Deep Search or Harvest: how we decided

Simon Moffatt
Ilaria Corda

*Discovery & Access – The British Library*
Overview

• Focussing on operational and functional considerations
  – Harvesting data and using Deep Search.

• Along the way…
  – Challenges of operating with a large data set
  – Demonstrate our new Web Archive search of 1.3bn web pages
Introductions

• We work in a team of six supporting and developing *Discovery and Access* at the British Library

• The British Library has reading rooms and storage in London and in Yorkshire. (We are based in Yorkshire)

• We are a UK Legal Deposit library
  – We collect *everything*
  – It can only be accessed in our reading rooms

• Our Ex Libris products
  – Aleph v20  Primo v4.x  SFX v4.x
  – All locally hosted
Deep Search or Harvest: how we decided

• Like most institutions, the remit of our search is expanding
  – New digital content
  – Migration from older technologies

• If you have a large number of new records to add, often you have two options:
  – Harvest the records into your main index
  – Deep Search to an external index
Harvest
Deep Search

![Diagram showing the process of Deep Search or Harvest, how we decided]

- User
- Primo Front End Server
- Primo Index
- Primo Back Office
- Aleph
- Sound Archive
- Articles
- NEW Index
- NEW
Two case studies

Recently we faced this Harvest vs Deep Search dilemma for two new sources of data

• A replacement to our Content Management System
  – we decided to harvest

• A search of our new Web Archive
  – we implemented a deep search
Content Management System - Harvested

The technology behind the BL website is being updated.

- There are around 50,000 pages
- It has its own index, updated daily

Work involved

- Creating Normalisation rules and Pipe
- Setting up daily processes

Not appropriate for deep search – more later….
Large datasets in Primo: Our experience

Background: Our data

- 13m Aleph records (Books, Journals, Maps etc)
- 5m SirsiDynix Symphony records (Sound Archive)
- 48m Journal articles (growing by 2m per year)
- 1m other records from five other pipes

Total: 67 Million records
Our topology

Load-balanced requests

FE1
Front End Server

FE2
Front End Server

FE3
Front End Server

SE1
Search Slice

SE2
Search Slice

SE3
Search Slice

SE4
Search Slice

SE5
Search Slice

SE6
Search Slice

SE7
Search Slice

SE8
Search Slice

DB
Database Server

BO
Back Office Server

Public

Private

Config, Web 2.0 etc

Config, PNXs etc

Search Results

Indexes

Harvests

IGeLU 2014 – Deep Search or Harvest: How we decided [14.5]
Service challenges with 67m records

Our index is ~100GB

- Indexing takes at least 3 hours; Hotswap takes 4 hours
  - Even if there is only one new record
  - Overnight schedules are tight
  - Fast data updates are impossible

- System restart takes 5 hours
  - Re-sync Search Schema is a whole day
  - Failover system must be available at all times

- Primo Service Packs and Hotfixes need caution

- Standard documented advice must be read carefully
Development challenges with 67m records

- Full re-harvest takes 7 weeks
  - Major normalisation changes only 3 times per year
  - Smaller UI changes can be made more often

- Primo Version upgrades affect 13 servers (or 56 in total)

- Implementing Primo enhancements
  - We must consider index size
  - Sort, browse etc all have an impact
Our development cycle

(not to scale)
But there are compensations…

- Speed of the index
- Control over the data
- Consistency of rules
- Common development skills
- A single point of failure

These are all important
Web Archive - Deep Search

**Web Archiving** collects, makes accessible and preserves web resources of scholarly and cultural importance from the UK domain

**Some Figures:**

- ~1.3 billion documents
- Regular crawling processes - E.g. BBC News

**And the infrastructure?**

- Index size is ~3.5TB spread across 24 Solr shards
- 80-node Hadoop cluster (where crawls are processed for submission)
What is Explore the British Library?

With Explore the British Library you can search, view and order items from our main catalogue of nearly 57 million records, or search the contents of the Library’s website.

Whilst Explore the British Library gives access to the majority of the Library’s collection, it does not yet include records from all the Library’s catalogues. Notable exceptions are:

- Archives and manuscript collections
- Register of Presentation Surrogates (RPS)
- British National Bibliography (BNB) (Explore the British Library continues to include BNB records for items held in our collections)
- Specialist catalogues

Quick Links

- Ask the Reference Team - for help with your research
- Locations and opening times - London and Boston Spa
- Reading Rooms - we have Reading Rooms on both sites
- Register for a Reader Pass - you need a Pass to use our Rooms
- Document supply - we have a number of document supply services
- Exhibitions and events - what’s on at the Library
- Online shop - books, audio and gifts
- All catalogues - we have over 20 online catalogues
- MARC records via Z39.50 - access our MARC 21 records

Notices

Collection items temporarily unavailable to readers
Some items are temporarily unavailable whilst being moved or digitised.

Reference Services Quick Chat

Use our Quick Chat service for your short enquiries and obtain an instant reply during our opening hours: Mon to Fri 10.00 to 16.00.

When the chat service is closed, you can contact us via Ask the Reference Team.

Booksellers Association

Books can be purchased or ordered from members of The Booksellers Association who can be found on the BA website. The database includes all the major chains and local independents and can be searched by bookshop name, town, county and postcode. The display will show full bookshop details including Google map locations and Street View and where the bookshop has a website, you can click straight through.
**Service challenges with Deep Search**

- Additional point of failure within Primo
  - Newly introduced Troubleshooting procedure
  - Service Notices for planned Downtime/Maintenance
  - Changes to the Solr Schema that could break our search

- Primo Upgrades & Hotfixes can affect the functioning of the Deep Search
  - Re-builds of the Client in case the Deep Search libraries (jars) change
Development challenges with Deep Search

- Significant **development work** to implement Primo/Solr integration
- Accommodate new Primo features (versioning control)
  - E.g. Multi-select faceting (Exclusion & Inclusion)
- Needs to **ensure consistency** across local and non-local collections
  - Seemingly NO difference from a UI/UX point of view
But there are compensations…

- Ideal for **large indexes** and **frequent updates**

- **Independent** indexing processes

- **Maintenance** of **existing** scheduled **processes**

- Leverage **existing Primo-built** in features

- **Existing** and **Extendable Solr APIs** / active community
Harvest vs Deep Search: how we decided

- ADDITIONAL INFRASTRUCTURE COSTS
- IMPACT ON CURRENT PROCESSES
- FREQUENCY & VOLUME OF UPDATES
- INDEX SIZE/NO. RECORDS
- DEV TIME / SPECIALIST SKILLS

Deep Search vs Harvest

www.bl.uk

IGeLU 2014 – Deep Search or Harvest: How we decided [14.5]
Thank you