Doing a Content Inventory (Or, A Mind-Numbingly Detailed Odyssey Through Your Web Site)

by Jeffrey Veen
June 18, 2002

I've spent the last year working with clients on a variety of information architecture and design problems. One of the most strikingly consistent issues, however, has been how many of these companies still haven't developed content management systems. I've spoken with enterprises in the Fortune 100 who find themselves sitting on top of 6 years' worth of Web content trapped in static HTML files. They know they need to get this stuff into database and redesign their site into a template-driven system. But their first question is inevitably, "So, uh, where do we start?"

If you're in a similar situation, your first step is to take stock of what you've got. This process, known as a *content inventory*, is a relatively straightforward process of clicking through your Web site and recording what you find. We've developed a simple Excel spreadsheet to help you structure your findings, and some tips on how to get through it.

Start at your home page. Identify the major sections of your site. For example, at adaptivepath.com, we've divided our site into these sections: team, services, workshops, publications, and contact. If I were doing an inventory of this site, I'd start with one of those sections, click in, and see what's linked from it. For each page that I visit, I'd record the information specified in the columns of the spreadsheet. I'd follow every link and navigate as far as I could through the site, making sure to gather data about every possible page on the site.

Here's a description of the things I look for:

	A	B	C	D
1	Link ID	Link Name	Link	Document Type
2	2.0.0	products	http://www.xyz.com/products/index.htm	collector page
3	2.1.0	software	http://www.xyz.com/products/software/i	collector page
4	2.1.1.0	internet software	http://www.xyz.com/products/internet/in	paragraphs
5	2.1.1.1.0	server products	http://www.xyz.com/products/servers/in	paragraphs
6	2.1.1.1.1	web server	http://www.xyz.com/products/servers/w	paragraphs
7	2.1.1.1.2	mail server	http://www.xyz.com/products/servers/m	paragraphs
8	2.1.1.1.3	portal server	http://www.xyz.com/products/servers/po	paragraphs
9	2.1.1.1.4	press releases	http://www.xyz.com/pressreleases/serv	paragraphs
10	2.1.1.1.5	events	http://www.xyz.com/events/2002.html	paragraphs

- **Link ID:** When doing a content inventory, we create a numbering system as we move through the site. It helps us refer back to particular sections and pages as we fill in the spreadsheet. In the sample Excel file, you'll see that "Products" is the second section of the site we've been analyzing, and the pages under the Product page are numbered accordingly. A system like this can prove invaluable later in the process when writing functional or interface specifications.
- **Link Name:** The content you are evaluating needs to be called something. We usually just use the title of the HTML doc, or if that's not specific enough, the headline from the content. It should be unique and descriptive.
- **Link:** It can be very useful to record the URL of the piece of content you're looking at - not only can you click and navigate from the spreadsheet, but you you've also captured the canonical location of the document on the Web server. Remember, the URL should point to the location of the actual HTML file, not a symbolic link or redirect.
- **Document type:** What template does the page use, or which should it? Is it a product page, or a legal brief, or a press release? Every site will have different types of documents, but most have fewer than a couple dozen.
- **Topics, Keywords:** What is the content about? View the source of the page and see what - if anything - is in the "keywords" meta tag. Ideally, you would develop a controlled vocabulary - a collection of approved keywords used to describe your content. This not only helps you choose the appropriate descriptive words, but also keeps your metadata in check. With a controlled vocabulary, you can avoid having half of your content creators labeling transportation stories with "trains" and the other with "locomotives". You can read more about this issue on the IA-Wiki.
- **Owner, Maintainer:** Who created this content? Who maintains it? If you run a smaller site, this may be you, so you might ignore this. With our corporate clients, we assign responsibility for every piece of content.
- **ROT:** This acronym stands for Redundant, Outdated, or Trivial. It's a tag we use to identify content that should possibly be removed from the site. Is there another copy of this content somewhere else? Is it no longer timely? Maybe it should never have been posted in the first place? If it doesn't belong on the site any more, make a note here. This is stuff that shouldn't make the jump to your new database.
- **Notes:** Anything else you may notice. Things to include here are issues like broken images, or other HTML problems.

Really just try to record anything you want to remember for later.

After you've filled in a couple hundred lines of the spreadsheet, you'll inevitably start to wonder if there is something — anything! — that can speed this process up. Surely technology can come to the rescue. Sorry. The best we've been able to do is enlist the help of a programmer to write us a script that will crawl a Web site and spit out the URLs it finds. And that merely ensures that we don't miss any pages. Even with this head start, we always go through the pages by hand. A content inventory is a decidedly human task. In fact, we find that the process can often be as valuable as the final spreadsheet. If you invest the time in scouring your Web site and deconstructing every page (or at least a good selection of pages), you will end up as the uncontested expert in how it all goes together. And that's invaluable knowledge to possess when redesigning your site.

Jeff Veen is the Director of Product Design and a founding partner at Adaptive Path. He specializes in innovative Web design techniques. You can learn more about Jeff at his personal site, Veen.com.

Download: content_inventory_template.xls (34K Excel file)

This article is part of an occasional series about techniques for doing user experience work. The previous essay in this series was "Setting Priorities" by Janice Fraser.

Published by **Adaptive Path** | 363 Brannan St. | San Francisco, CA 94107 | 1-415-495-8270 | http://adaptivepath.com/