## RÉSUMÉ


### Contexte

La numérisation à grande échelle constitue un effort important dans le secteur du patrimoine culturel, où une énorme quantité de documents originaux tels que des journaux historiques, des manuscrits ou encore des cartes ont été et continuent à être convertis en formats numériques. Au cours des dernières années, avec la mise à disposition d'une quantité massive de collections numériques, certains problèmes universels se sont manifestés. Tant l'investissement humain que certains défis techniques ont été clairement sous-estimés et dans de nombreuses initiatives de numérisation, le contrôle de la qualité reste un défi. Or, le contrôle de la qualité dans sa globalité est essentiel pour assurer l'intégrité, la cohérence et la conservation à long terme des fichiers et des données produits, ainsi que leur accès public.

### Objectifs & méthodologie

Le projet ADOCHS (2016-2021) est le fruit d'une collaboration entre le Centre d'Études Guerre et Société (CegeSoma, qui fait partie des Archives de l'État en Belgique), la Bibliothèque royale de Belgique (KBR), la Vrije Universiteit Brussel (VUB) et l'Université libre de Bruxelles (ULB). Ce projet visait à apporter des améliorations significatives en matière de contrôle qualité dans le cadre de la numérisation du patrimoine, en développant de nouvelles approches méthodologiques et en créant un ensemble de directives et d'outils directement utilisables en vue d'une approche étape par étape des projets de numérisation; avec comme objectif global **d'améliorer la qualité des images et des métadonnées produites dans le cadre de la numérisation patrimoniale et documentaire**. Trois points focaux ont émergé : l'amélioration de la qualité des images, des métadonnées et des processus de numérisation.

1/ En ce qui concerne **la qualité des images dans le secteur du patrimoine cultural**, trois volets ont été étudiés dans le cadre de ce projet, à savoir : la DIS (Document Image Segmentation), l'IQA (Image Quality Assessment) et la DDR (Document Damage Recognition). Ces trois sujets forment un tout cohérent et sont étroitement liés les uns aux autres dans le cadre du processus de la numérisation et de l'extraction d'informations. Ces trois volets ont été abordés d'un point de vue mathématique, en exploitant leur homogénéité. Cette recherche (initiée par le Dr Tan Lu) a conceptualisé et modélisé l'homogénéité de manière à caractériser les relations entre ces éléments et à pouvoir en distinguer des modèles uniques. Ce travail a permis en particulier de proposer un **modèle d'homogénéité probabiliste pour le volet de la DIS**, en définissant le concept d'homogénéité textuelle à partir le prisme des principes de Gestalt. Un modèle statistique exploitant l'homogénéité du texte a ensuite été formulé pour la classification de zones textuelles et non textuelles dans le contexte de la DIS. Cela a conduit au développement d'une nouvelle méthode, à savoir la *Document Segmentation with Probabilistic Homogeneity* (DSPH), qui a démontré des performances prometteuses sur des ensembles de données de référence.

Le problème de la DDR a été abordé à travers **une modélisation conjointe de l'homogénéité globale et locale,** où cette dernière a été exploitée à travers divers angles (graph modelling, wavelet

propagation). Ces différents modèles ont été intégrés dans un cadre bayésien pour la DDR. Les résultats de tests effectués sur des échantillons d'images issues du terrain ont démontré les performances encourageantes de la méthode proposée.

Enfin, un système unifié pour l'IQA a été conçu en exploitant l'homogénéité cross-domain entre les images naturelles et les images de documents. Sur la base de réseaux de neurones profonds/ et convolutifs et de l'apprentissage par transfert, les connaissances acquises sur le traitement des images naturelles a été progressivement exploité, d'abord pour l'évaluation des documents et ensuite pour l'évaluation de la qualité unifiée**. Un système unifié basé sur l'apprentissage génératif contradictoire a également été développé**, donnant lieu à des performances encourageantes sur le *blur noise* dans le cadre de deux ensembles de données de référence.

2/ Un deuxième grand point focal était la recherche doctorale menée par dr. Anne Chardonnens sur **les données d'autorité archivistiques dans le contexte des données ouvertes et liées.** Son travail s'est divisé en deux principaux volets : les méthodes d'analyse de la qualité et les méthodes d'amélioration de la qualité des (méta)données.

La première partie a été réalisée à l'aide d'une approche en trois étapes – analyse de la gestion des métadonnées/données d'autorité et de leur publication, analyse des besoins implicites ou explicites de l'institution, analyse des besoins implicites ou explicites des utilisateurs (une méthode semi-automatique a été appliquée à une centaine de milliers de requêtes d'utilisateurs effectuées sur les catalogues en ligne de KBR et du CegeSoma. Les résultats ont été présentés dans la thèse de doctorant de Anne Chardonnens (*La gestion des données d'autorité archivistiques dans le cadre du Web de données*).

La seconde partie de sa recherche doctorale, se concentrant axée sur le développement de nouvelles méthodologies pour améliorer la qualité des (méta)données, visait à explorer le potentiel d'une **gestion semi-centralisée des données d'autorité à l'aide du logiciel libre et *open source* Wikibase**. Les (méta)données du CegeSoma ont été stockées à l'aide d'une instance Wikibase, ce qui a impliqué une série de tâches essentielles parmi lesquelles la normalisation et le traitement d'un ensemble très hétérogènes de données data (stockées soit dans différentes listes nominatives et bases de données ou dans le système de gestion des collections (« Pallas), la mise en place de scripts de *record linkage,* l'alignement des 'entités personnes' avec des ensembles de données externes et des 'entités lieux' avec la liste d'autorité des Archives de l'État en Belgique, le développement d'un modèle de données et son implémentation dans une instance Wikibase, la personnalisation d'un programme pour l'importation massive de données à partir de fichiers CSV, et enfin, la documentation de l'installation et de la configuration, la création de requêtes SPARQL et la maintenance et mise à jour de l'instance Wikibase. Les résultats de cette « expérience de terrain » pionnière ont été analysés dans la thèse de doctorant susmentionnée.

3/ Un troisième et dernier livrable est le **guide de Contrôle Qualité**[1] rédigé par Chloé Brault avec le soutien de Anne Chardonnens. En 2017, Nicolas Roland (KBR) avait déjà réalisé une première étude

[1] https://www.cegesoma.be/fr/publication/guide-numérisation-qualité-cegesoma

de la littérature spécialisée, des normes internationales, des guides de bonnes pratiques et des monographies plus générales sur le sujet de la qualité. Son travail a été approfondi par Chloé Brault, qui a été recruté par le CegeSoma/les Archives de l'État en octobre 2020. Elle a identifié les problèmes de qualité les plus fréquents et a effectué en parallèle un travail de terrain en menant entre autres des entretiens avec les équipes de numérisation de KBR et des Archives de l'État. S'appuyant également sur la richesse des recherches menées par l'équipe du projet (en particulier des docteurs Anne Chardonnens et Tan Lu), elle a publié en septembre le guide des bonnes pratiques ADOCHS, disponible en français, néerlandais et anglais.

Le guide offre une approche de pointe de la qualité, tout en étant concret et visant à proposer des méthodes utilisables par une grande variété d'institutions conservant des collections. Il aborde successivement un bref rappel du contexte de la numérisation et des défis y afférents; un affinement de la notion de qualité en partant de la définition de la norme ISO-9001 et des réalités spécifiques du processus de numérisation; la qualité des images et des métadonnées; l'environnement de la numérisation et la bonne gestion des studios; et enfin, des fiches de synthèse – outils pratiques indispensables pour garantir la qualité d'un projet de numérisation. Le guide a été largement distribué parmi les nombreux réseaux institutionnels des instituts détenteurs de collections concernés.

**Conclusions et recommandations**

Ce projet de recherche fut une opportunité unique d'allier des méthodes de modélisation mathématique, de *machine learning,* de *natural language processing* avec des méthodes plus classiques basées telles que la conduite d'interviews. Il a servi d'impulsion dans les deux établissements scientifiques fédéraux partenaires du projet pour initier des réflexions et expériences sur les données issues de la numérisation des collections. Les tests et analyses réalisées sur la base des études de cas des collections du CegeSoma/Archives de l'État et de KBR permettent d'imaginer de nouvelles façons de répondre aux besoins et attentes des utilisateurs. Ils démontrent également la pertinence et l'importance des techniques de traitement des images et documents dans le contexte de la numérisation du patrimoine culturel, les connaissances et l'expertise humaine pouvant être transposées dans des algorithmes informatiques au service des flux de numérisation et de l'amélioration de l'exploitation des fruits de cette numérisation.

Outre l'enthousiasme suscité par les journées d'étude organisées dans le cadre du projet[2], l'un des plus grands succès de ADOCHS est probablement le fait que la dynamique initiée au cours du projet se poursuit à travers le lancement du projet ***Wikibase Résistance***[3] au sein des Archives de l'État et par le biais du développement d'un ***Data Science Lab***[4] à KBR. Nous espérons que ces projets, menés par plusieurs membres de l'équipe ADOCHS, s'appuieront sur les leçons apprises au cours du projet et sur les questions qui doivent être traitées, telles que les défis des relations avec les services informatiques,

[2] See http://adochs.be/linking/ and http://adochs.be/idpchs/
[3] https://cegesoma.be/fr/project/wikibase-résistance
[4] http://adochs.be/wp-content/uploads/2021/09/AnnDooms_FredericLemmers_KBRDataScienceLab.pdf

l'interrelation entre images et métadonnées, l'automatisation de l'amélioration de la qualité et l'intégration complète des résultats au sein des collections.

Enfin, le projet ayant montré l'importance d'adapter et de développer des méthodes et des modèles qui répondent aux spécificités du patrimoine culturel, nous nourrissons l'espoir qu'il sera l'un des premiers d'une longue série de projets mettant les technologies au service d'un meilleur accès aux collections.

**Mots-clés**
Qualité des données; numérisation; Web sémantique; Données ouvertes et liées; Traitement des images.