Preserving E-Prints: Scaling the Preservation Mountain

Sheila Anderson, Arts and Humanities Data Service
Stephen Pinfield, University of Nottingham

SHERPA

- **Acronym:** Securing a Hybrid Environment for Research Preservation and Access
- **Initiator**: CURL (Consortium of University Research Libraries)
- **Development Partners**: Nottingham (lead), Edinburgh, Glasgow, Leeds, Oxford, Sheffield, York, British Library, AHDS
- **Duration**: 3 years, November 2002 – November 2005
- **Funding**: JISC and CURL
- **Programme**: FAIR (Focus on Access to Institutional Resources)
- **Aims**:
  - to construct a series of institutional OAI-compliant e-print repositories
  - to investigate key issues in populating and maintaining e-print repositories
  - to work with service providers to achieve acceptable standards and the dissemination of the content
  - to investigate standards-based digital preservation e-prints
  - to disseminate learning outcomes and advocacy materials

SHERPA

‘E-prints’

- ‘E-prints’ = a digital duplicate of an academic research paper that is made available online as a way of improving access to the paper
- Document types:
  - ‘pre-prints’ (pre-refereed papers)
  - ‘post-prints’ (post-refereed papers)
  - conference papers
  - book chapters etc.
- Formats:
  - PDF
  - HTML
  - TEX/LATEX etc.

ahds

SHERPA

“~~How~~ should we preserve e-prints?”

“Forget about OAIS for now! The OAI-compliance of the Eprint Archives is enough for now.”

Stevan Harnad[1]

“An OAI system that complied with the OAIS reference model, and which offered assurances of long-term accessibility, reliability, and integrity, would be a real benefit to scholarship.”

Peter Hirtle[2]

Sources:
1. Stevan Harnad, September98 forum, 13 February 2003
2. Peter Hirtle, *D-Lib Magazine* 7, 4, April 2001

ahds

An Institutional Repository

"….is a set of services that an institution 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 organisational commitment to the stewardship of these digital materials, including long-term preservation where appropriate, as well as organisation and access or distribution."

Lynch, C., ARL Bimonthly Report 226, http://www.arl.org/newsltr/226/ir.htm

ahds

SHERPA

Why preserve e-prints?

Possible reasons:

- Preserving (open) access
- Where e-prints are commonly cited
- Where e-prints contain / sit alongside more than the conventionally published paper
- Where they form part of a specific collection
- Guarantees of preservation may attract authors to submit papers

ahds

SHERPA

What needs to be done?*

* James, Ruusalepp, Anderson and Pinfield, “Feasibility and Requirements Study on the Preservation of E-prints” 2003

- Preservation planning and actions
- Recognise the preservation risks of file formats
- Adopt open, standards based file formats wherever possible
- Plan for migrating rare and obsolete file formats
- Collect administrative and preservation metadata
- Define e-print preservation metadata
- Develop e-print preservation infrastructure

ahds

SHERPA

What is being done now?

AHDS-SHERPA study:

- Properties of e-prints
- Collection policies and procedures
- Selection and retention criteria
- Preservation metadata
- Preferred formats
- Rights issues
- Organisational models
- Cost / funding models and open access

ahds

SHERPA

A Good Start …… but this isn't 'doing' preservation

- Preservation storage layer needs to be added
- Preservation planning needs to take place
- Preservation and administration metadata needed
- Preservation processes and protocols in place and ready for action

ahds

SHERPA

SHERPA DP Project

- **Acronym:** Securing a Hybrid Environment for Research Preservation and Access: Digital Preservation
- **Development Partners**: AHDS (Lead), Nottingham + 3-4 SHERPA Partners
- **Duration**: 2 years, November 2004 – October 2006
- **Funding**: JISC and CURL
- **Programme**: JISC Digital Preservation and Records Management Programme

ahds

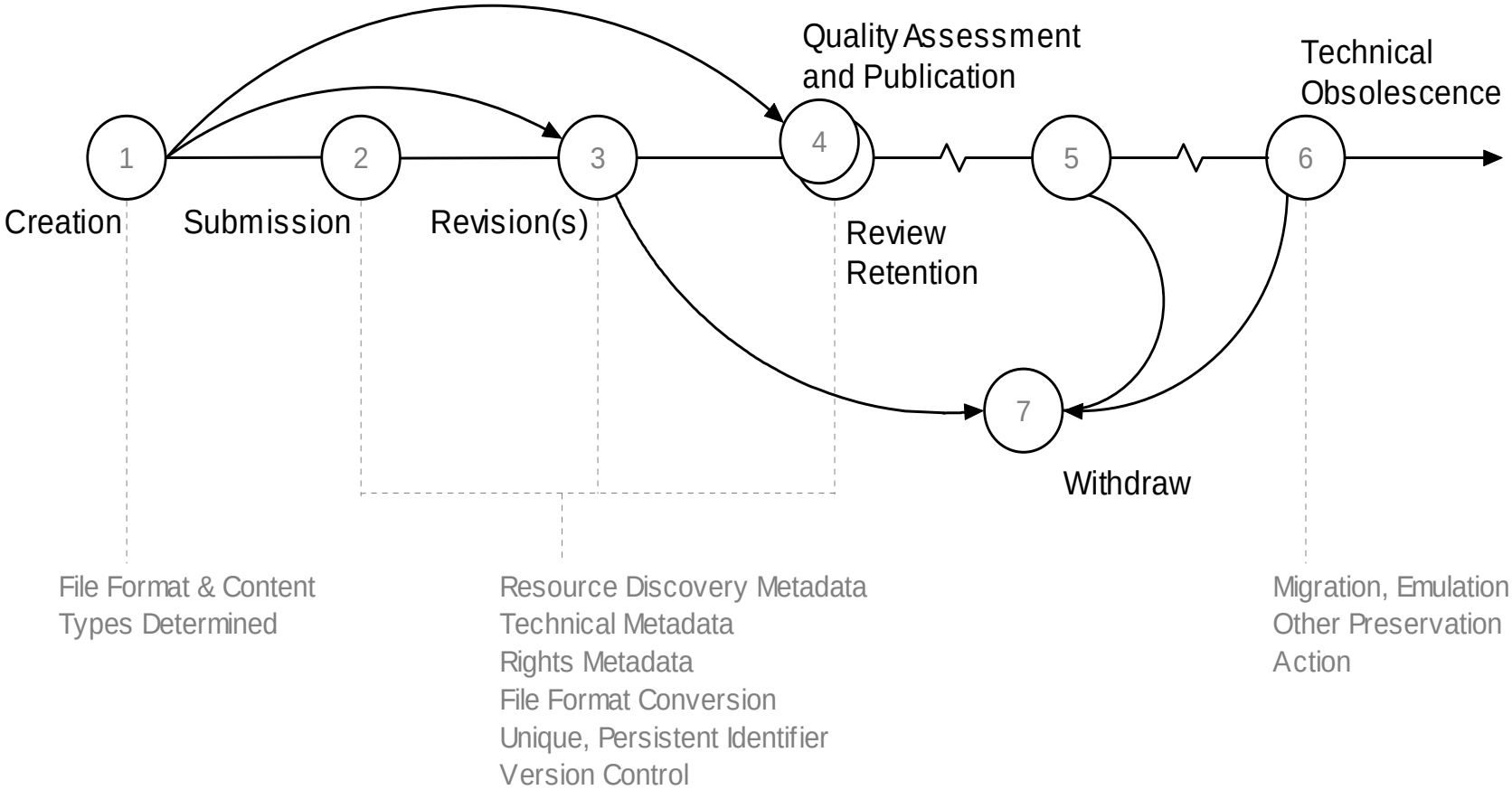
SHERPA

SHERPA DP Project

- **Aims**:
  - To develop a persistent preservation environment for SHERPA Partners based on the OAIS reference model, including a set of protocols and software tools
  - To explore the use of METS for packaging and transferring metadata and content
  - To explore the use of open source software and tools to add functionality to and extend the storage layer of repository software applications
  - To create a Digital Preservation User Guide

ahds

SHERPA

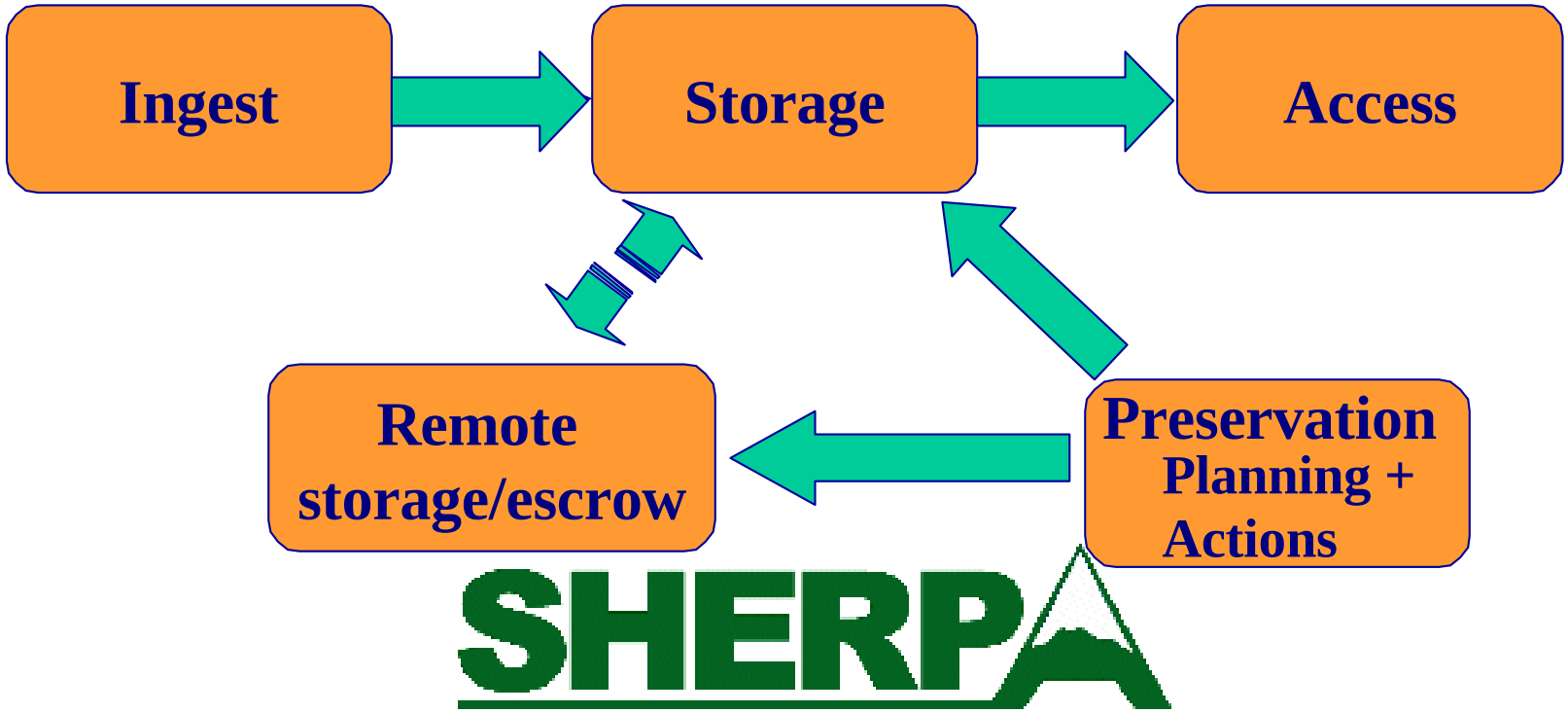

Quality Assessment
and Publication
Technical
Obsolescence
1
2
3
4
5
6
7
Creation
Submission
Revision(s)
Review
Retention
Withdraw
File Format & Content
Types Determined
Resource Discovery Metadata
Technical Metadata
Rights Metadata
File Format Conversion
Unique, Persistent Identifier
Version Control
Migration, Emulation
Other Preservation
Action

SHERPA

ahds

Disaggregated model:

- Institutional repository for access
- Supra-institutional preservation service



ahds

Preservation Planning

- Evaluate contents of archive and undertake risk assessment
- Recommend updates to migrate current holdings
- Develop recommendations for preservation standards and policies
- Monitor changes in technology environment, users' service requests, and knowledge base
- Develop detailed migration plans, software prototypes and test plans

ahds

SHERPA

Preservation Actions

- Provide a permanent storage facility
- Create and manage multiple copies of content, including off-site storage
- Manage storage hierarchy
- Refresh/replace media
- Provide disaster recovery capabilities
- Implement migration plans and migrate holdings as appropriate
- Manage version control

ahds

SHERPA

Why Disaggregated?

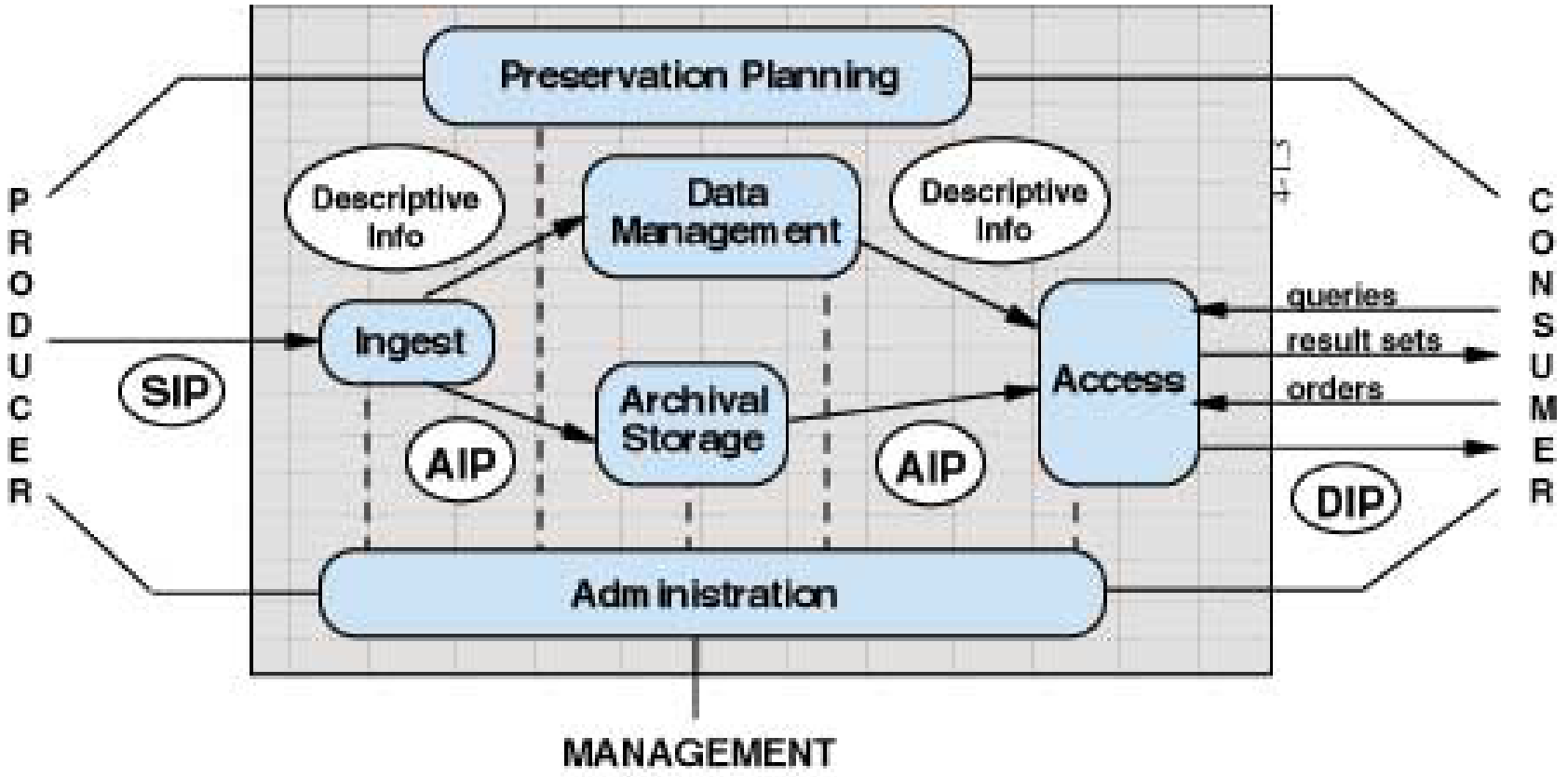
Reasons:

- Preservation is not inherent in most repository software
- DSpace and Eprints software primarily about submission, basic storage and access
- Scarcity of staff with necessary preservation skills and expertise
- Seeking to remove repetition of services
- Cost savings?

ahds

SHERPA

OAIS Functional Model



Applying the OAIS Reference Model

- Critical review of the OAIS Model
- Map OAIS functionality onto the proposed disaggregated model
- Identify rights and responsibilities of each party
- Identify and assign services and actions to be carried out and apportion these
- Review and refine AIPs, DIPs and SIPs
- Work up draft processes and procedures

ahds

SHERPA

Metadata and METS

- Review existing metadata captured by repositories against agreed administrative and preservation metadata set
- Identify additional metadata requirements and capture methods
- Review the potential for the use of METS within the SHERPA environment
  - As a framework for combining and packaging metadata
  - As a transfer mechanism for metadata and e-prints

ahds

SHERPA

Repository Archiving

- Investigate and implement automated transfers of data between institutional repositories and preservation repository
- Review DSpace and Eprint APIs, storage layers and module add-on capabilities
- Prototype and test SRB as a common storage medium
- Prototype and test API based access mechanisms
- Prototype and test external synchronisation mechanisms

ahds

SHERPA

Preservation Actions

- Investigate the processes required to enable changes and updates to e-print content that ensures their long-term integrity and preservation
- Create repository integrity checking and reporting services
- Create repository obsolescence checking, reporting and migration services
- Investigate remote alerting service capabilities
- Investigate mechanisms for automatic creation of new versions, or migration and redeposit

ahds

SHERPA

Implementation

- Preservation plans drawn up
- Risk assessment finalised
- Policies and procedures finalised
- Cost models and business case developed
- Implement services

ahds

SHERPA

Digital Repository Preservation User Guide

- The User Guide will recommend standards, best practice, protocols and processes that might be used in the management, preservation and presentation of e-print repositories
- Will draw on experiences of SHERPA and other relevant projects, and include case studies
- Will complement Beagrie and Jones “The Preservation Management of Digital Material Handbook”

ahds

SHERPA

http://www.sherpa.ac.uk

sheila.anderson@ahds.ac.uk

Stephen.Pinfield@Nottingham.ac.uk