The Infrastructure behind Web Archiving at Scale

Gil Hoggarth
The British Library Web Archiving team

IIPC Web Archiving Conference, Zagreb
June 7, 2019

Introduction

The UK Web Archive has collected UK web content since 2004 based on owner permissions.

After the Legal Deposit Libraries Act (2003) was extended in 2013 to non-print publications, UKWA also started to collect all UK web content* via annual domain crawls.

Total amount of crawled data in 2004: 2.8TB

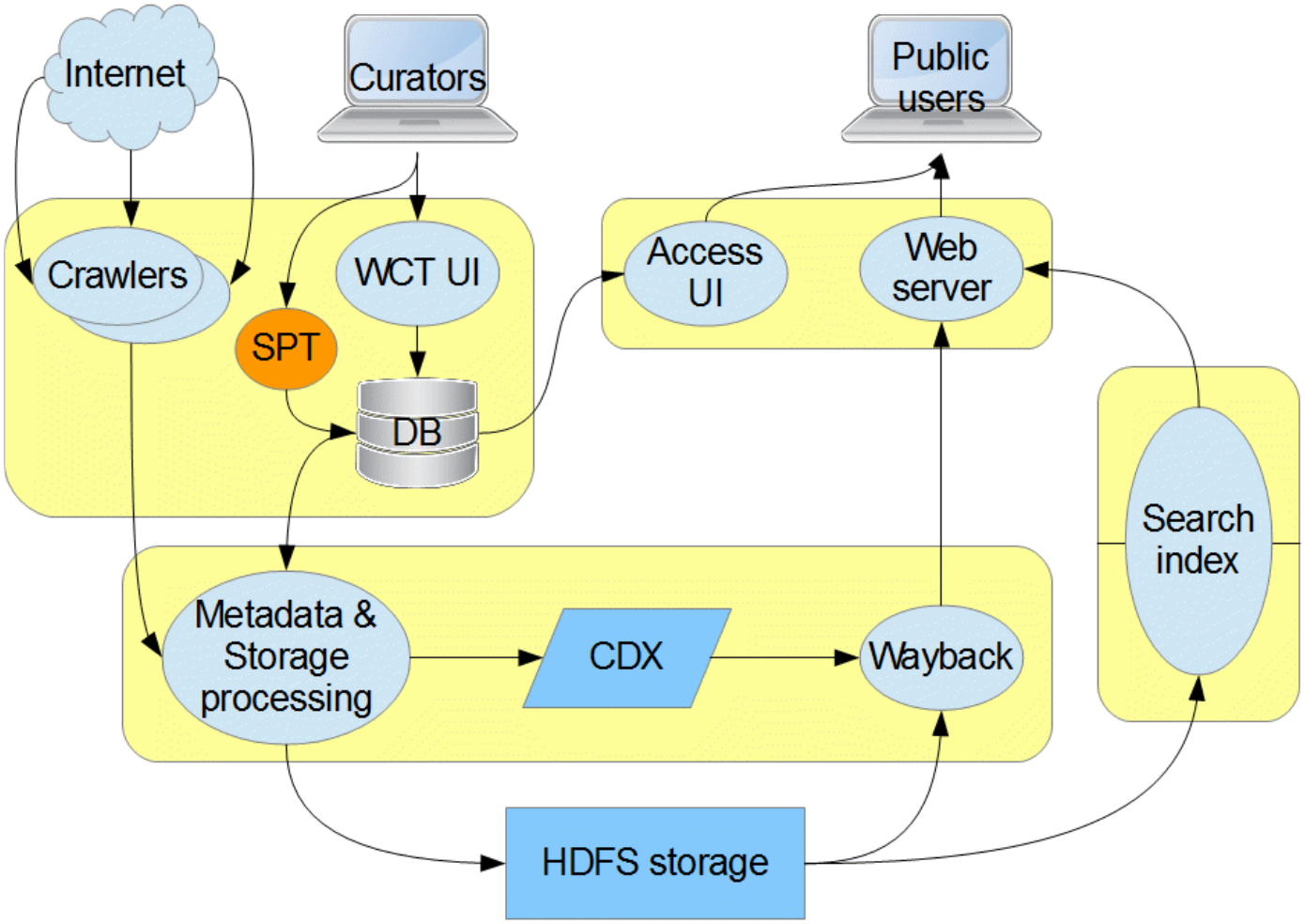
Total amount of crawled data in 2019: ~600TB

2004 – Beginning of UKWA web archiving

Selective crawling performed on hosted equipment running PANDORA Digital Archiving System (PANDAS), developed by the National Library of Australia:

- A RHEL4 archiving server, using HTTrack
  - 2 external disks
- An Oracle database
- A web server
- Tape backup facility

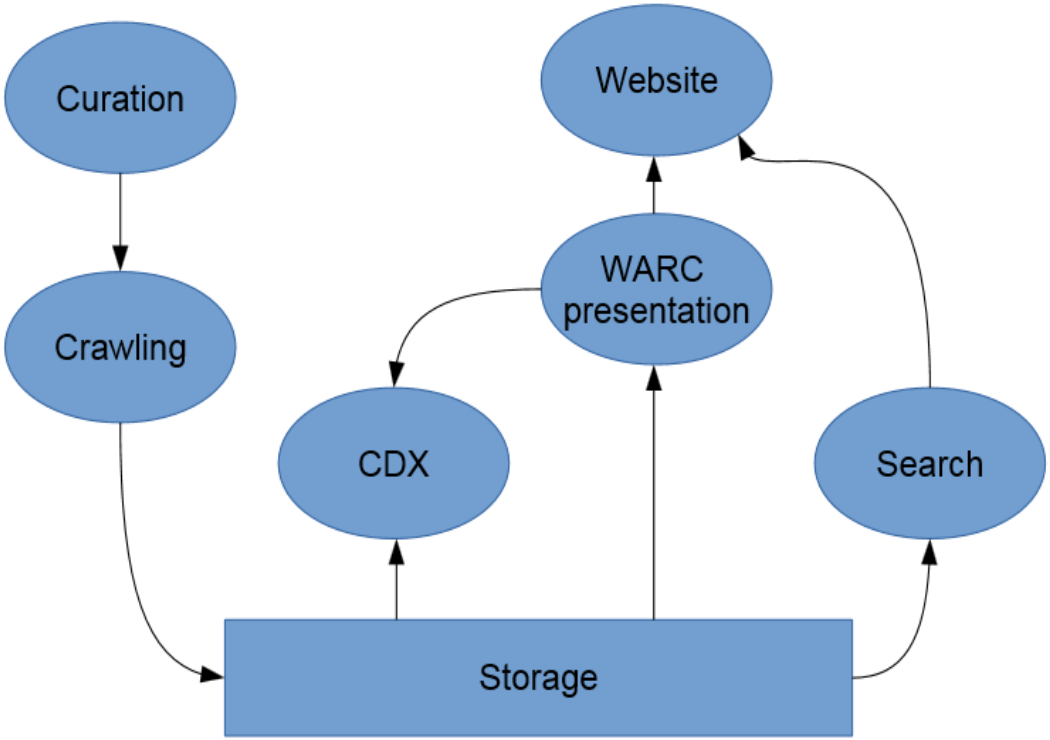
2008-2016 – Web Archiving Services



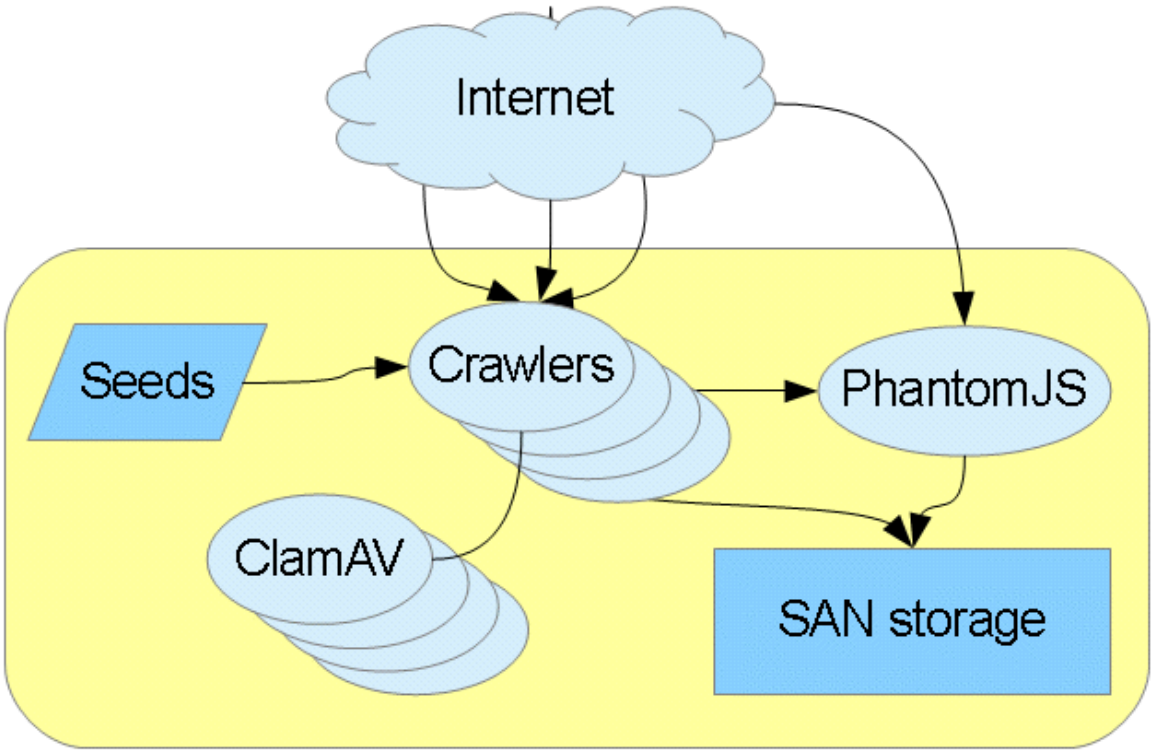
Generalised Services

- Curation service
- Crawling service
- CDX lookup service
- WARC presentation service
- Search service
- Public website
- Storage service
- Glue

Generalised Services Diagram



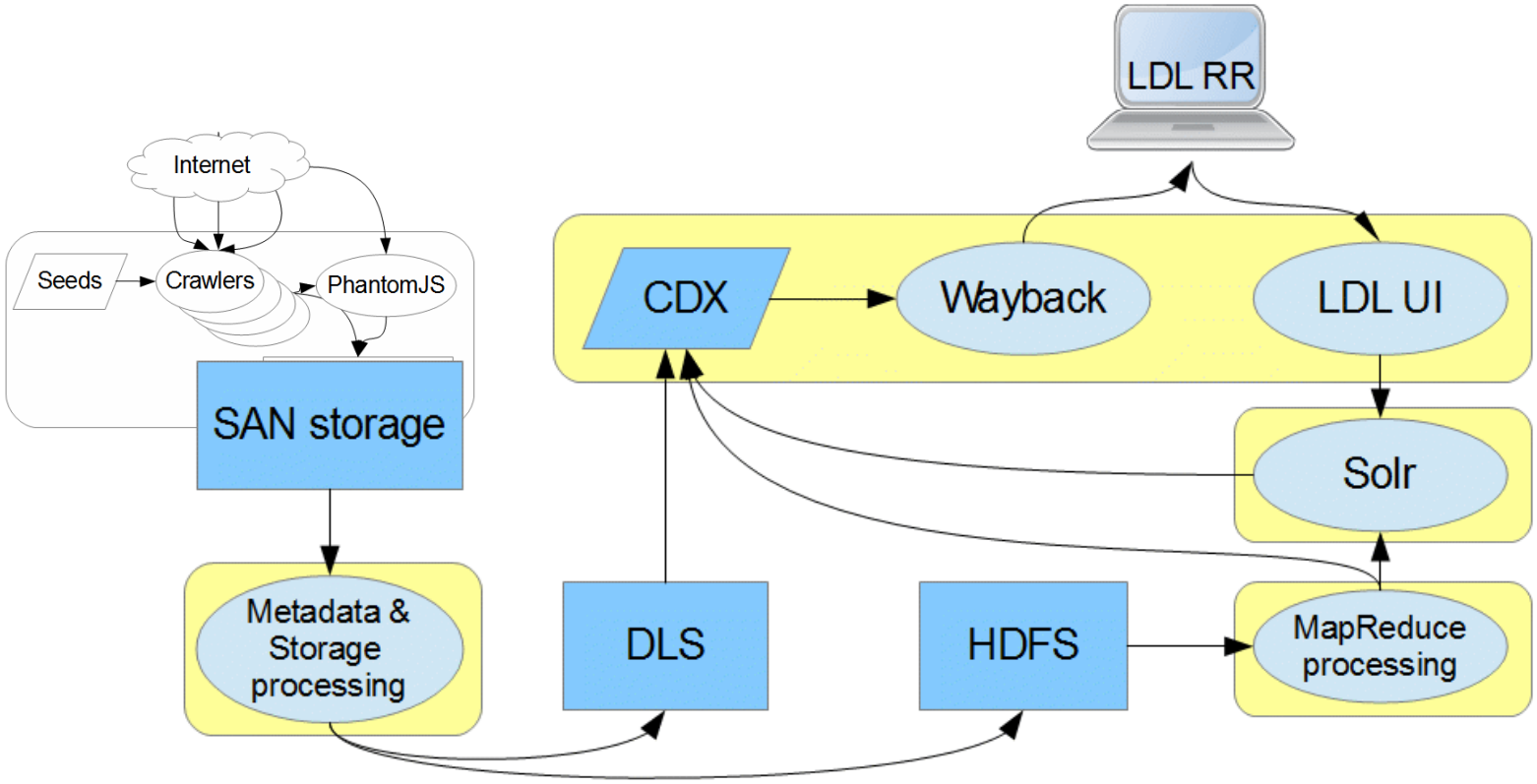
2013-2015
– Legal Deposit UK Domain Crawling



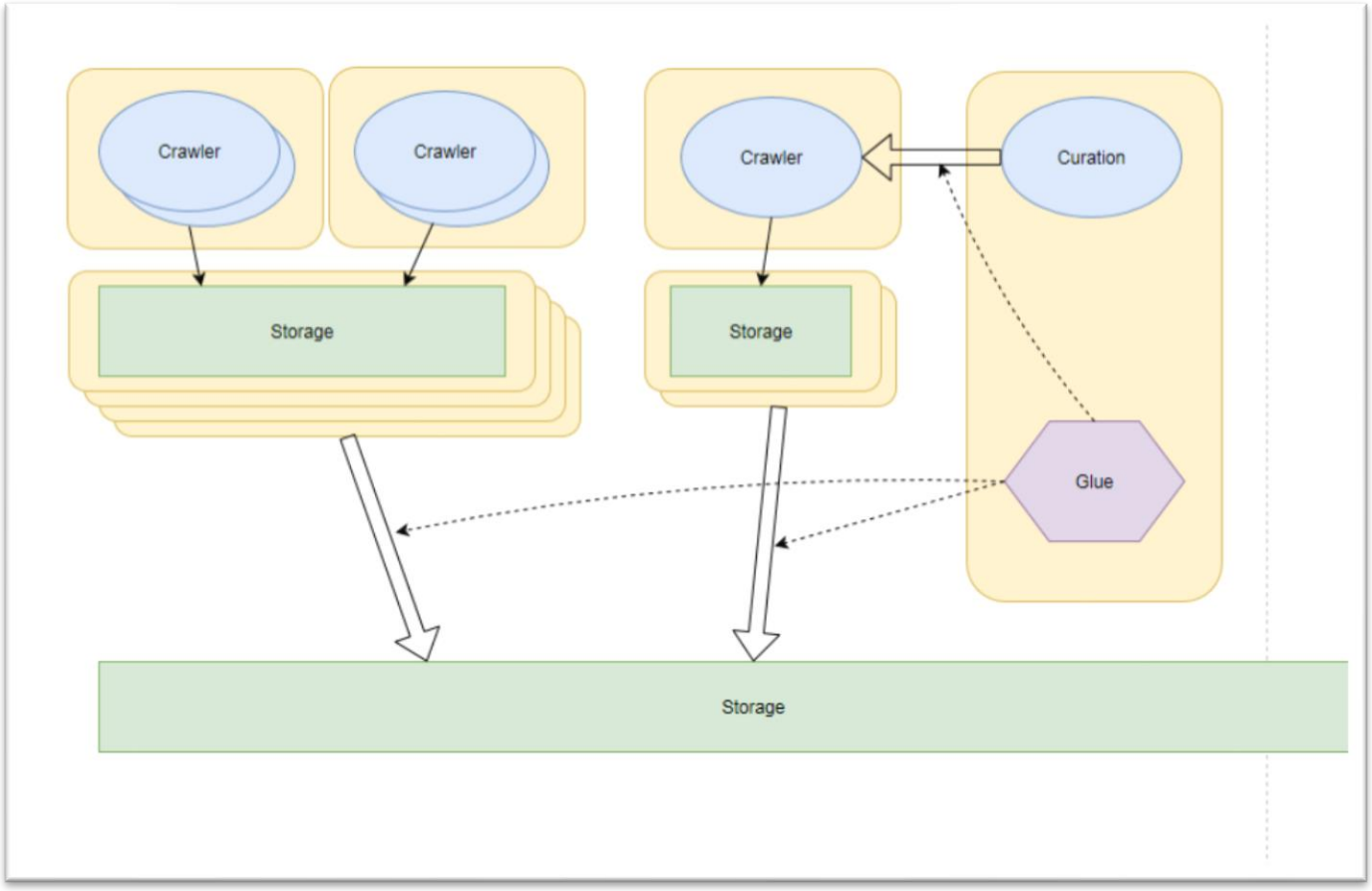
2013 – 2015 Domain Crawl Services

- List of seeds
- Multiple crawlers
- Virus scanning
- Homepage image rendering
- Storage
- Glue

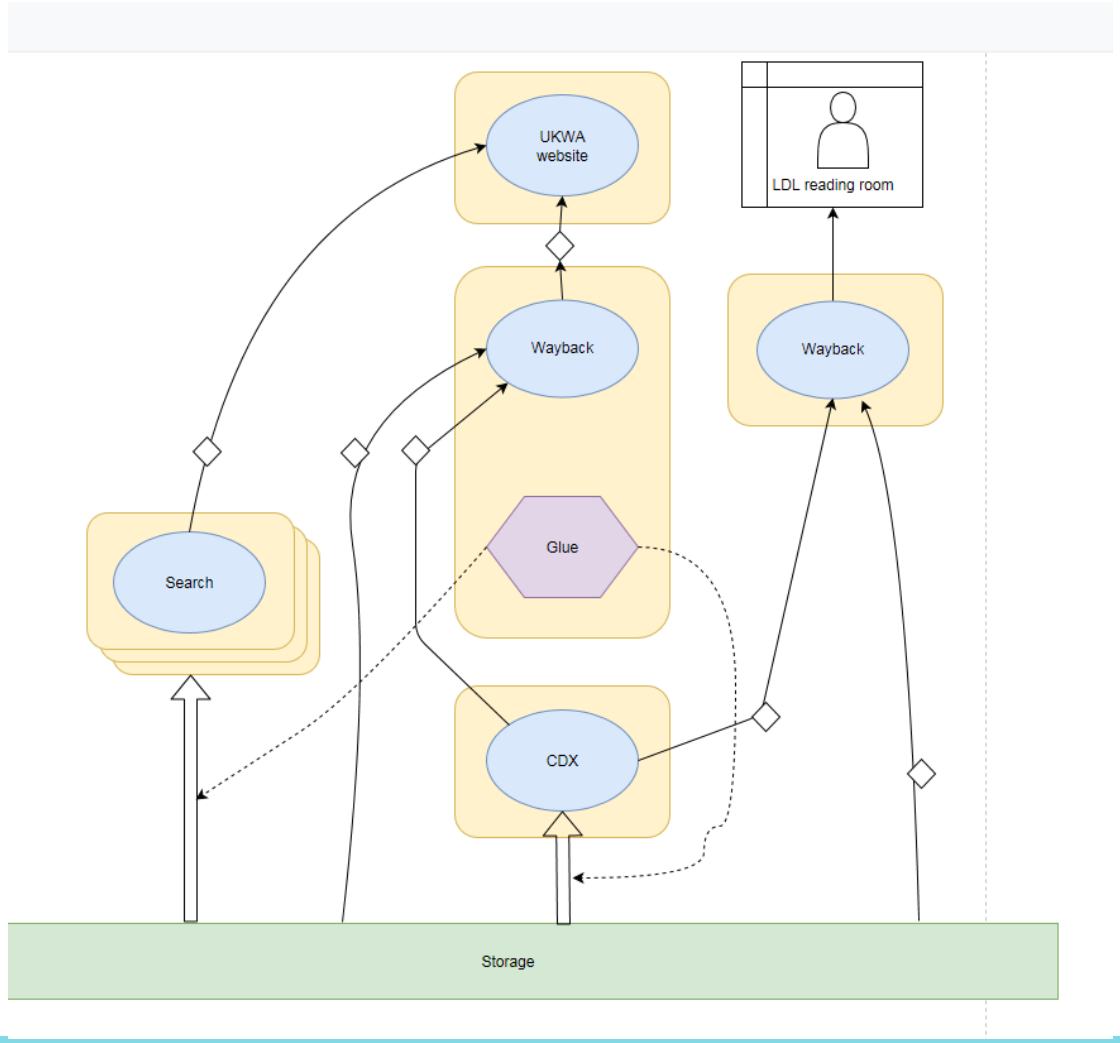
2013-2016
– Access to Legal Deposit crawl data



2017 - today
– Ingest Service Consolidation



2017 – today
- Access Service Consolidation



Suggestions

- Minimise large data movement
- Isolate services
- Use code libraries, even if they are your own
- Include service proxies
- Decide a Virus Policy
- Security

Real Components

- Curation service – W3ACT
- Crawling – heritrix 3
- CDX service – Outback CDX server
- WARC presentation – OpenWayback & pyWb
- Search – Solr
- Storage – Gluster and HDFS
- Glue – github.com/ukwa tools

Thank you

http://www.webarchive.org.uk/
http://www.webarchive.org.uk/blog/
@UKWebArchive