

Digital Library Research, 1995-2010

Candy Schwartz, Retired

It is a great honour to appear before this august society to present a review of the past 15 years of digital library research. Since my retirement for the Graduate School of Library, Information, Communication and Education Services, Systems, Studies and Science (GSLICESSSS) last year, I have had some time to reflect on our accomplishments, and I must admit that I am proud to have participated in some small way in the advancement of digital libraries. I would like to focus on 4 key areas: identification, metadata, visualization, and relevance.

Identification

In the late 1990s we were grappling with the problems of uniquely identifying each digital object in the universe. URNs, URCs, URIs, URLs, PURLs, SICIs, DOIs, you remember. Following the great reconstruction of the digital world, after Black Saturday, January 1, 2000, we gained the opportunity to assign a unique identifier to each object as it was recreated. Coinciding with the release of Windows 00 and the realization that one operating system for the world's computers was, after all, not unreasonable, each reconstruction kit was loaded with 00, MS-Navigator 6, and Bill 1.0. Bill, as you know, automatically sends an encrypted copy of each digital object as it is placed on any network to the UN's Online Union Catalog of Digital Objects in Columbus Ohio, managed by MS-OCLC. At the OLUC, a unique 26 letter 10 digit MS-OCL number is assigned to the object and is returned to the host machine, along with the metadata record created by the MS-OCLC metadata analysts. The OLUC monitors all the objects, recording new versions when content changes, and retiring the object when it is deleted. While there were, and still occasionally are, attempts to either change or erase the MS-OCLC number, or to disengage Bill, a visit from the MS-OCLC Special Branch has proven to be a strong deterrent, especially if the surgical removal of the offender's spinal cyberjack is invoked.

Metadata

As the great reconstruction was proceeding, the Travelling Metadata Road Show finally came to consensus, leaving its members to return to their institutions, where they were greeted by startled family members and colleagues (or were relocated into special homes if their origins could not be traced), and were slowly reintegrated into normal society. The final document was called the Sheraton Barstool, after the only meeting location any committee member could name with any certainty. This standard incorporated key elements of the Dublin Core, GILS, the FGDC, TEI, IAFA, and the Warwick Framework. In 2004, MS-OCLC hired 5000 metadata analysts, absorbing the entire output of the nation's library and information science programs in that year. Each item in the OLUC is now fully described. Metadata records are stored in the Online Normal Link Information ONLIMARC format, and MS-OCLC makes available, for a small fee, a variety of conversion tools, for example MARC->SGML->HTML->XML->VRML->PDF->GIF->AVI->MOV->WAV->AU.

Visualization

Advances in visualization research, coupled with developments in flat screen wall monitors, has made it possible for even the lowest income families to dedicate at least one wall in each room to a user friendly approach to the global digital world. While it is theoretically possible to summarize the entire OLUC database, most users choose to use USLC approved filters, which restrict the digital universe to only those objects deemed appropriate for the age, gender, and intellectual level of the user. For many users, this results in a much reduced dataset, which is amenable to Schneiderman's classic "overview > zoom > filter > details" approach. For scholarly research in digital collections, the LYBERVIBE visualization model has emerged as the leader after extensive testing with the professional user group. LYBERVIBE uses the

urban wayfaring model, showing the user an overview of information space as a cityscape. The user can zoom in to the relevant building, have a filtering conversation with an agent, and finally enter a familiar environment in which he or she can easily select relevant items. The introduction of the OLUC, coupled with LYBERVIBE, did lead to the demise of most search engine companies, and all of the online vendors, but this community has largely found gainful employment working for MS-OCLC, where they are principally engaged in hand-checking lost and redirected links. Their involvement in an earlier project will be described shortly

Relevance

I believe that this is where the greatest strides have been taken, and especially following the Great Reconstruction. One of the barriers to truly meaningful relevance research in large digital collections was the problem of obtaining realistic recall measures. This problem was resolved in 2005, when the aforementioned search engine and vendor staff persons assisted in the construction of complete relevance judgement sets for 200 queries across the entire OLUC database. The audience is reminded that MS-OCLC would be most grateful for contributions to the Fund for Relevance Veterans, which so far has funded three care providing centers where these valiant soldiers can pass the rest of their days in fibrefree environments.

In any event, once the basic relevance set had been created, we were able to test some of our basic assumptions about what makes an item relevant with respect to a particular query. I would remind you all of the seminal work undertaken prior to this time period, and presented at SIG/CON sessions, by luminaries such as Brett Butler (1980), Ev Brenner (1983 and an update in 1986), your own speaker in 1987 and 1989, the distinguished Dr. Janes in 1993, and both Dr. Janes and myself in 1994. Many of these papers are in print in the Journal of ASIS, vol. 16, number 0, October 1995. The fact that this worthy collection has not been digitized is a consequence of its fragility, and not of its lack of worthiness.

The concepts I am about to present are, of course, familiar to all of us, but for the newer information scientists in the audience, I would remind you that for those of us who have been in the field for half a century, these were very heady times.

Since the OLUC does maintain the international digital database in the same configuration as it appears on the global network, one of the first tasks was to derive a metric for the value of a site. This is, of course, the sum of the values of the individual objects in it, divided by the total number of objects minus 1 (this last element is to render a one page site completely valueless). Which begs the question, how is the value of an individual object assessed. Obviously different formulae have been identified for different formats, so let me just share the elements used to derive the value of a page in HTML (the actual formula is complex and proprietary).

?? d, depth

Length of path name in characters divided by number of slashes. A lower value is better.

?? l, linkout

of external links divided by number of internal links. A higher value is better.

?? s, syllabicity

of syllables in last name of creator. A higher value is better.

?? j, juvenility

Age of page in days divided by age of creator in years. A lower value is better.

?? di, distraction

Total number of , , and <BLINK> divided by number of words. A lower value is better.

?? t, timesink

Total square pixels of image divided by # of images. A lower value is better.

?? tc1, titular colonicity 1

Presence of a colon in dc.title. A 1 is better than a 0.

?? tc2, titular colonicity 2

Position of colon in dc.title. A lower value is better.

With this breakthrough, it has been possible to assign an absolute inherent value to each item in the international digital database, and therefore to every digital collection on the global network.

And finally, with respect to an individual query, the absolute value of the site is multiplied by the age of the user. This is popularly known as the Baby Boom Failsafe, and ensures that truly valuable information will only be found by those who are mature enough to appreciate it and who also have good credit ratings.

Concluding Remarks

In the time allowed, I have only been able to touch on several of what I consider to be the most important advances in SIG/CON research in the past decade and a half. I have had to ignore important work in user behaviour, especially among the higher primates, Dr. Janes area of expertise, and in bio-communications and its allied field techno-apparel, but I am sure that there will be other opportunities.