Towards a digital preservation policy

Presented by Ria Groenewald & Ina Smith
National Library IT Event
20 May 2008
# Agenda

<table>
<thead>
<tr>
<th>Part I</th>
<th>Digitization &amp; Preservation</th>
<th>Ria Groenewald</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td><a href="mailto:ria.groenewald@up.ac.za">ria.groenewald@up.ac.za</a></td>
</tr>
<tr>
<td>Part II</td>
<td>Preservation &amp; Trusted Digital Repositories</td>
<td>Ina Smith</td>
</tr>
<tr>
<td></td>
<td></td>
<td><a href="mailto:ina.smith@up.ac.za">ina.smith@up.ac.za</a></td>
</tr>
</tbody>
</table>
Part I: Digitization & Preservation

Ria Groenewald
• The birth of the web site
6 August 1991 - 2:56:20 pm

• The world’s first website was made available on the public internet - a creation of Tim Bernes-Lee at CERN
The WorldWideWeb browser

The first web browser - or browser-editor rather - was called WorldWideWeb as, after all, when it was written in 1990 it was the only way to see the web. Much later it was renamed Nexus in order to save confusion between the program and the abstract information space (which is now spelled World Wide Web with spaces).

I wrote the program using a NeXT computer. This had the advantage that there were some great tools available - it was a great computing environment in general. In fact, I could do in a couple of months what would take more like a year on other platforms, because on the NeXT, a lot of it was done for me already. There was an application builder to make all the menus as quickly as you could dream them up. there were all the software parts to make a wysiwyg (what you see is what you get - in other words direct manipulation of text on screen as on the printed - or browsed page) word processor. I just had to add hypertext, (by subclassing the Text object)

This is a (242kB) screen shot of the browser, taken when things had got to the point that Communications of the ACM was interested in an article, in 1993. The differences between this and the first edition (Christmas 1990) were:

- The whole thing would have been grey scale as NeXTs were at the time just grey scale;
- The inline images such as the world/book icon and the CERN icon, would have been displayed in separate windows, as it didn't at first do inline images.

See also:

- another screen shot, this one by JFG, later but grey scale

A quick tour of this screen to answer the FAQs:

In this shot I am making a link from the word "ATLAS" in the list of experiments to some web page.

The NeXTstep operating system put the menu for each application in the top left of the screen. The application is called WorldWideWeb, because the menus are in this block they windows are very unencumbered. A little like like the windows "start" menu later.

The Navigate menu had things like "back" and "next" and "previous", these last two were useful when you follows a link from a list of links - they meant "go back a step and then take the next link from the same page instead"

The document menu was like the "file" menu for windows I suppose. The "find" menu is fairly self-explanatory, as is "edit".

The "Link" menu you can see. "Mark all" would remember the URI of where you were. "Mark selection" would make an anchor (link target) for the selected text,
World Wide Web

The WorldWideWeb (W3) is a wide-area hypertext information retrieval initiative aiming to give universal access to a large universe of documents.

Everything there is online about W3 is linked directly or indirectly to this document, including an executive summary of the project, Mailing lists, Policy, November's W3 news, Frequently Asked Questions.

What's out there?
- Pointers to the world's online information, subjects, W3 servers, etc.

Help
- on the browser you are using

Software Products
- A list of W3 project components and their current state. (e.g. Line Mode, X11, Viola, NeXtStep, Servers, Tools, Mail robot, Library)

Technical
- Details of protocols, formats, program internals etc

Bibliography
- Paper documentation on W3 and references.

People
- A list of some people involved in the project.

History
- A summary of the history of the project.

How can I help?
- If you would like to support the web.

Getting code
- Getting the code by anonymous FTP, etc.

http://www.w3.org/History/19921103-hypertext/hypertext/WWW/TheProject.html
History to date

A few steps to date in the WWW project history are as follows:

March 1989  
First project proposal written and circulated for comment (TBL). Paper "HyperText and CERN" (in ASCII or WriteNow format) produced as background.

October 1990  
Project proposal reformulated with encouragement from CN and ECP divisional management. RC is co-author.

November 1990  
Initial WorldWideWeb prototype developed on the NeXT (TBL).

November 1990  
Nickola Pellow joins and starts work on the line-mode browser. Bernd Pollemann helps get interface to CERNVM "FIND" index running. TBL gives a colloquium on hypertext in general.

Christmas 1990  
Line mode and NeXTStep browsers demonstrable. Access is possible to hypertext files, CERNVM "FIND", and internet news articles.

February 1991  
workplan for the purposes of ECP division.

26 February 1991  
Presentation of the project to the ECP group.

March 1991  
Line mode browser (WWW) released to limited audience on priam vax, rs6000, sun4.

May 1991  
Workplan produced for CN/AS group.

17 May 1991  
Presentation to C5 committee. General release of www on central CERN machines.

12 June 1991  
CERN computer seminar on WWW.

August 1991  
Files available on the net, posted on alt.hypertext (6, 16, 19th Aug). Comp.sys.next (20th), comp.text.sgml and comp.mail.multi-media (22nd). Jean-Francois Greff joins the project.

October 1991  
VMS HELP and WAIS gateways installed. Mailing lists www-interest (now www-announce) and www-talk@info.cern.ch started. One year status report. Anonymous telnet service started.

December 1991  
Presented poster and demonstration at HT91. W3 browser installed on VMS/CMS. CERN computer newsletter announces W3 to the HEP world.

15 January 1992  
Line mode browser release 1.1 available by anonymous FTP. See news. Presentation to AHEP'92 at La Louque.

12 February 1992  
Line mode v 1.2 announced on alt.hypertext, comp.infosystems, comp.mail.multi-media, cern.sign, comparchives.admin, and mailing lists.

May 1992  
Presentation and demo at JENC3 (Innsbruck). Carl Barker joins the project.

June 1992
Internet Archive, Library of Alexandria (copy of the Internet) mirror site
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: Some duplicates are not shown. See FAQ.
Use of digital information

- It took only 5 years for 50 million people to use the internet compared to 25 years for 50 million people to use phones

- Preservation or permanent availability of digital information is one of the processes affected by the evolution towards an all digital world
Digital preservation is a broad term used to describe the continued accessibility and maintenance of a digital resource safeguarding it into the foreseeable and the distant future. Digital preservation is a vital part of the creation and management of any digital collection.

http://www.tasi.ac.uk/advice/delivering/digpres.html
Joost sê sy handtekening is ‘gecopy en gepaste’

Nellie Brand, Kaapstad

Joost van der Westhuizen se handtekening verskyn onderaan ‘n brief van Progressive Investment Holdings (PIH) wat na bewering verkeerde inligting en ‘n logo van ‘n ander batebestuurder bevat en aan klíënte gestuur is.

PIH en drie verwante ondernemings is vandeesweek onder voorlopige curatorskap geplaas weens onder meer die beweerd oortreding van die Wet op Finansiële Advies-en Tussengangerdienste (Fais).

Die Raad op Finansiële Dienste (RFD) het die aanse het die curatorskap ingedien na ’n jaarlange onderzoek na PIH se sake. Finansiële dienste is na bewering onregmag en gelever.

Van der Westhuizen is sedert begin Maart bedryfsohoof by PIH. By navaar het hy aan Sake in geë se dat dit wel sy handtekening is wat onderaan die brief verskyn, maar bygevoeg dit is ‘n afkroft daarvan wat op die brief “gecopy en gepaste”.

Volgens hom het hy toestemming gegee dat sy handtekening gebruik mag word solank hy ingelig word waarvoor dit gebruik word.

Oor die betrokke brief, gedateer 5 Mei, het Van der Westhuizen gesê die beslissingsdepartement van PIH het dit uitgestuur om aan mense te wys hy is “aan boord”.

Volgens hom het hy nie geweet dit word “namens hom uitgestuur nie”.

Van der Westhuizen het gesê hy weet nie wat die inhoud van die brief beteken nie. Hy het volgehou dat hy nog opgeleid word en net ‘n “student” is en niks van die onderneming se sake weet nie.

Hy ontvang opleiding van ‘n persoonlike opleier, mnr. Murray Kilgour, sodat hy in Januarie volgende jaar as uitvoerende hoof van die maatskappy kan oorneem.

Die pos van bedryfsohoof is intussen vir hom geskep deur mnr. Freddie Andalaf, huidige uitvoerende markingsdoelindes, nie vir kontakte nie,” het hy gesê.

Volgens hom word dit ook op pamflette en penne gebruik, asook be markingsbrieewe soos die betrokke een.

Die brief bevat ook ‘n handelsmerk van die batebestuurder Dynamic Wealth wat volgens mnr. Phillip Hattingh, bestuurende direkteur van die onderneming, onwettig gebruik is.

Hattingh het gesê hulle gaan regstappe doen daaroor. Volgens hom word alle briefe waarop hul handelsmerk verskyn, deur hul nakomingskantoor goedgekeur vanwee die streng vereistes wat deur die Fais-wet gestel word vir inligting wat aan klíënte gestuur word.

Die brief se inligting is ook verkeerd, volgens Hattingh.

In die brief word onder meer gesê dat van Dynamic Wealth se “aandeelplatform” gebruik gemaak word.

Volgens Hattingh het PIH nog nooit ‘n aandelerenking by Dynamic Wealth gehad nie.

Andalaf het gesê die logo is gebruik na aanleiding van ‘n brief van ‘n direkteur van Dynamic Wealth.

In daardie brief, gedateer April, skryf mnr. Kobus Zietsman, ‘n direkteur van Dynamic Wealth, dat hulle die effektruists van Progressive Asset Management (PAM) so spoedig moontlik gaan oorneem.

Hattingh het egter gesê niks is deur die direksie goedgekeur nie.

Hattingh het gesê dit is nie die eerste keer dat sulke ondernemings probeer om uit Dynamic Wealth se goeie naam munt te slaan nie.

Dit het ook al met ander bekende batebestuurders gebeur, het hy gesê.

Die RFD het intussen ‘n verkla ring uitgereik waarin gesê is die betrokke effektruists word deur Metropolitan Collective Investments besit.

Die RFD het gesê die effektruists is nie onder curatorskap nie. Dit is nie watter bewering be betrokkie is nie.
Future of academic libraries

No.1 assumption (ACRL, March 2007)

There will be an

– preserving digital archives, and
– increased emphasis on digitizing collections
– improving methods of data storage and retrieval

• The digitization of unique print collections may emerge as one of the primary missions of academic libraries in the 21st century

• Librarians should collaborate with disciplinary colleagues in the curation of data as part of the research process

Digital workflow of the Alexandria Library. Software for this workflow is available at http://wiki.bibalex.org/DAFWiki/index.php/Main_Page
Metadata Editor

Send to submitters via:
- email
- External hard drive
- DVD/CD/Flashdrive
- Internal server

UPSpace IR

Archival server

- Scan directly to archival server

Reviewer

- Copy from AS
- Quality Control
- Deskew/cleaning/derivating/filter
- Safe webready
- Final QC + Storage

Unique URI created for object
Standards (1)

- PREMIS (2005)
- OAIS (Open Archival Information System)
- Z39.87 - Standard for Technical Metadata for Digital Still Images (ANSI/NISO)
Standards (2)

• ISO (International Standards Organisation) defines “quality as the totality of characteristics of an entity that bear on its ability to satisfy stated or implied needs” (ISO8042:1994)

• or “the degree to which a set of inherent characteristics fulfils requirements” (ISO9000:2000)
Framework for research

- Outline the types of information that should be associated with an archived digital object
- The use of metadata to support the digital preservation process

The PREMIS (Preservation Metadata: Implementation Strategies Working Group -

- Develop a data dictionary of core elements for archived objects
- Guide the implementation of element sets in preservation systems
- Suggest best practice for populating the elements

The OAIS (Open Archival Information System) reference model was developed under the auspices of NASA’s Consultative Committee for Space Data Systems (CCSDS).

The OAIS reference model is a conceptual framework for a digital archive.

Regarded as the “standard” for digital object repositories.
Z39.87 is a standard which defines a set of metadata elements for raster digital images.

The purpose is to help in the development, exchange and interpretation of digital images.

The original DIG35 goals were adapted by the NISO group.
Scanning

- No set resolution can be selected for all projects
- Resolution for a master image range between 300 - 600 dpi
- Colour settings 8-bit greyscale; 24-bit colour
- The most widely adopted format for storing a preservation quality digital master is uncompressed TIFF
• A derivative is a manipulated image derived from the master image, to produce smaller file sizes

• Lossy file formats such as JPEG are used for derivative images

• Resolution ranges between 72 dpi and 150 dpi and up to 800 pixels in width

• ICC (International Colour Consortium) profiles
Reasons for preservation

• Updated versions of the file format
• Reading device become obsolete
• Updated versions of the software used to create, manage, or access digital content
• Changes in computers
• Movement at vendors level
• Unforeseen errors
Requirements of data protection

- Visibility/accessibility
- Regular quality control
- Authenticity
- Security
- Performance
- Ease of use
- Interoperability
- Cost of ownership
- Automation
Preservation methods

- **Refreshing:**
  Copy the same type of digital information from one long-term storage medium to another

- **Modified refreshing:**
  Copy information to another medium of a similar type

- Refreshing is part of a process or program

- Refreshing address issues such as decay and obsolescence
Migration and Emulation

- **Migration:**
  Move or adapt the objects to another platform

- **Emulating:**
  Environment will be adapted to new platform (the objects themselves will not be tampered with)
Preserve the usability of a .TIFF file

- A TIFF viewer, plus its formal specification and sufficient subsidiary documentation to explain how it work in practice must be preserved
- To run the TIFF viewer - an operating system must be preserved
- To run the operating system -
  - the original hardware will need to be preserved, or
  - emulation software that allows the old hardware to be emulated on new machines needs to be developed
Welcome to PRONOM

The online registry of technical information, PRONOM is a resource for anyone requiring impartial and definitive information about the file formats, software products and other technical components required to support long-term access to electronic records and other digital objects of cultural, historical or business value.

Tools and Services
Free PRONOM tools and services to support digital preservation, including DROID, the automatic file format identification tool, together with links to relevant external tools and services.

Contribute to PRONOM
Contribute new information to PRONOM via our online submission form.
Find out more about PRONOM's creators.

New to PRONOM?
- What is PRONOM?
- How do I search PRONOM?
- Who is PRONOM for?
- How do I find out more?
**JPEG File Interchange Format (1.00)**

The JPEG File Interchange Format (JIF) is a file format for storing JPEG-compressed raster images. It was developed by the Independent JPEG Group and C-Cube Microsystems, in the absence of any such format being defined in the JPEG standard, and rapidly became a de facto standard; this is what is commonly referred to as the JPEG file format. A JIF file comprises a JPEG data stream together with a JIF marker. It begins with a Start of Image (SOI) marker, immediately followed by a JIF Application (APP0). This is followed by the JPEG image data, which is terminated by an End of Image (EOI) marker. JIF supports up to 24-bit colour and uses lossy compression (based on the Discrete Cosine Transform algorithm). Other types of compression are available through JPEG extensions, including progressive image buildup, arithmetic encoding, variable quantization, selective refinement, image tiling, and lossless compression, but these may not be supported by all JIF readers and writers.

---

**JPEG File Interchange Format (1.01)**

The JPEG File Interchange Format (JIF) is a file format for storing JPEG-compressed raster images. It was developed by the Independent JPEG Group and C-Cube Microsystems, in the absence of any such format being defined in the JPEG standard, and rapidly became a de facto standard; this is what is commonly referred to as the JPEG file format. A JIF file comprises a JPEG data stream together with a JIF marker. It begins with a Start of Image (SOI) marker, immediately followed by a JIF Application (APP0). This is followed by the JPEG image data, which is terminated by an End of Image (EOI) marker. JIF supports up to 24-bit colour and uses lossy compression (based on the Discrete Cosine Transform algorithm). Other types of compression are available through JPEG extensions, including progressive image buildup, arithmetic encoding, variable quantization, selective refinement, image tiling, and lossless compression, but these may not be supported by all JIF readers and writers.

---

**JPEG File Interchange Format (1.02)**

The JPEG File Interchange Format (JIF) is a file format for storing JPEG-compressed raster images. It was developed by the Independent JPEG Group and C-Cube Microsystems, in the absence of any such format being defined in the JPEG standard, and rapidly became a de facto standard. A JIF file comprises a JPEG data stream together with a JIF marker. It begins with a Start of Image (SOI) marker, immediately followed by a JIF Application (APP0) marker and one or more optional application extension markers. This is followed by the JPEG image data, which is terminated by an End of Image (EOI) marker. JIF supports up to 24-bit colour and uses lossy compression (based on the Discrete Cosine Transform algorithm). Other types of compression are available through JPEG extensions, including progressive image buildup, arithmetic encoding, variable quantization, selective refinement, image tiling, and lossless compression, but these may not be supported by all JIF readers and writers.
## Details for: JPX (JPEG 2000 Extended)

### Summary

<table>
<thead>
<tr>
<th>Name</th>
<th>JPX (JPEG 2000 Extended)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Version</td>
<td></td>
</tr>
<tr>
<td>Other names</td>
<td>JPX</td>
</tr>
<tr>
<td>Identifiers</td>
<td>PUID: fmt/151</td>
</tr>
<tr>
<td>Family</td>
<td></td>
</tr>
<tr>
<td>Classification</td>
<td></td>
</tr>
<tr>
<td>Disclosure</td>
<td></td>
</tr>
</tbody>
</table>

### Description

JPX 2000 Extended File Format is an optional, lossless, extension of the JP2 format, the main file format for the JPEG 2000 international standard for image coding, created by the Joint Photographic Experts Group in 2000 for the compression of photographic images for storage or transmission. The standard specifies how an image and its metadata is transformed into byte stream data. This format is widely used for storing and transmitting photos and other compressed image data online. Although it is based on JP2 it can support multiple layers, animation and other features, and is a lossless image compression file. Like JP2 the format is made up of a contiguous sequence of boxes, beginning with a signature box and a file type box, which provides version and file type information.

### Orientation

Text

### Byte order

Big-endian (Motorola)

### Related file formats

None.
Preservation of the format

Digital formats contain texts, databases, still and moving images, audio, graphics, software and web pages. They are fragile and require purposeful production, maintenance and management to be retained.

- **Viability** - maintenance of the bitstream
- **Renderability** - viewable by humans and processible by computers
- **Understandability** - interpretable by humans

http://www.icpsr.umich.edu/dpm/dpm-eng/terminology/preservation.html
Part II: Preservation & Trusted Digital Repositories

Ina Smith
Institutional Repository

“A university-based institutional repository is a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members. It is most essentially an organizational commitment to the stewardship of these digital materials, including long-term preservation where appropriate, as well as organization and access or distribution.”

Clifford A. Lynch,
Digitally born & digitized material

**Clinical Cases**

**Awaking a sleeping epidemic**

Lies van Collie, Elise van Laerbeek, Jean Schilt, Rhiannon Lock, Michael Wahlgren, MG Dews

Two patients with African sleeping sickness (NSS) presented to the neurology unit, Pasteur Institute Hospital, in May 2006 and 2007. SS has shown a recent resurgence, with epidemics in the Sudan, Angola and the Democratic Republic of Congo. The number of infected people in Africa is currently estimated at more than 500,000. According to the World Health Organization (WHO), about 25,000 new cases are reported each year. The diagnosis of African trypanosomiasis is mainly made by clinical symptoms and by clinical examination of the CSF. Diagnosis of a relapse was made in one patient who presented a clinical picture of a relapse which was confirmed by a positive reaction in the ELISA test.

Five months after discharge, one patient presented to the hospital for a relapse of the disease. The patient was admitted to the neurology unit and the diagnosis of African trypanosomiasis was confirmed by a positive reaction in the ELISA test. The patient was treated with artemether and recovered.

**Case 1**

A 27-year-old woman presented with a 1-month history of fatigue, loss of appetite, intermittent fever, headache, and excessive daytime sleepiness. She had traveled to Malawi 3 months before admission. Her temperature was 38°C, she had a palpable lymphadenopathy and an unremarkable neurological examination. The patient had been diagnosed with malaria and treated with chloroquine and sulfadoxine-pyrimethamine. Five days after admission, the patient was admitted with a high fever, headache, and altered mental status. The CSF revealed a pleocytosis with a lymphocytic predominance and an elevated protein concentration. The patient was diagnosed with African trypanosomiasis and treated with artemether and sulfadoxine-pyrimethamine. The patient recovered and was discharged after 2 weeks.

**Case 2**

A 22-year-old woman presented with a 2-week history of fever, headache, and altered mental status. She was a nine-year-old girl from Zimbabwe where she had been treated for malaria without any clinical improvement. She had a history of multiple blood transfusions but did not have a diagnosis of African trypanosomiasis. On admission, her temperature was 38°C, she had a general examination, and was unremarkable. She was very weak but she was able to walk. The CSF revealed a pleocytosis with a lymphocytic predominance and an elevated protein concentration. She was diagnosed with African trypanosomiasis and treated with artemether and sulfadoxine-pyrimethamine. The patient recovered and was discharged after 2 weeks.

**Fig. 1.** Graph showing the correlation between trypanosome parasitemia and the peripheral blood.
Digitally born & digitized material
https://www.up.ac.za/dspace/
Benefits of an open access IR

- Research out quickly, worldwide
- Increases visibility, usage, impact of research
  “open access [material] are read more widely, and, therefore, cited more frequently. The consequence of this is that they have greater impact” (Jones, Andrew and MacColl 2006)
- Open access to all – also those who cannot afford subscribing
- Persistent URL
- Decentralised/ Distributed input
- E-workflow for quality control
- Full text - searchable
- Central archive of research
- Preservation function
11 Repositories National (ROAR)

- African Higher Education Research Online
- CSIR Research Space
- Durban University of Technology Institutional Repository
- Rhodes eResearch Repository
- Stellenbosch University Electronic Theses & Dissertations
- University of Cape Town Computer Science Research Document Archive
- University of Cape Town Lawspace
- University of Johannesburg Electronic Theses & Dissertations
- University of Pretoria Electronic Theses & Dissertations
- University of Pretoria Institutional Repository (UPSpace)
- University of the Western Cape Electronic Theses & Dissertations
1 200 Repositories Internationally

Univ. of Australia

Cornell University

And many more ….
Trusted Repository Defined

“One whose mission is to provide **reliable, long-term access** to managed digital resources to its designated community, now and in the future.”

(RLG-OCLC Report 2002)
Attributes of a Trusted Repository

• Compliance with the Reference Model for an Open Archival Information System (OAIS)
• Administrative responsibility
• Organizational viability
• Financial sustainability
• Technological & procedural suitability
• System security
• Procedural accountability

Source: Trusted Digital Repositories: Attributes and Responsibilities
An RLG-OCLC Report
OAIS Functional Model – Archival Storage

Source: http://public.ccsds.org/publications/archive/650x0b1.pdf
Archival Information Package (Digital item submitted)

Source: http://public.ccsds.org/publications/archive/650x0b1.pdf
Technologies for enabling trust & preservation
Digital Repository Software

- Proquest Digital Commons (*proprietary*)
- DSpace (*open source*)
- ContentDM (*proprietary*)
- Fedora (*open source*)
- E-Prints (*open source*)
- Greenstone (*open source*)
File formats

- Proprietary e.g. MSWord
- Open formats e.g. ASCII plain text – no restrictions
- Industry standard formats e.g. HTML, PDF
- Industry/Open standard format with proprietary extension e.g. Microsoft Version of XML
DSpace Commitment to Preservation

• 2 levels of preservation: Bit & Functional
• Three levels of preservation for a given file format:
  – Supported: The format will be fully supported and preserved using either format migration or emulation techniques.
  – Known: The format can be recognised by DSpace, but full support cannot be guaranteed.
  – Unsupported: The format cannot be recognised by DSpace; these will be listed as "application/octet-stream", aka Unknown.

• Bit-level preservation will be done so that digital archaeologists of the future will have the raw material to work with if the material proves to be worth that effort.
E.g. Adobe PDF, XML, Text, HTML, MSWord - Known
**UPSpace Policy for file formats**

- **Everything** put in UPSpace will be **retrievable**
- As many **files formats** as possible will be **recognised**
- As many known **file formats** as possible will be **supported** through UPSpace
- **Formats and techniques** will be continuously **monitored** to ensure needs can be accommodated as they arise
- The **size of a bitstream** allowed for submission is currently **unlimited**, but this will be revised over time
- The same **file** can be submitted in **more than one format**, of which one must be **pdf** (does not apply to media files)
Preserving items in DSpace

Metadata + Bitstream

Relationships stored between components in a bundle
Metadata

- Data about data
- Qualified Dublin Core Metadata Schema
- DSpace supports the Open Archives Initiative’s Protocol for Metadata Harvesting (OAI-PMH) v2.0 as a data provider
- Enhance Descriptive Metadata
- Capture Administrative Metadata (incl. preservation metadata)

“Preservation metadata is the information necessary to maintain the viability, renderability, and understandability of digital resources over the long-term.”

Source: Feasibility and Requirements Study on Preservation of E-Prints/ Hamish et al.
## Preservation Metadata

<table>
<thead>
<tr>
<th>identifier</th>
<th>uri</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td><a href="http://hdl.handle.net/2263/4726">http://hdl.handle.net/2263/4726</a></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>description</th>
<th>abstract</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>In this paper, an attempt will be made to examine the concept of ataraxia as it appears in the works of Pyrrho of Elis, Sextus Empiricus, and others.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>description</th>
<th>provenance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Scanned in 24-bit descreened colour 100% on DigiBookRGB1000 scanner at 400 dpi.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>description</th>
<th>provenance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Submitted by Ina Smith (<a href="mailto:ina@ais.up.ac.za">ina@ais.up.ac.za</a>) on 2008-03-13T11:55:46Z</td>
</tr>
<tr>
<td></td>
<td>No. of bitstreams: 1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>description</th>
<th>provenance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Approved for entry into archive by Julene Vermeulen (<a href="mailto:julene.vermeulen@up.ac.za">julene.vermeulen@up.ac.za</a>) on 2008-03-14T06:03:00+02 (GMT)</td>
</tr>
<tr>
<td></td>
<td>No. of bitstreams: 1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>description</th>
<th>provenance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Made available in ESpace on 2008-03-14T09:00:22Z (GMT)</td>
</tr>
<tr>
<td></td>
<td>No. of bitstreams: 1</td>
</tr>
<tr>
<td></td>
<td>PHV1_Wilkinson-009.pdf: 145337 bytes, checksum: 0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>format</th>
<th>extent</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>145337 bytes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>format</th>
<th>mimetype</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>application/pdf</td>
</tr>
</tbody>
</table>
Checksums in DSpace

Checksum generated by Checksum software:

<table>
<thead>
<tr>
<th>Name</th>
<th>Size</th>
<th>Checksum\State</th>
</tr>
</thead>
<tbody>
<tr>
<td>smithgroenewald_karm08.pdf</td>
<td>3.663 KB</td>
<td>E5A95F725F982A4B08B3C35DA7DC60C</td>
</tr>
</tbody>
</table>

Identical to Checksum generated by DSpace (UPSpace):

<table>
<thead>
<tr>
<th>File</th>
<th>Size</th>
<th>File Format</th>
<th>Checksum</th>
</tr>
</thead>
<tbody>
<tr>
<td>smithgroenewald_karm08.pdf</td>
<td>3,751,421 bytes</td>
<td>Adobe PDF (known)</td>
<td>e5a95f72b5f982a4b08b3c35da7dc80c (MD5)</td>
</tr>
</tbody>
</table>
Digital Signatures

• **Digital signatures** added to full text
• Compute a digital signature for digital masters & store signature in technical metadata of object
• Compute signature for complete item and store externally to repository

**AIP – Preservation Info – Fixity**
Persistent Identifiers

- Web references are untrustworthy; telephone numbers, IP addresses, Social Security numbers share properties of PID’s – more trustworthy
- **Persistent Identifiers**: globally unique name assigned to a digital object that can be used in perpetuity, to refer to and to retrieve the digital object
- CNRI Handle System
Persistent Identifiers

Please use this identifier to cite or link to this item: http://hdl.handle.net/2263/5103

Title: Eustachian tube diverticulum chondroids and neck abscessation in a case of Streptococcus equi subsp. equi

Inquiries: ann.carstens@up.ac.za

Author/s: Furniss, C. Carstens, A. Cilliers, I.

LC Subjects: Streptococcus equi

Keywords: Carrier Chondroids Equine Streptococcus equi subsp. equi

Issue Date: Sep 2007

Publisher: South African Veterinary Association

Citation: Furniss, C, Carstens, A & Cilliers, I 2007, 'Eustachian tube diverticulum chondroids and neck abscessation in a case of Streptococcus equi subsp. equi', Journal of the South African Veterinary Association
NON-EXCLUSIVE DISTRIBUTION LICENSE

In order for UPSpace to reproduce, translate and distribute your submission worldwide your agreement to the following terms is necessary.

By submitting this license, you (the owner of the rights) grants to the University of Pretoria the non-exclusive right to reproduce, translate (as defined below), and/or distribute your submission (including the abstract) worldwide in print and electronic format and in any medium, including but not limited to audio or video.

You agree that the University of Pretoria may, without changing the content, translate the submission to any medium or format for the purpose of preservation.

You also agree that the University of Pretoria may keep more than one copy of this submission for purposes of security, back-up and preservation.

You represent that you own the original work, and that you have the right to grant the rights contained in this license. You also represent that your submission does not, to the best of your knowledge, infringe upon anyone's copyright.

If the submission contains material for which you do not hold copyright, you represent that you have obtained the unrestricted permission of the copyright owner to grant the University of Pretoria the rights required by this license, and that such third-party owned material is clearly identified and acknowledged within the text or content of the submission.

If the submission is based upon work that has been sponsored or supported by an agency or organization other than the University of Pretoria, you represent that you have fulfilled any rights of review or other obligations required by such contract or agreement.

The University of Pretoria will clearly identify your name(s) as the submitter of the submission, and will not make any alteration, other than as allowed by this license, to your submission.

All items in the UPSpace collection are subject to copyright. For more information on the SA Copyright, visit the SA Copyright Act No. 98 of 1978 (as amended) available at http://www.buya.co.za/publications/cyberlaw/CopyrightAct.htm

I Grant the License

I Do Not Grant a License
Deposit Licence

Permissions given to the repository

The IR shall distribute electronic copies of the work for the lifetime of the repository, or based upon an agreed time span, and translate it as necessary to ensure it can be read by computer systems in the future.

The Repository’s Rights and Responsibilities

The IR may electronically store, translate, copy, or re-arrange the work to ensure its future preservation and accessibility, unless notified by the depositor that specific restrictions apply.

Storage Management

• Storage hardware is a key component of a repository
• SAN (Storage Area Network) vs NAS (Network Attached Storage)
  – Increased scalability: up to 16 million devices can be added
  – All other participants on SAN can connect and see each other
  – High-speed throughput: carry traffic between devices at 2 Gb/s
  – Independent of other network operations – functions separate from any LAN
Preservation Policies & Tools

- PADI – Digital Preservation Policies
- erpaTool – Digital Preservation Policy Tool
- Cornell Digital Preservation Tutorial
  [www.icpsr.umich.edu/dpm/dpm-eng/contents.html](http://www.icpsr.umich.edu/dpm/dpm-eng/contents.html)
- DRAMBORA [http://www.repositoryaudit.eu](http://www.repositoryaudit.eu)
  Digital Repository Audit Method Based on Risk Assessment Toolkit
Institutional Repository Workshop

A to Z of digital preservation within an Institutional Repository
Business Plans, Policies, Digitization, Metadata, Implementation,
Marketing & Buy-in and many more …

1 – 3 October 2008
University of Pretoria

www.library.up.ac.za/irtoolbox/workshop.htm

OR

E-mail:
ria.groenewald@up.ac.za
ina.smith@up.ac.za
Join our IRSpace CoP!

E-mail us:

ina.smith@up.ac.za
Will your work withstand the test of times to come?

Questions?
Sources

11. NISO. http://www.niso.org/kfile_download?pt=RkGKiXzW643YеЩaUaYUqZ1BFwDhIG4-24RJboZBWg8uE4vWdpZsJDs4RjLz0t90_d5_ymGsjiKVa86hjP37r_hM9t9qad1BrORLqssvegis%3D. Retrieved May 2008
13. PLANETS http://www.planets-project.eu/