Trust and the Web:

Can the audit criteria apply to Web Archives?

Gerard Clifton
Manager, Digital Preservation
National Library of Australia

APSR Forum on Long-Term Repositories
National Library of Australia, 31 August – 1 September, 2006

Overview

- Approaches to Web archiving
- Users & uses
- Data collection & management
- TDR compliance issues for Web
- Ways forward?

Aims of Web Archiving

- Collect & preserve online documentary heritage
- Usually by National Libraries or Archives
  - Examples:
    - PANDORA – National Library of Australia & 9 other partner agencies
    - MINERVA – Library of Congress, USA
    - Kulturarw[3] – National Library of Sweden
    - WARP (Web Archiving Project) – National Diet Library, Japan
    - WebArchiv – National Library of the Czech Republic
    - Bibliothèque nationale de France
    - Groups: Nordic Web Archive, UK Web Archiving Consortium
      International Internet Preservation Consortium (IIPC)
    - Internet Archive

More information – PADI – Web archiving topic: http://www.nla.gov.au/padi/topics/92.html

Approaches to Web Archiving

- Comprehensive ('Whole Domain')
  - Whole domain snapshots
  - Large volumes, automated, low QA
    - *Examples*: Internet Archive, Kulturarw[3] (Sweden)

- Selective
  - Focused, selected harvests, high QA
    - Documents, publications, sites
    - *Examples*: PANDORA (NLA), UK Web Archiving Consortium
      MINERVA (LoC) (Thematic)

- Combined
  - Mix - comprehensive, continuous (10%), selective, thematic
    - Bibliothèque nationale de France

Users & Uses of Web Archives

- No ‘typical user’
    - Anyone with a Web browser & access
- Uses[*]
    - General uses
    - Evidence for civil or criminal cases
    - Patent searches for prior art
    - Researchers
        - Historians (of technology, Internet)
        - Data mining (specialist)

(* IIPC use cases - http://netpreserve.org/publications/iipc-r-003.pdf)

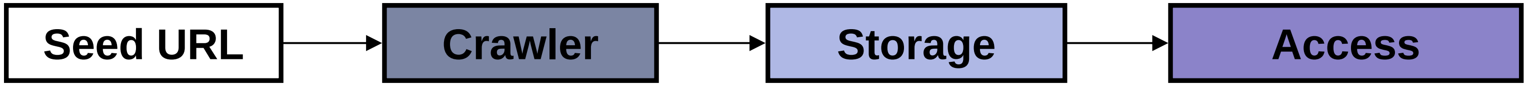
Users & Uses of Web Archives

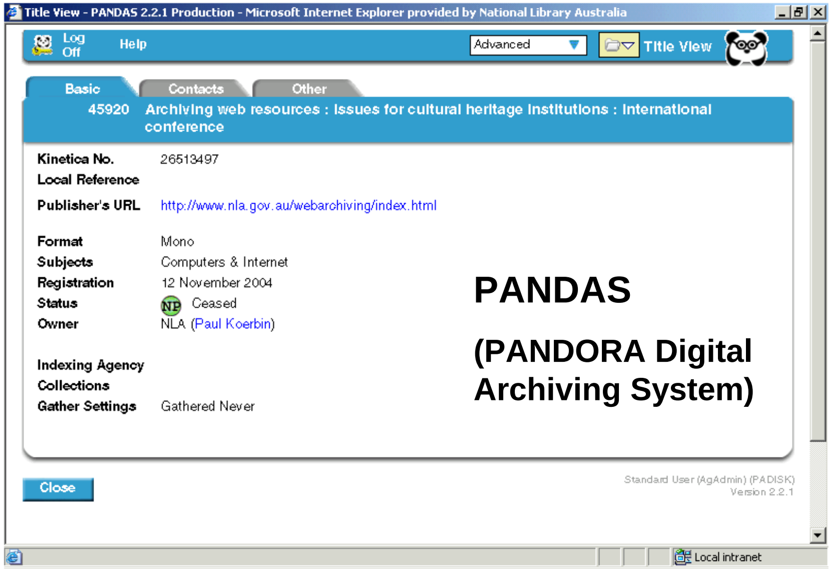
- General uses
  - Finding things that have disappeared from the live Web
    - PANDORA:
      - First families 2001 (http://nla.gov.au/nla.arc-10421)
      - Sydney Olympics (http://nla.gov.au/nla.arc-10194)
  - Finding things that have changed
  - Persistent citation
    - Indexing agencies

User Expectations

- Stability
    - Persistence of identifiers etc.
- Authentic reflection of what was…
    - Time / Date snapshot
    - Separation
    - Completeness
    - (degree of) Functionality
- …Availability into the future

Data Collection & Management




Title View - PANDAS 2.2.1 Production - Microsoft Internet Explorer provided by National Library Australia
Log Off
Help
Advanced
Title View
Basic
Contacts
Other
Close
Local intranet

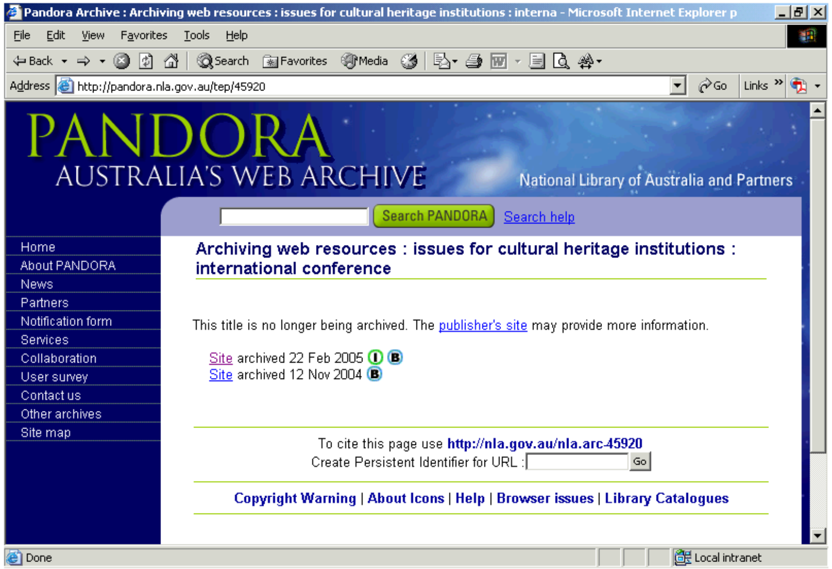
45920 Archiving web resources : issues for cultural heritage institutions : international conference

Kinetica No.	26513497
Local Reference	
Publisher's URL	http://www.nla.gov.au/webarchiving/index.html
Format	Mono
Subjects	Computers & Internet
Registration	12 November 2004
Status	NP Ceased
Owner	NLA (Paul Koerbin)
Indexing Agency	
Collections	
Gather Settings	Gathered Never

Standard User (AgAdmin) (PADISK)
Version 2.2.1

PANDAS

(PANDORA Digital Archiving System)

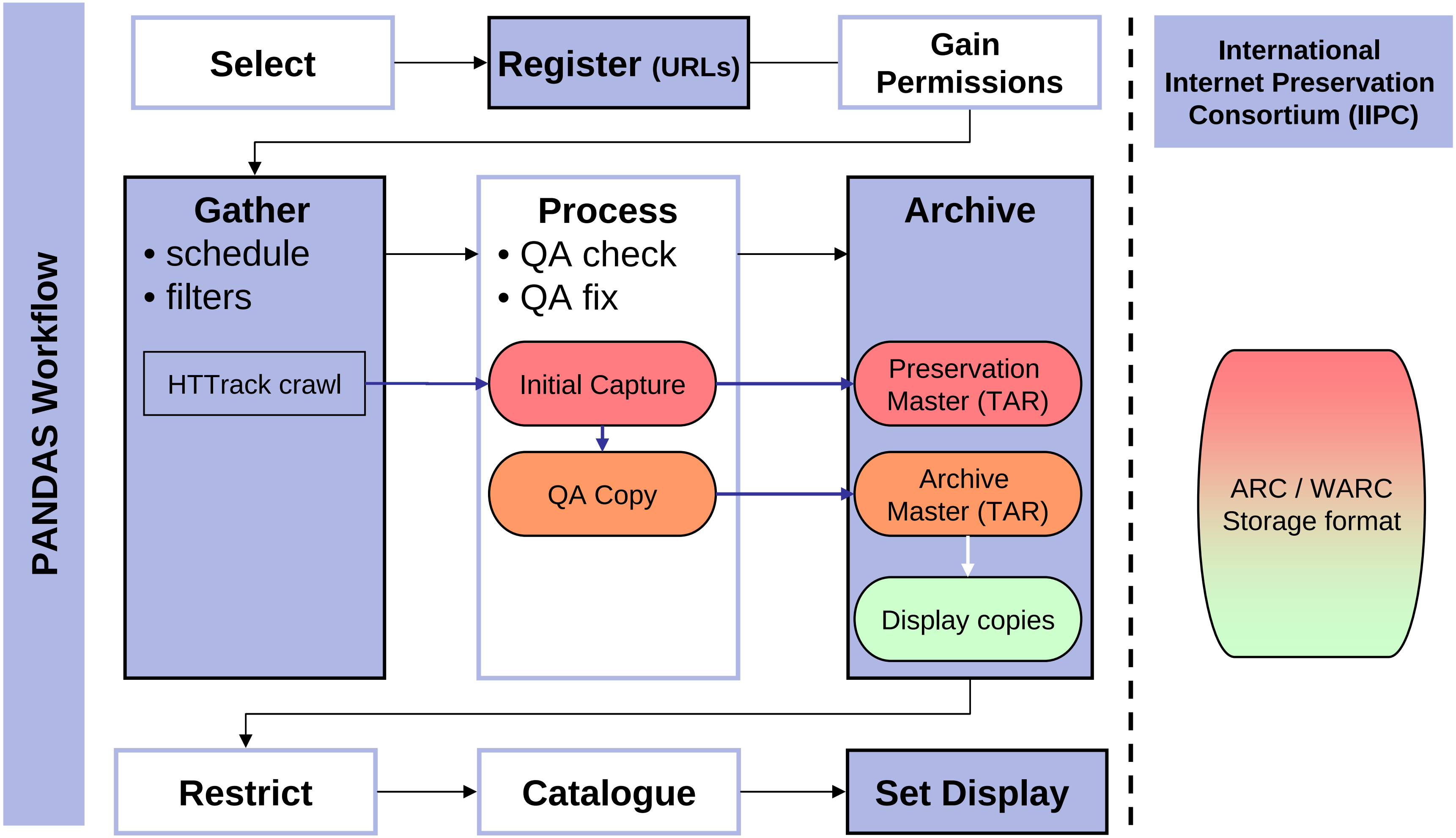

Pandora Archive : Archiving web resources : issues for cultural heritage institutions : interna - Microsoft Internet Explorer p
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://pandora.nla.gov.au/tep/45920
Go Links
PANDORA
AUSTRALIA'S WEB ARCHIVE
National Library of Australia and Partners
Search PANDORA
Search help
Home
About PANDORA
News
Partners
Notification form
Services
Collaboration
User survey
Contact us
Other archives
Site map
Copyright Warning | About Icons | Help | Browser issues | Library Catalogues
Done
Local intranet

Archiving web resources : issues for cultural heritage institutions : international conference

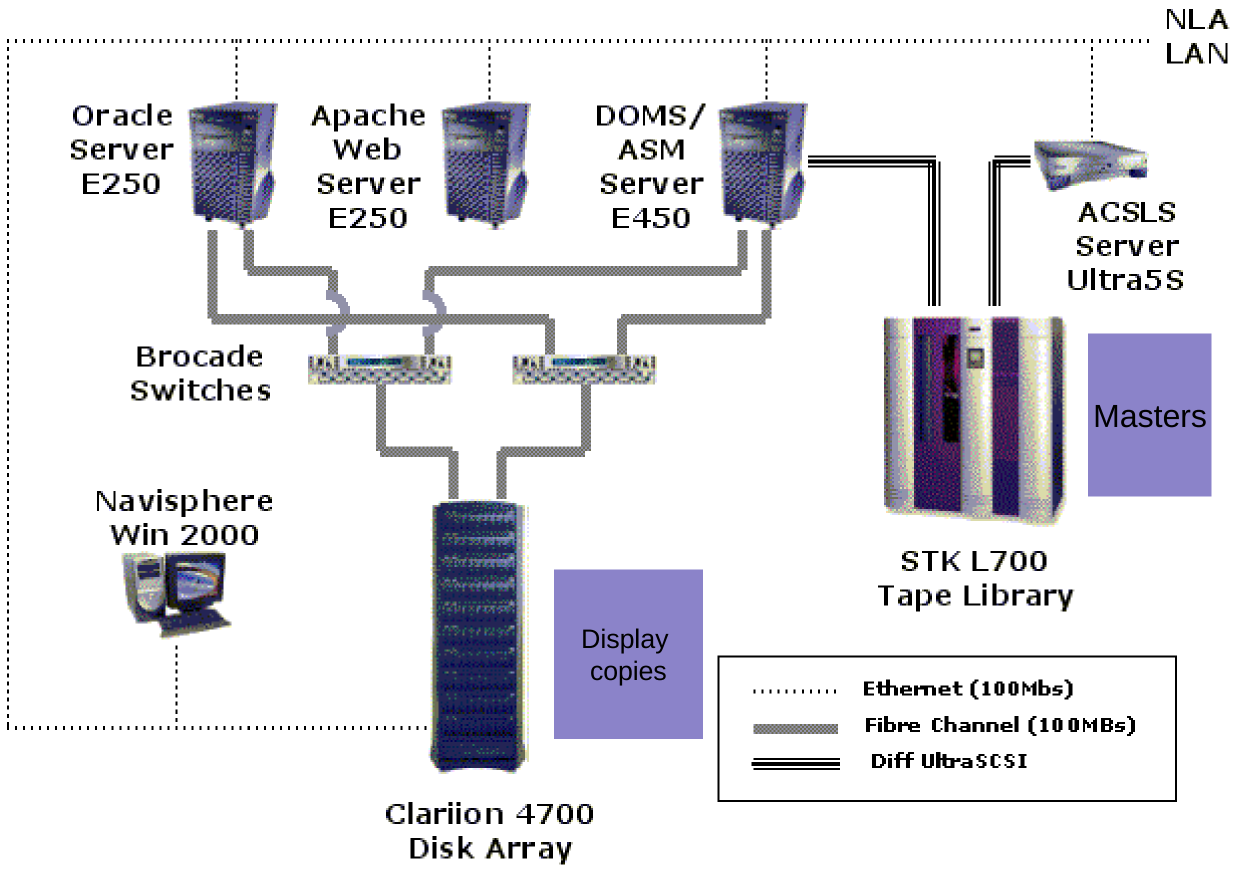
This title is no longer being archived. The publisher's site may provide more information.

Site archived 22 Feb 2005 I B
Site archived 12 Nov 2004 B

To cite this page use http://nla.gov.au/nla.arc-45920
Create Persistent Identifier for URL : Go


PANDAS Workflow
Select
Register (URLs)
Gain Permissions
International Internet Preservation Consortium (IIPC)
Gather
• schedule
• filters
HTTrack crawl
Process
• QA check
• QA fix
Initial Capture
QA Copy
Archive
Preservation Master (TAR)
Archive Master (TAR)
Display copies
ARC / WARC Storage format
Restrict
Catalogue
Set Display

DOSS System Architecture



Digital Object Management

- Administration
  - Management of works, copies, relationships, metadata
- Data Management
  - Redundant storage and backup
  - Refreshment cycles
  - Restrictions on access
  - User authentication
- Delivery
  - Persistent citation and access
  - Online delivery

The Audit Checklist for TDR

A. Organisation

1. Governance & viability
2. Structure & staffing
3. Procedural accountability & policy framework
4. Financial sustainability
5. Contracts, licenses, liabilities

B. Functions, Processes

1. Ingest / Content acquisition
2. Archival storage & management
3. Preservation planning
4. Data management
5. Access management

C. Designated community

1. Documentation
2. Appropriate descriptive metadata
3. Use and usability
4. Verifying understandability

D. Technical infrastructure

1. System infrastructure
2. Appropriate technologies
3. Security

The Audit Checklist for TDR

A. Organisation

1. Governance & viability
2. Structure & staffing
3. Procedural accountability & policy framework
4. Financial sustainability
5. Contracts, licenses, liabilities

B. Functions, Processes

1. Ingest / Content acquisition
2. Archival storage & management
3. Preservation planning
4. Data management
5. Access management

C. Designated community

1. Documentation
2. Appropriate descriptive metadata
3. Use and usability
4. Verifying understandability

D. Technical infrastructure

1. System infrastructure
2. Appropriate technologies
3. Security

TDR – Issues for Compliance

- Flexibility in interpretation
  - Level of granularity has large effects for compliance
- Web archives don't follow deposit model
  - Agreements don't always fit
- Complexity & volume – makes compliance difficult for some criteria
  - Ingest verification, metadata collection
  - Preservation process demonstration
- 'Designate community' not easily defined
  - Affects demonstrations of understandability

TDR – Issues for Compliance

- A5.1. Appropriate deposit agreements

- Rights, responsibilities, expectations
- Mainly for third-party preservation
- Can be less formal
- Conditions should be notified to all depositors

- A5.2 Agreements specify preservation rights

- Written policies & agreements transferring preservation permission to repository
- Acceptable to ingest, then follow up later

TDR – Issues for Compliance

- A5.1 Appropriate deposit agreements
- A5.2 Agreements specify preservation rights

- Harvest model, especially for comprehensive, may not include agreements

Possible remedies

- Post statements of responsibility etc. for central access
- Send automated notifications at time of crawl

TDR – Issues for Compliance

- B1.3 Written definition for each SIP (& AIP)

- Written inventory of agreement specifies what is transferred

- B1.6 Verify SIP for completeness & correctness

- Completeness of data transfer (no truncation)
- Complete set of material

- Correctness of files transferred – received what was expected

TDR – Issues for Compliance

- B1.3 Written definition for each SIP (& AIP)
- B1.6 Verify SIP for completeness & correctness

- Difficult to specify 'boundary' or full set of files (esp. if no contact)
- Harvest = crawler view – cannot know what you don't have
- May be items that are not crawlable (Flash, JavaScript, DBs)
- Web servers not always accurate – MIMEs misreported

Possible remedies

- Definitions of SIP / AIP
  - Classes, sites, pages, files
  - Acceptable 'generic' specifications of what is 'complete'
- Metadata collection during crawl – transactions, checksums
- Further development of tools for analysis & verification

TDR – Issues for Compliance

- B1.5 Sufficient physical control of objects

- Analysis of digital content
- Verification, analysis and metadata creation
  - Detailed technical metadata
- AIP creation & association with metadata

TDR – Issues for Compliance

- B1.5 Sufficient physical control of objects

- Level of detail required may not be possible for large heterogeneous collections
    - Limitations of current tools
    - Too labour intensive for manual creation

Possible remedies

- Definitions of SIP / AIP
- Tools may verify & analyse >95% of materials (HTML, JPEG, GIF)
- Target additional formats for tools
- Web metadata set – collection during crawl
- AIP creation - WARC format includes metadata with content

TDR – Issues for Compliance

- B3. Preservation planning & strategies

- Level of detail required difficult for large heterogeneous collections

Possible remedies

- Definitions of SIP / AIP – reduce scope
- Event recording – logs etc.
- PANIC / AONS automated monitoring against registries

TDR – Issues for Compliance

- C1.1 Definition of Designated Community

- ‘General user’
  - ‘General English-reading public educated to high school and above, with access to a Web browser (HTML 4.0 capable)’

- C4 Verify understandability

- Documented process for testing understandability to Designated Community

- Verification of testing

TDR – Issues for Compliance

- C1.1 Definition of Designated Community
- C4 Verify understandability

- General user – broad group, limited contact
- Heterogeneous material - what extent needs to be verified as understandable? How many tests?

Possible remedies

- Central definition & commitment etc. statements
- Reasonable definition of test scope
  - (e.g. range of browsers, range of material)
- Representative testers

Moving Forward

- Define scope for Web SIP / AIPs
- Recast criteria for Web archives – reduce uncertainties about compliance
- Levels for compliance?
- Improve tools, metadata collection
- Find the middle ground