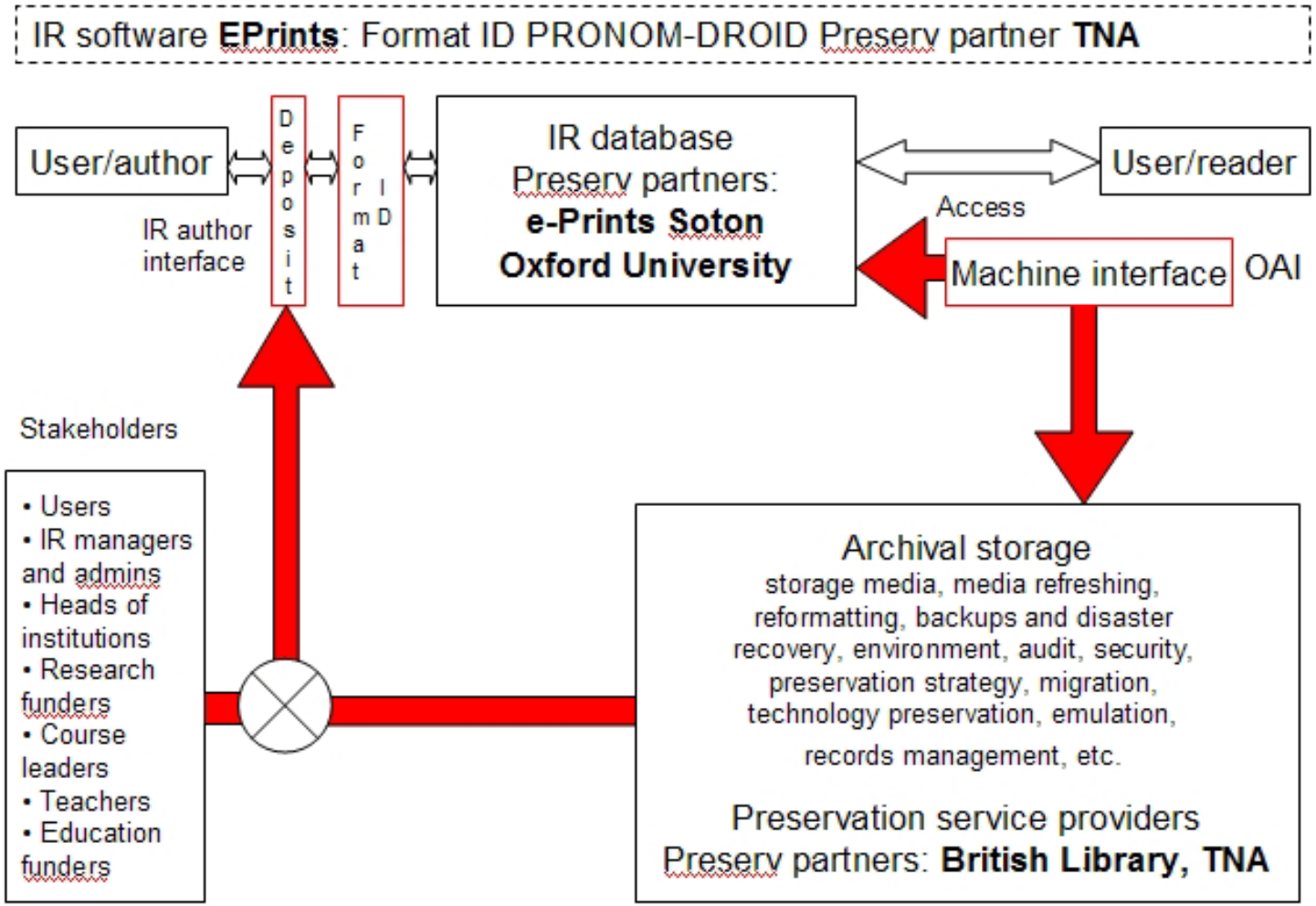
Repository models and policies for preservation

Steve Hitchcock

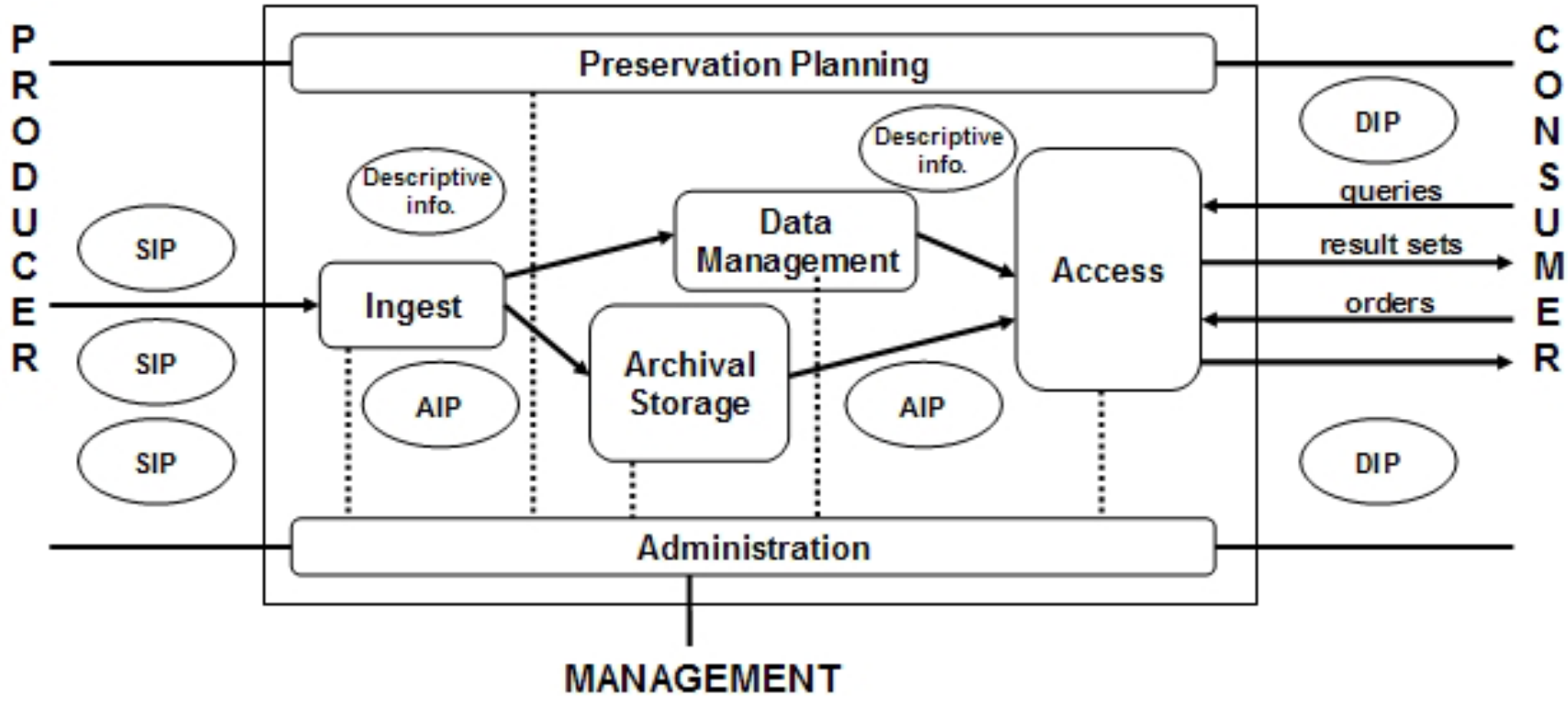
Preserv Project
Intelligence Agents Multimedia Group,
School of Electronics and Computer Science (ECS),
Southampton University

DPC Briefing on
Policies for Digital Repositiories: models and approaches
British Library, London, 5 July 2006

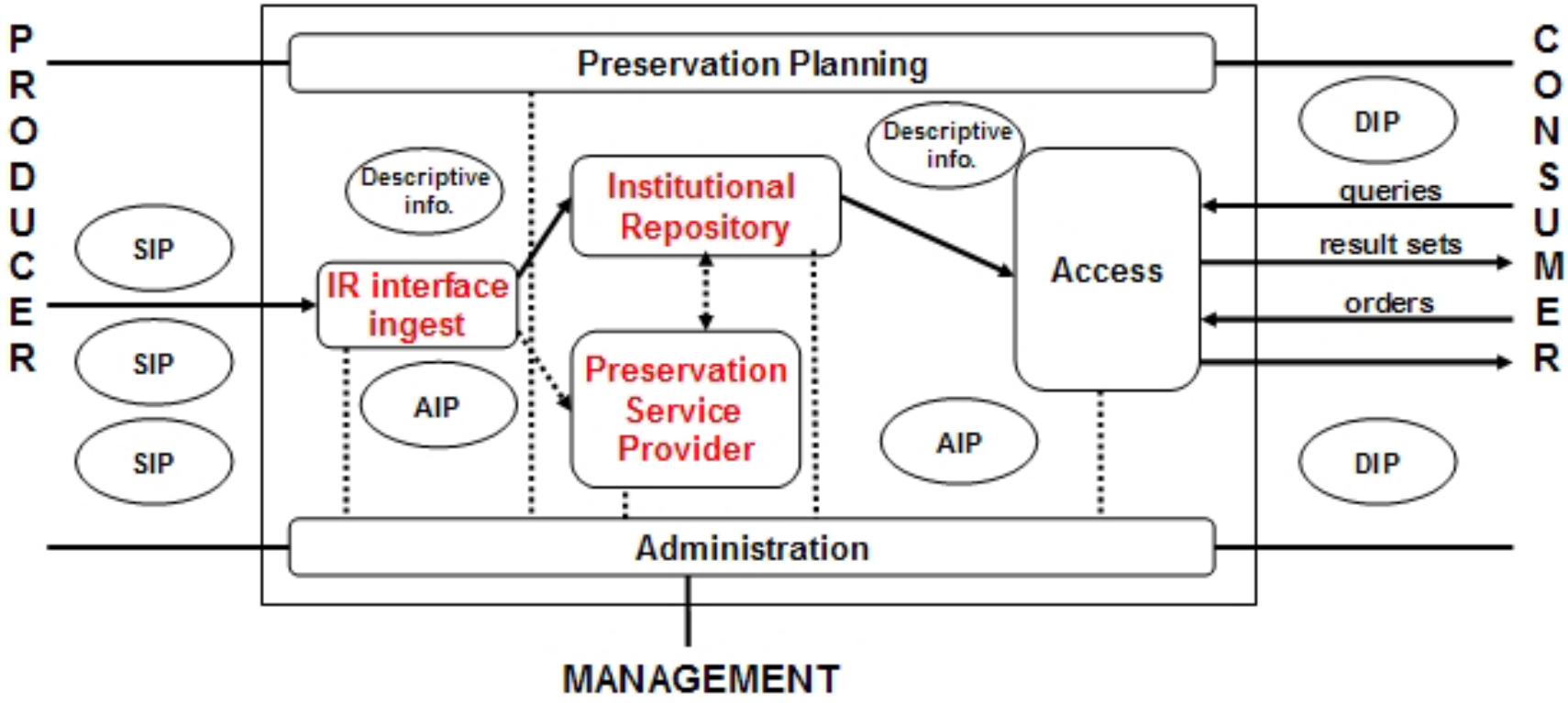

Preserv preservation service provider schematic



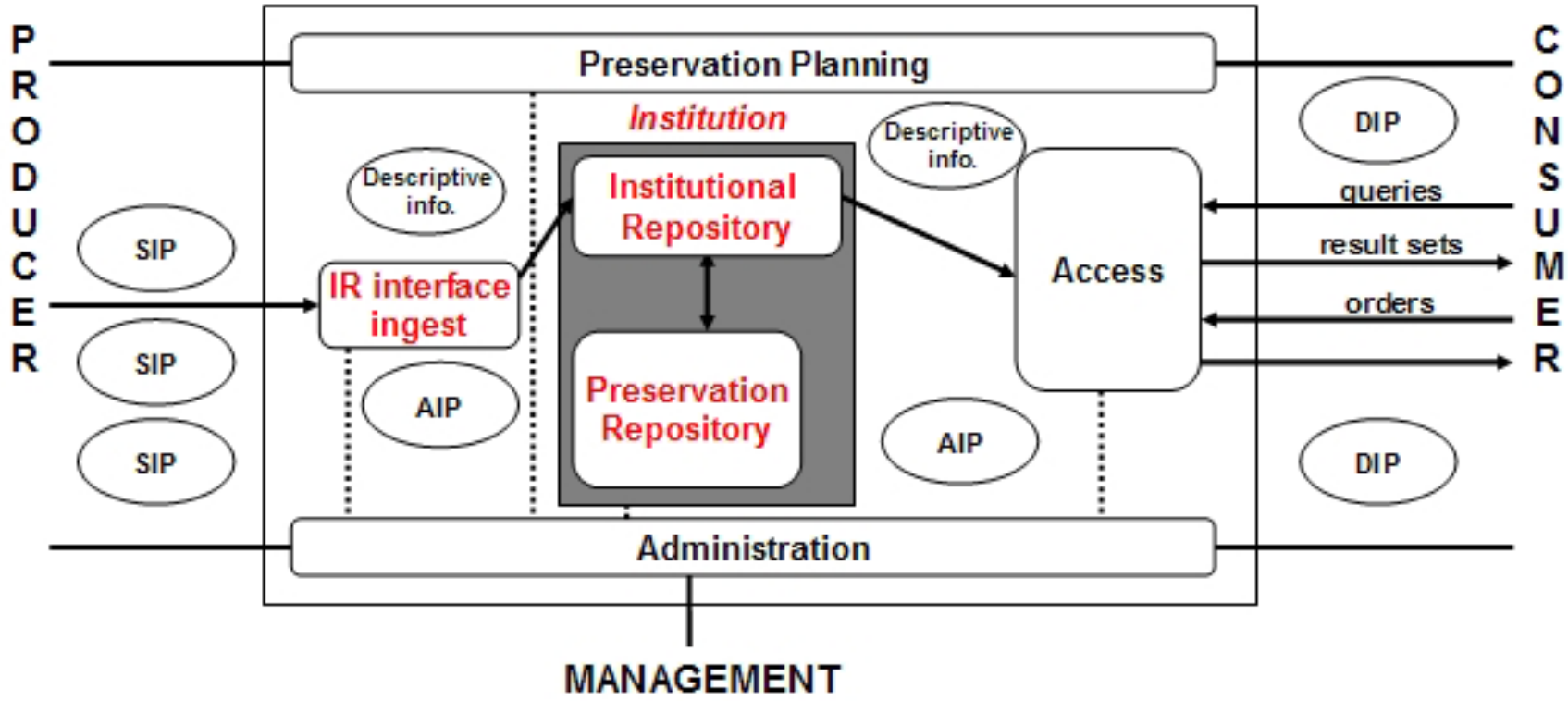
OAIS functional model



Service provider model



Institutional model



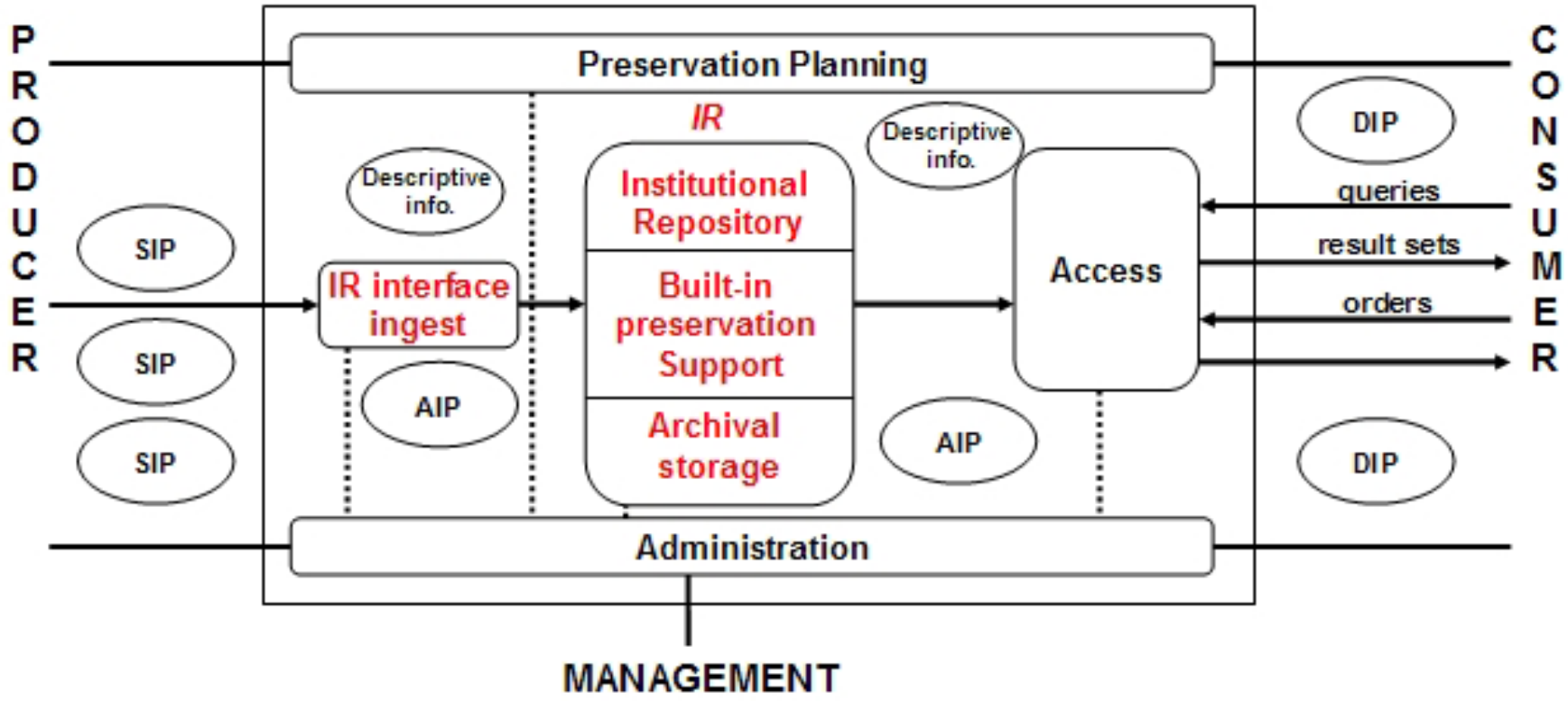
“Augmenting repository interoperability”

- Repositories as active nodes
- Cross-repository workflow
- Compound objects, e.g. objects within objects + datastreams
- Surrogates

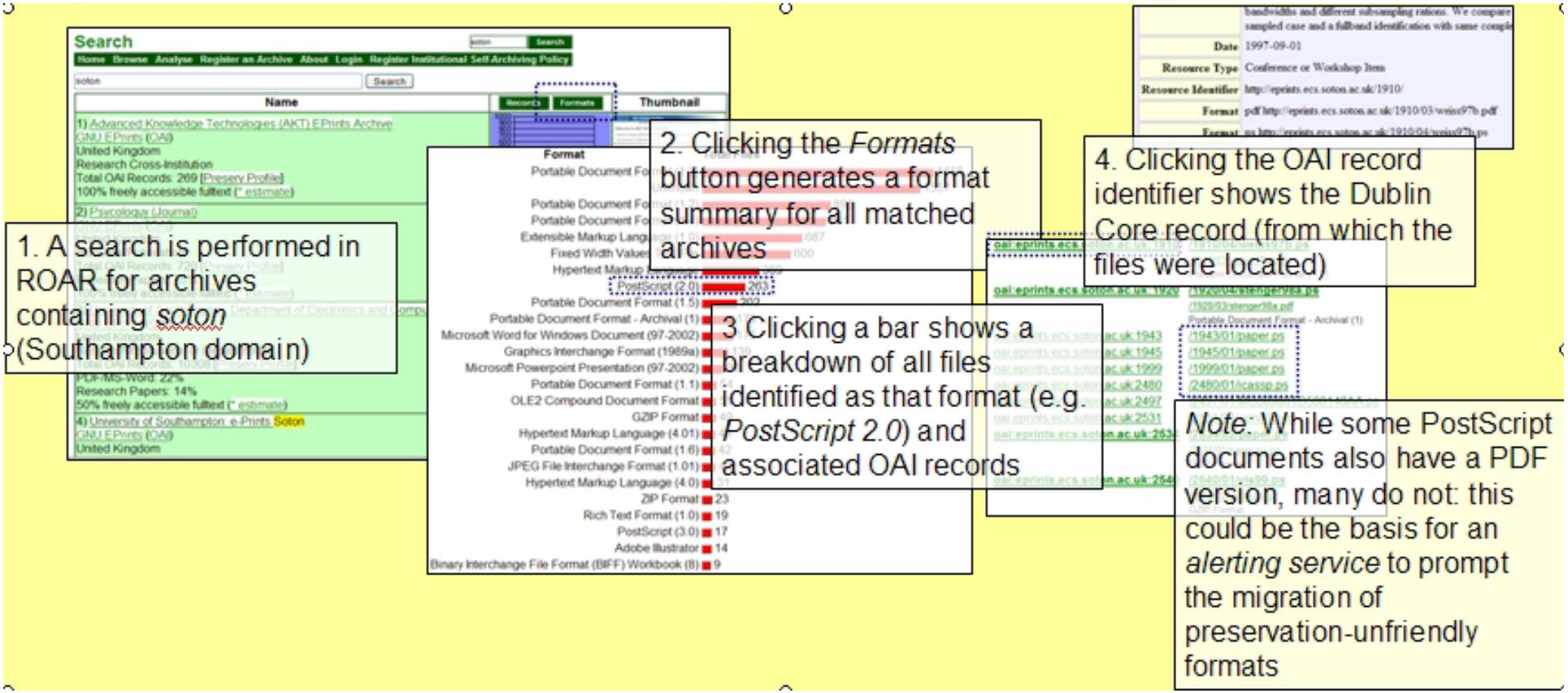
from Herbert Van de Sompel, Southampton workshop on repository interoperability, 4 July 2006

Also to be presented at JISC/CNI Conference, York, 6-7 July 2006

Repository model



Format profiling using PRONOM and ROAR



ROAR http://archives.eprints.org/

Survey of repository policies

Selected repository administrators invited to participate, based on availability and size of ROAR profile

Original test sites for profiling and survey included Oxford University, e-Prints Soton, ECS EPrints (Soton)

Series of questions, based on analysis of preservation metadata for Preserv model

	EPrints	DSpace	Both
Accepted/sent	22	11	2
Returned	13	4	2

Does the repository have any existing policy on preservation?

Does the repository have any existing policy on preservation?

Yes 1 No 18

Example policy
http://www.rub.ruc.dk/rub/selvbetjening/projektbiblioteket_eng.shtml

Does the repository implement any preservation measures, internally or with external agents/services?

Byte preservation **Y 8 N 9 No reply 2**

Transformation **Y 3 N 14 No reply 2**

Rendering **Y 1 N 13 No reply 5**

Emulation **Y 0 N 14 No reply 5**

Other: backup, mirroring, geographic cluster backup

Partnerships: Sherpa-DP, MetaArchive NDIIPP, dissertation copies at German National Library

Does the repository have a policy on submission file formats?

Y 11 N 4 No reply 4

- prefer PDF / DOC / PPT / HTML
- recommend using PDF or HTML
- PDF (Sherpa policy)
- accept all formats, text documents should be at least be pdf preferred pdf/a
- Use DSpace supported, known, and unknown formats (x3)
- Rendering software must be free, i.e. Acrobat, text, PostScript, HTML

Are there any restrictions (on formats) for submitting authors?

Y 6 N 11 unclear 2

- Word, PDF, Postscript, ascii, html, LaTex, PowerPoint (for conference posters), jpeg (images of book jackets).
- PDF required, other formats optional
- PDF only
- Accept HTML, ASCII, PDF
- ask authors to submit pdf
- majority of deposits PDF; allow other formats, e.g. html, rtf

Does the repository transform submitted formats in any way?

Y 14 N 2 no reply 3

- convert Word docs to PDF
- Most files (e.g. Word, Postscript, PowerPoint) converted to PDF
- transform source files to pdf
- Proprietary formats usually converted to PDF
- transform textual documents to pdf/a
- maths Latex, PS to PDF
- author option to convert to supported format
- video or graphic files zipped

Does the repository require the original source version from the author?

Y 0 N 13 no reply 6

- source file formats (e.g. Word, TeX, WordPerfect) can be deposited
- keep the original sources if deposited
- Authors asked to deposit a copy of "own final corrected draft version" rather than "publisher's formatted version"

Author agreements

Does the repository have any explicit agreement with authors on rights?

Y 16 N 3 ? 0

Examples:

DSpace license with minor changes

“Yes, but mediated deposit means depositing authors don't see it”

Does the agreement refer to rights for preservation?

Y 4 N 10 ? 5

Summary: policy before preservation

Repositories don’t know what they want and are looking for guidance on preservation

Don’t assume one-size-fits-all service will be sufficient

Repositories embrace different institutional, cultural and social constraints that will shape policy, including preservation, when they get round to defining it!

Propose a hierarchical series of preservation service models so repositories can choose which one suits

Repositories are already taking actions on ‘preservation’ that might compromise preservation services