digirati
DIGITAL BUSINESS INNOVATORS

digirati

DIGITAL BUSINESS INNOVATORS

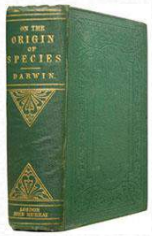
Tom Crane, Technical Director
tom.crane@digirati.co.uk

Existing infrastructure

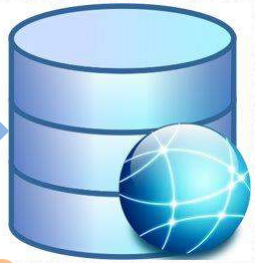
- Encore (discovery)
- Catalogue
- Membership

The Wellcome Library did not want to introduce an off-the-shelf “silo” digital library product

sierra innovative



cat record: b18021013



Digital Asset Management System (SDB / Preservica)

Digitisation Workflow (Goobi)

METS

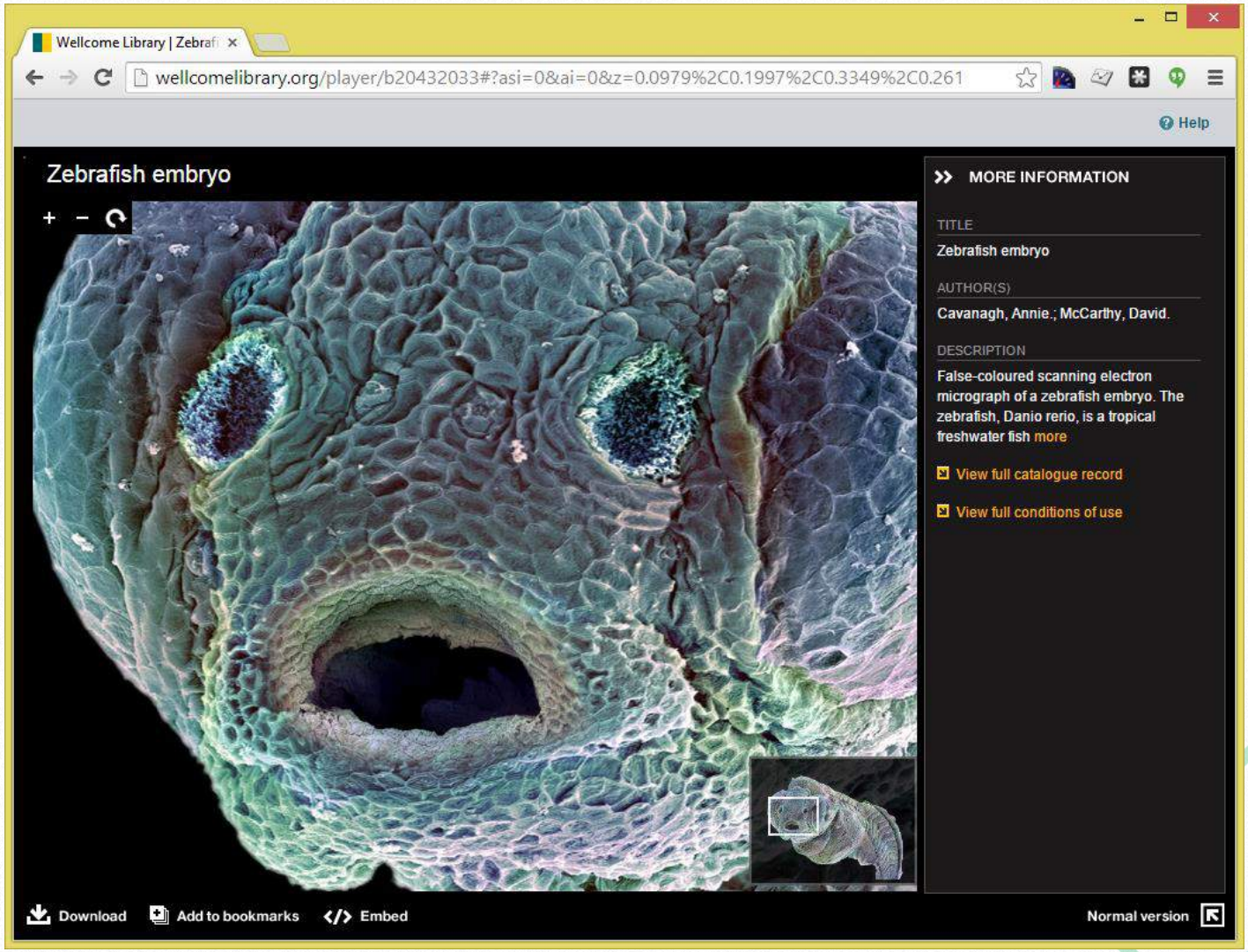
b18021013.xml

METS-ALTO

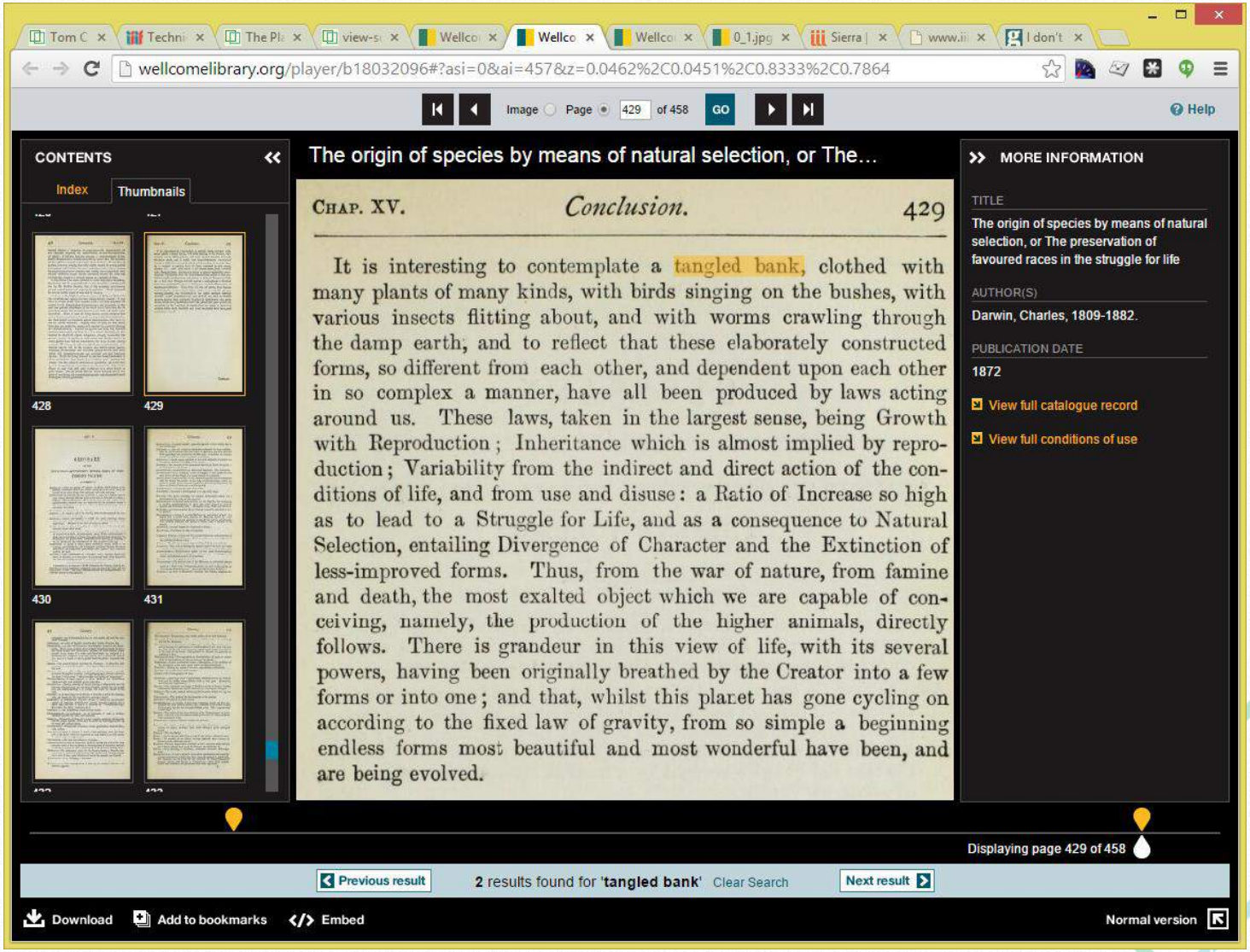
b18021013_001.xml
b18021013_002.xml
b18021013_003.xml
…

digirati

How do we get to this?



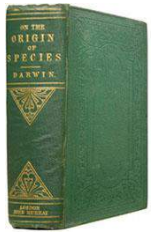
Or this?


wellcomelibrary.org/player/b18032096#?asi=0&ai=457&z=0.0462%2C0.0451%2C0.8333%2C0.7864
Image Page 429 of 458 GO
Help
CONTENTS
Index
Thumbnails
428
429
430
431
The origin of species by means of natural selection, or The…
CHAP. XV. Conclusion. 429
It is interesting to contemplate a tangled bank, clothed with many plants of many kinds, with birds singing on the bushes, with various insects flitting about, and with worms crawling through the damp earth, and to reflect that these elaborately constructed forms, so different from each other, and dependent upon each other in so complex a manner, have all been produced by laws acting around us. These laws, taken in the largest sense, being Growth with Reproduction; Inheritance which is almost implied by reproduction; Variability from the indirect and direct action of the conditions of life, and from use and disuse: a Ratio of Increase so high as to lead to a Struggle for Life, and as a consequence to Natural Selection, entailing Divergence of Character and the Extinction of less-improved forms. Thus, from the war of nature, from famine and death, the most exalted object which we are capable of conceiving, namely, the production of the higher animals, directly follows. There is grandeur in this view of life, with its several powers, having been originally breathed by the Creator into a few forms or into one; and that, whilst this planet has gone cycling on according to the fixed law of gravity, from so simple a beginning endless forms most beautiful and most wonderful have been, and are being evolved.
MORE INFORMATION
TITLE
The origin of species by means of natural selection, or The preservation of favoured races in the struggle for life
AUTHOR(S)
Darwin, Charles, 1809-1882.
PUBLICATION DATE
1872
View full catalogue record
View full conditions of use
Displaying page 429 of 458
Previous result
2 results found for 'tangled bank' Clear Search
Next result
Download
Add to bookmarks
Embed
Normal version

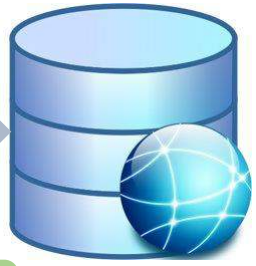
Features

- Deep Zoom
- Search within
- Bibliographic metadata
- Access control
- Thumbnails
- Standalone player, can be embedded
- Audio, video, Born-digital

- http://wellcomelibrary.org/player/b18032096
- http://blog.wellcomelibrary.org/2014/10/a-victorian-lunatic-asylum-begins-to-reveal-its-secrets/



cat record: b18021013



Digital Asset Management System

Digitisation Workflow

METS

b18021013.xml

Read METS

Read JP2, Audio, Video

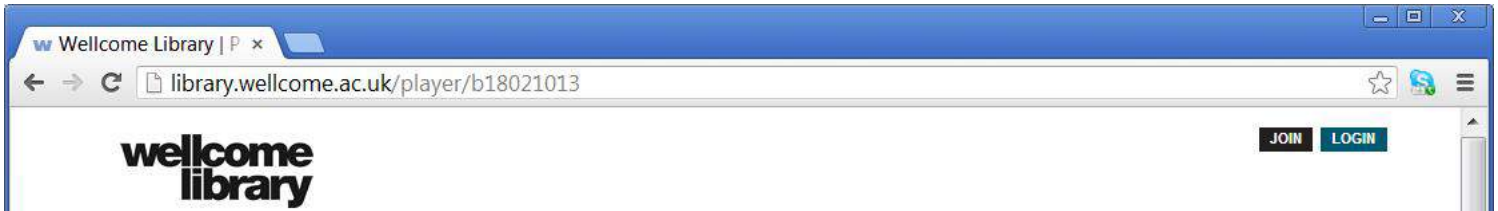
DDS

Request data about b18021013

Serve JSON data

Tile request

Serve JPG tile generated from JP2



The DDS

- Takes information from METS, METS-ALTO, Catalogue to generate **presentation** data
- Provides Image services to generate deep zoom tiles, arbitrary crops, and PDFs
- Provides APIs for search within, authentication and authorisation and biblio data

- Package: http://wellcomelibrary.org/package/b20432033 (will be IIIF eventually)

- Tiles: http://wellcomelibrary.org/dz/b20432033/0/d7b80202-a1d7-4f88-ab57-0f6b369fe41b.jp2_files/12/2_3.jpg

- Services:

wellcomelibrary.org/service/autocomplete/b18032096/0?term=ban


```
[
  "band",
  "bank",
  "banish",
  "bantam",
  "bankiva",
  "bantams",
  "banished",
  "bandicootsand"
]
```


wellcomelibrary.org/service/search/b18032096/0?t=tangled

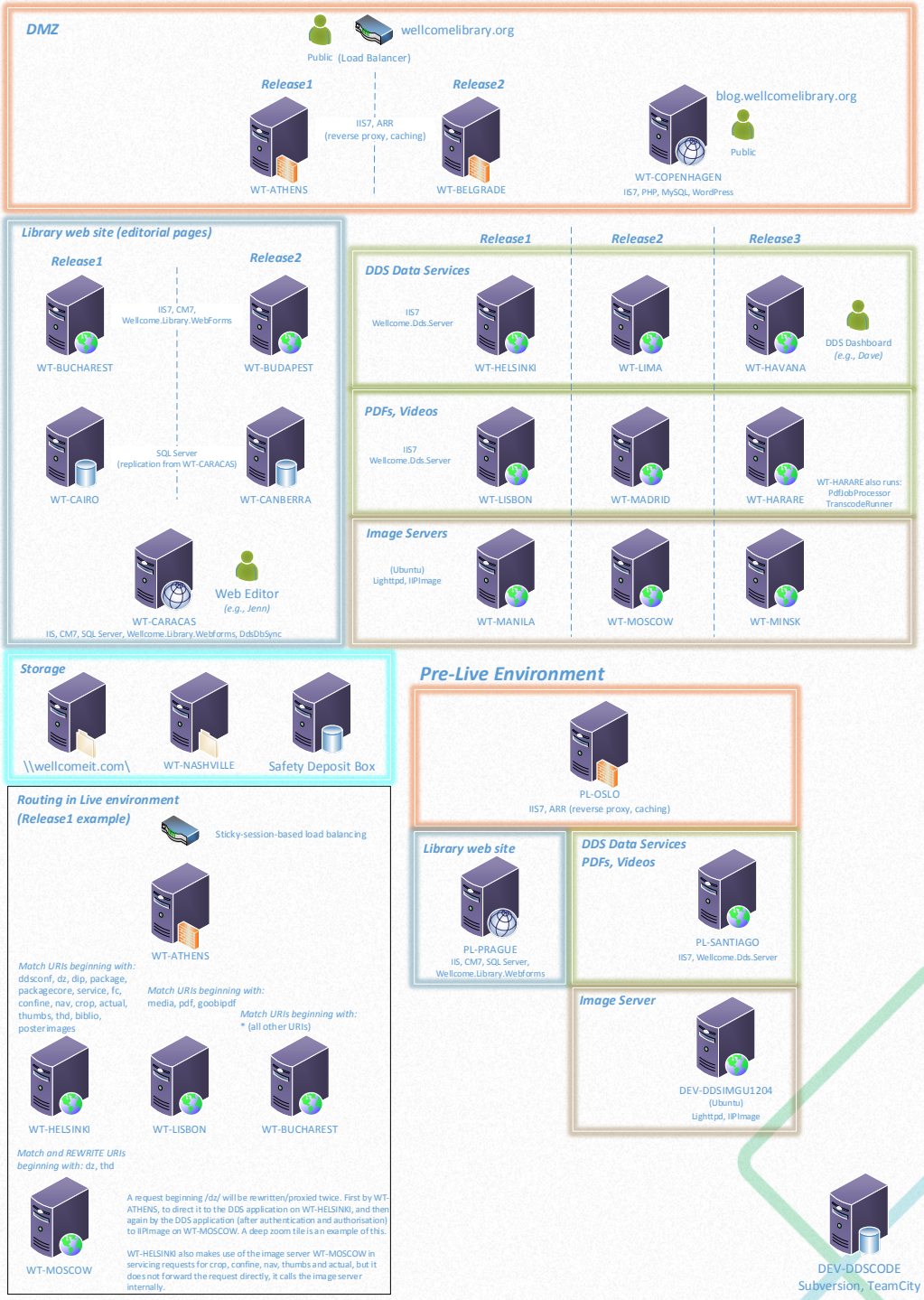

```
      "y": 1074,
      "w": 188,
      "h": 44,
      "hit": 0,
      "before": " cases it can be shown that widely-different checks act on the same species in different districts. When we look at the plants and bushes clothing an ",
      "word": "entangled",
      "after": " bank, we are tempted to attribute their proportional numbers and kinds to what we call chance. But how false a view is this! Every one has heard that"
    }
  ]
 },
 {
  "index": 222,
  "rects": [
    {
      "x": 1041,
      "y": 304,
      "w": 203,
      "h": 45,
      "hit": 1,
      "before": "ng like long oars, they swept a branch rapidly 0 across the object-glass of my microscope. When a branch was placed on its face, the vibracula became ",
      "word": "entangled,",
      "after": " and they made violent efforts to free themselves. They are supposed to serve as a defence, and may be seen, as Mr. Busk remarks, “to sweep slowly and"
    }
  ]
 },
 {
  "index": 457,
```

The DDS

- A combination of existing software and bespoke application

- IIPImage for generating jpeg tiles and other images from archival JPEG2000 files
- Bespoke software for generating JSON packages and other API

The DDS


DMZ
wellcomelibrary.org
Public (Load Balancer)
Release1
Release2
IIS7, ARR
(reverse proxy, caching)
WT-ATHENS
WT-BELGRADE
blog.wellcomelibrary.org
Public
WT-COPENHAGEN
IIS7, PHP, MySQL, WordPress
Library web site (editorial pages)
Release1
Release2
IIS7, CM7,
Wellcome.Library.WebForms
WT-BUCHAREST
WT-BUDAPEST
SQL Server
(replication from WT-CARACAS)
WT-CAIRO
WT-CANBERRA
Web Editor
(e.g., Jenn)
WT-CARACAS
IIS, CM7, SQL Server, Wellcome.Library.Webforms, DdsDbSync
Release1
Release2
Release3
DDS Data Services
IIS7
Wellcome.Dds.Server
WT-HELSINKI
WT-LIMA
WT-HAVANA
DDS Dashboard
(e.g., Dave)
PDFs, Videos
IIS7
Wellcome.Dds.Server
WT-LISBON
WT-MADRID
WT-HARARE
WT-HARARE also runs:
PdfJobProcessor
TranscodeRunner
Image Servers
(Ubuntu)
Lighttpd, IIPImage
WT-MANILA
WT-MOSCOW
WT-MINSK
Storage
\\wellcomeit.com\
WT-NASHVILLE
Safety Deposit Box
Routing in Live environment
(Release1 example)
Sticky-session-based load balancing
WT-ATHENS
Match URIs beginning with:
ddsconf, dz, dip, package,
packagecore, service, fc,
confine, nav, crop, actual,
thumbs, thd, biblio,
posterimages
Match URIs beginning with:
media, pdf, goobipdf
Match URIs beginning with:
* (all other URIs)
WT-HELSINKI
WT-LISBON
WT-BUCHAREST
Match and REWRITE URIs
beginning with: dz, thd
WT-MOSCOW
A request beginning /dz/ will be rewritten/proxied twice. First by WT-ATHENS, to direct it to the DDS application on WT-HELSINKI, and then again by the DDS application (after authentication and authorisation) to IIPImage on WT-MOSCOW. A deep zoom tile is an example of this.
WT-HELSINKI also makes use of the image server WT-MOSCOW in servicing requests for crop, confine, nav, thumbs and actual, but it does not forward the request directly, it calls the image server internally.
Pre-Live Environment
PL-OSLO
IIS7, ARR (reverse proxy, caching)
Library web site
PL-PRAGUE
IIS, CM7, SQL Server,
Wellcome.Library.Webforms
DDS Data Services
PDFs, Videos
PL-SANTIAGO
IIS7, Wellcome.Dds.Server
Image Server
DEV-DDSIMGU1204
(Ubuntu)
Lighttpd, IIPImage
DEV-DDSCODE
Subversion, TeamCity

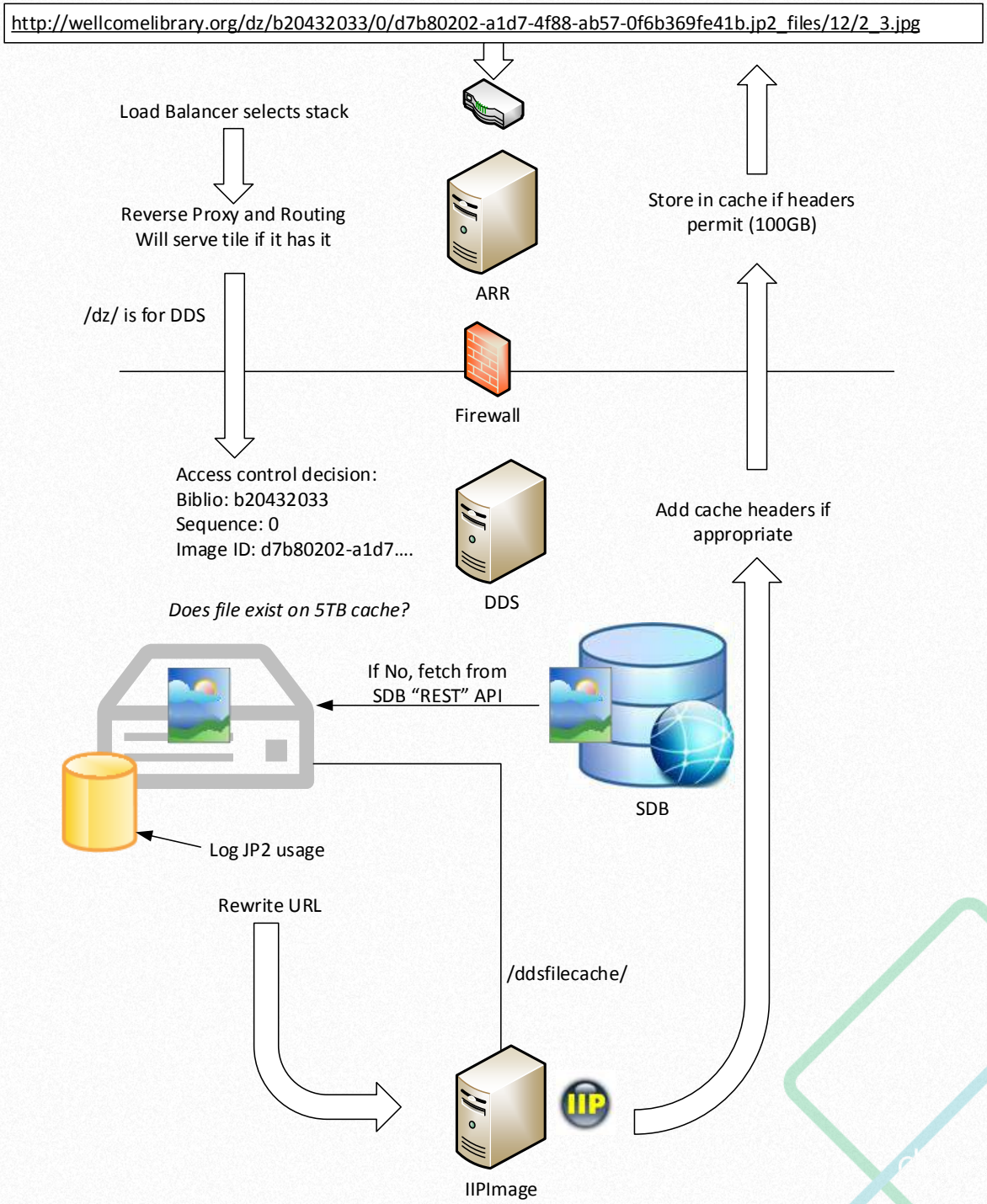
Challenge: tiles at scale

- SDB (DAM) is a long way from the public facing web application
- SDB provides one way to get a JPEG2000:
  http://sdb.wellcome.ac.uk/sdb/rest/digitalFileContents/{0}

- IIPImage needs a local file path, not a URL – can't have fast random access to file like that
  http://WT-MANILA/cgi-bin/iipsrv.fcgi?DeepZoom=/path/to/image.jp2_files/12/2_3

/path/to/image.jp2 – needs to be a local file system on the IIPImage server

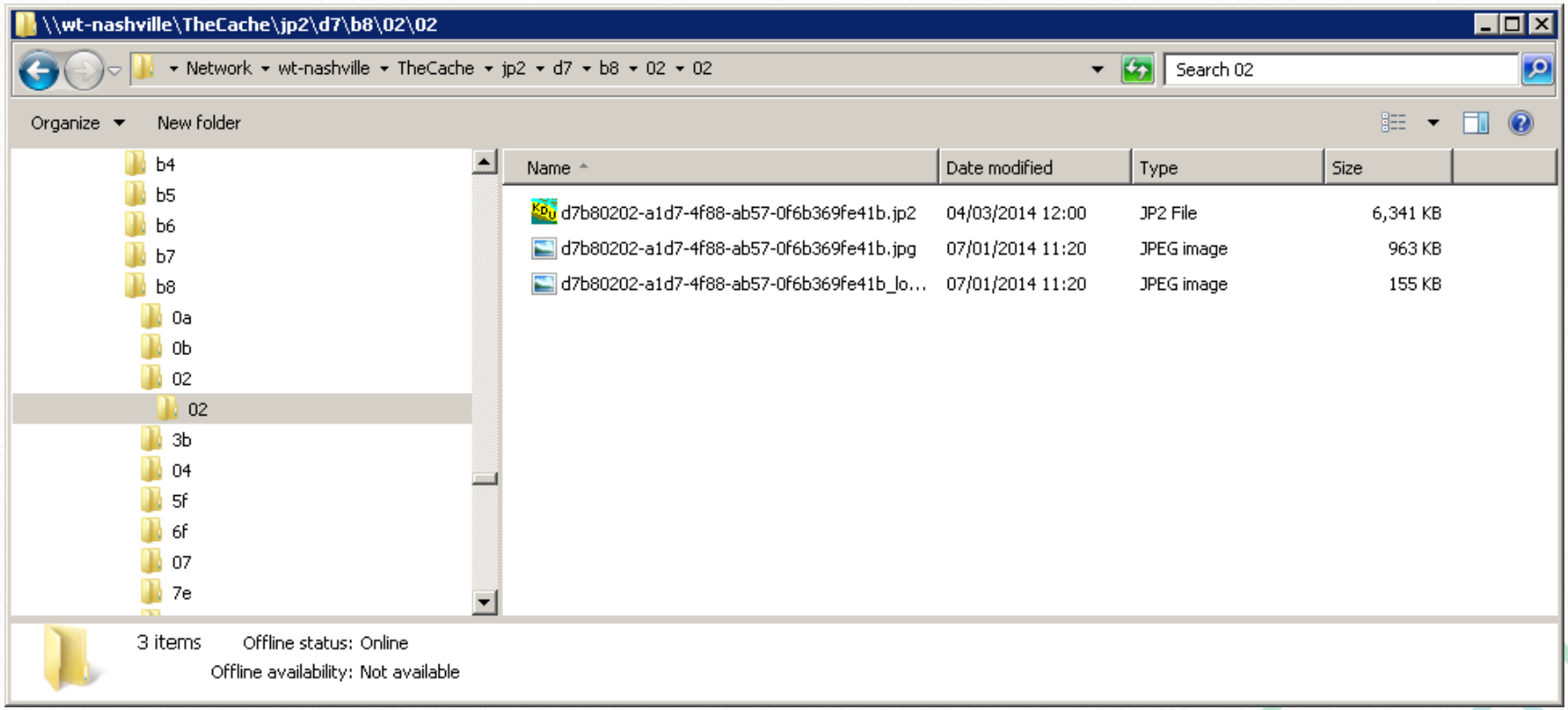
Tile Request



http://wellcomelibrary.org/dz/b20432033/0/d7b80202-a1d7-4f88-ab57-0f6b369fe41b.jp2_files/12/2_3.jpg

```
<rule name="Reverse Proxy to tile">
   <action type="Rewrite"
     url="http://WT-MANILA/cgi-bin/iipsrv.fcgi?DeepZoom=/ddsfilecache/jp2/{R:3}/{R:4}/{R:5}/{R:6}/{R:3}{R:4}{R:5}{R:6}{R:7}.jp2_files/{R:8}" />
</rule>
```

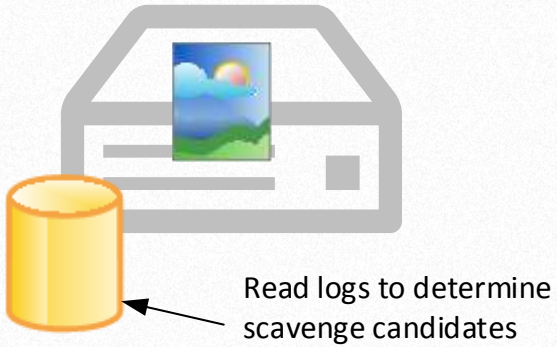
http://WT-MANILA/cgi-bin/iipsrv.fcgi?DeepZoom=/ddsfilecache/jp2/d7/b8/02/02/d7b80202-a1d7-4f88-ab57-0f6b369fe41b.jp2_files/12/2_3.jpg



5TB cache

- The 5TB of JP2s that IIPImage needs most are available on fast disk

- This cache is maintained by a “scavenger” process



DDS Dashboard

PDF Generation

b20649691 , Manifestation 0
Musaeum Tradescantianum,
Start: 2014-10-31 12:40
222/222
Ended: 2014-10-31 12:48
Warnings: no warnings PDF

b1863042x , Manifestation 0
[General account book: bills received and cash disbursements]
Start: 2014-10-30 13:20
293/293
Ended: 2014-10-30 13:38
Warnings: no warnings PDF

b21557354 , Manifestation 0
Photograph of Henry Wellcome's camp in Maine, 1886
Start: 2014-10-29 11:40
1/1
Ended: 2014-10-29 11:40
Warnings: no warnings PDF

b21417428 , Manifestation 0
[WMS Arabic 394]
Start: 2014-10-28 18:40
212/212
Ended: 2014-10-28 18:53
Warnings: no warnings PDF

b20295212 , Manifestation 0
[WMS Arabic 90]
Start: 2014-10-28 15:40
79/79
Ended: 2014-10-28 15:45
Warnings: no warnings PDF

b21557299 , Manifestation 0
Advertising panel portraying a pregnant

View Queue

Cache status

Number of JP2s registered in cache
523308
Storage root
\\wt-nashville\thecache
Image Server URL mapped to cache
http://WT-MINSK/cgi-bin/iipsrv.fcgi?FIF=/ddsfilecache
Database
Server=wt-havana;Database=DdsInstrumentation;user=ddsinst;password=**********
Required free space
279 GB *(300000000000)*
Current free space
277 GB *(297782067200)*
Number of files per deletion batch
100

Cache threshold exceeded. This app will delete files if run now with the scavenge parameter.

Please be careful with the following link... it needs to look at ALL the METS files
Launch METS report

Unfinished PDFs

PDFs with warnings

Other warnings

Recent cache activity

View top 100 scavenge candidates

View LAST 100 scavenge candidates

Date	Event	Description
2014-10-30 18:31	ScavengeEnd	Finished scavenging, 1300 files deleted. freeBytes=300010692608, freeSpaceRequiredBytes=300000000000
2014-10-30 18:31	ScavengeSetEnd	END Scavenge set 13. freeBytes=300010692608, freeSpaceRequiredBytes=300000000000
2014-10-30 18:31	ScavengeSetEnd	END Scavenge set 12. freeBytes=299273994240, freeSpaceRequiredBytes=300000000000
2014-10-30 18:31	ScavengeSetStart	Scavenge set 13. freeBytes=299273994240, freeSpaceRequiredBytes=300000000000
2014-10-30 18:31	ScavengeSetEnd	END Scavenge set 11. freeBytes=298618503168, freeSpaceRequiredBytes=300000000000
2014-10-30 18:31	ScavengeSetStart	Scavenge set 12. freeBytes=298618503168, freeSpaceRequiredBytes=300000000000
2014-10-30 18:31	ScavengeSetEnd	END Scavenge set 10. freeBytes=297818652672, freeSpaceRequiredBytes=300000000000
2014-10-30 18:31	ScavengeSetStart	Scavenge set 11. freeBytes=297818652672, freeSpaceRequiredBytes=300000000000
2014-10-30 18:31	ScavengeSetEnd	END Scavenge set 9. freeBytes=297022697472, freeSpaceRequiredBytes=300000000000
2014-10-30 18:31	ScavengeSetStart	Scavenge set 10. freeBytes=297022697472, freeSpaceRequiredBytes=300000000000
2014-10-30 18:31	ScavengeSetEnd	END Scavenge set 8. freeBytes=296294600704, freeSpaceRequiredBytes=300000000000
2014-10-30 18:31	ScavengeSetStart	Scavenge set 9. freeBytes=296294600704, freeSpaceRequiredBytes=300000000000
2014-10-30 18:30	ScavengeSetEnd	END Scavenge set 7. freeBytes=295499403264, freeSpaceRequiredBytes=300000000000

IIPImage

- Keep IIPImage boxes completely vanilla – just lighttpd and IIPImage
- Caching everywhere
- Check for file heavily optimised
- Fetch from SDB the weak point

- Only JP2s cached this way (volatile) – everything else is permanently available

Infrastructure Review

- Caching/scavenging strategy
- Is 5TB enough? “long-tail”
- Can this problem be solved a different way?