By STEVE CAYZER

Semantic Blogging
AND DECENTRALIZED KNOWLEDGE MANAGEMENT

Tapping into the structured metadata in snippets of information gives communities of interest effective access to their collective knowledge.

In the Semantic Web research group at Hewlett-Packard Laboratories, Bristol, we frequently circulate items of interest (such as news articles, software tools, links to Web sites, and competitor information). We call them snippets, or information nuggets, we would like to store, annotate, and share. Email is not the ideal medium for these tasks; its transient nature means the snippets are effectively lost over time. Yet the risk from using a more formal process, like a centralized database, is that it is both cumbersome to use (a barrier to entry) and overly rigid in its data model (not amenable to storing different types of information). Our need illustrates what I call decentralized, informal knowledge management [5].

ILLUSTRATION BY GARY CLEMENT

We are looking for a system capable of aggregating, annotating, indexing, and searching a community's snippets. The challenges we would face in developing such a system include:

Ease of use and capture. Capturing the snippets should be easy at a useful level of detail while causing minimal disruption to users' normal activities;

Decentralized aggregation. Though snippets are likely to be scattered throughout an organization in a variety of locations and stored in a variety of formats, it should be possible to integrate them and perform some global search over the result;

Distributed knowledge. Information consumers can add value by enriching snippets at the point of use by, say, adding ratings, annotations, relationships, and categories;

Flexible data model. Snippets are polymorphic; depending on the task, people may reasonably want to capture email, Web pages, documents, text fragments, and images;

Extensible. It should be possible (post hoc) to not only enrich snippets but extend the snippet data schema to model the changing world; and

Inferencing. It should be possible to infer new metadata from old; to take a trivial example, a machine should be able to "know" that a snippet about a particular HP Photosmart model is about a digital camera?

Some commentators in the technology media have suggested blogs make an ideal tool for this kind of knowledge management [10]. But blogging offers only half the solution (roughly the top three capabilities just outlined). For the rest, we must look elsewhere, first at today's blogging technologies. For example, one popular blogging site—Google's www.blogger.com—defines blogging as "push-button publishing for the people." Indeed, blogging tools provide several routes to speedy publishing, ranging from online Web forms, to email, to the mobile phone. So in one sense, blogs are a method for the quick creation of Web pages. But with the help of Really Simple Syndication (RSS) and Atom (both XML-based syndication formats for blogs), blogs can publish machine-readable summaries of their content, allowing individual or community aggregators to collect, merge, sort, and index this data. In addition, hyperlinks, trackbacks (essentially reverse hyperlinks that identify who is talking about me, rather than who I am talking to), and recommended blogs (blogrolls) provide both coarse- (blog-to-blog) and fine-grain (item-to-item) link elements for the blogosphere. Blogging is thus a powerful tool for establishing and maintaining an online community.

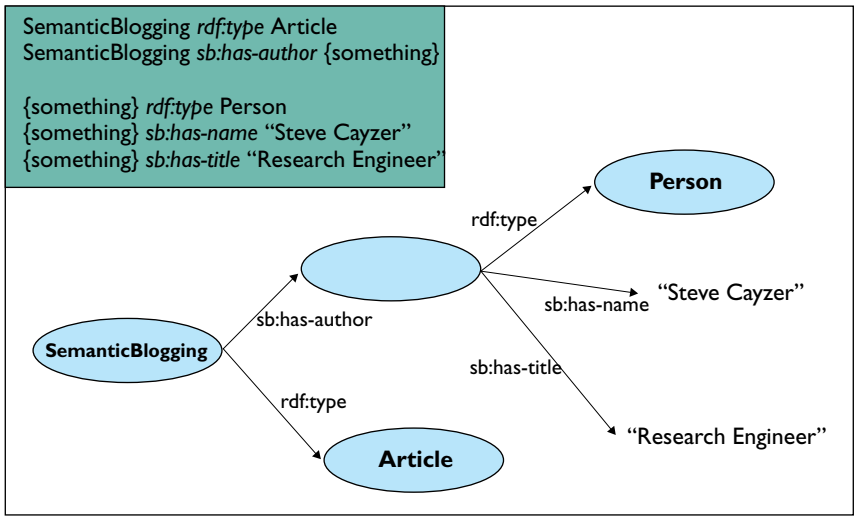

Figure 1. RDF graph encoding information concerning this article, including its type and author; sb and rdf are namespace abbreviations. Modeling the author as a "resource" (a node in the graph) allows anyone to attach other information (such as job title), perhaps post-hoc. The strings (or literals) are end points in the graph, to which no further information can be added.

Blogging's greatest benefit is social, not technological. First, ease of use makes it likely that more people will publish and publish more often, and that more information will be communicated. The structure of the information is often different from more static home pages, more like online journals, or (at a higher level) a series of snippets. Because blogs are interlinked, they lend themselves to a sort of digital ecology [1]. The structured nature of RSS allows bloggers and readers alike to integrate and search information feeds (such as BBC News).

Are these desirable capabilities enough to address and overcome the knowledge-management challenges outlined earlier? No, because in traditional blogging, metadata is used only for headline syndication. Metadata is not extensible, not linked to a rich, flexible data model, and certainly not capable of supporting vocabulary mixing and inferencing. Metadata can be extended and is what the existing blog standard RSS1.0 aims to do (web.resource.org/rss/1.0/).

How might Web developers improve and extend the way information is handled on the Web? One contender is the Semantic Web, a common frame-

work that allows data to be shared and reused across application, enterprise, and community boundaries; information is given well-defined meaning, better enabling computers and people to cooperate [3]. RSS1.0 is a Semantic Web vocabulary that provides ways to express and integrate with rich information models (web.resource.org/rss/1.0). The Semantic Web standard Resource Description Framework (RDF) specifies, in essence, a Web-scale information-modeling format (www.w3.org/RDF/). The key element in RDF is the triple—a simple subject-predicate-object construct—that can be joined to create a graph-like structure, with subjects (or objects) as their links, or arcs. Figure 1 outlines a simple example, representing statements concerning this article, expressing roughly the following assertions: "Semantic Blogging" is an article whose author is a person with name Steve Cayzer and title Research Engineer.

RDF provides several useful features for rich information processing:

- Like XML, it is an open standard designed by the World Wide Web Consortium specifically for Web use;
- Arcs are both directional and labeled, while nodes may be labeled or blank; the identities of blank nodes are unimportant and can be identified through their properties. Labels take the form of globally unique identifiers, or uniform resource identifiers (URIs), allowing graphs with shared URIs to be merged. (A subset of the URI is the familiar Web address, or URL.) URIs have a namespace (roughly a vocabulary identifier) that allows synonyms (such as "has-title" in Figure 1) to be defined in different vocabularies without risk of accidental clashes. The result is that data integration across the Web is more tractable; and
- RDF vocabularies facilitate arbitrary enrichment of blog entries, perhaps even post-hoc (by bloggers and readers alike). Examples include core blogging vocabularies (such as RSS), international metadata standards (such as Dublin Core), social links (such as the Friend of a Friend project, or FOAF), personal information (such as virtual business cards), and event schemes (such as Apple Computer's desktop calendar iCal); and
- RDF metadata is backed by an ontology [6] that allows useful inferences to be drawn and implicit information to be derived.

Semantic Web technologies provide a useful way to address the last three of these desiderata. Consider capturing, say, an announcement (a snippet) posted on a Web page by a chief technology officer of a particular company, then retrieve it by querying on any of the relevant attributes (such as "this month's announcements" and "snippets related to company X"). Now imagine that extra information about the company could be added by a third party or integrated from an external source, and that it could be used to retrieve the snippet. Rich information models allow inference-enabled querying, perhaps employing background ontologies (such as "Find all snippets on this company's competitors").

The Semantic Web is best viewed as an enabling technology. Although my colleagues and I have shown it is possible to build dedicated Semantic Web snippet-management applications [2], the technology is likely to gain more traction when set within a familiar and lightweight context. In particular, a well-known problem in the Semantic Web is bootstrapping, or creating enough initial metadata to generate the network effect. In other words, how can we give information providers enough immediate value to encourage them to produce consistent, high-quality metadata in the first place? Blog software developers must do three things:

- Make it easy to get data into the information ecosystem;
- Use automatic mechanisms to assist metadata creation; and
- Provide mechanisms for individual bloggers and their readers to add value through third-party annotations and community enrichment.

Each can be explored by embedding Semantic Web technology within a blogging framework.

Combined Features

Semantic blogging is an attempt to use the desirable features of both blogging and the Semantic Web; examples include:

- Using the blog-creation process to expedite the lightweight capture of snippets and other items of interest;
- Using an RDF vocabulary (such as RSS1.0) to represent and export blog metadata;
- Using consistent URIs for snippets, so information can be integrated across multiple blog entries;
- Enriching snippet metadata using published vocabularies (such as Dublin Core and FOAF); and
- Creating, publishing, and sharing specialized domain-specific vocabularies.

Blog software developers must keep snippet capture lightweight, yet allow snippets to be stored in an accessible knowledge repository. Post-hoc enrichment provides adaptability and future proofing. As the carriers of snippets (think of a blog entry as an annotation attached to a snippet), blog entries can facilitate the linking of semantic blog output with other snippet data sources.

Semantic blogging demonstrator. I built a simple prototype application in 2003 as part of a pan-European project called Semantic Web Advanced Development-Europe (SWAD-E), a European Union-funded effort to bring Semantic Web technologies to a broad developer community. I set the prototype in the domain of a small group's bibliography management effort while retaining the essential characteristics of a more general snippet manager. My vision of semantic blogging is that Semantic Web technologies will enable new blogging modalities that would be difficult or impossible otherwise. For the purpose of the demonstrator, I chose three types of functionality:

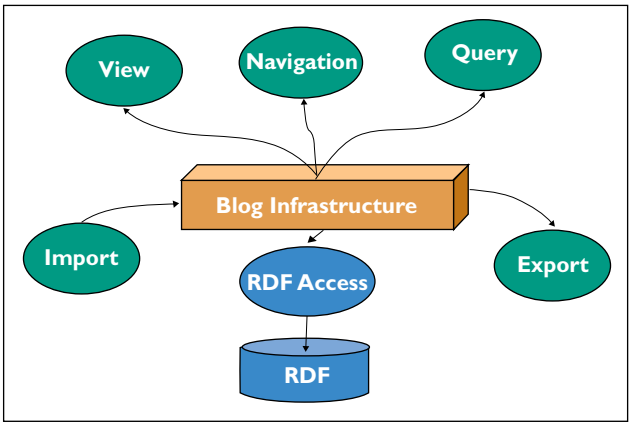

Figure 2. High-level architecture of the semantic blogging demonstrator prototype application.

Semantic view. Semantic metadata is used to enhance the snippet sharer's view of rich content. For example, bibliographic items could be viewed as record cards, arranged in tables, or grouped in clusters. The semantic view is chosen by the snippet reader, not the provider, on the basis of the metadata. Readers could thus arrange someone else's blog entries according to their personal categorization scheme;

Semantic navigation. Semantic metadata is used to more efficiently find blog items of interest; examples are the use of dynamically constructed tree-type interfaces to browse related items and follow labeled links (such as "agrees with" and "part of"). Another option is defining a semantic similarity measure in order to "find similar entries;"

Semantic query. Semantic metadata can help build rich queries for accessing a community's collective knowledge (such as "Find all blog entries about a paper written by this author" and "Find all blog items about my friends").

Figure 2 outlines the demonstrator's basic structure in which a semantic metadata store is built beneath a blog infrastructure (see www.semanticblogging.org/blojsom-devt/blog/). Input and output mechanisms complete the metadata pipeline, over which the semantic capabilities described earlier are built. I built over a Java blogging platform called blojsom (blojsom.sf.net/), using the HP Labs Semantic Web toolkit Jena (www.hpl.hp.com/semweb/jena.htm). A full description of the demonstrator's design is in [4].

Figure 3 is a demonstrator screenshot, including simple navigation and query options in the left-hand pane. Note that queries can be about blog entries (annotations), as well as about the underlying items. That is, the author of an article is not the same as the author of a blog entry about the article, although it is reasonable to want to search on either type of author. The panel is schema-controlled, so the exact form (both presentation and content) can be customized at runtime. The right-hand pane contains a table-based view of the search results. Again, the exact form of the view returned is schema-controlled, providing a mechanism for personalization.

Metadata control means that different forms can

How can we give information providers enough immediate value to encourage them to produce consistent, high-quality metadata in the first place?

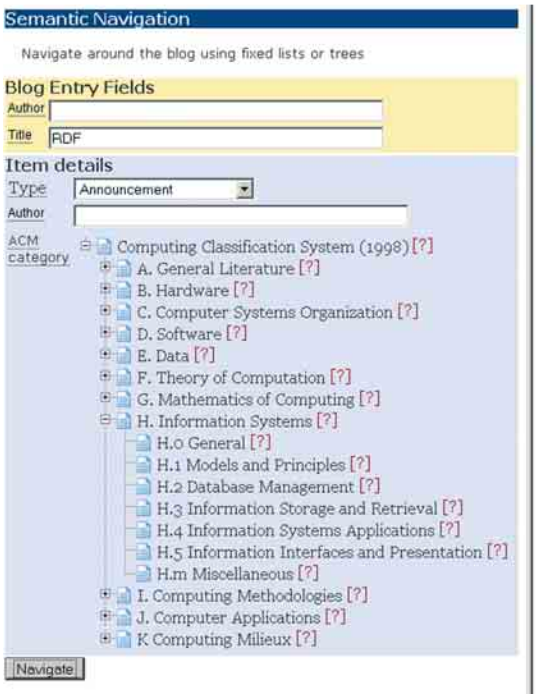

Figure 3. Snippets can be searched for, either through their own attributes (such as "I'm interested in snippets about HP") or through the attributes of their attached blog entry (such as "I'm interested in snippets captured by Bob"). Here, the query is summarized and the results are returned in a plain table format.

be used for different types of input, whether papers, conference proceedings, or news articles. The demonstrator also provides a simple ontology-backed mechanism to help users create metadata. Using the Simple Knowledge Organisation Systems schema, an RDF schema for thesauri and related knowledge-organization systems [7], make it possible to represent a group of categories as a hierarchical set of concepts. A user sharing snippets can thus associate a set of ranked (preferred and nonpreferred) indicator terms with each concept. When creating blog entries, the demonstrator analyzes the text using a simple stemmer [9], enabling the snippet sharer to rank categories that might be suitable for the blog post.

Genuinely useful tool. The demonstrator was a simple prototype, designed to illustrate Semantic Web values. My colleagues and I are now planning to exploit its ideas and technologies for several purposes:

Semantic blogging for knowledge management. Semantic Web (and blogging) values play well in decentralized knowledge management. We are exploring how to exploit them both for HP internally and for HP's customers;

Departmental portal. My peer group within HP needs a community snippet-manager application for information items ranging from articles read to technical tips and news items. We are assessing semantic blogging as a lightweight solution to this need; and

Community tools. The RDF developer community is rich and vibrant, and a number of developers have begun to take code, components, and ideas from semantic blogging for such diverse purposes as collaborative storytelling, educational frameworks, and social networks.

For the latest on semantic blogging, see my blog at www.semanticblogging.org.

It is fair to ask whether other approaches might bring about these same benefits. For example, do we need the Semantic Web at all? Even without semantic metadata, much can be done to move blogging toward structured knowledge management. For example, one system—www.bloglines.com—uses aggregated blogs to recommend new, interesting blogs to its subscribers. Another—called Meme Streams, www.memestreams.net/—tracks "memes," or whatever is copied from one person to another on the network, as they spread across the blogosphere. Yet another—called Waypath, www.waypath.com—uses blogged URLs as a way of linking blog entries "about" the same item. However, there is a limit to how far these mechanisms can go before they need a richer information model.

The idea of blogging for knowledge management is not new [10], but blogs are not yet widely deployed as corporate intranet solutions.[1] Other solutions (such as email and threaded discussion boards) are used, along with their attendant limitations of post-hoc access (for email) and system lock-in (for discussion boards). Blogs are imperfect [12]. Consider, for example, the problem of the signal-to-noise ratio in content; more

[1] For example, a recent survey by the American Productivity and Quality Center found that blog use ranked 2.38 on a scale of 1 (not used) to 7 (used extensively); www.apqc.org/site/images/Guerrilla_Technologies_Proposal.pdf.

information inevitably means more irrelevant, incomplete, or inaccurate content. A key challenge is finding a way to filter, sort, and navigate through the blogosphere. The Semantic Web can help here, too, by enriching blogs with richer, structured metadata and by providing mechanisms for recommendation networks.

Structure is, perhaps, a more troublesome issue. Consider, for example, the needs of shared categorization so you can link other people's information into a conceptual scheme that makes sense to you. Here again, the Semantic Web can help by providing a standard knowledge-representation format, together with tools for decentralized ontology creation and linking.

The Semantic Web is not the only way to embed such structure in XML. Topic Maps [8] are an information-modeling technique that allows conceptual maps to be represented, linked, and shared. Indeed, much of the value of semantic blogging might also be implemented over topic maps. However, RDF does offer one interesting capability—inferencing over a rich information model—thus enabling the creation of implicit metadata.

The idea of decentralized ontology creation is not unique. Systems like the Topic Exchange (topicexchange.com), del.icio.us (del.icio.us for social bookmarking), and Flickr (www.flickr.com for online sharing of photographs) allow a community of users to collaboratively build up a knowledge structure (actually a list of tags). However, in all of these systems the ontology lacks semantics and are both centralized and universal. The Semantic Web may be better served by precise, local, domain-specific vocabularies that are loosely coupled, rather than by a one-size-fits-all central ontology, no matter how collaborative.

Some technologies can be used to help power semantic blog applications. A blogger might, for example, build semantic capabilities into Cascading Style Sheets to provide rich, user-specific customizations [11]. One might also use data formats like Easy News Topics (ENT) (matt.blogs.it/specs/ENT/1.0/), a lightweight metadata-creation tool and central portal (the k-collector) to provide a collaborative view of a community's blog postings. The metadata-creation tool is a useful way of providing machine-assisted metadata creation, one I extended in the demonstrator application. The k-collector is much more flexible than the Topic Exchange but not as flexible as an RDF model.

The semantic blogging demonstrator is complete and the lessons learned documented [4]. Meanwhile, other groups continue to apply semantic blogging ideas to communal blogs (such as Planet RDF, www.planetrdf.com), wikis (such as Platypus, platypuswiki.sourceforge.net/), aggregators (such as the semblog platform, www.semblog.org/wiki/), and authoring tools (such as Compendium, www.compendiuminstitute.org).

Conclusion

Semantic blogging is not yet a paradigm but is already more than a tool. I have sought to present its key ideas, describe an initial implementation, and point to ongoing work. I look forward to the future with interest. C

References

1. Adar, E. and Adamic, L. Tracking information epidemics in blogspace. Posted 2003; www.hpl.hp.com/research/idl/papers/blogs2/index.html.
2. Banks, D., Cayzer, S., Dickinson, I., and Reynolds, D. *The ePerson Snippet Manager: A Semantic Web Application.* HP Labs Technical Report HPL-2002-328; www.hpl.hp.com/techreports/2002/HPL-2002-328.html.
3. Berners-Lee, T., Hendler, J., and Lassila, O. The Semantic Web. *Scientific American 284,* 5 (May 2001), 34–43.
4. Cayzer, S. *Semantic Blogging: Lessons Learnt.* Tech. Rep. SWAD-E 12.1.8, Sept. 2003; www.w3.org/2001/sw/Europe/reports/demo-lessons-report/.
5. Davenport, T., and Prusak, L. *Working Knowledge: How Organizations Manage What They Know.* Harvard Business School Press, Boston, 2000.
6. Gruninger, M. and Lee, J. Ontology applications and design. *Commun. ACM 45,* 2 (Feb. 2002), 39–41.
7. Miles, A., Rogers, N., and Beckett, D. *SKOS-Core 1.0 Guide.* W3C Draft, Mar. 2004; www.w3.org/2001/sw/Europe/reports/thes/1.0/guide/.
8. Park, J., Ed. *XML Topic Maps: Creating and Using Topic Maps for the Web.* Addison-Wesley Professional, Boston, 2003.
9. Porter, M. An algorithm for suffix stripping. *Program 14,* 3 (1980), 130–137.
10. Roll, M. Business Weblogs: A pragmatic approach for introducing Weblogs in medium and large enterprises. In *BlogTalks,* T. Burg, Ed. Herstellung: Books on Demand GmbH, Norderstedt, Germany, 2003.
11. Udell, J. The Semantic Blog. O'Reilly xml.com, Apr. 15, 2003; webservices.xml.com/pub/a/ws/2003/04/15/semanticblog.html.
12. Weiss, A. The last word: Your blog? Who gives a @*#%! *ACM netWorker 8,* 1 (2004).

Steve Cayzer (steve.cayzer@hp.com) is a research engineer in the Semantic Web research group at Hewlett-Packard Laboratories, Bristol, U.K.