COMET Project Overview

Ed Chamberlain
Cambridge University Library
Cambridge Open Metadata

Funded by the JISC Discovery Programme

What?
- Cambridge, back in 2010 ...

- OKFN - Open Bibliography project (2010-2011)

- Debate around re-use of catalogue records from vendors (not just OCLC)

- CUL already provides public APIs

- Increasing interest in linked data

- FAST / VIAF

- Lorcan
“The initial aim of this project will be to identify and release a substantial record set to an external platform under an open license” ...

“For OCLC-derived bibliographic records data will be released in a fashion compliant with their WorldCat Rights and Responsibilities for the OCLC Cooperative” ...

“The project aims to then deploy and test and number of technologies and methodologies for releasing open bibliographic data including XML, RDF, SPARQL, and JSON” ...
Cambridge University Library
- Metadata conversion
- Development
- Project management

CARET
- Infrastructure support

OCLC
- Licensing consultancy
- FAST / VIAFF enrichment

Who?
• Value for money – Taxpayers

• Open data = affiliate marketing for our collections

• Drive innovation - vital buy-in from non library developer communities

• One of many open data projects at the time
“Library catalogues have imposed on them librarian or supplier-made decisions about what can/can’t be searched and in what way. Some of these decisions are limited by current cataloguing rules, but not all; often the data is recorded, but not in a usable way, or is there but isn’t tapped by the interface. For example, in most catalogues you can limit by publication type to newspapers, but you can’t limit by frequency of the issues.”

“Releasing data means that people can start to use it in the way they want to.”

Katie Birkwood – Rare books specialist

Most of the catalogue (3 million +)
- Bulk downloads of RDF triples
- Query-able ‘endpoints’
- Fast / VIAF enriched
- Snapshot – not enlivened

RDF conversion tools

Working model and code to decide on MArch21 record origin

Codebase for ‘library centric’ RDF publishing website

SPARQL tutorial

Verbose blog

The results ...

Data and code at http://data.lib.cam.ac.uk
Welcome to the Cambridge University Library Open Data service. This is a work in progress.

This service is being initially developed through the COMET project. COMET is a JISC funded collaboration between Cambridge University Library and CARET, University of Cambridge. It is funded under the JISC Infrastructure for Resource Discovery programme.

COMET will release a large sub-set of bibliographic data from Cambridge University Library catalogues under Open Data Commons licensing. It has also explored a number of technologies and methodologies for publishing RDF.

Current areas of development:
- Data quality and enhancement - correct datatypes for some literals not yet in place
- Encoding issues -
- Linking and enrichment - Fast/ VIAF /LOC subject URI's not yet in place
- Backend scaling and performance review

Send any more to emc59@cam.ac.uk

Blog entries:
- Back with more data!
- Final post
- Where exactly DOES a record come from?
- More data and status of future updates
- Friday update ...
- Cost benefits
- Project update and following in our footsteps
- And now for something completely different ...
- Two more updates ...
- data.lib.cam.ac.uk now live
- On licensing ...

Useful links:
- Discovery programme
- Cambridge University Library
- CARET
- JISC Open Bibliographic Data Guide
- OKFN Open Bibliography
- UK RDF metadata guidelines
- JISC
- RLUK
- OCLC

Built with the ARC2 library and ARC2 starter pack
Record licensing
- Examine contracts with major vendors
- Contact them and decide on re-use conditions
- Deduce record origin from Marc21 fields

What could we publish?
Several places in Marc21 where this data could be held (015,035,038,994 ...)

Logic and hierarchy for examination

Attempt at scripted analysis

Marc21 fails at ‘IPR’

Potential down the line for problem to persist if attribution is not handled correctly in future formats

Where does a record come from?
Several places in Marc21 where this data could be held (015, 035, 038, 994);

Logic and hierarchy for examination;

Attempt at scripted analysis;

Marc21 fails at ‘IPR’;

Potential down the line for problem to persist if airbnb is not handled correctly in future formats.

What does a come from?
Need the right license!

- Most vendors happy with permissive license for ‘non-marc21’ formats
- RLUK / BL B.N.B. – Public Domain Data License
- OCLC – ODC-By Attribution license with community norms
  - Recently voted upon

No good reason not to re-publish
RDF data
Try to track the emerging consensus on bibliographic description

Followed work on Open Bibliography

Eventually tracked BL

Victory for pragmatism?
Marc21 to RDF problems

- Binary encoding
- Numbers for field names
- Punctuation in field data
- Mixed content fields (020)
- Bad characters
- Replication of data in fields
- PHP script to match text against LOC subject headings
  - Using label lookup over httpd
  - Enrich with LOC GUID

- FAST / VIAF enrichment courtesy of OCLC
  - Sent Marc21 files
  - Received augmented Marc for conversion
<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of records:</td>
<td>3,658,384</td>
</tr>
<tr>
<td>No. of records with LCSH headings:</td>
<td>2,709,878</td>
</tr>
<tr>
<td>Percentage with LCSH headings:</td>
<td>74%</td>
</tr>
<tr>
<td>No. of subject headings found:</td>
<td>5,889,048</td>
</tr>
<tr>
<td>No. of subject headings skipped:</td>
<td>45</td>
</tr>
<tr>
<td>Valid FAST subjects:</td>
<td>8,134,230</td>
</tr>
</tbody>
</table>

Fast Linking

(OCLC enrichment stats)
ARC2 - Lightweight MYSQL / PHP solution

- Good fit for a six month project
- Great for around 3-500 k records
- Not so good for 1 million plus
- 20 million + ?

How –storage and access
Triplestores are cumbersome without the right resources / skills

SPARQL alone does not do the trick – need a better index

High entry barrier to RDF is partly a result of these accompanying technologies

Supporting tech - What did we learn?
Advantages:
- Skills retained in house - no consultancy ‘black-box’
- Understand local data issues – quicker development
- Faster project runaround - six month experimental project

Dis-Advantages:
- Pain
- Learning from scratch
- Mistakes made in final output

Going it alone ...
Marc / AACR2 cannot translate easily to semantically rich formats

Libraries need to better utilise modern container / transfer standards (not necessarily RDF)

No ‘one size fits all’ approach for future

RDF - What did we learn?
Karen Coyle criticises the Marc21 Bibliographic Framework Transition Initiative for not including museums, publishing, and IT professionals ...

She argues that our data is not just for us to consume alone ...

“The next data carrier for libraries needs to be developed as a truly open effort. It should be led by a neutral organization (possibly ad hoc) that can bring together the wide range of interested parties and make sure that all voices are heard. Technical development should be done by computer professionals with expertise in metadata design. The resulting system should be rigorous yet flexible enough to allow growth and specialization.”

The future?
Lightweight approach to sharing bibliography now its open ...

- Bottom up, community led software called Bibserver
- Wikimedia for bib data
- JSON as a container format – flexible, able to cope with different structures, vocabularies etc.
  - BibJSON
  - JSON-LD
- Engagement with UK PubMed Central
Charting new approaches to traditional library workflows (copy cataloguing) using open data

- Using rich open data from a variety of sources to enrich bare bones data
- NOSQL database technology
- APIs as key deliverables

C.L.O.C.K.
(Cambridge/ Lincoln open cataloguing knowledgebase)
Beyond bibliographic

Bibliographic

Language

Place of publication

FAST subject headings

LCSH subject headings

Creator / entity

Special collections

Archives

Libraries

Holdings

Course lists

Transactions

Librarians
- Anonymous usage data from circulation systems
- Aggregated from several University Libraries
- API feed
- Available openly (CC-BY )

Copac Activity Data Project (SALT2)
- It becomes (even) easier to go to Amazon
- Our status as authoritative data providers will be (further) eroded
- Assume we can
- Assume we should (where we can)

“No-one will want to play with us if we cannot learn to share” ...
- http://discovery.ac.uk

- http://okfn.org - Open Knowledge Foundation

- http://data.lib.cam.ac.uk
Questions?

- Ed Chamberlain
  - @edchamberlain
  - emc59@cam.ac.uk
  - http://www.slideshare.net/EdmundChamberlain/