SUMMARY

Context

Large scale digitisation has become an important effort in the cultural heritage sector, where a huge amount of original materials such as historical newspapers, manuscripts, maps, etc. were and are continuously being converted to digital formats. With a massive amount of digital collections now available, some universal problems have become apparent these last years. Both human investment and certain technical challenges were clearly underestimated and in many digitisation efforts, quality control remains a challenge. Nevertheless, quality control in a holistic sense is essential to ensure the integrity, the consistency, and the long-term preservation of files and data produced, as well as its public access.

Objectives & methodology

ADOCHS (2016-2021) was a collaboration between the Centre for Historical Research on War and Contemporary Society (CegeSoma, part of the Belgian State Archives), the Royal Library of Belgium (KBR), the *Vrije Universiteit Brussel* (VUB) and the *Université libre de Bruxelles* (ULB). The project aimed to make significant improvements in **quality control for the digitisation of cultural heritage** by developing new approaches in terms of methodology and to create a set of practical guidelines and applicable tools for a step-by-step approach to digitizing projects, with the overall objective to improve the quality of images and metadata produced during heritage and documentary digitisation projects. Three focal points emerged : **improving the quality of images, of the metadata, and of the processes of digitisation.**

1/ Concerning **image quality in the cultural heritage sector**, three image processing subjects, namely DIS (Document Image Segmentation), IQA (Image Quality Assessment) and DDR (Document Damage Recognition), were investigated in this project. These three subjects are all coherent parts of a digitisation- and information extraction workflow and are closely related to each other. These three subjects were approached from a mathematical point of view, by exploiting homogeneity. One research (by dr. Tan Lu) conceptualised and modelled homogeneity in such a way that relations between elements can be characterised and unique patterns can be described. In particular, the research proposed a probabilistic homogeneity model for DIS, where the concept of text homogeneity was defined from the perspective of Gestalt principles. A statistical model exploiting text homogeneity was subsequently formulated for text and non-text classification in the context of DIS. This led to the development of a **new DIS method,** namely **document segmentation with probabilistic homogeneity (DSPH**), which demonstrated promising performance on benchmark datasets.

The problem of DDR was approached with a joint modelling of global and local homogeneity, where the local homogeneity was exploited from different perspectives using **graph modelling and wavelet propagation**. These different models were integrated in a Bayesian framework for DDR. Test results on real-life image samples demonstrated encouraging performance of the proposed method.

Lastly, we ventured towards a unified IQA framework by exploiting cross-domain homogeneity between natural and document images. Based on deep convolutional neural networks (DCNNs) and transfer learning, knowledge acquired on natural image processing was exploited progressively first for document and then unified quality assessment. A **unified framework based on generative adversarial learning** was also developed, where encouraging performance is obtained on blur noise across two benchmark datasets.

2/ A second large focal point was the PhD research by dr. Anne Chardonnens on **archival authority data in a Linked Open Data context.** Her work has been structured around two parts: quality analysis methods and methods for improving the quality of (meta)data. The first part was carried out using a three-step approach – analysis of metadata/authority data management and their publication, analysis of the implicit or explicit needs of the institution and analysis of the implicit or explicit needs of the users (a semi-automated method was applied to a hundred thousand user requests made on KBR and CegeSoma digital catalogues). Scholarly results were presented in the PhD **dissertation** by Anne Chardonnens (*La gestion des données d'autorité archivistiques dans le cadre du Web de données*).

The second part of her PhD research, focusing on the development of new methodologies to improve the quality of (meta)data aimed at exploring the potential of a **semi-centralised management of authority data** using the free and open-source **Wikibase** software. The CegeSoma (meta)data were stored via a Wikibase instance, which entailed a series of essential tasks among which were the standardisation and processing of a set of highly heterogeneous data (stored either in several nominative lists/databases or in the collections management system ("Pallas"), setting up record linkage scripts, a reconciliation of person entities with external datasets and of place entities with the Belgian State Archives authority list, the development of a data model and its implementation in a Wikibase instance, the customisation of a program for the massive data import from csv files, and last but least: the documentation of the installation and configuration, the creation of SPARQL queries and the update and back up of the Wikibase instance. The results of this pioneering 'field experiment' were also analysed in the abovementioned PhD dissertation.

3/ A final and third deliverable is the **Quality Control Guide**[1], authored by Chloé Brault supported by Anne Chardonnens. In 2017, Nicolas Roland (KBR) already carried out an initial study of the specialised literature, international standards, good practice guides and general monographs on the subject of quality. His work was further developed by Chloé Brault, who was recruited by CegeSoma/State Archives in October 2020. She identified the most frequent quality problems and did additional fieldwork, amongst others interviews with the State Archives and KBR digitisation teams. Also drawing from the wealth of research carried out by the project team (and in particular dr. Anne Chardonnens and dr. Tan Lu), she published ADOCH's 'best practice' Guide in September 2021 in French, Dutch and English.

[1] https://www.cegesoma.be/en/publication/digitisation-quality-guide-cegesoma

The quality guide offers a state-of-the-art approach to quality, while also being practical and aimed at concrete methods usable by a wide variety of collection holding institutions. It successively tackles a brief review of the context surrounding digitisation and the associated challenges; a refining of the notion of quality according to the ISO-9001 international standard in line with the specific realities of the digitisation process; the quality of images and associated metadata; the digitisation environment and good studio management; and finally essential 'summary sheets' as a practical instrument to guarantee the quality of a digitisation project. The Guide was widely distributed amongst the many institutional networks of concerned collection holding institutes.

Conclusions and recommendations

This research was an unique opportunity to combine mathematical modelling, machine learning and natural language processing methods with more traditional methods such as interviewing. It served as an impetus in both partner federal scientific institutions, to initiate reflections and experiments relating to data processes from collection digitisation. The tests and analyses carried out on the basis of case studies of CegeSoma/State Archives and KBR collections enable to imagine new ways of meeting the needs and expectations of users. They also demonstrate the relevance and importance of document and image processing techniques in the context of cultural heritage digitisation, where human knowledge and expertise may be imparted in computer algorithms to assist digitisation workflows, and to improve the exploitation of the digitisation products.

In addition to the enthusiasm generated by the study days organised within the framework of the project[2], one of the greatest successes of ADOCHS is probably the fact that the dynamics initiated during the project continues through the launch of the ***Wikibase Resistance* project**[3] within the State Archives and the development of a ***Data Science Lab***[4] at KBR. We are hopeful that these projects, led by several members of the ADOCHS team, will build on the lessons learned during the project and issues that still need to be addressed, such as the challenges of relationships with ICT, the interrelationship between metadata and images, the automation of quality improvement and the full integration of results within collections.

Finally, as the project has shown the importance of adapting and developing methods and models that take respond to the specificities of cultural heritage, we nurture hope that it will be one of the first in a long series of projects using technology to improve access to collections.

Keywords
Data quality; Digitisation; Semantic Web; Linked Open Data; Image Processin

[2] See http://adochs.be/linking/ and http://adochs.be/idpchs/

[3] https://cegesoma.be/en/project/wikibase-resistance

[4] http://adochs.be/wp-content/uploads/2021/09/AnnDooms_FredericLemmers_KBRDataScienceLab.pdf