

The use of metadata and preservation methods for continuous access to digital data



African Digital Scholarship
and Curation
12-14 May 2009

Amelia Breytenbach and Ria Groenewald

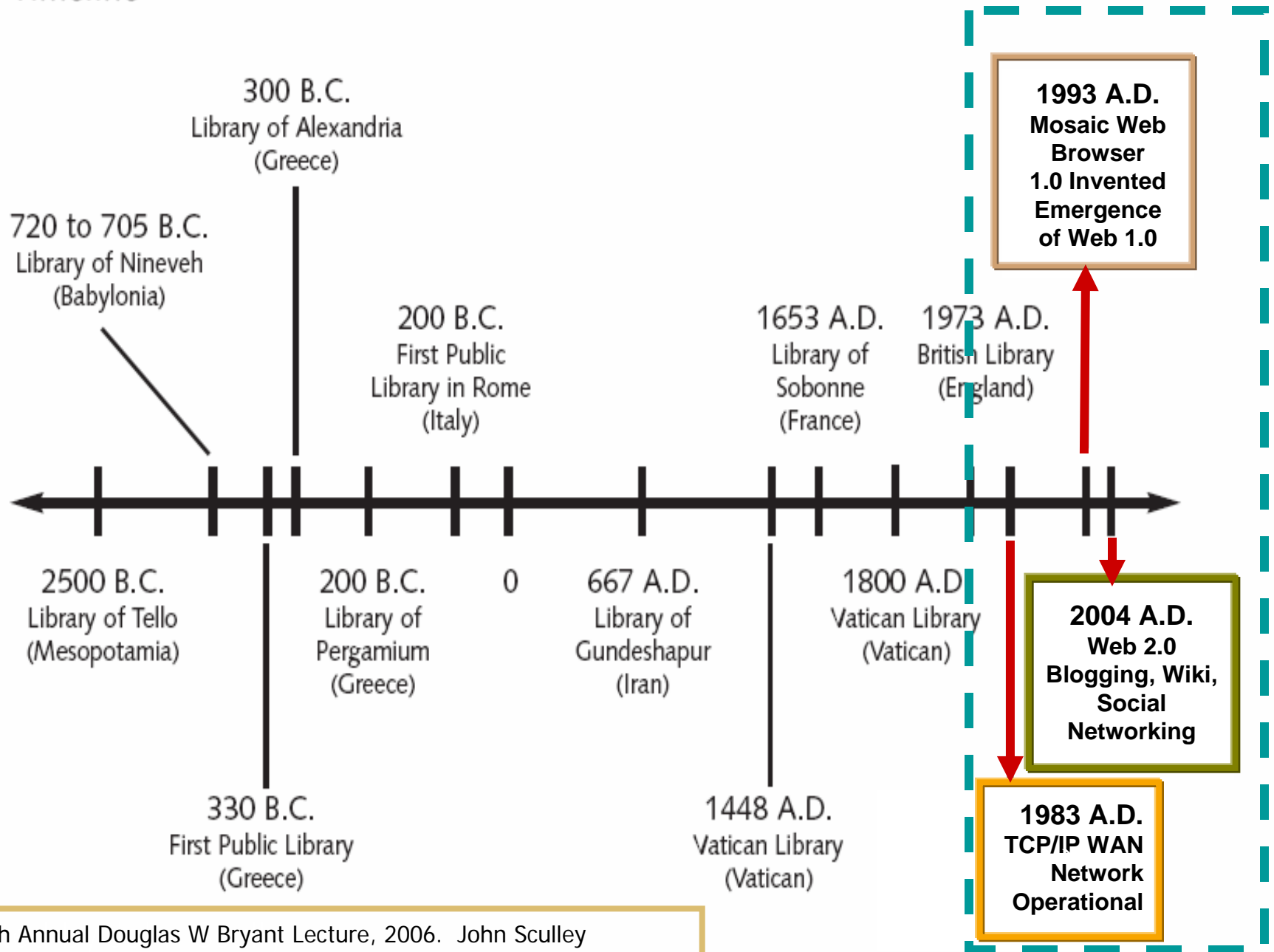
Department of Library Services

University of Pretoria



UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA

Timeline



2001, September 11

Terrorist attacks in the USA

http://portal.unesco.org/ci/en/ev.php_URL_ID=3618&URL_DO=DO_TOPIC&URL_SECTION=201.htm



http://www.cathousechat.com/cathouse_chat/WindowsLiveWriter/TwinTowers911.jpg

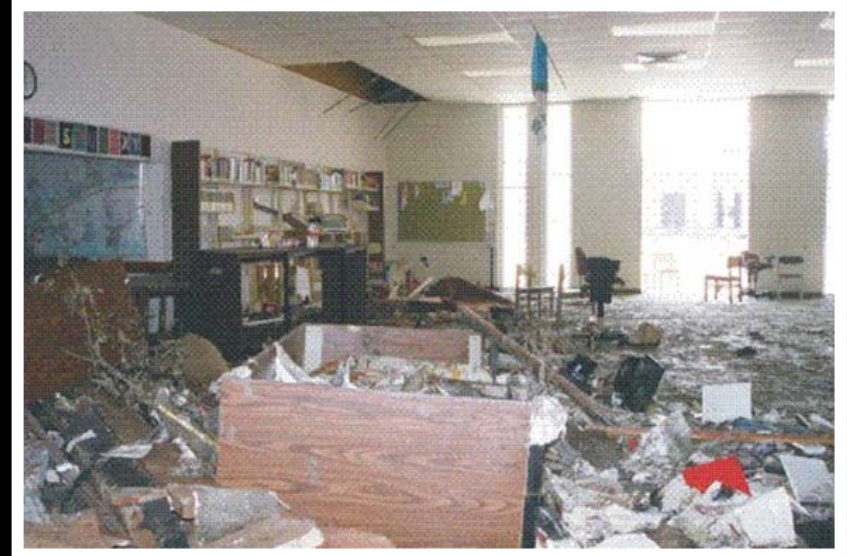
http://www.alarmingnews.com/archives/Twin-Towers-Reflected_1.jpg

www.arsenalofhypocrisy.com/.../image015.jpg

2005, August

Hurricane Katrina

- http://www.regent.edu/general/library/about_the_library/news_publications/images/katrina1.png
- http://msnbcmedia2.msn.com/j/msnbc/Components/Photos/060622/060622_library02_hmed_6p.h2.jpg
- <http://www.ala.org/ala/online/currentnews/newsarchive/2005abc/September2005abc/katrina10.htm>



Digital preservation



- Digital preservation
 - is a term used for storage and the
 - ongoing action to
 - protect digitally born or digitally created material

Safeguarding digital material

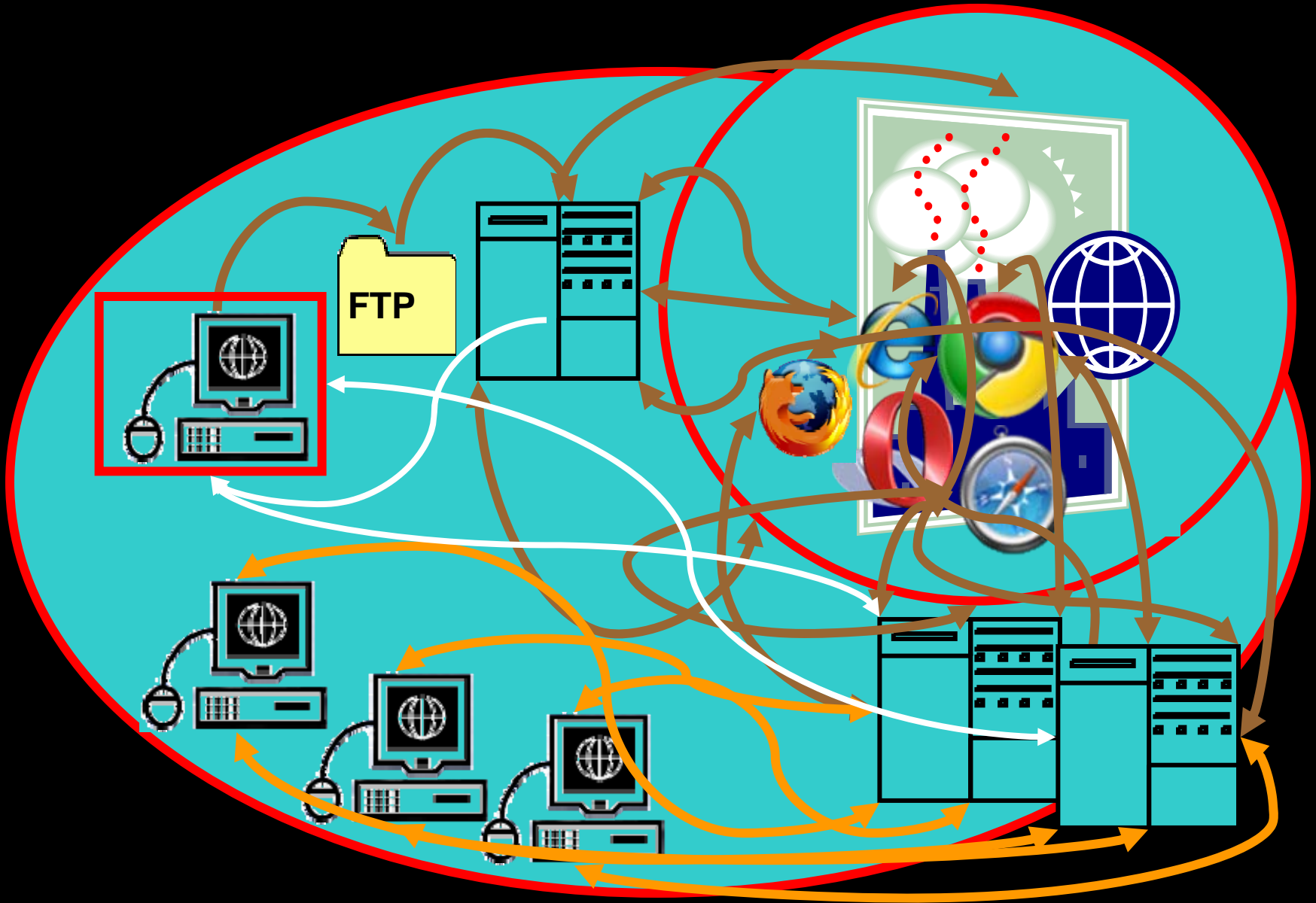
- For future access and usage
- Maintenance in the form of refreshing
- Migration of data to a different format to ensure its immunity from hardware and software obsolescence
- Enabling data to be accessed and reviewed by interesting parties over a wide range of space and time

Why preserve?



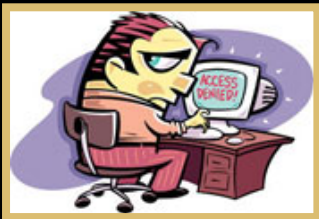
- Fragile digital documents
- The records of the entire present period of history are at risk
- To ensure the readability and interpretation of digital objects, preservation actions need to be taken at creation
- You don't know you have a gap in your digital records until you try to review something

Why preserve? Chaos of the 'structured web'



Risk analysis for digital objects

- Hard drive failure
- URL error – linked broken
- Storage medium failure
- Loss of information/data
- Human error and memory
- Hackers



www.fotosearch.com



cwa0023 www.fotosearch.com

Digital preservation - Questionnaire

- Questionnaire circulated on email list serves
- Lack of knowledge about preservation methods
- Indication of digital material stored on personal computers
- Format types – no unique formats used
- Awareness and training needed
- Marketing material should be created



Preservation strategy



- A preservation strategy is needed to safeguard digital content for future access
- A true preservation strategy must put planned-out business rules behind storage migration
- It should address
 - the risk factors involved
 - the actions needed for digital preservation
 - estimated time to do so
 - responsible entities

Actions required for digital preservation



- Store multiple copies
- Characterize and validate
- Ensure data remains authentic, reliable and usable
- Data cleaning, assigning preservation metadata and representation information
- Ensuring acceptable data structures or file formats
- Allocating unique persistent identifiers with comprehensive metadata
- Develop and execute preservation plans
- Implement a comprehensive technology watch mechanism
- Develop or acquire tools for preservation actions

Successful digital preservation



- Longevity
- Interoperability
- Total cost of ownership
- Technology obsolescence protection
- Back-up and recovery support

MoSCoW approach



- **M** : Things you/institution must preserve
- **S** : Things you should preserve, if at all possible
- **C** : Things you could preserve, if it does not affect anything else
- **W** : Things you won't preserve

Digital curation

- Digital curation includes data archiving and digital preservation as well as active management and appraisal of data over the life-cycle of scientific interest
- Effective data curation is a requisite component of multi-scale integration and re-use

Heidorn, P.B., et al.

Standards

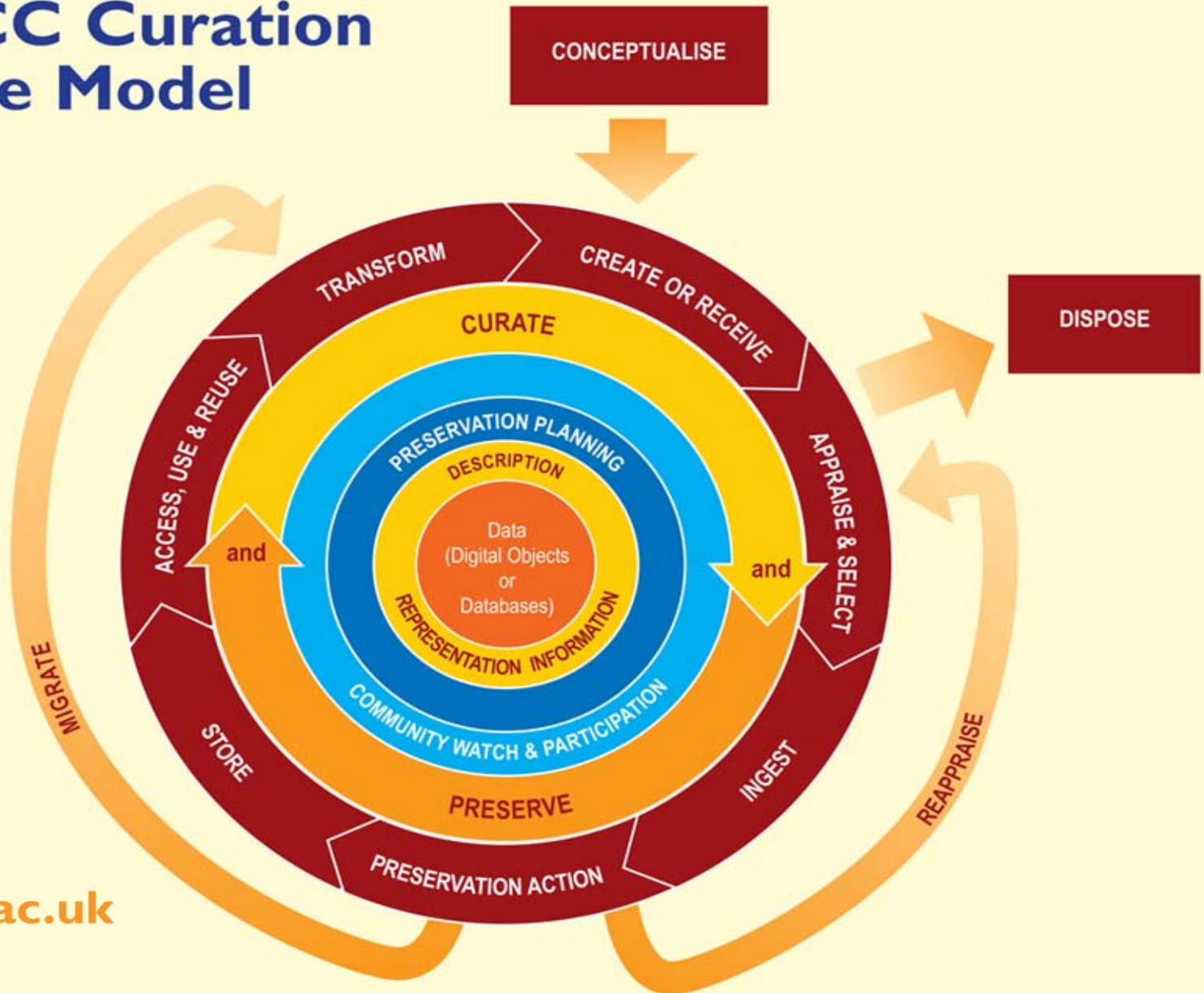
- OAIS (Model widely adopted as starting point in digital preservation efforts)
(http://en.wikipedia.org/wiki/Open_Archival_Information_System)
- PREMIS (Preservation metadata)
(<http://www.oclc.org/research/projects/pmwg/premis-final.pdf>)
- ISO standards - Z39.87 (Technical metadata)

<http://www.lockss.org/lockss/OAIS>





The DCC Curation Lifecycle Model

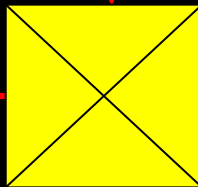


PREMIS MODEL



Intellectual entity (photo)

Converted to digital object



TIFF image file

Rights = Object -
instructed user what
it represent

Transform to JPEG
for web display



Preserve for
interoperability,
access and readability

Agent:

- The role of the person undertaking the event (name/organization)
- Software name and version no.
- OS type

Object:

- File size
- Date created
- File format
- Creating application

Rights:

- License agreement
- Exact permissions granted over preservation of the object

Definition of metadata

Conventional metadata

- Is data about a digital object
- Explain the technical method of creation and administrative data
- Contain the descriptive information

Preservation metadata

- Include data to authenticate the provenance of a digital object; and
- Contain a complete dataset to ensure the future usage of a digital object

Preservation metadata categories

Preservation Description Information (PDI)

- Reference
- Context
- Provenance
- Fixity

Creators of preservation metadata

- ❑ Creators of digital resources
- ❑ Digitization projects
- ❑ Fixity information



+

Descriptive
and
Preservation
Metadata

=



Object becomes useful and can be preserved for future usage

Preservation

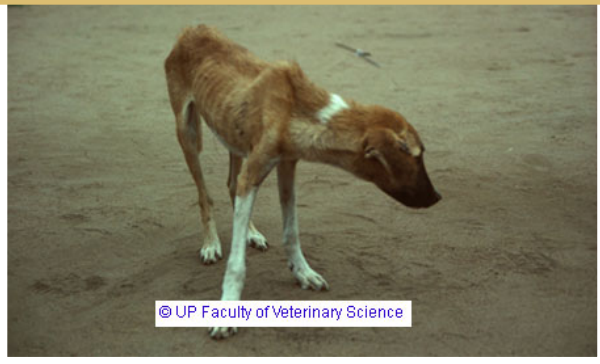
Description:

Colour photos.
Original scanned
size: 26.29 kb

Photo 2: Original scanned size: 221 kb JPEG, 600 dpi. Final web-ready size: 24.86 kb. Estimate download time: 10 sec. @ 28.8 kbps. Original TIFF file housed at the Dept. Veterinary Tropical Diseases, University of Pretoria. Metadata assigned by Prof. R.C. Tustin, Professor Emeritus: DVTD. His academic and professional experience includes: veterinarian for 54 years, senior lecturer at UP for 7 years, head of Department at UP for 17 years and Veterinary Council for 3 years.

Description Provenance:

Original scanned in at 600dpi 100% TWAIN scanning program using a EPSON 1640XL scanner. Downsized to 400 pixels in width, resolution 300dpi. Web version done automatically by PhotoShop 7 Software. Downloading time shift between 4 to 7 seconds. Date done April 2005.



authorized users
Edit Profile

Help
About UPSpace

Description:
Colour photos. Photo 1: Original document size: (w)7 x (h)4.64 cm. Original scanned size: 274 kb JPEG, 600 dpi. Final web-ready size: 26.29 kb. Estimate download time: 10 sec. @ 28.8 kbps. Photo 2: Original document size: (w)7 x (h)4.62 cm. Original scanned size: 221 kb JPEG, 600 dpi. Final web-ready size: 24.86 kb. Estimate download time: 10 sec. @ 28.8 kbps. Original TIFF file housed at the Dept. Veterinary Tropical Diseases, University of Pretoria. Metadata assigned by Prof. R.C. Tustin, Professor Emeritus: DVTD. His academic and professional experience includes: veterinarian for 54 years, senior lecturer at UP for 7 years, head of Department at UP for 17 years and Veterinary Council for 3 years.

More Info Available Online: <http://en.wikipedia.org/wiki/Nagana>
URI: <http://hdl.handle.net/2263/4802>

Rights: ©University of Pretoria. Dept of Veterinary Tropical Diseases (Original and digital) Provided for educational purposes only. It may not be downloaded, reproduced or distributed in any format without written permission of the original copyright holder. Any attempt to circumvent the access controls placed on this file is a violation of copyright laws and is subject to criminal prosecution. Please contact the collection administrator for copyright issues.

Type: Image
Language:
Appears in Collections: [Faculty of Veterinary Science, University of Pretoria](#)

Files in This Item:

Need for Preservation Metadata

- Identifies the
- Determines w
- Details the co
- Puts the reco
- Provides tech
- Provides know
- and how it wa

Submitted by Amelia Breytenbach

(abreyten@op.up.ac.za) on 2007-10-11T07:47:42Z

No. of bitstreams: 6

AT_fotoalbum6.jpg: 18703 bytes, checksum:
b34593b272cb8fc200ade463859ed82d (MD5)

AT_fotoalbum5.jpg: 17862 bytes, checksum:
ad287e77f2a71ffaa8eeb45acf7bfd27 (MD5)

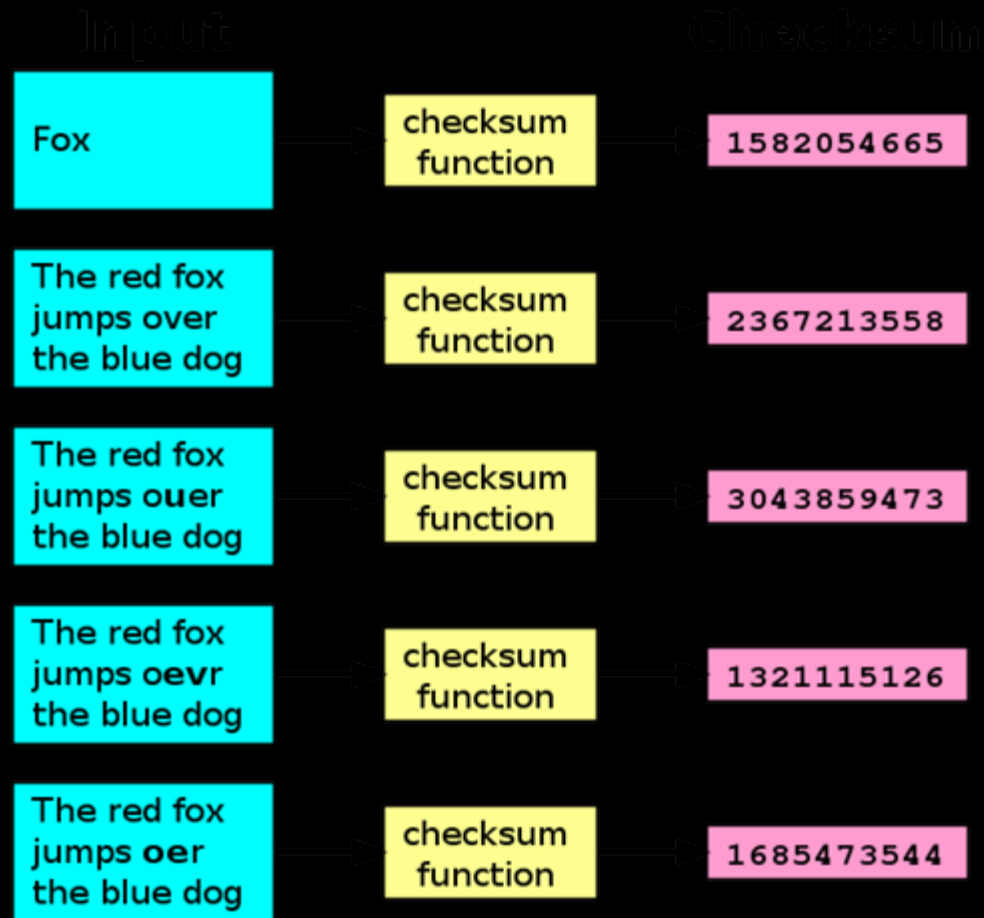
AT_fotoalbum4.jpg: 17339 bytes, checksum:
4ce10ea57074f0b69daf1ad64923ac72 (MD5)

AT_fotoalbum3.jpg: 21651 bytes, checksum:
9b2d12d4355caa0a60033bd4e2d2b9a9 (MD5)

AT_fotoalbum2.jpg: 14891 bytes, checksum:
ab557dffa30175fde5f6ccf35b1ea690 (MD5)

AT_fotoalbum1.jpg: 12779 bytes, checksum:
fefefe9f90df44fd483d7fe8322755f7 (MD5)

Effect of a checksum function



Example of a digital document with metadata

The African Elephant: a digital collection of anatomical sketches as part of the University of Pretoria's Institutional Repository - a case study

Authors:

Amelia Breytenbach; Metadata specialist
Department of Library Services
University of Pretoria, South Africa
Email: amelia.breytenbach@up.ac.za

Ria Groenewald; Digitization Coordinator
Department of Library Services
University of Pretoria, South Africa
Email: ria.groenewald@up.ac.za

Abstract

Purpose – Although several collections have been digitized and made available in the University of Pretoria's Institutional Repository, a pilot study has not been done to measure the project management and workflow. The collections available in the repository at the time of this project were all long-term projects. There was a need to identify a project small enough to conform to normal project management requirements to use as an example to establish the planning and workflow of future projects. The purpose was to determine the outcome and quality of the final web-ready institutional repository product against specific digitization project goals.

Design/methodology/approach – A collection of anatomical sketches in the custody of the Faculty of Veterinary Science, Department of Anatomy and Physiology was identified as a possible collection that could comply with the above criteria. The different sketches in the Elephant collection could be digitized in phases, making it an ideal project for future comparison. In each phase a number of tasks were identified which the various role players should complete during the workflow process. Each phase would be compared to the previous completed phases to measure the outcomes and progress made in quality and time. Through successful interaction and collaboration between the Library and the Department of Anatomy and Physiology during the digitization process, valuable tacit knowledge could be preserved for future use in the field of Veterinary Science.

Findings – The completed project delivered on key areas such as the electronic availability of the collection through metadata description. Basic preservation of the physical collection was undertaken as necessary and the physical as well as the digital collections were archived for future use. The conclusion will describe the lessons learned and how it can be applied in future projects to the advantage of the institution.

Practical implications – The paper provides a very useful case study for other academic libraries who want to develop their own digital collections.

Originality/value – This paper offers practical help to libraries starting with digitization. It supplies valuable information for project management, planning of workflow and estimate time frames for completing a specific task in the digitization process.

Article Type: Case study

Example of metadata
accompanying a
MSWord document

Example of metadata
accompanying a PDF-
document

Example of metadata created for a digitally born, MSWord document

Document Title	The African Elephant: a digital collection of anatomical sketches as part of the University of Pretoria's Institutional Repository - a case study	
Authors	Breytenbach, Amelia and Groenewald, Ria	
Description	Although several collections have been digitised and made available in the University of Pretoria's Institutional Repository, a pilot study has not been done to measure the project management and workflow. The collections available in the repository at the time of this project were all long-term projects. There was a need to identify a project small enough to conform to normal project management requirements to use as an example to establish the planning and workflow of future projects. This paper offers practical help to libraries starting with digitisation, it supplies valuable information for project management, planning of workflow and estimate time frames for completing a specific task in the digitization process.	
Date created	2007/09/28 -	
Rights	The authors. Document can be migrated for future usage.	
Type	Article	
Access	Own use <input type="checkbox"/>	<input type="checkbox"/> Social network
	Journal <input type="checkbox"/>	<input type="checkbox"/> Repository
Format	MS Word 2003 (.doc)	
Format extent	3.62 MB (3,796,480 bytes)	
File name	2007_gro_bre	
Language	English	
Keywords	Digital storage ; Collections management ; University libraries ; Anatomical drawings ; South Africa	
<u>Document History</u>		
Version	Date	Comments
1	2007/09/28	Document created by authors
2	2007/11/20	Document edited by authors
3	2007/11/30	Final edit and submission to Journal



Example of embedded metadata created for a transformed document to PDF format

FinalUPSpace.pdf

Description
Advanced

Description

Document Title: The African Elephant: a digital collection of anatomic.

Author: Breytenbach A.; Groenewald R.

Author Title:

Description: Although several collections have been digitized and made available in the University of Pretoria's Institutional Repository, a pilot study has not

Description Writer:

Keywords: Digitization; Metadata; Digital collections; Preservation; Anatomical sketches

Commas can be used to separate keywords

Copyright Status: Copyrighted

Copyright Notice: Emerald Group Publishing Limited

Copyright Info URL: www.emeraldinsight.com/1065-075X.htm




Go To URL...

Created: 2009/03/31 09:03:21 AM
Modified: 2009/03/31 09:27:53 AM
Application: Acrobat PDFMaker 9.0 for Word
Format: application/pdf

Powered By **xmp**

OK Cancel

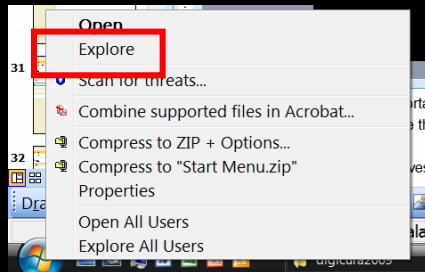
Side-car spreadsheet containing metadata of a digital collection

Elephant	Date scan	Size of original canvas (cm) (w x h)	dpi	scanned sized (w x h) - KB	%	Date Conv	format	(w) x (h) pixels	ppi	Photo shop 8 - work on image	format	size in pixels (w x h)	General info (Dimensions w x h)	ppi	Save for web	Download information	Adobe Acrobat	Password	Description
oli_001	16-Mar-07	35.5 x 25.5	610	8520 x 6400 pixels 156M	100	24-03-07	jpeg	840 x 525	300	Mode: RGB 8-bits channels Grag Scale Clean canvas: white area Crop Rotate canvas: 0.4 mm - cw	gif	550 x 344	9.31 cm x 5.83 cm 100% dither; selective pallett, 256 colours	150	✓	87.93 k 32 sec @ 28.8 kbps	Standard Size 86 kb		oli_001: Medial view of elephant skull, juvenile 
oli_002	16-03-07	35.5 x 25.5	610	8424 x 6352 pixels 153.1M	100	24-03-07	jpeg	840 x 548	300	Mode: RGB 8-bits channels Rotate canvas: 0.4 mm - cw Crop Clean canvas: white area	gif	550 x 359	9.31 cm x 6.08 cm 100% dither; selective pallett, 256 colours	150	✓	83.28 k 31 sec @ 28.8 kbps	6.0 Standard		oli_002: Lateral view of elephant skull, juvenile 
oli_003	16-03-07	35.5 x 26.0	610	8640 x 6048 pixels 149.5M	100	24-03-07	jpeg	840 x 638	300	Mode: RGB 8-bits channels Crop Clean canvas: white area	gif	550 x 421	9.31 cm x 6.08 cm 100% dither; selective pallett, 256 colours	150	✓	111.8 k 41 sec @ 28.8 kbps	6.0 Standard		oli_003: Caudal view of elephant skull, juvenile 
oli_004	16-03-07	41.0 x 28.0	610	8640 x 6656 pixels 164.6 M	100	24-03-07	jpeg	840 x 643	300	Mode: RGB 8-bits channels Crop Clean canvas: white area	gif	550 x 421	9.31 cm x 6.08 cm 100% dither; selective pallett, 256 colours	150	✓	107.5 k 39 sec @ 28.8 kbps	6.0 Standard		oli_004: Ventral view of elephant skull, juvenile 
oli_005	16-03-07	35.5 x 26.0	610	6224 x 7728 pixels	100	06-04-07	jpeg	840 x 916	300	Crop Mode: RGB 8-bits channels Clean canvas: white area Rotate 90° CW	gif	550 x 420	9.31 cm x 7.12 cm 100% dither; selective pallett, 256 colours	150	✓	101.2 k 37 sec @ 28.8 kbps	6.0 Standard Size 103 kb		oli_005: Dorsal view of elephant skull, juvenile 

Search functions

- Important role in the retrieval of electronic documents
- MicroSoft Explorer search tool
- The CDS (Copernic Desktop Search) * tool was found to be valuable for desktop searching

*<http://www.copernic.com>



Search Clear

African elephant case study ▼ ➔

The Web All Emails 10 Files Music Pictures Videos Organizer Contacts Favorites History 1

Refine Clear

Name: ▼

Type: ▼

Size: ▼

Date: ▼

Folder: ▼

18 matching documents - [More on the Web](#) Standard | Folder

hogsproceed.doc	My Documents\Goats\	2002/02/27 11:07:12 AM
Folder: My Documents\HICSA\		
HICSA_1d.ppt	My Documents\HICSA\	2008/11/19 01:06:24 PM
HICSA_finaal.ppt	My Documents\HICSA\	2008/11/19 01:19:09 PM
Folder: My Documents\Metadata\Artikel_elephant\		
OCLC_article-Final-1.doc	My Documents\Metadata\Artikel_...	2007/11/30 01:26:28 PM
OCLC_article-UPSpace.doc	My Documents\Metadata\Artikel_...	2009/03/25 01:19:28 PM
OCLC article.doc	My Documents\Metadata\Artikel_...	2007/06/27 01:02:36 PM
Folder: My Documents\Metadata\Artikel_elephant\OCLC\		
Breytenbach_African(2008).pdf	My Documents\Metadata\Artikel_...	2009/03/31 09:27:54 AM
FinalUPSpace.doc	My Documents\Metadata\Artikel_...	2009/03/31 08:52:26 AM
OCLC_article-Final.doc	My Documents\Metadata\Artikel_...	2007/11/30 11:08:01 AM

OCLC_article-Final.doc
2007/11/30 11:08:01 AM

Find: African elephant case study

OCLC_article-Final.doc

Type: Microsoft Word Document
 Size: 3.62 MB
 Date: 2007/11/30 11:08:01 AM
 Author: UP User
 Folder: My Documents\Metadata\Artikel_elephant\O

The African Elephant: a digital collection of anatomical sketches as part of the University of Pretoria's Institutional Repository - a case study

Authors:
 Amelia Breytenbach : Metadata specialist
 Department of Library Services
 University of Pretoria, South Africa
 Email: amelia.breytenbach@up.ac.za

Search [Clear](#)

elephant

Refine [Clear](#)

Name:

Dimensions:

File Type:

Location:

Folder:

Skip images smaller than 32x32

- The Web
- All
- Emails 17
- Files 150
- Music
- Pictures**
- Videos
- Organizer
- Contacts
- Favorites 1
- History 4

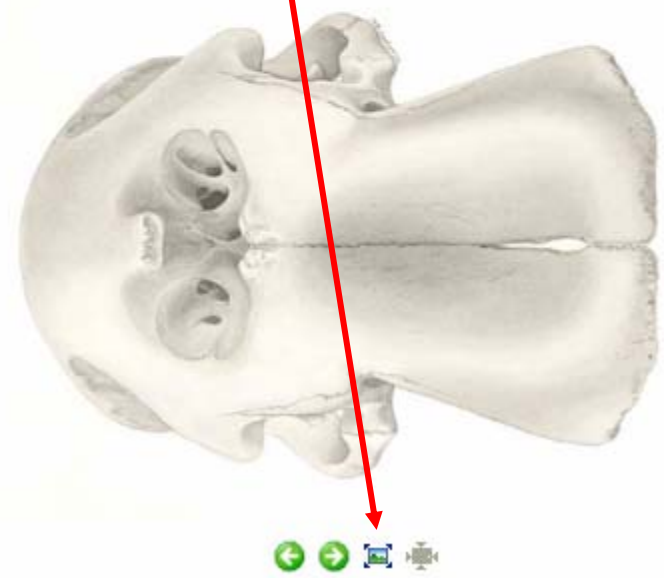
8 matching documents - [More on the Web](#) Thumbnails | 2 Folder

- Folder: My Documents\SANVR Slideshow\
 - sanvr-eleph...
 - sanvr-eleph...
 - sanvr-eleph...
 - sanvr-eleph...
- Folder: My Documents\up_ovi webpage\images\
 - ...

sanvr-elephant_2.jpg My Documents\SANVR Slideshow\ 19 KB

.jpg
2009/04/06 09:38:12 AM

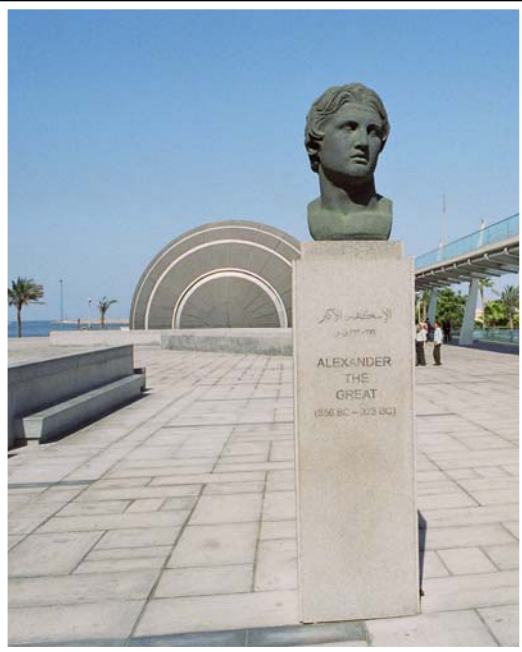
Find: Elephant Open Open Folder



Research done

Library of Alexandria

Wellcome Library



Preservation Tools



- Content Management System, i.e. Joomla
- Archival software - Alfresco
- Format Preservation - Xena
- Web 2.0 - Web Curator Tool
- Search Facilities – Copernic and Windows Explorer
- Back-up systems
- Spreadsheets or inter-active list containing separate metadata
- DCC digital object life-cycle
- Standardization

Conclusion



- Preservation awareness in SA
- Negligence on format specifications and standardisation
- Storage and preservation of digital data
- Knowledge of preservation methods should be promoted
- Training in the preservation of digital content and the actual delivery of plans and policies
- Metadata, as a consistent, logical manner to keep documents accessible and usable

Technical developments

- The technical challenges, like the perceptual issues, will be overcome with time
- Bandwidth will increase
- Processors will get faster
- Memory will continue to grow
- The two major technical hurdles that need to be addressed are -
 - the tools that are available for digital preservation
 - interoperability



Amelia Breytenbach
Metadata Specialist
Veterinary Library
University of Pretoria
Email:

amelia.breytenbach@up.ac.za

Tel: 012 x 529-8391



Ria Groenewald
Digitization Coordinator
Merensky Library
University of Pretoria
Email:

ria.groenewald@up.ac.za

Tel: 012 x 420-3792

References

- Katrina http://waynecountyreadingcouncil.com/yahoo_site_admin/assets/images/katrina_library.236181047_std.jpg
- John Sculley – 11th Annual Douglas W Bryant Lecture under the auspices of the Eccles Centre for American Studies: 2006 BL <http://www.bl.uk/eccles/bryant2006.html> or <http://www.bl.uk/eccles/pdf/dwbryant/2006dwb.pdf>
- Audit of Digitization Initiatives in South Africa – Request for Information. Unpublished document
- Digital Insurance for Information at Risk. A Strategic Overview of Digital Preservation. H. Andrew Lawrence, Worldwide Product Marketing Manager Document Imaging, Eastman Kodak Company. <http://digitalpreswhitepaper.pdf>
- 61st IFLA General Conference - Conference Proceedings - August 20-25, 1995. Libraries Are Not for Burning: International Librarianship and the Recovery of the Destroyed Heritage of Bosnia and Herzegovina. Andras Riedlmayer, Harvard University
- An Introduction to Digital Preservation (TASI: <http://www.tasi.ac.uk/advice/delivering/digpres.html>)
- June 2002 Report by the OCLC/RLG Working Group: <http://www.oclc.org/research/pmwg/>
- PREMIS (<http://www.oclc.org/research/projects/pmwg/premis-final.pdf>)
- Heidorn, P.B., et al. Data Curation Education and Biological Information Specialists. University of Illinois, Urbana-Champaign <http://www.digitalpreservationeurope.eu/what-is-digital-preservation/>
- Paynter, G.; Joe, S.; Lala, V.; Lee, G. (2008) 'A Year of Selective Web Archiving with the Web Curator at the National Library of New Zealand', D-Lib Magazine, vol. 14, nr. 5/6
<<http://www.dlib.org/dlib/may08/paynter/05paynter.html>>
- McKnight, D. (2003) *DPI: The Digital Preservation Imperative*, Power Point Presentation: Access 2003 Conference, October 2, 2003, Vancouver, BC
- Caplan, P. (2009) *Understanding PREMIS*, Library of Congress Network Development and MARC Standards Office. (www.loc.gov/standards/premis/understanding-premis.pdf)
- University of London Computer Centre; UKOLN; JISC (2008) *PoWR: The preservation of web resources handbook*. <<http://jiscpower.jiscinvolve.org/handbook>>
- Day, M. (2005) *DDC/Digital curation manual instalment on metadata*. HATII, University of Glasgow; University of Edinburgh; UKOLN, University of Bath; Council for the Central Laboratory of the Research Councils. <<http://www.dcc.ac.uk/resource/curation-manual/chapters/metadata>>
- OAIS model. Paradigm project, *Workbook on Digital Private Papers*, 2005-7 <<http://www.paradigm.ac.uk/workbook>> [April 2009].