

ADOCHS

Contrôler les données produites par les processus de numérisation dans le secteur du patrimoine culturel

DUREE
15/12/2015 – 31/12/2020

BUDGET
596.652 €

DESCRIPTION DU PROJET

Depuis le milieu des années 90, les institutions culturelles sont incontestablement entrées dans l'ère du numérique. En Belgique, le gouvernement a adopté en 2004 un premier plan de numérisation pour une période de dix ans, qui a conduit à la réalisation de neuf projets de numérisation dans les institutions scientifiques fédérales. Ces projets de numérisation ont nécessité d'importantes ressources humaines et financières, entre autres pour pallier aux imprévus. Une deuxième phase a été lancée en 2014, permettant aux institutions de poursuivre les activités de numérisation entamée précédemment. L'expertise accumulée au cours de la première phase a permis de concevoir les nouveaux projets de manière plus réfléchie. C'est précisément dans ce contexte que s'inscrit cette proposition.

La question du contrôle de qualité fut l'un des obstacles majeurs de la première phase de numérisation. En effet, il est apparu que de nombreux projets avaient sous-estimé l'ampleur de cette étape, tant sur le plan humain que technique, dans l'ensemble du processus de numérisation. Dans la plupart des cas, les équipes ont été confrontées à un manque de normalisation méthodologique et d'outils d'automatisation pour effectuer le travail. Ils ont dû souvent travailler manuellement, sans directives de procédure adaptées à leurs besoins spécifiques. Il est pourtant évident que le contrôle de qualité constitue un élément essentiel à chaque étape d'un projet de numérisation si l'on souhaite garantir l'intégrité et la cohérence des fichiers et des données générés, ainsi que leur conservation à long terme. Cela vaut tant pour les projets de numérisation externalisés que pour les projets de numérisation réalisés en interne. En abordant la question du contrôle de qualité, ce projet vise à accélérer le processus de numérisation dans son ensemble tout en minimisant les coûts et à augmenter la valeur des données produites dans le cadre de futurs projets de numérisation.



Bien que les institutions scientifiques fédérales en soient les principaux bénéficiaires, le but de cette recherche est de répondre à un besoin beaucoup plus vaste qui concerne toutes les institutions patrimoniales en Belgique et à l'étranger. L'objectif est d'aborder la problématique en deux temps. Premièrement, en se concentrant sur les aspects méthodologiques du contrôle de qualité, en proposant des lignes directrices applicables au traitement des collections patrimoniales. Ensuite, en développant des outils techniques qui automatisent les tâches liées au contrôle de qualité ou prennent en charge la vérification manuelle. Les aspects techniques (résolution d'image, formats d'intégrité de fichier, etc.) et les métadonnées de contenu (descriptions des collections, conformité avec les schémas XML, etc.) seront pris en compte. En plus de la littérature scientifique internationale, les chercheurs peuvent se référer à l'expertise acquise par les différentes institutions au cours de la première phase de numérisation, mais également aux besoins des nouveaux projets. Dans cette perspective, les chercheurs travaillent sur deux types de collections : les collections numérisées de la Bibliothèque royale de Belgique qui traitent des problèmes liés aux documents textuels et les collections photographiques du CEGESOMA qui fournit un ensemble de documents iconographiques.

ADOCHS

Bien que les objets diffèrent, leur processus de numérisation a beaucoup en commun. Il est donc possible d'identifier que les erreurs de qualité génériques suivantes – qui seront abordées durant le projet ADOCHS – se sont produites lors de la première phase de numérisation et resteront présentes dans le cadre des nouveaux projets si aucune procédure de contrôle de la qualité (CQ) n'est mise en place :

- Numérisation incomplète (des pages sont manquantes ou une partie d'une page est manquante)
- Erreurs dans l'ordre des scans
- Numérisation couleur au lieu de numérisation en niveaux de gris et inversement
- Changements dans la résolution ou le type de fichier
- Erreurs formelles lors de la transcription et du codage manuel des métadonnées
- Incapacité à produire des métadonnées structurées
- Images floues (tout le document ou seulement des parties)
- Représentation des couleurs non uniforme
- Images mal ou non recadrées
- Qualité de l'OCR (erreurs de conversion de texte et de détection de zone, par exemple, titre, numéro d'édition, etc.)

Ces erreurs peuvent être classées en deux catégories : 1) celles résultant d'erreurs de manipulation pure et 2) celles qui constituent une sortie logicielle insatisfaisante. Dans ce dernier cas, celles-ci peuvent provenir d'une combinaison de: a) encore des erreurs de manipulation, b) les limites du logiciel utilisé ou c) des dégradations de l'objet lui-même. L'objectif est d'atténuer les problèmes de qualité mentionnés en fonction de leur origine.

Un doctorant de la VUB, du département d'électronique et d'informatique (ETRO), est responsable du contrôle de qualité des images (engagé à temps partiel par la VUB et à temps partiel par la KBR), tandis qu'un doctorant de l'ULB est responsable du contrôle de qualité des



métadonnées (engagé à temps partiel par l'ULB et à temps partiel par la KBR et à temps partiel par la CEGESOMA). Ces chercheurs sont engagés partiellement par les universités et par les institutions car leurs recherches scientifiques doivent être envisagées directement en lien avec les problèmes rencontrés in situ. Dans ce contexte, les résultats de la recherche sont régulièrement confrontés à des études de cas.

Deux chercheurs supplémentaires sont également prévus pour une période d'un an chacun: l'un au début du projet et l'autre à la fin. La première tâche consistera à dresser un inventaire des bonnes pratiques en matière de procédures de contrôle de la qualité, qu'il s'agisse d'images numériques ou de métadonnées (voir la description de la tâche). Le second sera responsable de l'intégration des résultats de la recherche à la fin du projet afin de définir une procédure claire de contrôle de la qualité à suivre dans les institutions patrimoniales.

COORDONNEES

Coordinateur

Florence, Gillet
Archives de l'Etat en Belgique / CegeSoma
florence.gillet@arch.be

Partenaires

Lemmers Frederic
Bibliothèque royale de Belgique
frederic.lemmers@kbr.be

Ann Doods
Vlaamse Universiteit Brussel (VUB)
Département d'Electronique et Informatique (ETRO)
Multimedia Forensics Team
ann.doods@vub.ac.be

Seth Van Hooland
Université Libre de Bruxelles (ULB)
Département en Sciences et Technologies de l'Information
et de la Communication
svhoolan@ulb.ac.be

LIENS

<http://adochs.be/>

