1. I attended the Imaging Science & Technology (IS&T) Archiving 2005 conference at the Washington Hilton. This is my report on the conference.

2. Washington is quite a long way away – home to hotel was about 20 hours with hotel to home about 18 hours. This needs to be borne in mind when planning travel to such a conference and return to work - the body needs time to recover.

3. The conference itself started on Tuesday, 26 April with a number of tutorials. I attended the Long-Term Archiving of Digital Images tutorial – see attached summary. The conference proper ran from Wednesday 27 April – Friday 29 April, kicking off at 0830 each morning (and finishing at 1700 on Wednesday and Thursday and 1500 on Friday). Wednesday featured a 40-minute keynote address and 15 20-minute sessions; Thursday featured a 40-minute keynote address, 10 20-minute sessions and approximately 20 90-second poster previews followed by the opportunity to visit the poster presentations. Friday featured a 40-minute keynote address and 10 20-minute sessions. I felt that there were too many sessions, cramming too much into a short space of time.

4. I did not like the layout of the main conference room – only about half the attendees had a table in front of them. (I prefer to have a table for note-taking.) The air-conditioning was unpleasant in that people sitting along the sides of the room were subjected to icy blasts of air. The catering arrangements also left a little to be desired – there were no scones, muffins or biscuits for morning and afternoon breaks. Indeed there was no tea or coffee for the Thursday afternoon break.

5. As for the conference itself, the main themes that interested me were:
   - skills, experience, interests of camera operators;
   - TIFF v JPEG2000; emergence of JPEG2000;
   - digital object management system (as opposed to storage);
   - role of microfilm;
   - automating the capture of metadata;
   - preservation v access
   - digital preservation: volumes and costs – practicality and sustainability.

6. The conference focused as much on imaging as digital preservation (which is not quite what I was anticipating). Most talks were academic/research based rather than case study based. I came away with much less practical exposure to digital preservation than I had been expecting. However, I did note a number of themes, see point 5 above, that PRONI will have to consider in the near future.

7. The following pages contain my notes on some of the sessions that I found interesting and note-worthy.

Hugh J Campbell
Public Record Office of Northern Ireland

Tutorial: Long-Term Archiving of Digital Images, Rudolf Gschwind, University of Basel

The emphasis in the tutorial was on images. RG looked at record keeping in the past and considered the implications for the digital world.

Objects in the physical world will decay at some time – the decay process is inexorable but it can be decelerated (conservation). It is impossible to save everything from decay. Photographs in archives and museums are part of our cultural heritage, providing a visual documentation of the 20th century. Examples of decay include: black and white - silver mirroring, yellowing, vinegar syndrome, sour paper; colour – colour dye fading.

The benefits of digitisation are:

- Only digital techniques will allow sensible access, and
- As a means of addressing the accumulation of decay factors.

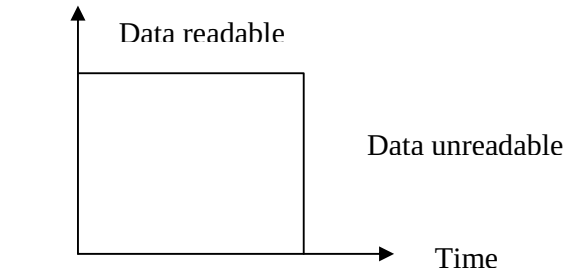
Parts of our cultural heritage in the past have survived:

- in symbolically encoded form;
- through regular copying;
- through dissemination.

For example, the Bible has been copied and disseminated 'digitally' many times. Content (thoughts, language) is medium independent (papyrus, chiselled in stone, paper). However, the quality of the medium is of secondary importance as long as the code can be decoded. The disadvantage of digital code is that it can only be decoded by machines. The advantages of digital code are that it can be copied and/or immaterially transferred with no information loss and there is no difference between the copy and the original.

RG proposed that, as analogue information always decays, in principle, only digital code can be archived.

Decay curve



Maintaining readability of digital material introduces the 'digital cliff'. Analogue material, e.g. photographs would have a decay curve – slowly but inexorably. Possible causes of the digital cliff:

- data medium defective (ageing, wear);
- no (hardware) devices, e.g. tapes but no devices to read them;
- no software – RG feels that software is a greater problem than hardware;
- faulty operation, e.g. back-up failure.

There are a number of implications with the digital way:

- is the state of the art solution;
- storage media will also deteriorate;
- storage media will be outdated soon;
- there are sources of information loss.

There are six areas of concern of information loss:

Key data	Technical data, e.g. how the medium is organised
Metadata	Without image descriptions etc image data are almost worthless
Data formats	No software to read obsolete image formats, e.g. new version not being able to read older versions.
Data medium formats	File system formats which are no longer supported.
Reading devices	Obsolete media, e.g. DLT1. Past cycle in computing has shown that things are on the way to being outdated as soon as they are released. The current lifetime cycle is about seven years.
Data medium	Damaged media

If you want to archive digital information, you have to fight against information loss.

Archiving means migration today. An archiving strategy needs to consider a number of issues: periodical verification; copying and migration; optimal storage conditions; several copies at different locations; documentation; metadata; long-term secured financing; open standard formats.

RG felt that there were a lot of issues (most of which were more organisational than technical):

- create a list of all files to be archived/copied;
- process a manageable batch size (depends on the storage media used);
- classify according to subordinate criteria is possible, e.g. folder structure;

- include metadata (technical, descriptive);
- produce copies on at least three different media (hard disk, DLT or TTO tape);
- store in different locations;
- check readability of tape data on a different tape drive (not the one used for writing);
- use checksums;

Storage medium issues:

- CD: medium 2-30 Years; format 10 years; fairly difficult in terms of handling; low capacity; difficult to distinguish between good and bad CDs.
- DVD-R(AM): less stable than CD; medium < 10 years; format 2-5? Years; problematic technology for archiving – too fragile, error prone and unreliable
- Digital Linear Tape (DLT, LTO): medium 30 years; format 5 years (is a problem); suitable for relative longevity; high capacity.
- Magneto-Optical Disks (MO-Drives): medium > 30 years; format 5 years; sensible longevity but expensive.
- Hard-Disk: medium < 10 years?; format 5-10 years; high capacity.

Helical tape – operation must be very precise as any mis-alignment can result in problems trying to read on a different machine.

Recommendation for archiving is Digital Linear Tape (DLT, LTO) and hard disk.

Image file format recommendations:

- use an open, well documented, uncompressed file format for archiving;
- TIFF – open, widespread, readable and writable by virtually all known imaging applications. Note that this can produce large file sizes.
- Do not use lossy formats (e.g. JPG) or compressed image formats (e.g. PNG) for archiving.
- JPEGs are good for presentation. Always produce JPEG files from TIFF archive files.
- JPEG2000 is too complicated and there is not enough software for it.

Curation and Preservation: Re-thinking Roles and Responsibilities in the Distributed Digital Environment, Sheila Anderson, King's College London

Started off mapping traditional archiving practices and made assumptions, e.g. collections were complete and static, depositors were happy to allow access, objects were in a manageable form, AHDS would take on archival curation and preservation processes. Worked for a while but is now changing. Key issues that have come to the fore:

- key tension between preservation and presentation. More than just objects and metadata; building up layers of information; interpretation.
- Sustainability – digital mode has ushered in a new way of sharing scholarly knowledge, e.g. web accessible databases with content and metadata; searchable digital library of text documents, catalogues and links – with links to other resources. Not always clear what is being curated and who is responsible.

Carried out a review and gap analysis of digital repositories, primarily UK based:

- what is a digital repository? Not just major institutions: scholars building own websites; use of peer-to-peer software to share.
- Collections cease to be a single entity hosted in a single place.
- Culture change required. Need to stand back and ask if someone could do this better – global collaboration.

Exploring Strategies for Digital Preservation for DSpace@Cambridge, Jim Downing, Cambridge University and Massachusetts Institute of Technology

Aim

- to develop an institutional repository for Cambridge using DSpace software. Currently over 30,000 items now in the repository, including e-prints, video, images.
- To reduce users' overheads in using the repository – make it as easy as possible to use thus removing a potential barrier that might prevent people from using it.

File type identification is an area where they are looking to automate. Currently use file extensions to identify file formats. Problem – lots of files have for example .doc extensions. Looked at some tools for automating file format identification. Hope to use JHOVE for file type identification, validation and characterisation. Another area for automation is metadata extraction. One of the problems that can occur is separation of items and metadata, i.e. items with no metadata, metadata with no items. Looking at how best to integrate (technical) metadata into an existing system. Looking at adding ongoing automated verification of checksums.

The development of preservation plans is an essential part of a quality repository service. Need action plan and background report for each format – this should include long-term preservation strategy and short-term actions.
Starting to work with JHOVE as part of the ingest process.

Policy development was every bit as hard (if not more so) than building the DRS – constantly being updated and revised:

- what can go in? – has a controlled list of file formats (only accepts library-like material)
- metadata requirements – each object has administrative and technical metadata. (Descriptive metadata held separately.)

Policy (creation and implementation) is the difference between storage and preservation.
DRS requires people as well as technology. Staff are in a constant state of learning.

Digital Repository Planning and Policy, Lee Mandell, Harvard University

Digital Repository Service (DRS) – electronic storage of digital objects by Harvard.
Issue to note – storage was on Unix file system rather than database because of the time required to perform a back-up, 2 – 3 days.
As technology evolves, the DRS has to do the same – do not build it and levee it.

Building A Dark Archive in the Sunshine State, Priscilla Caplan, Florida Centre for Library Automation

Strategy based on re-formatting. OAIS based. Build in redundancy for risk management – multiple master copies of files; always retain the original file as submitted; stores metadata in the Archival Information Package as XML and in RDBMS; virus check and checksum. Use normalisation – creating a copy of a file in a different format, one that is likely to survive longer. E.g. PDF file would be turned into page-image TIFFs. Would 'preserve' original and normalised versions – gives a greater chance of preserving over time.
Preservation can only happen when there is a documented action plan for format.

Hope to release under open source licence.

A Performance Model and Process for Preserving Digital Records for Long-Term Access, Andrew Wilson, National Archives of Australia

National Archives of Australia has about 450 staff, a budget of about $A65m, about 350 shelf km, about 90m+ items and only about 1TB of digital holdings. However, staff and funding in this area have grown. Hope to have an operational digital process by summer 2005.

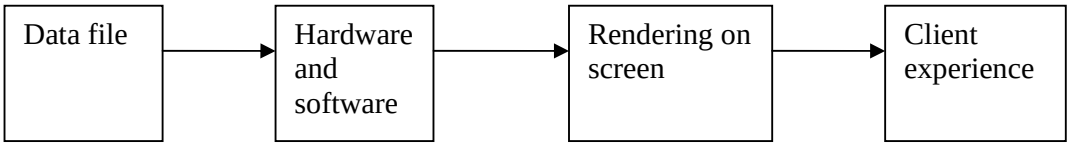
Digital records are the result of the mediation of hardware, software and data – unlike traditional, physical, unique records.

Performance model:

- Source is the recorded bitstream, ones and zeroes, by themselves unintelligible.
- Process is the technology (hardware and software) used to interpret the bitstream
- Performance is the experience of mediation of source and process – display on a screen, paper printout, audio through speakers.

We can keep the source but we cannot keep the process.

Digital records performance model:



Approach 1
Passive access - keep a master copy of every source take in. Researcher gets access to copy of original as bitstream, not the performance.
Approach 2
Active intervention to re-create the performance. The source object becomes less important than the performance. You could change any component in the source or process and it would not matter as long as the performance is the same – defined as the essence of the performance.

Designing Effective Retrieval Systems for Digital Archives of Historic Documents, Andy White, University of Ulster

Focus on two digitisation projects. Digital archive is non-linear. This underpins everything. Often we are putting linear material onto an anti-linear platform.

Case Study 1: Troubled Images: About 3,500 images, mainly posters, not catalogued. Selection excluded duplicates and those not relevant to NI politics; next stage was cataloguing. Also digitised other objects such as flags and murals. Used Filemaker Pro database – populated with images and associated catalogue information. Also wrote annotations of about 50 words for each image – e.g. allowed search using ‘civil rights’ to find something that did not appear in a poster. Annotation provides a full text search in addition to other searches, e.g. author, date. Also contains some audio interviews. Quality assurance was relatively easy as there were only a few hundred thousand words.

Case Study 2: Act of Union Virtual Library: www.actofunion.ac.uk Bibliographic elements in MS Access; 400dpi TIFF images; ascii text tagged with page delimiters; data fed into MySQL.
Struck balance (between speed and quality/legibility) in selecting presentation image – tested using laptop with 56k modem.

Conclusions:
- design to exploit non-linear narrative;
- use other techniques and explore other ways – don't just present a book;
- quality assurance is critical.

Archiving, Stewardship, Curation: From the Personal to the Global Sphere, Clifford Lynch, Coalition for Networked Information

Digital preservation is becoming a concern of a lot of ordinary people, not just academics, researchers, institutions – this is very significant. There has been an enormous investment in the digitisation of cultural material: Google's proposal to digitise (out of copyright) material from libraries and make it available on-line: Brewster Kahle's proposal to digitise one million books and make them available on-line. History shows us that bad things happen to unique collections. Whilst a digital copy is not a replacement, it would be good to have a surrogate if something happened to the original. People are starting to hold sizeable digital collections, e.g. music, photographs. The question is: how long before this cannot be migrated the next format? The music industry view seems to be that you just buy it again. People are moving to third, fourth PC – question of how to migrate from one to the other. Starting to see sharing of this material with others, e.g. family members. Also starting to see digital objects appearing in wills. A fundamental problem may be that in future (for historians) the amount of source material may be over whelming. Starting to see companies offering to store digital material (although not promising long-term storage – almost certain that some of these will go out of business.

RIT American Museums Survey on Digital Imaging for Direct Capture of Artwork, Mitchell Rosen, Munsell Colour Science Laboratory and Rochester Institute of Technology

Survey www.art-si.org – follow links to benchmarking.
Feels that there is a huge gap between best practice and common practice.

Imaging Workflows, Reports from the Field: Panel discussion

Erik Landsberg, MoMA

Digitise as part of the acquisition process now. Have had a recent re-structuring – imaging a core function. All copy distribution is now digital. Typical file size is now 100MB, 8 bit. Now starting to consider 16 bit. Starting to consider Adobe DNG format but don't want to be the first to jump. Could spend up to an hour editing a single the colour in an image to get it right.

David Remington, Harvard College Library

When digitising a collection, one option is to do a few samples and apply standard (colour correction) to the entire collection. High quality would require colour correction to be applied to each image.

Stuart Snydman, Stanford University Libraries

Have robotic page turning, scanning device – not for rare or fragile material. Primarily greyscale scanning of text documents – with a high premium on automation in the workflow. Create files for viewing, printing, searching, archiving. Digitise about 3,000 images per day. Typically 400dpi greyscale. Adding 6-8TB per year. Would like 600dpi colour TIFF, uncompressed master and corrected master but unable to do so yet due to space considerations. Quick check of every page – hope to go towards random selecting for QA. Not doing any correction of OCR. Averaging about 99% character accuracy on text. Would like to save raw images but not in a proprietary format.

Design for the long Term: Authenticity and Object Representation, Adam Farquhar, British Library

The British Library has a duty to preserve non-print material now. Have taken in about 1.5-2TB of material under voluntary deposit scheme (in preparation for flood to come). Also have digitised material. Also have sound archive (15TB/50 years). Also started to archive UK web domain. Planning for 500 TB per year.

Have developed Digital Object Management (DOM) program to enable UK to preserve and use its digital intellectual heritage.

DOM system:

- needs to be disaster tolerant – needs robust, multi-site design and clear security;
- scalable 100s TB and millions of objects;
- cost effective;
- OAIS compatible.

Hugh J Campbell
Public Record Office of Northern Ireland

Key issues – integrity and authenticity. Relatively easy to detect when paper has been damaged or modified, e.g. tears, stains, folds, visible signs of change. More complicated with digital material. Use checksum or digest to detect damage. Digitally sign every digital object using public key cryptography. Also use a hardware solution that is tamper resistant and tamper evident. Use a trusted time stamp. Use a layered approach. Make minimal assumptions. Uniquely identified bitstreams linked to metadata. Use METS framework for metadata – includes link from metadata to bitstream.

www.bl.uk/about/policies/dom/homepage.html

PREservation Metadata: Implementing Strategies, Rebecca Guenther, Library of Congress

Implementation survey report 2003/04 – findings:

- little experience with digital preservation – a lot in the planning stage;
- lack of common vocabulary and conceptual framework;
- some metadata was being recorded but it varied – could not assess adequacy.
- Trying to server both goals of preservation and access;
- METS
- OAIS as a framework and starting point;
- Maintaining multiple versions;
- Choosing multiple strategies for digital preservation

Portable Image Archiving: Annotation, Search and Data Retrieval, Vladimir Misic, Rochester Institute of Technology

Problems of digital archives:

- too big (MB, GB, TB);
- image quality (decreases at high compression ratio);
- content redundancy and data access – multiple versions, master lossless, master colour corrected etc;
- portability – bibliographic information not part of the image file.

Is JPEG2000 the answer?

- superior performance at high compression;
- continuous tone and bi-level compression;
- lossless and lossy compression in same codestream;
- progressive transmission – image built as transmission takes place;
- error resilience;
- metadata handling – can have metadata relating to different parts of an image – can search and pull back just part of an image.

Hugh J Campbell
Public Record Office of Northern Ireland

JPEG2000 is defined as a collection of boxes, some mandatory, some optional. There are optional XML boxes where any metadata could be stored. Could preserve the raw original in one of the boxes.

JPEG2000 v TIFF

- comparable on some features: image bit depth, lossless compression;
- TIFF advantage – established practice;
- JPEG2000 advantage – embedded additional content (not just related to whole image); supports multi resolution (eliminates need to store same image at different resolution, compression; random codestream access (image can be reorganised on the fly).

Real Time, Deep Time, Life Time: Spanning Digital and Traditional Collections Life Cycles, Helen Shenton, British Library

Life cycle management – breaking into different stages – way of looking at management of collection. What are expected life of paper?, digital? What are the costs of maintaining collections? Work carried out to define costs (over life of collections). BL identified 8 stages: selection; processing; cataloguing; initial preservation; handling; longer-term preservation; storage; retrieval and replacement. Time span: Year 1 (initial costs, e.g. acquisition, cataloguing); Year 10 (technology change may create costs); Year 100 – have to preserve in perpetuity. Analysed costs at these 3 time periods. Costs shift from selection/acquisition to retrieval, storage and preservation over time.

BL digital collection (includes voluntary deposit). Digitised images – best practice – digitise at high quality and retain.

Mapped stages to digital environment. Addressing issues associated with cost of storing digital material. Digital Object Management programme underway.

Also looking at risk associated with digital material.

BL creating new conservation centre: state of the art; public access; tours.

Earliest New Testament: held in different locations; project to unite virtually.

Microfilm: A Preservation Technology for the 21st Century? Stephen Chapman, Harvard University

NARA 1985 identified 530m documents as high risk 'Hayes' Study, 1986, found 11m unique volumes in libraries at risk. NARA review endorsed preservation and use of microfilm. Goal

Hugh J Campbell
Public Record Office of Northern Ireland

(post Hayes) - to preserve 3m volumes. From 989 to date: spent about $84m, preserved on microfilm just over 1m volumes. (This would suggest to me that institutions should focus on the digitisation of a small number of significant archives rather than the impossible task of digitising millions of original documents.)

Association of Research Libraries published a paper in 2004 recognising digitisation as a preservation reformatting method.

Points to note in managing obsolescence:

- not found;
- not useable;
- not supported (by continuous funding stream). What happens if custodians stop paying? – migration is a financial issue as well as a technical issue;
- not convenient.

Format independent criteria for use, sustainability and affordability for preservation.

Use – obligation to preserve level of service as well as the content; deliver to users in preferred formats?
Sustainability – metrics for sustainability need to be defined.
Affordability – underlying business model must ensure that it is perpetually sustainable.

Computed costs for storing Women Working texts in best preservation environments offered by Harvard		
	$ per title	$ collection
35mm microfilm	0.21	436
Printed volumes (average 189 pages/volume)	0.22	456
Digital images (average 687KB)	.62	1,273

Microfilm costs are close to paper because of decision to store second master in film vault rather than standard vault. If second master was moved to standard vault, microfilm cost would be reduced by approximately 29%.

Conclusion: microfilm is a good preservation format; is a good delivery format (for communities satisfied with reader/printers).