Fading away: The problem of digital sustainability

Danny Kingsley from the Australian Partnership for Sustainable Repositories explores the challenges of keeping our memories alive in a digital age.

The recent past is littered with unsuccessful attempts to store information for the longer term. The problem is technology changes incredibly quickly. Some readers will remember microfiche – small sheets of film viewed through a large light box – at libraries. How many computers have a slot for 3.5 inch floppy disks any more? And forget about 5.25 inch disks. Regardless of the format you are currently storing your precious digital baby photos on, they are likely to be unreadable in 10 years time.

But while your digital baby album may be a personal priority, the issue of digital sustainability is a massive problem on a worldwide scale. Consider the necessity of maintaining crime scene evidence (increasingly stored digitally as images or instrument readings) for a statutory period of time, or the far more serious problem of long-term records of nuclear waste. On a more mundane level, government records, articles in academic online-only journals and the data that was gathered in the name of research all need to be looked after for an extended period of time.

There is a new emphasis on the reuse of data, with the perception that the initial government investment in research should not be 'single use'. Potentially the same data can be used by other researchers for different research later down the track or to check the accuracy of analysis. The changing focus of research means that while information today is usually collected in digital format, previous data collected in analogue form needs to be digitised for preservation or analysis.

One group, based at ANU, is taking the issue of digital sustainability very seriously. The Australian Partnership for Sustainable Repositories (APSR) is working towards a future where all major public institutions have a digital repository – not only as their back-up plan, but as a way of enabling research. APSR was established in 2002 with a brief to share software tools, expertise and planning strategies; participate in international standards; maintain a technology watching brief; and support national teaching and research with technical advisory services. A partnership between the University of Sydney, University of Queensland, ANU and the National Library, APSR has developed several practical open source software tools to assist the use and running of repositories.

The data problem

A UK report released last year showed that less than 20 per cent of UK organisations surveyed have a strategy in place to deal with the loss of or degradation to their digital resources.

Of the large amount of data collected in academic contexts, much of it is unavailable after it is used for analysis. This is partly because it may be sitting in a box under a researcher's desk or because it is in a format that only an old computer can read. Indeed, as Margaret Henty and Kevin Bradley discovered when interviewing academics for a survey on data habits, some researchers keep old computers in their offices so they can read old disks that are storing data from previous research.

But this does not have to be the case. There are data archives which look after data in a sustainable way. The Australian Social Sciences Data Archive (ASSDA), based at ANU, was set up in 1981 with a brief to collect and preserve computer-readable data relating to social, political and economic affairs and to make the data available for further analysis. Great care was put into the Archive when it was established, explains Margaret Henty, the National Services Program Coordinator at APSR. "The use of internationally recognised standards has meant that data can be exchanged and sustained with the minimum of intervention, even though the hardware and operating systems have changed over time," she says. "This emphasis on standards has been maintained, and the ASSDA has been an important contributor to international work in this field."

Software - so last millennium

Software incompatibility is a serious problem for long-term data storage. Some software programs are simply no longer used, such as the first word processing program - WordStar (with

the exception of several hundred members of the WordStar Users Group Mailing List). Even earlier versions of current software are sometimes unreadable. It all comes down to backward compatibility – the ability of newer versions of programs to convert older files into the new format.

The 2007 versions of Microsoft Word, Excel and PowerPoint have been designed so they will not automatically read earlier versions of the same programs. That is, they are not backwardly compatible. People who are buying the new versions of the software will need to also obtain a 'compatibility pack' to allow the migration of these files across.

This is a problem for digital repositories. APSR is developing a program called Automated Obsolescence Notification System (AONS), which scans the repository and gives an alert to the repository manager if there are items that have become, or are about to be, obsolete. "The scanning runs overnight and is configured by the repository manager. Scans can occur daily, weekly or monthly," explains Peter Raftos, Project Manager in Scholarly Technology Services at ANU. "We won't know until the program is operational how often things will become obsolete. The terms 'obsolete' and 'obsolescent' depend as much on context as format."

Considering those home albums once again, the sustainability problem goes further than having hardware able to read your storage medium. What about the format the images have been saved as? Images are commonly saved as JPEG files, but this is simply a standard method of compression of images created by the Joint Photographic Experts Group committee - hence the name. (The video equivalent MPEG was created by the Moving Picture Experts Group.) There is no guarantee these standards will remain in the future as digital image requirements change.

404 error - document not found

The internet began in the 1980s and the World Wide Web software was released with the first web server in 1991. The World Wide Web Consortium was created in September 1994. The internet was originally designed to withstand disaster by having multiple nodes so if one node was wiped out, the information would be sitting elsewhere. But the internet provides its own challenges.

The web in itself is not a way to archive anything. For example search engines regularly rewrite the past by updating their indices, overwriting web pages with new ones. The 'Wayback Machine' (www.archive.org) is one attempt to create a record of what has passed, by taking snapshots of the web at given points in time, and unlike web pages these are properly stored in the Internet Archive in California.

Academic publishing is increasingly moving onto the web. The issue of URL permanence gains importance as more articles include online citations. The phenomenon of URLs disappearing over time has been described as 'link rot'. A recent study showed that over a four year period more than 37 per cent of online citations of top refereed communication journals had disappeared.

An attempt to address the problem of link rot is to use a globally unique persistent identifier. "These are in theory everlasting," explains Scott Yeadon, a DSpace Committer based at ANU, which means he's one of the few people in the world who can access the code behind the research-focused repository system. "As long as the identifier resolution service and object repository keeps running it's not a problem." There are several persistent identifier programs, with the Handle System being the most well known. Handles provide a single URL based at a separate server which points to a document. The fairly well-known Digital Object Identifier (DOI) System is a subset of the Handle System with a cost attached. "These systems are arguably better than using an arbitrary URL," explains Yeadon. "They are external and the URL is meaningless, which avoids any problems when the meaning of the object changes."

Another approach is by a group that started out of Stanford University called LOCKSS (Lots of Copies Keep Stuff Safe) and involves creating a distributed, self-repairing, robust, digital preservation system. LOCKSS uses a process called format migration which converts material to a newer format that the browsers do understand.

Show me the money

Organisations and governments worldwide are grappling with this problem and working towards solutions is proving costly. On an institutional scale there is the general cost of setting up repositories. Even repositories based on open source software have set-up costs, and all

repositories have costs associated with their running and maintenance.

The Federal Government has recognised the need to address sustainability issues and has started investing considerable sums into the area. The Systemic Infrastructure Initiative was announced in 2001, funding several projects including APSR, Australian Research Repositories Online to the World (ARROW), the Digital Theses Project, the Meta Access Management System (MAMS) to the tune of tens of millions of dollars. Future funding for sustainability will come out of the National Collaborative Research Infrastructure Strategy (NCRIS).

But there is only so much money in the bucket and this is causing discomfort in research circles. “Researchers are concerned that funds will be taken away from their research to cover sustainability,” explains Henty.

Solutions?

There are no easy solutions to the problem of large scale digital sustainability, but there are things that you can do to look after your own digital information. Run regular back-ups and make sure that when you upgrade your computer you also update all of your files. Put versions of your work into your own institutional repository. And those baby photos? Print them out and put them in an album.