Michelle Light, University of California, Irvine
EAD @ 10, August 31, 2008

The endangerment of trees

Last year, when I was participating on a committee to redesign the Online Archive of California, many of us wanted to enable browsing finding aids by subject. Even better, we imagined using subject facets to help users explore or limit their search results. We discovered, however, that only about 60% of the finding aids used controlled access tags. If we really wanted this feature, would we have to ask institutions to go back and index their finding aids? In an age where "More Product, Less Process" and the mantra of more, faster, and cheaper has hit archival processing, would this even be tenable? We decided no, and the OAC designers have come up with some other innovative ways to enhance access with the data and level of encoding that is there. However, this got me to thinking…there's a lot of information in OAC, and as expected in a big consortium, this information is not always encoded consistently in ways that allow for much manipulation, other than searching and viewing the document as encoded. What use is all that information in the finding aid? Are we really creating description in order to maximize the accessibility of collection materials to our users? What really is necessary in an online finding aid? How can we encode more intelligently? Is "more" necessarily better, or could we do "more" with "less" if we did "less" strategically? In other words, how can we get more bang for our buck with archival description?

For the sake of answering these questions, I'll posit that we do archival description in order to facilitate user's discovery of materials. (Note that I'm deliberately NOT saying that we do archival description to represent the work we do, or to represent the exact physical state of a collection. We create finding aids to help people find things.) We can measure success in many ways, but I'll focus on 2: whether in response to a user's query, a finding aid is surfaced that describes materials he or she needs; and the finding aid directs the user within 1-4 linear feet of the needed material. According to a 2003 OCLC user study, 89% college and university students use search engines to begin an information search, while only 2% start on a library's website. When I examined the web server statistics for UC Irvine's finding aids, these numbers seem about right. About 90% landed on our finding aids directly from search engines. (Very few stay for more than 30 seconds, but that is a different can of worms.) Whether we like it or not, most potential users will find our finding aids through a back door opened by a search engine.

How can we support this? Search Engine Optimization is the art, craft, and science of driving web traffic to web sites. One of the first tricks of the trade is to look at the keyword density in a page. One of many ways search engines rank pages or try to evaluate the nature of a page's content is to look at how often keywords repeat, or how often the same or similar words are used on a page. Out of curiosity, I wanted to see how a search engine might interpret what our finding aids were about, so I used some free keyword density analyzers. According to these sites, my finding aids were about boxes, folders, materials, undated, circa, correspondence, and individual dates, such as 1993 or 1986. In a few finding aids, "California" made the top 10 as did the last name of an individual, but in not many.

So, the most important concepts in the finding aid were NOT represented by a significant number of keywords for the search engine's index. Why? Probably because one of the most fundamental, international principles of archival description, that description should proceed in a top-down manner from the most general level to the most specific, branching like a tree. At each hierarchical level we record the information pertinent to that level and strive not repeat information from level to level. So, we deliberately do not repeat the keywords that represent the most important content in our collection. While this makes absolute sense from a management perspective or when humans are our primary readers -- why repeat the same information over and over -- does it continue to make sense for delivering finding aids on the web, when search engines evaluate the key content according to the presence of keywords?

Luckily, search engines such as Google also use many other factors to rank pages. Some of you may be thinking, my finding aids regularly appear within the top 10 of Google searches. One reason may be the long-tail keyword. The long tail is a type of statistical distribution where a high-frequency population is followed by a low-frequency population which gradually "tails off". This tail is often long, meaning that the total number of infrequently used keywords outnumbers the total of the top ten keywords. So, most traffic to a site is not the result of users typing in commonly used keywords, but rarely used ones. Search engines often assign a greater value to keywords that don't appear that often in their index, so finding aids with a lot of unusual words or names will rise to the top. In a way, our container lists, which may contain long lists of materials with unique names of people, places, organizations, or topics, help us capture a niche market on the Internet. Interestingly enough, since I've had my processors use DACS to construct titles for series, subseries, and folders, the numbers of long tail, unique keywords in the finding aids have really increased, largely due to the name segments of the titles. But the irony is, in a search engine's eyes, our finding aids are probably better at

surfacing the content of the folders than they are at surfacing the content of the entire collection or its series.

Don't despair yet though. Search engines assign greater value to words in the title of an html page, or in heading elements, or the words towards the top of the page. As expected, search engines also like pages that contribute to the web of interconnected pages and aren't just dead ends. Pages that are linked to by authoritative sources -- and yes, Wikipedia is considered as such by Google and others -- are given higher ranks. Pages that link outwards to these authoritative sources also are preferred.

How should we respond in this search engine environment, with these trends in user behavior and the drive to process collections more efficiently? Realize that you are creating finding aids for the web and that Google the most used index to your site. Efficient description in the online environment must communicate the main content of collection and surface some significant or unusual content lower down. So….

- The most important content must be represented toward the top of the page.
- When summarizing the content of the collection, use meaningful key words that users would likely search for.
- Maybe you should bend some archival rules about repeating key names and concepts.
- And what I emphasize a lot when training new processors is that your finding aids shouldn't contain a lot of unnecessary noise. That is, don't repeat non-unique words that don't add value to the findability or understandability of the finding aid. Not only does it take more work to describe more, you're making your finding aids harder to use.
    - For example don't create long folder lists of materials types, such as a long scrolling list that repeats "correspondence" or "president's reports" 25 times, each with a different date.
    - Say it with less, and only list what is necessary to get someone within 1-4 boxes of what they want.
    - Sometimes folder lists are very useful enterprises and worth it if, WITHOUT arranging items within or across folders, the folder titles will yield some significant long-tail keywords, such as the names of important individuals, organizations, places, or important subjects or topics. In these cases, folder-level description may lead users to you, but that is not always the case. Think about the value of the description in an online environment before investing in the description.

> What will it really take to get users within 1-4 feet of the needed material from a search engine?

Now that we're in the digital environment, I've sometimes quipped that we should do more description, less arrangement. Description is often thought of as a consequence of arrangement, ideally done at the same level as arrangement or preservation. I wonder if description in the online environment will allow us to shift the burden of arrangement, or intellectual sense-making, to users. For example, what if we did our finding aids in such a way that users can arrange materials intellectually based on searching or other means, and then through their sequence of call slips, arrange materials in their own way?

The book *Everything is Miscellaneous* by David Weinberger has made me question the nature of archival description in the digital environment too. Weinberger contrasts systems of organization in the physical and digital worlds. In the physical world, we are limited by the fact that an object can only be in one place at one time. To organize the physical world, humans like to categorize things, or nest information in classification systems, such as the Dewey Decimal System. The shape of this knowledge has assumed the shape of a tree. In the digital environment, we don't have the same limitations as in the physical environment. The best form of order in the digital world, according to Weinberger, is to make everything miscellaneous. Put everything in a big pile, provide a lot of metadata so that a single leaf could possibly co exist on many branches, and allow users to help describe materials through tagging. Machines then will do the work of sorting through the information and will deliver to users an organization that makes sense for their specific needs.

For example, he discusses the BBC broadcasting archive's dilemma in providing access to digitized recordings. Initially, in order to access the programs online, you had to navigate through a structure that organized the files according to which station produced the recording and the station's schedule. But BBC found that people didn't care where the program originated or when it was aired, they just wanted the program. So, BBC tore apart this hierarchical structure. They are making the digital objects "miscellaneous." It is miscellaneous because *how* the content is actually arranged does not determine how users can access that content. It is miscellaneous because users don't need to know the inner organization, and the system gives users the flexibility to order the pieces the way they want. If you have information in miscellaneous piles (with lots of metadata), then it's like having a Dewey Decimal system written to order for each user.

Archival finding aids were perfected for managing information housed in one physical place, for providing objects with an intellectual context and a physical map, of sorts, for their retrieval. EAD is the hierarchical tree structure that Weinberger calls outdated, where components usually have one place, one context. That works for the physical world, but will that way of seeing and organizing the world work with the digital records of our future? Do we really need to perpetuate the tree metaphor in the online environment? Last month I was checking out the new delivery system for entirely digitized collections at the Smithsonian's Archives of American Art. It uses the traditional finding aid very effectively in a digital environment; the finding aid structure endows the digitized objects with intellectual and physical context. But what about born-digital items that don't necessarily have one physical home? What about born-digital records that are collaboratively produced so that there are multiple ways to understand context? Weinberger's idea of a big, messy pile (where objects have lots of metadata to provide context and meaning) may be a more suitable way to understand the digital landscape, where users can access materials through searching and generate their own meaning and views.

I wonder, in the future will our EAD and finding aid structures be generated on demand to represent materials and their contexts, rather than being a static way we manage our holdings? Will our finding aids merely exist in the realm of multiple possibilities, waiting to be assembled by a user query? Will our top-down way of organizing information be the way of the future, or will our description focus more on lower levels, filling out contextual information and relationships to enhance retrieval possibilities and enable our users to put the pieces back together? At this conference from Anne Sauer's research poster about OAI-ORE (object reuse and exchange), I learned that this might not be science fiction. OAI-ORE defines standards that will facilitate the identification and description of aggregations of web resources, the parts of which may have various relationships to each other and together make up a whole. It gives compound web objects a clear boundary and defines their internal and external relationships. It allows you to reference an object and know its context. Anne's diagrams showed how this might work among collections, series, items, creators, and other contextual information.

I know that I will continue to work for some time in a mixed world, where physical and digital live side by side, and I will continue for now to manage them in ways that have always made sense in the physical environment. But I used to think that I could apply the same archival principles to any archival materials, no matter their manifestation in print or digital form. I am increasingly less sure. I'm inclined to believe the grand statement of Weinburger, that the shape of our knowledge is changing. The hierarchical trees that I use every day to make sense

of the archival collections under my care are being called into question. I am excited, and maybe a little apprehensive, to see how we as a profession will respond to these changes.