

APSR and the National Information Infrastructure Questnet 2004



Towards an Australian Partnership for
Sustainable Repositories



The APSR Program

Is a response to:

- the burgeoning growth in digitised publications and resources supporting research and teaching;
- the challenge of creating a repository fabric for those materials given their diversity, complexity and primacy and the impact of cost, timing, rights and access arrangements;
- the need to preserve materials over their valued 'life-times'.

APSR Mission

The Partnership seeks to:

- foster digital continuity and sustainability practices wrt a national data repository fabric;
- maintain strong international linkages and collaborations;
- establish a demonstrator program within partner organisations;
- establish a Sustainable Digital Collections Centre to run partnership programs;
- coordinate and integrate activities with related infrastructure initiatives.

Presentation Structure – the APSR Context

1. Research Information Infrastructure
 - eResearch - new but enduring
 - engagement with research methods
 - publishers all – discovery and dissemination
 - institutional planning – funding - organisation
2. Data, information & scholarship
 - whose information – what quality?
 - sustainability challenges
 - APAC data-grid program
3. Repository technologies - services
 - repositories – collections – functions
 - pilots – trials (ANU)
4. APSR - structure/objectives
 - Sustained, valued digital collections
 - sector-wide development

e-Research – A new paradigm

The invention and exploitation of advanced IT

To generate, curate and analyse research data

- From experiments, observations and simulations
- Quality management, preservation and reliable evidence

To develop and explore models and simulations

- Computation and data at extreme scales
- Trustworthy, economic, timely and relevant results

To enable dynamic, distributed virtual organisations

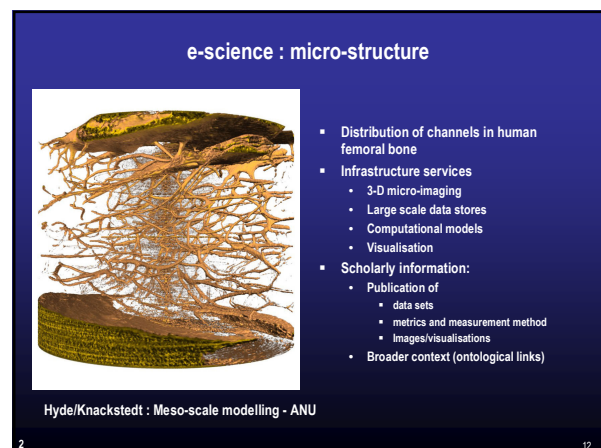
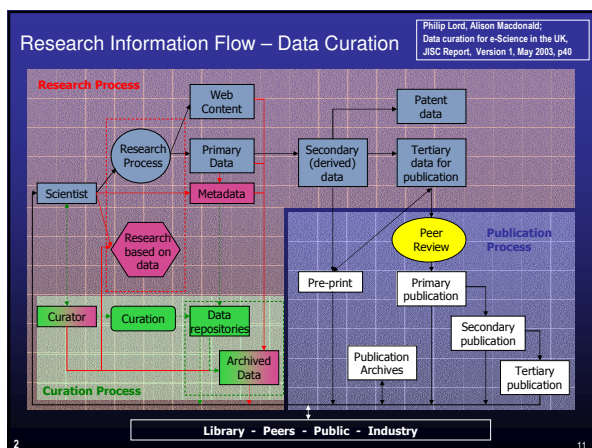
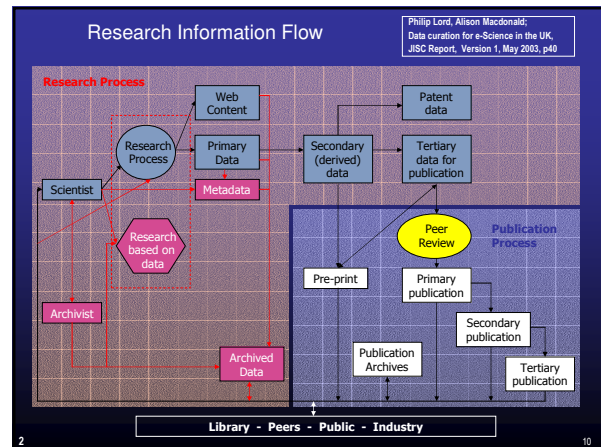
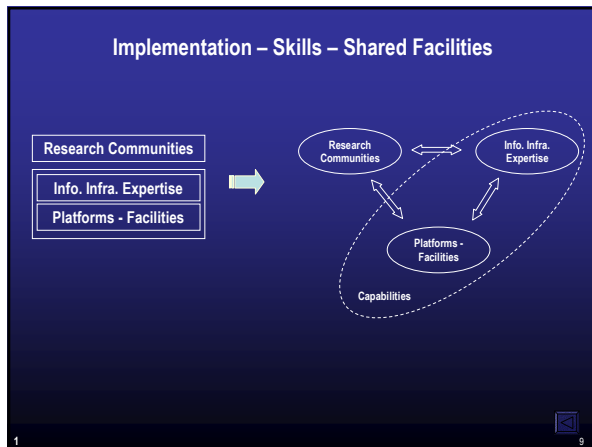
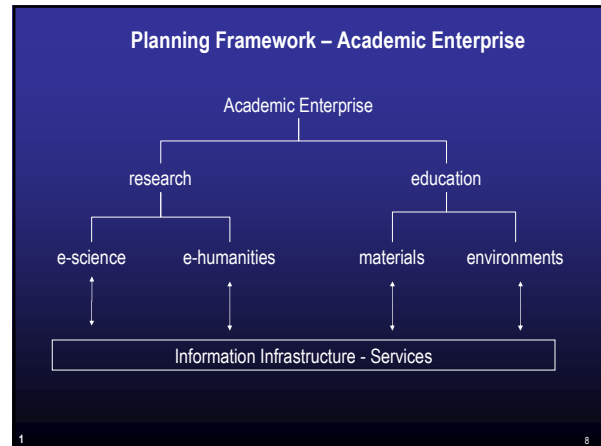
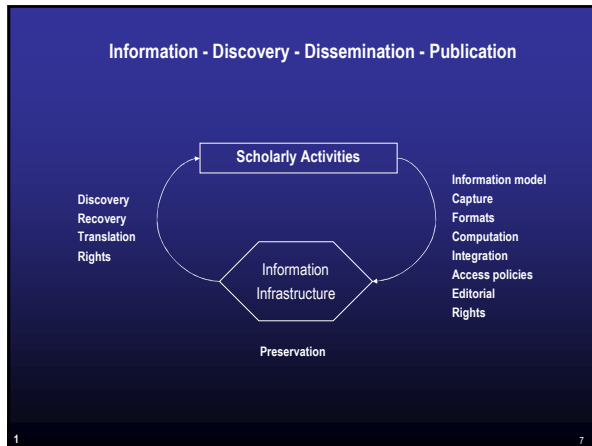
- Facilitating collaboration with information and resource sharing
- Security, reliability, accountability, manageability and agility

[Tony Hey – Dir. UK e-Science]

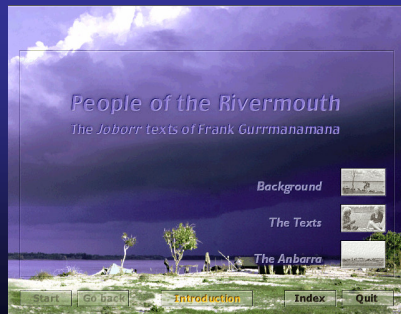
Community Declarations (open access)

- ... We define open access as a comprehensive source of human knowledge and cultural heritage that has been approved by the scientific community.
- ... the future Web has to be sustainable, interactive, and transparent. Content and software tools must be openly accessible and compatible.
- ... Open access contributions include original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material.
- Open access contributions complete version of the work and all supplemental materials, ... in an appropriate standard electronic format is deposited (and thus published) in at least one online repository using suitable technical standards ... supported and maintained by an academic institution, scholarly society, government agency ... and long-term archiving.

Berlin Declaration



Typical Challenge : People of the Rivermouth



Sponsored by
 • National Museum of Aust.
 • ANU
 • AIATSIS
 • AASA
 • Bawinanga Aboriginal Corporation

Hiatt/McKenzie/... (ANU ...)

2

13

APAC (National) Data Grids

- Potentially large scale archives
 - 10's-100TB
- Complex, non-bibliographic data
 - Astronomy, Linguistics, Particle Physics, Arts, Bioinformatics, GIS, ...
- Derivative products
 - Browse data, rights limited, ...
- Network performance critical
 - Delivery to multiple computing, visualisation and collaboration resources

2

14

Example: PARADISEC

- Linguistics archive of unique and endangered material from Asia-Pacific
 - 6000 languages in the world today
 - Only a few hundred left at the end of this century
 - 2500 in Asia Pacific
 - 800 in Papua New Guinea
- Sustainability is critical
 - Many old analogue tapes, many TB of data
- Complex rights management
 - Cultural constraints (eg. gender-based)
 - Time varying (speakers die, period of mourning)
- Complex and rich metadata

2

15

Example: Virtual Observatory

- Astronomical data
 - Much now born digital
- Complex and various data formats
 - Images, spectra, multi-wavelength, ...
- Highly federated
 - At various observatories, universities, ...
- GIS-like coordinate alignment issues
- Rich data fusion requirements

2

16

Institutional Repositories



Preservation of digital scholarship 1994 - 2004

3

17

Institutional Setting (typical)

- Recognition of digital scholarship
- Preservation problems
 - Legal
 - Administrative
 - Technical
 - Philosophical
- Devolved structure
- No culture of curation of collections
- Long-term stewardship of the output of 'non-institutional' funded research projects
- Policy on retention of research data – (between 5 and 15 years)

3

18

Implementation setting (ANU)

- Born-digital scholarly material
 - Research data, scholarly writing, learning objects
- Digitised collections
 - Macho1, Paradesic, ePrints, Artsserve, Chinese Digital Archive 1966-1976, Historic Photographic Collections, ...
- Proliferation of systems
 - ePrints, CONTENTdm, 4D, Filemaker pro, MS Access, mysql, php ...
- Need for broad spectrum repositories
- Acquisition
 - Specific purpose application - custom built
 - Open source
 - DSpace, Fedora, i-TOR, MyCoRe, ePrints, CDSware

3

19

About Repository Platforms (Dspace)

- Open-source software development - sponsored by MIT & Hewlett-Packard.
- Low barrier to submission.
- Captures
 - Digital research material (any format)
 - Directly from creators (e.g. faculty)
 - Large-scale, stable, managed long-term storage
- Describes
 - Descriptive, technical, rights metadata
 - Persistent identifiers
- Distributes
 - Via WWW, with necessary access control
- Preserves

3

20

DSpace Information Model

- Communities
 - Research units of the organisation
- Collections (in communities)
 - Distinct groupings of like items
- Items (in collections)
 - Logical content objects
- Receive persistent identifier
 - Bitstreams (in items)
- Individual files
 - Receive preservation treatment

MacKenzie Smith, Associate Director for Technology MIT Libraries
DSpace: An Open Source Institutional Digital Repository

3

21

Long-term accessibility and preservation

- Enable discovery
- Persistent identifier scheme (CNRI Handles)
- Curate content for easy retrieval
- Ensure integrity of data (md5 checksum)
- Encourage use of preservable file formats
- Migrate the content to new formats if necessary
- Collection Policies - Levels
 1. Collect metadata only
 2. Save binary files and metadata
 3. Preserve content but not look & feel
 4. Look & feel intact
 5. Complete

"If you want it to be collected, build it to be collectable"

Supporting Digital Scholarship, University of
Virginia's Institute for Advanced Technology, 2004

3

22

DSpace format support

- Known/supported
 - TIFF, SGML/XML, AIFF, PDF, TXT
- Known/unsupported
 - Microsoft Word, PowerPoint (common, proprietary)
 - Lotus 1-2-3, Visicalc, WordPerfect (less common)
- Unknown/unsupported
 - One-of-a-kind software program

3

23

DSpace Federation

- Governance of the code
 - An initial group of system architects and developers
 - Copyright and licensing of code & DSpace trademark
- Foster new and innovative uses of the platform
- Institutional repository forum
 - Preservation
 - Access
 - Advocacy, marketing, assessment
 - SIGs

3

24

Strategic Alliances

- **Strategic Alliances**
 - Global Digital Format Registry
 - SIMILE
 - CNRI Handles
 - Creative Commons
 - Google
- **eg Google pilot project**
 - Google, OCLC & DSpace community
 - Improve access to scholarly material using Google's search technology
 - Mechanism not yet developed however goal to enable new services based on harvested scholarly collections

3

25

ANU DSpace Demonstrators (Advanced)

- **Image collection - tiff**
 - ANU Archives
 - Business and trade union activities in Australia
 - Early Canberra
- **Text collection - XML**
 - Anthropology thesis
- **Audio collection - wav**
 - Australian Anthology of Music

3

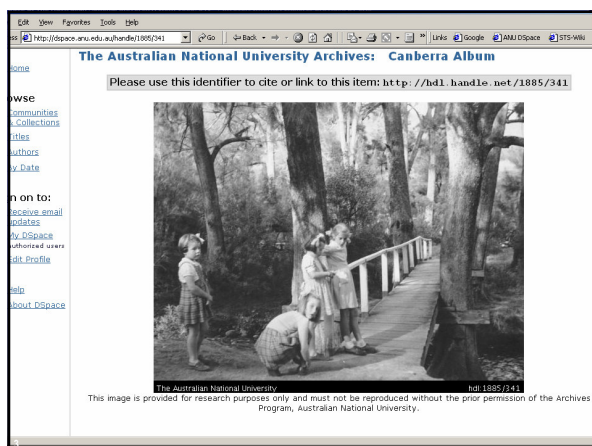
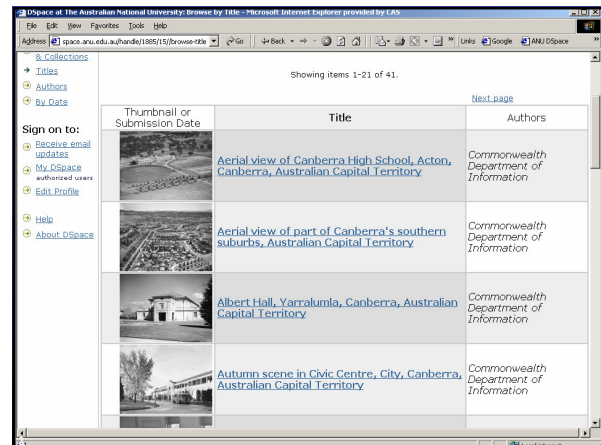
26

Image demonstrator

- **Images : high quality tiffs**
 - 600 dpi
 - 6000 px longest side
- **Auto-generation of derivatives - thumb and access image**
- **Contribute metadata to PictureAustralia**
 - Federated metadata repository
 - Harvested using OAI-PMH
- **Brand images – ANU, persistent identifier**
 - 20 px strip added to bottom of image
 - Text – ANU and persistent identifier
 - ImageMagick
- **Auto-generation**
 - Access image and thumb generated from tiff
 - Generated after deposit

3

27

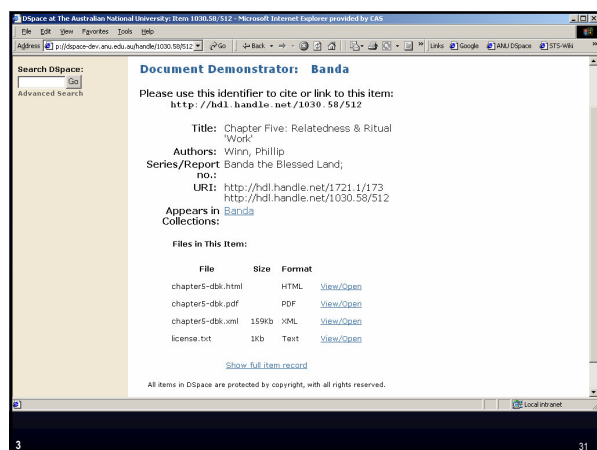


Text Demonstrator

- **Thesis**
 - Banda Islands in Central Maluku
 - MS Word format
 - Styles consistently applied
- **Conversion to XML**
 - Use Upcast V4.0.2
 - Convert to DocBook
 - XSLT to convert HTML and PDF
 - RDF - hierarchical information about the content
- **Upload into DSpace**
 - Thesis is a collection
 - Each chapter is an item
 - RDF an item in the collection
 - Images in the chapter bundle

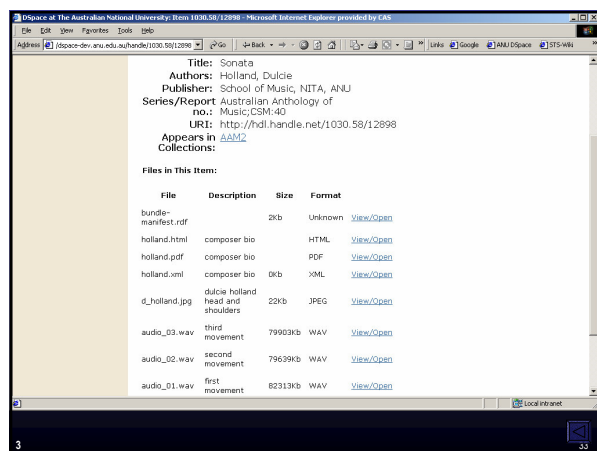
3

30



Audio Demonstrator

- Australian Anthology of Music
 - Created annually by the School of Music
 - Australian composers
- Structure
 - CD made up of a number of compositions
 - Composition made up of a number of movements
- Users views
 - CD
 - Composition
 - Composer
- CD at the collection level
- Items include:
 - Movements - .wav
 - Composer bio - XML
 - Image of composer - JPG
- RDF used to store complex relationships
 - Each composition has its own RDF file describing the components



The APSR Program

APSR Partners

Have institutional commitments to a national repository based infrastructure;
 Support an 'open partnership' framework linked to 'communities of interest' and focussed on long term goals

Partners [Initial]

- Australian Partnership for Advanced Computing
- The Australian National University
- The National Library of Australia
- The University of Sydney
- The University of Queensland

The Four Programs

- Core: Digital Continuity and Sustainability Program (NLA focus)
- Practices and Test-beds Program
 - ANU - Integrated repository framework
 - Usyd - E-humanities and continuity
 - UQ - eScholarship Australia
 - APAC - E-research data management
- International Linkages Program
- National Services Program

Finis

Engagement with research methods

Sustainability

Test-beds

Technologies

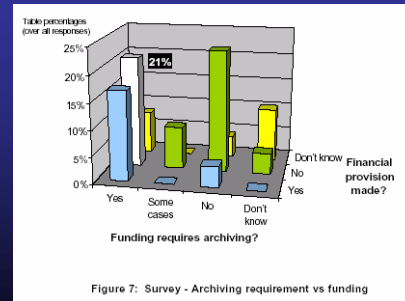
Australian Government
 Department of Education,
 Science and Training

Acknowledgements

- Markus Buchhorn - ANU
- Michele Huston - ANU
- Brian Molinari - ANU
- John O'Callaghan - APAC
- Scott Yeadon - ANU

37

Funding Issue



Philip Lord, Alison Macdonald;
Data curation for e-Science in the UK,
JISC Report, Version 1, May 2003, p37

38

Community Declarations (mission)

- The Internet has fundamentally changed the practical and economic realities of distributing scientific knowledge and cultural heritage. For the first time ever, the Internet now offers the chance to constitute a global and interactive representation of human knowledge, including cultural heritage and the guarantee of worldwide access.
- ... Obviously, these developments will be able to significantly modify the nature of scientific publishing as well as the existing system of quality assurance.
- ... to promote the Internet as a functional instrument for a global scientific knowledge base and human reflection and to specify measures which research policy makers, research institutions, funding agencies, libraries, archives and museums need to consider.

Berlin Declaration

1

39

APSR Context (Summary)

E-Research Data Management

Digital to digital research methods – digitised primary data
Trusted data availability over the life-time of its value

National Infrastructure Development

Systemic Infrastructure Funding:
Networks – AREN - ARENAC
APAC-2
Research Information Infrastructure - ARIIC

International Data-Grid Focus

Physics, astronomy, social sciences
Digital Library Programs (US)
Data Curation Centre (UK)

APAC-2 Data Grid

Uniform access to distributed large scale data sets
Establishing archives of animations from digital models
Searching/analysing distributed large scale research data sets wrt Col

4

40