Microformats

A Pragmatic Path to the Semantic Web

Rohit Khare, Ph.D
CommerceNet Labs

Tantek Çelik
Technorati, Inc.

CommerceNet Labs Technical Report 06-01
January 2006

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 position paper introduces the ideas behind microformats and gives examples; compares to similar efforts in the Semantic Web; and compares their prospects 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

Authorship

This report also draws upon the first author's column for *IEEE Internet Computing,* "Microformats: The Next (Small?) Thing on the Semantic Web?" Additional portions have been submitted to the 15th International World Wide Web Conference.

Mr. Çelik is Chief Technologist for Technorati, Inc. He can be reached at 665 3rd Street, Suite 207, San Francisco, CA 94107, or as <tantek@technorati.com>.

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 [29]*

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 part of the original appeal of the Semantic Web [3]. To a lesser degree, it was the original appeal of XML as well [24].

Nor surprisingly, several early applications of microformats parallel related projects in the Semantic Web and XML communities. This paper takes closer look at the strengths and weaknesses of the microformat alternatives in arguing our position that aggressive evolution of XHTML authoring practices can go a long way to achieving the goals set forth for the Semantic Web without invoking the complexity of the technology stack traditionally associated with it.

At the same time, it is also important to acknowledge the limits of the microformats community's approach. While it can encode explicit information to aid machine readability, microformats do not address implicit knowledge representation, ontological analysis, or logical inference. They are, indeed, a pragmatic path towards the full vision of the Semantic Web, a new on-ramp that make semantic markup more usable for authors and developers that both communities ought to embrace together.

1.1 Usability matters for markup, too

Explicitly designing new formats for ease of authoring follows and leverage HTML's extremely successful adoption model. In addition, by modeling microformats on *existing* publishing behavior, these formats are well-adapted to what humans do today with their content, rather than asking them to adapt to a new format and rethinking their content according to new abstractions.

Providing standards for marking up existing implicit content building blocks lowers barriers for creating, assembling, mixing, presenting, and sharing semi-structured information across a variety of systems and purposes. Furthermore, what makes a document into a hypertext is the addition of links; microformat extensions to HTML's `rel` and `rev` attributes provide particu-

larly fertile ground for enhancing link analysis along the axes of social relationships, topicality, licensing, and enclosures.

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.

Our use of these terms is based on the classic *No Silver Bullet* essay [6] 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 resides in an external 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.

That's largely 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 using microformats. In this section, we'll consider three kinds of such information: calendar events (hCalendar), contact information (hCard), and typed hyperlinks (rel-tag, XFN).

2.1 Calendar Entries

To publicize an upcoming lecture, 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 [18, 20, 38].

```
BEGIN:VCALENDAR
BEGIN:VEVENT
SUMMARY:Microformats: What the Hell Are They and Why Should I Care?
DTSTART:20050926T000000Z
LOCATION:Balder Room
DTEND:20050926T010000Z
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.
END:VEVENT
END:VCALENDAR
```

Figure 1: *An example of an event in vCalendar format.*

```
<rdf:RDF
 xmlns:rdf='http://www.w3.org/1999/02/22-rdf-syntax-ns#'
 xmlns='http://www.w3.org/2002/12/cal/ical#'>
 <Vcalendar>
  <prodid>-//kanzaki.com//RDFCal 1.0//EN </prodid>
  <version>2.0</version>
  <method>PUBLISH</method>
  <component>
   <Vevent>
    <dtstart rdf:parseType='Resource'>
     <dateTime>2005-09-26T00:00:00Z</dateTime>
    </dtstart>
    <dtend rdf:parseType='Resource'>
     <dateTime>2005-09-26T01:00:00Z</dateTime>
    </dtend>
    <summary>Microformats: What the Hell Are They and Why Should I Care?
    </summary>
    <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.</description>
    <location>Balder Room</location>
    <dtstamp>20051012T061505Z</dtstamp>
    <uid>1129097705622@kanzaki.com</uid>
   </Vevent>
  </component>
 </Vcalendar>
</rdf:RDF>
```

Figure 2: *The same event in RDF Calendar format.*

An example taken from [25] shows that all the announcer has to do is link to a `myEvent.vcs` file as shown in Figure 1.

Hopefully, a `.vcs` file is machine-readable, because it certainly isn't very human-readable. Still, modern software tools favor angle-brackets over colon-delimited header/value pairs. Therefore, an even "better" answer using the Semantic Web is to use RDF Calendar [16]. Figure 2 shows the output from RDFical-a-matic [22], a web service that accepts vCalendar files as input.

Regardless of the file format for machine consumption, both of these two alternatives require the announcer to hyperlink to an external resource. Figure 3 is an example of HTML anchor text that might describe the event described by the file it refers to.

```
<a href="/myEvent.vcs">
  <b>Microformats: What the Hell Are They and Why Should I Care?</b>
  <p>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>
  <small>September 25th, 2005, 5-6PM in the <i>Balder Room</i></small> </a>
```

Figure 3: *The same event in presentational HTML format.*

This example merits the epithet "presentational HTML" because it uses inline formatting tags like <small> that originated in the willy-nilly growth of HTML tags during the Netscape-Microsoft Browser Wars from 1995-1998. A "better" answer for modern browser accessibility is to use Cascading Style Sheets (CSS, [28]). As shown in Figure 4, adding class attributes to HTML elements allows an external stylesheet to define its look and feel.

```
<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 4: *The same event in microformatted XHTML.*

The class names chosen (highlighted in boldface in Figure 4) were not chosen at random. The payoff for choosing them as we did is that the announcer doesn't need a separate file in the first place.

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 [30] to "abbreviate" the machine-readable ISO8601 timestamps [26] that correspond to natural-language phrases in the original (human-readable) description.

2.2 Contact Information

Similarly, hCard [15] takes as its starting point an IETF specification for using vCard in email [17]. To add hCard microformatting to an existing Web page, the first step is to ensure that the most appropriate XHTML elements are being used. Then, each of the data fields defined in vCard are mapped on to the most appropriate element.

Thus, a vCard's `URL` data field for a homepage is best captured as an ordinary hypertext anchor, but denoted by the url class name: `<a class="url" href="...">...</a>`. By extension, the vCard `EMAIL` data field is also represented by a link, but one that uses a `mailto:` URL scheme instead of HTTP. A link to a person's photograph, however, is not mapped onto the generic <A> anchor; instead PHOTO becomes a class on `<IMG>` images.

Some data fields can occur more than once, or have further internal structure. Singular keys like a formatted name (`fn`) are resolved by using only the first matching descendant element; but since a person can have many telephone numbers, each and every instance of a descendant element with class `tel` should be preserved, with its additional flags such as `home`, `work`, `fax`, or `pref` ("preferred").

Finally, the results of the transformation so far have to be evaluated for how well they balance human- and machine- readability. For example, the information about whether a phone line supports faxing or goes to the person's residence ought to be kept visible. Putting it in a class attribute hides it from the reader, so an additional layer of indirection had to be added using the class `type`, whose element text is interpreted according to the vCard list of values. Similarly, since a telephone number now has to be decomposed into its attributes as well as the number itself, the class `value` is introduced to make that separation explicit. It also follows from the rule about keeping critical information user-visible, as shown in Figure 5, rather than hiding `home` as a class attribute after `tel` (as was the case in earlier revisions of hCard).

```
<span class="tel">
    <span class="type">home</span>:
    <span class="value">+1.415.555.1212</span>
</span>
```

Figure 5: hCard keeps the "home telephone" qualifier visible.

The astute reader with an eye toward localization might note that this overloads the use of a previously machine-readable flag. The four-letter string "home" might also be meaningful in English, but we're still waiting for real-world experience with the complications of, say, German: `<span class="type home">Haupttelefon</span>`, perhaps?

To complete the mapping of vCard to XHTML, there are other loose ends to tie up. Certain elements lost their meaning, such as `prodid`, `version`, and `source`. Over time, hCard users also gained enough experience to suggest certain optimizations, like using the words in a formatted name as the implied `given-name` and `family-names` of the compound name property, `n`; preferring `organization-name` when an `organizational-unit` isn't mentioned; and assuming an hCard represents corporate contact information when `fn` and `org` have precisely the same value, e.g. by appearing as class names on the same element.

2.3 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. According to Technorati, within the first six months of introducing a way to tag blog posts the way photos were on Flickr, they were tracking 20 million blog posts. Today over a third of all entries are tagged [36].

RelTag [9] couldn't be much simpler: it's a value for the REL attribute of a hyperlink. To indicate that some entry relates to ice cream, all one has to insert is `<a href= "http://technorati.com/tags/ice-cream" rel="tag"> Ice Cream!</a>`. The last component of the URL path is used as the tag name for further indexing, so users can cite or create any tag vocabulary they like. The simplicity of merely adding a link relation to make blog posts much easier to search for and correlate with each other resulted in dramatic adoption, as illustrated in Figure 6.

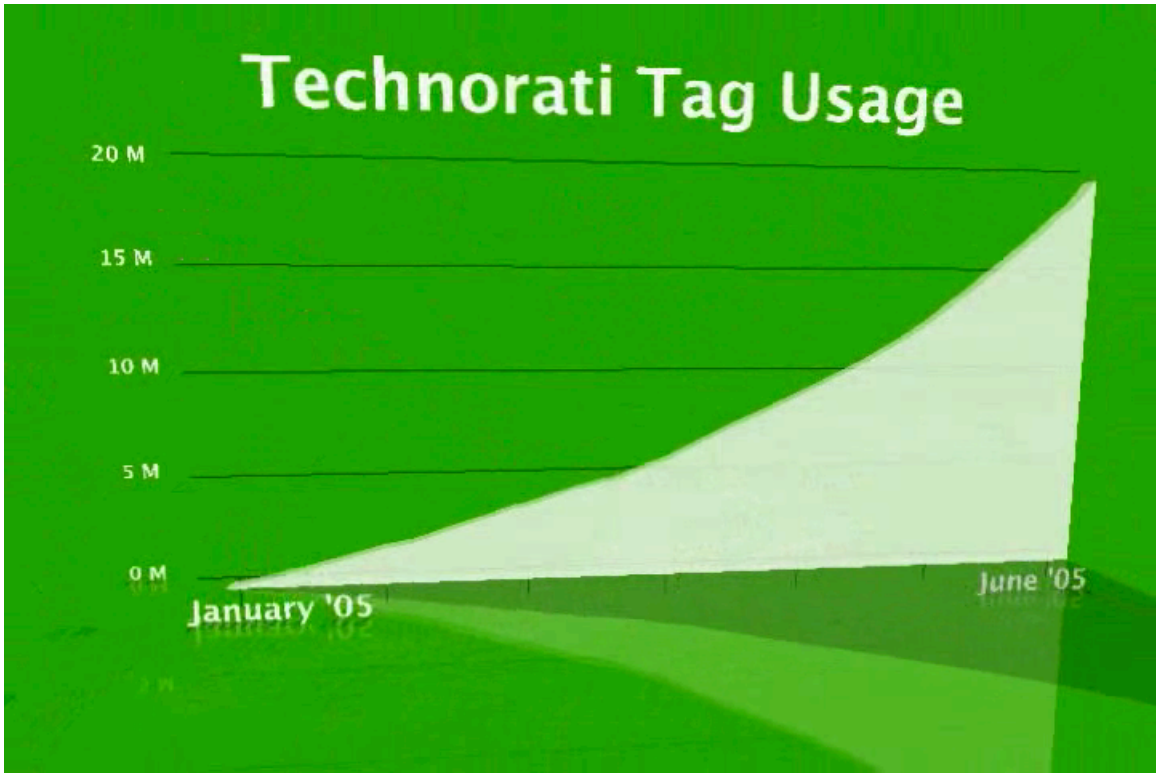

Figure 6: *The relTag microformat was adopted rapidly by the blog community, growing to 20M tagged posts in 6 months. Taken from a video visualization by Carnegie Mellon researchers [2].*

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 [5]), the XHTML Friends Network (XFN, [11]) 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, as shown in Figure 7.

Friendship	`contact`, `acquaintance`, `friend` (*pick one)*
Physical	`met` (*presumed symmetric)*
Professional	`co-worker`, `colleague`

Geographic	`co-resident`, `neighbor`
Family	`child`, `parent`, `sibling`, `spouse`, `kin` (*pick one)*
Romantic	`muse`, `crush`, `date`, `sweetheart` (*not always symmetric!)*
Identity	`me` (*excludes all other types)*

Figure 7: *The (deliberately limited) vocabulary of the XHTML Friends Network.*

The last of those (`rel="me"`) is the most intriguing. It may seem superfluous, but in a world of fragmented digital identity across multiple isolated websites, it provides a pivot point for future integration.

3. The Full Potential of XHTML

> *"But a web full of XML documents of arbitrary application; 'plain XML'? That future never happened."*
>
> *—David Janes [21]*

Now this process may have seemed *ad hoc*, but there is a fairly principled transformation for encoding event metadata into XHTML. This section will describe how it works; we'll return to the "why" in the next section.

When XML was new, CSS was scarce, and the Browser Wars raged, HTML was often cast as a hopeless muddle. Instead, the "Web of HTML" might have given way to a "Web of XML" where each publisher used their own tags and their own presentation logic within a new generation of browsers [24]. Now, it happens to be that in 2005 users have access to fairly full-featured XML+XSL browsers on the desktop, but it's past too late. Like Java, XML seems to have found its niche on the server rather than the client.

In the meantime, HTML grew up and became a proper XML application, offering all of the rigor and modularization an information architect could ask for, while CSS support matured to the point that authors and designers adopted it broadly. This was the key ecological change that triggered the resurgence of experimentation with "plain old HTML."

If XML's essential strength – decentralized evolution of new tag sets — was also its essential weakness, there's little to be gained by simply renaming the problem of Babel by encouraging random mutation of new HTML class names. Technically, this does add a degree of freedom insofar as each XHTML element can have multiple classes (it's a space-separated list), where as 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 [14] attempts to reuse the names, objects, properties, values, types, hierarchies, and constraints from RFC2445 iCalendar. It doesn't even interpose a clever prefix: it may be called hCalendar, but it uses class names spelled `vcalendar` and `vevent` because that's the case-insensitive transliteration of the labels in the original specification.

Mechanically, spaces become dashes, and plurals are made singular (hence, `additional-name` in hCard). The latter rule is because lists of values are expanded out into multiple sibling elements in the DOM. Similarly, hierarchical containment relationships are reflected by the nesting of XHTML elements. Particularly human-unreadable data is indirected through the `<abbr>` construct.

Each of these rules elaborates one basic theme: use the most semantically appropriate XHTML element in the first place. In the examples above, there were `<div>`s and `<span>`s, but those are actually the last resort. Better choices are to use existing list, dictionary, link, or quote constructs.

Predating the class name based compound microformats, the extensions to the rel/rev attributes emerged as the simplest valid extensions to (X)HTML.

These link-based microformats emerged first in the form of the social relationship standard XFN, later extending to VoteLinks, the nofollow link (to avoid influencing search engines' ranking algorithms), and license links. Today, these are called elemental microformats, while the kind of inline semantic text markup presented earlier (as well as hReview, [10]) are called compound microformats, which often make use of the elemental microformats.

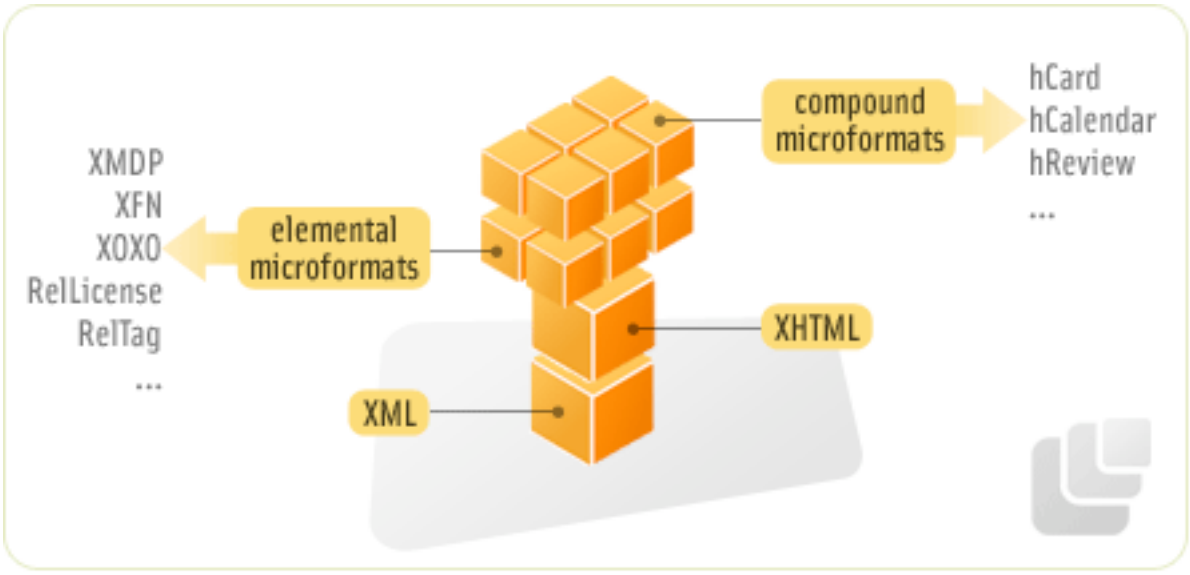

Figure 8: *How compound microformats relate to elemental ones.*

Historically, the stack illustrated above is a roughly accurate timeline. Like a sedimentary fossil record, successful adoption at each layer has been essential for the later diversification and growth.

3.1 Philosophy

A skeptic might well note that the two examples presented so far are just two facets of the same specification. A cynic might go further and ask whether microformats are simply a matter of slapping an "h" in front of an existing specification: call it the "h* Effect."

Microformats advocates may want to celebrate such criticism, since it only serves to underscore the philosophy of "reduce, reuse, and recycle." It's shorthand for a number of design principles that contrast strongly with existing standards bodies and their processes.

Reduce. The community and process serves to focus attention on a *specific problem* ("how can we point to licensing terms for weblog posts?") and favors the *simplest solutions.*

Reuse. Never proceed from *a priori* reasoning alone; work from experience and favor examples of current practice. Always keep in mind Picasso's dictum, "lesser artists borrow; great artists steal" — avoid NIH and embrace any existing, widely-adopted interoperably implemented schemas.

Recycle. Make sure the results make it easy to decentralize innovation by encouraging modularity and embeddability. By making sure microformats are always valid XHTML, they can be carried in blog posts, ATOM and RSS feeds, and anywhere else the Web can be accessed.

In fairness, though, an XML advocate would readily salute the same flags. It may be easy to create a new DTD, but anyone would prefer reusing existing standards. The key point of departure between the Semantic Web and the "lowercase semantic web" community is the rallying cry "design for humans first, machines second."

Presentable *and* **Parseable.** Ruby's postulate states, "The accuracy of metadata is inversely proportional to the square of the distance between the data and the metadata" [35]. Combined with the venerable maxim "out of sight, out of mind," these are the reasons why the microformats community insists on keeping semi-structured information in-band and visible.

3.2 Don't Repeat Yourself (DRY)

In general, duplicating data is bad; that only multiplies the opportunities for errors to creep in over time. Ironically, both classical metadata efforts (use of <meta> keywords and other hidden Dublin Core properties to duplicate information already in the document), and newer XML/RDF efforts (hiding markup and duplicated content in comments or <script> tags) are both egregious violations of this principle.

An evolutionary process controlled by natural selection presumes that innovation is always occurring in fits and starts. Microformats are not the only alternative to the intelligent design of the Semantic Web. The original vision of XML is also adapting to the environment of weblogs, in the form of RSS, Atom, and experiments such as Structured Blogging [40].

Rather than treating the XHTML content as authoritative and weaving metadata around it, the original Structured Blogging proposal embedded 'plain XML' within a `<script>` tag, and then uses additional Dynamic HTML techniques to present it to end-users in a browser.

This has the advantage of complete decentralization for creating new vocabularies, with an additional level of schema constraints for enumerated values and other basic data types. The disadvantage is that the semi-structured information is invisible to ordinary browsers, screen readers for the disabled, and creates less incentive for sharing a common vocabulary and modularization.

A subtler consequence is that Structured Blogging forces all the structured information on a Web page into a single redundant ghetto, an island of XML within the larger HTML document.

Microformats can be added to more complex HTML structures, such as table layouts for agenda grids or formatted presentation of bibliographic records.

For all of these reasons, the Structured Blogging community recently announced that all of their tools will produce microformatted XHTML wherever the appropriate specifications exist, including for syndication in stuctured feeds.

3.3 Standardization

The "standards process" for microformats has been growing along with the community.

Originally, XFN was a product of GMPG, a self-proclaimed club of a few designers whose name was borrowed from a science fiction novel.[1] Consequently, the same team defined a format for describing future standards like XFN. XHTML Meta Data Profiles (XMDP, [13]) enumerate class names and rel/rev link attribute values a particular microformat uses. XMDP declarations are linked in from the lesser-known `profile` attribute of the <head> element in HTML 4.01.

XMDP is the foundation for microformats, but it is not as ambitious as other, more powerful schema description languages. It's more of a human-readable help file than a machine-readable set of rules for automating parsing and validation. This sort of 80/20 solution, once again, is why microformats are making headway as a simple authoring solution while more complete Semantic Web description languages have been less widely adopted.

The microformats.org community includes an open wiki, mailing list, and IRC channel that has proven remarkably scalable and accommodating. The only restriction on participation is that copyrights and patents on the resulting specifications must be openly published and entirely royalty-free, respectively. The community also values research into existing standards, which helps damp the tendency to promote too many narrow innovations. "Ruthless self-criticism" is actually one of its published values.

3.4 Developing hReview

Early in 2005, this process rapidly developed a new format for publishing reviews. Whether of books, movies, restaurants, or myriad other items, reviews are a common idiom in weblog postings, and developers of those tools wanted a common way to share them with search engines that could aggregate community opinions. hReview [10] was the result, jointly authored by individuals from AOL, Microsoft, Yahoo!, Six Apart, and others.

---

[1] Global Multimedia Protocols Group first appeared in Chapter 3 of *Snow Crash* by Neal Stephenson [12].

This was a watershed because hReview did not appeal to the authority of a pre-existing standard [23]. Instead, a survey of related work uncovered such widely divergent standards as the Platform for Internet Content Rating Services (PICS, [27]) and the idiosyncratic layout of popular e-commerce sites.

The critical break from the past was making hReview independent of the items being reviewed; it contains nothing to specify the book, movie, or restaurant. Instead the `item` property is merely a formatted name, link, image, or an hCard (for reviewing a person or corporation).

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* [34]. 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).

It is by no means the only framework for analyzing technology adoption; much less a foolproof system for predicting the future, but it is instructive to compare how the latest attempt to add semi-structured information to the Web's global hypertext compares to earlier hopes for XML, and current work on RDF.

4.1 Relative Advantage

> *Relative advantage is the degree to which an innovation is perceived as better than the idea it supersedes. [It] may be measured in economic terms, but social prestige, convenience, and satisfaction are also important factors.*

As with all of these factors, the key is an individual's *perception* of advantage. Many technologies with clear, quantitative advantages can fail; indeed, that is the starting point for this entire field of research.

To apply this definition, we must identify what is being "superseded" by microformats or its alternatives. Furthermore, we also ought to quantify advantages for distinct social subgroups: authors, developers, and readers. We presume that all of the individuals involved in the diffusion process favor the automated analysis of semi-structured information on the Web; that is, we stipulate the relative advantage of a more semantic Web.

For authors, the relative advantage of adopting microformats appears to be that publishing metadata once and in-line with the data lowers the cost of maintaining the accuracy of that data. We might go further and posit a cultural argument that authors, by definition, work with HTML everyday and perceive some satisfaction in *not* recasting themselves as "knowledge engineers."

For developers, the case is more mixed. For example, the Document Object Model (DOM, [39]) makes it easy to search for elements by tag name, but doesn't even include an operation to search by class attributes. Regardless of how easy that function is to write — and the implementation is quite simple — it's not part of the existing landscape.

Developers familiar with XML and RDF may both perceive a distinct *dis*advantage to having to parse all of the additional degrees of freedom in XHTML — not just having to ignore extraneous formatting and scripts, but actively having to acknowledge the role of XHTML structuring constructs like dictionary lists. For Web site developers, though, easy access to microformat presentation through CSS is much easier to use.

Finally, for readers, the comparative advantages of these markup approaches may be less distinct. They want to be able to extract addresses, schedule meetings, and plot maps; and the ease or difficulty of doing so depends mainly on the quality of the tools developers build for them.

In fact, this results from a deliberate decision in the microformats community to favor ease of authoring, specifically to break the chicken-and-egg 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 adoption of an incompatible innovation often requires the prior adoption of a new value system, which is a relatively slow process.*

The most important aspect of this definition to note is that compatibility is defined culturally, not technologically. To the degree microformats are preferable because they are completely compliant XHTML, it affects the next three factors, not this one.

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 tools for making sense of the Web. For hypertext authors, who create Web content in the first

place, these are all fairly 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… New ideas that are simpler to understand are adopted more rapidly than innovations that require the adopter to develop new skills…*

Incremental improvement is one of the best ways to reduce perceived complexity. Given that 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 without positing a new technology distinct from HTML.

Furthermore, the use of the class attribute was not chosen in a vacuum, either: the broad interest in CSS design and broad support for CSS rendering ensured that it remains credible to sell semantics as an extension of style.

Forbidding invisible metadata also reduces complexity. There are no "hidden" inference rules at work to extract semi-structured data. To the degree that microformats do push user data into hidden markup elements, it is still anchored (literally) by a visible string, as with `<abbr>`.

4.4 Trialability

> *Trialability is the degree to which an innovation may be experimented with on a limited basis. New ideas that can be tried on the installment plan will generally be adopted more quickly …. [and] learn by doing.*

By definition, all Web content management systems make it possible to use the full power of XHTM. There is no similar confidence for widespread access to storage, data entry, and analysis in alternative markup formats.

For example, any blogger with a text-input area can use a simple Javascript form (hCard Creator, [7]) to fill in contact information that gets pasted in. Using Greasemonkey [31], a browser can even be 'reprogrammed' to always offer that editor when posting a new article.

Because the class attribute is a space-separated field, microformats do not interfere with any document's existing structure and appearance – it is a purely additive step that can be adopted at any scale, from a paragraph to a page to an entire site.

Having to point to an external XML or RDF file in its own vocabulary is a not-insignificant barrier to casual experimentation.

4.5 Observability

> *Observability is the degree to which the results of an innovation are visible to others… Such visibility stimulates peer discussion of a new idea, as friends and neighbors of an adopter often request innovation-evaluation information about it.*

The ultimate benefit of adopting any more-semantic format is realized by clever user-facing applications that *do something* with the data. The Semantic Web community has made very impressive demonstrations of its power in the right vertical areas, but the observed benefits are also limited to those verticals.

At this early stage, many microformats-aware applications take advantage of the Web browser as a development and delivery platform. User scripts such as Mark Pilgrim's MagicLine [32] and Monkey Do [33] are already detecting, parsing, storing, sharing, and searching snippets of structured data captured from web pages.

Another deployment mechanism is online Web services, like an an XSLT transformation that exports hCards it finds on the Web into `.vcf` files suitable for import into any standard address book application[2] (X2V, [37]).

In the long-term, the most compelling observable benefit of adoption will come from search engines. Today, all search engines are HTML-aware, at a minimum to distinguish between headings &c. Blog search engines were critical to the rapid adoption of tags, and are currently driving the growth of calendaring. Discovering related content is a compelling and observable benefit of modifying one's markup.

5. The lowercase semantic web

A classic joke is that once an inventor was showing off the latest gadget, when a scientist comes by and asks, "yes, yes, it works in practice — but does it work in *theory?*" Comparisons between the nascent lowercase semantic web and the Semantic Web tend to raise the same question.

---

[2] Though it's not quite seamless — users still have to beware of whether their tools support either UTF-8 or UTF-16 encoding for international character sets and choose the right translation mode. This sort of annoyance was supposed to go away with the advent of XML, but vCard is still a legacy format that predates the wide availability of Unicode.

At the most recent Web conference, the organizers of Developer's Day held a panel discussion with advocates from both communities, and at least two distinct responses emerged. On one hand, a lowercase semantic web, microformatted HTML approach could be a case of "worse is better" [19], where Gresham's law might apply to crowding out "complete" solutions like FOAF in favor of XFN. On the other, it could be the breakthrough on-ramp that the Semantic Web needs, a seedbed of common personal information that grows alongside 600,000-word ontologies for oncology.

While the elegant and painstakingly interlocked edifice of technologies such as RDF, XML, and query languages are growing powerful enough to attack massive information challenges in disciplines such as bioinformatics [4], incremental, decentralized innovation is continuing within some niches of Web like blogging which are driving simple iteratively evolved microformats.

In the meantime, it may be more illustrative to quote what microformats are *not:*

> *Microformats are not a new language; infinitely extensible and open-ended; an attempt to get everyone to change their behavior and rewrite their tools; a whole new approach that throws away what already works today; nor a panacea for all taxonomies, ontologies, and other such abstractions.*
>
> *— Microformats: Evolving the Web [8]*

6. Acknowledgments

We would like to thank our colleagues at CommerceNet and Technorati for their support, as well as the broader microformats.org community. For a clearer understanding of the potential of connecting microformats and the vision of a Semantic Web, we would also like to thank Danny Ayers, Dan Connolly, and Jim Hendler.

7. 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). http://knowledge.wharton.upenn.edu/index.cfm?fa=printArticle&ID=1247

[2] Art and Computer Science group. *Technorati tags (60 sec.).* Carnegie Mellon University, 23 July 2005. http://www.ourmedia.org/node/37881

[3] Berners-Lee, T., Hendler, J. and Lassila, O. *The Semantic Web* in *Scientific American*, May, 2001. vol. *284* (5), pp. 34-43.
http://scientificamerican.com/print_version.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21

[4] Bostrom, J. *The Key to the Semantic Web* in *Bio-IT World*, 10 June, 2005.
http://www.bio-itworld.com/issues/2005/June/expo-special-report-berners-lee

[5] Brickley, D. and Miller, L. *FOAF Vocabulary Specification.* 27 July 2005. http://xmlns.com/foaf/0.1/

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

[7] Çelik, T. *hCard Creator.* 2005. http://tantek.com/microformats/hcard-creator.html

[8] Çelik, T. *Microformats: Evolving the Web.* 4 October 2005.
http://tantek.com/presentations/2005/10/microformats-evolution/

[9] Çelik, T. *Rel-Tag Specification.* Microformats.org, 10 January 2005. http://microformats.org/wiki/reltag

[10] Çelik, T., Diab, A., McAllister, I., Panzer, J., Rifkin, A. and Sippey, M. *hReview Specification (Draft).* 2005.
http://microformats.org/wiki/hreview

[11] Çelik, T. and Meyer, E. *XHTML Friends Network (Poster)*, in *ACM Hypertext 2004*, (Santa Cruz, CA, 9-13 August 2004). http://www.gmpg.org/xfn/intro

[12] Çelik, T., Meyer, E. and Mullenweg, M. *GMPG History to date.* March 2003. http://gmpg.org/history

[13] Çelik, T., Meyer, E. and Mullenweg, M. *XHTML Meta Data Profiles (Poster)*, in *WWW2005*, (Chiba, Japan, 12 May 2005). http://gmpg.org/xmdp/description

[14] Çelik, T. and Suda, B. *hCalendar Specification.* 2004. http://microformats.org/wiki/hcalendar

[15] Çelik, T. and Suda, B. *hCard Specification.* 2004. http://microformats.org/wiki/hcard

[16] Connolly, D. and Miller, L. *RDF Calendar - an application of the Resource Description Framework to iCalendar Data.* W3C Interest Group Note, 29 September 2005. http://www.w3.org/TR/2005/NOTE-rdfcal-20050929/

[17] Dawson, F. and Howes, T. *RFC 2426: vCard MIME Directory Profile.* IETF, September 1998.

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

[19] Gabriel, R. P. *Patterns of Software: Tales from the Software Community*. Oxford University Press, 1998. 256pp.
http://dreamsongs.com/NewFiles/PatternsOfSoftware.pdf

[20] Internet Mail Consortium. *IETF-Calendar Mailing List.* 1996-present.
http://www.imc.org/ietf-calendar/index.html

[21] Janes, D. P. *XML - what is it good for?*, 4 October 2005.
http://blog.davidjanes.com/mtarchives/2005_10.html#003410

[22] Kanazaki, M. *RDFical-a-matic.* 2005. http://www.kanzaki.com/docs/sw/rdfical-a-matic.html

[23] Khare, R. *hReview in Review*, in *WWW2005 Developer's Day*, (Chiba, Japan, 14 May 2005).
http://cnlabs.commerce.net/~rohit/hReview-in-Review/

[24] Khare, R. and Rifkin, A. *The Origin of (Document) Species* in *Computer Networks and ISDN Systems*, 1998, *30*. pp. 389-397.

[25] King, R. *Microformats: What the Hell Are They and Why Should I Care?*, in *Webzine 2005*, (San Francisco, CA, 24-25 September 2005). http://theryanking.com/presentations/2005/webzine/

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

[27] Krauskopf, T., Miller, J., Resnick, P. and Treese, W. *PICS Label Distribution Label Syntax and Communication Protocols, Version 1.1*. W3C Recommendation, 31-October-96 1996. http://www.w3.org/TR/REC-PICS-labels-961031

[28] Lie, H. W. and Bos, B. *Cascading Style Sheets : Designing for the Web* 3rd ed. Addison-Wesley Professional, 2005. 416pp.

[29] Microformats.org. *Wiki, Blog, and Mailing Lists for the Microformats community.* 20 June 2005. http://Microformats.org/

[30] Pemberton, S. and et al. *REC-xhtml1: XHTML™ 1.0 The Extensible HyperText Markup Language (Second Edition), A Reformulation of HTML 4 in XML 1.0*. World Wide Web Consortium, Cambridge, MA, 1 August 2002. http://www.w3.org/TR/2002/REC-xhtml1-20020801/

[31] Pilgrim, M. *Dive Into Greasemonkey: Teaching an old web new tricks.* 9 May 2005. http://diveintogreasemonkey.org/

[32] Pilgrim, M. *MagicLine: Collecting page metadata & microformats in the browser.* 2 August 2005. http://www.mozdev.org/pipermail/greasemonkey/2005-August/004738.html

[33] Pilgrim, M. *Monkey Do: Automating del.icio.us posting of microformatted pages.* 17 August 2005. http://www.mozdev.org/pipermail/greasemonkey/2005-August/005030.html

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

[35] Ruby, S. *Attractive Nuisance*, in *Applied XML Developers Conference 5*, (Stevenson, WA, 20-21 October 2004). http://intertwingly.net/slides/2004/devcon/68.html

[36] Sifry, D. *State of the Blogosphere, August 2005, Part 3: Tags.* 2005. http://www.technorati.com/weblog/2005/08/37.html

[37] Suda, B. *X2V: hCa* to vCard/iCalendar converter.* 2005. http://suda.co.uk/projects/X2V/

[38] versit Consortium. *vCalendar: The Electronic Calendaring and Scheduling Exchange Format (Version 1.0).* 18 September 1996. http://www.imc.org/pdi/vcal-10.doc

[39] Wood, L. and et al. *REC-DOM-Level-1: Document Object Model (DOM) Level 1 Specification*. World Wide Web Consortium, Cambridge, MA, October 1998. http://www.w3.org/TR/1998/REC-DOM-Level-1-19981001

[40] Wyman, B., Canter, M. and et al. *What is Structured Blogging?*, 2005. http://structuredblogging.org/