Oxford Digital Library Infrastructure

Matt McGrattan, Digitisation Service Manager,
Bodleian Digital Library Systems & Services.

Outline

- Background
- Workflow
  - Images
  - Storage
  - Delivery

Background / Who am I?

- BDLSS: Bodleian Digital Libraries Systems and Services.
    - Digitisation.Foundations project: to create a new digitisation workflow, from image capture to long-term storage, for new digitisation projects.
    - Digital.Bodleian project: to migrate all of our legacy digitised collections to a common repository, common set of file formats, and a common search and discovery interface.
- Image capture / digitisation is a different department.

Workflow (1)

- Formerly bespoke (and baroque) mix of:
  - MS Access & MySQL databases
  - MS Access & browser based ‘Ajax’ front ends
- With:
  - Imagemagick for image processing
- Scripted via a mix of PHP, Python, Perl and … VBScript.
- Historically constrained by local storage, network bandwidth, and CPU on the image processing server.

Workflow (2)

- Currently moving to Goobi for project orders (approx. 500,000 images a year)
- ‘Legacy’ system still in use for small commercial orders although due for replacement.
- Move to Goobi accompanied by a new hardware infrastructure with dedicated server cluster; 1G and 10GigE networking, and 40TB of local working storage.

Images (TIFFs)

- Formerly TIFFs only
- Stored on tape with checksums
- Approx. 60TB
- TIFFs as preservation masters, not delivery.

Images (JPEG2000)

- Goobi outputs stored as lossless jpeg2000s
- Legacy TIFFs being converted to lossless jpeg2000s - approx. 5-10% done.
- Files converted using Kakadu
- Single file for preservation and delivery

JPEG2000 (2)

- Profile:
  -rate - Creversible=yes Clevels=6 "Cprecincts={256,256},{256,256},{128,128}" Corder="RPCL" ORGgen_plt=yes ORGtparts="R" Cblk="{64,64}" Cuse_sop=yes Cuse_eph=yes

- Lossless, precincts, RPCL order, no tiling.

Colour Management (1)

- First 5 -10 years of digitisation, little or no colour management. But …
- Targets were shot and images kept.

We would be keen to explore automated methods of generating profiles from our legacy targets.

Colour Management (2)

TIFF workflow:

- Gretag Macbeth Digital Colorchecker SG
- Custom ICC profile
- Files converted to Adobe RGB on ingest into the HFS tape archive.

Colour Management (3)

JPEG2000s:

- As with tiffs, but files converted to sRGB.

Issues:

- Gamut
- Best-practice: source vs. display profiles; ICC in JP2.

Keen to explore standardisation of approaches.

Technical metadata

- Goobi:

Store XMP as a sidecar XML file in the Databank dataset for each image.

- Legacy content (Digital.Bodleian):

Currently, whatever the JP2 XMP box contains after conversion with Kakadu.

TIFFs still exist, so technical metadata can be extracted or re-embedded.

Keen to explore standardisation of approaches to storage of technical metadata.

Storage

- ‘HFS’
  - Centrally university managed
    - Tape-based
- Databank
  - Bodleian ‘repository’ / preservation store

‘HFS’: 1

- Bodleian have been digitising since the 1990s.
- Oldest image archives date to around 2000, and are still in use.
- Centrally managed tape store with approx. 60TB of TIFF images.

‘HFS’: 2

- In-house browser UI for scripted item retrieval
- Multiple, largely comprehensive but not completely commensurable indexes of the content.
- Scripted ingest of content into the archive on a nightly basis from studio servers.
- Slow access to material.

‘Databank’: 1

- ‘Spinning disk’-based.
- BDLSS managed / developed.
- REST API.
- Silos / Datasets / Items.

‘Databank’: 2

- Supports versioning.
- Micro-services based model.
- ‘REST’ model and well-documented API makes it easy to develop new applications that push into Databank (e.g. from Goobi, using our python framework).
- ‘REST’ model, on the other hand, makes image delivery via JP2s problematic.

‘Databank’: Pairtree

- https://confluence.ucop.edu/display/Curation/PairTree

- Maps identifiers to file system paths in pairs, e.g.

  ID: 00581637-cda7-4c34-86fe-c454361450e5

  Filepath: /[root]/00/58/16/37/-c/da/7-/4c/34/-8/6f/e-/c4/54/36/14/50/e5/

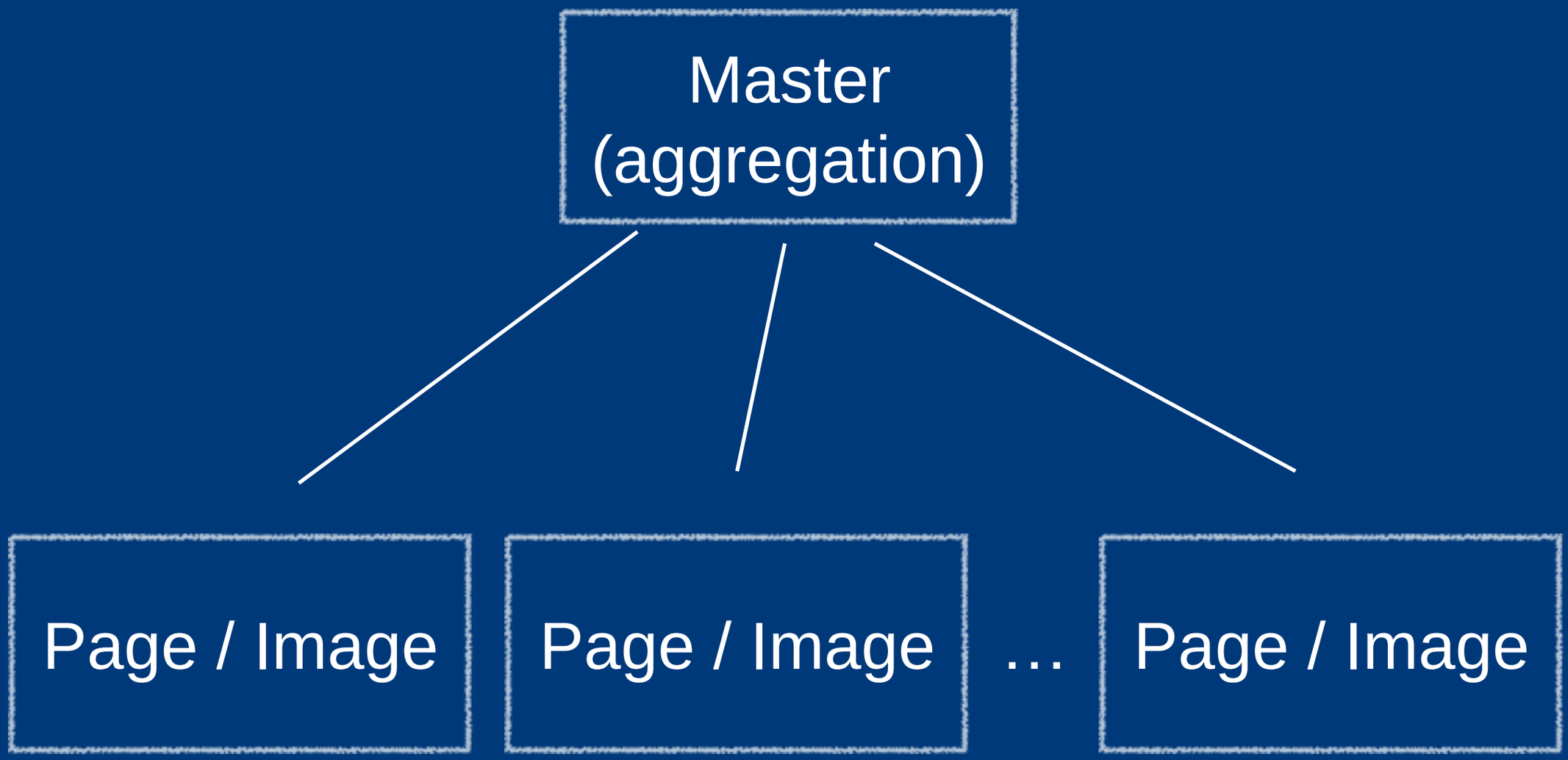
Image Delivery from Databank

- ‘REST’ API makes retrieving parts of files difficult
- Solution: read-only mounts of file store on VMs with no public access
- Pairtree requires an additional layer to pass file paths derived from IDs to image servers
- Solution: python ‘shim’ with Apache mod_rewrite / proxy

Databank in use

- Currently our Goobi workflow feeds content directly into Databank
- The Digital.Bodleian project is migrating legacy content into Databank.
  - Approximately 150,000 of 3,000,000 images done.
  - Migration ramping up now.
- iNQUIRE / Digital.Bodleian is delivering content from Databank.

Data structure



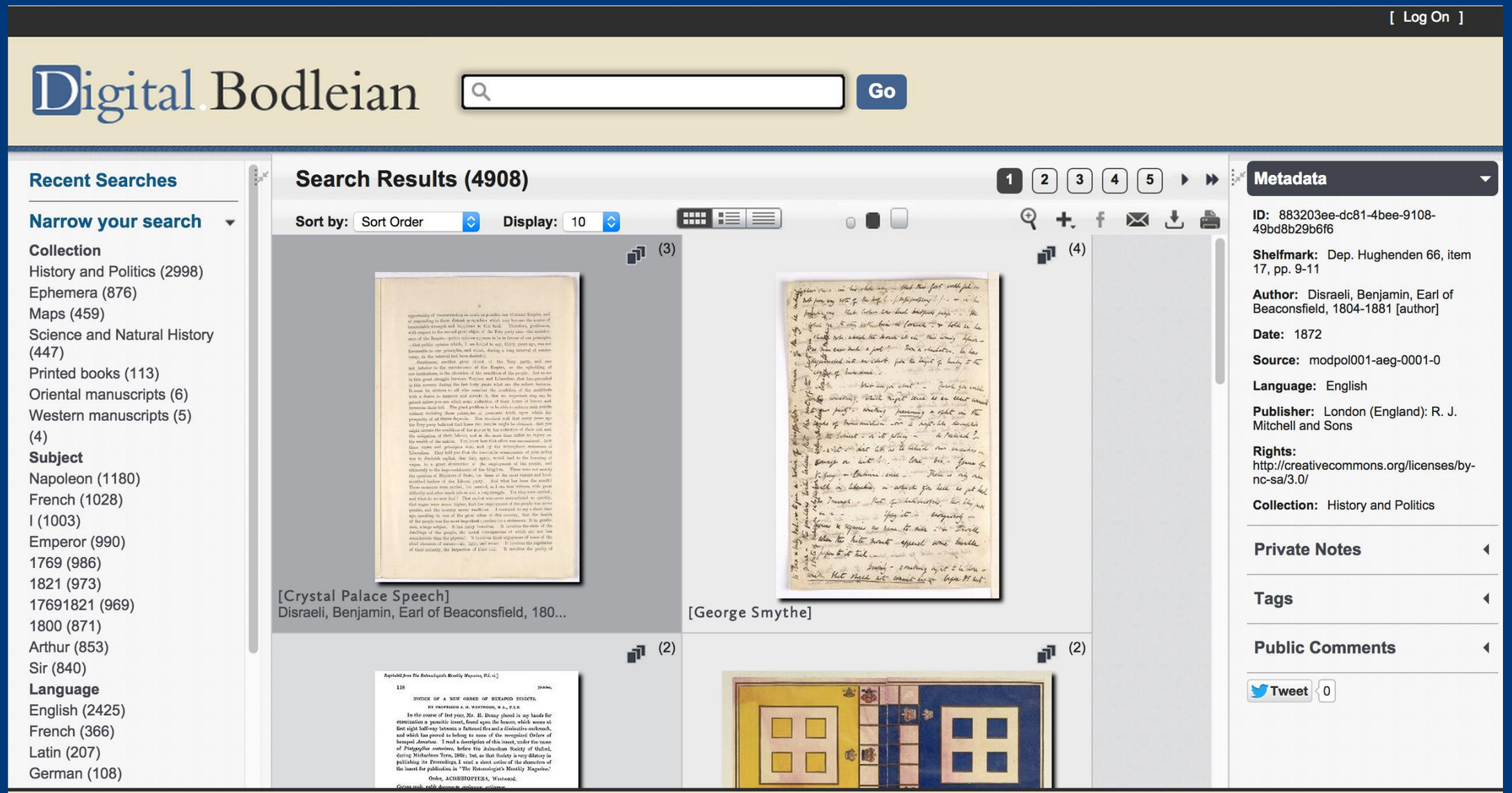
Typical ‘master’ dataset

- https://databank.ora.ox.ac.uk/digital.bodleian/datasets/75433b44-be24-4dbc-ac79-e49937d2d499
- Contents:
  - manifest (RDF)
  - DC (for Digital.Bodleian ingest)
  - DC (for OAI)
  - METS (source)
  - jpeg
  - thumbnail
  - lossless jp2

Image Delivery

Image Delivery (Background)

- Lots of incompatible sites using a wide-range of application methods including Zoomify, Djatoka, static jpegs, and Luna.
- Moves to standardise on Digital.Bodleian as the primary entry point for our general image collection, with search and discovery and rich user experience.
- Standardising on Viewer.Bodleian as the primary method for serving up single items, or small collections with no search/discovery.
- Exploring Mirador (2.0) and possibly the Wellcome player as we move increasingly to a IIIF API based environment.



Digital.Bodleian

Legacy content, and search and discovery.

Digital.Bodleian Frontend

- http://digital-d2v.bodleian.ox.ac.uk/inquire_1.14/
- iNQUIRE (currently 1.14) from Armadillo Systems.
- HTML5 / Ajax based
- Commercial, not open-source
- Windows/ ASP.NET ’stack’

Digital.Bodleian Backend

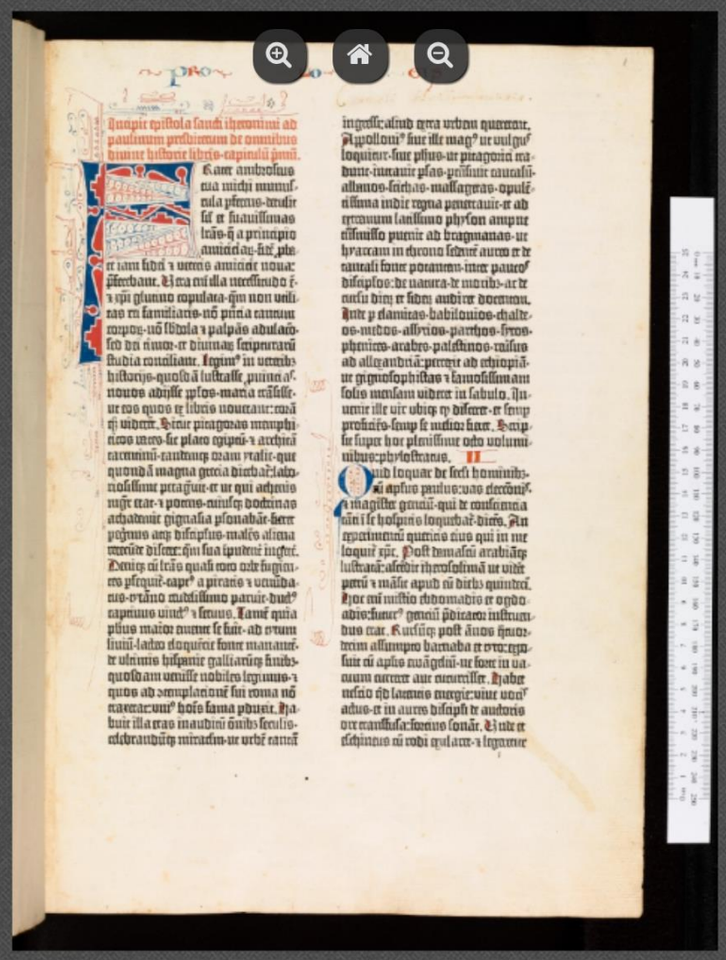
- Solr 4.10 for Search / Indexing.
- IIP Image for tiled image delivery.
- Kakadu 7.2.
- Varnish cache / memcached.
- Apache mod_rewrite / mod_proxy layer with Python ‘shim’ to convert UUIDs to absolute file paths & handle multiple ID to file path mapping schemes.

Digital.Bodleian Metadata

- Legacy collections mapped to DC + a few local Oxford-specific fields.
- Source metadata, e.g. METS, retained in the Databank dataset for the master ‘parent’ record.
- Currently, data is harvested out of Databank and ingested into Digital.Bodleian.
- In future, the two will be more tightly coupled together.

Digital.Bodleian Status

- Approx. 150,000 live images served to Europeana.
- Major launch with a complete UI overhaul planned for January - N.B. iNQUIRE 1.14 is a development / beta release.
- Performance enhancements with load-balancing, better caching, and additional IIP based tile servers.
- Ingest of approx. 200,000 more images before January.
- ‘Live’ pipe directly from digitisation workflow to iNQUIRE for out of copyright material in near future.



Viewer.Bodleian Frontend

- http://viewer.bodleian.ox.ac.uk/icv/page.php?book=ms._kennicott_3
- In-house development
- PHP, Javascript, OpenSeadragon
- Easy to style, quick to deploy.
- Initially developed for Polonsky Foundation funded collaboration with the Vatican.

Viewer.Bodleian Backend

- Simple JSON files for structure / labelling / metadata / links
- DeepZoom with pre-generated image tiles (via vips from TIFFs)
- Varnish cache + load balancing
- Semi-automatic ‘hook’ into our HFS tape archive for quick delivery of existing content

Viewer.Bodleian Status

- Currently in regular use for Vatican/Polonsky project and a large Chinese digitisation project.
- Also being offered to partner institutions within Oxford with content hosted on Bodleian servers.
  - http://viewer.bodleian.ox.ac.uk/christchurch/page.php?book=ms._92
- Switch to dynamic serving of image tiles via the same IIP-based ‘stack’ used for Digital.Bodleian.
- Switch from Bodleian-specific JSON format to IIIF Presentation API (or a subset of the IIIF Presentation API)
- Open-source it

Public & Internal Image Service via APIs

- Content in Databank can be delivered to applications (using IIP) via:
  - IIIF
  - DeepZoom
  - IIP
- and via Djatoka

IIIF endpoints to be made public once more robust file-level authorisation and authentication mechanisms are in place.

Summary

- Goobi-based workflow
- Lossless JPEG2000s for preservation and delivery
- ‘Databank’ as repository
- Image delivery via IIP (and some legacy Djatoka uses)
- Image viewing via iNQUIRE (Digital.Bodleian) and Viewer.bodleian