Hydra at Oxford

Neil Jefferies
R&D Project Manager, Bodleain Libraries
Neil.Jefferies@bodleian.ox.ac.uk

An important bit of background...

We don't really use FEDORA

- FEDORA **was** a strategic choice in 2007
  - VTLS Vital to provide institutional repository (ora.ox.ac.uk)
  - Project failed, ORA public interface subsequently rewritten from scratch
  - VTLS Valet (open source) continues to provide deposit workflow
  - ORA the only digital content system still based on FEDORA
- What we liked about FEDORA
  - Flexible generic object model – multiple datastreams with versioning
  - Semantic model – RELS-INT/RELS-EXT
  - REST API
  - Storage abstraction
- What we didn't like about FEDORA
  - Unnecessary wrappering – FOXML not good for preservation of active objects
    - Active-use is the major economic justification for preservation
    - No, we don't like METS either!
  - Lack of modularity – external triple-store grief
  - Feature bloat – content models, XACML etc. - made worse by lack of modularity
  - Installation (need I say more)
  - Silent periods

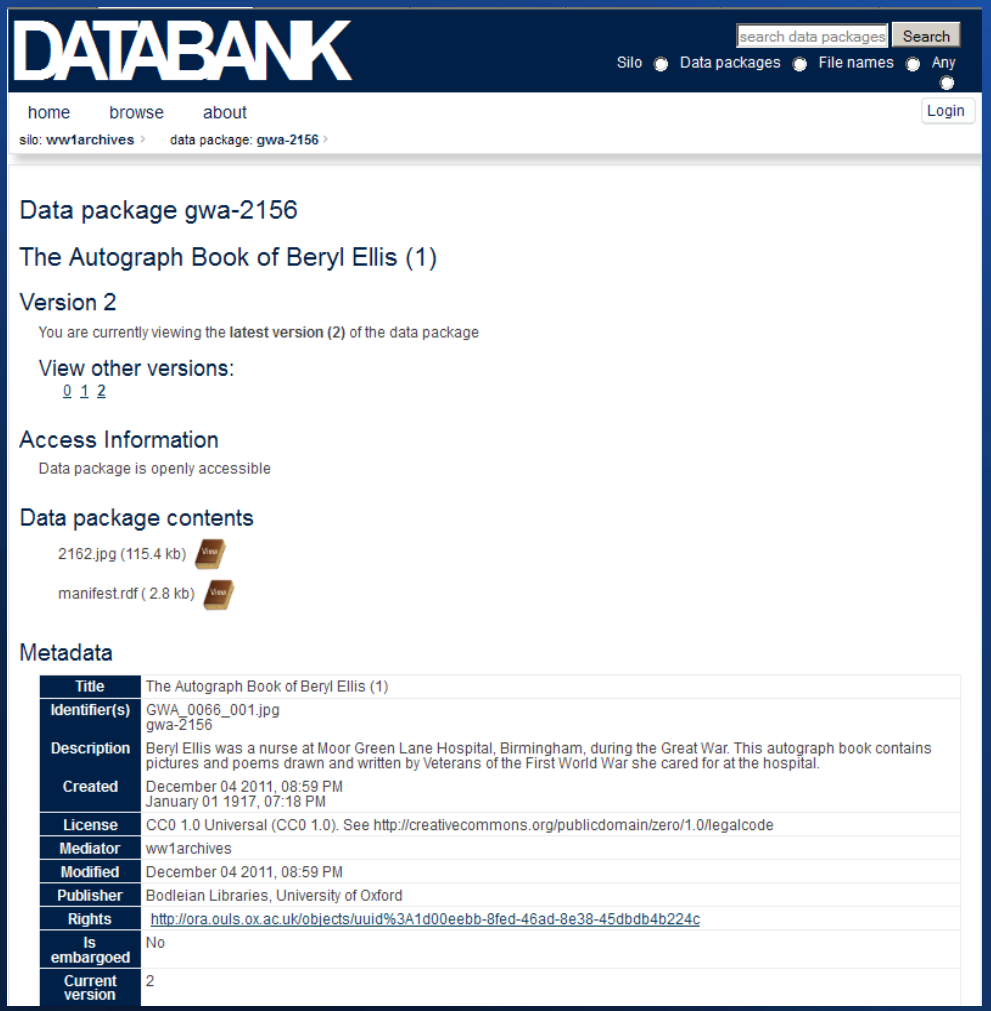
CDL Microservices

- CDL “What's the minimum amount of code you need to add to a filesystem to make it look like a repository?”
  - Pairtree, Namaste, BagIt
  - Pluggable Web Services provide bulk of repository/preservation functionality
- Ben O'Steen “What's the minumum amount of code we need to add to a Micro-services respository to make it look like a FEDORA repository”
  - Enough like a FEDORA repository to do the things we like
  - But retaining the scalability and modularity of the Microservices model

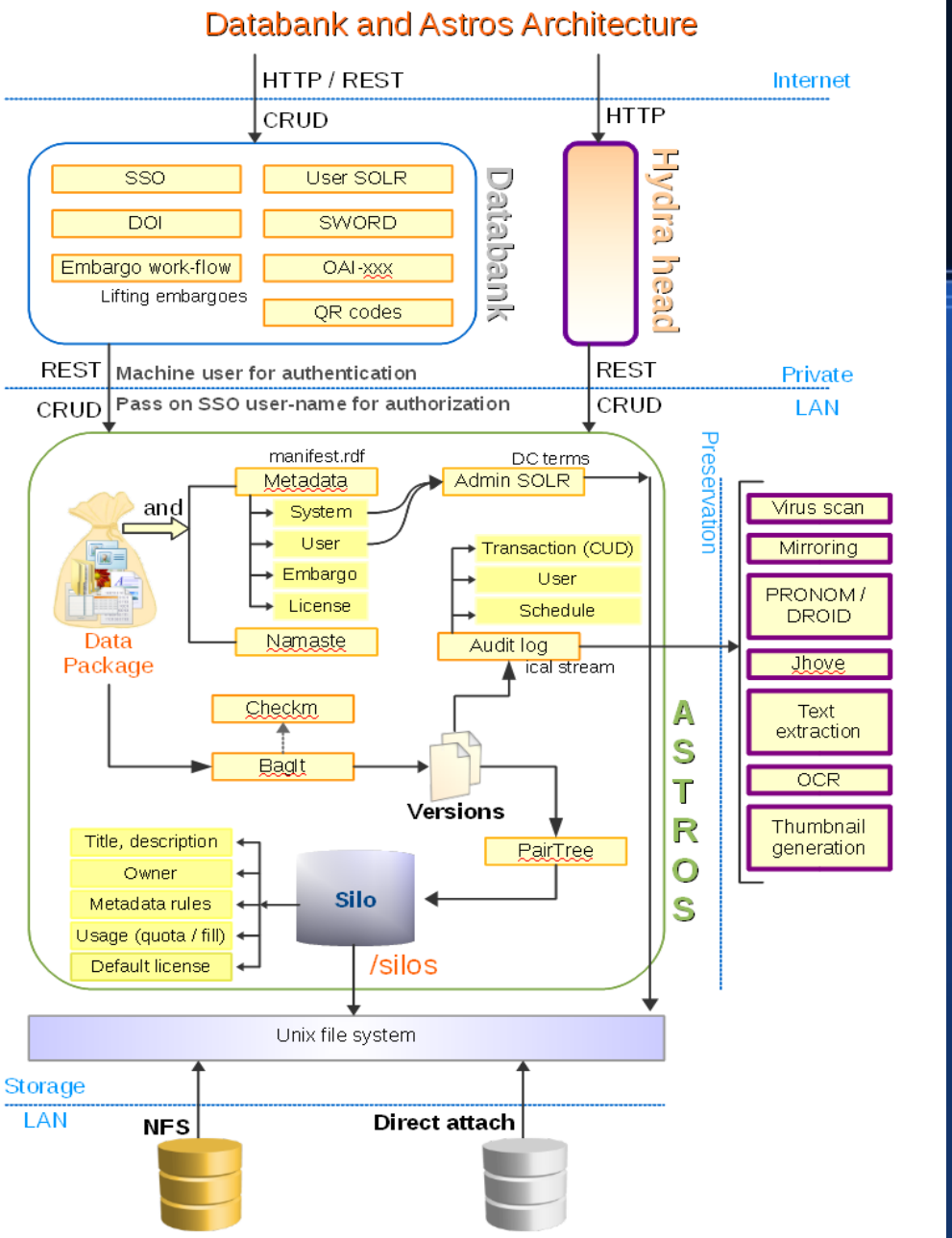
...rather less than we thought!

- Databank came into being in 2008
- **2010-12** JISC/HEFCE (UMF) funded Admiral/DataFlow
  - Prototype/productionise DataStage/DataBank
  - Libraries, Computing Services, OeRC, IBRG, UKOLN, Canonical
  - Lightweight data management/archiving

DataBank



- Bodleian Data Repository (in dev since 2008) parallels ORA
- “Data” currently defined as “Research outputs that don't fit in ORA”
- File and metadata format agnostic
  - supports packages (zip & tar)
  - component subaddressing
- Built on “FEDORA-Lite” object model
- Assigns DataCite DOI's
- Manages embargos
  - Secure, dark archive is segregated
- Manual and SWORD2 deposit
- REST API
- Debian Packages or OVF



Architecture

- Microservices orchestrated using message queues
  - Event streams act as schedule, log and provenence
  - Queues can be exposed externally
- Search/browse interface
  - SOLR built-in
  - REST API provides meaningful responses to requests for text/html
- Multi-streamed RDF Metadata
  - Segregated by type/accessibility
  - Other XML metadata also supported
- Split authentication/authorization
  - Systems integration
- Unix file system semantics
  - Less abstraction?
  - LTFS

So...why Hydra?

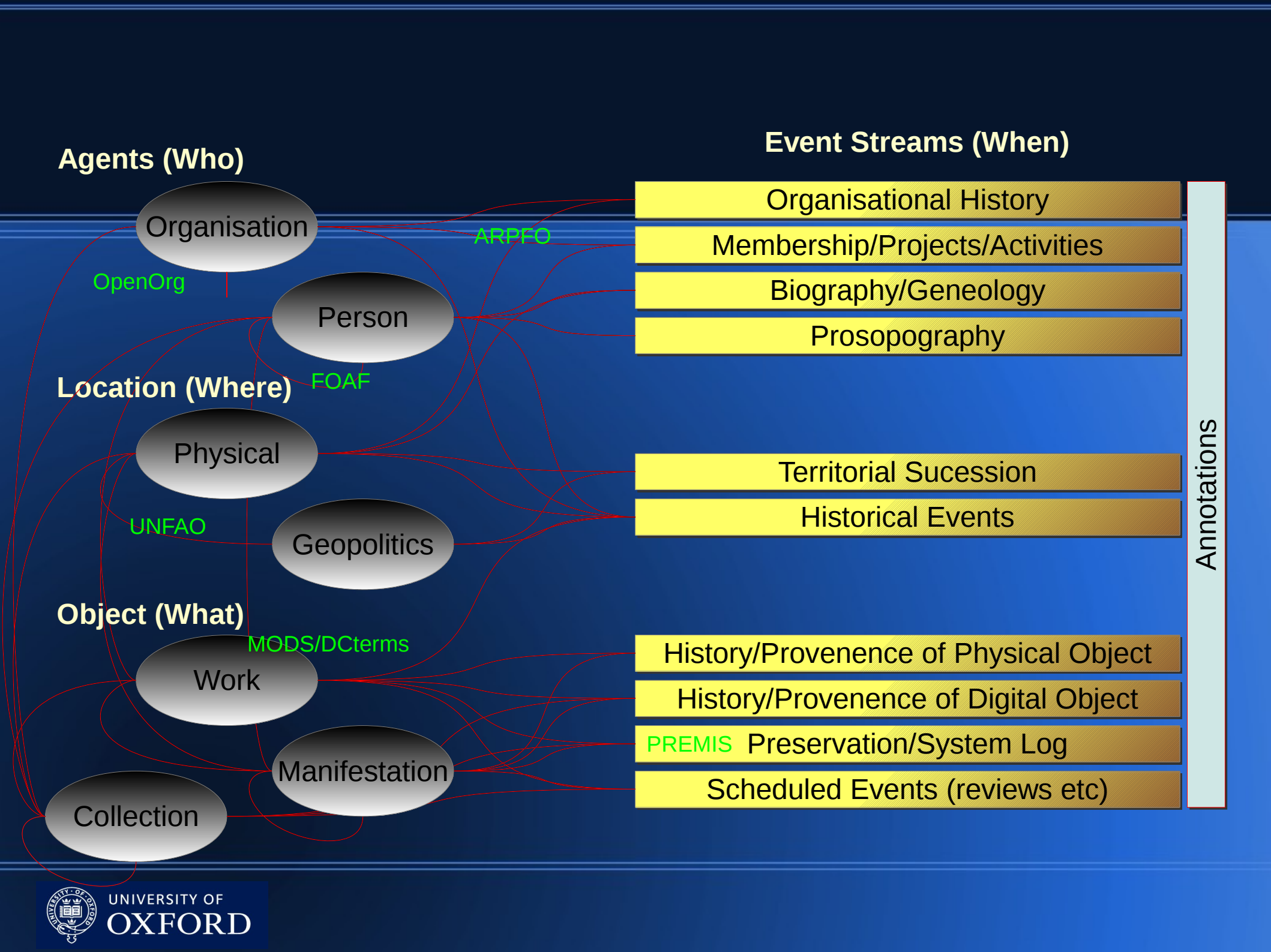
The subset of FEDORA functions implemented by DataBank/ASTROS almost exactly matches the subset used by (most) Hydra heads. This subset, in essence, characterises a generic semantic object store. We can use this for pretty much everything...

Hydra in the near future

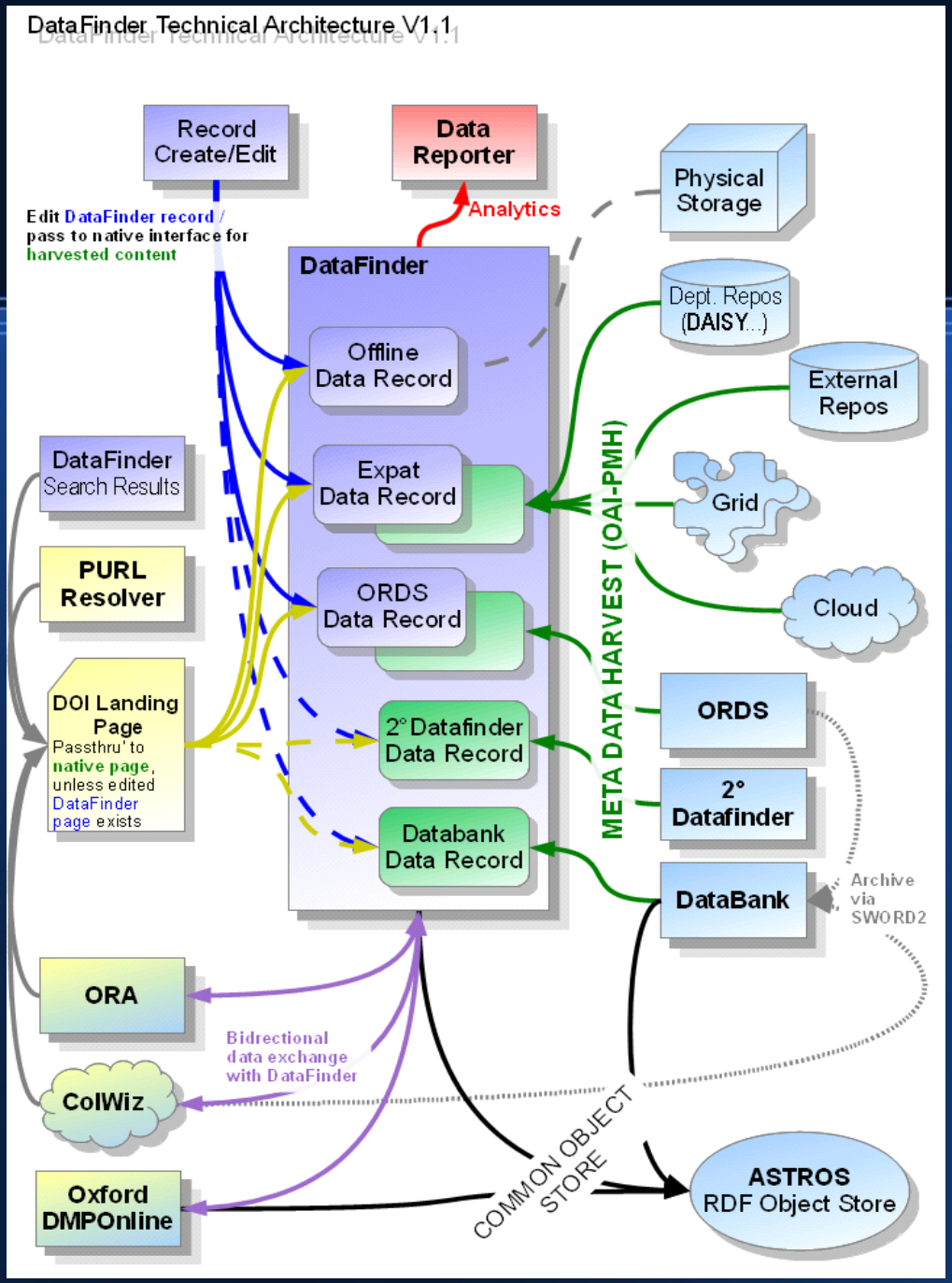
- Replace VTLS Valet for ORA ingest
  - Still running over FEDORA at this point
  - Migration off FEDORA still planned
- Consolidate legacy digitized materials
  - Migrate into ASTROS from many scattered servers/websites
  - Publish through Digital.Bodleian (Armadillo)
  - Metadata is somewhat variable MARC->MODS->METS
    - Need a MODS editor
- Archival materials
  - Physical, hybrid and electronic
  - EAD is problematic
- Shared Canvas/IIIF
  - Viewer encapsulated as a Hydra head
  - Annotation/transcription tools

Catalogue 2.0

- Objects have a context from which much of their meaning is derived
  - Inlcude context objects representing people, places, events etc.
  - Catalogues become contextual skeletons fleshed-out by “traditional” digital objects
  - Authority lists become prosopographical and biographical resources
  - Geopolitical and temporal information
  - Aggregations become a key structural element
  - Should reflect actual knowledge – conflict with cataloguing practice!
- Annotations and files can be attached to any object
  - Context objects hold content...what is the difference?
- Objects are not static – preservation challenge...and benefit!
- This model works for almost any content from any period



UNIVERSITY OF OXFORD



DataFinder

- Catalogue/registry of research data
  - Wherever and whatever it is!
  - OAI-PMH harvesting of external data stores
  - Manual record entry for non-electronic or non-harvestable data
- Search/browse interface
- DataReporter module
  - CERIF compatible
  - Analytics as well as content statitics
- Core Metadata schema based on DataCite
- Interfaces with many systems
  - “Hub” Of RDM activity
- Hierarchical architecture
  - Local catalogues, subjects specific or inter-institutional catalogues possible

Questions?

Neil Jefferies
R&D Project Manager, Bodleain Libraries
Neil.Jefferies@bodleian.ox.ac.uk