Ph.D. Theses Mass Digitization at ULB

Anthony Leroy, Benoit Pauwels
Université libre de Bruxelles
Brussels, Belgium
{anleroy, bpauwels}@ulb.ac.be

In 2012, our library initiated a mass-digitization project of all the Ph.D. theses produced at the university since its creation in 1834.
The goal was twofold: improving the visibility and accessibility of our scientific research and freeing space for the development of our learning centers. Over ten thousand volumes and three million pages needed to be digitized.
We decided not to outsource the project to a vendor but to perform it fully in-house. This choice offered several advantages: we kept full control over the production, it was less expensive and, once in place, the infrastructure could be reused for other projects.
After three consecutive years of intensive work, our digitizing project is now about to end.
This poster describes the main characteristics of our digitization workflow custom-designed to fit our dissemination and preservation requirements.
The workflow handles each step from the collection of paper volumes on bookshelves to the online release of the corresponding digitized objects in our institutional repository and their storage in our preservation repository.
While the most prestigious theses have to be digitized with a non-destructive book scanner, other paper volumes are to be sent for destruction and recycling. Most theses can thus be digitized with a production feeder scanner after trimming the book binding.
Thesis volumes are duplex-scanned twice with ultrasonic multi-feed hardware detection and multiple page count verification.
The first pass produces raw uncompressed TIFF images dedicated for preservation and keeping all the features of the original paper version.
The second pass produces JPEG images optimized for dissemination by automatic image quality improvement filters (background whitening, sharpness, color calibration).
Quality assurance was paramount in the design of our workflow. It is fully automated to lower the risk of human error: metadata is directly extracted from the library catalog to generate QR identification barcode. All generated files are automatically named based on the barcode content.
Quality assurance is also supported by custom-designed quality control software. Quality control is always performed with the original in hand. The application allows the operator to report quality issues both in the digitized object and in the original paper object.
Every week, this workflow generates up to half a terabyte of raw images which would be too costly to preserve. For preservation purposes, raw images are thus migrated to visually lossless JPEG 2000 format for one tenth the original image size with the same quality of experience. The image quality degradation due to compression is automatically controlled with image comparison metrics.
For dissemination purposes, ocerized PDFs are generated from the JPEG files with ABBYY OCR 11 with mixed-raster content technology to provide the best trade-off between image quality and file size.
Dissemination and preservation copies are referenced in the institutional repository using a custom-designed METS profile.
The workflow outlined in this abstract is designed to be generic and thus can be reused for our other in-house digitizing projects.