Microformats: a Pragmatic Path to the Semantic Web

Rohit Khare
CommerceNet Labs
169 University Avenue
Palo Alto, CA 94301
+1 650 714 5529
rohit@commerce.net

Tantek Çelik
Technorati
665 3rd Street, Suite 207
San Francisco, CA 94107
+1 415 896 3000
tantek@technorati.com

ABSTRACT

Microformats are a clever adaptation of semantic XHTML that makes it easier to publish, index, and extract semi-structured information such as tags, calendar entries, contact information, and reviews on the Web. This makes it a pragmatic path towards achieving the vision set forth for the Semantic Web.

Even though it sidesteps the existing “technology stack” of RDF, ontologies, and Artificial Intelligence-inspired processing tools, various microformats have emerged that parallel the goals of several well-known Semantic Web projects.

This poster compares their prospects to the Semantic Web according to Rogers’ Diffusion of Innovation model.

Categories and Subject Descriptors

I.7.2 [**Document and Text Processing**]: Document Preparation – *markup languages, hypertext/hypermedia.*

H.3.5 [**Information Storage and Retrieval**]: Online Information Services – *Web-based services.*

General Terms

Design, Standardization, Human Factors

Keywords

Microformats, Semantic Web, Decentralization, HTML, CSS

1. INTRODUCTION

> *Designed for humans first and machines second, microformats are a set of simple, open data formats built upon existing and widely adopted standards. —microformats.org*

By taking full advantage of the existing XHTML facilities such as `class` attributes, microformats can make existing Web pages easier to recycle into new services and applications. This is a key aspect of the original appeal of the Semantic Web [2]. To a lesser degree, it was the original appeal of XML as well [7].

Several early applications of microformats can be compared to related projects in the Semantic Web and XML communities. At the same time, it is also important to acknowledge the limits of the microformats community’s approach.

While microformats can encode *explicit* information to aid machine readability, microformats do not address *implicit* knowledge representation, ontological analysis, or logical inference. They are a new on-ramp for the vision of a Semantic Web that make semantic markup more usable for authors and developers that both communities ought to embrace together.

WWW 2006, May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-323-9/06/0005.

2. MICROFORMATS

Innovation in software can sometimes be ascribed to overcoming ‘accidental’ or ‘essential’ challenges — and it is worth acknowledging at the outset that the case for microformats may owe more to accident than essence.

These terms come from the classic essay, *No Silver Bullet* essay [4] to distinguish between practical limitation of our tools at a moment in time, rather than a gap in our theoretical understanding of the problem. A straightforward example is that a separate file format for machine-readable information, however powerful, may not succeed simply because it uses another file.

It turns out that in most weblogging tools it can be complicated or even impossible to upload a file attachment. Even images introduce new difficulties such as off-site hosting on a separate photo service, much less uploading an RDF file or proprietary metadata and linking to it from the “plain text”

That’s admittedly an accidental consequence of our tools, which also make it that much easier for a writer with some knowledge of HTML to encode additional semantic information such as calendar events (hCalendar), contact information (hCard), and typed hyperlinks (rel-tag, XFN) using microformats.

2.1 Calendar Entries

To publicize an upcoming lecture, for example, one must clearly state its time, place, duration, and speaker. The first few concerns are so broadly applicable that international calendaring and scheduling standards already address them. Issues such as timezones, recurrences, organizers, performers, and locations are a few of the debates settled by vCalendar and its Internet-specific successor, iCalendar [6, 11].

The class names highlighted in Figure 1 were not chosen at random. The payoff for choosing those is that the announcement no longer requires a separate `.vcf` file in the first place.\

```
<div class="vcalendar vevent">
 <span class="summary">Microformats: What the Hell Are
They and Why Should I Care?</span>
 <p class="description">Ryan King will explain why
microformats are important and how you can mark up specific
kinds of content in ways that make it easier for the right people
to find your stuff.</p>
 <abbr class="dtstart" title="20050926T050000-
0700">September 25th, 2005, 5</abbr>–
 <abbr class="dtend" title="20050926T060000-
0700">6PM</abbr>
 in the <span class="location">Balder Room</span>
</div>
```

Figure 1: An event in microformatted XHTML.

The inline style information is sufficient to encode the same information that the other formats did — especially when combined with a of the lesser-known element in the XHTML specification to "abbreviate" the machine-readable ISO8601 timestamps [8] that correspond to natural-language phrases in the original (human-readable) description.

2.2 Typed Hyperlinks

> *"If one web site links to another, the link doesn't carry any information about why the sites are linked. But what if it did?"* —Knowledge@Wharton [1]

The most successful microformat for human generated content is tagging. Within the first six months of introducing a way to tag blog posts, Technorati was tracking 20 million blog posts; and today over a third of all entries include tags [10].

Typed link relations are a mainstay of hypertext theory, but have generally been overlooked on the Web. Consider the social networking phenomenon of "blogrolls": lists of one author's favorite blogs to read, presented as a list of links in the margin. While more abstract efforts exist to represent the combination of contact/profile information and social network relationships (e.g. FOAF, the RDF friend-of-a-friend format [3]), the XHTML Friends Network (XFN, [5]) took the approach of focusing only on the social relationship aspect, by adding link relationships to existing blogrolls. The vocabulary chosen was based on a study of common (the "80%") relationships that bloggers indicated publicly on their web logs. This is incomplete in the theoretical sense — but still solves "80%" of the problem.

3. COMPARISONS TO XML

If XML's essential strength – decentralized evolution of new tag sets — was also its essential weakness, then there would be little to be gain by simply renaming the problem of Babel by encouraging random mutation of new HTML class names. Technically, classes do add a degree of freedom, insofar as each XHTML element can have multiple classes (it's a space-separated list), whereas an XML element is limited to a single tag name.

Socially, however, the key insight is that microformats appeal to authority by migrating existing standards or codifying common practices. Rather than creating a new calendaring specification out of thin air, hCalendar attempts to reuse the names, objects, properties, values, types, hierarchies, and constraints from RFC2445 iCalendar. It doesn't even interpose its own clever prefix: it may be called hCalendar, but it uses class names spelled `vcalendar` and `vevent` because those are the case-insensitive transliteration of the labels from the original specifications.

4. DIFFUSION OF INNOVATION

In 1962, Everett Rogers published the first edition of his seminal text on the sociology of technology adoption, *Diffusion of Innovations* [9]. It introduced terms such as "early adopter" and studied innovations both in the form of objects and as practices, in fields as diverse as farmers evaluating new strains of seeds to the introduction of videogame systems (in later editions)..

4.1 Relative Advantage

> *Relative advantage is the degree to which an innovation is perceived as better than the idea it supersedes.*

As with all of these factors, the key is an individual's *perception* of advantage. For authors, publishing metadata once and in-line with the data is cheaper to maintain. This results from a deliberate decision to favor ease of authoring to break the deadlock of adopting new formats.

4.2 Compatibility

> *Compatibility is the degree to which an innovation is perceived as being consistent with the existing values, past experiences, and needs of potential adopters…*

The basic framing of this debate is "compatible for *whom?"* For AI-influenced researchers and developers, technologies that explicitly reference ontologies, rules, and structure are desirable. For hypertext authors, these are unfamiliar concepts that place "knowledge management" beyond the bounds of their discipline.

4.3 Complexity

> *Complexity is the degree to which an innovation is perceived as difficult to understand and use…*

Part of the original promise of XML was to enhance the Web so that strings that looked like prices actually were prices; microformats promise a similar improvement by incrementally adapting XHTML with constructs familiar from CSS.

4.4 Trialability

> *Trialability is the degree to which an innovation may be experimented with on a limited basis.*

The full power of XHTML is almost always available in Web content management systems, without requiring new tools support, or linking to external resources.

4.5 Observability

> *Observability is the degree to which the results of an innovation are visible to others…*

Early applications have taken full advantage of the Web browser as a platform for detecting, parsing, storing, sharing, and searching snippets of structured data captured from web pages; there is complementary enthusiasm for semi-structured search, particularly with tagging.

ACKNOWLEDGMENTS

We would like to thank Danny Ayers, Dan Connolly, and Jim Hendler for a clearer understanding of the potential of connecting microformats to the vision of a Semantic Web.

REFERENCES

[1] *What's the Next Big Thing on the Web? It May Be a Small, Simple Thing — Microformats* in *Knowledge@Wharton*, 2005. (27 July).

[2] Berners-Lee, T., Hendler, J. and Lassila, O. *The Semantic Web* in *Scientific American*, May, 2001. v.*284* (5), pp. 34-43.

[3] Brickley, D. and Miller, L. *FOAF Vocabulary Spec.* 2005.

[4] Brooks, F. P. *No Silver Bullet: Essence and Accidents of Software Engineering* in *IEEE Computer*, 1987.

[5] Çelik, T. and Meyer, E. *XHTML Friends Network (Poster)*, in *ACM Hypertext 2004*, (Santa Cruz, CA, 9-13 Aug. 2004).

[6] Dawson, F. and Stenerson, D. *RFC 2445: Internet Calendaring and Scheduling Core Object Specification (iCalendar).* IETF, November 1998.

[7] Khare, R. and Rifkin, A. *The Origin of (Document) Species* in *Computer Networks and ISDN Systems*, 1998, *30*.p.389-97

[8] Klyne, G. and Newman, C. *RFC 3339: Date and Time on the Internet: Timestamps.* IETF, July 2002.

[9] Rogers, E. M. *Diffusion of Innovations* 4th ed. Free Press, 1995. 518pp.

[10] Sifry, D. *State of the Blogosphere, Aug 2005, Part 3: Tags.*

[11] versit Consortium. *vCalendar: The Electronic Calendaring and Scheduling Exchange Format (Ver. 1.0).* 18 Sep 1996.