



# Information Overload

Issue 47, July 2006

---

## Welcome:

Welcome to the July edition of Information Overload, this month we take a look at the issues surrounding archiving of web pages. If according to recent definitions a record is "any record of information however recorded and includes –

- (a) any thing on which there is writing or Braille;
- (b) a map, plan, diagram or graph;
- (c) a drawing, pictorial or graphic work, or photograph;
- (d) any thing on which there are figures, marks, perforations, or symbols, having a meaning for persons qualified to interpret them;
- (e) any thing from which images, sounds or writings can be reproduced with or without the aid of anything else; and
- (f) any thing on which information has been stored or recorded, either mechanically, magnetically, or electronically." *West Australian State Records Act 2000. State Records Principles and Standards 2002 as published by the Government Gazette, Tuesday 5<sup>h</sup> March 2003, No 38.*

If that is the case then it stands to reason that web pages should also be included in the records management mix and be subject to the same laws and considerations as all other electronic information, whether born digital or not especially if they are "in pursuance of legal obligations or in the transactions of business." *AS/ISO 15489*. The question is – is anyone managing their web resources as "records" yet?

We would like to thank you in advance for forwarding this edition onto friends, colleagues and other interested readers. Please note that all back issues of this edition, as well as our registrant resources edition can be read and/or downloaded from our web site – <http://www.iea.com.au> should any of the topics be of interest and use. The July edition of the Registrant Resources edition looks at the benefits (or not) of life in the contract arena. As we say – "with the changes to the Industrial Relations laws – does anyone have a "permanent" job anymore?".

Lorraine Bradshaw  
Marketing Coordinator and Projects Officer

---

© IEA 2006. All rights reserved. You are free to use material from the Information Overload newsletter in whole or in part, as long as you include acknowledgement of source.

**Information Enterprises Australia Pty Ltd**  
Unit 4, Upper Level, 201 High Street, FREMANTLE WA 6160  
Tel: 08 9335 2533 Fax: 08 9335 2544 e-mail: [training@iea.com.au](mailto:training@iea.com.au)

## In this Issue we will be looking at:

- Version control and the web
- How old is your web site? And does it matter?
- Archiving strategies
- Does web harvesting breach copyright?
- Book review
- A Thought to ponder.

## Version control and the web

Version control is not something that is usually discussed with relation to an organisations web site. Yet each day changes can be made to the structure and content of the pages that make up your online presence. New courses are added to the calendar, new jobs added to the employment sections, changes to product pricing are made, pages are “tweaked” to ensure a better ranking on the major search engines and so on. The question is – do you keep a copy of each change that is made so that you can prove evidence to “how things were?” The answer is – probably not. The question is why not?

Well there are a couple of reasons I can think of –

- Most organisations appear to be struggling with capturing all the other electronic records it creates or receives each day; they don’t have time or resources to worry about the web site yet.
- The “web” is managed by the IT department or “web manager” neither of whom may understand the importance of record keeping.
- Some web creation packages don’t appear to have an archive function let alone a version control function – but you don’t have time to check.

And then there is the minor problem associated with security.

The beauty and therefore the main problem with “the Internet” is that anyone with a little bit of knowledge can create a web site (and probably has). Unfortunately some web sites lack basic protection allowing sensitive information to be found. And even with some security measures in place, people with some time on their hands, not to mention access to password cracking software can get into a considerable number of web sites and can become an unauthorised editor of your organisations online face.

The problem as you may know – is not the surfer who manages to stumble onto the “good stuff” (search google for the term “google hacking” and you will find information on how to get into parent directories, locate password files and even credit card information) but those people who deliberately set out to appropriate information for their own needs – industrial espionage notwithstanding, there are some people who delight in their ability to change web pages and delete information at will.

Do you have the correct measures in place to ensure this can’t happen to you? Did you know that when most (if not all servers) are delivered they have default passwords installed so that the person setting up the system can log on and add users etc? Unfortunately a lot of these default passwords are not changed so systems are wide open to attack from someone with a pre-determined list of accounts and passwords. The most widespread case documented, involved military websites across numerous countries – so it they don’t batten

---

© IEA 2006. All rights reserved. You are free to use material from the Information Overload newsletter in whole or in part, as long as you include acknowledgement of source.

down the hatches – what chance the rest of us? *Stoll, Clifford: The Cuckoo's Egg: Tracking a spy through the maze of computer espionage.*

## How “old” is your website? And does it matter?

It has been said that the “average lifespan for Web sites is just 44 days, according to James Billington, Librarian of Congress.” *Coughlin, Kevin; Archivists say computers have no sense of history. NJ.com Thursday 19 June 2003. <http://www.nj.com>*

I can hear the “so what” from here.

Well going back to the version control issue just for a moment – if you remove pages from your web site, or change the information contained on those pages, and you don't maintain a copy of the changes – how will this affect your ability to produce a “document” in its entirety should you need to, especially in the face of litigation.

Take for example a document that you wrote that contains links to your own web site – and documents contained within the web site. If the web site is changed (for whatever reason), or documents have been removed you may not be able to access or find the material it referred to.

If you take those documents that contain embedded objects and hypertext links and remove them, can you re-create the “experience” for the end user in the same way that you viewed the document /item/ page if it has been isolated from the originating software? For example, a PDF is a snapshot of how a document “looked” at a particular point in time. Like its other electronic counterparts, a PDF document may also be overwritten or deleted (accidentally or otherwise). Even with security measures in place, the person who attaches the security to the record can remove it. Throwing into doubt the records reliability, authenticity and accuracy. And so it is with archiving web pages, or not as the case may be.

Whilst it is relatively easy to PDF a document and store it someplace “safe” can the same be true of web pages and in particular entire web sites. The answer is – no, not yet.

Static web sites can be printed out and placed on file, which can solve some of the problems associated with how the site looked at a particular time – (and yes it does seem rather like a backwards step doesn't it) it can hardly be deemed to be the best solution. Pages can also be copied into another electronic medium eg., word and then “archived” in whatever way an organisation currently archives its material. Recent discussions on the various list serves have also talked about taking snapshots of the page, wrapping the associated metadata around it and logging it into the EDRMS. However, this predisposes that the person making the changes remembers to take a copy of the before and after.

But what about those dynamic web sites? Dynamic web sites often draw their content from a database (or databases) or content management systems. Pages are constructed only after a question is posed. Does the answer to this problem lie partly with ensuring the changes to the content management system or database are captured by the software that operates and manages it? And is there a technological trigger that can be used to ensure that the odd spelling correction made to the web site is also archived. Spelling mistakes? Yes I am very serious – can you imagine going to a web site that contained incorrect spelling of drug names? Your doctor takes the information on face value and prescribes the medication – and you get sick... Perhaps a little bit of an extreme example perhaps, but is it any less real than the issues surrounding the deletion of emails we have seen in cases such as British Alcohol and Tobacco. The answer is no.

---

© IEA 2006. All rights reserved. You are free to use material from the Information Overload newsletter in whole or in part, as long as you include acknowledgement of source.

## Archiving strategies

The National Library of Australia has looked at the problems of providing long-term access to electronic information via the Internet. Entitled PANDORA (Preserving and Accessing Networked Documentary Resources of Australia), the National Library of Australia selectively chooses web pages to archive, based on a pre-determined selection criteria (about Australian interests or by Australian authors) and "harvesting" schedule. The National Library acknowledges that there are disadvantages with the selective approach., as they are making subjective judgments about the value of resources and what researchers of the future may find useful. For example they may only choose to archive a small percentage of any particular web site, or as was the case of the web site belonging to the Sydney Olympics – the web site was harvested on a daily basis. It appears that the National Library of Australia has to make its decisions on a case-by-case basis. Needless to say things may fall through the cracks.

In building its print collections, the National Library relies on the [legal deposit](#) provisions of the [Copyright Act 1968](#), which requires publishers to deposit one copy of each edition of a work with the Library. These provisions do not yet cover electronic publications. Most of the States also have legal deposit legislation, some of which include provisions for physical format electronic publications, such as CD-ROM and DVD. Of the deposit libraries contributing to PANDORA, only the Northern Territory Library has legislation which specifically includes online publications. In the absence of legal deposit provisions for online publications and web sites, most PANDORA participants are obliged to seek the permission of publishers before copying a title into the Archive. <http://pandora.nla.gov.au/overview.html>

Whilst PANDORA covers the collection of websites with Australian interests, it should be noted that it does not appear to take into consideration the recordkeeping requirements of the organisation, and in particular the Retention and Disposal of material from the organisation. What happens when a "document" / web page or pages need to be removed from a site and "disposed of" – does PANDORA make provisions for the destruction of material from its archive as well? And if it doesn't – does this make a mockery of the various State Records Acts and other legal provisions for the retention and disposal of records?

Whilst we are on the subject of collectors or harvesters of web sites and pages, we also need to consider the biggest collector of all – The Internet Archive.

The Internet Archive <http://www.archive.org/index.php> has been collecting material for 10 years. Founded in 1996, the archive attempts to capture as much of the "web" as possible on a regular basis (currently listed as 55 billion web pages) in a bid to ensure we don't succumb to the digital dark ages. Unfortunately it relies on robot technology and has no way of knowing whether it has captured everything that is available (and doesn't).

It is also interesting to note that it actually collects web pages without asking permission first, so any site that says "don't index me" will not be collected. Nor can it yet collect material in parts of the "deep web". As with most (if not all the search engine crawlers), any site that is stuck behind password controlled front ends, has links that dead end, or has material contained within databases will not be archived by the Internet Archive. For instance our own web site cannot be archived beyond the first page because of the way the first page was written – you have to hit a big button in the centre of the screen to gain access (and the site is currently frames based), needless to say the robots and spiders have bypassed us. Whilst the site – <http://www.iea.com.au> is listed on the Internet Archive – all you see is the first page!!

Preserving Access to Digital Information (PADI) <http://www.nla.gov.au/padi/topics/92.html> from the National Library of Australia has an overview of some of the major archiving initiatives across the world. As with all things electronic there doesn't seem to be a one size fits all strategy.

## Does web harvesting breach copyright?

Whilst PANDORA asks for permission to harvest material, the Internet Archive does not. And feels that it has gotten around the issue of copyright by not indexing those pages that say "do not index me". However, there is an interesting twist to this particular tale.

It has been determined that "As a group of files, which form a unity, (web sites) belong among author works and can therefore be included into the Copyright law. The Copyright Licencing Agency in the Great Britain states that "The World Wide Web is subject to copyright, and Web pages are themselves literary works" (<http://www.cla.co.uk/copyrightvillage/internet.html>). Any collecting or archiving of these materials, without the permission of the author or copyright holder is against the law. Exceptions are individual data, government publications which belong into the public domain and other publication in which it is explicitly stated that their reproduction is allowed. Any further use of such publications requires the citing of the source. The rights of authors are limited to their life plus 70 years after the death of the author." <http://www.ifla.org/IV/ifla68/papers/116-163e.pdf>

So of the 55 billion pages collected on the web site – it has to be said that because copies have been made – the Internet Archive has well and truly breached the copyright laws.

OK, I can hear a collective "who cares" after all they have succeeded in capturing more of the Web than any other organisation or group of organisations across the world. And it is not like they are using the information for nefarious purposes.

And I have to agree – but it is also interesting to note that the Internet Archive is not the only site that uses search engine technology to capture information. Every search engine breaks this rule. In fact every search engine and almost every computer takes snap shots of web pages / sites for ease of retrieval at a later date. Check out how long the Web History says on your computer and you will begin to realise how impossible it would be to police this particular problem.

As you can imagine this is only a very small introduction to the problems and the issues surrounding the archiving of web pages. But we hope as always it has given you some food for thought, and comments will be made available in later editions.

## Book Review

The following was listed on the UK Records Management List Serv on 19<sup>th</sup> July 2006. IEA have yet to view a copy of the book and cannot endorse the contents. But for those of you who are interested in obtaining a copy....

***Archiving Websites: A practical guide for information management professionals by Adrian Brown*** has now been published by Facet Publishing and is available from - (<http://www.nationalarchives.gov.uk/bookshop/>), or from other retailers.

---

© IEA 2006. All rights reserved. You are free to use material from the Information Overload newsletter in whole or in part, as long as you include acknowledgement of source.

"This is the first book to offer practical guidance to information-management professionals seeking to implement web archiving programmes of their own. It is essential reading for those who need to collect and preserve specific elements of the web - from national domains or individual subject areas to an organization's own website.

Drawing on the author's experience of managing The National Archives' web-archiving programme, together with lessons learned from other international initiatives, this book provides a comprehensive overview of current best practice, together with practical guidance for anyone seeking to establish a web-archiving programme. It assumes only a basic understanding of IT and web technologies, although it also offers much for more technically oriented readers.

Contents include the development of web archiving; selection; collection methods; quality assurance and cataloguing; preservation; delivery to users; legal issues; managing a web-archiving programme and future trends.

Written to address audiences from the whole spectrum of information-management sectors, this book is essential reading for three types of reader: policy-makers, who need to make decisions about establishing or developing an institutional web archiving programme; information-management professionals, who may be required to implement a web-archiving programme; and website owners and webmasters, who may be required to facilitate archiving of their own websites."

We hope you have a great week.

---

### **A Thought to Ponder:**

"It is the greatest nuisance that knowledge can only be acquired by hard work"

**W. Somerset Maugham (1874 – 1965)**

English Writer

---

Your comments and suggestions on the subject of this newsletter are most welcome. Or if you would like to see other issues covered in future editions, please email me at [training@iea.com.au](mailto:training@iea.com.au). Please feel free to pass on this newsletter to your colleagues' friends and associates. To subscribe they should send an e-mail to [training@iea.com.au](mailto:training@iea.com.au) with "subscribe newsletter" in the subject line.

If you have any suggestions as to what should be included in future editions, then please send an email to [training@iea.com.au](mailto:training@iea.com.au). If you would prefer not to receive this newsletter, please send an email to [training@iea.com.au](mailto:training@iea.com.au) with "unsubscribe newsletter" in the subject line.

---

© IEA 2006. All rights reserved. You are free to use material from the Information Overload newsletter in whole or in part, as long as you include acknowledgement of source.

**Information Enterprises Australia Pty Ltd**  
Unit 4, Upper Level, 201 High Street, FREMANTLE WA 6160  
Tel: 08 9335 2533 Fax: 08 9335 2544 e-mail: [training@iea.com.au](mailto:training@iea.com.au)