

Characterising content in web archives with Nanite

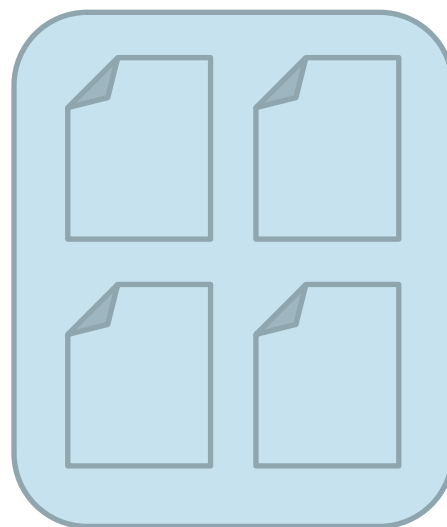
William Palmer

SCAPE Information Day

British Library, UK, 14th July 2014

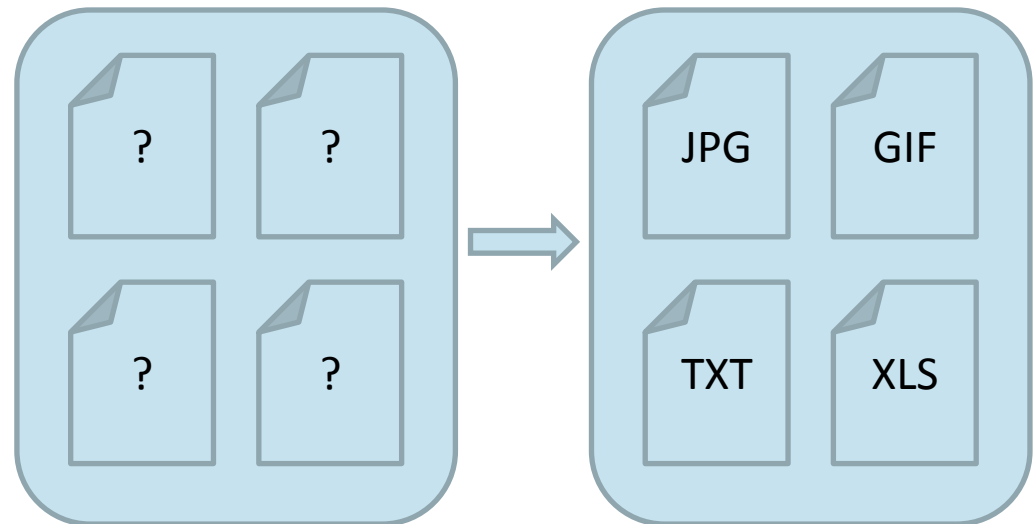
- When web sites are harvested they are stored in a container format
- The main web archive container formats are ARC and WARC (an ISO standard)
- They are effectively analogous to a zip file

WARC Container

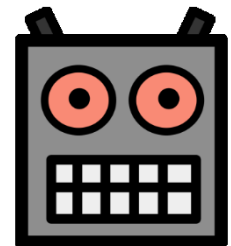


Characterisation

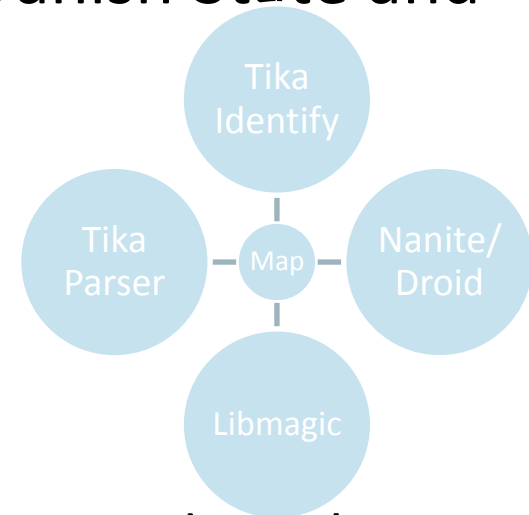
- Web archives can hold billions of individual records
- To answer deeper questions you have to determine what data is held
- Not the same as a homogenous collection of images
- Can contain everything and anything
 - Correctly formed files
 - Malformed files
 - Viruses
 - Unknown files?
 - You name it



- Nanite is formed of two main modules
- **nanite-core**: a Java API for the UK National Archives' Droid
- **nanite-hadoop**: WARC content characterisation using Hadoop
 - Apache Tika (Detector), Nanite-core & libmagic-jni ('file')
 - Optionally use Tika (Parsers); data output to sequence files
 - Also list server content type & file extension
 - Reuses: **warc-hadoop-recordreaders** (partially SCAPE)



- **Fast:** for 1TB, 14k warcs, 93m files; mimetypes detected in 17 hours
- Nanite has also been used at the Danish State and University Library
 - 7.3TB data, 80k ARC files, 261m files
 - Identification using Droid and Tika
 - Characterisation using Tika
 - ...in 32 hours
 - Same platform but using FITS (not using Hadoop, but parallelised):
 - 12TB data, 100k ARC files, 400m files
 - An entire year of processing (8760 hours)



- 1370 different MIME types reported by the original servers
- Tika detected 342
- DROID detected 319
- Additional information in this blog post:
<http://www.openplanetsfoundation.org/blogs/2014-05-28-weekend-nanite>

Visualising Characterisation Information

- Nanite has an option to output C3PO compatible outputs
- They can be directly loaded into C3PO for visualisation

