Digitisation, domain harvesting and discussion

Kerry Webb

kerry.webb@alianet.alia.org.au

More Google digitisation

There's no doubt that Google is serious about getting into the digital library space. They announced in November that they are contributing a significant amount to the Library of Congress' World Digital Library (WDL) project. LC is currently looking to develop a plan for the project, which will be concentrating on digitising unique items, such as manuscripts. Unlike other projects where Google has an involvement, the Library of Congress will get special permission to include works in the World Digital Library that aren't in the public domain. Other projects that Google and LC have previously worked on include digitising about 5000 public-domain books, and the scanning of works considered of historical value from the Library of Congress' Law Library.

And, not to be left behind

Microsoft, in association with the Internet Archive, announced an agreement with the British Library to scan 25 million pages from the library's collection to be made available on the MSN Book Search site later in 2006. The BL will be contributing around 100 000 books from its collection, all of which are no longer under copyright restrictions. Microsoft won't be paying the library for access, but the BL will benefit, as it increases the extent of its digital collection.

Out of the box

The National Library of Australia completed the first whole Australian domain harvest during June and July 2005. The Internet Archive (http://www.archive.org) undertook the crawl was on behalf of the Library. Around 185 million unique documents were sourced from 811 000 hosts, consisting of about 6.69 terabytes of content. The report of the project, by Paul Koerbin, has now been made publicly available at http://pandora.nla.gov.au/documents/domain_harvest_report_public.pdf.

A busy little bee

I mentioned a couple of months ago that I was reading John Battelle's *The Search*, and a fine book it is. He writes engagingly about the history of search facilities like Excite, Yahoo, AltaVista and of course Google. He also has a lot of references to useful resources like ResearchBuzz. This is a site written by Tara Calishain, covering everything about Internet research, such as search engines, new data management software, browser technology, general information, web directories, and so on. The site design is minimalist and it has a lot of good stuff. See it at http://www.researchbuzz.com/.

Where phrases come from

There's a neat little site that's devoted to phrase meanings and origins, at http://www.phrases.org.uk. It not only has a comprehensive listing of phrases, but there's a very good discussion forum where you can ask questions, challenge received wisdom or just engage in chewing the fat (and they're not sure where that one originated). The site also features a couple of quizzes on phrase origins and Shakespearean quotations.

No humbug at all

International consultancy firm Booz Allen Hamilton was commissioned by the UK Cabinet office to study world's best practice in e-Government, and have produced the report at http://extfile.bah.com/livelink/livelink/151607/?func=doc.Fetch&nodeid=151607. It's remarkably good. The study, *Beyond e-Government*, looked at several countries (including Australia) and analysed where each one was in the various stages of development, and what lessons could be learned from their experiences. One observation was that users stayed away if their initial experiences didn't match the hype, but in many cases the hype was where the expenditure had been, rather than in the back-office procedures that needed improvement to get rid of bottlenecks.

Meanwhile, somewhere in Germany

I attended a presentation in November that I thought would be better than it turned out to be. Professor Dr Uwe Kamenz from the German ProfNet Institute for Internet Marketing spoke on a study that he had undertaken in 2004 on 1750 government websites in 204 countries. Using a complicated set of criteria, he had ranked websites and countries, and presented the results to a group of people in Canberra with an interest in such matters. I don't think that we were being overly defensive, but we had trouble understanding why Australia was ranked so low. Some of this had to do with his choice of criteria and the way that they were weighted. For instance, if you had bios and pictures, your site scored well. Which may be one reason why the Governor-General's site was higher up the rankings that that of the Department of Health and Ageing. He also brought certain cultural biases to the study, suggesting that we need to put more pictures on our home pages. He wasn't completely convinced when it was pointed out

Google is serious about getting into digital library space

that Australia's rural population probably suffered more from bandwidth limitations than people in Germany. It's always useful to get feedback and see how we rank against other countries, but you have to be careful with the criteria you choose.

The last piece of the jigsaw

A few years ago, the National Archives started work on the 'Documenting a Democracy' site, as part of the Centenary of Federation celebrations. In doing this, they gathered the documents (in real or digital form) that contributed to the foundation of our nation. Documents of relevance to all parts of the country were sought in libraries, repositories, archives, government registries and anywhere else where they may have been hiding. All were digitised and published on the site at http://www.foundingdocs.gov.au. Now, with the ACT's contribution finally in place, the picture is complete. And a pretty picture it is.

The dangers of unverified information

Wikipedia is in the news again, and for the wrong reasons. A former administrative assistant to Robert F Kennedy was said to have been linked to the assassinations of both of the Kennedy brothers. The claims were made in a bogus biography submitted (as was common until now) by an anonymous contributor. Following a protest by the injured party, Wikipedia now requires new submissions to be made by a registered user, but this isn't much of a deterrent, as users do not have to provide any identifying information. The most unfortunate aspect of the whole story is that the incorrect information has been picked up by two other sites, reference.com and answers.com. It's now been fixed — sort of. See http://en.wikipedia.org/wiki/John_Seigenthaler_Sr.

Enough, already

And if you haven't been turned off completely by the diatribes against Wikipedia, Matthew White's Wikiwatch at http://users.erols.com/mwhite28/wikiwoo.htm gives you one man's view on this publishing phenomenon.

Shout out loud and be proud

And just to show that there are some Wiki things that are working well, have a look at 'Library Success: a Best Practices Wiki' at http://www.libsuccess.org/. It's a one-stop shop for great ideas for librarians from all over the world. It seems there's a fair bit of activity on the site, with several additions and updates to entries each day. A quick scan through the various categories reveals a number of posts about projects in Australia and New Zealand, as well as other countries of course.

What Phil wants

Phil Bradley is an internet consultant, based in the UK. He has a very useful pages that addresses many general topics about the Net — choosing search engines, sites where you can publish photos, recovering from a crashed hard drive and so on. One of the unusual parts of the site is the 'I want to...' section, which lists a series of resources that you can use to share research with colleagues, set up bookmarks that can be used anywhere, do various things with web pages, and many others. It's at http://www.philb.com/iwantto.htm.

A free toolkit

James Robertson of Step Two Designs has announced the release of their 'Intranet Review Toolkit', a free set of guidelines and heuristics for conducting an intranet expert review. The toolkit provides intranet managers and designers with a simple method to assess the strengths and weaknesses of their services. It allows them to conduct a detailed intranet review, focusing on a wide range of functionality, design and strategy. It's released under a Creative Commons license, and is free to download and use. It's at http://www.steptwo.com.au/products/irtoolkit/index.html.

For better or verse

No matter what you think of poetry (and even if you never think of poetry at all), there's something special about hearing a poet reading his own work. On the other hand, I suspect that Spike Milligan and Edith Evans probably gave better renditions of his work than William McGonagall ever could; but it would have been good to hear him try. Anyway, the Poetry Archive at http://www.poetryarchive.org tries to make sure that recordings by poets are provided whenever possible. Make up your own mind.

Well, that's their aim ...

I've written before about WebAIM, the facility that provides — among other things — a discussion list for technical accessibility matters. Starting early in 2006, they're publishing *WebAIM Monthly* at http://webaimmonthly.org, featuring cross-disciplinary articles about the techniques and theory of web accessibility. Go to the site for an overview of upcoming themes and submission guidelines.

Let there be light

There's been plenty of discussion over the years about the 'invisible web', that huge collection of web content that exists somewhere, but is not accessible to search engines, or indeed to most casual users. The term usually refers to material buried in databases that can't be accessed except through specific searches. In an article at http://www.searchengineposition.com/info/Articles/%20InvisibleWebStillExists.asp Rob Sullivan refers to another definition, alluding to the poor choice made by many web managers and designers in choosing techniques that are not friendly to search engines. He points out a few simple ways of ensuring that all of the content on your site can be opened up. ■

Poor design creates an invisible website: make sure your site can be opened up