ADOCHS

Auditing Digitization Outputs in the Cultural Heritage Sector

DURATION
15/12/2015 – 31/12/2020

BUDGET
596.652 €

PROJECT DESCRIPTION

Since the mid-nineties, cultural institutions have undoubtedly entered the digital age. In Belgium, the government adopted in 2004 a first digitization plan for a ten-year period which has led to the realization of nine digitization projects in the federal scientific institutions. These digitization projects required substantial human and financial resources in order to overcome unforeseen difficulties. As a result, in 2014, a second phase was launched, allowing institutions to continue the digitization activities of the past decade. The expertise accumulated during the first stage of digitization can help us to engage future projects in a better-thought-out manner. It is precisely in this context that we introduce this proposal.

The issue of quality control was one of the major obstacles in the first phase of digitization. Indeed, it appeared that many projects had underestimated the extent of this step both in human and technical terms in the overall process of digitization. In most cases, teams were faced with a lack of methodological standardization and automation tools to perform the job. They therefore often had to work manually, without procedural guidelines adapted to their specific needs. However, it is clear that quality control is an essential component to every stage of a digitization project if you want to ensure the integrity and consistency of files and data produced, as well as their long-term conservation. This is true both for outsourced or internal digitization projects. In addressing the issue of quality control, this project aims to speed up the whole digitizing process while minimizing the costs and also to increase the value of the data produced in the framework for future digitization projects.

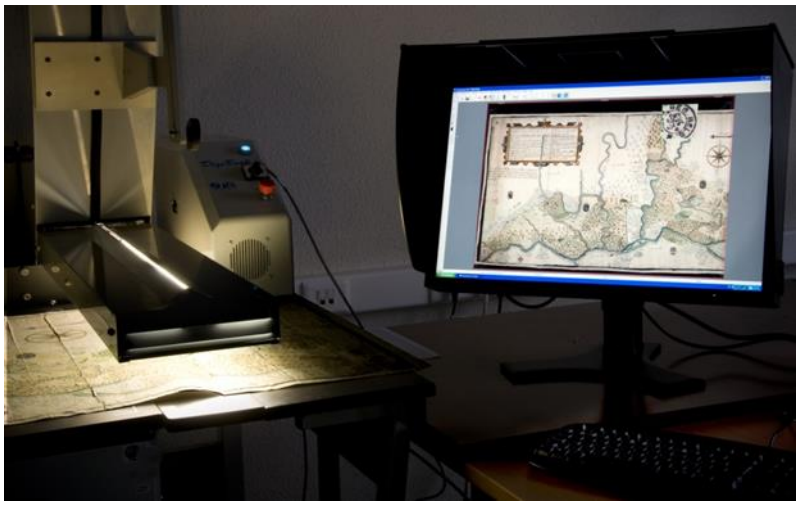
While federal scientific institutions will be the main beneficiaries, the purpose of this research addresses a much broader need that affects all heritage institutions in Belgium and abroad. The goal is to address the issue in two stages. First, by focusing on methodological aspects, proposing guidelines applicable in the treatment of heritage collections. Then, by developing technical tools that automate tasks related to quality control or support the manual check. Both technical aspects (image resolution, file integrity formats, etc..) and content metadata (descriptions of collections, compliance with XML schema's, etc..) will be taken into account. In addition to the international scientific literature, researchers can refer to the expertise acquired by the various institutions during the first phase of digitization but also to the needs of the new projects. In this perspective, the researchers are working on two types of collections: the digitized collections of the Royal Library of Belgium that address the problems with textual documents and the photographic collections of the CEGESOMA which provides a set of iconographic documents. Though the objects differ, their scanning process has a lot in common and hence we can say that the following generic quality errors occurred in the first phase of the digitization campaign and will remain to occur in the workflow if no quality control (QC) procedure is developed - hence they will be addressed in the context of this project:

incomplete scanning's (pages are missing or part of a page is missing)
mistakes in scan ordering
color scan instead of grey scale scan and vice versa
changes in the resolution or file type
formal errors with manual metadata transcription and encoding
inability to produce structured metadata
unsharp images (whole document or just parts)
non-uniform color representation
badly or uncropped images
OCR quality (mistakes in text converting and zone detection, e.g. title, issue number etc)

These errors can be classified into two categories: 1) those arising from pure manhandling mistakes and 2) those which are an unsatisfactory software output. In the latter case, these can come from a combination of a) again manhandling errors, b) limitations in the used software or c) degradations in the object itself. We therefore mitigate the mentioned quality problems in relation to their origin.

A VUB PhD student from the Department of Electronics and Informatics (ETRO) is responsible for the image quality control (appointed part time by the VUB and part time by the KBR) while an ULB PhD student is responsible for quality control of the metadata (appointed part time by the ULB and part time by the CEGESOMA). They are both hired part-time by the university and part-time by the institutions because the scientific work should be considered in connection with the problems encountered in situ. In this context, the research results is regularly confronted with case studies.

Two additional researchers are also provided for a period of one year each: one at the beginning of the project and the other at the end. The first task will be to draw up an inventory of good practices in quality control procedures, whether it concerns digital images or metadata (see the task description). The second will be responsible for integrating the research results at the end of the project in order to define a clear procedure of quality control to be followed in the heritage institutions.



CONTACT INFORMATION

Coordinator

Florence, Gillet
State Archives of Belgium / CegeSoma
florence.gillet@arch.

Partners

Lemmers Frederic
Royal Library of Belgium
frederic.lemmers@kbr.be

Ann Dooms
Vlaamse Universiteit Brussel (VUB)
Department of Electronics and Informatics (ETRO)
Multimedia Forensics Team
ann.dooms@vub.ac.be

Seth Van Hooland
Université Libre de Bruxelles (ULB)
Department of Sciences and Technologies of Information and Communication
svhoolan@ulb.ac.be

LINKS

http://adochs.be/

BELGIAN SCIENCE POLICY OFFICE
Louizalaan 231 Avenue Louise • B-1050 Brussels
Tel. +32 (0)2 238 34 11
http://www.belspo.be/brain-be/ • Email : BRAIN-be@belspo.be