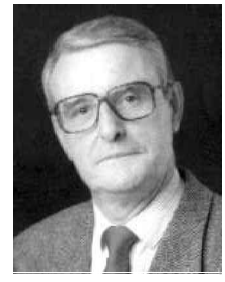
Key issue

The Invisible/Deep Web



BARRY MAHON
Executive Director
ICSTI, Paris
and Associate
Consultant, TFPL,
London

"I love the fact that students now have access to the riches of online content… The question is whether their online searching skills are limited to finding MP3 files and pictures of Britney Spears…."

Mary Ellen Bates, *Econtent*, June 2002

The Deep Web is that part of the net which is generated by web sites 'on the fly', i.e. that cannot be found by direct searching the so-called Surface Web. Typically, Deep Web activity is generated when you ask a question on a site, for example when searching a knowledge base on a support site, e.g. ESRI (http://support.esri.com) or on one of the many data sets sites, e.g. The Protein Data Bank (http://www.rcsb.org/pdb).

It is claimed that most search engines, such as Google, only index from one third to one half of the public documents. CompletePlanet (http://www.completeplanet.com) estimates the Surface Web at 1-2 billion documents, and the Deep Web at 550 billion documents. However you try to measure it, the deep material that is not accessible through these search engines represents a huge gap. As search engine robots can be blocked from a web site or only allowed to index a selected level of pages on the site, some data sites may not allow 'Deep Indexing'.

This gap has led to a veritable explosion in activity designed to mine this area of the net. Sites such as Deep Web Search Tools at: http://www.bhsu.edu/education/edfaculty/lturner/Deep%20Web%20Search%20Tools.htm have listings for places where you can dip into the Deep Web and associated places such as Search Engine Directory sites. Deep Web search products do not cover the many subscription-based services like most A&I databases or the growing volumes of articles in e-journals, but that is another story.

The general impression one gets from perusing Deep Web search tools is that the coverage is variable. Many of them sub-divide the world by country, which is as good a way as any, but it may not be the best way for the search you have in mind. Others, such as the Open Directory project (http://dmoz.org/about.html), attempt to classify sites using a conventional classification scheme like UDC or Dewey:

> 'The Open Directory provides the means for the Internet to organize itself. As the Internet grows, so do the number of net-citizens. These citizens can each organize a small portion of the web and present it back to the rest of the population, culling out the bad and useless and keeping only the best content.
>
> The Open Directory follows in the footsteps of some of the most important editor/contributor projects of the 20th century. Just as the *Oxford English Dictionary* became the definitive word on words through the efforts of a volunteers, the Open Directory follows in its footsteps to become the definitive catalog of the web.'

There is also significant Deep Web activity surrounding business use of intranets. There are companies like Bright Planet (see http://www.brightplanet.com/technology/deepweb.asp for the White Paper on the Deep Web) who market Lexibot, a search tool with capabilities to search multiple sites in one pass. A similar capability is offered by metasearch engines. See: http://www.metasearchguide.com, which offers a guide to the capabilities of these tools. The opening page offers 14 options!

In the field I know best, Scientific and Technical Information (STI) is going through an interesting change. Serials, which were the backbone of

scientific communication, are rapidly changing from subscription-based, print on paper distributed, rather formally structured artefacts to interactive, electronic-only productions, in many cases free at the point of use (e.g. BioMed Central at www.biomedcentral.com). This change is driven in part by a movement away from the formal schemas – write/referee/publish, towards – write/load on a server/announce its availability/receive reactions/rewrite, etc. As a piece of the Invisible Web this presents challenges, not least in identifying the material, and then qualifying it as a 'work in progress', which in many cases it is. In addition, there is the risk that when the research project finishes, the server gets switched off, so where is the material archived? Life was easy when all you did was send the journal set to the binders…The other big challenge is that users (i.e. research authors) are now realising that they can augment their 'papers' with sound, video, and interactive programs so that 'readers' can check the data for themselves if they wish.

Interestingly, the major science publishers are realising that even their own sites do not necessarily provide the simplest access to all their material. Recently, Elsevier have started to offer help and advice to their users via a site: http://www.webresearch.sciencedirect.com/

The site says:

'Elsevier has prepared the 'Web Research Guide' to help students, faculty members, authors and research scientists find the information they need on the web. The guide covers a broad range of topics, including expert tips on how to:

use search engines effectively

focus your research on quality STM information only

find hidden scientific information online

locate peer-reviewed, subject-specific directories

set up subject-specific alerts that automatically e-mail you the latest news.'

The challenge of the Deep/Invisible Web for information professionals is to make it less deep and more visible. In fact, the challenge of the web in general is to make useful, valid and available information that the user actually needs and will benefit from.

The Deep Web does present issues, but the role of information professionals has always been to provide the access. What we need to do in the case of the Deep Web is identify what is where, point the user towards it and adapt ourselves to the new forms and formats.