauermann (2005)

[26] Steve Cayzer and Paolo Castagna. How to build a Snippet Manager In: Proceedings of the 1st Workshop on The Semantic Desktop. 4th International Semantic Web Conference (Galway, Ireland) eds Stefan Decker, Jack Park, Dennis Quan, Leo Sauermann (2005)

[27] Cayzer, Steve. Semantic Blogging and Decentralized Knowledge Management. *Communications of the ACM* 47(12): 47 - 52 (December 2004)

[28] Tim Clark, June Kinoshita, Yong Gao, Elizabeth Wu, Ryan Lee and Eric Miller. 2006 *SWAN*: *A Distributed Knowledge Infrastructure for Alzheimer Disease Research* Journal of Web Semantics, Volume 4, Number 1, January 2006 222-228

[29] Sebastian Ryszard Kruk, Stefan Decker. Semantic Social Collaborative Filtering with FOAFRealm In: Proceedings of the 1st Workshop on The Semantic Desktop. 4th International Semantic Web Conference (Galway, Ireland) eds Stefan Decker, Jack Park, Dennis Quan, Leo Sauermann (2005)

[30] Oscar Celma, Miquel Ramírez, Perfecto Herrera: Foafing the Music: A Music Recommendation System based on RSS Feeds and User Preferences In: ISMIR 2005 (6th International Conference on Music Information Retrieval) London, UK 11 - 15 September 2005 (pp. 464-467) http://www.semanticaudio.org/foafin-the-music/ accessed 13 October 2006

[31] Tim Finin, Li Ding, Lina Zhou, and Anupam Joshi. Social Networking on the Semantic Web. The Learning Organization 2005 5 (12): 418-435

[32] Eytan Adar and Lada A. Adamic. Tracking Information Epidemics in Blogspace. In: Web Intelligence 2005, Compiegne, France, Sept. 19-22, 2005.

[33] Gilad Mishne, Experiments with Mood Classification in Blog Posts. In: Style2005 – the 1st Workshop on Stylistic Analysis of Text for Information Access, at SIGIR 2005, August 2005.

[34] R. Navigli & P. Velardi. Structural Semantic Interconnections: a knowledge-based approach to word sense disambiguation. IEEE Transactions on Pattern Analysis and Machine Intelligence (27-7), 2005.

[35] Peter Mika. "Ontologies are us: A unified model of social networks and semantics" In: Proceedings of the 4th International Semantic Web Conference, Galway, Ireland 6-10 November 2005.

[36] http://www.w3.org/RDF/ accessed 13 October 2006

[37] Ferrara, A., Ludovico, L. A., Montanelli, S., Castano, S., and Haus, G. 2006. A Semantic Web ontology for context-based classification and retrieval of music resources. *ACM Trans. Multimedia Comput. Commun. Appl.* 2, 3 (Aug. 2006), 177-198. DOI= http://doi.acm.org/10.1145/1152149.1152151 accessed 13 October 2006

[38] Stephen Downes (2005) Semantic Networks and Social Networks *The Learning Organization Journal* 12(5) 2006 411-417

[39] David Newman, Debra Logan 2006 Spotlight on Enterprise Information Management Gartner Report 2 June 2006

[40] Corporate Weblogs: deployment, promotion and measurement Jupiter Research, June 2006

[41] Christopher Harris-Jones 2006 Collaboration software market analysis Ovum Report 11 April 2006

[42] SIOC Ontology Specification (27 Sep 2006) http://rdfs.org/sioc/spec/ accessed 13 October 2006

[43] Tim Finin, Li Ding, Lina Zhou and Anupam Joshi (2005) Social Networking on the Semantic Web *The Learning Organization* 5 (12) 2005 418-435

[44] Joanne Lustig 2006 *Playing Tag in the Enterprise* Outsell Now September 1 2006

[45] Thomas Gruber (2005). Folksonomy of Ontology: A Mash-up of Apples and Oranges. First on-Line conference on Metadata and Semantics Research (MTSR'05). http://mtsr.sigsemis.org/ Paper available at http://tomgruber.org/writing/mtsr05-ontology-of-folksonomy.htm accessed 13 October 2006