LIBRARY COMMITTEE

Wednesday 10 October 2018

PhD Digitisation Project update

Description of paper
1. This paper provides an update on the Library’s project to digitise its entire collection of PhDs, dating from the late 16th century to the present day.

Action requested
2. Library Committee is invited to acknowledge and comment on the paper

Recommendation
3. Library Committee is asked to note this update and circulate it to colleagues.

Background and context
4. For the last two years, the Library has undertaken a project at the University Collections Facility (South Gyle) to digitise its entire collection of PhD theses. The scanning element of this project was completed in July 2018 and a small amount of post-processing work remains to be done. The project is on track to enable the Library to provide access to the entire collection on-line by the end of 2018.

This project is the first mass digitisation project undertaken by the Library and complements the existing digitisation activity undertaken in the Digital Imaging Unit by placing a greater focus on throughput while ensuring quality standards.

5. The project is a key element in the Library’s digitisation strategy and feeds directly into ISG’s strategic vision, particularly Student experience and the unique Edinburgh offer (the theses are unique to Edinburgh); and online and distance learning leadership (theses can be accessed online).

6. A short video was created to summarise the project: https://youtu.be/8uk5gkB9yQ

Progress
7. The project aimed to digitise 17,000 theses not already in digital format, making this vast collection of unique Edinburgh research available open-access. It sought to create MARC records for uncatalogued theses and provide conservation treatment for any items which were unstable and in danger of being damaged during the digitisation process. At the end of the project, every doctoral-level University of Edinburgh thesis would exist both in physical and digital format, be adequately described and, where necessary, have received conservation treatment.

8. In total, the project has delivered the digitisation of 16,787 doctoral-level theses volumes. Approximately half of the collection (8,233 volumes) had duplicate hard-copy volumes while the rest had only one copy. Of these unique items, the digitization of approximately 4,000 was outsourced.
9. To date, 13,520 theses have been uploaded to ERA, with the remaining 3,267 to be uploaded in monthly batches prior to the end of 2018.

**Thesis types**
10. Throughout the course of the project, we have digitised a remarkable range of thesis types, ranging from single-page 16th century Latin disputations, through small A5 printed Latin volumes to 19th century handwritten MDs to modern typewritten or word-processed PhDs. **Dating from 1599, we believe our collection is the largest and the longest continuous run of any institution in the UK.**

11. These different formats have provided challenges for digitisation, especially large fold-outs, maps and computer printouts which have required flexibility and creativity in our approach. Where we have found material that we cannot digitise ourselves, we have worked with colleagues in the DIU (slide-scanning), School of Scottish Studies (audiovisual material) and the NLS (very large maps and foldouts) to ensure all material is in digital format.

12. Over the course of the project we have also come across “curiosities” ranging from the revolting (laminated slices of human lung), to the comical (a bullet which the author appeared to have accidentally shot himself with) and the poignant (photographs of buildings in Syria and Iraq which have been destroyed in the recent conflict in that region). Our project blog provides more information on these interesting discoveries: [http://libraryblogs.is.ed.ac.uk/phddigitisation/](http://libraryblogs.is.ed.ac.uk/phddigitisation/)

13. We will also be holding an exhibition in the CRC in January 2019 to tell the story of the Edinburgh thesis and to celebrate the end of the project.

**Notable alumni**
14. We have digitised the theses of several notable individuals during the course of the project, including Alexander McCall Smith’s *In defence of duress*, the thesis of Wong Fun, the first Chinese graduate from a European University, and the MD of Isabel Emslie Hutton, a doctor in the First World War who became a Serbian national heroine.

15. As part of the cataloging element of the project, the Metadata team have worked to identify notable individuals and match them up to Library of Congress authority records so that the digitised theses can be linked to other works created by that author. This work has identified notable authors who do not yet have a presence on Wikipedia, enabling us to run Wikipedia training sessions throughout July 2018 for Library staff which resulted in the creation of 20 new articles about authors which link back to their digitised thesis.

**Digitisation**
16. The approach to digitisation varied depending on whether the item was unique or duplicate and depended on the physical condition of the book being scanned. For duplicate volumes, we digitised one copy destructively, meaning its spine was removed using a guillotine and the pages were then fed through a 100-page-per-minute Kodak i4250 document scanner. In this way, we were able to scan a typical, straightforward 350 page modern thesis in under five minutes.
17. Where there was only one copy, we scanned non-destructively using Copibook Cobalt scanners and V-shaped scanners at the DIU and New College when bindings were especially tight.

18. Following digitisation, images were then processed using LIMB software to ensure no pages were missing or blurry, all text lines were straight and images sharp and that all typed, printed and word-processed theses were OCR-ed and therefore keyword searchable. The final, optimized file was then uploaded to ERA by the Scholarly Communications team.

19. Although the theses are still in copyright, we have decided that there is a very low risk of complaint by the author and have developed a takedown policy stating our willingness to remove a thesis if its author requests it. We have received overwhelmingly positive feedback about the project to date and have only had two takedown requests, both of which were resolved immediately.

**Outsourcing**

20. We outsourced approximately 4,000 theses to our digitisation partner Autodocs, who were appointed after a thorough procurement exercise in 2017. The items outsourced were modern theses in robust physical condition. We decided only to outsource this type so that we could focus on damaged, fragile or more awkward items in-house.

21. Aside from some initial communication and handling issues, the outsourcing of digitization work was a great success and was completed on time and to budget.

**Download statistics**

22. We keep track of downloads on a monthly basis and compare our statistics with those of other repositories across the UK. At the end of August 2018, theses downloaded as part of this project had been downloaded a total of 46,636 times, with the most popular items outlined in the table below:

<table>
<thead>
<tr>
<th>Title</th>
<th>Author</th>
<th>Year</th>
<th>No. downloads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Social differentiation of English in Norwich</td>
<td>Trudgill, P.J.</td>
<td>1971</td>
<td>615</td>
</tr>
<tr>
<td>Business ethics and the corporate manipulation of expressions</td>
<td>Song, Ji-Hwan</td>
<td>1996</td>
<td>327</td>
</tr>
<tr>
<td>Study of the interpersonal dimension of narrative fiction with specific reference to power and control in Muriel Spark's Memento mori and its implications for the teaching of English literature in a TEFL context</td>
<td>Myo-Myint</td>
<td>1987</td>
<td>281</td>
</tr>
<tr>
<td>Theory of al-Masalih al-mursalah in Islamic law</td>
<td>Mtupah, Juma Mikidadi Omari</td>
<td>1990</td>
<td>208</td>
</tr>
</tbody>
</table>
23. The graph below shows the comparative number of downloads on a monthly basis since the initial upload in July 2016.

24. For the last 12 months, ERA is seventh behind the White Rose group, UCL, Exeter, Birmingham, Manchester and Cambridge (note the figures below are for total thesis downloads, not just those scanned as part of this project). This is a drop from fifth place in February 2018.
Next steps
25. Following the completion of the scanning element in July 2018, a small amount of work is being undertaken to finish post-processing and complete outstanding cataloguing. This work will be completed by the end of October and all remaining files will then be uploaded to ERA prior to the end of 2018 project deadline.

26. Of the five digitisation team members, two have gone on to higher grade roles within the University, one has left the University to take a professional qualification and the final two have moved over to the Scottish Session Papers digitisation project.

27. We are actively exploring ways that the equipment and software purchased for the project can be used in the future. There are a number of possible projects for this including potential document scanning of Moray House BRE index cards and digitisation of the School of Scottish Studies transcriptions catalogue. In the short term, the scanners could be repurposed for ILL and Ereserve scanning by the Collections Lifecycle Management teams.

28. Added to this, we are looking to develop a proposal to complete the digitisation of the remaining non-doctoral theses such as the collection of MScs and MPhils.

29. We are seeking avenues to fully exploit this digitised collection using advanced research methods: the fully OCR-ed PDFs have excellent potential as a source for text mining, image recognition and other digital scholarship techniques.

Lessons learned
30. The PhD digitisation project was the Library’s first attempt to digitise collections at this scale, and we have learned a great deal from the experience.

31. The final lessons learned and closure reports are still to be written (the project does not finish until January 2019, when all theses have been uploaded to ERA) but some of the main lessons learned, which we have begun to implement in subsequent projects, are outlined below:

- Conservation: it is vital that conservation staff are involved directly in all discussions around workflows, the purchase of equipment and outsourcing to ensure that no damage is done to the collections during digitisation.
- Digitisation: all barriers to efficiency should be removed prior to digitisation, so that the operator can focus exclusively on image capture without distraction. This means items for digitisation should be pre-selected with metadata ideally created before digitisation.
- Staff motivation: there needs to be an element of variety within scanning activity as staff productivity can be negatively affected during repetitive work. For the thesis project this meant rotating staff regularly between destructive scanning, non-destructive scanning and processing. Non-scanning tasks such as writing blog posts, managing social media and reshelving also help to break up the day. Giving staff specific responsibilities (for example one team member was appointed the 'Conservation Champion'), providing clear targets, while engaging the team in the wider work of the University through training and attending meetings all aided motivation and productivity. With the team
based at the South Gyle it was important to take steps so they felt included with the wider team.

- Project management: it is crucial that the whole project team is involved from the start of the project to agree budgeting, planning, risk analysis, milestones and other project elements, including quality assessment and project assurance.
- Infrastructure: important to ensure the correct network infrastructure is in place before any digitisation work begins. This project was based at the University Collections Facility at South Gyle, which had a very slow network connection with the central area until the start of 2018. This caused difficulties with file backups, working remotely and sharing resources across the team.
- Temporary staff: towards the end of the project we hired several temporary staff from Pertemps and were very lucky to have received high quality staff. There is great value in bringing in temporary staff who are motivated to come in and perform a job to a high standard in a short, defined period of time.

St Andrews thesis digitisation
32. As well as undertaking digitisation of our own collection, we also digitised approximately 1,400 thesis volumes for the University of St Andrews, who approached us with a request to digitise their collection as part of the Scottish Digital Library Consortium’s (SDLC) shared service agreement.

33. It was an interesting challenge to juggle the completion of our own project while providing a service for another institution, and we learned a lot from being a digitisation service provider.

34. This project has now been successfully completed and enabled us to retain digitisation staff in the gap between the end of our project in May 2018 and the start of the Scottish Session Papers project in August 2018.

35. Feedback from St Andrews was very positive about the management of the service and the quality of digitisation provided.

Resource implications
36. There are currently no resource implications related to this update.

Risk Management
37. There are no risks associated with this update.

Equality & Diversity
38. An EqIA for the project, and for digitisation activity more generally, is in development.

Next steps/implications
39. Any agreed amendments to the update will be made and circulated to the Library Committee via email.

Consultation
40. This paper is for discussion at Library Committee. It has not been circulated to any other committee for revision.
Further information
41. Further information can be supplied by:

Author
Gavin Willshaw
Library and University Collections
10 September 2018

Presenter
Jeremy Upton
Director of L&UC

Freedom of Information
42. This paper is open.