Presenting multilingual archival data online

Introduction

This poster details the technical underpinnings of the Qatar Digital Library website (www.qdl.qa). It also shows how the data is extracted from source formats including common archival standards such as METS, EAD and ALTO, stored in intermediate formats, and finally rendered online.

In October 2014 the Qatar National Library launched www.qdl.qa, a bilingual website that presents digitised copies of hundreds of thousands of historical archive documents, along with detailed metadata. The source content and structured metadata was produced by the British Library according to common archival standards including METS, EAD and ALTO. The final website, created to present this information in an engaging and highly usable way, was produced by Cogapp.

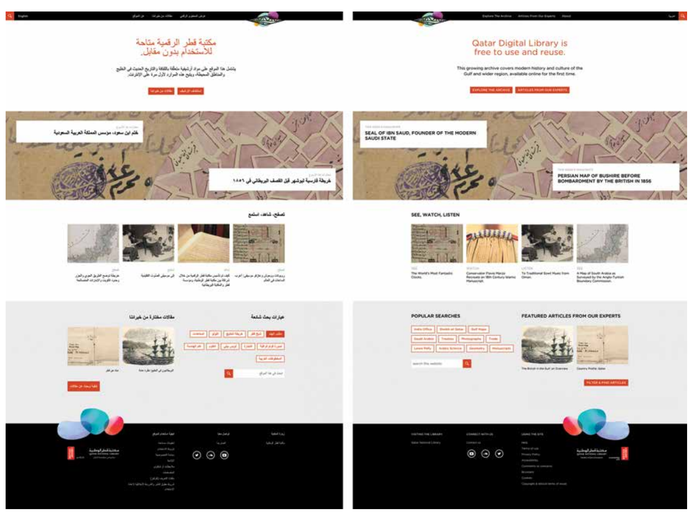

Figure 1: bilingual Arabic / English homepage

Data sources

The Qatar Digital Library is created from two content sources:

1. Archival data in the form of Submission Information Packages (SIPs)
2. Editorial content entered via a web-based CMS (Drupal)

The SIPs are multi-gigabyte zip files that include the following types of source content:

- SIP structure in Metadata Encoding and Transmission Standard (METS) format
- SIP classification and categorisation in Encoded Archival Description (EAD) format
- High-resolution scans of each physical page in JPEG 2000 format
- Transcription via Optical Character Recognition (OCR) in Analyzed Layout and Text Object (ALTO) format

Data delivery

We have developed bespoke software that takes these SIPs and transforms the data so that it is more readily usable online. To this end, all relevant parts of the textual content (METS, EAD, ALTO) are submitted to Apache Solr for storage and indexing. This data is further queried by code running on a Drupal server, and combined with templates and presentation information to deliver web pages in HTML format to the site visitor.

Meanwhile, all JPEG2000-format images are copied to a server running IIPImage, which is able to transform these master images into scaled derivatives and tiles in standard JPEG format for delivery to the visitor's browser.

This can be represented like so:

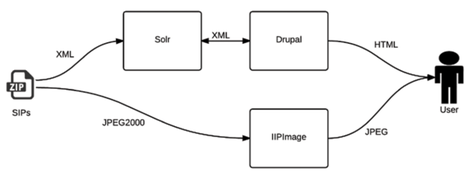

Figure 2: information flow, from source to delivery

Data transformation

The site is fully bilingual in both Arabic and English, and this presented multiple challenges.

Firstly, because the METS format does not take into account the possibility of several languages, several workarounds were used. These included embedding two different EAD records per METS Descriptor, and, for parts of the METS record that could not be duplicated, to add multiple text strings for a single attribute such as the ORDERLABEL in the physical structure map section, and the TYPE in the logical structure map section.

Secondly, because the Arabic-language version of the site uses Eastern Arabic numerals, numerical information such as the ORDER attribute in the physical structure map section had to be pre-processed to convert from Western Arabic numerals to Eastern Arabic prior to rendering.

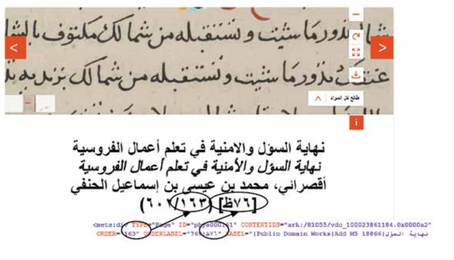

Figure 3: mapping of bilingual content from METS to online representation

Finally, we carefully tuned the site's Cascading Style Sheets (CSS) to deliver an identical visual experience for users in left-to-right / English as well as right-to-left / Arabic. At the same time we accounted for the challenges presented when displaying and laying out text and images in both these languages, as shown in Figure 1 above.

Data storage

As mentioned above, the archival information is stored exclusively in Apache Solr. In order to enable visitors to interact with the material in the fullest possible manner, we use three distinct types of field in the Solr schema:

- searchable fields
- categorisation fields
- structured data fields

The structured data fields contain verbatim copies of the EAD and ALTO records, which means that the presentation can be updated easily without having to reprocess all source SIPs. Some examples of how these different field types are used on the site are given below.

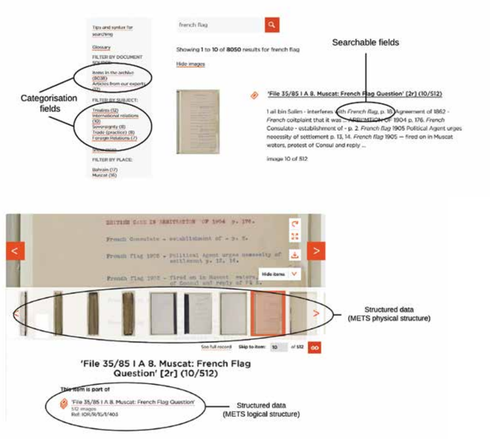

Figure 4: mapping of fields in Solr and METS/EAD to representation online

System scalability and resilience

To ensure that the system is able to handle spikes in traffic, and to make sure there is no single point of failure, we created a highly redundant architecture.

Servers were deployed on Amazon Web Services infrastructure in two different Availability Zones, and all front-end application servers (Drupal, IIPImage) use autoscaling rules to automatically add new server instances to the load-balanced clusters based on CPU load.

We use GlusterFS, a clustered file store, in "distributed and replicated" mode to firstly ensure that the risk of data loss or interruption to service is small, and secondly so that we can easily add more capacity to the multi-terabyte image store as more and more documents are digitised.

We also use Solr 4 in SolrCloud mode, to ensure that the indexes are distributed across several servers.

The backend data sources (Solr, GlusterFS, Drupal database stored in Relational Database Service (RDS) are also all stored on a private subnet to reduce the risk of unauthorised access, as shown below.

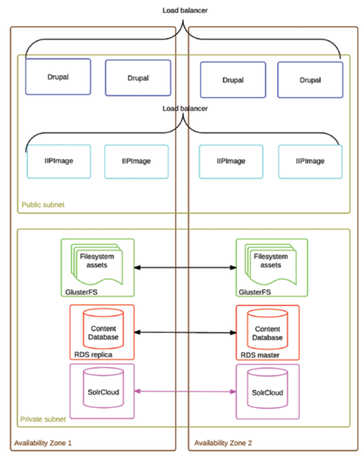

Figure 5: server configuration, employing clustered, redundant servers at all points

Data growth

The digitisation programme at the British Library is providing more and more items each day, and the aim is to have over one and a half million documents online by the end of 2018.

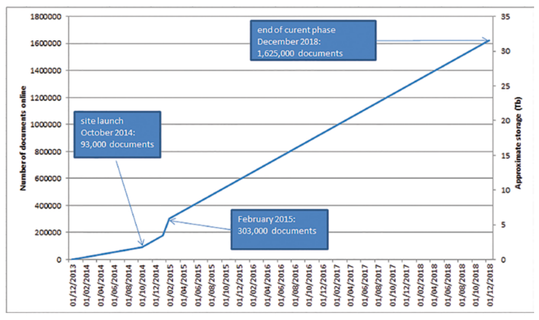

Figure 6: current and projected growth in number of physical items and associated data storage

Further information

Qatar Digital Library www.qdl.qa
Cogapp www.cogapp.com
BL/QNL Partnership Programme www.bl.uk/qatar
Apache Solr lucene.apache.org/solr
IIPImage iipimage.sourceforge.net
GlusterFS www.gluster.org

Dr Tristan Roddis
Head of Web Development
Cogapp
tristanr@cogapp.com

cogapp