PASIG
Directions & Issues

Richard Boulderstone
Director eStrategy & Programmes
April 2011
PASIG June 2007: Concluding comments

- The British Library has a substantial long term vision for electronic resources
- We believe that we have the correct processes & design principles to build the capability we need
- We have made good initial progress although not without setbacks
- We have designed a highly scalable, fault tolerant, redundant, self-validating architecture
- We are seeking partners to collaborate in the more complex areas of the architecture, such as Digital Preservation, Digital Rights Management, Resource Discovery, Trusted Repositories, Metadata Strategies, Dataset Management and Preservation, ...
Roadmap for introducing IT capabilities to support Library’s business goals (BL IT Strategy 2011-15)
The definition of a long term target IT architecture provides strategic direction for the evolution of our systems infrastructure. However, delivery of the architecture is achieved through business-led programmes. Therefore, investment priorities will be determined by business need. In this context, the execution and delivery of our business-led programmes will determine the development of the IT architecture over the period of the 2011-15 strategy.

- Focus on continuous improvement
- Requires extension to achieve aspirations
- Requires major investment to achieve aspirations
Digital Library System Principles

- **Security**
  - Layered Security
  - Provable Authenticity
  - Single Point of Failure

- **Preservation**
  - Intolerant of bit loss
  - Never be down to your last copy
  - Preserve Metadata
  - Self checking, self healing

- **Implementation**
  - Vendor Independence
  - Commodity Hardware
  - Purchased as needed

- **Scalable**
  - Highly Scalable
  - Extensible

- **Software**
  - Transactional and re-entrant
DLS Network

4 Nodes (Complete Copies)
- British Library, St Pancras
- British Library, Boston Spa
- National Library of Wales
- National Library of Scotland

Additional Access Points
- Bodleian Library, Oxford
- Cambridge University Library
- Trinity College Library, Dublin
DLS Server Architecture

Boston Spa

National Library of Wales

St Pancras

National Library of Scotland
Replication

- Ingest into BSP
- BSP notifies all sites
- All sites update holdings
- NLS and LON request object from BSP
- BSP replicates to NLS
- NLS notifies all sites
- All sites update holdings
- BSP replicates to LON
- LON notifies all sites
- All sites update holdings
- NLW requests object from LON
- LON replicates to NLW
- NLW notifies all sites
- All sites update holdings
Cryptographic Services

- Every object signed as part of Ingest

- Digital Provenance
  - Detect corruption
  - Detect tamper

- Inter-site Trust
  - Sites have no trust model between them
  - Signature verified on receipt

- Every object checked on a regular basis
Digital Library System - Ingest

Store now contains:
1.25 Million Items
140 Terabytes of Content

A Terabyte is 1,000,000 Megabytes or $10^{12}$ bytes
Object Size Distribution
Issues - Content

- Content
  - eJournals
    - Huge Range of Formats (PDF Still Dominant)
    - Large Number of Publishers (A Few Large, Many Small)
    - Supplementary Information
    - Updates
    - Open Access
  - Images (Digitisation)
    - Format Standards (Now using JPEG 2000)
    - Resolution Debate (How Many Megabytes / Page?)
    - Viewer
  - Web Archives
    - Legal Deposit Legislation
    - Domain Crawl Size
    - Automated vs. Manual Selection
  - Many Others
    - Datasets (Selection, Curation, Size, etc)
Issues – Metadata & Preservation

- Metadata
  - Use Cases (Access & Preservation)
  - Persistent Identifiers (Datacite Service)
  - Level of Granularity
  - Synchronisation Across Stores
  - Standards (METS, MODS, etc)
  - Schema vs. RDF
  - Access vs. Preservation Design Considerations
  - External References

- Preservation
  - Proprietary File Types
  - Dynamic Content
  - Warc Files
  - Image Errors
Issues – Access & Infrastructure

- Access
  - API Definitions to Support Discovery & Navigation Of Rights Controlled Content
    - Authentication
    - Rights
    - Discovery
    - Metadata
    - Image, Video, Audio, Document,
    - Edit
  - Viewer for Range of Content Types

- Infrastructure
  - Buy / Build / Open Source
  - Modularity of Components
  - Scalability & Security are never solved
The British Library has a substantial long term vision for electronic resources

We believe that we have the correct processes & design principles to build the capability we need

We have made good initial progress although not without setbacks

We have designed a highly scalable, fault tolerant, redundant, self-validating architecture

We are seeking partners to collaborate in the more complex areas of the architecture, such as Digital Preservation, Digital Rights Management, Resource Discovery, Trusted Repositories, Metadata Strategies, Dataset Management and Preservation, ...