Increasing Access to Chinese & Indonesian Collections at British Library through Crowdsourcing

Sara Chiesura - DADH conference 2015, Taipei
Paper and project contributors

- **Sara Chiesura**, Lead Curator, East Asian Collections (Chinese), Asian & African Collections, British Library

- **Nora McGregor**, Digital Curator, Digital Scholarship, British Library

- **Alex Mendes**, Project Support Officer, Asian & African Collections, British Library

Contributions from across the Library: collections management, imaging studio, IT, operations and so on
Chinese and Indonesian collections at the British Library: overview

Chinese language collections:
About 100,000 printed items
About 2,500 periodical titles
About 20,000 manuscripts

Highlights:

2015 DADH
Chinese and Indonesian collections at the British Library: overview

Indonesian language collections:
- About **20,000** printed items
- About **50** periodical titles
- About **600** manuscripts

Highlights:
Background: collection catalogues
Background: collection catalogues

Need for readers to physically be in London and search through thousand of card catalogues

Consequences: limited access, more work for reading room team, more external enquiries for curators, more time needed to find items, limited research possibilities and so on…
<table>
<thead>
<tr>
<th>Acquisition period</th>
<th>Method of cataloguing</th>
<th>Number of items</th>
<th>Accessibility</th>
<th>Action</th>
<th>Items priority</th>
</tr>
</thead>
<tbody>
<tr>
<td>Before 1966</td>
<td>Microfiches (Wade Giles transliteration)</td>
<td>20,000</td>
<td>Reading Room</td>
<td>Retro-conversion</td>
<td>high</td>
</tr>
<tr>
<td>1966 - 1993</td>
<td>Cards catalogues (Pinyin)</td>
<td>40,000</td>
<td>Reading Room</td>
<td>Retro-conversion</td>
<td>high</td>
</tr>
<tr>
<td>1993 - 2013</td>
<td>Allegro system</td>
<td>20,000</td>
<td>UK Union Catalogue</td>
<td>Migration</td>
<td>medium</td>
</tr>
<tr>
<td>Current acquisitions</td>
<td>backlog</td>
<td>20,000</td>
<td>Non accessible to readers (suppressed minimal record)</td>
<td>Deriving</td>
<td>medium</td>
</tr>
</tbody>
</table>
Idea!

For material published after relatively recently and likely to be already in the OCLC database (Worldcat)
Crowdsourcing

The act of a company or institution taking a function once performed by employees and outsourcing it to an undefined (and generally large) network of people in the form of an open call (Howe, 2006)

... becoming increasingly common in museums, libraries, archives and the humanities as a tool for digitising or computing vast amounts of data (Ridge, 2014)

Usually applied to tasks like: analysing images, creating tags and metadata, subtitling, transcribing, correcting OCR text, identifying geographic locations, classifying data, matching data.
With **50,000** records in the Chinese pinyin card catalogue alone the cost is simply prohibitive for internal budgets. To hire a fulltime cataloguer with the required language skills, assuming they could create an average of 2500/3000 records per year based on previous internal figures, it would take **nearly 20 years to finish the task**.

Crowdsourcing

- More records in less time
- Direct engagement with audiences
- Better use of technologies
- Data available for further use, public domain information
<table>
<thead>
<tr>
<th>Acquisition period</th>
<th>Method of cataloguing</th>
<th>Number of items</th>
<th>Accessibility</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>1966 - 1993</td>
<td>Cards catalogues (Pinyin)</td>
<td>40,000</td>
<td>Reading Room</td>
<td>Retro-conversion</td>
</tr>
</tbody>
</table>

First step: **scanning** of the cards
<table>
<thead>
<tr>
<th>Acquisition period</th>
<th>Method of cataloguing</th>
<th>Number of items</th>
<th>Accessibility</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>1966 - 1993</td>
<td>Cards catalogues (Pinyin)</td>
<td>40,000</td>
<td>Reading Room</td>
<td>Retro-conversion</td>
</tr>
</tbody>
</table>

Second step: **dividing in batches (by date and MS vs PRINTED)**

<table>
<thead>
<tr>
<th>PINYIN CARDS: Chinese catalogue of acquisitions from 1006 to 1993</th>
<th>Suitable for crowdsourcing platform</th>
<th>Suitable for sending to Taiwan National Central Library (pre-1011 printed)</th>
<th>Leftovers - difficult cards</th>
<th>Total pinyin catalogue</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3 PY_1</td>
<td></td>
<td></td>
<td></td>
<td>1312</td>
</tr>
<tr>
<td>4 PY_2</td>
<td></td>
<td></td>
<td></td>
<td>1488</td>
</tr>
<tr>
<td>5 PY_3</td>
<td></td>
<td></td>
<td></td>
<td>1475</td>
</tr>
<tr>
<td>6 PY_4</td>
<td></td>
<td></td>
<td></td>
<td>1394</td>
</tr>
<tr>
<td>7 PY_5</td>
<td></td>
<td></td>
<td></td>
<td>1479</td>
</tr>
<tr>
<td>8 PY_6</td>
<td></td>
<td></td>
<td></td>
<td>1431</td>
</tr>
<tr>
<td>9 PY_7</td>
<td></td>
<td></td>
<td></td>
<td>1788</td>
</tr>
<tr>
<td>10 PY_8</td>
<td></td>
<td></td>
<td></td>
<td>1779</td>
</tr>
<tr>
<td>11 PY_9</td>
<td></td>
<td></td>
<td></td>
<td>1822</td>
</tr>
<tr>
<td>12 PY_10</td>
<td></td>
<td></td>
<td></td>
<td>1727</td>
</tr>
<tr>
<td>13 PY_11</td>
<td></td>
<td></td>
<td></td>
<td>1602</td>
</tr>
<tr>
<td>14 PY_12</td>
<td></td>
<td></td>
<td></td>
<td>1774</td>
</tr>
<tr>
<td>15 PY_13</td>
<td></td>
<td></td>
<td></td>
<td>1815</td>
</tr>
<tr>
<td>16 PY_14</td>
<td></td>
<td></td>
<td></td>
<td>872</td>
</tr>
<tr>
<td>17 PY_15</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18 PY_16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19 PY_17</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20 PY_18</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21 PY_19</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22 PY_20</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23 PY_21</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24 PY_22</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25 PY_23</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26 PY_24</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27 PY_25</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28 PY_26</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30 TOTAL</td>
<td></td>
<td></td>
<td></td>
<td>21250</td>
</tr>
</tbody>
</table>

2015 DADH
The Convert-a-Card process

1. Contributions have been made from all over the world. A map displayed on our statistics page provides current location data.

2. By asking three volunteers to complete each task, we can be more confident about the accuracy of the records selected.

3. Datasets of all tasks and contributions, along with the image sets used, are openly available from our data page.

4. Search queries submitted to WorldCat are filtered so that only records from a set of trusted institutions are returned.

5. Cards for which there isn't an appropriate record in WorldCat will be fed back into the system to be used in alternative projects.

6. Data is checked by both curators and computers, to ensure that records are of a high enough quality to ingest.

7. Contributions that don't make it this far will not be wasted; they will just require more thorough checking to verify accuracy.

8. The ingest process is largely automated, which is certainly helpful when creating thousands of new catalogue records.

9. Each WorldCat holding includes a unique OCLC control number, which can be used to locate the correct record.

10. Catalogue records previously only searchable at the British Library in St. Pancras are now searchable worldwide.

A successful match is achieved when 2 volunteers select the same WorldCat® record.
LibCrowds – launched 8th of June 2015.

Pinyin Card Catalogue: Drawer Three

Transform this printed card catalogue into a searchable online database.

Indonesian Card Catalogue: Drawer Three

Transform this printed card catalogue into a searchable online database.
LibCrowds – launched 8th of June 2015.

http://www.libcrowds.com/
LibCrowds – launched 8th of June 2015.

http://www.libcrowds.com/
Technology

• Built on top of PyBossa, an open source framework for crowdsourcing projects.

• A custom theme was developed.

• Various plugins have been developed, some to handle library-specific features (e.g. Z39.50 communication)

• All code is open source and available at github.com/LibCrowds
Convert-a-Card

Welcome to Convert-a-Card

2015 DADH
The LibCrowds statistics page provides visual representations of some of the aforementioned data. This page is updated hourly.

http://www.libcrowds.com/statistics
The LibCrowds open data repository contains all images used and anonymised versions of all datasets produced via the platform. Everything available on this page is made available for reuse without copyright restrictions, under CC0.

http://www.libcrowds.com/data
Working with crowds: Quality control

- Use search queries that will only return records from ‘trusted’ institutions
- Ask three people to perform each task
- Generate spot check reports from random samples
- Various global changes made to the derived records
Post production

When drawer is complete, Curator downloads Grouped Data Zip which contains:

- **Matched**: Ready to go to Metadata Services team for upload to catalogue.
- **Ambiguous**: Denotes conflicting responses to records which are now in a special Libcrowds Task “For Further Inspection by Staff”.
- **Not Matched**: Denotes difficult records which will go back into a Transcription Libcrowds Task Workflow or be sent to a cataloguer.
- **Spot Check**: Curator can review a small selection of the matched files to verify.
Post production

{project_name}_matched.csv file sent to BL Metadata Services to create derived MARC records for catalogue
New catalogue record!

Title: Comparative grammar of the Tamil language / by P. S. Subrahmanya Sastri.

Author: P. S. Subrahmanya Sastri, 1890-1978.

Subjects: Dravidian languages -- Grammar, Comparative -- Tamil, Tamil language -- Grammar, Comparative -- Dravidian.

Publication Details: Trichinopoly, 1947.

Language: English

Identifier: System number 012505984
Recap: Convert-a-Card workflow

Catalogue card is scanned → Scanned cards are divided in batches → Suitable cards are uploaded on Flickr and LibCrowds

LibCrowds platform

Fully matched → Ingest into Library system

Ambiguous/Partial Match → Choose the right match

Unmatched → Manual input by cataloguer

Data cleaning

Spot check → Ingest into Library system

BL staff

Crowd

BL staff

Public
Working with crowds: Results

- 20,000 contributions
- 300 volunteers
- 29 countries
- 4 projects completed
- 75% success rate
- 4,000 records created

More at libcrowds.com/statistics
Statistics report

- We can study how our users engage with us
- We can produce material for further internal or external analysis and research
- We share the data in open access

Libcrowds enables research uncovering collection items but it is also producing data worth being analysed – it produces material for research.
Successful?

Early experimentation + ongoing maintenance + development of new features = **500 hours** spent by BL staff so far

Furthermore: while ongoing work will be required, the platform is now firmly established and the technical process largely automated, as such the overall number of staff **hours spent** on the project has already **decreased**.

If converted with an external contractor: £20-30 = **£80,000 to £120,000** for the 4,000 records created so far.

Plus:
- External engagement
- On-going dialogue
- Data production
Working with crowds: Engagement

- Twitter
- AAS blog
- Platform blog
- Other social media? Weibo?
- Community chat on platform
- Conferences, publications, bulletins...

Continuous engagement
Working with crowds: Engagement

Effects of a successful blog post

Future success will largely depend on consistent promotion and keeping volunteers engaged
Our aim

To provide **free and easy access** to the Library collections catalogues and items to worldwide researchers, students and the general public via Explore BL on-line catalogues.
Living Knowledge

2023 vision

six core statements of purpose:
custodianship, research, business, culture, learning and international

we contribute to research, culture, education and economic prosperity, for the benefit not just of the UK but of users and partners around the world.

To make our intellectual heritage accessible to everyone, for research, inspiration and enjoyment

The British Library to be the most open, creative and innovative institution of its kind by 2023.

In this framework, the LibCrowds experiment of retro-converting our printed card catalogue collections through crowdsourcing has proven a success.
Thank you!

謝謝

Contacts:
sara.chiesura@bl.uk
nora.mcgregor@bl.uk
alex.mendes@bl.uk
crowdsourcing@bl.uk