UNIVERSITY OF OXFORD

Long Tail Data

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

UNIVERSITY OF OXFORD

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

UNIVERSITY OF OXFORD

Digital Library

- Digital Asset Management System (DAMS)
  - Common infrastructure for DL applications
  - Legacy digitisation projects (~20TB)
  - 1M digital surrogates (~5TB)
  - “1M” Google Library Project (?TB)
  - Digitised maps data (100TB)
  - Current digitisation programmes (10's of TB)
  - Born-digital materials (20-40TB)

UNIVERSITY OF OXFORD

Is it really about “Long Tail Data”?

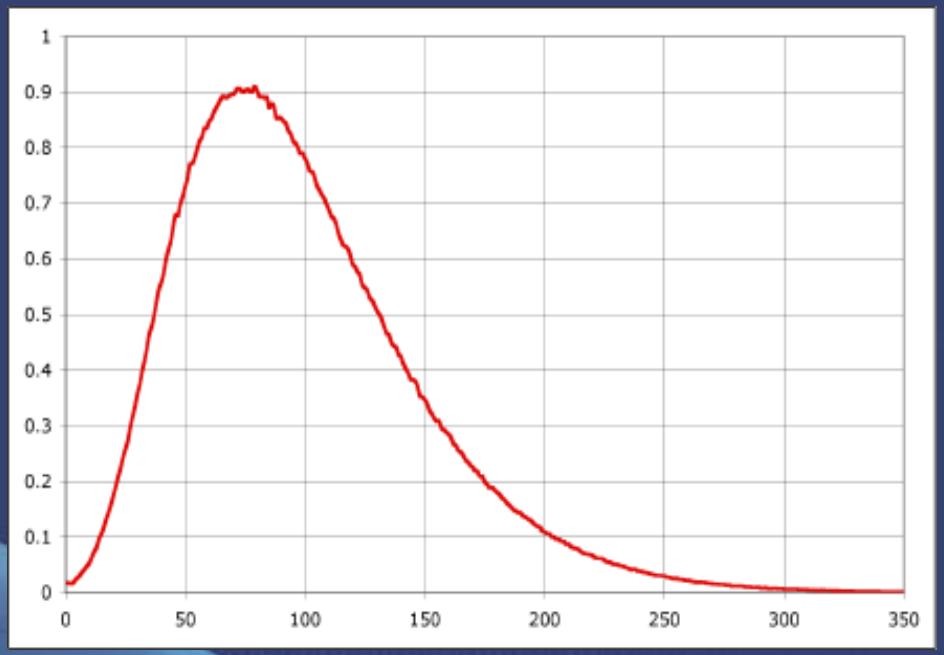
“Artifacts of human discourse”

What are the implications as we move on from storing “book-like” objects and start to operate more fully in the digital domain?

UNIVERSITY OF OXFORD

“Long Tail” Data?

Datasets ordered by Size



Frequency vs Size

UNIVERSITY OF OXFORD

Characteristics of Data

- A good test of repository architecture!
  - No Standardised Formats
    - Collections of files
    - Homegrown formats
  - No Standardised Metadata
    - Domain specific - if at all (http://www.disc-uk.org/datashare.html)
    - Implicit knowledge makes discovery hard
- Strongly bounded
- Restricted **intended** audience

UNIVERSITY OF OXFORD

What are the challenges?

- Context is very important
  - Discovery
    - Metadata - Who, when, where, why?
  - Delivery
    - Visualisation - Relationships
  - Preservation
    - Curation by addition/Sheer curation
      - (http://oxfordrepo.blogspot.com/2008/10/modelling-and-storing-phonetics.html)
    - Format preservation/documentation
    - What are the significant properties?

UNIVERSITY OF OXFORD

The Oxford Approach...

- Content Objects
  - Encapsulate the Dataset/Book/Whatever
- Context Objects
  - People – authors, players, populations?
  - Places – GIS
  - Epochs – wide variety of timescales
- Organisational Objects
  - Collections, Data Abstracts, File Organisation
- Administrative Objects

UNIVERSITY OF OXFORD

In Practice...

- FEDORA Object Model
  - Objects composed of multiple datastreams
    - Content + metadata
  - Composition determined by a Content Model
    - Used for Ingest, Validation, Discovery and Delivery
- Objects related through RDF
- All objects have URI's/UUID's
- Not restricted to FEDORA/Oxford

UNIVERSITY OF OXFORD

An example of what happens...

- A Person...
  - Institutional Repository (http://ora.ouls.ox.ac.uk/)
    - People are Authors and Depositors
  - eAdmin (http://brii.medsci.ox.ac.uk)
    - People have departmental/project affiliations (http://vocab.org/participation), FOAF profiles
  - Correspondence archive (http://futurearchives.blogspot.com/)
    - People have time-based locations, relationships
  - Historical archive
    - Documents provide evidence for personal relationships

UNIVERSITY OF OXFORD

Summary

- Objects (not just datasets) have much more meaning and utility when considered in conjunction with their context
- Object and Semantic techniques give us a way of expressing this coherently and systematically
- Relationships need be qualified in terms of domain and evidence (http://vocab.org/evidence)
  - Relationship become an event/process
    - e.g. Authoring, membership

UNIVERSITY OF OXFORD

Stop right there...

Any questions on this bit

UNIVERSITY OF OXFORD

What is a Book/Paper?

- In an academic context:
  - Captures the state of knowledge at a particular time within a particular domain
- Ontologically
  - Facts and predictions are related to hypotheses and models
  - Relationships are qualified by evidence and domain mediated by debate
- Look familiar?

UNIVERSITY OF OXFORD

It's a wrapper format...

- Human friendly
- But...
  - Not machine parsable
  - Not easy to update
  - Often snapshotted at the wrong time!
  - Context missing
  - Eyeball-limited as a dissemination medium

UNIVERSITY OF OXFORD

A thought...

There are already projects where the primary output is an online resource and this is growing. In the current (2008) review of HE funding in the UK there have been queries about how to submit online databases for review.

UNIVERSITY OF OXFORD

What about VRE's?

- Virtual Research Environments
  - Enable collaborative working on research
  - Predicated on “project” approach
    - Final outputs and project shutdown
- Knowledge does not proceed like that
  - Make the VRE the output
    - ...and the input to the next phase
  - Visualise and capture the process of the evolution of ideas

UNIVERSITY OF OXFORD

What is a VRE?

- Discourse takes place in many places already: email, twitter, wikis
- Use the object store directly
- Already have a semantic model for evidence
- Cultures of Knowledge
- (http://www.culturesofknowledge.org)
- Mediaeval Libraries of Great Britain

UNIVERSITY OF OXFORD

But,,,

There's no reason why a book/paper couldn't be created as a snapshot...

Peter Murray Rust has demonstrated the generation of animations and documentation from raw synthesis data, described here

The SPIDER project looks at semantically enhancing online papers

(http://imageweb.zoo.ox.ac.uk/pub/2008/plospaper/latest)

(http://imageweb.zoo.ox.ac.uk/wiki/index.php/Spider_Project)

UNIVERSITY OF OXFORD

What about Big Science?

- BID Project (http://www.ouls.ox.ac.uk/sers/bid)
  - Link to SRB Grid datasets
    - Harvest metadata to create objects representing datasets
    - Data abstracts developed – but not used
- EIDCSR (Embedding Institutional Data Curation Services for Research)
  - http://e2idcsr.blogspot.com
  - More generic, HFS-based, Fedora Registry

UNIVERSITY OF OXFORD

Questions

Neil Jefferies

neil.jefferies@sers.ox.ac.uk

www.sers.ox.ac.uk

Oxford Research Archive

ora.ouls.ox.ac.uk

Developer's Blog

oxfordrepo.blogspot.com

Google Code

look for: python fedora-commons

UNIVERSITY OF OXFORD

Generics over Specifics

- Identify generic operations in applications
  - Push into the tools layer
  - Simplifies future application development
  - Replacement can upgrade all applications
- A generic capability is much less likely to become obsolete than a specific one
  - Protects investment in development

UNIVERSITY OF OXFORD

Capability Focus

- Every design decision should be reviewed...
  - What does it allow us to do?
  - What might it prevent us from doing that we will really regret later?
- Scalability will always be an issue
- Short workflows
  - Decouple actions as much as possible

UNIVERSITY OF OXFORD

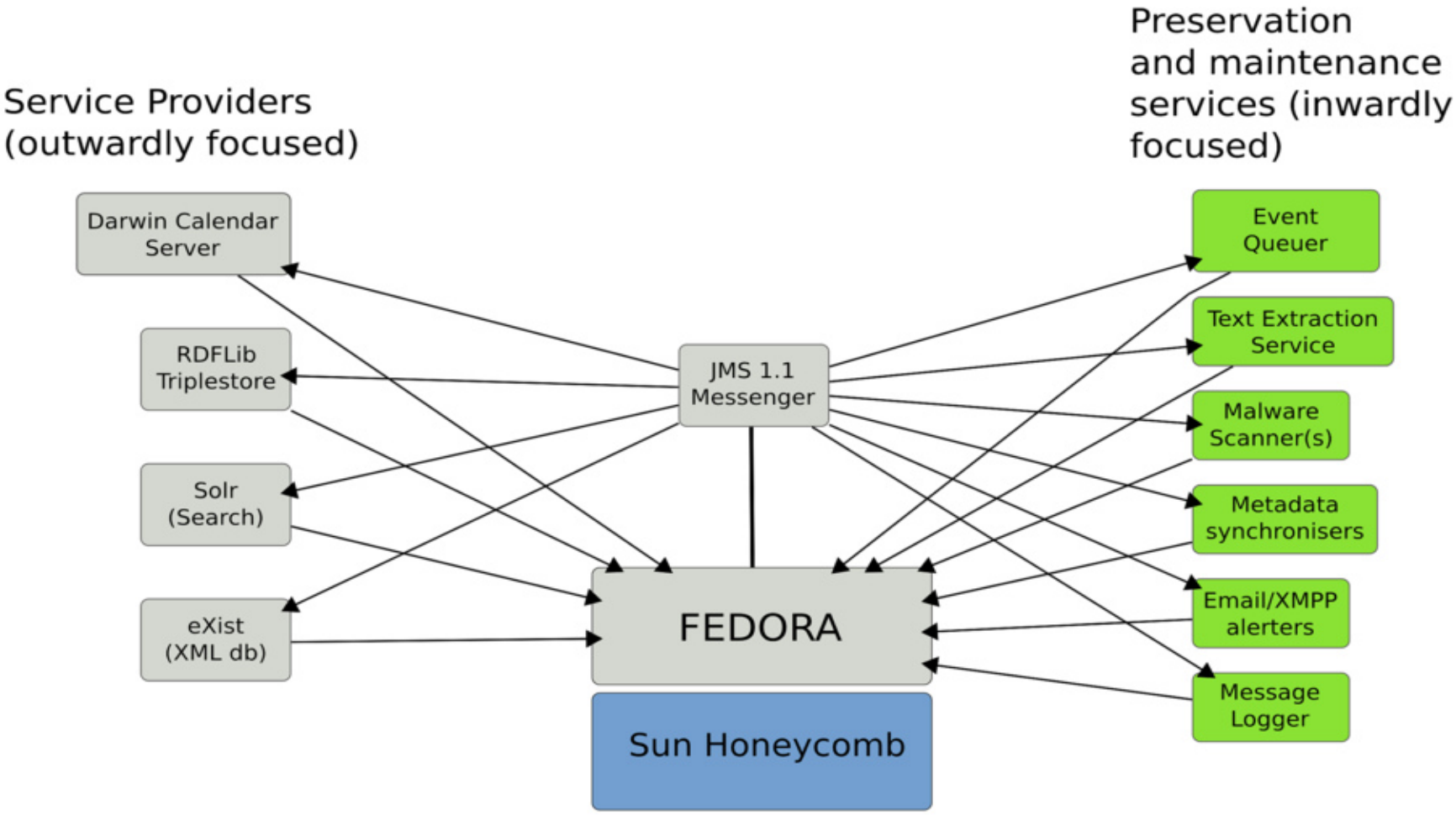
DAMS Architecture

- The practice...

Oxford University Research Archive
ora.ouls.ox.ac.uk

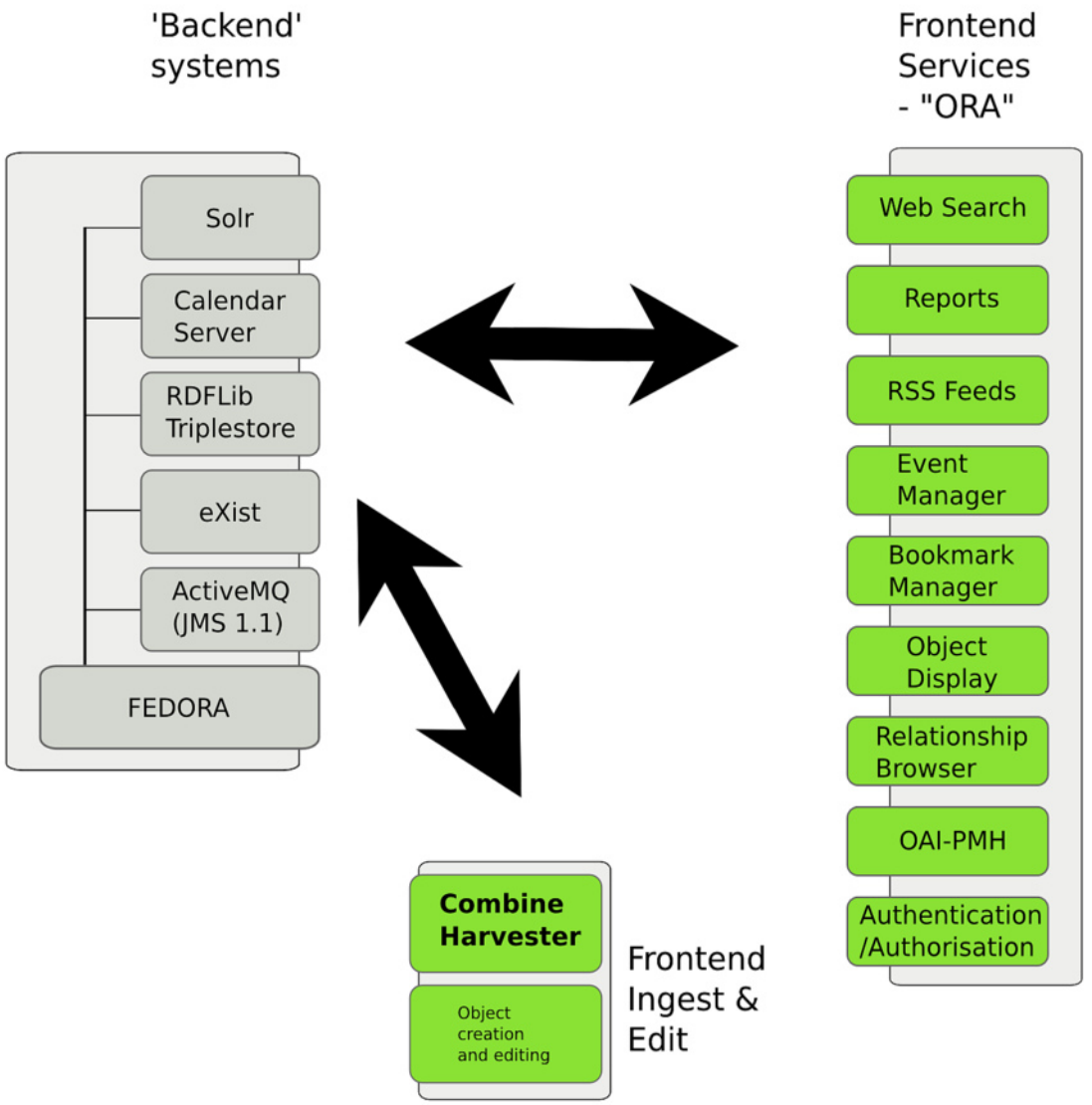
UNIVERSITY OF OXFORD

'Backend' systems - service provision view



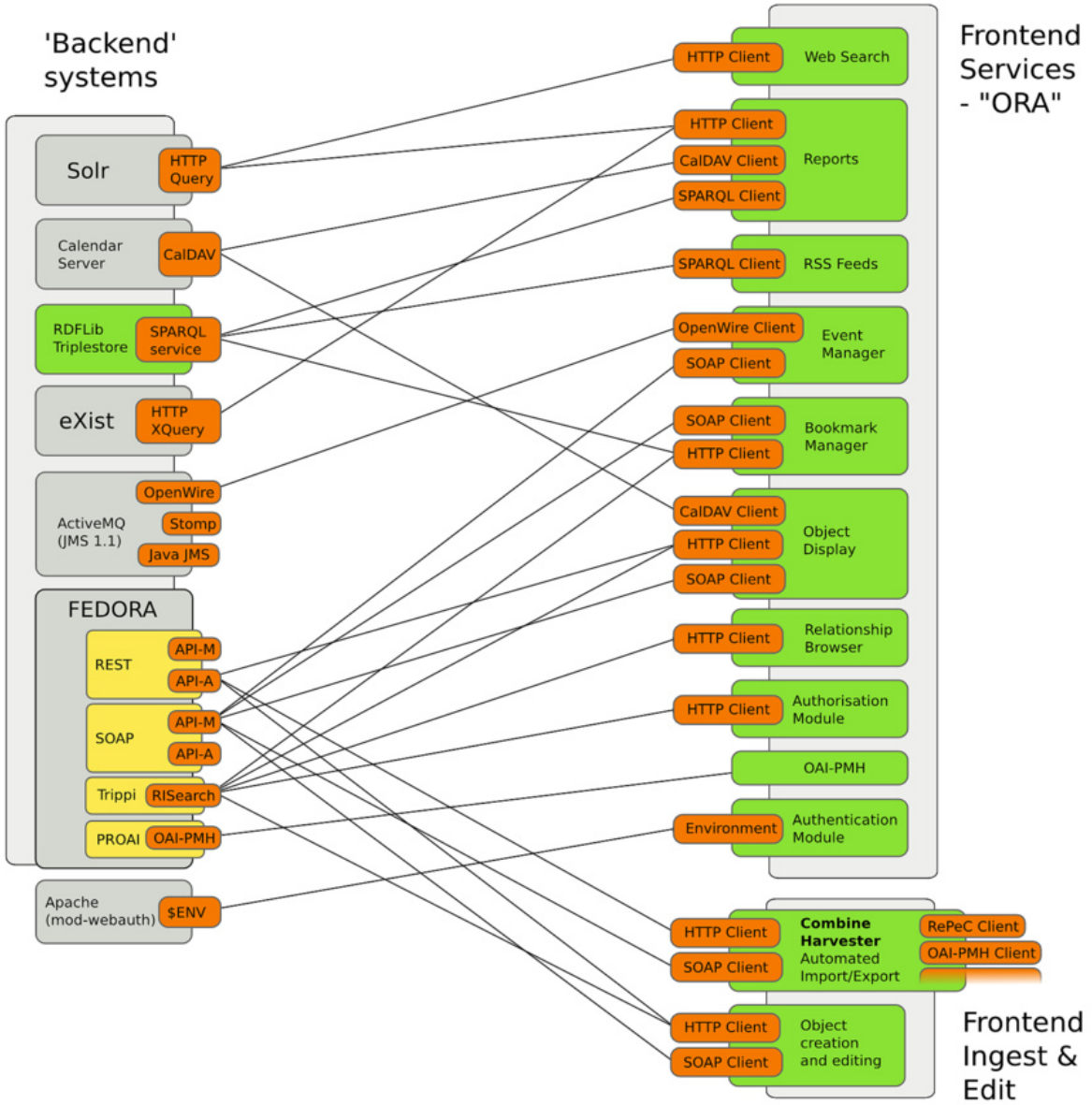
UNIVERSITY OF OXFORD

Frontend to Backend - service view



UNIVERSITY OF OXFORD

Frontend to Backend connections



UNIVERSITY OF OXFORD

Example Design Decisions

- Fedora Commons (www.fedora-commons.org)
  - Flexible object/metadata model
- UUID's preferred over handles for uniqueness
  - Scalability, dependencies
- No OAIS/METS!
  - Longevity, efficiency, scalability
- Calendaring
  - Generic capability, additional benefits

UNIVERSITY OF OXFORD

Interoperability

- BID (Bridging the Interoperability Divide)
  - Oxford Research Archive (Fedora)
  - Virtual Learning Environment (Sakai)
  - Grid Datasets (SRB)
- BRII (Building a Research information Environment)
  - Captures details of interests, funding, projects and links to publications and data
  - http://www.flickr.com/photos/oxfordrepo/2615764936/in/photostream/

UNIVERSITY OF OXFORD

Preservation

- PRESERV2
  - Southampton, The National Archive
  - Preservation tools/services for repositories
  - OR08 – Whole repository migration
- SHERPA DP2
  - Multiple UK Institutions and Arts and Humanities Data Service
  - Provision of outsourced preservation service

UNIVERSITY OF OXFORD

Expansion

- DISC Datashare
  - Bringing datasets into repositories
  - Metadata standards
- PARADIGM/CAIRO
  - Complex objects, personal digital archives
  - Security and longevity are key
- Digitisation: Google, John Johnson...

November 2008

PASIG Fall 2008