intranet librarian

Darlene Fichter (University of Saskatchewan)
With Jeff Wisniewski (University of Pittsburgh)

Microformats and the Search for Meaning

How often have you struggled to find a fragment of data buried in mountains of full text? Whether the information is on the web or your intranet, it can be frustrating to locate exactly what you need without having some context for the content. Whether you're looking for an address, a phone number, an ingredient in a recipe, or an event location, searching for that information can result in a lot of false positives. One way around this is for someone to create the mother of all databases for events, reviews, items for sale, contact information, and all other information under the sun. Assuming this might be technically feasible, imagine trying to convince everyone to add their content to this site and keep it updated in addition to adding this content to their own sites. It's not going to happen.

> How can we make discovering needles in haystacks more like finding elephants in a zoo?

How can we make discovering of needles in haystacks more like finding elephants in a zoo? One way is by publishing content on our sites that it is meaningful to search engines, specifically by using microformats. According to the microformats website (http://microformats.org/about), microformats are "a set of simple open data formats" that humans can read and from which computers can also read and extract information. They're "small bits of HTML that represent things like people, events, tags, etc. in web pages." Developed collaboratively and through open discussion, microformat development is guided by the following principles:

- Solve a specific problem.
- Start as simple as possible.
- Design for humans first, machines second.
- Reuse building blocks from widely adopted standards.
- Aim for modularity and embeddability.
- Enable and encourage decentralized development, content, and services.

Think of it as making information understandable to computers, not just machine readable.

FINDING BURIED TREASURE

The web is filled with millions of pages with buried treasure—pieces of information that describe people, events, and things. What if we could mine this information from full-text documents and know when a name is a name and an event location really is an event location? We really *could* build the mother of all databases of events, recipe ingredients, reviews, or whatever, after the fact. Microformats encode meaning into full text by taking advantage of the class element in HTML to label content as a review, an address, or title. Once information is encoded according to a microformat standard, search engines and other tools know "this is a review" or "this is an event."

For those of you already worrying that this sounds like more work to publish your site and how you can possibly train all your contributors, relax. More and

more tools are supporting microformat creation. These spare you the time and effort of hand coding.

When you glance at a webpage that contains microformats in the code, it looks like any other HTML page. But there is a prize inside this page: Information in a structured format that search engines can extract, and tools that you use, such as calendars and address books, can digest this information with one click instead of the tedious "copy and paste" approach.

Darlene Fichter's contact information encoded in hCard format

Looking at the accompanying figure of Darlene Fichter's hCard, let's pop open the hood and look at the code. Take note of the first line that indicates the class "vcard" and other class names such as street address, locality, region postal code, country name, organization, and other elements. Note that there's nothing terribly exotic here, just some HTML. What makes this code unique, however, is the fact that the div and span names are part of the hCard standard, so any webpage, browser, or application that understands microformats will know what the information in each of these spans and divs is supposed to be.

```
<div id="hcard-Darlene-Fichter" class="vcard">
 <img style="float:left; margin-right:4px" src="http://library2.usask.ca/~fichter/images/fichterborder.jpg" alt="photo of Darlene Fichter" class="photo"/>
 <a class="url fn" href="http://library2.usask.ca/~fichter/">Darlene Fichter</a>
 <div class="org">University of Saskatchewan</div>
 <a class="email" href="mailto:darlene.fichter@usask.ca">darlene.fichter@usask.ca</a>
 <div class="adr">
  <div class="street-address">3 Campus Drive</div>
  <span class="locality">Saskatoon</span>
  <span class="region">SK</span>
  <span class="postal-code">S7N 5A4</span>
  <span class="country-name">Canada</span>
 </div>
 <div class="tel">306-966-7209</div>
 <p style="font-size:smaller;">This <a href="http://microformats.org/wiki/hcard">hCard</a> created with the <a href="http://microformats.org/code/hcard/creator">hCard creator</a>.</p>
 </div>
```

Where to Learn More

100+ Articles and Resources by Jessica Hupp (www.virtualhosting.com/blog/2008/microformats-university-100-articles-and-resources)

Citation microformat efforts (http://microformats.org/wiki/cite)

Microformats wiki (http://microformats.org/wiki)

Microformats and libraries (http://thecorporatelibrarian.com/2006/10/13/microformats-and-libraries)

SEMANTIC WEB

Now imagine this same markup without the semantic tags, no "org," no "email," no "country-name." You're smart, so if you saw this information on a webpage, you'd know it was contact information and, if you wanted to keep it, would know to copy and paste it into your address book or some other application. It might take some looking to locate this "needle in a haystack," but you could do it. Computers, on the other hand, aren't so smart. Without the semantic tag indicating that Saskatoon is a "locality," a computer would no more think it was a place than the name of Japanese animé character.

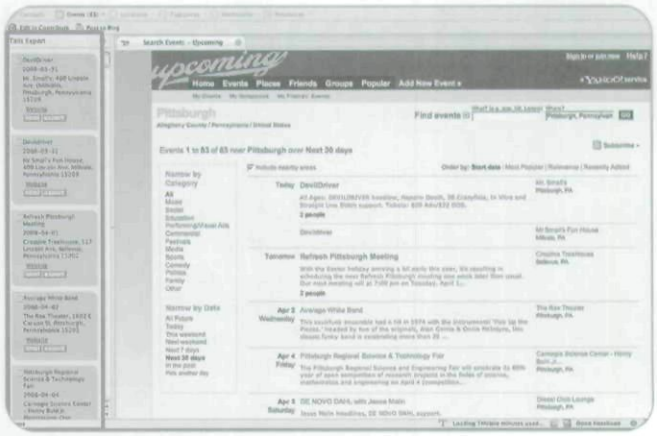
Microformats are one of the key building blocks of the semantic web. Only 3 years old, they recently got a major boost when Yahoo! announced its added support for microformats in search results. Technorati (www.technorati.com) uses microformats for tags, lists, contacts, events, and calendars. Additionally, there is strong indication that an upcoming release of the Firefox browser will support microformats natively.

Semantic markup is nothing new for libraries and librarians, MARC being a major case in point. We have worked for decades with standards that describe data, with 245 being the place for the title and the 020 as the place for the ISBN. In fact, it's so ingrained in everything we do that it's surprising that the massive database we call the web could have grown and prospered without meaningful markup. But indeed it has, and the idea of semantic markup in general, and microformats specifically, are attempting to address this shortcoming.

There's also an ease of use issue. To get the contact information from a webpage into your address book would require a cut and paste effort. With encoded information, and a microformat aware tool such as the next generation Firefox browser, the same contact information could be automatically exported from the webpage into your address book at the click of a button. That's better customer service.

You may have already received microformat encoded

information, such as flight information with a link to add it to your calendar. Multiply this by all of the soccer, football, hockey, dance, and music programs your family attends. Now think about adding the library's book club meetings or the family film night to your calendar. On the intranet, consider microformatting special events, staff training sessions, and your contact database information. Why not talk to library catalog vendors about encoding due date information in the event format? Library patrons could pop open their personal calendars and see what's due when. Just put on your creative thinking cap and you'll see lots of ways that microformats can enhance ease of use.



Using a Firefox Extension to view/export microformats from a Yahoo! webpage (http://upcoming.yahoo.com/place)

GETTING STARTED WITH MICROFORMATS

As you can see from the hCard example, if you know HTML, getting started with microformats is easy. There are a number of different accepted microformats currently in use, including the hCard for contact information and hCalendar for events. There are also a number of draft specifications for encoding latitude and longitude, for resumes and CVs, and for reviews. Creating a citation microformat is a work in progress.

Some ways to use these might include encoding your library staff's contact information, but be creative! Why not make hCards for your "Contact Us" or "Ask-a-Librarian" information? How about using hCalendar to encode your listings of community events, training sessions, or book club meetings? Looking at a next generation library OPAC? If it offers the ability for users to submit reviews, ask if it supports reviews encoded in hReview, so that your reviews can be shared with the world.

CREATION TOOLS

You don't need to worry about learning all of the appropriate divs and spans if you don't want to. There are several tools available to automate the process. Both the hCard (http://microformats.org/code/hcard/creator) and hCalendar (http://microformats.org/code/hcalendar/creator) creators are simple, online forms into which you can enter your contact or event information and have the generator create the code for you to paste into your pages.

Likewise, the free Dreamweaver Microformats Extension (www.webstandards.org/action/dwtf/microformats) provides a list of microformats, pops up a dialog box into which you enter your information, and voilà—you're microformatting.

MICROFORMAT HANDLERS

Currently the two best microformat handlers are both extensions to Firefox, and they perform similarly. Both install icons onto your browser. Whenever they detect the presence of microformat encoded information on a page, the icons "activate" and offer options for exporting the information either directly to sites such as Google Calendar or into the appropriate program on your computer, such as iCal or Outlook. The Events button on the menu bar is highlighted and the upcoming events are displayed in a side pane.

These are two handlers:

- Tails Export: https://addons.mozilla.org/en-US/firefox/addon/2240
- Operator: https://addons.mozilla.org/en-US/firefox/addon/4106

MICROFORMAT SPECIFICATIONS

Much of the early work on microformats was pioneered on Technorati's Developer Wiki (http://developers.technorati.com/wiki) and championed by Tantek Çelik, chief technologist at Technorati. Technorati has passed the microformats development and control to a collaborative and open community at microformats.org.

Anyone is free to participate in the discussion and to propose and develop new microformats, so we librarians are welcome (and bring years of experience with semantics to this party). Be sure to take a look at "So You Wanna Develop a Microformat" at http://microformats.org/wiki/process.

Microformats offer endless possibilities for making the information on your webpages more usable and remixable, which ultimately makes life easier for both you and your users. It's simple to get your toes wet with microformats, so go ahead and give them a try. They're easy to create, easy to use, and, most importantly, useful. As microformats (and the semantic web in general) continue to develop and evolve, this kind of intelligence that now is novel and delightful will one day become a common attribute of the online universe. It is nice to know though that at least for now, we're still smarter than computers.

Darlene Fichter *(fichter@lights.com) is data library coordinator, University of Saskatchewan Libraries.* ***Jeff Wisniewski*** *(jeffw@pitt.edu) is web services librarian, University Library System, University of Pittsburgh.*

Comments? Send email to the editor (marydee@xmission.com).