Oxford Update

Neil Jefferies
R&D Project Manager
Systems & eResearch Services (SERS)
Oxford University Library Services (OULS)
Introduction to SERS/OULS

- Bodleian one of ~110 libraries
- 32 OULS Libraries
- 29 Non-OULS Libraries
- 46 College Libraries
- And a few others!

OULS
- 750 Staff
- £25M budget
- 11 million items
- 156 shelf miles (250 km)
- SERS provides all electronic services
Digital Library

• Digital Asset Management System (DAMS)
  – Common infrastructure for DL applications
  – Digitised “book-like” materials
    • Legacy projects – Greenstone, Luna, Custom Apps (~20TB)
    • 1M digital surrogates from imaging service – HFS (~10TB)
    • “1M” Google Library Project (?TB)
    • Digitised maps data (100-300TB+)
    • Current digitisation programmes (10's of TB)
  – Born-digital materials
    • Institutional Archives, Data, Web Servers, Personal Digital Effects, Digital Legal Deposit, Knowledge Management
What are the Challenges?

- Flexibility - We don't know what's coming
- Scalability - ...but it's going to be big...
- Longevity - ...and it's not going to go away...
- Availability - ...and people will want it...
- Sustainability - ...so we've got to deal with it...
- Interoperability - ...and work with everyone else.

“Big bang system implementations are lossy - we really, really don't want to have to do them - ever”
Flexibility

• The FEDORA Object Model
  – Object composed of multiple data & metadata streams
  – RDF used to describe structure and relationships
    • Semantic Web/Linked-data ideas are crucial
• Don't make unnecessary/premature decisions
  – Formats not proscribed (besides RDF)
• Componentised Web Services approach
  – Organic system growth and evolution
• Favour Open Interfaces/Standards and Tools
  – De facto standards matter – availability of tools
Scalability

• Volume Scaling “Billion file problem'
  – Object store technologies
  – Distributed Systems – Scale through replication
• Resilience of Large Scale Systems
  – Live Replicas rather than backup
    • Versioning instead of overwriting.deletion
  – Self-healing systems
  – Safety-critical systems approach - heterogeneity
• Scalability of Access/Discovery
Longevity

• Content and Principles persist
  – People and Technologies will change

• Principles of System Design for Longevity
  – Decoupling of Functions
    • Components can be replaced without impact on the rest of the system
    • Simple Interfaces & well-define functions - easy to reimplement/shim
    • Minimise dependencies – open source, alternate implementations
  – Support Heterogeneity at all levels
    • Manage evolution and obsolescence
    • Resolvers/Abstraction layers
Availability

- Basic IT availability (Access)
  - Network, Power, Physical Environment

- Archival recoverability (Content)
  - In the long term, unlikely events become significant
  - Preservation more important than availability
  - Recovery from bare storage – everything is an object

- Digital Preservation (Usability)
  - Bit-level, Conversion, Emulation
  - Meta-Preservation – Collections, Archive, Contextual Data
Sustainability

• Budget and cost as a conventional library
  – Human costs remain – cataloguing, curation...

• Proactive involvement with projects/research
  – Data management and preservation are costed at the outset
    • Deposit and publication no longer distinct

• Leverage content to generate income

• Communities survive
  – Do not develop in isolation (minimise custom code)
  – Migrate and disseminate skills
Interoperability

- Interoperability is an ongoing process
  - Support for emerging and established standards
  - The Web/Semantic Web will persist
- Persistent, stable, well-defined interfaces
  - Dependencies beyond the organisation
- Ideally implement interfaces bidirectionally
- Abstraction is vital
  - The system can evolve without compromising interoperability
  - Low-level access limited to specific, time-bound cases
The DAMS

- The DAMS is an infrastructural component
  - Storage, Object Management and Tools layers
  - Provides the framework for Application development

- The DAMS is not visible to end-users
  - Applications: ORA, FMO etc. are
  - As the DAMS evolves the visible applications do not need to
  - Application development is decoupled

- The DAMS is primarily constructed from off-the-shelf components
In Practice

• Physical system (Hardware)
  – Virtual machine environment for scalability, resilience
    • On demand deployment too
  – Segregated networks for security
  – Duplication functions for resilience
    • Multi-site for disaster recovery

• Logical system (Software)
  – “Cloud” approach for tools and services
    • Load balancing and failover
  – MDICS
DAMS (Physical)

- **Public**
  - VMWare ESX
    - Data Image Store (22TB)
    - Root Image Store (2TB)
  - MDICS
- **Private**
  - VMWare ESX
    - Root Image Store (2TB)
    - Data Image Store (22TB)
- **Service**
  - ZFS Replication
- **Storage**
  - Split Writes & Crosschecks

- **HC 1**
- **HC 2**
- **Open Storage 1**
- **Open Storage 2**

- **SAS**
- **NFS**
- **ZFS**

October 2009

PASIG
DAMS (Logical)

Storage
- Object Store
- Bulk File Store
- HFS

Tools & Services (Cache)
- Fedora+
- Fedora/Resolver+
- PairTree+
- Fedora+
- Fedora+

Applications
- Search
- Reporting
- Scheduling
- Job Mgt
- ORA
- Digitised Books
- Ingest
- Virus Scan
- Text Extract

October 2009

PASIG
Projects and Applications

That's all very nice but what happens in reality?
Basic Services

• Oxford University Research Archive (ora.ouls.ox.ac.uk)
  – “Institutional Repository” of Research Outputs
  – Where it all started...

• Managed EPrints Service (not DAMS!)
  – Provides departments with their own repositories
    • Allows devolved control and collection policies
    • Content can be selectively harvested into ORA

• vocab.ox.ac.uk
  – Publishes vocabularies/ontologies supporting University Semantic applications
Content Expansion

• Current Digitisation Programmes
  – Electronic Ephemera - www.bodley.ox.ac.uk/eejjc/
  – Blockbooks - www.bodley.ox.ac.uk/csb/blockbooks.html
    • image library vs formal presentation
  – Google Libraries Project – volume bypass (MDICS)

• Digital Migration
  – Oxford Digital Library collections - www.odl.ox.ac.uk

• On-Demand Digitisation
  – Imaging Studio – external service
  – Option to replace delivery from remote storage
Interoperability

• BRII (JISC) - brii.ouls.ox.ac.uk
  – Building a Research Information Infrastructure
  – Add Research Information as context/metadata for ORA
  – Becoming a resource in its own right
    • Provides a logical route for holding long-tail data
• DART Europe, NEEEO, ETHOS
  – Etheses with different scopes and projects
• EIDCSR (JISC) - eidcsr.oucs.ox.ac.uk
  – Embedding Institutional Data Curation Services in Research
  – DAMS provides metadata registry for externally held data
Preservation

• Accessioning Web Servers
  – Material received as whole servers or even clusters
  – Initial action is to host/virtualise
  – Extraction and preservation of content is an “interesting” problem

• P2N
  – Oxford/Southampton collaboration
  – Distributed storage/preservation network appliance
  – Pure storage focus – works with standard repository software
  – More later
Extension

- FutureArch (Mellon) - futurearchives.blogspot.com/
  - Complex objects, personal digital archives
    - Whole disk images
    - Security and longevity are key
- Cultures of Knowledge (Mellon) - www.history.ox.ac.uk/cofk/
  - Enhanced catalogue of C17 letters, people, locations, dates
    - Direct capture of knowledge semantically
- Medieval Libraries of Great Britain 3 (Mellon)
  - Traces extant manuscripts back to original libraries
    - Adds evidence-qualified links to semantic model
Things we have learnt

• The transition to digital is very rapid

• We need to preserve before curation
  • The “Bit-Bucket” preservation store

• There is an emergent model...
  • People, places, events/activities, artefacts

• There is no single route for service delivery

• How do we run the system operationally?
  – DAMS combines operations and development, no exit

October 2009

PASIG
“Convergent evolution”
Questions

Neil Jefferies
neil.jefferies@sers.ox.ac.uk
Oxford Research Archive
ora.ouls.ox.ac.uk

Developer's Blog
oxfordrepo.wordpress.com

Google Code
look for: python fedora-commons