# Mass digitisation at the University of Edinburgh

Gavin Willshaw
Digital Curator
University of Edinburgh
gavin.willshaw@ed.ac.uk

## Introduction

The University of Edinburgh library is currently undertaking a large-scale in-house theses digitisation project. Digitisation at the library has, until recently, largely focused on the creation of high-quality digital images of collection items in response to reader requests, for project work or for documentary purposes. Two full-time professional photographers have, over the last ten years, built up a collection of approximately 40,000 digital images, of which many can be viewed in high resolution online and downloaded from http://images.is.ed.ac.uk.

In the past the library would outsource large-scale digitisation projects, the largest of which was the digitisation of 5,000 duplicate science and engineering theses in 2015–16. This resulted in approximately one million digital images.

## The project

The library holds in the region of 27,000 PhD theses dating from the early 17th century to the present day. Approximately 10,000 are in digital format already. The new project seeks to digitise the remaining 17,000 so that this entire collection of unique Edinburgh research is available for anyone, anywhere in the world, to download free of charge.

*Fig. 1 The collection*

To carry out the digitisation, the library has invested in equipment and software, and has recruited a team of digitisation assistants to scan, process and upload 13,000 volumes; a further 4,000 will be outsourced. The in-house element will be completed by May 2018, with a target of having all theses online through ERA (Edinburgh Research Archive), our digital repository, by the end of 2018.

Of the 13,000 scanned in house, around 10,000 currently exist in duplicate form in print, meaning they can be scanned destructively; the remaining 3,000 are unique items which require careful, non-destructive digitisation. Duplicate theses have their boards and spines removed before being fed through a Kodak i4250 document scanner, while unique theses are scanned using a manual i2s Copibook Cobalt scanner. All scanned theses are then processed, made keyword searchable and quality reviewed with LIMB server processing software before they are uploaded to ERA.

# Mass digitisation at the University of Edinburgh

*Fig. 2 Non-destructive scanning*

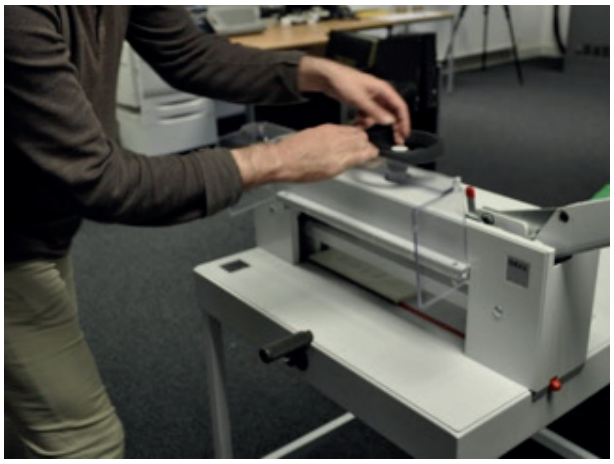

*Fig. 3 Spine removed for destructive scanning*

*Fig. 4 Document scanner for destructive scanning*

## Conservation and cataloguing

Image capture is only one component of the project. Around 2,000 PhD volumes require conservation treatment and another 4,000 uncatalogued theses require catalogue records.

## Conservation

Conservation treatment ranges from basic cleaning and dusting, a task performed by the scanning assistants and volunteers, through to more specialist conservation work to reattach boards, repair damaged volumes, loosen tight spines and, on some occasions, remove theses entirely from their bindings and re-house them in boxes. A full-time Projects Conservator has been appointed to complete this work and to train scanning assistants

# Mass digitisation at the University of Edinburgh

in handling rare and unique collections. The Projects Conservator provides conservation guidance and expertise as and when required.
The inclusion of a qualified conservator in the project has been vital in ensuring that best practice in handling is adhered to and that no damage is done to volumes as they are scanned. The vital role the conservator plays in the team highlights how conservation must be a fundamental element of any digitisation project.

## Cataloguing

4,000 PhD theses in the collection have no MARC catalogue records, their bibliographic information being accessible only through the paper catalogue held in the Centre for Research Collections (CRC). For these theses, metadata is being created at the point of scanning and converted into basic MARC records by cataloguing staff.

By incorporating conservation and cataloguing into the project, we shall ensure that at project end, all University of Edinburgh PhD theses will exist in physical and digital format, will have undergone conservation work and will be catalogued and discoverable on our library Discovery system DiscoverEd.

## Benefits of our approach

Arguably, it might have been simpler and cheaper to outsource the entire collection of 17,000 theses – an approach adopted by colleagues at a number of universities. Anecdotal feedback and our own experience of outsourcing have suggested that this approach has its own complications and is far from straightforward.

We decided for a number of reasons to undertake a hybrid approach, scanning some theses in house and outsourcing the rest. First, there is a strong digitisation agenda among research libraries, as exemplified through the National Library of Scotland's goal to have a third of its collections in digital format by 2025, and the University Library is keen to develop its skills and experience in this area in order to inform future projects. Secondly, the software, equipment and materials used on this project will all be available for future mass and high-quality digitisation initiatives.

While this project is large in scale, in many ways it is also a pilot project as we explore our options and develop the best approach to mass digitisation within the organisation.

## Thesis scanning service

Alongside this project, the library offers a thesis scanning service that allows readers to pay for a thesis to be digitised on demand. While our project will ultimately negate the need for such a service, demand continues to be strong for this direct service. Since the project team took on responsibility for the service in October 2016, there have been 35 requests, which is in line with the hundred or so requests per year before the project began. Undertaking this request service as part of the project has provided the team with a customer-focused dimension to their work, and service turnarounds have shortened considerably thanks to our more specialist scanning equipment.

## Progress

The scanning team is on course to complete the in-house scanning element by May 2018, and it is anticipated that the first batch of outsourced theses will be collected in early April and returned by July 2017. As of 3 February 2017, the in-house team had scanned 4,500 volumes (out of an in-house total of 13,000), 3,500 University of Edinburgh theses now being available through ERA.

# Mass digitisation at the University of Edinburgh

## Use of digital theses

There is strong demand for digitised PhD theses among the academic community: digital copies are downloaded from our repository an average of thirty times each per month, whereas the total number of physical volumes consulted in our reading room rarely surpasses a hundred per year. To put that in perspective, one individual digital thesis is accessed more times in three months than the entire collection of physical theses is accessed in a year.

That is not to say that creating digital files is the end point. Now that our scanning processes are embedded and workflows established, we are considering ways in which we can increase awareness and use of the collections. We have begun working with the university's Wikimedian in Residence to explore how the PhD thesis collection can be linked to author pages in Wikipedia, particularly making sure that a link to their thesis is included in the author's Wikipedia info box (which appears at the top of Google search results). One example is the Wikipedia page of Ernest Bashford, whose page was created three months ago and has been viewed 214 times. At this stage the process is entirely manual, although there is scope to explore whether it can be automated.


*Fig. 5 Wikipedia infobox for Ernest Bashford*

We are also exploring uploading older theses into Wikisource, Wikimedia's online library of out-of-copyright texts. Again this is still a largely manual task, but it would be interesting to explore whether any of the processes can be automated. At February's History of Medicine Wikipedia editathon, we imported T.S. Jehu's thesis 'Some problems in variation and heredity' into Wikisource as a test for future works.

In addition, we intend to start exploring advanced research techniques such as data visualisation, text mining and image recognition and to look at areas such as geolocation and the demographic breakdown of authors.

## Future projects

We hope that the expertise and equipment gained from this project will put the library in a strong position for future mass digitisation initiatives. While this project specifically deals with the digitisation of PhD and other doctoral level theses, the library holds several thousand non-doctoral theses, including a visually diverse collection of dissertations from Edinburgh College of Art, which could form a standalone project. Scoping work is under way for a large-scale project to digitise the Scottish Session Papers held by the University of Edinburgh, Signet Library and Advocates Library, and there is potential for the mass digitisation of out-of-copyright general collections. We are also beginning to investigate the possibility of using our document scanners to provide an internal digitisation service for university departments.

## Conclusion

The PhD digitisation project is the library's first attempt at mass digitisation and has provided many opportunities for learning and development. Digitisation activity is on schedule and the creation of 17,000 digital theses will provide a huge corpus of text for traditional researchers and for those who wish to employ advanced research techniques.

Keep up to date by visiting our project blog: http://libraryblogs.is.ed.ac.uk/phddigitisation/