Web Metadata Standards: Improving Website Visibility

Lisa Goddard
Memorial University Libraries Web Team
August 2009

Librarians are well aware that most users will begin their information search not at the library website, but in a major search engine like Google. It is vital that the library website be well-indexed in search engines, that our hit list results are well formatted and contain useful information, and that our pages receive high rankings when they are particularly relevant to an information need. Metadata and tag structure is also used by screen readers and other accessibility tools. Finally, the way in which a web site is structured will greatly impact its ability to incorporate emerging metadata standards, and the availability of applications which can parse and re-format data for export to other types of applications.

When we talk about metadata on a website, we are talking about much more than simply meta tags. Many things will affect the ability of a third party application to parse, harvest, or re-use website data, including the structure of the page, the tags used, the text in title and H1 tags, class names, and compliance with web standards. The topic of web metadata is therefore rather broad in scope. This report attempts to outline the existing standards that impact the parsing and indexing of web sites. Where appropriate I also note some of the modules in Drupal and Joomla that may be useful for improving page structure, semantic interoperability, and search engine optimization.

Semantic HTML

Semantic HTML refers to the practice of creating documents with HTML that contain only the author's intended meaning, without any reference to how this meaning is presented. Semantic HTML gives search engines and screen readers more accurate metadata from which to draw conclusions about the meaning of a site's content. The list below comprises some of the major principles of Semantic HTML.

Make your structure clear

Resist the temptation to lay your page out in non-standard ways: you want it to be very clear to the search engine where the navigation is, where the content is, and where the headings are. As a rule, put navigation first in your page.

Search engines cull most of their information from the title and <h1> tags.

Always use the heading tags (h1, h2, etc.) for headings and sub-headings. Header tags go from h1 to h6, in order of importance, with the h1 being the most important.

The core block elements in HTML are: div, h1 – h6, p, blockquote

HTML markup should not contain information about presentation

All presentation should be handled by Cascading Style Sheets (CSS), not done in HTML (this means Tables should be used only for tabular data, not for page layout; similarly, spacer GIFs should never be used)

Use all semantic HTML tags, not just the most common ones.

Avoid using generic span and div tags and only making things clear to the user through CSS font sizes: instead, use every 'semantic' HTML tag that applies to your content. If you're quoting someone, use the blockquote tag; if you're posting program code, use the code tag. Search engines love this.

The main disadvantage is that HTML does not contain enough markup tags to describe every single conceivable description or meaning. As such, people will typically use the division ( ) tag along with a set of pre-defined classes or IDs to properly mark up text for their intended meaning. If the designer has a glut of sections or meanings that don't fit well with HTML's markup, they may be forced to use a lot of division ( ) tags, which could easily obfuscate the code.

Use good semantic class names in HTML/CSS

HTML class names should be deliberately chosen to reflect the meaning (semantics) of the content being marked up, rather than the presentation (style). Semantic class names have been a part of modern web design since 2002.

Bad Name	Good Name
border4px	warning
lighttext	submenu
prettybackground	downloadableImage
foobar	booktitle
lisafave	spotlight

Good semantic class names describe what a certain element represents, and they are not likely to change. A warning will always remain a warning, but you may decide to change the width of the border around that message, and suddenly your class name won't be accurate. A submenu will always be a submenu, but maybe the next version of your website will use dark text for the submenu instead of light text. An advantage of using CSS is that you won't have to change much in order to change the looks of your website. If you have to change all light text into dark text, and thus change all classes lighttext to darktext in all your HTML pages, you're likely to miss a few.

Over and above the general housekeeping advantages of semantic class names, they are also the beginning of a major transition into machine-readable content. In order for search engines to perform more accurate and granular indexing, search, and retrieval, they will require clearly structured, semantically meaningful tags.

Keep Keywords Consistent

It's not usually worth deliberately saturating your content with keywords in hope of a higher search ranking – the engines have pretty much wised up to this tactic – but do make sure that your keywords

appear consistently when they occur naturally. Use consistent spelling in your content e.g. 'website' throughout, as suddenly writing 'web site' instead would bring down your search engine rankings.

HTML and Javascript

It's worth noting that search engines read HTML, but they don't, in general, read Javascript. That means that using Javascript to insert text into your page is a bad idea if you want search engines to see the text. On the other hand, you might want to have just the text in HTML and insert all the other parts of the page with Javascript: this will tend to make your page appear more focused, although you should be careful not to insert navigation links this way if you want the search engines to follow them.

Relevant CMS Modules: Drupal & Joomla

Both Drupal and Joomla have various tools and utilities that can help you to control tags and syntax. There are tools that restrict the use of certain tags, tools that validate input to make sure that standards are met, and tools that will attempt to do HTML clean-up like closing tags and quoting attributes.

HTML corrector (Drupal)
http://drupal.org/project/modules?text=HTML%20corrector
Creates a filter to close any unclosed tags when a node is submitted.

Filter: Input formats for user content (Drupal Core)
http://drupal.org/handbook/modules/filter
The filter module allows you to configure formats for text input for your site. For example, you may want a filter to strip out malicious HTML from users' comments. Despite the name "filter," the module not only lets you keep out text you don't want but also lets you enhance the text you let in. So, for example, you can use a filter to turn ordinary line breaks into HTML paragraph tags.

Title Manager (Joomla)
http://extensions.joomla.org/extensions/site-management/seo-a-metadata/3521
This plugin manages browser titles and makes them customizable. It lets you use site name in titles before or after page title with a custom separator. It also accepts an optional site name to use in titles.

Useful Resources:

Guide to Semantic Use of HTML Elements
http://www.joedolson.com/articles/2008/04/guide-to-semantic-html/

Traditional HTML Semantics
http://microformatique.com/?p=83

XHTML

XHTML consists of all the elements in Semantic HTML, combined with the strict syntax of XML. XHTML is not as tolerant of sloppy code as HTML, and will not render properly if tags are nested incorrectly, or if they are not closed properly. Clean, validated code is important if you want to be able to port your web data across different applications and transform it into different formats. For example, browsers that

run on mobile phones or other small devices do not have the resources or power to interpret a "bad" markup language.

All XHTML documents must have a DOCTYPE declaration. The html, head, title, and body elements must be present.

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Transitional//EN"
"http://www.w3.org/TR/xhtml1/DTD/xhtml1-transitional.dtd">
```

Attributes must be written formally as name="value".

Wrong:

```
<option selected>Sydney</option>
```

Right:

```
<option selected="selected">Sydney</option>.
```

The checklist in **Appendix A** outlines some of the most important steps we can take to make our current web pages more XHTML compliant.

Note: IE does not support XHTML, though it can render XHTML documents authored with HTML compatibility principles and served with a text/html MIME-type.

The leap into XHTML from HTML is not great, so long as the HTML code is already well-structured, and valid.

There are two flavours of XHTML, “transitional” and “strict”. Using “strict” implies that anything deprecated (condemned) or made obsolete is not allowed, whereas “transitional” allows deprecated elements in the markup.

Webmasters often find that the move from (X)HTML transitional to (X)HTML strict is actually more difficult than the move from HTML to XHTML. Many webmasters make the move gradually from "street HTML" to "almost valid" HTML to validated HTML transitional to validated HTML strict to validated XHTML 1.0 strict.

Several emerging metadata standards do require XHTML, so we should clean up our code to make it as compliant as possible. This will facilitate a future move to incorporate Microformats, Dublin Core RDFa, or other semantic web formats which rely on XHTML.

Useful Resources:

XHTML validator: http://validator.w3.org/check/referer

CSS validator: http://jigsaw.w3.org/css-validator/check/referer

XHTML Tag Reference: http://www.w3schools.com/tags/default.asp

Metatags

Meta tags have been so badly abused by spammers that search engines pay no attention to them any more when it comes to ranking your site, but they're still important in one way: the meta description tag is still often used to decide what text search engines' users see when they find your site in their results! This can be just as important as the ranking itself – write something here that will look useful to the searcher, and you're more likely to get them to click-through. Don't forget that, while search engines are just machines and algorithms, the end result of it all does involve a human decision: to click, or not to click?

```
<title>Irish studies collection page</title>

<meta name="description" content="A summary of the main strengths of the
Irish studies collection of the Queen Elizabeth II Library." />

<meta name="keywords" content="Political Radical Demography Family history
Newspapers Journals Irish Studies" />
```

Some meta tags are intended as directions to browsers on how to render content:

```
<meta http-equiv="Content-Type" content="text/html; charset=UTF-8" />
```

Some meta tags are used to direct the way in which search engines harvest the page:

```
<meta name="googlebot" content="noindex, noarchive, nosnippets" />
```

is to be placed in the HEAD section of the pages that are to be excluded from the index.

```
<meta name="googlebot" content="nofollow" />
```

instructs Googlebot not to crawl the URLs that the page on which this directive is found links to.

Relevant CMS Modules: Drupal & Joomla

Both Joomla and Drupal have core modules that allow for the central management of metadata keywords and description, as well as a number of additional add-on modules to extend functionality.

Missing Metadata module (Joomla)
http://extensions.joomla.org/extensions/site-management/seo-&-metadata/2846/details
The Missing Metadata module adds a panel to the Control Panel of the Joomla Administrator interface that lists any published articles that have empty metadescription or metakeys fields.

SEO-Generator (Joomla)
http://extensions.joomla.org/extensions/site-management/seo-&-metadata/7171/details
This native Joomla 1.5 plugin automatically generates keywords and description by pulling text from the title and/or the content to help with SEO.

Meta Tags aka Nodewords (Drupal)
http://drupal.org/project/nodewords

This module allows you to set some meta tags for site, content types and nodes. You can also turn off indexing for a particular content type. More attention to the important metadata such as keywords and description on some of your nodes allows you to get better search engine positioning

Yahoo Terms (Drupal)
http://drupal.org/project/yahoo_terms
The Yahoo Term Extractor allows your site to extract key terms out of a piece of content. These terms can be used as metadata for the page, can be used to help find related content, and can be used in conjunction with other services. The Yahoo Terms module provides drupal integration with this service.

Dublin Core

The Dublin Core Standard is used heavily in libraries. There is little proof that search engines pay any particular attention to Dublin Core tags, but as an accepted standard for encoding citation metadata, they may provide a means of exporting page information to other applications. Most of the information sent to OAI harvesters, for example, uses the DC metadata standard, and DC had already been adapted for several flavours of RDF. Dublin Core can be expressed in a number of different ways on a website. Typically, we have seen DC included within meta tags, but DC has been adapted for many different standards, and can be written using more than one syntax.

DC Embedded as Metatags

```
<Meta NAME="DC.Title" CONTENT="Dublin Core Meta Tags">
<Meta NAME="DC.Creator.1" CONTENT="Lisa Goddard">
<Meta NAME="DC.Creator.Address" CONTENT="lgoddard@mun.ca">
<Meta NAME="DC.Creator.2" CONTENT="Wendy Rodgers">
```

DC Embedded as XML:

```
<?xml version="1.0"?>
<metadata
  xmlns="http://example.org/myapp/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://example.org/myapp/
  http://example.org/myapp/schema.xsd"
  xmlns:dc="http://purl.org/dc/elements/1.1/">

  <dc:title>    UKOLN  </dc:title>
  <dc:description>
      UKOLN is a national focus of expertise in digital information
      management. It provides policy, research and awareness services
      to the UK library, information and cultural heritage communities.
      UKOLN is based at the University of Bath.
  </dc:description>
  <dc:publisher>    UKOLN, University of Bath  </dc:publisher>
  <dc:identifier>    http://www.ukoln.ac.uk/   </dc:identifier>
</metadata>
```

DC as XHTML

```
<div xmlns:dc="http://purl.org/dc/elements/1.1/"
  about="http://www.example.com/books/wikinomics">
```

```
  <span property="dc:title">Wikinomics</span>
  <span property="dc:creator">Don Tapscott</span>
  <span property="dc:date">2006-10-01</span>
</div>
```

DC as RDF/XML

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:dcterms="http://purl.org/dc/terms/">

 <rdf:Description rdf:about="http://example.org/123">
  <dcterms:title xml:lang="en">Learning Biology</dcterms:title>
 </rdf:Description>

</rdf:RDF>
```

DC as RDFa

```
<p xmlns:dc="http://purl.org/dc/elements/1.1/"
   about="http://www.example.com/books/wikinomics">
  In his latest book
  <cite property="dc:title">Wikinomics</cite>,
  <span property="dc:creator">Don Tapscott</span>
  explains deep changes in technology,
  demographics and business.
  The book is due to be published in
  <span property="dc:date" content="2006-10-01">October 2006</span>.
</p>
```

Relevant CMS Modules: Drupal & Joomla

Although Drupal developers have begun many projects to create specific modules for generating Dublin Core, none have come to fruition. Joomla has one module designed specifically to generate well formatted Dublin Core meta tags.

Dublin Core Extended (Joomla)
http://extensions.joomla.org/extensions/site-management/seo-&-metadata/6900/details
Simply upload, add the required fields as part of the plugin parameters and publish for all pages.

Useful Resources:

Expressing Dublin Core metadata using the Resource Description Framework (RDF)
http://dublincore.org/documents/dc-rdf/

Dublin Core metadata generator: http://www.ukoln.ac.uk/metadata/dcdot/

Drupal & Dublin Core
http://openconcept.ca/blog/mgifford/adding_dublin_core_metadata_to_drupal

PDF metadata

Customize your PDF properties. The title in the properties becomes the title tag, not the title of the file, so be sure to write an optimized title for your PDF properties. This is what searchers will see in the results of their favorite search engine.

Make sure your PDFs are text-based. Creating a PDF from a Microsoft Word document should be ok, but when you get into other programs like InDesign or Quark, then you run into text-based issues. And, as always, optimize the copy to reflect the targeted keywords.

Pay attention to Accessibility. The accessibility tools can be found under the advanced options. With them, you can order the priority of text on a page and create alt tags for images.

Don't forget about usability. Optimization of a PDF is only good if web users can read it. Be aware that not everyone has the latest version of Adobe Reader. Save your PDF at a version lower than the most recent update. Also, make sure file sizes are as compressed as possible and use properties to see if you have a "fast web view." If not, go to your General Settings panel under preferences to make it web-friendly.

Specify the reading order. This helps search engines know which part of the text is most important on your page. Also, if your first paragraph wouldn't make a great description tag, consider creating a footer and making it number 1 in the reading order.

Build links into PDFs (and be sure to verify them). Keep links to PDFs closer to the root level of the site's file structure. Use keyword-rich anchor text to link to PDFs.

Relevant CMS Modules: Drupal & Joomla

File Framework (Drupal)
http://drupal.org/project/fileframework
File Framework is one option for extracting metadata from PDF files. It is a collection of modules which allow uploading and displaying different media type files as Drupal nodes. It uses the distributed content-addressable storage (CAS) system Bitcache for a file storage and the RDF module for metadata storage. Metadata is extracted from uploaded files using EXIF, getID3, pdfinfo and stored as Resource Description Framework (RDF).

PDF Indexer (Joomla)
http://extensions.joomla.org/extensions/search-a-indexing/site-search/741
Allow PDFs to be searched via the Joomla/Mambo search module. This component allows you to index PDFs on your site and the corresponding plugin (mosbot) allows that index to be searched. Directories that contain PDFs can either be set to public or registered.

Useful Resources:

Optimization Tips for PDFs: Great Advice for B2B Search Marketers
http://blog.searchenginewatch.com/080404-100034

Tagging

User-generated tags can serve as an alternate internal search, can allow users to apply terms that are meaningful to themselves, or can provide staff & librarians with a secondary means of grouping together pages from all over the site. Another means of increasing your site's visibility is to provide widgets that make it easy to tag your pages in other social networking sites like del.ici.ous, or post pages to networks like facebook.

Tag pages can also be indexed in Google e.g. "technorati ebooks" will bring http://technorati.com/r/tag/ebook back as the first hit.

Relevant CMS Modules: Drupal & Joomla

Tags for Joomla (Joomla)
http://extensions.joomla.org/extensions/search-&-indexing/tags-&-clouds/7718/details

The Ultimate Social Bookmarking Plugin (Joomla)
http://extensions.joomla.org/extensions/communities-&-groupware/social-bookmarking/4416/details
This plugin will add social bookmarking buttons to your content, making it easy for your visitors to submit your articles and build traffic to your website.

Active Tags (Drupal)
http://drupal.org/project/active_tags
Active Tags adds a new option to free tagging taxonomies. If selected the taxonomy widget is replaced by a new jQuery enabled tag entry widget. Enforces correct syntax for entering tags (e.g. comma separated).

Suggested Terms (Drupal)
http://drupal.org/project/suggestedterms
This module provides "suggested terms" for free-tagging Taxonomy fields based on terms already submitted. It replaces the description field on free-tagging fields with a clickable list of previously entered terms.

AddThis (Drupal)
http://drupal.org/project/addthis
Provides an addthis button to let your users share your content to social network sites.

Microformats

Microformats are simple conventions for embedding semantic markup in human-readable documents. They are simple ways to add information to a web page using mostly the `class` attribute (although sometimes the `id`, `title`, `rel` or `rev` attributes too). The class names are semantically rich and describe the data they encapsulate.

hCal uses “class” and “id” attributes of XHTML elements

```
<span class="vevent">
    <abbr class="dtstart" title="2006-12-05">December 5-</abbr>
    <abbr class="dtend" title="2006-12-07">7th</abbr>
    </b> At <b><span class="summary">;XML 2006</span></b>
    <span class="location">Boston, MA USA</span>
    for a presentation on "Social Semantic Mashups".
</span>
```

Geo is a microformat used for marking up WGS84 geographical coordinates (latitude;longitude) in (X)HTML. Use of Geo allows parsing tools (for example other websites, or Firefox's Operator extension) to extract the locations, and display them using some other website or mapping tool, or to load them into a GPS device, index or aggregate them, or convert them into an alternative format.

```
<div class="geo">Belvide: <span class="latitude">52.686</span>; <span
class="longitude">-2.193</span></div>
```

rel-license relationship links using “rel” and “rev” attributes on <a> and <link> elements.

```
<a href="http://creativecommons.org/licenses/by/2.0/" rel="license">cc by 2.0</a>
<a href="http://www.apache.org/licenses/LICENSE-2.0" rel="license">Apache 2.0</a>
```

rel-home relationship links using “rel” and “rev” attributes on <a> and <link> elements.

```
<a href="http://technorati.com" rel="home">Home</a>
```

Microformats can be mashed-up and used in-line. E.g. hreview, hcard

```
<div class="hreview">
  <p><span class="item"><strong>Blast 'Em Up   </strong></span>Review</p>
  <p>by <span class="reviewer vcard">
    <p>by <span class="fn">Bob Smith</span>,
```

```
          <span class="title">Senior  Editor</span>at
          <span class="org">ACME Reviews</span>
  </span><p>
  <p><span class="description">This is a great game. I enjoyed it from the
  opening battle to the final showdown with the evil aliens.</span></p>
  <p><span class="rating">4.5</span></p>
</div>
```

Search engines can easily parse web documents to look for microformats and extract them. Google and Yahoo! Search parse almost every defined microformat. Microformats render snippets of data portable, harvestable, and re-useable.

Web services can use this semantic information to create very specific search services:

- Upcoming.org - extracts hCalendar definitions of events,
- Yahoo! Tech – finds product reviews using the hReview format
- Creative Commons Search – finds pages with specific Creative Commons licenses attached

Relevant CMS Modules: Drupal & Joomla

Although I couldn't find any Joomla microformat modules, Drupal has modules to create several different kinds of microformats.

vCard (Drupal)
http://drupal.org/project/vcard

hCard (Drupal)
http://drupal.org/project/hcard

Calais Marmoset (Drupal)
http://drupal.org/project/calais_marmoset
Marmoset adds microformats to web pages on the fly during search engine indexing. When a search robot is identified, Marmoset invokes the OpenCalais Web service and retrieves rich semantic data for the requested page. It then injects the resulting Microformats into the original Web page and returns the result to the search robot.

Creative Commons (Drupal)
http://drupal.org/project/creativecommons
The Creative Commons module allows users to select and assign a Creative Commons license to a node and any attached content. Additionally, the site admin can select a license to assign to the entire site.

Linked Data (aka Semantic Web)

The semantic web is a set of W3C recommended standards that support the creation of very semantically rich web content that describes groups of related resources and the relationships between these resources. The Semantic Web is also known as the "Web of Data", because the intention is to create explicit many-to-many relationships between data from disparate sources all across the web.

The main building block of the semantic web is Resource Description Framework, or RDF, which encodes meaning as sets of Triples (e.g. Shakespeare wrote Macbeth). There are a number of different ways to encode or "serialize" triples, including RDFa, JSON, RDF/XML, N-Triples, and Turtle.

The Semantic Web is still in its infancy, but major search engines like Yahoo & Google have already announced support for RDFa. Drupal7 is supposed to incorporate RDF as a core module, but there are several modules for Drupal6 that provide support for RDF. These modules will be useful to help us explore RDF, learn what can be done, and identify future uses for linked data on our site.

Relevant CMS Modules: Drupal & Joomla

RDF (Drupal)
http://drupal.org/project/rdf
Drupal 7 is on the path to have RDFa integration out of the box. RDFa provides a method for sharing semantic information right in your xhtml pages. But, for those of us developing sites now Drupal 7 isn't an option. Thanks to the RDF module we don't have to wait. This module integrates RDF into your site. Modules like Open Calais take advantage of the RDF module to share your metadata with search engines and other bots.

Calais (Drupal)
http://drupal.org/project/opencalais
The Calais Web Service automatically creates rich semantic metadata for the content you submit – in well under a second. Using natural language processing, machine learning and other methods, Calais analyzes your document and finds the entities within it.

SIOC (Drupal)
http://drupal.org/project/sioc
SIOC (Semantically-Interconnected Online Communities) project is an open specification for describing communities using online discussion forums or blogs, leading to what some may term "distributed conversations". At the moment, online communities are islands that are not interlinked, and the SIOC ontology has been proposed to not only link these communities but to leverage data in ways that were previously unknown.

Search Engine Optimization

No matter what Content Management system we eventually choose, Search Engine Optimization will be important, and cannot be managed with software alone. The site manager should be aware of the frequency with which the site is crawled, the errors are returned by crawlers, the way in which search results appear in various search engines, pages that are being missed by crawlers, etc. The best way to gather this information is to verify the site with Google, Yahoo, and any other major engines. You can then use the Webmaster Tools provided by those services to find out more about how major search engines are handling the site.

Relevant CMS Modules: Drupal & Joomla

Both Drupal and Joomla provide a number of modules that are specifically intended to help with Search Engine Optimization.

SEO Compliance Checker (Drupal)
http://drupal.org/project/seo_checker
The SEO Compliance Checker checks node content on search engine optimization upon its creation or modification. Whenever a publisher creates or modifies a node, the module performs a set of checks and gives the user a feedback on the compliance of the rules.

SEO Checklist (Drupal)
http://drupal.org/project/seo_checklist
This module provides a checklist of good Drupal SEO (Search Engine Optimization) best practices. Maximize the presence of your Drupal website in the major search engines like Google, Yahoo, Bing, etc. It provides a checklist that helps you keep track of what needs to be done. First, it will look to see what modules you already have installed. Then, all you have to do is go down the list of unchecked items and do them. When all the items are checked, you're done!

Google Verify (Joomla)
http://extensions.joomla.org/extensions/site-management/seo-&-metadata/2796/details
Google verify is a small plugin to make Google Webmaster Tools site verification a bit easier. Works with the Meta Tag method and no template adjustments are needed. This is ideal for website with multiple templates or demo sites, no templates adjustments needed when you add a new template this plugin will make sure you stay verified.

Yahoo.com Webmaster Verification (Joomla)
http://extensions.joomla.org/extensions/site-management/seo-&-metadata/6056/details
YahooVerify is a new plugin for Joomla! 1.5 that allows the easy verification of your Joomla website when you register it for indexing on https://siteexplorer.search.yahoo.com/.

Useful Resources:

Google Webmaster Tools
http://www.google.com/support/webmasters/bin/topic.py?topic=8464
Gives you access t information about how Google Will provide crawl stats, crawl errors, Google Page rank, top search queries, other pages that link in to the site, etc.

How to Register with Google, Yahoo, and MSN Webmaster Tools
http://www.semwisdom.com/blog/webmaster-tools
These tools from the major search providers are very useful in telling you what is wrong with your website and pointing out what you can do to improve your website for better search engine consumption.

Joomla SEO web site
http://www.joomlaseo.net/

Internal Search

Both Drupal and Joomla come with built-in site search modules that are highly configurable. There are also modules available that will allow us to offer a more advanced, field-based search engine, or faceted

search. Many different file types can be supported, and the search engine will use metadata as well as full-text content to generate search results. It is also possible to search for tags that have been applied to the pages by end-users or page maintainers.

Field Indexer (Drupal)
http://drupal.org/project/field_indexer
The Field Indexer module indexes field data into Drupal's search index. Each field enabled for indexing becomes a type of index entry. Then, with an appropriate search module, users may perform keyword searches restricted by field.

Faceted Search (Drupal)
http://drupal.org/project/faceted_search
The Faceted Search module provides a search API and a search interface for allowing users to browse content in such a way that they can rapidly get acquainted with the scope and nature of the content, and never feel lost in the data. More than a search interface, this is an information navigation and discovery tool.

Taxonomy (Drupal)
One of Drupal's major strengths is the dynamic taxonomy management module that empowers non-technical users to define and change the structure of their websites. No other CMS tool integrates a dynamic taxonomy module – not even Microsoft SharePoint, which is a fairly new product.

Nodes can be organized into categories, also called taxonomies. Forums are an example of content nodes organized by category. Categories can be hierarchical, where one parent category contains multiple child categories. The thing that makes Drupal Taxonomies different from other category systems is that they can be used as tags (set them when you write a page) and you can have the same page in more than one category at a time (some content systems don't have this option).

- taxonomy is the "system"
- vocabulary are "facets"
- terms are "categories or tags"
- tags are then applied to nodes

Taxonomy Menu (Drupal)
http://drupal.org/project/taxonomy_menu
Transform any of your taxonomy vocabularies into menus easily!

Synonyms (Drupal)
http://drupal.org/project/synonyms
Synonyms is a small module that makes it possible to search for taxonomy term synonyms via the built-in search module.

PixSearch (Joomla)
http://extensions.joomla.org/extensions/search-a-indexing/site-search/3547
The PixSearch module is trying to mimic a search like the one on apple.com. It will perform a backend AJAX Query to the core component com_search and display the results instantly while typing in an inputbox.

Advanced Search (Joomla)
http://extensions.joomla.org/extensions/search-a-indexing/site-search/3849
The Advanced Search extension suite brings professional search features to the Joomla! platform for the first time. Advanced Search extends the power of the default Joomla search engine and gives you the ability to narrow your search results to specific content types such as categories, articles, events, etc...

Content Search SEO Plugin (Joomla)
http://extensions.joomla.org/extensions/search-a-indexing/site-search/6671
A very useful Joomla search SEO Plugin, it is based on the official search content plugin, and add some SEO features. It will generate page title, meta description and meta keywords dynamically according to the search keywords and search results.

PDF Indexer (Joomla)
http://extensions.joomla.org/extensions/search-&-indexing/site-search/741/details
Allow PDFs to be searched via the Joomla/Mambo search module. This component allows you to index PDFs on your site and the corresponding plugin (mosbot) allows that index to be searched.

Useful Resources:

Drupal and the New Paradigm for Content Management
http://digitalsolutions.ph/couchkamotereviews/newCMS

Drupal Taxonomy: the power to organize and reorganize
http://digitalsolutions.ph/couchkamotereviews/power_drupal_categories

Recommendations for Discussion:

We should make efforts to ensure that we are using the principles of POSH (plain Old Semantic HTML) to ensure that presentation is completely divorced from content, and that we are using good semantic class names.

The site currently has a doctype of XHTML transitional, but there are many errors (17 on the main page alone when you run it through the W3C validator). For the next migration we should be aiming to have the site validate as XHTML transitional with very few errors. Several emerging metadata standards do require XHTML, so we should clean up our code to make it as compliant as possible. This will facilitate a future move to incorporate Microformats, Dublin Core RDFa, or other semantic web formats.

We should move gradually towards XHTML strict, which will likely be necessary to support emerging applications.

We need to identify the major types of errors on our current site, and determine the best way to handle those in the migration to a new CMS. The checklist in Appendix 'A' can be used as a guideline for checking our current (X)HTML structure and mark-up.

The web team should arrange for some or all members to receive training in writing XHTML and CSS.

That we explore the core metadata modules in Drupal and Joomla to determine which additional modules are desirable.

That we verify the site with Google and Yahoo in order to use Webmaster Tools to track our site's performance in Search Engines.

Appendix 1 : Checklist for (X)HTML Cleanup

The following guidelines will improve page structure to make it clearer to search engines, ensure that presentation is completely divorced from content, and move the site towards XHTML compliance.

HTML tags should be written in lower case letters.

The first tag on every page should be <html> and the last tag should be </html>

The opening <head> tag must appear only once on every page.

The <title> </title> tags must appear only once on every page, and they should be nested within the <head></head> section.

Each page should have a unique title that concisely describes the page content or purpose.

The <body> element must be present in every page. It should appear directly after the </head> element.

The </body> element must be present in every page. It should appear directly before the </html> element.

The <h1> tag is supposed to signify the main idea of the page and, given this, there should only be one per page and it should be unique to the page.

All presentational tables should be identified and eliminated. Tables should not be used for purely presentational purposes, but should be used only when presenting tabular data.

All spacer GIFs should be identified and eliminated. Spacer GIFs should not be used, but presentation should be done in CSS.

Emphasis tags, such as <b>, are presentational, so should be omitted. The CSS should be used to make chunks of text bold if necessary. The tag should only be used to add semantic emphasis to meaning. Screen readers use to add enunciation, so it will read a <b> tag in a normal voice, but it will emphasize a tag with a "bigger" voice.

The paragraph tag <p> or header tags (<h2>, <h3>, etc) should be used instead of
.

All links (anchor tags) should point somewhere—a link which has an empty href attribute (linking to nowhere) should not be used.

All tags must be closed. Include explicit close tags for elements that permit content but are left empty (for example, , not).

Use the empty-element syntax only for elements specified as empty in HTML. These are:

area, base, basefont, br, col, frame, hr, img, input, isindex, link, meta, param

Include an extra space in empty-element tags:

A horizontal rule: <hr />
An image:
A meta tag: <meta name="keywords" content="Political Radical Demography Family history Newspapers Journals Irish Studies" />

Attribute names must be in lower case.

Attribute values must be quoted.

Change the image “name” attribute to “id”.

This is wrong:
This is correct:

Pages should be validated, which can be done automatically via the W3C Markup Validation Service.