# LOCKSS UK with a focus on reporting experience

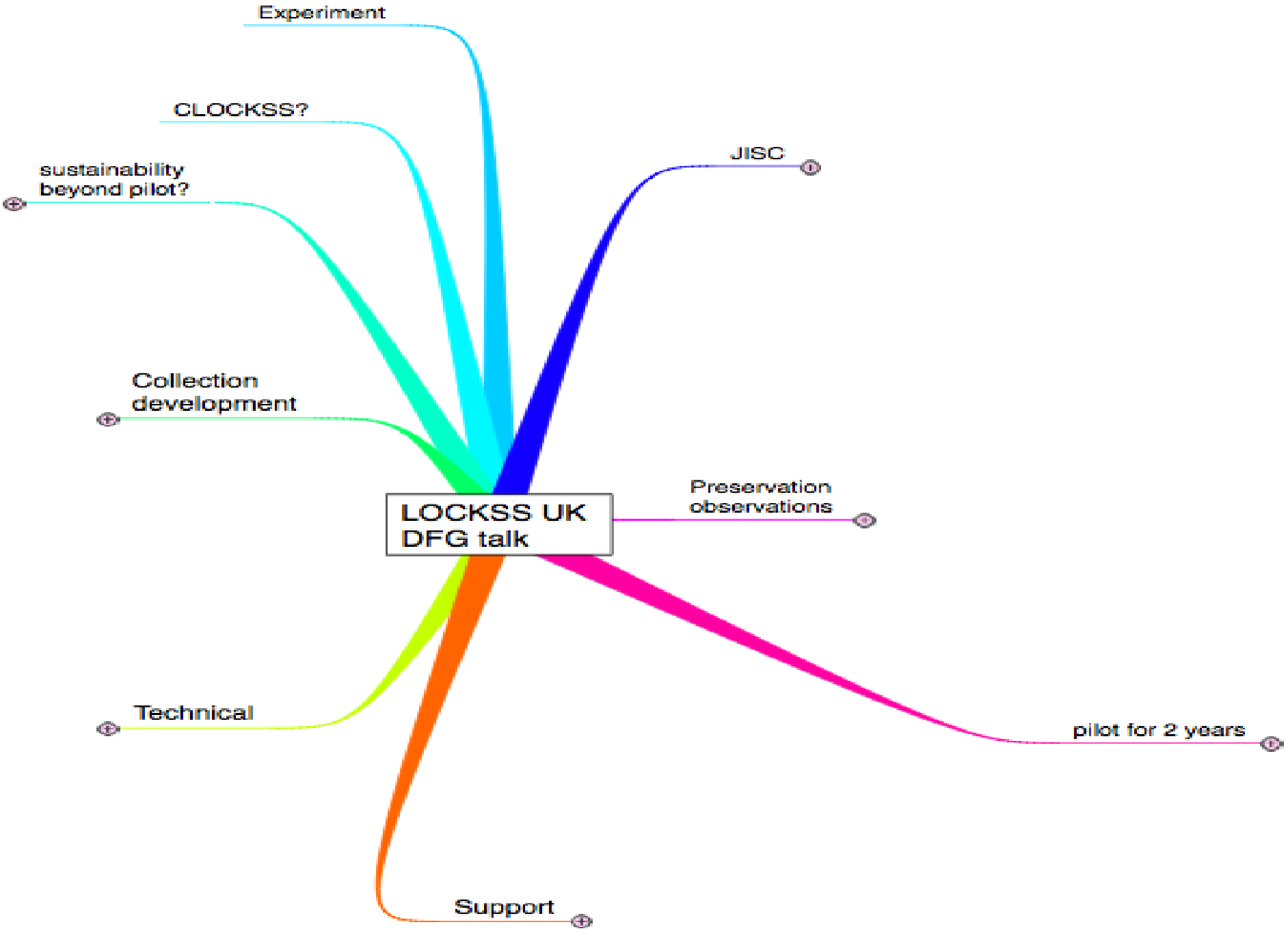
Chris Rusbridge

11 June 2007

Berlin

Slides from Helen Hockx-Yu, Paul Yarwood, William Nixon, Adam Rusbridge

Funded by:
JISC e-Science


Experiment
CLOCKSS?
sustainability beyond pilot?
JISC
Collection development
LOCKSS UK DFG talk
Preservation observations
Technical
pilot for 2 years
Support

# Joint Information Systems Committee (JISC)

- Strategic coordination role
- Development programmes: innovation
- National services
- National information provision role
  - Databases initially (ISI)
  - EJ national site licences (Model Licence)
  - Digitisation & other collections
- Funding top-sliced + partial cost recovery

# JISC & preservation

- Two Warwick workshops in 1990s
- Digital Archiving WG with publishers
  - 7 studies & reports
- eLib Projects
  - CEDARS and CaMiLeON ++
- Founder with BL of Digital Preservation Coalition
- Digital Preservation Strategy 2000-2005
- Small scale development programme (4/04)
- Digital Curation Centre 2004

# Role of JISC in EJ preservation?

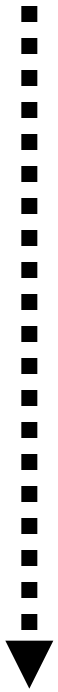
- Liaison between publishers, libraries & 3rd party service providers
  - explore options for practical implementations of archiving clauses in JISC Model Licence
- Provide community reassurance on archiving of e-journals
- Help community build experience with emerging approaches for informed decisions
- Work with national, deposit and institutional libraries towards national approach to e-journal archiving
- … and/or provide a centralised UK service? ⌛

# My objectives…

- Generate real library involvement & responsibility in preservation
  - Responsibility is more than paying a bill
- Develop a distributed approach that works for UK
- Address IPR, building on Model Licence terms
- Promote diversity of preservation approaches
  - But no real alternative at first
  - Legal deposit alternatives limited to on-site use
- Technical, organisational, licence, low cost & appliance approaches of LOCKSS attractive
- Work to get libraries involved in further types of content including scientific data…

# Preservation risks

- Not caring enough to try
- No permissions to do it (or don't know what permissions we have!)
- Insufficient contextual information to interpret
- Human error
- Media failure
- Lack of money
- Policy failure
- Deliberate attack
- Obsolescence of format



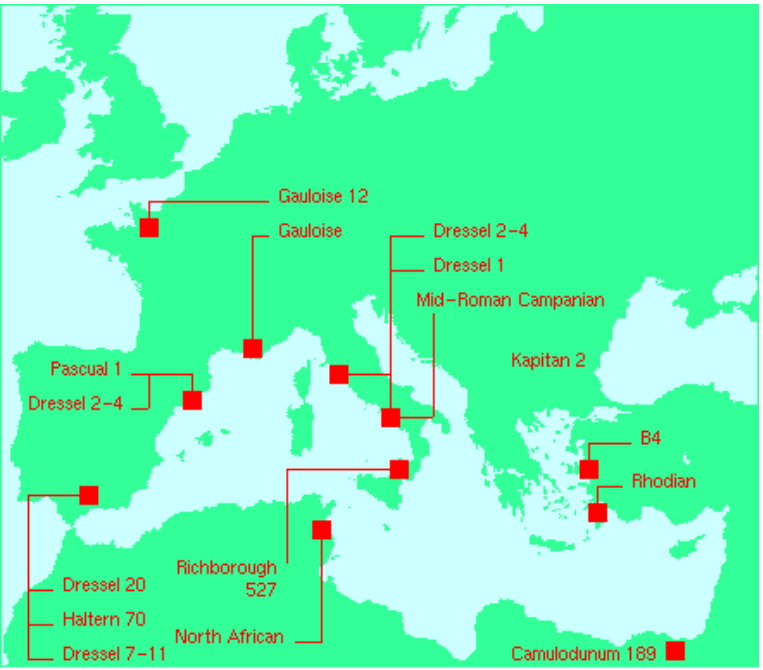
# Risks on 100 year timescale

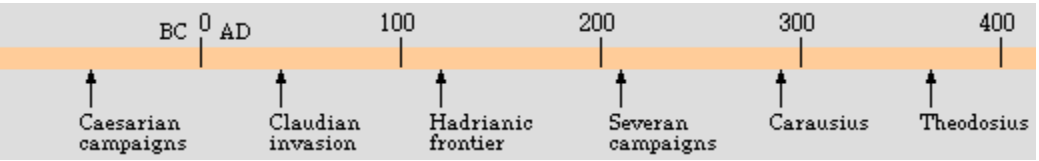
- Money and policy issues become critical
- Technology changes certain and impossible to predict
- Take a reasonable (10-year?) timescale
  - Keep in good order so you can hand over to successors
- OAIS not really any help!
  - Neither Representation Information nor Designated Community are adequately defined (or definable?)

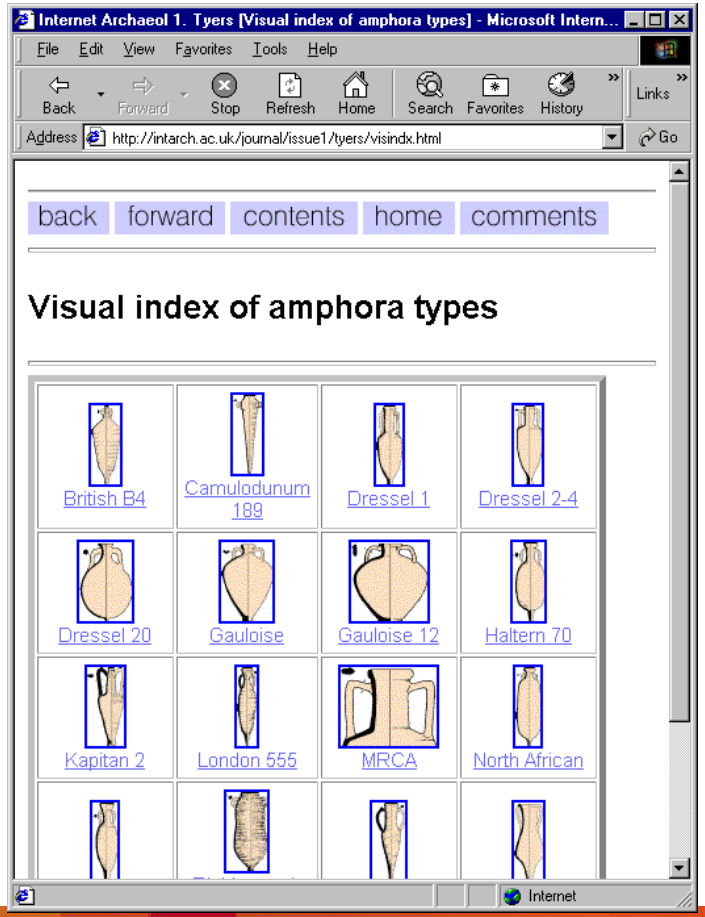
# Migration versus emulation

- False dichotomy!
- Is it migration or emulation if I…
  - View a Mac Powerpoint 4 slideshow with a Windows XP Office 2000 suite?
  - Write a PDF with GhostScript on Linux & read it with Adobe Reader on Mac?
  - Need a closer analysis of the object model (data structures plus methods)
- In e-journals, obsolescence is very rare
- Greatest risk in supplementary materials

# Internet Archaeology: publication with data



# Preservation of licensed content needs

- Technical approach
- Organisational & support structures
- Economic & business model that works
  - NB AHDS problem
- Compliance with IPR laws & licences
  - Legal deposit?
  - Licence **agreements**

# Response?

- Diversity of approaches
- Diversity of technology
- Diversity of funding
- Diversity of political context

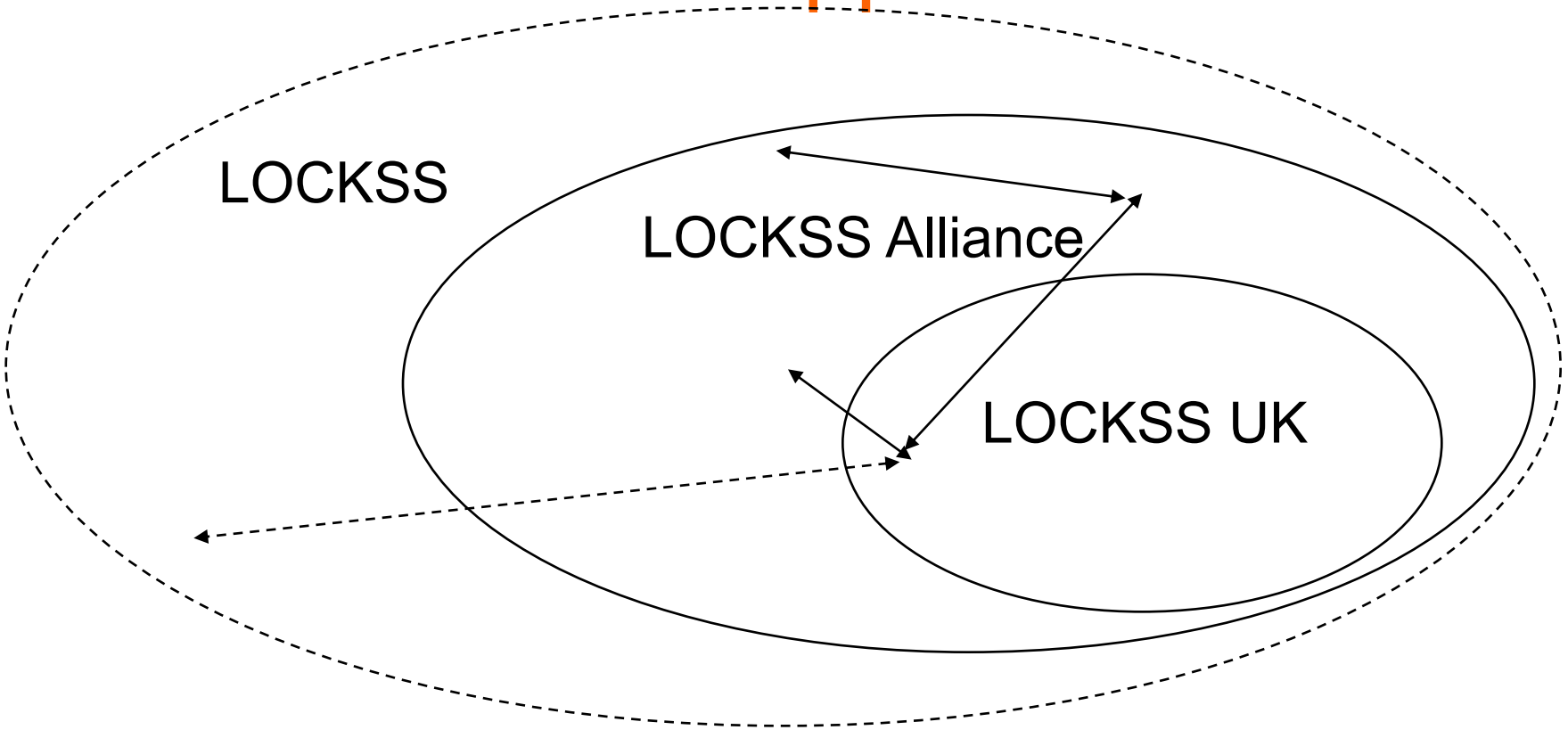
# JISC LOCKSS-UK Pilot Objectives

- Raise awareness of the LOCKSS initiative
- Seed a self-sustaining base of LOCKSS users in the UK
  - provide practical help to get started
  - develop the skills needed beyond Pilot.
- Trial LOCKSS technology in an operational environment
  - Investigate challenges associated with collective preservation of e-journals in common use in the JISC community.
- Build a centre of expertise outside the US, feeding the lessons learned back for the benefit of the international LOCKSS community.
- Allow the JISC community to make an informed assessment regarding future use of LOCKSS versus other alternatives.

# UK LOCKSS Pilot

- Two year pilot, launched in February 2006.
  - Funded by JISC and CURL
- 30 HE institutions in total
  - 24 initial pilot programme participants (funded by JISC)
  - 6 associate members (self-funded)
- Programme components include:
  - LOCKSS Technical Support Service (LTSS)
  - Publisher negotiation & legal appraisal of the archiving clauses in Model Licence
    - Led by Content Complete
  - Collection development at programme and institutional level
    - OpenLOCKSS
  - Collective UK membership at LOCKSS Alliance

# LOCKSS UK approach




LOCKSS UK a proper subset of LOCKSS Alliance

# Support

- 1st line technical support
  - Technical backup; mediate local library & technical staff
  - Particularly network, security, firewall issues etc
- Plain language support!
  - LOCKSS documentation often technically oriented
- Central coordination
  - Eg re aggregators
- Workshops
  - Training
  - Advocacy
  - Awareness
  - Issue feedback

# Support Overview

- 105 helpdesk queries handled
- Common Issues include:
  - Access problems (eg firewall issues) major problem at first
  - Many installation and setup queries
    - System Architecture, Configuration Details
  - Collection Errors
    - Crawl Errors
    - Crawl Window Closed
  - How to serve content
  - Range of available content

# Technical support service

- Machine acquisition
  - Bulk purchase 24 machines @ £500 (Dell)
  - 1*250 MB SATA drive per machine
  - Associate members bought their own
- Technical staff member: plan
  - Installation & support
  - Future development (eg blogs)?
  - Second implementation?
  - Plug-ins & publisher liaison
  - Proxy trials
- Experience feedback

# Machine Status

- 30 machines up and running correctly (6/07)
  - 2 hard disk failures
  - 1 difficult BIOS problem
- Content has been added correctly and successfully
  - Gla 1244 AUs: 47 GB
  - KCL 642 AUs: 60GB
  - Issues highlighted related to:
    - Aggregators
    - Transfer of License (journal moves between publishers)

# System Development

- Plugin Development: 5 in progress
  - Annual Reviews in testing
  - Royal Society of Chemistry finalising
  - Taylor & Francis to begin
  - Cambridge University Press: plugin development begun
  - Open Source titles: beginning
- System Development
  - Jhove Integration?
    - Make available on demand format verification and validation
    - Integrated into the LOCKSS user interface

# Plugin scalability?

- LOCKSS Alliance release notices over 8 weeks
  - ~54 volumes per week
  - ~3 new titles per week

# Content issues in journal publishing

- Assets of scholarship increasingly in control of publishers
  - No guarantees of perpetuity
  - Copyright law restricts archiving programmes; publishers resistant to change
- The NESLi2 license offers some leverage for libraries
  - NESLi2 is national initiative for the licensing of electronic journals
  - Model Licence = agreement between institution and publisher, containing terms and conditions of access, use and service
  - The Model Licence **includes** archiving clauses and **requires** continued access following termination of licence **without charge**
    - however adherence to clauses remains at the discretion of publishers!

# Publisher Negotiations

- Led by Content Complete Ltd
  - Negotiation agent for JISC
- Permissions required from publishers
  - Crawl permission in form of manifest page used by LOCKSS crawler
  - licence or terms of conditions for libraries
- Negotiations started with:
  - 7 NESLi2 publishers, 10 non-NESLi2 publishers
  - Open access negotiations begun (OpenLOCKSS)

# Negotiation status?

- Annual Reviews 🗐
- Cambridge University Press 🗐
- Royal Society of Chemistry 🗐
- Taylor & Francis 🗐
- (Oxford University Press 🗐)
- American Psychological Association ⌛
- Emerald 🗐?
- Lippincott Williams & Wilkins 🗐?
- British Psychological Society 🗐?
- Palgrave MacMillan ⌛?

# Forthcoming content

- Annual Reviews
  - 41 journals, 1300 AUs
  - 1 AU=45 MB (sample of 2) ➔ total ~60GB??
- RSC
  - 27 journals, 182 AUs
- CUP
  - 256 journals, 2800 AUs
- T&F
  - ~ 100 journals?

# Some observations

- For publishers, decisions regarding preservation and archiving of their content are not taken lightly
- The increased activity in this area (CLOCKSS, Portico) has heightened awareness of the issue but slowed us down
- Implementing LOCKSS competes heavily with other strategic issues:  OA, digitising backfiles, acquisitions, platform changes

# OpenLOCKSS Collection Development

- Led by Glasgow Library for JISC & CURL
  - Core objective for the Programme
  - Supported by Technical Support Service
- Priority NESLi2 content
  BUT
- Surveys indicate other (Open Access) titles
- Priority SHOULD perhaps be small publisher closed access?
  - Median number of journals/publisher = 1!


Establishing a UK LOCKSS Pilot Programme, Helen Hockx-Yu
http://eprints.rclis.org/archive/00007354/

# Surveying the Scene

- LOCKSS Open Access Title Survey
  - OA Titles published in the UK
  - Based on DOAJ listings
  - 97 titles but scope for more
- Polling the Community
- Permission from Publishers
- Plug-in Development
- Preservation Critical Mass (≥ 6 participants)

# Problems?

- Some acceptance (10 titles)
- Some valuable titles declined (eg D-Lib Magazine)
- Do people think Internet Archive does enough?
- Should we focus on long tail of small closed access publishers?

# LOCKSS UK sustainability

- Pilot ends February 2008
- Outline plan for subscription service
  - UK Technical Support
  - UK Librarians negotiating small titles
  - LOCKSS Alliance group membership
  - NESLi2 licence negotiations from JISC
- Bid for funding to under-write for 2 years
  - Reasonable cost for ≥~ 30 subscribers?

# CLOCKSS

- Controlled LOCKSS
- Almost identical approach & technology base, but content dark
- Small group of participant libraries
  - Edinburgh only non-US so far?
  - Recruiting? Aim for geographic & political variety?
- Larger group of participant publishers
- Publisher/library funding
- Agreement on trigger events?
  - Responsibility & load post-trigger not clear to me!

# LOCKSS failure experiment

- James Currall of Glasgow…
- Set up LOCKSS machine
- Ran for some time
- Replaced disk with empty disk
  - Equivalent to total disk failure
  - Added list of relevant AUs
- Disk slowly re-filled with content
  - No operator intervention needed!

# LOCKSS research questions?

- LOCKSS is a single software implementation
  - How to build (and trust) additional software implementations to the software design?
- LOCKSS is distributed but not de-centralised
  - How to prevent Stanford team being a single failure point for LOCKSS?

# Non-LOCKSS approaches?

- Several science data archives have been going for >25 years with high reliability (we think!)
  - Ie 20 years before OAIS
- Use domain scientists to define contextual metadata requirements
- Work as “community proxy”

# What kinds of data?

- Observations
  - eg UARS (Upper Atmosphere) Level 0: telemetry
  - UARS Level 1: measured physical parameters (post calibration?)
- Derived data
  - UARS Level 2: calculated geophysical? profiles
  - UARS level 3: gridded, interpolated?
- Combined data
- Crafted data
  - Eg annotated gene/protein databases
- Descriptive (meta)data

# StORe: Source data formats

| | | |
|---|---|---|
| CAD/GIS: | | 39 |
| Extensible mark-up language (XML): | | 35 |
| Database files (e.g. Access, MySQL): | | 117 |
| Flat files (e.g. FITS): | | 66 |
| Hypertext mark-up language (HTML): | | 60 |
| Image files (e.g. .jpg, .tif, .bmp, .gif): | | 228 |
| Plain text (.txt): | | 179 |
| Portable document format (.pdf): | | 156 |
| Rich text files (.rtf): | | 53 |
| Spreadsheets (e.g. Excel/.xls): | | 220 |
| Statistical software: | | 75 |
| Tables/catalogues: | | 102 |
| Word processed files (e.g. Word/.doc): | | 220 |
| Other *(please specify)*: | | 76 |

# StORe: the *other* data formats?

They said the 76 other formats included:

+latex+.cc source code, .cif (crystallographic data), .pdb, .mtz, .pool, .root, .raw, .swf, .fla, .raw, .mpg, binary files, chemdraw cdx, xwin nmr files, .ps files, .fla, .swf, masslynx files, derived data in PAw-format ntuples, raw mass spectrometry data, X-ray diffraction data, kaleidagraphs, Atlas/ti hermeneutic unit files, C++/shell scripts, Fourier induction decay files, etc., etc., etc., etc……..

# StORe: the *other* data formats - more

They also said such things as:

> "It is stored in a database, but nothing so simple as an Access file! It's one of the largest databases in the world! The format is Kanga/Root and previously was Objectivity. I think it's of the order of Picobytes in size."

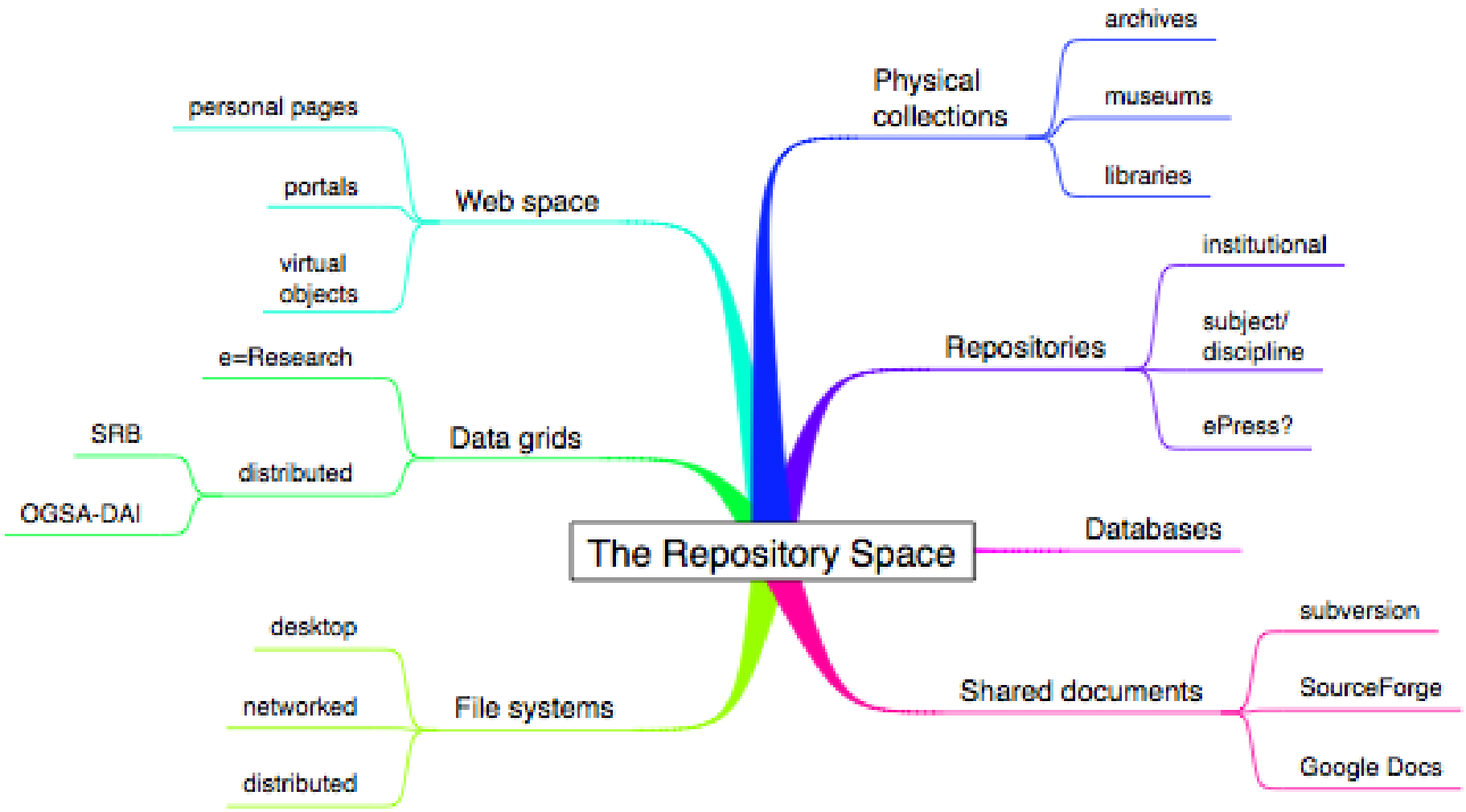
And:

> "God preserve us from idiots who archive data in proprietary commercial formats (Excel spreadsheets and MS-word documents)!"


• Slide from StORe project

# Registry/Repository Of RepInfo

- Attempting to implement OAIS Representation Information
  - In a registry and repository that itself should be OAIS compliant
- We have LOTS of internal “discussions” about RepInfo!
- Just precisely what RepInfo is needed to preserve a particular object for 100 years?
  - RepInfo is different from format information… but how? And is it scalable?


The Repository Space
Physical collections
archives
museums
libraries
Repositories
institutional
subject/ discipline
ePress?
Databases
Shared documents
subversion
SourceForge
Google Docs
File systems
desktop
networked
distributed
Data grids
e=Research
distributed
SRB
OGSA-DAI
Web space
personal pages
portals
virtual objects

# Preservation research questions

- Representation information
  - Show me some that's useful!
- Context information
  - Particularly for data (e-Journals are comparatively easy)…
- Designated community
  - How to define, how to monitor?
  - How to handle multiple simultaneous?
- Obsolescent format handling tools
- Understanding authenticity through format change
- Mashups
- What are the effects of very large size?
- What is the affordable amount of diversity?

# Thanks

Chris Rusbridge

Digital Curation Centre

University of Edinburgh

c.rusbridge@ed.ac.uk