Ph.D. Theses Mass Digitization at ULB
Anthony Leroy, Benoit Pauwels

To cite this version:
Anthony Leroy, Benoit Pauwels. Ph.D. Theses Mass Digitization at ULB. 19th International Symposium on Electronic Theses and Dissertations (ETD 2016): "Data and Dissertations", Jul 2016, Villeneuve d'Ascq, France. hal-01430995

HAL Id: hal-01430995
https://hal.univ-lille.fr/hal-01430995
Submitted on 10 Jan 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
PhD Theses Mass Digitization at ULB
Anthony Leroy, Benoit Pauwels

Abstract
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 of the project was twofold: improving the visibility and accessibility of our scientific research and freeing space for the development of our learning centers. Over 10000 volumes and 3 million pages needed to be digitized.

To Outsource or Not to Outsource? That Is the Question.
We decided to insource this mass digitization project.
- total control of the production
- less expensive
- once in place, the workflow can be reused
- extra human resources needed
- longer time to start
- taking full responsibility

We developed a custom-designed digitizing workflow
After trimming the book binding of incoming theses, each volume is duplex-scanned twice. Quality control is then performed with the original in hand. The paper volume is finally boxed for storage using a custom cloud application.

Automated & manual cross-checks ensure low error risk
automatic filenaming with QR code generated from our catalog
duplex scanning even for simplex print
both raw and automatic quality improved images
double scan with iMFF: portrait and landscape
4 page count verifications
software-assisted quality control

Quality Assurance is supported by custom software
Quality control is always performed with the original in hand. The client-server application allows the operator to report quality issues both in the digitized object and in the original paper object.

Digitizing has two aims: dissemination and preservation

Dissemination (mixed-raster content ABBYY OCR PDF, ALTO, txt)

Preservation (JPEG 2000)

Thesis representations in our institutional repository
The submission representation of the thesis (TIFF files) is automatically transformed into an archival representation (JPEG 2000 files), from which various dissemination representations are generated (PDF, ALTO, txt). These representations are referenced in our institutional repository using a custom-designed METS profile.

Digitized theses are preserved in SAFE PLN
The SAFE Private LOCKSS Network is an international distributed preservation repository based on the LOCKSS technology operated by seven universities.

Each collection of archives is automatically replicated in every box constituting the network, thus ensuring a broad geographical data redundancy.
Automatic data integrity monitoring and comparison is performed regularly through the network via secured connections. Bit-rot and other forms of data corruption can thus be quickly detected, reported and corrected.