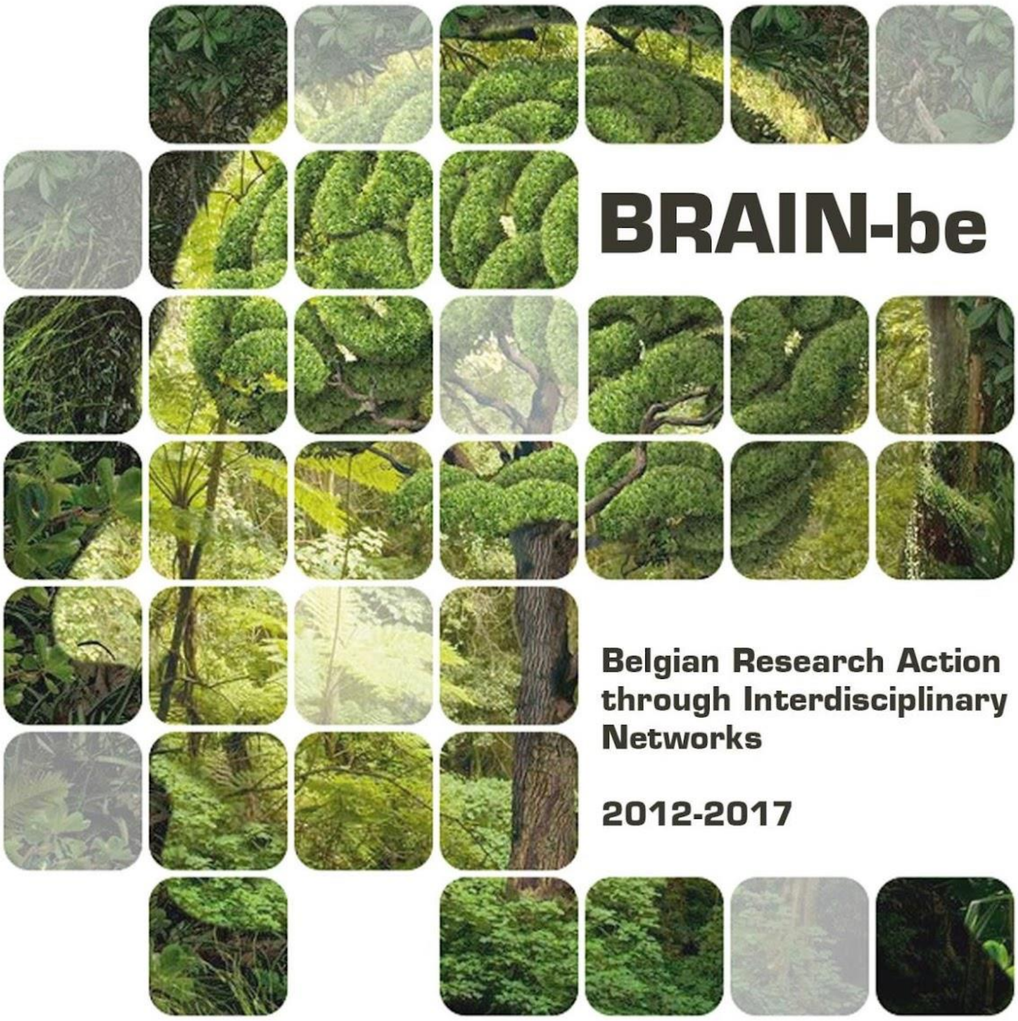


ADOCHS

Auditing Digitalization Outputs in the Cultural Heritage Sector

Chloé Brault (State Archives in Belgium/CegeSoma) – Anne Chardonnens (ULB & State Archives in Belgium/CegeSoma) – Ann Dooms (VUB) – Tan Lu (VUB & KBR) – Frédéric Lemmers (KBR) – Nico Wouters (State Archives in Belgium/CegeSoma) - Florence Gillet (CegeSoma/State Archives) - Seth van Hooland (ULB)

Axis 6: Management of collections

NETWORK PROJECT

[ADOCHS]

[Auditing Digitalization Outputs in the Cultural Heritage Sector]

Contract - BR/154/A6/ADOCHS

FINAL REPORT

PROMOTORS: Florence Gillet/Nico Wouters (State Archives in Belgium/CegeSoma, Square de l’Aviation 29, 1070 Brussels)
Frédéric Lemmers (KBR, Boulevard de l’Empereur 4, 1000 Brussels)
Seth van Hooland (ULB, Av. Franklin Roosevelt 50, 1050 Brussels)
Ann Dooms (VUB, Bd de la Plaine 2, 1050 Brussels)

AUTHORS: Chloé Brault (State Archives in Belgium/CegeSoma)
Anne Chardonnens (ULB & State Archives in Belgium/CegeSoma)
Ann Dooms (VUB)
Tan Lu (VUB & KBR)
Frédéric Lemmers (KBR)
Nico Wouters (State Archives in Belgium/CegeSoma)
Florence Gillet (CegeSoma/State Archives)
Seth van Hooland (ULB)

Published in 2022 by the Belgian Science Policy
Avenue Louise 231
Louizalaan 231
B-1050 Brussels
Belgium
Tel: +32 (0)2 238 34 11 - Fax: +32 (0)2 230 59 12
http://www.belspo.be
http://www.belspo.be/brain-be

Contact person: Helena Calvo Del Castillo
Tel: +32 (0)2 238 36 15

Neither the Belgian Science Policy nor any person acting on behalf of the Belgian Science Policy is responsible for the use which might be made of the following information. The authors are responsible for the content.

Chloé Brault, Anne Chardonnens, Ann Dooms, Tan Lu, Nico Wouters, Florence Gillet, Seth van Hooland (ULB),. ***Auditing Digitalizing Outputs in the Cultural Heritage Sector***. Final Report. Brussels: Belgian Science Policy Office 2021 – 60 p. (BRAIN-be - (Belgian Research Action through Interdisciplinary Networks))

TABLE OF CONTENTS

ABSTRACT 5

1. INTRODUCTION 5

2. STATE OF THE ART AND OBJECTIVES 6

3. METHODOLOGY 14

4. SCIENTIFIC RESULTS AND RECOMMENDATIONS 21

5. DISSEMINATION AND VALORISATION (BIBLIOGRAPHY) 49

6. PUBLICATIONS 51

7. ACKNOWLEDGEMENTS 53

ABSTRACT

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 creating 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 by heritage and documentary digitisation projects. Three focal points emerged: improving the quality of images, the metadata, and the processes of digitisation. These three dimensions have been studied using a multidisciplinary approach combining mathematical modelling, machine learning, and natural language processing methods with more traditional methods such as interviewing.

The project lead to two doctorate dissertation studies: a) by Anne Chardonnens (ULB-CegeSoma) about (meta)data quality with a focus on archival authority data in a Linked Open Data context (title: *La gestion des données d'autorité archivistiques dans le cadre du Web de données*) and b) by Tan Lu (VUB-KBR) about different mathematical models to address image processing problems pertaining to segmentation, damage recognition, and quality assessment (title: *Homogeneity Models for Image Processing in the Cultural Heritage Sector*). A third major deliverable was a Quality Control Guide (in Dutch, French, and English) authored by Chloé Brault, about approaches, methods, and guidelines for a step-by-step development of heritage digitisation projects.

The results, presented at a final online conference on 14 September 2021 attended by over a hundred participants, consist of new ways of meeting the needs and expectations of users by improving the data processes from collection digitisation. They demonstrate the relevance and importance of document and image processing techniques in cultural heritage digitisation, where human knowledge and expertise may be imparted in computer algorithms to assist digitisation workflows and improve the exploitation of digitisation products.

1. INTRODUCTION

1.1 The Digital Era

Since the mid-nineties, cultural institutions have certainly entered the digital age. In Belgium, the federal government adopted a first digitisation plan in 2004 for a ten-year period which has led to the realisation of nine digitisation projects in the federal scientific institutions. These digitisation projects, devoted to massive digitisation of original documents such as newspapers, books, manuscripts, handwritings, drawings, and paintings, prints, historical maps, coins, medals and audio collections, 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 digitisation activities of the past decade with the expertise accumulated during the first stage of digitisation. Belgium is part of a broader European trend where multiple projects were funded to increase access to digitised content in the cultural heritage sector. These projects focused initially on large-scale digitisation, but today also increasingly on technological innovation pertaining to document processing.

1.2 The Need to Develop Methods to Improve Data Quality

In 2016, the *Centre for Historical Research on War and Contemporary Society (CegeSoma*, part of the *Belgian State Archives*) and the *Royal Library of Belgium* (KBR) entered a BRAIN-partnership with the *Vrije Universiteit Brussel* and the *Université libre de Bruxelles,* intending to a research project on the audit of digitisation outputs in the cultural heritage sector. Indeed, the issue of quality control was one of the major obstacles in the first phase of digitisation: it appeared that many projects had underestimated all the implications of this aspect - both in terms of the human investment and the technical challenges – within the overall process of digitisation. In most cases, teams found themselves confronted with a lack of methodological standardisation and automation tools. They often had to work manually, without procedural guidelines adapted to their specific needs. However, quality control is an essential component of every stage of a digitisation project to ensure the integrity, consistency, and long-term preservation of files and data produced, as well as the public access to them. This is true both for outsourced and internal digitisation projects.

The ADOCHS project has been designed to investigate exactly this challenge. ADOCHS aims to make significant improvements in quality control for the digitisation of cultural heritage, develop new approaches in terms of methodology, and create a set of guidelines and tools to follow digitizing projects step by step.

1.3 An Interdisciplinary Approach

The ADOCHS project was launched in November 2016 under the coordination of the CegeSoma/State Archives to work on the improvement of the quality control process concerning digitised heritage collections. One of its main characteristics was the multidisciplinary approach. The project team combined both the expertise of a Digital Mathematics researcher and the skills of an Information Sciences researcher. Whilst the first concentrated its efforts on image quality assessment, the other focussed on metadata quality assessment, intending to identify problems encountered by institutions in order to develop new quality metrics and quality enhancement methodologies. They were supported by another researcher during the first year of the project, who reviewed the state-of-the-art tools and methods for quality control, and a second researcher who combined the project's research results into a practical quality control guide, during the last year of the project.

2. STATE OF THE ART AND OBJECTIVES

2.1 Quality

Quality is defined by the standard ISO 9000 as "*all the features and characteristics of a product, process or service that bear on its ability to meet identified or implied needs*" (International Organization for Standardization, 2015). The same standard defines quality assurance as "*a set of* *activities whose purpose is to prove that an entity meets all quality requirements*". For a digitisation project, two approaches are possible: one focuses on the quality of deliverables and the other on the quality of processes.

The first approach is to ensure the quality of a digitisation project by controlling the deliverables. The guide written by the National Information Standards Organization's (NISO) believes that "*There is a direct correlation between the production of a digital object quality and the ability and flexibility with which this object can be used, reused and migrated through platforms. Thus, the creation of objects at an appropriate level of quality can pay off in the long run because the objects are usable and accessible*" (National Information Standards Organization, 2007).

Quality should be considered in a "*fitness for use*" approach to ensure the adequacy of the results and objectives of the institution. It is, therefore, necessary to define very precisely in the preparatory phase of a digitisation project the ambitions and the desired quality levels. The requirements results will vary according to the needs, the type of documents and the resources available. During or after a digitisation phase, quality control is the evaluation of the quality of images, metadata, and file integrity. Although it represents a significant cost in a digitisation project and often produces a non-tangible product, it is critical if the institution wants to increase its return on investment and avoid extra costs in the future. Depending on the resources available and the desired level of quality, the control will occur more or less frequently in the production chain. Its presence both at the test phase and in the end – during the validation of the work done – is nevertheless indispensable. Manual control can be envisaged for some types of data, handle exceptions, or accompany an automated control. However, the manual approach must be minimised given its high costs. In all cases, it is impractical on a large mass of documents and, therefore, involves work on sampling. The ideal is to combine the two types of controls: manual and automatic.

But ensuring the quality of a digitisation project consists not only of controlling the deliverables a posteriori. It also helps to ensure the implementation of processes to meet the requirements fixed by the project. This step is all the more fundamental as upstream errors can directly lead to quality problems downstream. For example, badly captured images can hamper optical character recognition (OCR). For this reason, Frédéric Baillard suggests in his scanning manual to deploy a quality management system (Claerr and Westeel, 2011). In this context, the implementation of good project management that takes into account the human, technical and procedural aspects is central.

2.2 Image Quality

Indeed, to be able to fully exploit these digital collections, various image processing functionalities are required. For example, to allow users to look for certain keywords in a digital collection, images of the collections (e.g. historical newspapers) need to be parsed such that text regions can be recognised. This process relies on image processing algorithms known as document image segmentation (DIS). In the meanwhile, towards the assessment of the quality of digitisation production, quality assessment (IQA) on different types of images (in particular, natural and document images) shall be investigated. Lastly, to the benefit of both DIS and IQA, recognition of damages in documents (DDR) can be discussed from an image processing perspective. In this project, these three problems, namely DIS, IQA and DDR, were investigated using mathematical modelling as well as machine learning, where various image processing algorithms/systems were developed and evaluated. The results obtained in this project demonstrate the relevance and importance of image processing techniques in the context of cultural heritage digitisation, where human knowledge and expertise may be imparted in computer algorithms to improve digitisation workflows and maximize the exploitation of the digitisation products.

2.2.1 Document Image Segmentation (DIS)

The objective of DIS is to parse, based on only the image of a document, the layout of the document such that different regions (e.g. paragraphs, titles, captions, tables, images, etc.) can be recognised and that relevant textual information can be automatically extracted. DIS is a well-known research topic where different methods and algorithms were published, even prior to the start of ADOCHS. However, DIS remains an open problem where methods that were published mainly operate on rule-based implementation of human knowledge on document structures. These methods are in general classified into three categories: bottom-up, top-down and hybrid.

Bottom-up methods follow an agglomerative approach where letters or words are first extracted and then merged progressively into text lines and paragraphs. These methods (O'Gorman, 1993; Mao and Kanungo, 2001; Kise et al., 1998) rely on the processing of a set of basic units, which are known as connected components (CC), that are extracted from binarised document images. On the other hand, top-down methods (Lee et al., 2001) start from the whole image and make use of global information (e.g. large foreground or background strips) to segment the image iteratively into smaller regions until region homogeneity is reached. Defined as the histogram of the accumulated foreground (or background) pixels in the horizontal or vertical direction across an image region, projection profiles are often employed in this process. Lastly, hybrid methods offer a constructive combination of bottom-up and top-down techniques. These methods (Chen et al., 2013; Tran et al., 2017) normally combine the analysis of foreground (e.g. CCs) and background (e.g. whitespace rectangles) components of an image when extracting different text and non-text regions.

2.2.2 Image Quality Assessment (IQA)

IQA is a special subject in the general field of quality-of-experience (QoE), where quality is interpreted from the perspective of perception (Brunnström et al., 2013). In particular, IQA concerns the objective characterisation (normally through mathematical modelling) of human perception of especially distorted images. Different types (e.g. natural scene, document and screen content) of images have been studied in the field of IQA, where primary focus has been given to natural and document images which are also the two types of images that are mostly seen in the cultural heritage sector.

In general, IQA methods can be divided into three categories, known as full-reference, reduced-reference and no-reference (NR), as shown in Figure 1, where the former two types of IQA methods (Wang and Bovik, 2009; Wang et al., 2004; Rehman and Wang, 2012) require damage-free (so-called pristine) images as a reference in a quality assessment process. Without having to refer to a pristine copy when assessing the quality of a given image, NR methods have quickly become mainstream and various NR models were developed in the meantime. For NR IQA, quality of an image is characterised using objective perceptual scores such as mean opinion score (MOS) or differential MOS (DMOS), which are obtained through subjective experiments (Sheikh et al., 2006).

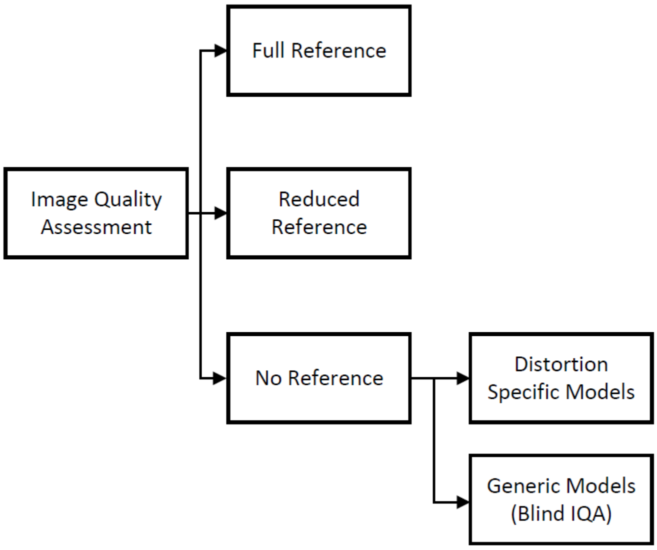


Fig.1 Different types of IQA models.

While some of the NR IQA models (a.k.a. damage-specific models) operate under certain assumptions imposed on the characteristics of the damages contained in images, most of the NR IQA models (a.k.a. general-purpose models) are designed to function in a blind manner where little or no prior information of the damages is available in the process of quality assessment (Cai et al., 2019). Specifically, the task of general-purpose NR IQA was first recognised as one of agglomeration, where different damage-specific subsystems (Moorthy and Bovik, 2010; Moorthy and Bovik, 2011) or various quality-aware feature extractors (Zhang et al., 2015) are stacked to form a general-purpose system. Conventional learning-based structures such as ‘support vector regressors’ (SVRs) are normally incorporated into these systems to enhance generalisation. However, the extraction of effective features that correspond to the human perceptual scores, turns out to be difficult and hinders the further improvement of such systems. On the other hand, the development of convolutional neural network (CNN) based NR IQA models, where the extraction of quality descriptive features is automated through the use of convolutional layers together with back propagation, provides an alternative solution. Therefore, it has received great attention in most recent studies (Kim et al., 2017; Kang et al., 2014a; Li et al., 2016; Sebastian et al., 2018; Talebi et al., 2018; Kim et al., 2019).

Similar to IQA for natural images, much attention has been devoted to documenting image quality assessment (DIQA) where optical character recognition (OCR) accuracy is employed in most of the DIQA models as the quality metric. As the pristine versions of the document images under quality assessment are likely to be unavailable in practical scenarios, NR models are the most relevant for DIQA. The extraction of effective features for OCR accuracy prediction, which is the key component in DIQA, has been approached using different types of image characteristics (Ye and Doermann, 2013). While early works (Stamatopoulos et al., 2011) focused on character-level hand-crafted features (such as font size, stroke thickness, white speckle, broken character, etc.), recently more attention has been given to learning-based feature extraction frameworks, especially to end-to-end models using CNNs (Kang et al., 2014b; Li et al., 2017).

In this project, we focused on deep learning (DL) based IQA models, which have become the main focus in the field of IQA for both natural and document images.

2.2.3 Document Damage Recognition (DDR)

Compared to the conventional types of noise, such as Gaussian, Poisson, speckle, etc., that are mostly addressed in the image 'de-noising' literature, physical damages usually manifest themselves with a remarkably different pattern highly concentrated in local regions, instead of spreading throughout the whole image. This seriously undermines the application of existing mainstream image de-noising techniques, where the underlying noise models are formulated to characterize the fluctuations of pixel values across the whole image. However, the behaviour of physical damages may very well deviate from the assumptions behind these noise removal techniques.

In particular, a large group of de-noising algorithms fall under the paradigm of coefficient shrinkage. These algorithms are developed to exploit the discrepancy between the characteristics of normal image content and the (e.g. additive white) noise in a transformed domain. Specifically, by applying certain transformations (e.g. wavelet, curvelet, shearlets, discrete cosine, etc.), regular image content is compressed into a few large coefficients which coincide with major spatial activities (e.g. edges, corners, etc.) in the image. On the contrary, during the same transformation process, noise gets to spread over all coefficients with relatively smaller magnitudes (Pižurica, 2017). Subsequently, shrinkage estimators can be constructed either to directly derive an estimation of the noise-free coefficients using an optimisation framework or to shrink certain coefficients (normally with small magnitudes) while maintaining the large ones, such that noise is suppressed in the transformed domain and the 'de-noised' image can be obtained by applying the reverse transformation.

Unfortunately, such discrepancy does not hold between normal image content and physical damages, where the damage pixels resemble those of letters. These non-conventional abnormalities have been approached using a miscellaneous collection of target-specific techniques. For example, content obstructing objects in document images may be removed by applying exemplar-based 'inpainting' together with spectral regularisation (Wu and Hou, 2018). Undesired marginal noise (which can be either textual or non-textual content) appearing along the borders of a document may be identified and eliminated based on projection profile analysis of both foreground and background pixels (Shafait and Breuel, 2009). Iterative algorithms (Shah and Gandhi, 2018) were proposed to address shading damages in document images, where an initial estimation of shadow regions is obtained using binarisation techniques. A shading map is then derived by replacing foreground pixels with interpolated local background pixels. Subsequently, the reflectance (foreground) of the image is derived by calculating a ratio between the original image and the shading map. This process can be executed repeatedly until shadow regions are completely removed from the image. Similar to border noise, removal of punched holes in document images is also discussed (Meng et al., 2007), where a pipeline for detecting circular regions corresponding to punched holes is developed based on a combination of resolution reduction, heuristic filtering and Hough transformation. Upon detection, regions corresponding to punching holes in the document image can be inpainted by fitting a bi-linear blending Coons surface which interpolates along the four edges of the noisy regions. In a more general way, a two-phased process for the removal of clutter noise (such as punched holes, ink seeps, ink blobs, etc.) in document images was proposed (Agrawal and Doermann, 2009). The authors first obtain a half-residual image by thresholding a distance map computed from the original image. A set of CC features are then extracted from the half-residual image and are fed into a two-class SVM for clutter classification. Meanwhile warping and perspective damages in document images have also been approached, where a coarse-to-fine strategy was proposed (Stamatopoulos et al., 2011). Words and text lines are first detected in their process to rectify a distorted document image in a coarse scale, before individual words are further refined in finer detail using base-line correction. Striving to deal with multiple document damages, integrated frameworks were also developed (Zhang et al., 2009). The authors first extract the background layer from a document image using total variation (TV) inpainting. By fine-tuning this process, background damages such as shading and bleed-through can be removed. They further apply radial basis function (RBF) based smoothing to extract a smooth shading image, which then helps to reconstruct the surface of the document from perspective and geometric damages.

However, there are still differences between physical damages and these abnormalities. For example, punched holes (or similarly border noise) usually appear at the margins of a page, while torn-offs can be present at arbitrary locations in a document. Meanwhile, the specific assumptions (e.g. circular shape for punched holes or the thickness of clutter noise) cannot be easily generalised to characterize physical damages. It is not uncommon to see that the stroke width (or the size) of a scratch might be similar to or even smaller than that of the large fonts in the same document. Furthermore, compared to warping or perspective damages with global impact on pages, physical damages only affect local regions in a document. Meanwhile, unlike foreign objects (such as pens or cables) which are normally placed on top of a document, physical damages often embed themselves in the text regions of a document. That is, even from the perspective of human perception, the pixels in damages appear to be the same as the ones we find in the textual content in a document image.

In this project, we approach the problem of document damage recognition using mathematical modelling of global and local homogeneity in document images, as will be explained in Section 3.2.

2.3 Metadata Quality

2.3.1 Quality Metrics

The literature review shows that interest in (meta)data quality has grown in the 2000s. While Boydens and van Hooland (2011) have shown that the empirical nature of metadata and the absence of an “absolute’ reference make it difficult to examine their quality, some quality indicators have nevertheless been established. One seminal framework in the digital library context has been published more than 15 years ago by Bruce and Hillmann (2004). Shortly after the emergence of the Open Archives Initiative Protocol for Metadata Harvesting (OAH-PMH), whose implementation has revealed difficulties to aggregate heterogeneous data, they have outlined seven general characteristics of metadata quality: completeness, accuracy, provenance, conformance to expectations, logical consistency and coherence, timeliness, and accessibility. If some variations exist in the different lists of data quality dimensions that have emerged over the following years, a literature review conducted in 2009 by Batini et al. (2009) has shed light on a “*basic set of data quality dimensions, including accuracy, completeness, consistency, and timeliness*”.

Within the cultural heritage sector, the subject has thus been covered extensively over the last decade, leading to a whole series of metrics, data-profiling software, and quality assurance frameworks (see, for example, the extensive literature review and work of Király (2019), leaving little room for new original contributions.

This observation led us to focus on the principle of fitness of use (Boydens, 1999) expressed in the 9001 ISO standard for Quality Management System. The 2005 version of this standard describes quality as “*the totality of features and characteristics of a product, process or service that bears on its ability to satisfy stated or implicit needs”* (ISO, 2005). However, although this metric based on the principle of ‘fitness for use’ is frequently mentioned, to the best of our knowledge no in-depth study had been carried out to confront ‘supply and demand’ in terms of metadata, in the context of cultural heritage. One of the objectives of the ADOCHS project is therefore to develop new methods to identify implicit and explicit needs of categories of users using these metadata.

The other concept mobilised in the framework of our metadata quality analysis has been introduced by Kevin Clair: “*the technical debt as an indicator of metadata library quality”* (Clair, 2016). This metaphor of the technical debt was first used in the 1990s by Ward Cunnigham in software development (Cunningham, 1992): when maintenance activities (such as code rewriting and documentation) are not performed optimally due to lack of time and/or money, this generates a technical debt. This means that the lack of quality will make any modification or extension of the system more difficult and expensive, and, over time, it will require more and more extra efforts, which can be seen as interest paid on the debt. Clair has transposed this metaphor in the library context, where the code is replaced by metadata management: « *the labour, or lack thereof, required to ensure sufficient metadata for a properly functioning system can be thought of as a down payment toward the relief of that technical debt* » (Clair, 2016). He thus distinguished five types of technical debt that can be identified in the context of library metadata management and can arise both intentionally and unintentionally: code debt; design and architectural debt; environmental debt; documentation debt; requirements debt. This typology will be used for the method described in 3.3.1.

2.3.2 Quality Enhancement Methodologies

The development of metadata quality enhancement methodologies has been impacted by the emergence of new conceptual models to describe collections in the cultural heritage sector and the rise of Linked Open Data initiatives.

These conceptual models are developed in parallel in the three types of institutions conserving cultural heritage: museums, libraries and archives. The CIDOC Conceptual Reference Model (ICOM/CIDOC Documentation Standards Group, 2021) is a conceptual model for describing museum objects developed since 2006 by the International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM). It aims to make available all the information necessary to document and manage cultural heritage by providing a common and scalable semantic framework, with which all cultural heritage information can be matched. It is still evolving, currently includes 86 classes and 137 properties and is centred around the notion of events, namely the type of relationship that unites an object and its characteristic - such as its creation date, for example.

Developed in 2017 by the International Federation of Library Associations and Institutions (IFLA), the IFLA LRM provides a new model for libraries (IFLA, 2017). It is focused on the needs of users in the sense that it promotes the structuring of data for searching bibliographic information on the Web to enable users to identify works or places as a priority before locating a resource. To achieve this objective, the model is based on three concepts: the entity, namely the object described, the attribute, which corresponds to the different characteristics of that object, and the relationships. These relationships concern an entity’s relationships with its own attributes and relationships between several entities of one or more systems. In the long term, the aim is to replace bibliographic notices with a network of relationships between entities and thus promote their visibility on the Web and their use by machines.

Records in Contexts (RiC) is a project to overhaul archival description standards initiated in 2016 by the International Council on Archives (ICA, 2016). The objective is to improve and increase the interoperability of the archival descriptions published on the Web by creating a model that brings together existing archival description standards – ISAD(G), ASAAR(CPF), ISDF and ISDIAH.Like the two models aforementioned, RiC also proposes the establishment of a relational system between entities and their characteristics, as well as between entities.

The value of these new metadata models lies in their approach to digital objects, considered not as isolated elements to which a set of descriptive characteristics should simply be assigned, but as a heterogeneous set of objects that only really makes sense through the relationships linking these objects to each other. By bringing together existing standards and relying on a standard such as RDF (Resource Description Framework) (W3C, 2014), these new models are fully in line with the transition to the Semantic Web.

This interconnecting dimension - between entities of the same information system or connections made with external resources - relies on the use of entities unambiguously identified by URI (Uniform Resource Identifier) (Berners-Lee et al., 1998). The crucial role played by the couple [entity and URI] raises questions about the use and the future of authority data and traditional controlled vocabularies. For example, projects such as the Open Memory Project (Brazzo and Massini, 2015), the Social Network Archival Context (Larsson et al., 2014), Linking Lives (Stevenson, 2012), or the War Sampo Project (Hyvönen, 2020), have explored and shown how personal names can be displayed as Linked Open Data, so as to rationalize and improve access to the collections of cultural institutions.

The metadata component of the project, therefore, saw in these recent developments an opportunity to improve the overall quality of (meta)data. Indeed, the maintenance of coherent authority data, accessible in multilingual and machine-readable formats, through unique and permanent identifiers, is pivotal to describing (without ambiguity), sharing or interconnecting the collection items of an institution such as the CegeSoma, which keeps mostly private archives and many personal files. The creation of links to external identifiers appears effectively to be a key step towards a better contextualisation of digitised outputs and a sharing of the workload through 'mutualisation' of the descriptive work. This mutualisation of the workload is facilitated by the existence of "central hubs" such as the Wikidata Knowledge Base (Neubert, 2017), which on the one hand, serves as a pivot between several thousand external identifiers, and on the other hand, makes it possible to document in a collaborative manner all facts known about "notable" people. While these entities could include other types such as institutions, places or subjects, the scope in the context of the ADOCHS will be limited to physical persons for both conceptual and practical reasons.

Furthermore, this transition from data silo practices towards the publication of Linked Open Data raises the question of how to combine the maintenance of authority files for which the institution possesses the expertise, with the reuse of data shared by other institutions, in order to avoid reinventing the wheel. Organisations such as the *Bibliothèque nationale de France* and the ABES in France (ABES, 2019), the *Deutsche Nationalbibliothek* (Ohlig, 2019) in Germany, an the OCLC (Godby et al., 2019) are currently testing the potential of the Wikibase open-source software, which is behind the Wikidata Knowledge Base, and facilitates the use of federated queries between Wikibase instances (Diefenbach et al., 2021).

3. METHODOLOGY

3.1 Overall Methodology

The ADOCHS project was created with the aim of auditing the digitisation outputs in the cultural heritage sector. As explained in the introduction, this question has been studied from two perspectives: a) image quality; b) metadata quality.

In order to achieve the project objectives and deal with this double-pronged approach, two PhD researchers worked part-time in the two participating Federal Scientific Institutions and part-time in the two partner universities. While one researcher was working both at KBR and the VUB, the second was working at the CegeSoma/State Archives and the ULB. These round trips between theory and field have proven to be pivotal for a nuanced and practice-supported approach of the research question. The two researchers were supported by a KBR researcher during the first of the project to collect the state-of-the-art tools and methods for quality control, and a CegeSoma researcher joined the team to synthetise the research results into a procedure to be followed for quality control in a concrete digitisation chain during the last year of the project.

3.2 Image Quality

3.2.1 Document Image Segmentation (DIS)

Despite recent advances, the segmentation of documents with complicated layouts remains an open problem. One particular challenge lies in the recognition of text regions from versatile document structures, where a solution to text and non-text classification is normally formulated heuristically using rigid comprehension of human knowledge of text structure. This is difficult to generalize especially when processing versatile layouts where, for example, illustrations with dimensions comparable to large fonts are placed next to text paragraphs.

Meanwhile, humans can easily recognize text regions from complicated layouts. This capability was studied in this project using the Gestalt theory, which states that human perception is sensitive to homogeneous structures which are understood as a whole rather than the sum of its parts. With this principle, human perceptual text recognition can be attributed to the special visual pattern displayed in text regions. This observation was further exploited by conceptualizing the special Gestalt pattern displayed in text regions as text homogeneity: *"Text homogeneity is the homogeneous pattern displayed in text regions, which consists of proximately and symmetrically arranged units with similar morphological and texture features".*

One important principle of Gestalt perception is the reciprocal dependency between whole and parts, which states that the whole defines parts as much as the other way around. We incorporate this principle and exploit the text homogeneity pattern for the discrimination of individual components. In particular, we observe that the likelihood of a component being text (e.g. a letter) correlates with the formation of the text homogeneity pattern in a local neighbourhood of the component. Hence by characterizing this local text homogeneity pattern, we derive a probabilistic formulation for text and non-text classification. In particular, based on a set of CCs extracted from a document image, we represent a document in a Gestalt domain where each CC is recognised as an individual Gestalt and is characterised using a set of morphological and texture features. We construct a neighbourhood graph in the Gestalt domain by encoding the geometrical relation between Gestalten, where the edges of the graph are weighted based on a probabilistic description of homogeneity between Gestalten. This enables us to characterize the local text homogeneity based on a novel probabilistic local text homogeneity (PLTH) formulation which simulates a random walk-and-check process on the neighbourhood graph. To compute the homogeneity probability between Gestalten, we further propose a cue integration model where the homogeneity probability is evaluated based on a multi-aspect analysis using three different cues corresponding to one set of morphological and two sets of texture features. Meanwhile, under a specific cue, a Bayesian framework is employed for homogeneity characterisation, where similarity between Gestalten is encoded using likelihoods, while proximity and symmetry are encoded using a location prior. The proposed PLTH, therefore, serves as a generic formulation where various primitives, such as geometrical configuration, morphological features, texture characterisation and location priors, are integrated into one computational model. We develop a DIS framework based on the proposed PLTH model, where a text and non-text classification is obtained by thresholding the PLTH map, prior to any grouping process. We tested the resulting DIS framework, which we call document segmentation with probabilistic homogeneity (DSPH), using multiple benchmark datasets. Experimental results demonstrate that DSPH outperforms the state-of-the-art, while ablation and auxiliary tests further demonstrate the potential of the proposed PLTH model.

3.2.2 Unified Image Quality Assessment (UIQA)

In this project, we proposed a unified approach for IQA on natural and document images by exploiting the cross-domain homogeneity between these two types of images. In particular, we first proposed a document IQA (DIQA) model using a transfer learning approach, where the knowledge base of a previously trained deep convolutional neural network (DCNN) is exploited for optical character recognition (OCR) accuracy prediction. Using a two-phase training scheme, the knowledge base, which was trained for natural image processing, is re-tuned for OCR accuracy prediction in the first phase. Subsequently, in the second phase, a task-specific segment mainly consisting of FC layers was incorporated and trained from scratch to facilitate the application of the transferred knowledge base on the new task of DIQA. Experimental results on a benchmark dataset demonstrated that the proposed DIQA model exploiting knowledge on natural image processing performs competitively with the state-of-the-art DIQA models. This suggests that there is indeed exploitable cross-domain homogeneity between natural and document images for a unified IQA approach. Encouraged by this result, we developed a first UIQA model based on a similar transfer learning approach. The performance of this UIQA model was tested extensively using two configurations, where experimental results show that the proposed UIQA model performs competitively to content-specific IQA models simultaneously on natural and document images. Meanwhile, balanced performance was observed within the UIQA model on these two types of images. The performance of the proposed UIQA model was further evaluated using cross configuration tests, where encouraging results were obtained. This demonstrated the stability of the proposed model and encourages further exploration of a unified approach. Nevertheless, the cross-domain homogeneity was exploited inexplicitly in this UIQA model, where the process of learning a common representation of natural and document images is mingled with that of regressing the common representation towards the respective quality scores of these two types of images. To address this problem, we proposed an innovative contractive generative adversarial network (C-GAN) formulation, which operates to contract different distributions towards each other. This enables a generalisation of image samples from heterogeneous sources in a latent domain, where the generalised samples have similar characteristics and distributions, as if they were drawn from a same source. We developed a second UIQA framework based on the proposed C-GAN model, where the C-GAN was first applied to learn explicitly a common representation of natural and document images in a latent domain. The learned common representation was then regressed towards respective quality scores, where the regressor operates as if it is processing a same type of images. In this second UIQA model, the process of learning a common representation was clearly separated from the process of regression.

This allowed direct access to the generalised samples and enabled targeted investigation on either the generalisation or regression. The proposed UIQA model was tested on blur damage across natural and document images, where promising results were obtained. This encourages future work on a unified approach for IQA of images from heterogeneous sources.

3.2.3 Document Damage Recognition (DDR)

To pursue an automatic recognition of irregular structures (i.e. damages) in documents, we exploit the same text homogeneity pattern which was introduced in Section 3.2.1. In particular, the proposed PLTH model was employed as a global homogeneity measure for DDR. This is essentially a Gestalt oriented formulation where a group of pixels (i.e. a CC) is processed together as one single unit. However, a pixel-wise approach was desired for accurate recognition of damage pixels, especially from adjacent text pixels. Hence two local homogeneity measures were proposed based on Markov random field (MRF) and wavelet approximation propagation respectively. In particular, MRF offers a convenient way to address a global labelling problem based on the characterisation of local interactions. We proposed an MRF modelling of image patches containing mixed damage and text pixels. By considering first and second-order cliques on the MRF graph, grayscale similarity and pair-wise smoothness were encoded using unary and binary potentials respectively. This allowed us to separate damage and text pixels within the image patch based on a maximum a posterior (MAP) MRF inference, which was solved through energy minimisation using a graph cut algorithm. To derive the parameterisation of the unary potential, we proposed a peak-searching algorithm based on a random-walk simulation. The effectiveness of this MRF based local homogeneity measure was demonstrated based on experimental results obtained on image patches. However, energy minimisation on large-scale graph models is computationally intensive. This limited the application of this MAP-MRF model on entire images.

We further proposed a second local homogeneity measure, which was derived based on a multi-resolution analysis (MRA) characterisation of the neighbourhood transition of individual pixels. Thanks to text homogeneity, the local neighbourhood of a text pixel exhibits similarity across different scales. Hence, when a set of photos is taken while zooming out on a text pixel, the neighbourhood transition appears to be smooth and quickly stabilizes. However, the pattern in the local neighbourhoods of a damage pixel changes drastically from one patch to another. This leads to a more volatile neighbourhood transition which stabilizes only when a large neighbourhood is included. This difference in terms of neighbourhood transition can be effectively modelled using the propagation of approximation coefficients of a stationary wavelet transform (SWT), which we formulated as propagation of wavelet approximation (PWA) and propagation of cone-of-influence wavelet approximation (PCWA). Experimental and simulation results demonstrated the effectiveness of both PWA and PCWA for the characterisation of damage pixels, where the former was chosen as a significance measure given its comparable performance to the latter and its computation efficiency. We then proposed a generic DR method based on a Bayesian framework, where global and local homogeneity were modelled together using a joint likelihood formulation, while the MRF modelling was incorporated as a prior. We solved the MAP inference based on Markov chain Monte Carlo

(MCMC) sampling, where a Metropolis sampler was deployed to obtain an estimation of the optimal set of labels corresponding to the recognition of damage pixels. We tested the proposed method on a set of real-life document image samples containing physical damages of different size and shapes. Both quantitative and qualitative evaluation demonstrated the encouraging potential of the proposed DDR method.

3.3 Metadata Quality

3.3.1 Quality Metric

To extend metadata quality analysis beyond metrics offered by data-profiling tools and to elaborate a new metric based on the notion of fitness for use (see 2.3.1), a new method has been developed. This method aims to analyse cultural heritage institutions' metadata quality in light of the implicit and explicit needs of categories of users. Our scope being limited to authority data (see 2.3.2), we based our approach on the analysis conducted by the IFLA working group on Functional Requirements and Numbering of Authority Records (FRANAR, 2010) which identifies as users both "*authority data creators who create and maintain authority data",* and *"users who use authority information either through direct access to authority data or indirectly through the controlled access points (...) in catalogues, national bibliographies, other similar databases, etc*." (IFLA/FRANAR, 2009). Our method, based on the study of empirical data through case studies (Flyvbjerg, 2006), consists of three steps:

1. Analysis of metadata/authority data management and publication
2. Analysis of implicit or explicit needs of the institution
3. Analysis of implicit or explicit needs of users (in general)

The first step consists of a case study based on a typology developed by Clair (2016), which analyses the metadata quality through five types: code debt; design and architectural debt; environmental debt; documentation debt; requirements debt (see 2.3.1).

The second step is to identify and analyse the needs of the institution (metadata/authority data creators). Since these needs can be expressed in a variety of ways, several types of resources can be reviewed (for instance working documents and more formal publications such as annual reports).

The third step is to collect/identify and analyse user needs. Several sources can be used: information and findings from sources available internally or published in recent years, as well as user queries made in the digital catalogue of the institution. Although we did not conduct specific interviews due to lack of time, we were able to rely on raw data and reports from MADDLAIN, a BRAIN project dedicated to the study of user needs, in which KBR and CegeSoma were involved (Hungenaert and Gillet, 2017).

The semi-automated method developed for user queries analysis has been applied to more than 80 000 queries from the BelgicaPress online interface (KBR interface to access digitised newspapers) and described in detail in Chardonnens et al. (2018), while the whole method – including the three aforementioned steps – has been applied to CegeSoma and presented in a comprehensive manner in a PhD dissertation (Chardonnens, 2020).

3.3.2 Quality Enhancement

The development of new methodologies to improve the quality of metadata aims to explore the potential of semi-centralised management of authority data using the Wikibase software (see 2.3.2). A Wikibase instance has been installed to store the CegeSoma (meta)data – more precisely the person entities – and study the possibilities and limits of this tool:

1. To integrate and edit (manually using the graphical interface or massively using dedicated tools) authority data and other nominative databases about persons who are creators or 'subjects' of the CegeSoma collections (mainly related to WWII)
2. To publish it as Linked Data & link it to other (external) repositories through Uniform Resource Identifiers
3. To integrate new datasets and offer new crowdsourcing possibilities
4. To offer new perspectives on data through data mining (by formulating complex queries and using visualisation tools)
5. To exploit versioning to monitor/enhance data quality
6. To combine “in-house” data with external data through federated SPARQL queries

Before going into details, it has to be noted that the relevance of running a Wikibase instance rather than importing data directly into Wikidata has been taken into consideration. According to our knowledge and the Wikidata notability criteria (Wikidata, 2020), the best scenario seems to start by using Wikibase and, at a later stage, favour a joint use of both. Indeed, due to the difficulty of removing the ambiguity of several people's names, the presence of many "unknowns" (whose only "notable fact" was to be interviewed about daily life during the occupation, for example), the need for the institution to keep control of the data and the possibility to model it according to its needs, it seems more appropriate to opt for Wikibase.

Conducting this experiment involved various tasks (which are described in detail in a doctoral thesis and its appendices, see Chardonnens, 2020) such as:

- Creation and configuration of a Wikibase instance hosted on a State Archives of Belgium web server (including customisation and installation of extensions and various scripts intended to improve the user experience)
- Cleaning, standardisation and processing of CegeSoma nominative data, which were characterised by the that they were messy, more or less complete, sometimes ambiguous,

and stored either in several nominative lists/databases or in a deprecated collections management system ("Pallas")

- Setting up record linkage scripts (based on The Python Record Linkage Toolkit) intended to link and 'de-duplicate' records related to (supposed) same individuals coming from the same or different datasets
- Reconciliation of person entities with external datasets (Wikidata, Viaf, SNAC)
- Reconciliation of place entities (written natural language) with the State Archives authority list, Wikidata and GeoNames
- Development of a data model inspired by Wikidata and taking into account the constraints linked to Wikibase, in collaboration with the scientific staff of CegeSoma
- Implementation of the properties and basic elements of this data model in a Wikibase instance
- Customisation of a program (based on Wikibase-Edit) allowing the import massive amount of new data from CSV files
- Documentation of the installation and configuration
- Creation of SPARQL queries
- Update and back up of the Wikibase instance.

3.4 Digitisation guide

The Digitisation & Quality guide is the result of two years of work divided between two researchers. In 2017, Nicolas Roland (KBR) carried out an initial study of the specialised literature, international standards, good practice guides and general monographs on the subject of quality. The guide is thus based on :

- Standardised reference frameworks on quality and digitisation, and more specifically on ISO standards.
- International methodological guidelines and good practice guides for digitisation.
- General monographs on digitisation, quality and its management, digital imaging and metadata.
- Research conducted during the ADOCHS project by Tan Lu and Anne Chardonnens on improving image quality and metadata.

In parallel, Nicolas Roland carried out an analysis of the digital content produced by the KBR and the CegeSoma, respectively from their newspaper collections and their photographic archives. These case studies aimed, on the one hand, to support the work of the doctoral students of the ADOCHS project, and on the other hand, to identify precisely the errors and technical limitations encountered by the digitisation teams of the two institutions. By identifying the most frequent quality problems, it was then possible to develop a coherent methodology upstream of the project and to better deal with unforeseen problems during the digitisation process. After a second study of the specialised literature, Chloé Brault continued this fieldwork in 2020, conducting a series of interviews with the AGR and KBR digitisation teams.

In order to reflect both theoretical and practical approaches, the guide is divided into six chapters. The first chapter provides a brief review of the context surrounding digitisation and the associated challenges. It discusses the most common objectives of digitisation and how they are changing in the light of the increasing computerisation of our modern societies. These changes result in institutional changes that are themselves based on the changing expectations and needs of users of archives and libraries. The second chapter defines the notion of quality according to the ISO-9001 international standard. This general definition is then adapted and refined in line with the specific realities of the digitisation process. The strategic and intellectual development of any project is addressed here through three fundamental documents: the data management strategy, the digitisation policy and lastly the specifications.

Chapters three and four cover the quality of deliverables, and, more specifically, images and associated metadata. These two notions are in turn explained and characterised in order to identify the criteria that can be used to define them as good quality. The fifth chapter deals with the digitisation environment and good studio management. Heritage digitisation has the particular characteristic that it deals with precious objects – whether modern or antique – which require special measures to avoid damage during handling. Lastly, the sixth and final chapter provides a series of summary sheets that can be used to guarantee the quality of a digitisation project. These sheets summarise the key concepts for each quality control step and follow the chronological sequence of the digitisation chain. Each sheet also provides practical recommendations, tools and tips along with additional information sources to enable the completion of the control phase described.

4. SCIENTIFIC RESULTS AND RECOMMENDATIONS

4.1 General Conclusion

While the following paragraphs highlight concrete results in terms of images, metadata and digitisation process, it should first be noted that overall, ADOCHS 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 enabled institutions to imagine new ways of meeting the needs and expectations of users. This required contact with digitisation staff, collection management staff, scientific staff, and IT departments, generating cross-cutting reflections in the institutions, whether on the contributions of artificial intelligence, quality control methods, or the evolution of models and standards for describing collections. In addition to these internal dynamics

– which were made concrete in particular through the launch of the *Wikibase Resistance* project within the State Archives (see 4.3.3) and the development of a Data Science Lab at KBR (see 4.2.4) – the project has also made it possible to establish contacts at an international level with other players in the cultural sector and the academic world working on similar issues, for example in the context of the study days organised by the ADOCHS team in 2019 and 2021.

4.2 Image Quality

4.2.1 Document Image Segmentation (DIS)

Human perceptual recognition of text regions from complicated layouts, although having been linked to Gestalt visioning for years, had yet to be fully explored in the process of document segmentation. In this project, we exploit text homogeneity in the context of text and non-text classification for document segmentation. In particular, we conceptualize text homogeneity as the Gestalt pattern displayed in text regions, which consists of proximately and symmetrically arranged units with similar morphological and texture features. Inspired by the Gestalt two-sided dependency between whole and parts, we recognize that the likelihood of a component being a text component correlates with the formation of the text homogeneity pattern in its local neighbourhood. Hence by characterizing this local text homogeneity, we derive a probabilistic formulation for an accurate text and non-text classification prior to the extraction of higher-level text and non-text structures.

To characterize local text homogeneity, we represent a document image in a Gestalt domain based on a connected component (CC) analysis, where each CC is recognised as a Gestalt and is characterised using a set of morphological and texture features. By encoding the geometrical relation between Gestalten, a neighbourhood graph is constructed where the edges of the graph are weighted using a probabilistic description of the homogeneity between Gestalten. Based on the weighted graph, we propose a novel PLTH formulation, which characterizes the local text homogeneity pattern by simulating a random walk-and-check on the neighbourhood graph.

To derive the weights of the edges of the neighbourhood graph, we propose a cue integration model for the evaluation of the homogeneity probability between Gestalten. This allows us to characterize the homogeneity between Gestalten from three different aspects using morphological as well as texture features of the Gestalten. Meanwhile, under a specific cue, the homogeneity probability is evaluated based on a Bayesian framework, where similarity between Gestalten is encoded using likelihoods, while proximity and symmetry between Gestalten are encoded using a location prior.

The proposed PLTH serves as a generic formulation, where various primitives, such as geometrical configuration, morphological features, texture characterisation and location priors, are integrated in one computational model. This allows a reliable text and non-text classification preceding any grouping process, which is currently missing in DIS systems.

We developed a DIS system based on the PLTH model, where text and non-text Gestalten are classified by thresholding the PLTH map. Higher-level text and non-text structures are then extracted from text and non-text Gestalten respectively. The framework of the developed software, which is referred to as Document Segmentation with Probabilistic Homogeneity (DSPH), is demonstrated in Figure 2 below.

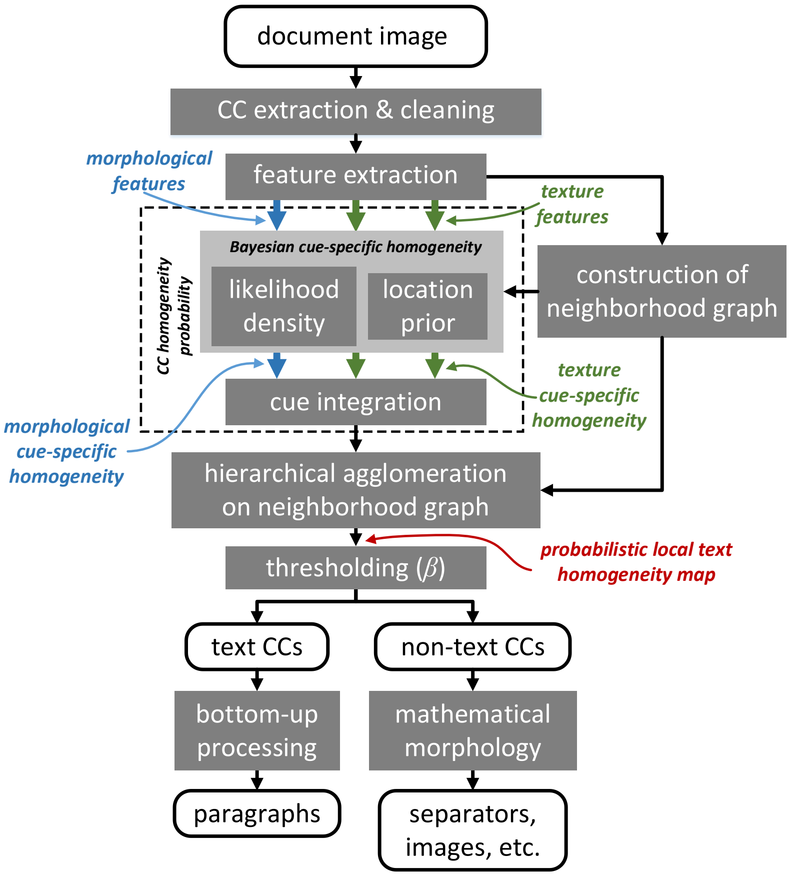


Fig.2 The framework of the Document Segmentation with Probabilistic Homogeneity (DSPH) framework that has been developed based on the proposed probabilistic local text homogeneity (PLTH) for parsing the content of document images.

To evaluate performance, DSPH has been tested, using different evaluation scenarios and metrics, on various datasets consisting of a wide range of documents with complex layouts. The results show that our method performs consistently across all datasets, and improves upon the state-of-the-art. Furthermore, ablation tests show that the incorporation of the text and non-text classification formulated based on the proposed PLTH model indeed improves the segmentation performance upon state-of-the-art. A visualisation of the PLTH map on documents with complex layouts is provided in Figure 3 where it can be seen that the proposed PLTH model operates effectively in distinguishing text and non-text content in documents with complex layouts (e.g. contains flexible text structures and rich non-text content such as images, graphics, etc.).

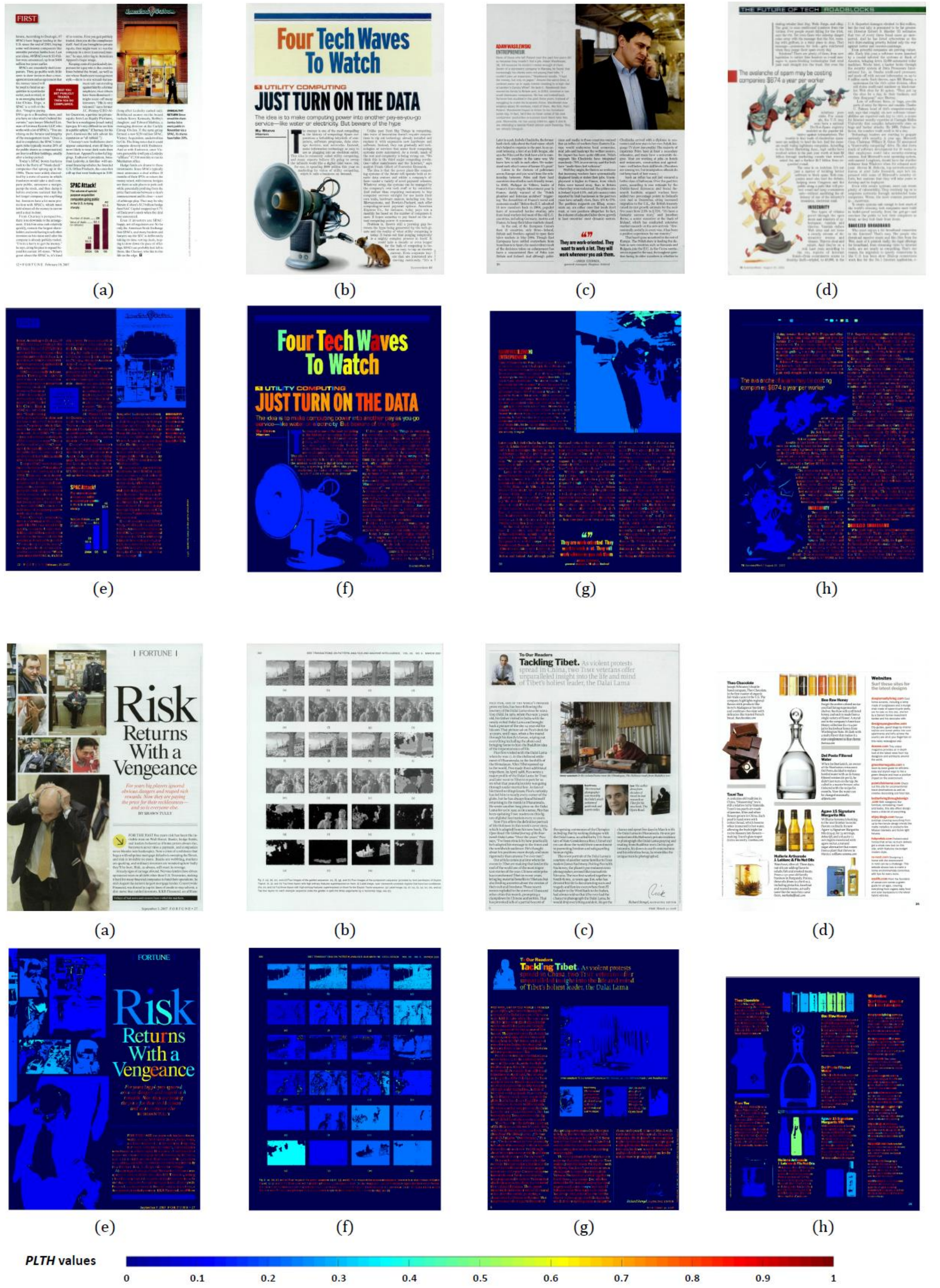
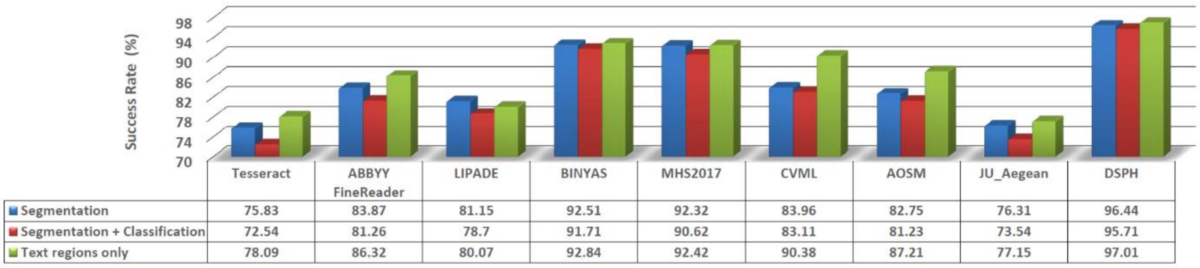


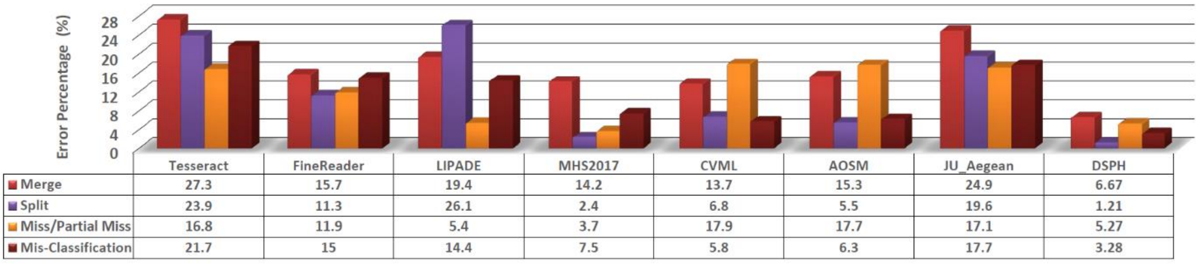
Fig.3 A visualisation of the operation of the proposed PLTH model. (a)-(d), documents with complex layouts; (e)-(h), PLTH maps generated by the proposed PLTH model on documents. It can be seen that PLTH gives more attention (in terms of probability) to text regions, therefore allowing a more reliable recognition of text components and enabling an accurate parsing of the content of documents with complex layouts.

Meanwhile, auxiliary experiments show that the proposed method is able to maintain a stable performance over a relatively wide range of parameter configurations. The quantitative evaluation results of DSPH are presented in Figure 4, while a visualisation of the output of DSPH on documents with different layouts is depicted in Figure 5.



	Tesseract	ABBYY FineReader	LIPADE	BINYAS	MHS2017	CVML	AOSM	JU_Aegean	DSPH
🟦 Segmentation	75.83	83.87	81.15	92.51	92.32	83.96	82.75	76.31	96.44
🟥 Segmentation + Classification	72.54	81.26	78.7	91.71	90.62	83.11	81.23	73.54	95.71
🟩 Text regions only	78.09	86.32	80.07	92.84	92.42	90.38	87.21	77.15	97.01

(a) Accuracy of DSPH when evaluated using scenario-driven document layout assessment metrics namely ‘Segmentation’, ‘Segmentation + Classification’ and ‘Text regions only’.



	Tesseract	FineReader	LIPADE	MHS2017	CVML	AOSM	JU_Aegean	DSPH
🟥 Merge	27.3	15.7	19.4	14.2	13.7	15.3	24.9	6.67
🟪 Split	23.9	11.3	26.1	2.4	6.8	5.5	19.6	1.21
🟧 Miss/Partial Miss	16.8	11.9	5.4	3.7	17.9	17.7	17.1	5.27
🟫 Mis-Classification	21.7	15	14.4	7.5	5.8	6.3	17.7	3.28

(b) Breakdown of segmentation errors under the ‘Segmentation + Classification’ scenario.

Fig.4 Evaluation of DSPH on a public bench-mark dataset (Recognition of Documents with Complex Layouts 2017, RDCL 2017) and comparison between DSPH and other DIS algorithms and software solutions. A more detailed analysis of the performance of DSPH can be found in Lu and Dooms (2021b).

In the latest RDCL2019 competition, DSPH has been declared the winner (Clausner et al., 2019). A more detailed demonstration of DSPH can be found in Lu and Dooms (2021b). In the meanwhile, the development of DSPH has led to the filing of an international patent application (Lu and Dooms, 2021a).

Automatic recognition of document content is of fundamental importance to digitisation workflows, where improved performance of DIS algorithms not only is beneficial for obtaining optimal OCR output but also is critical for enabling downstream applications such as keyword searching or content indexing. While there exist different methods and software solutions, some of the most fundamental challenges (e.g. discriminating text and non-text content in documents with complex layouts) remain unsolved and the performance of existing methods and solutions tend to suffer when incorporated in digitisation workflows where complex documents are processed. In this project we took the initiative to address some of the fundamental challenges and to develop a new DIS framework with original models and algorithms. Promising results were obtained concerning the performance of the proposed DIS method. In the meanwhile, we note that the performance of DSPH can still be improved especially when processing historical document images where additional challenges such as noisy binarisation and local and global skew are often seen. It is of practical interest to continue the development of DSPH to address such challenges, and to valorise relevant algorithms and systems.

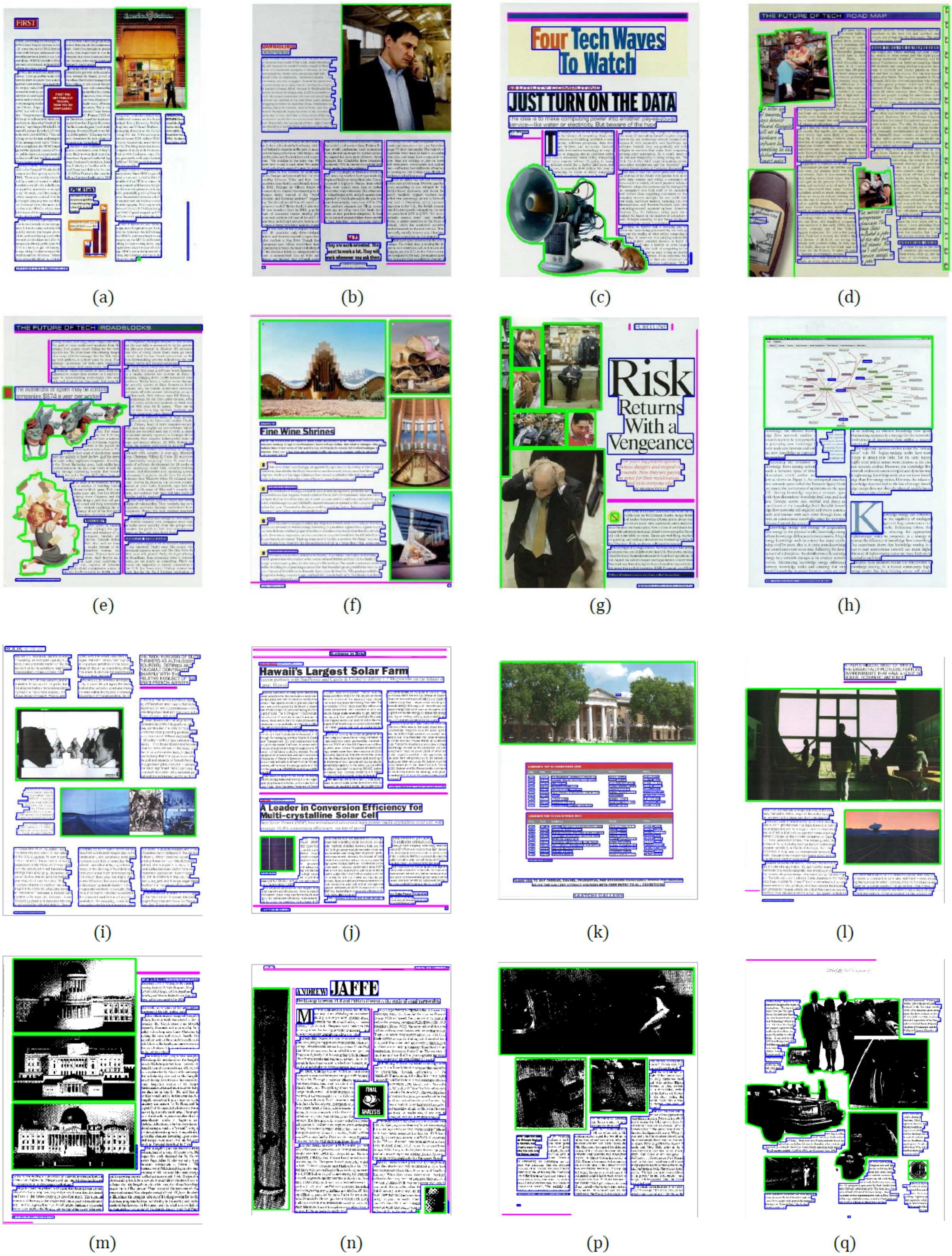

(a) (b) (c) (d) (e) (f) (g) (h) (i) (j) (k) (l) (m) (n) (p) (q)

Fig.5 Visualisation of the outputs of DSPH on documents with different style of layouts. Blue, green, blue-violet, orange, magenta represent text, image/graphic, table, chart and separator regions respectively. It can be seen that DSPH operates to derive an accurate parsing of the content of documents with complex layouts.

4.2.2 Image Quality Assessment (IQA)

Natural scene and document images are processed differently when it comes to quality assessment. This is because existing IQA models are content-specific, where different models are developed for different types of images. Unfortunately, there are potential limitations pertaining to such a divided approach. One immediate example involves the scenario where a set of mixed natural and document images are presented to an IQA model. Given that the IQA model is developed to process one specific type (e.g. natural) of images, its output on the other type (e.g. document) of images is not properly defined. In such cases, a unified approach, where different types of images can be processed simultaneously, is an advantage.

In this project, we investigated unified IQA models by exploiting the cross-domain homogeneity between natural and document images. Specifically, we first proposed a DIQA model using a transfer learning approach, where the knowledge base of a previously trained DCNN was exploited. The proposed DIQA model was trained in two phases. In the first phase, the knowledge base of an existing DCNN model, which has been trained on natural image processing, was fine-tuned for OCR accuracy prediction. Subsequently, in the second stage, a task-specific segment consisting mainly of FC layers was introduced and was trained to facilitate the application of the transferred knowledge base on the new task of DIQA. Experimental results showed that the proposed DIQA model exploiting knowledge on natural image processing performs competitively to state-of-the-art DIQA methods. We subsequently proposed a first UIQA model based on a similar transfer learning approach, where the original knowledge base of AlexNet was exploited to assess the quality of natural and document images simultaneously. The framework of the proposed UIQA system is presented in Figure 6.

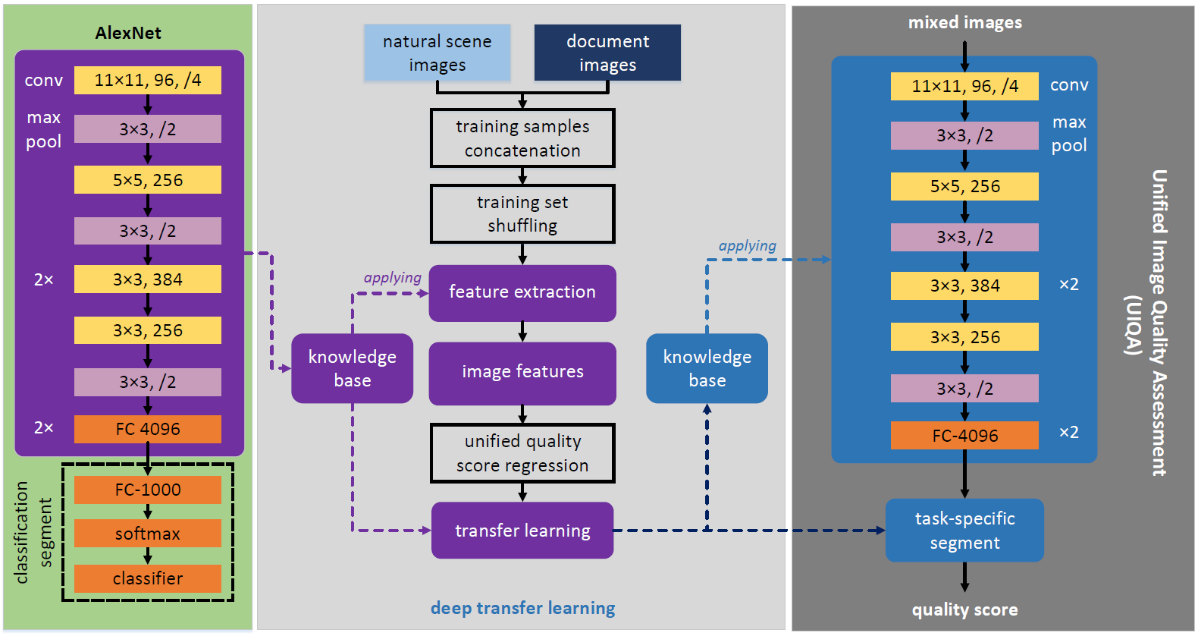
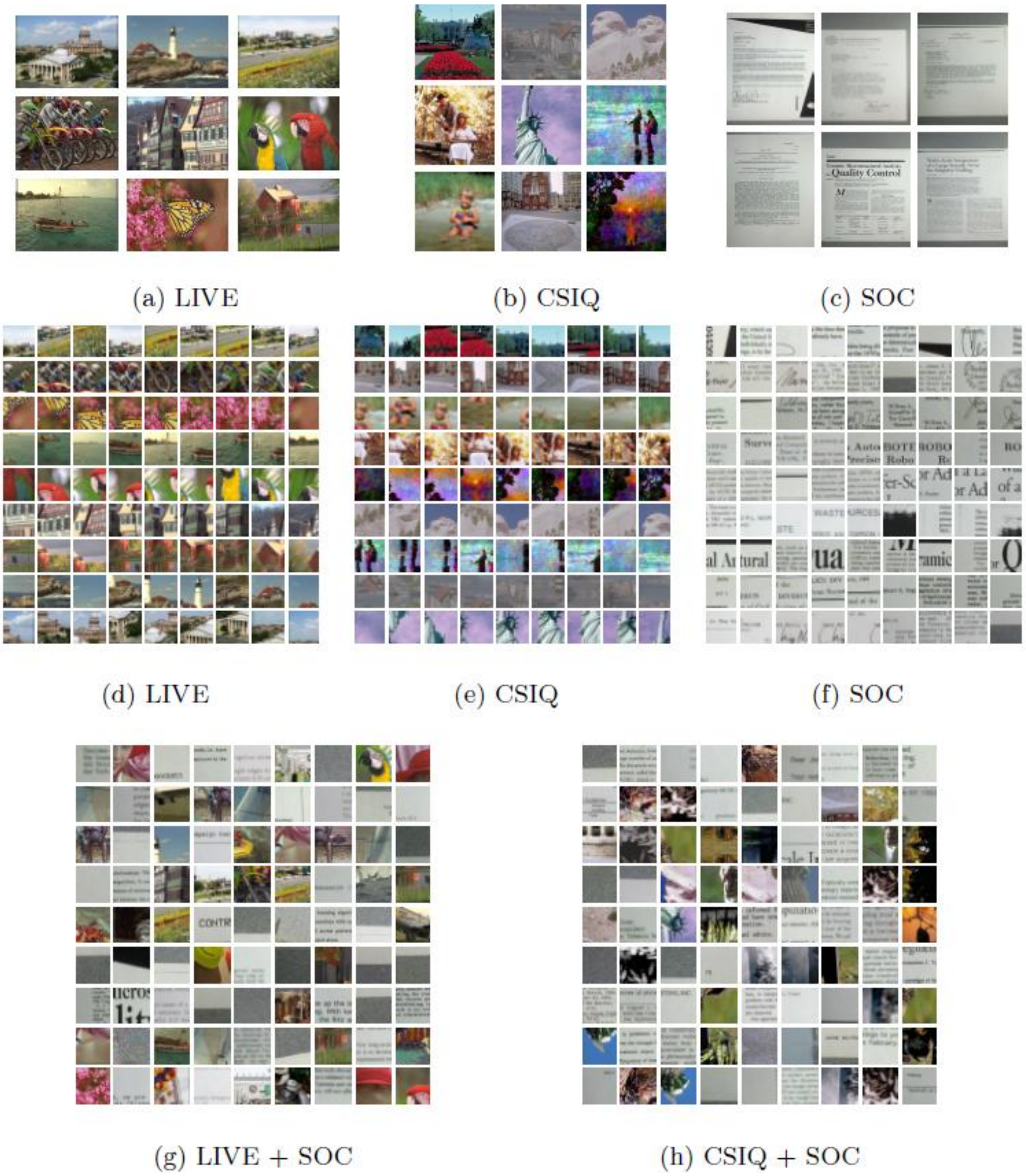


Fig.6 Structure of the proposed UIQA model using transfer learning. To train the proposed UIQA model, samples of both natural and document images are concatenated and shuffled to form the training space. The knowledge kernel of AlexNet

is exploited for simultaneously predicting the quality of natural and document images. In this process, natural images are regressed towards perceptual scores while document images are regressed towards OCR error rates.

A visualisation of the training space is depicted in Figure 7. We conducted extensive experiments to investigate the performance of the proposed UIQA model. Several benchmark IQA and DIQA datasets, namely the dataset from the Laboratory for Image and Video Engineering (LIVE), the Categorical Subjective Image Quality (CSIQ) dataset and the Sharpness-OCR-Correlation (SOC) dataset, were used for performance evaluation. Encouraging results were obtained as shown in Table I and Table II. It can be seen that the proposed UIQA model achieved good performance comparable to state-of-the-art simultaneously on natural and document images. Meanwhile, a balanced performance was observed on these two types of images. This further demonstrates the potential of generalizing IQA models across natural and document images.



(a) LIVE

(b) CSIQ

(c) SOC

(d) LIVE

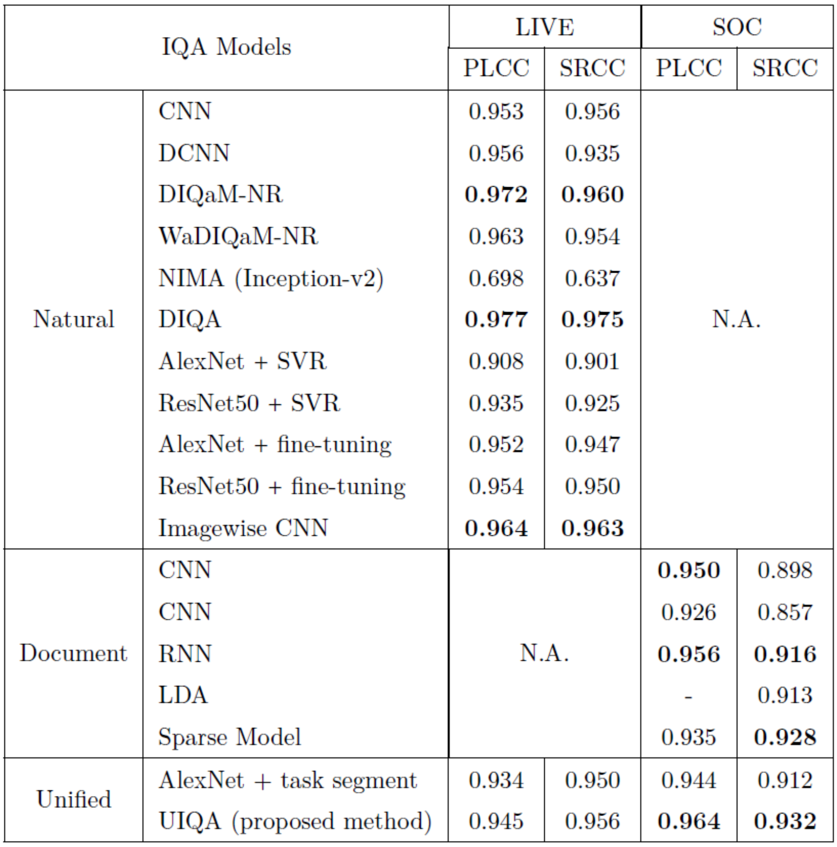
(e) CSIQ

(f) SOC

(g) LIVE + SOC

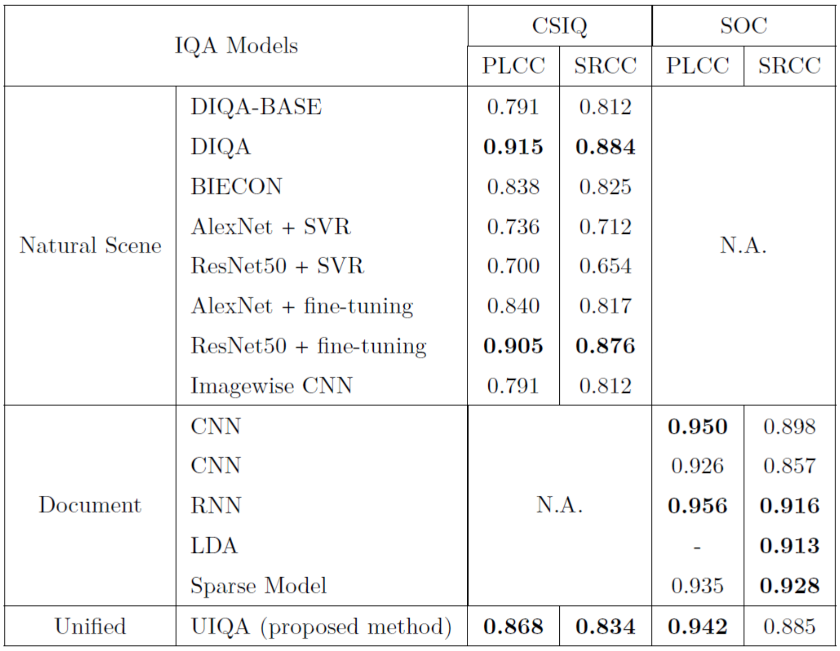
(h) CSIQ + SOC

Fig. 7 Visualisation of the training space of the UIQA model: (a) - (b) sample images from IQA datasets; (c) sample images from a DIQA dataset; (d) - (f) patches extracted from sample images; (g) - (h) sections of the training sets under two different experimental configurations.



IQA Models		LIVE		SOC	
		PLCC	SRCC	PLCC	SRCC
Natural	CNN	0.953	0.956	N.A.	
	DCNN	0.956	0.935		
	DIQaM-NR	**0.972**	**0.960**		
	WaDIQaM-NR	0.963	0.954		
	NIMA (Inception-v2)	0.698	0.637		
	DIQA	**0.977**	**0.975**		
	AlexNet + SVR	0.908	0.901		
	ResNet50 + SVR	0.935	0.925		
	AlexNet + fine-tuning	0.952	0.947		
	ResNet50 + fine-tuning	0.954	0.950		
	Imagewise CNN	**0.964**	**0.963**		
Document	CNN	N.A.		**0.950**	0.898
	CNN			0.926	0.857
	RNN			**0.956**	**0.916**
	LDA			-	0.913
	Sparse Model			0.935	**0.928**
Unified	AlexNet + task segment	0.934	0.950	0.944	0.912
	UIQA (proposed method)	0.945	0.956	**0.964**	**0.932**

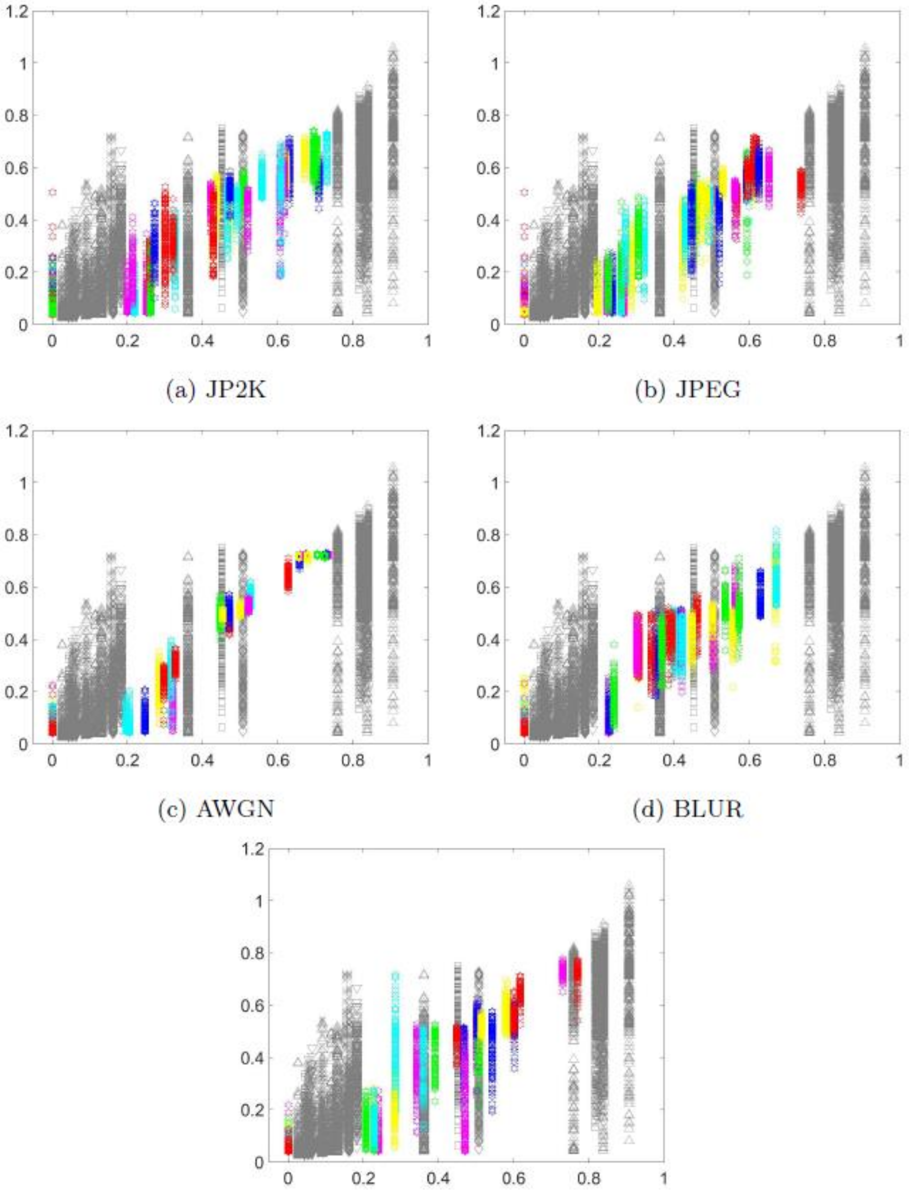
Table I Performance of the proposed UIQA model under LIVE + SOC configuration. The performance is measured based on correlation (namely Pearson linear correlation coefficient, PLCC, and Spearman's Rank Order Correlation Coefficient, SROCC) between the prediction of the model and the ground truth. The performance of the UIQA model is compared with state-of-the-art IQA and DIQA models.



IQA Models		CSIQ		SOC	
		PLCC	SRCC	PLCC	SRCC
Natural Scene	DIQA-BASE	0.791	0.812	N.A.	
	DIQA	**0.915**	**0.884**		
	BIECON	0.838	0.825		
	AlexNet + SVR	0.736	0.712		
	ResNet50 + SVR	0.700	0.654		
	AlexNet + fine-tuning	0.840	0.817		
	ResNet50 + fine-tuning	**0.905**	**0.876**		
	Imagewise CNN	0.791	0.812		
Document	CNN	N.A.		**0.950**	0.898
	CNN			0.926	0.857
	RNN			**0.956**	**0.916**
	LDA			-	**0.913**
	Sparse Model			0.935	**0.928**
Unified	UIQA (proposed method)	**0.868**	**0.834**	**0.942**	0.885

Table II Performance of the proposed UIQA model under CSIQ + SOC configuration. The performance of the UIQA model is compared with state-of-the-art IQA and DIQA models.

Furthermore, we conducted in-depth analysis of the performance of the proposed UIQA model across different types of distortions, where relevant results are presented in Figures 8 and 9.



(a) JP2K

(b) JPEG

(c) AWGN

(d) BLUR

(e) FF

Fig. 8 Prediction versus ground truth of the UIQA model simultaneously on natural scene and document images. Different colour represents different natural scene images from the LIVE dataset, while different marker shape in grey colour represents different document images from the SOC dataset. The analysis is conducted for different types of distortions that are contained in LIVE, namely JP2K compression distortion, JPEG compression distortion, added white Gaussian noise (AWGN), blur and fast-fading Rayleigh channel distortion (FF).

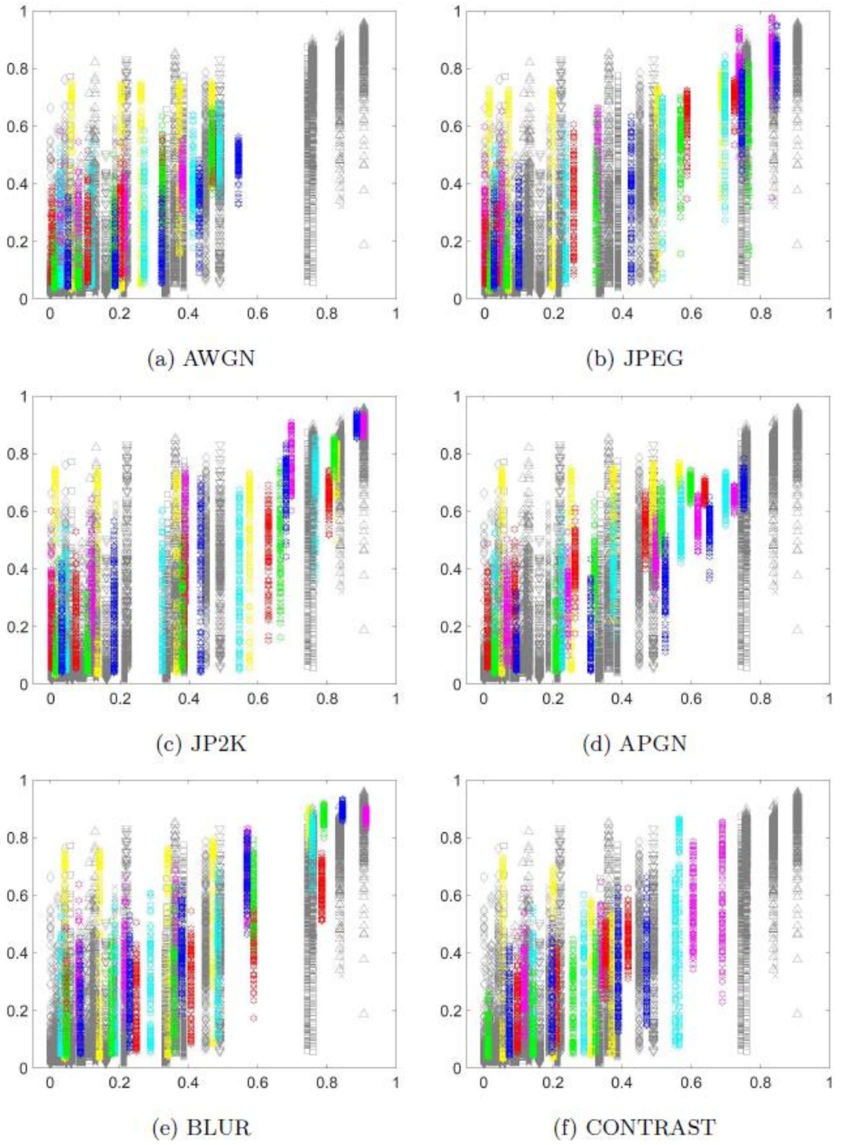
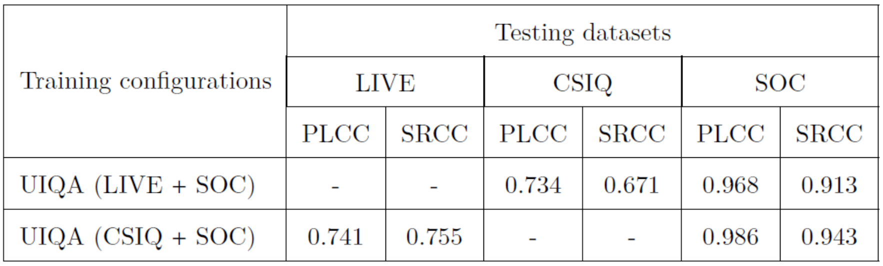


Fig. 9 Prediction versus ground truth of the UIQA model simultaneously on natural scene and document images. Different colour represents different natural scene images from the CSIQ dataset, different marker shape in grey colour represents different document images from the SOC dataset. The analysis is conducted for different types of distortions that are contained in the natural and document images, namely added white Gaussian noise (AWGN), JPEG compression distortion, JP2K compression distortion, additive pink Gaussian noise (APGN), blur and global contrast decrements (CONTRAST).

To further evaluate the stability of the proposed UIQA model, cross-datasets evaluations were also conducted, where the UIQA model trained under one configuration (e.g. LIVE + SOC) is tested using a different configuration (e.g. CSIQ + SOC). The results, as presented in Table III, demonstrate that the proposed UIQA model is able to generalize and maintains a relatively stable performance when trained and tested in a cross dataset manner.

However, in this first UIQA model, the cross-domain homogeneity between natural and document images was exploited inexplicitly, where the process of learning a common representation of these two types of images is mingled with the process of regressing the common representation onto quality scores. It is therefore difficult to interpret and investigate the exploitation of the cross-domain homogeneity. To address this problem, we proposed a novel C-GAN formulation, where a contractive process is modelled such that the distributions of input signals are pulled towards each other. Based on the C-GAN model, a common representation of natural and document images can be learned in a latent domain, where the generalised samples appear to have similar characteristics and distributions as if they were drawn from the same distribution. We propose a second UIQA model where C-GAN is first applied to generalize natural and document images in the latent domain. A regressor is then installed to map the generalised samples to their respective quality scores. Hence the process of exploiting the cross-domain homogeneity and that of regression are well separated. Test results on blur damages demonstrated the effectiveness of the UIQA model. For a more detailed explanation on this work, we refer to Lu and Dooms (2019a, 2019b, 2020b). In the meanwhile, to the best of our knowledge, this is the first time that a unified IQA approach generalizing across natural and document images was being proposed and evaluated. Given the results that have been obtained in the ADOCHS project, we consider UIQA as an attractive direction not only for future research work, but also for perspective implementation in relevant digitisation workflows.



Training configurations	Testing datasets					
	LIVE		CSIQ		SOC	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
UIQA (LIVE + SOC)	-	-	0.734	0.671	0.968	0.913
UIQA (CSIQ + SOC)	0.741	0.755	-	-	0.986	0.943

Table III Generalisation performance of the UIQA model measured across the two different experimental configurations.

4.2.3 Document Damage Recognition (DDR)

Physical damages such as torn-offs and scratches are commonly seen in historical documents. Their presence sabotages the operation of image processing frameworks such as DIS and OCR. This not only results in a reduced amount of information that can be automatically retrieved from a document but also deteriorates the performance of other document processing algorithms that rely on layout analysis or content recognition. A visual demonstration of physical damages in historical newspapers is depicted in Figure 10.

Hence automatic recognition of physical damages in document images is desirable. On one hand, based on DDR, document images can be inpainted in a pre-processing step. This ensures a smooth operation of relevant algorithms such as DIS and OCR and therefore facilitates a better automatic information extraction from document images. On the other hand, DDR can be applied to advise the DIQA process such that document images having similar global noise levels can be further discriminated based on the presence of physical damages.

Despite its necessity, little attention has been paid to DDR. This may be attributed to the fact that physical damages have different characteristics as compared to other types of noises (e.g. the additive Gaussian noise) which have been well addressed in the literature. Furthermore, physical damages may have arbitrary size and shape and may appear at any random locations in a document.

This renders difficulties for rule or exemplar-based algorithms, which are normally applied when dealing with abnormalities in documents. In fact, to the best of our knowledge, no method was available for recognizing general physical damages from document images.

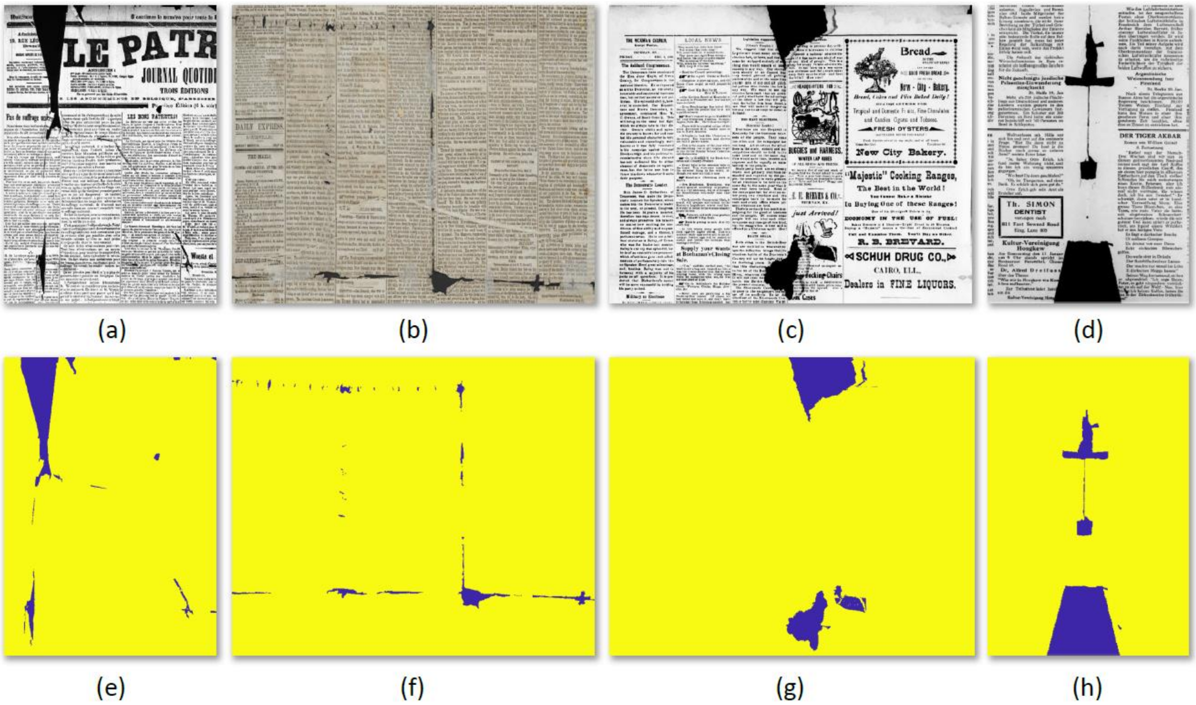


Fig. 10 Physical distortions in historical documents: (a)-(d), historical documents with physical distortions; (e)-(h), ground-truth of distorted document images. In general, physical distortions have varying size, and are randomly located. This makes it difficult to recognize their presence using heuristic approaches.

In this project, we consider a pixel-wise approach for developing a generic DDR method, where we exploit text homogeneity using two different models. Specifically, we first conceptualize the previously developed PLTH model as a global homogeneity measure that can be applied to reveal irregular structures in documents using a CC based manner, as demonstrated in Figure 11.

To derive an accurate recognition of the damaged pixels, we further propose a second local homogeneity measure by exploiting the text homogeneity pattern around individual pixels. In particular, when zooming out on a pixel from a text region, a smooth neighbourhood transition is observed, where the change of the neighbourhood pattern quickly stabilizes. On the other hand, the neighbourhood transition of a damage pixel appears to be volatile with more dramatic change from one patch to another and only stabilizes when a larger (compared to that of a text pixel) neighbourhood is included. This difference between damage and text in terms of neighbourhood transition can be effectively modelled using the propagation of the approximation coefficients of a stationary wavelet transform (SWT) across different scales, which we formulate as propagation of wavelet approximation coefficients (PWA) and propagation of cone of wavelet approximation coefficients (PCWA). Experimental and simulation results demonstrate that both PWA and PCWA can be applied for distinguishing pixels from regular and damaged regions of a document. A visualisation of PWA as a local homogeneity measure for damaged pixels recognition is depicted in Figure 12.

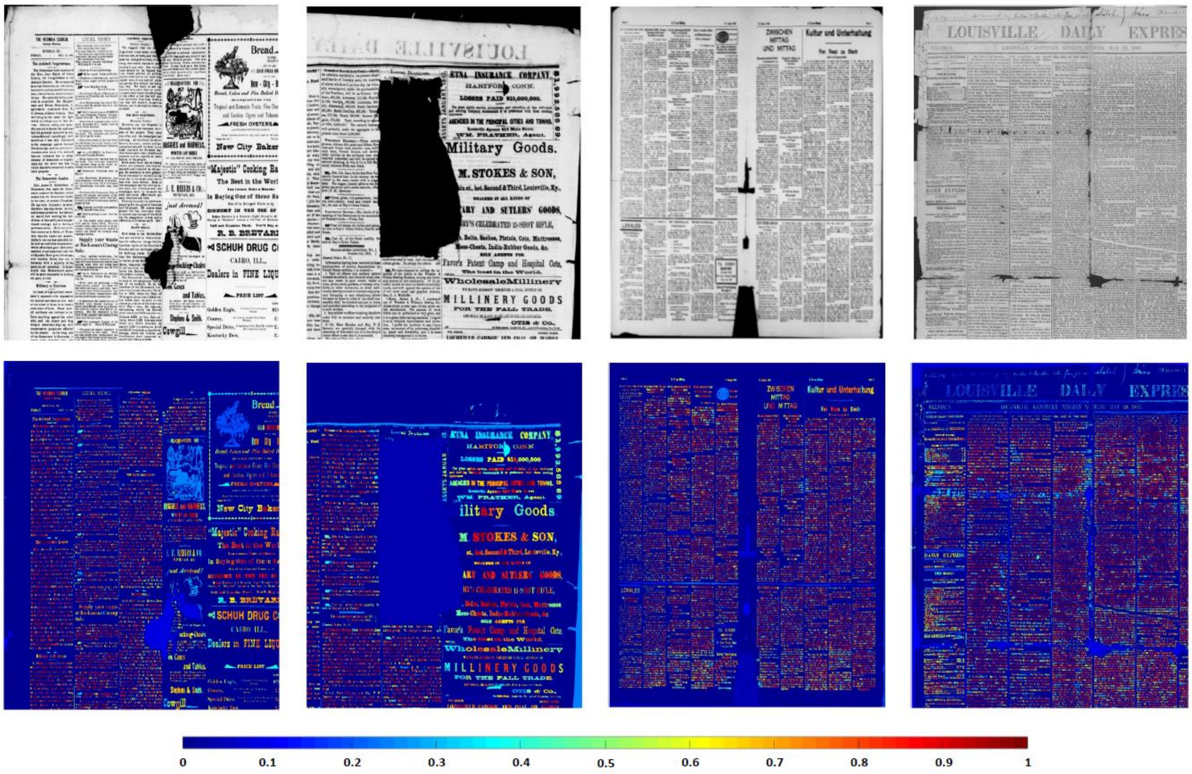


Fig. 11 Using the previously developed PLTH model as a global homogeneity measure for recognizing damages in document images. In this case, damaged regions (in terms of CCs) are flagged by low values on the PLTH map computed from document images.

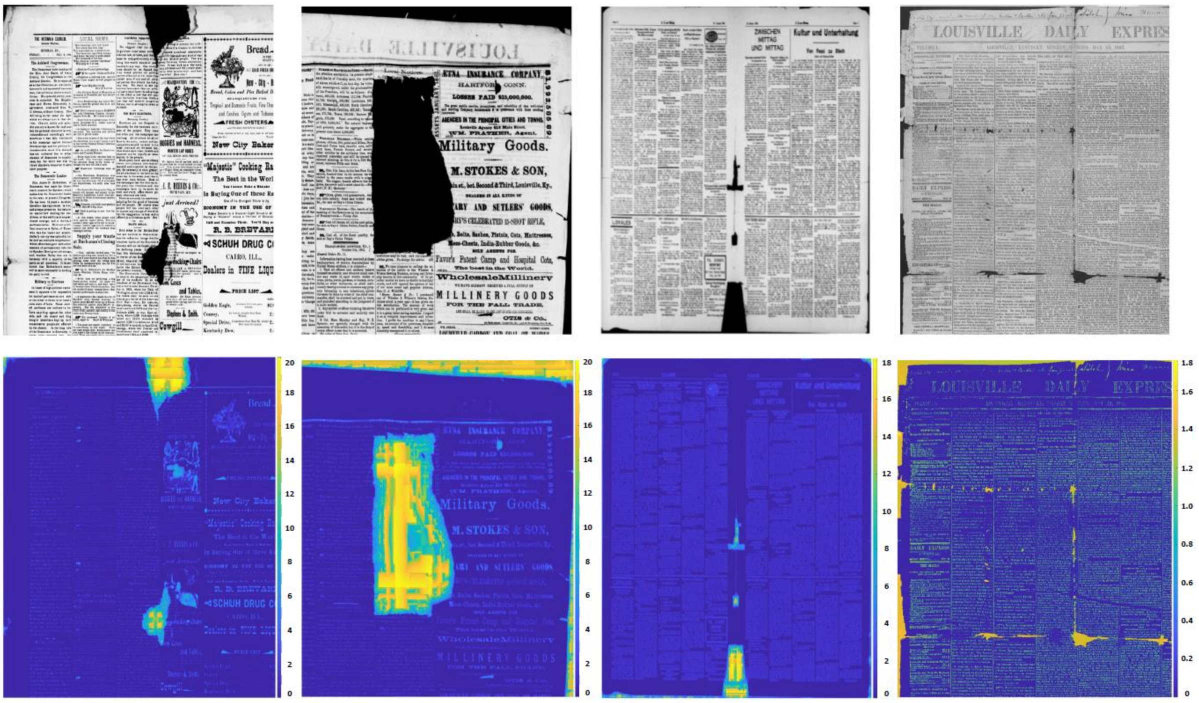


Fig. 12 Proposed PWA as a local homogeneity measure for recognizing damaged pixels from documents. The intensity of the pixels represent the value of PWA. It can be seen that damaged pixels in general yield higher PWA values as compared to pixels from regular content. This enables a pixel-based recognition of physical damages in document images.

Based on these different homogeneity measures, we propose a generic Bayesian DDR method where global and local homogeneity are simultaneously modelled using a joint likelihood density. As depicted in Figure 13, the joint distribution of the global and local homogeneity measures represents an effective classification of damaged and regular pixels.

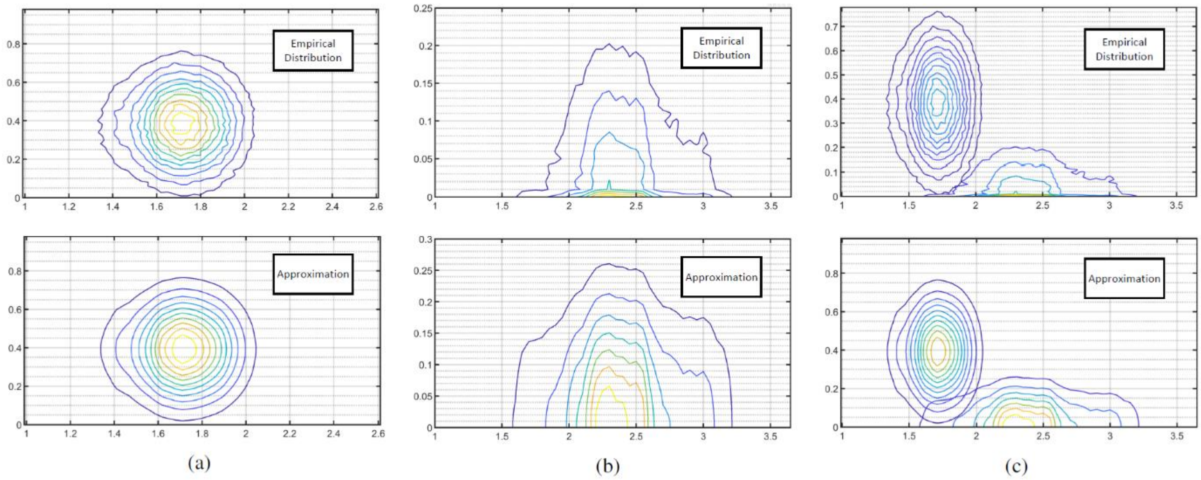


Fig. 13 Empirical evaluation of the joint global and local homogeneity likelihoods for text and damage pixels. Global and local homogeneity measures are plotted along the y and x axes respectively. (a) Upper: empirical joint global and local distribution computed from text pixels; lower: approximation of joint distribution on text pixels. (b) Upper: empirical joint global and local distribution computed from damage pixels; lower: approximation of joint distribution on damage pixels. (c) Comparison of the joint likelihoods for text and damage pixels. The global homogeneity values are re-sampled from a Gaussian distribution which was fit to the original empirical values. It can be seen that a good separation of text and damage pixels can be obtained using the proposed joint global and local homogeneity model.

Meanwhile, we formulate damage recognition as a Bayesian maximum a posterior (MAP) problem where the likelihood is calculated using the joint global and local homogeneity model while the prior is computed using a Markov Random Field (MRF) model. We solve this Bayesian MAP inference using sampling, where a Metropolis sampler is employed to derive an approximation to the optimal set of labels corresponding to damage recognition.

The proposed framework is evaluated on a set of real-life document samples containing damages of varying shapes and sizes. The performance of the algorithm is evaluated using both F-measures and intersection-of-union (IoU), where results are demonstrated in Table IV. Meanwhile, a visual demonstration of The operation results of the proposed DR framework is depicted in Figures 14 and 15. For a more detailed explanation on the relevant work and results, we refer to Lu and Dooms (2021c, 2018, 2020a).

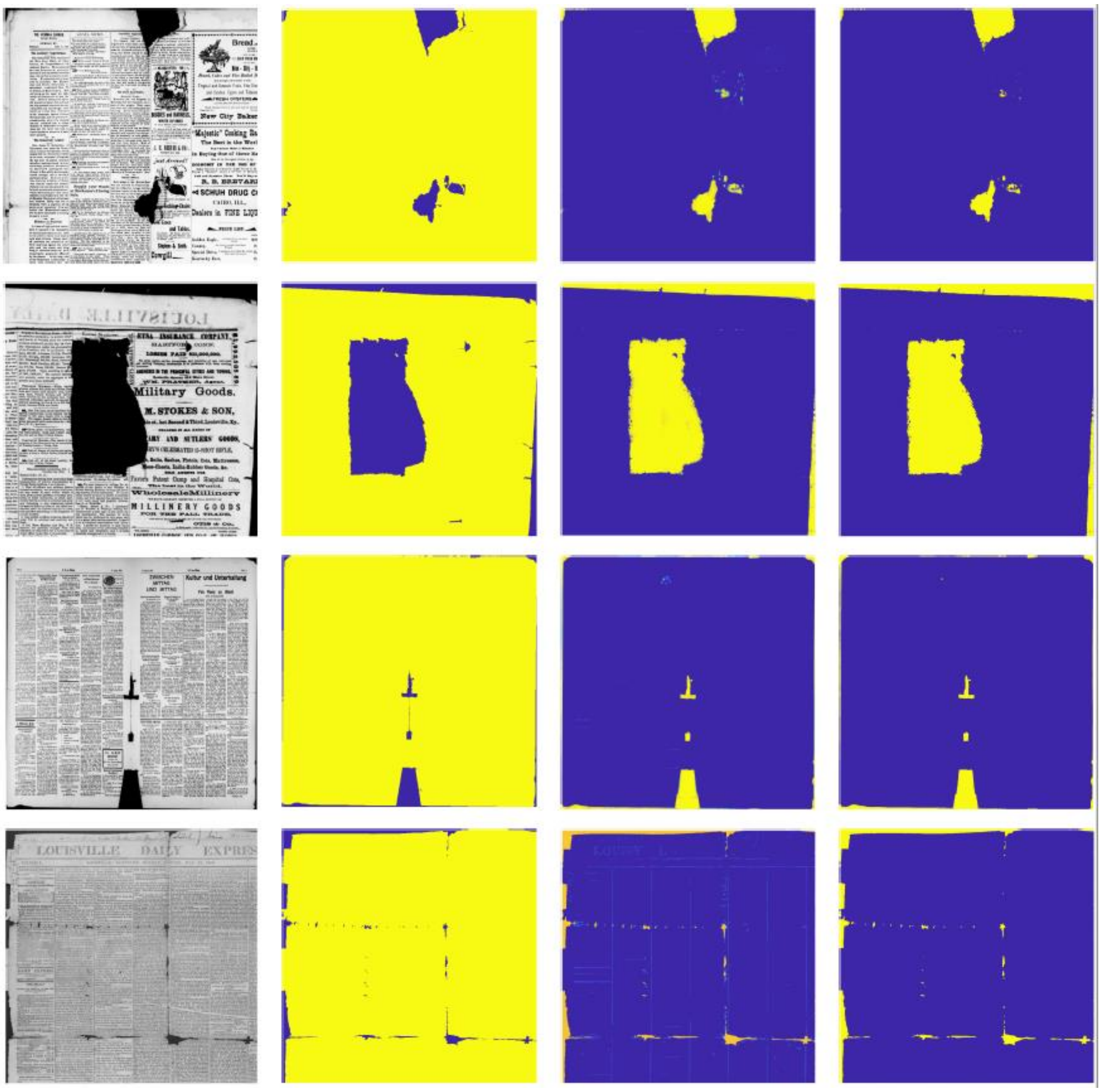


Fig. 14 A visual demonstration on the operation of the proposed damage recognition framework. First column, historical document images (corresponding to samples 1 - 4 in Table IV) with various physical damages; second column, ground truth of the document images; third column, heatmaps of the posterior probability obtained upon convergence of a sampling process; last row, outputs of DDR by thresholding the probability maps.

Samples	Precision		Recall	F-Score	IoU	
	Damage Pixels	Text Pixels			Damage Pixels	Text Pixels
1	0.9731	0.9948	0.8438	0.9038	0.8245	0.9941
2	0.9993	0.9863	0.9413	0.9694	0.9407	0.9861
3	0.9843	0.9965	0.8654	0.9210	0.8536	0.9961
4	0.9411	0.9992	0.9727	0.9566	0.9169	0.9976
5	0.9295	0.9997	0.8628	0.8949	0.8098	0.9995
6	0.9132	0.9960	0.8885	0.9007	0.8193	0.9930
7	0.9838	0.9950	0.9690	0.9763	0.9538	0.9924
8	0.9360	0.9989	0.9429	0.9394	0.8858	0.9976

Table IV Quantitative results of the evaluation of the proposed damage recognition framework on real-life documents.

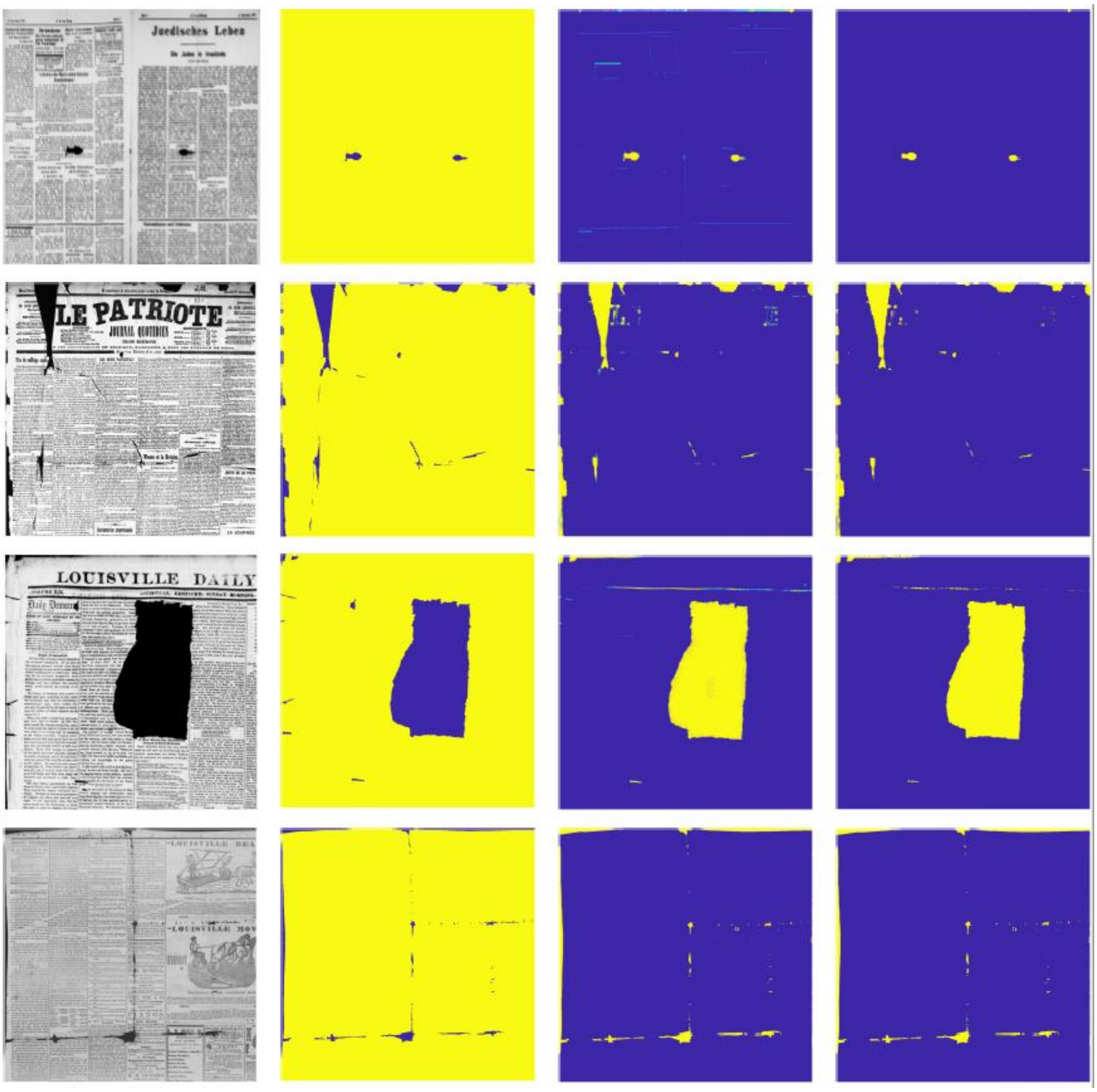


Fig. 15 A visual demonstration on the operation of the proposed damage recognition framework. First column, historical document images (corresponding to samples 1 - 4 in Table IV) with various physical damages; second column, ground truth of the document images; third column, heatmaps of the posterior probability obtained upon convergence of a sampling process; last row, outputs of DR by thresholding the probability maps.

4.2.4 Future Work

In this project, three topics that are pertaining to image quality in the cultural heritage sector, namely document image segmentation (DIS), image quality assessment (IQA) and damage recognition (DR), have been studied. The relevant results obtained, as well as the experiences accumulated in this project, have already led to new developments and initiatives. In particular, the proposed DIS solution DSPH is currently being used as a baseline for the development of new document parsing software. This development is specifically supported by Geert Laurent Braeckman who joined the ADOCHS team in the last year to work on the optimisation and valorization of DSPH.

In 2021, a VUB IOF project named *CAPTCHA 2.0 - Completely Automated Processing of scanned Text documents by teaching Computers how Humans Analyze them* was funded to promote the valorisation of the continuous development of DSPH. Meanwhile, the performance of DSPH has attracted attention from the industry, where a joint research project (based on DSPH) between the research group DIMA and an industry partner was funded by VLAIO in 2021. In 2022, a new project proposal aiming at developing DSPH into an advanced intelligent document parsing solution was submitted to Innoviris.

In the meanwhile, based on the results derived in ADOCHS, a long term collaboration between VUB and KBR has been established. This long term collaboration is established as a new research lab, named the KBR Data Science Lab (https://www.kbr.be/en/projects/data-science-lab/), which is successfully funded by the Belspo FED-tWIN program. The KBR Data Science Lab serves as a research and development hub where the primary objective is to bring together inspiration, expertise and resources for data intelligence in the cultural heritage sector. The Data Science Lab has two main goals: 1) Facilitating both fundamental and applied research in areas such as mathematical modeling, image and natural language processing; 2) Promoting the implementation of the relevant research outputs in digitization workflows. The results obtained in the ADOCHS will be used as the foundation of the future development in the KBR data science lab. For example, the DSPH framework lays the ground for future work on improving OCR performance on KBR collections. Another example is that the UIQA framework developed in the ADOCHS project will be used as a baseline for future work on automating digitization workflows. Thus, the exploration of the topics such as DIS, IQA and DDR will not end upon the completion of the ADOCHS project: they will be continuously explored in a more systematic way, and on a larger scale, in the context of the KBR data science lab.

4.3 Metadata Quality

4.3.1 Quality Metrics

The main contributions in terms of metadata quality metrics are both practical and theoretical.

Firstly, a method to analyse metadata through the prism of the “fitness for use” criteria has been developed and illustrated through a case study (Chardonnens, 2020). This method (to analyse metadata through the “fitness for use” critera) is based on particular on an analysis of user queries (see 3.3.1) which has been presented in detail in Chardonnens (2018). This paper presents a large-scale case study conducted at the Royal Library of Belgium in its online historical newspapers platform BelgicaPress (https://www.belgicapress.be). Launched in 2015, BelgicaPress provides online access to more than two million pages of digitised Belgian newspapers spanning the period 1831-1950. The user interface offers functionalities such as full text searching across the OCR’ed pages. Other search parameters include time ranges, specific dates, newspapers titles and languages (French, Dutch or German).

The object of the study is a dataset of 83 854 queries resulting from 29 812 visits over a 12-month period. By making use of information extraction methods, knowledge bases and various authority files, this research presents the possibilities and limits to identify what percentage of end users are looking for person- and place names. It demonstrates in an empirical manner how user queries can be extracted from a web analytics tool and how named entities can then be mapped with knowledges bases and authority files, in order to facilitate automated analysis of their content.

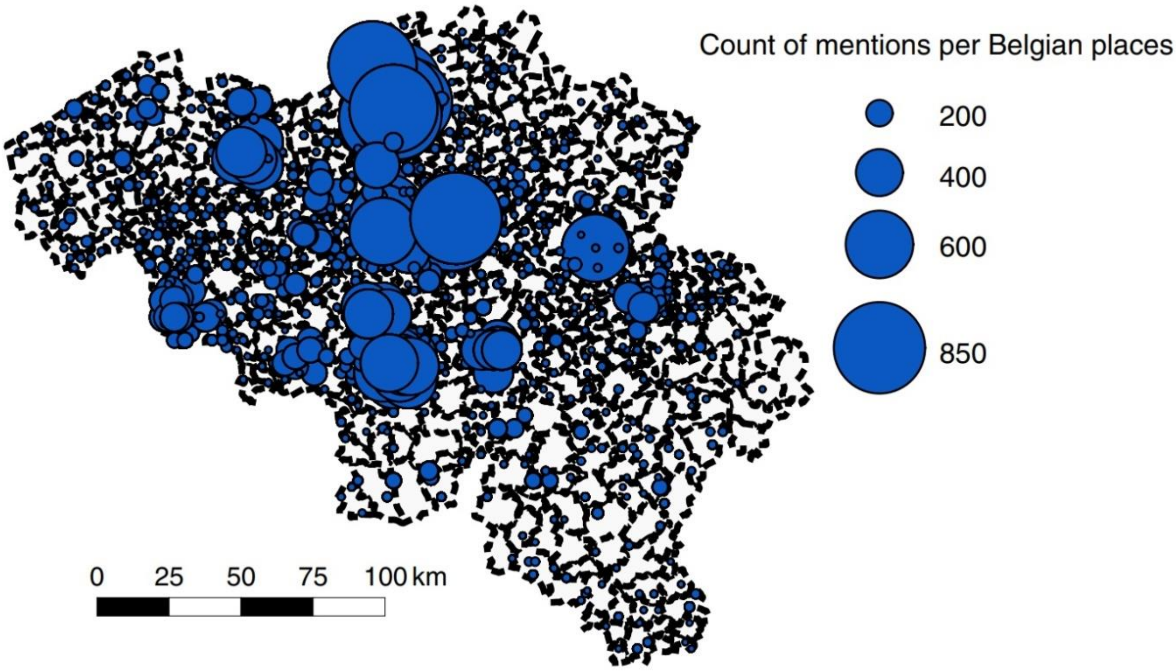


Fig. 16 Mapping of the number of mentions in BelgicaPress user queries, per municipality

As shown in Figure 16, the use of Named Entity Recongnition for user queries analysis allows, for example, collection holders to visualise the possible imbalances in the geographical distribution of the areas of interest. By extension, these quantitative data might be combined to more classical priority indicators for deciding which corpora should be digitised first.

This study (Chardonnens, 2018) also revelas limits related to the authority data used to retrieve Named Entities. For example, Figure 17 demonstrates how the percentage of retrieved personal names crucially depends on the size and the scope of knowledge bases used. For example, one of the user query contains the personal name “Lodewijk Vander Schopen”, a Belgian Writer of the nineteenth century. Fortunately, the Dutch National Library shares its data with the Virtual International Authority File (VIAF), making it possible to identify this individual. If this were not the case, Lodewijk Vander Schopen would not have been retrieved, given that the Dutch National Library is the only institution mentioning him in VIAF and that Wikidata, the other knowledge base used, contains no entry for this writer. This example, as well as the overlap between Wikidata and VIAF being only about one-third (see Figure 17), emphasises how these knowledge bases complement each other and highlights the strategic importance of the type of linked data which will be used for the data reconciliation.

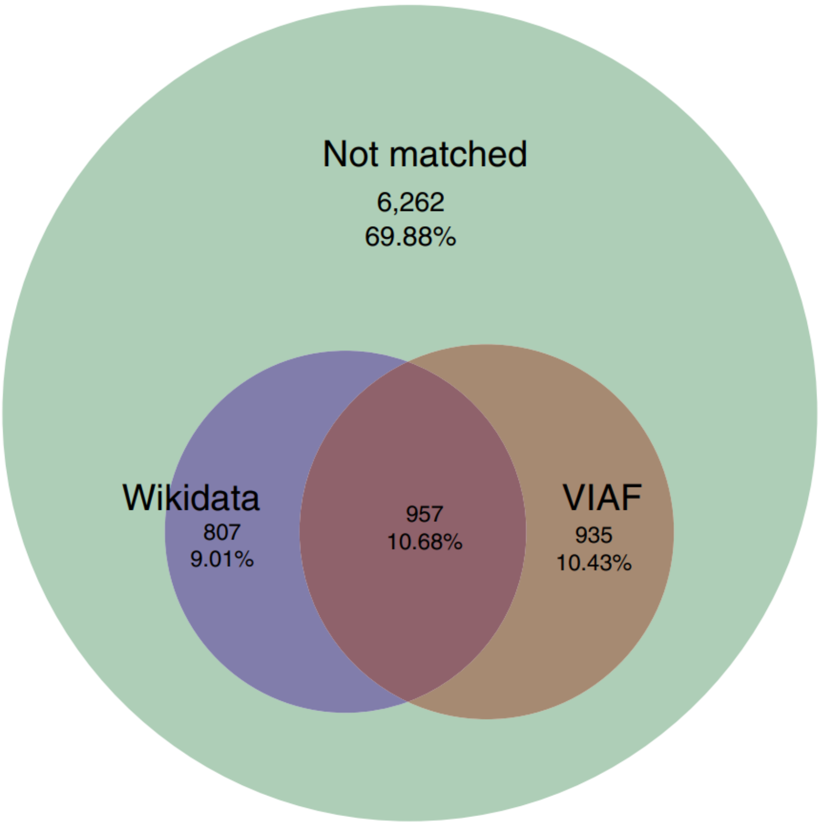


Fig. 17 Breakdown by number and percentage of total candidate names (diagram to scale) in BelgicaPress user queries

As demonstrated in detail in Chardonnens (2018), our method of matching user queries with place and personal names provides salient results. Given that it is freely available for other cultural heritage institutions (in order to favour an open data approach, the code has been made freely available on GitHub: https://github.com/ulbstic/BelgicaPress), the tool is perfectly suited to assist other Belgian institutions to perform a similar analysis, and sufficiently generalisable to be customised by libraries, archives and museums outside Belgium. For the identification of place names, the Geonames values specific to the country should be used. Regarding the identification of personal names, an institution should identify its own local authority list of first names and configure the list with tokens typically included in family names, such as "van" or "von" for Dutch or German.

However, it has to be noted that, although this method based on a quantitative assessment can successfully identify the majority of person- and place names from user queries, a limited amount of queries remained too 'ambiguous' to be treated in an automated manner, due to the specific nature of user queries. In addition, the absence of other preexisting studies on this topic makes it difficult to compare our result (i.e. the ratio of queries concerning place names or persons).

Besides our method to analyse metadata through the prism of the "fitness for use" criteria, we have introduced, in line with literature on the concept of technical debt (see 2.3.1), the concept of "semantic debt" (see Chardonnens, 2020), which refers to quality problems due to the presence of descriptive metadata available only in natural language and therefore subject to ambiguity problems. We consider this form of technical debt as a subset of the code debt category described by Clair (2016). Indeed, just as the source code of an application would be written in a sub-optimal way and would increase the maintenance efforts required to ensure the proper functioning of this application, the descriptive metadata – such as place names, for example – containing literal text that is not machine-readable will increase the maintenance effort required in the future. The price to be paid by the end user will be a sub-optimal search that may generate noise or silence, while the "interest" to be paid by the institution will be in the form of higher processing time for such metadata. For example, any attempt to geolocate data through a dedicated application will first require cleaning and disambiguation work, in order to be able to associate strings of characters with a unique and persistent identifier to which geographic coordinates can be associated.

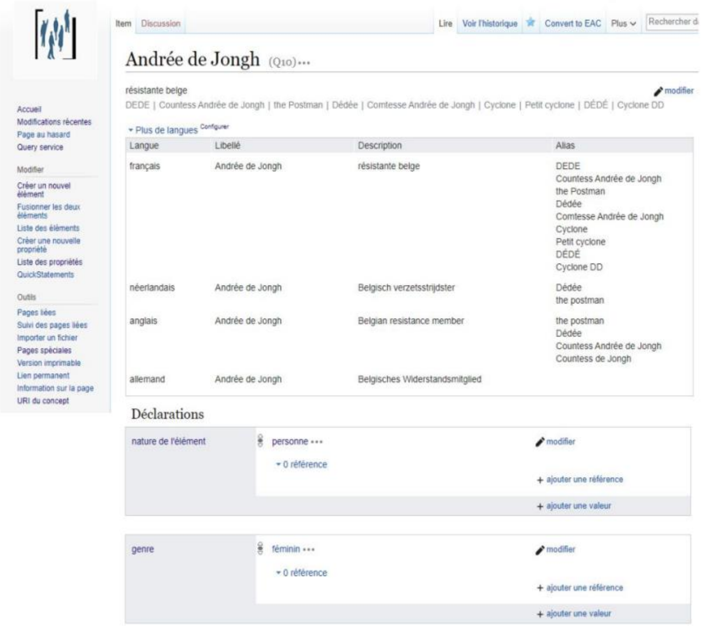
With this contribution, we have underlined both the need to take measures in the context of encoding new data to stop the growth of this technical debt, and the fact that the treatment of the existing debt itself requires resources be devoted to it, making it important that this type of work ceases to be invisible like so many other similar maintenance tasks.

4.3.2 Quality Enhancement Methodologies

Overall, serious efforts are needed in the area of metadata standardisation. In the case of person entities in particular, it is not uncommon to find duplicates or family names preceded only by an initial. Institutions would benefit from developing a coherent and systematic policy for databases containing persons or others types of entities, especially by anticipating the pitfalls of project-based management (different formats/standards, duplicates, redundancy, inconsistencies, etc.). However, we have observed that the development and maintenance of authority data – within the institutions participating in the ADOCHS project – was either non-existent or problematic, which explains some of the quality problems, complicates the creation of new metadata, and undermines the quality of access to the collections. As the importance of authority data continues to grow with the development of the semantic web, it is necessary for institutions to adopt a long-term view on this matter and to consider the creation of permanent identifiers for digital resources, but also for authority records, in order to improve access to collections and preservation in the long term.

Whatever the infrastructure chosen to store authority data (Wikibase – for reading the SWOT analysis on the relevance of such a tool for a cultural heritage institution like the CegeSoma, see the sixth chapter of the Anne Chardonnens' PhD dissertation, 2020; an internally developed tool; any other type of infrastructure adapted to metadata describing cultural heritage), a progressive structuring and semantisation of this data is necessary, whether they come from KBR, CegeSoma/State Archives or another cultural heritage institution. The creation of repositories based on complete and coherent data associated to an URI seems indeed a crucial and indispensable step for these institutions to make their data more accessible, reusable, searchable by humans as well as by machines, but also connected to the resources of other cultural institutions. It is a first step towards the semantic web that will allow these different institutes to join the initiatives towards the general transition of bibliographic and archival metadata (see 2.3.2.).

Let us illustrate this with an example related to CegeSoma collections: Andrée de Jongh, also known as Dédée, Petit Cyclone or even the Postman. This Belgian young woman co-founded the Comete Line, a Belgian resistance network which has helped during the Second World War many English airmen to reach their country, crossing France, the Pyrenees, Spain. Figure 18 shows the (Wikibase) web page displaying, in a human-readable way, the pieces of information which have been associated to the Uniform Resource Indicator (URI) identifying Andrée de Jongh.



Andrée de Jongh (Q10)•••

résistante belge

DEDE | Countess Andrée de Jongh | the Postman | Dédée | Comtesse Andrée de Jongh | Cyclone | Petit cyclone | DÉDÉ | Cyclone DD

▾ Plus de langues

Langue	Libellé	Description	Alias
français	Andrée de Jongh	résistante belge	DEDE Countess Andrée de Jongh the Postman Dédée Comtesse Andrée de Jongh Cyclone Petit cyclone DÉDÉ Cyclone DD
néerlandais	Andrée de Jongh	Belgisch verzetsstrijdster	Dédée the postman
anglais	Andrée de Jongh	Belgian resistance member	the postman Dédée Countess Andrée de Jongh Countess de Jongh
allemand	Andrée de Jongh	Belgisches Widerstandsmitglied	

Déclarations

nature de l'élément: personne •••
▾ 0 référence

genre: féminin •••
▾ 0 référence

Fig. 18 The Wikibase web page (displayed here in French) for the concept URI of Andrée de Jongh: https://adochs.arch.be/entity/Q10.

The same content can be displayed as well in the form of a knowledge graph, as shows in Figure 19.

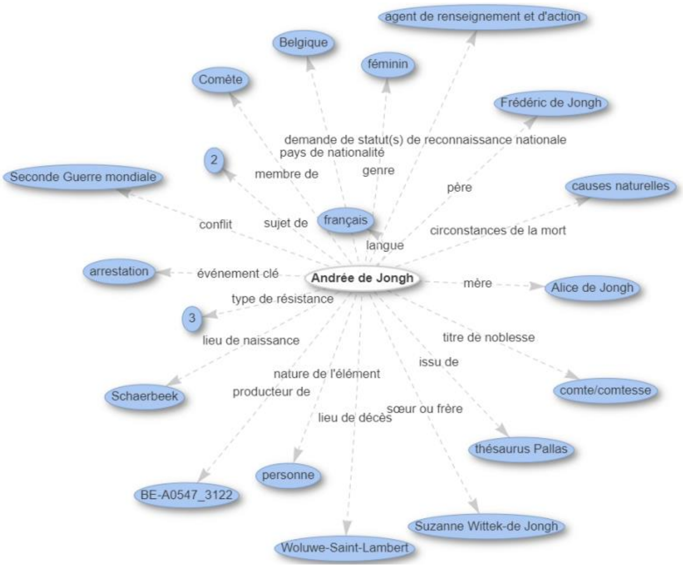
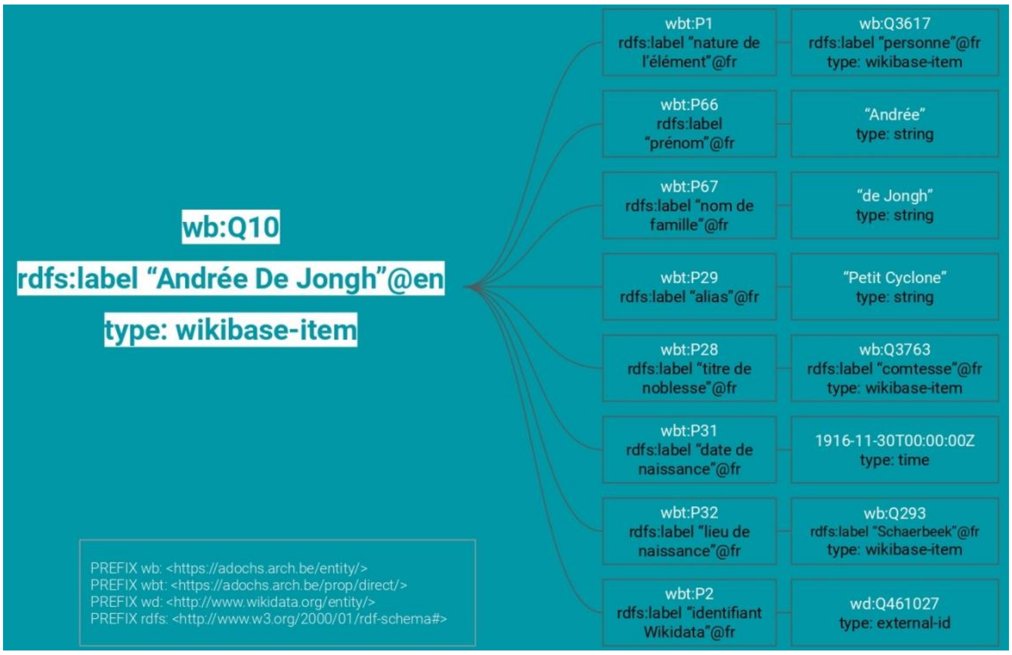


Fig. 19 The graph view (displayed here in French) for the concept URI of Andrée de Jongh: https://adochs.arch.be/entity/Q10.

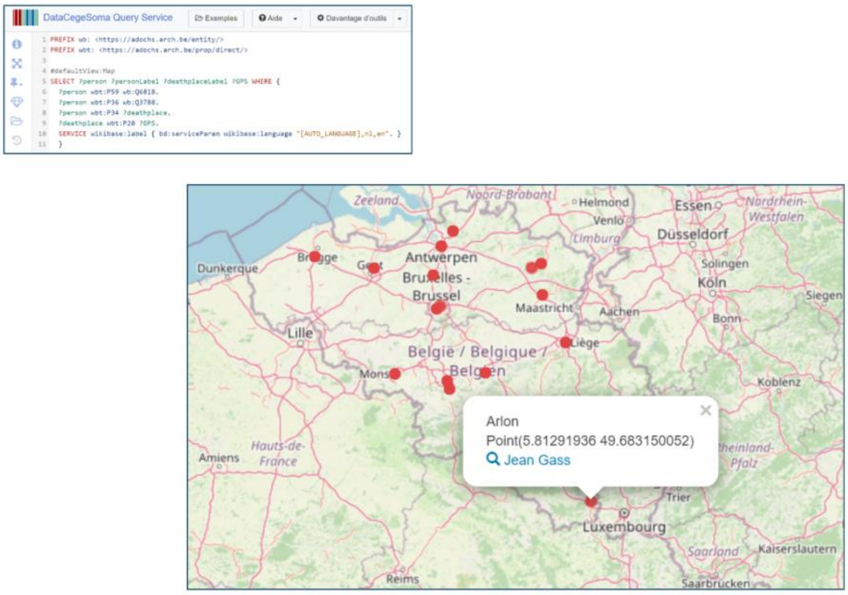
Figure 20 shows the data “behind” Figures 18 and 19: i.e. the result obtained when the “semantic metadata debt” (see 4.3.1) has been paid. The metadata about Andrée de Jongh are no longer strings with potentially some ambiguity. They are structured, multilingual, linked, human- and machine-readable data, which can be queried through an API or a SPARQL endpoint.


wb:Q10
rdfs:label “Andrée De Jongh”@en
type: wikibase-item
wbt:P1 rdfs:label "nature de l'élément"@fr
wb:Q3617 rdfs:label "personne"@fr type: wikibase-item
wbt:P66 rdfs:label “prénom”@fr
“Andrée” type: string
wbt:P67 rdfs:label "nom de famille"@fr
“de Jongh” type: string
wbt:P29 rdfs:label "alias"@fr
"Petit Cyclone" type: string
wbt:P28 rdfs:label "titre de noblesse"@fr
wb:Q3763 rdfs:label "comtesse"@fr type: wikibase-item
wbt:P31 rdfs:label "date de naissance"@fr
1916-11-30T00:00:00Z type: time
wbt:P32 rdfs:label "lieu de naissance"@fr
wb:Q293 rdfs:label "Schaerbeek"@fr type: wikibase-item
wbt:P2 rdfs:label "identifiant Wikidata"@fr
wd:Q461027 type: external-id

```
PREFIX wb: <https://adochs.arch.be/entity/>
PREFIX wbt: <https://adochs.arch.be/prop/direct/>
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
```

Fig. 20 Overview of the data model used to describe the person entity “Andrée de Jongh”.

Beyond the fact that it restricts ambiguity, it also offers the possibility to improve user experience by creating new way of accessing data: for example by looking for all resistance members related to a specifical archival record and seeing on the map the places where they were shot during the war, using a SPARQL query, as depicted in Figure 21.

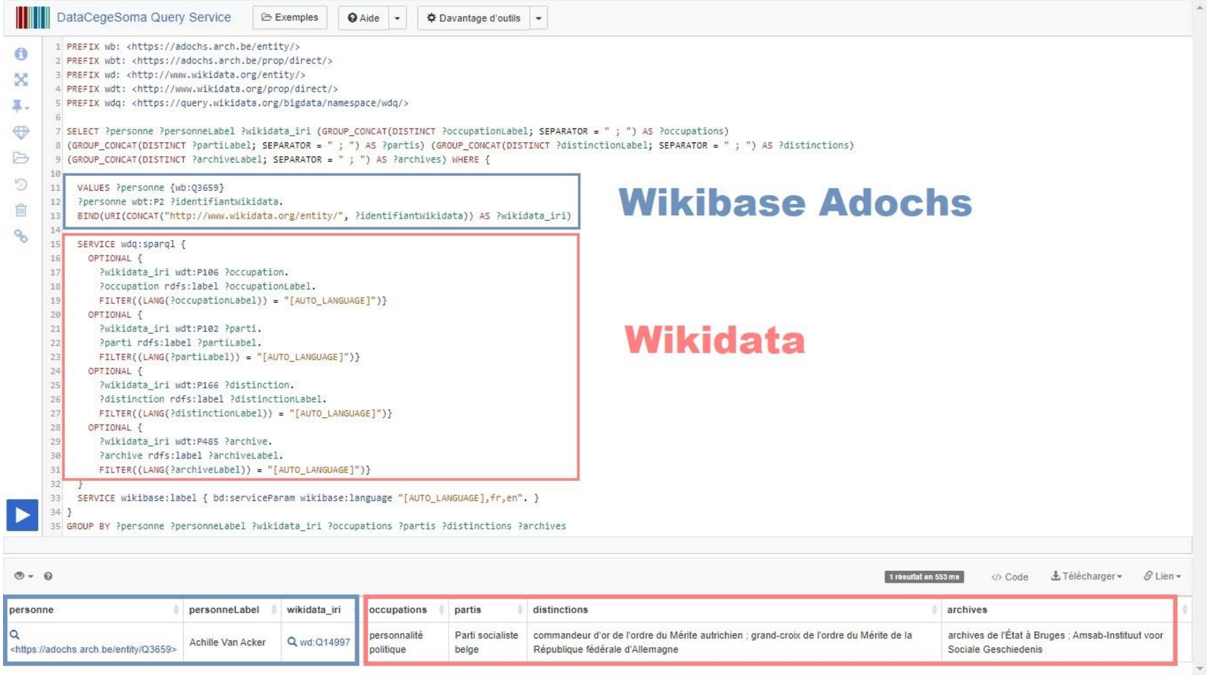


```
PREFIX wb: <https://adochs.arch.be/entity/>
PREFIX wbt: <https://adochs.arch.be/prop/direct/>

#defaultView:Map
SELECT ?person ?personLabel ?deathplaceLabel ?GPS WHERE {
  ?person wbt:P59 wb:Q6818.
  ?person wbt:P36 wb:Q3788.
  ?person wbt:P34 ?deathplace.
  ?deathplace wbt:P20 ?GPS.
  SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],nl,en". }
  }
```

Fig. 21 Mapping of the places of execution of the persons associated with the Cegesoma AA2346 archive, including Jean Gass, shot in Arlon on 7 July 1943.

Finally, one of the most promising results is undoubtedly the possibility to combine data from several repositories. Thus, by using federated SPARQL queries, we have shown how we can retrieve, for a given person from our ADOCHS Wikibase, any information present on Wikidata about his occupation(s), affiliation(s) to a political party, distinction(s) received(s) or about institutions with records about him, as illustrated in figure 22.



```
PREFIX wb: <https://adochs.arch.be/entity/>
PREFIX wbt: <https://adochs.arch.be/prop/direct/>
PREFIX wd: <http://www.wikidata.org/entity/>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
PREFIX wdq: <https://query.wikidata.org/bigdata/namespace/wdq/>

SELECT ?personne ?personneLabel ?wikidata_iri (GROUP_CONCAT(DISTINCT ?occupationLabel; SEPARATOR = " ; ") AS ?occupations)
(GROUP_CONCAT(DISTINCT ?partiLabel; SEPARATOR = " ; ") AS ?partis) (GROUP_CONCAT(DISTINCT ?distinctionLabel; SEPARATOR = " ; ") AS ?distinctions)
(GROUP_CONCAT(DISTINCT ?archiveLabel; SEPARATOR = " ; ") AS ?archives) WHERE {

  VALUES ?personne {wb:Q3659}
  ?personne wbt:P2 ?identifiantWikidata.
  BIND(URI(CONCAT("http://www.wikidata.org/entity/", ?identifiantWikidata)) AS ?wikidata_iri)

  SERVICE wdq:sparql {
    OPTIONAL {
      ?wikidata_iri wdt:P106 ?occupation.
      ?occupation rdfs:label ?occupationLabel.
      FILTER((LANG(?occupationLabel)) = "[AUTO_LANGUAGE]")}
    OPTIONAL {
      ?wikidata_iri wdt:P102 ?parti.
      ?parti rdfs:label ?partiLabel.
      FILTER((LANG(?partiLabel)) = "[AUTO_LANGUAGE]")}
    OPTIONAL {
      ?wikidata_iri wdt:P166 ?distinction.
      ?distinction rdfs:label ?distinctionLabel.
      FILTER((LANG(?distinctionLabel)) = "[AUTO_LANGUAGE]")}
    OPTIONAL {
      ?wikidata_iri wdt:P485 ?archive.
      ?archive rdfs:label ?archiveLabel.
      FILTER((LANG(?archiveLabel)) = "[AUTO_LANGUAGE]")}
  }
  SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],fr,en". }
}
GROUP BY ?personne ?personneLabel ?wikidata_iri ?occupations ?partis ?distinctions ?archives
```

personne	personneLabel	wikidata_iri	occupations	partis	distinctions	archives
<https://adochs.arch.be/entity/Q3659>	Achille Van Acker	wd:Q14997	personnalité politique	Parti socialiste belge	commandeur d'or de l'ordre du Mérite autrichien ; grand-croix de l'ordre du Mérite de la République fédérale d'Allemagne	archives de l'État à Bruges ; Amsab-Instituut voor Sociale Geschiedenis

Fig. 22 Federated SPARQL query combining data from the ADOCHS Wikibase and Wikidata.

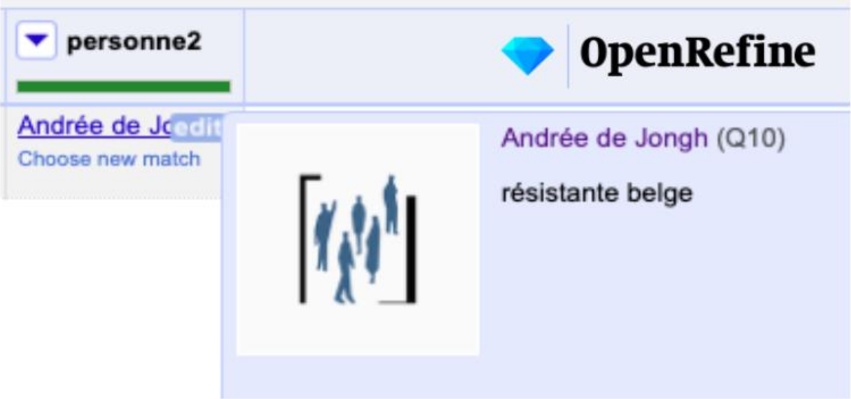


Fig. 23 Reconciliation of a named entity (here "Andrée de Jongh") associated to digitised content by connecting the OpenRefine reconciliation service to the ADOCHS Wikibase

One of the main achievements is certainly that the framework and tools are now in place to improve the quality of metadata describing digitisation projects: whether the entities (persons, places, etc.) are manually encoded for indexing images or extracted automatically by applying Named Entity Recognition techniques to OCRised documents, it is now possible to reconcile them directly with the Wikibase entities (*i.e.* the institution's authority data). Thus, a reference to "Andrée de Jongh" (or even her code name "Little Cyclone") can be associated with the URI identifying her identifying it in a permanent and unambiguous way. We have successfully tested a reconciliation system based on OpenRefine, connected to the Wikibase instance created in the framework of the ADOCHS project, as shown in Figure 23.

The ADOCHS project has been an unique opportunity for experimentation based on empirical data, resulting in the creation of a Wikibase prototype publicly accessible online but also in a PhD thesis answering these three research questions:

1. To what extent is it possible to rely on automation processes to reduce the semantic debt on authority data?
2. How can the functionalities offered by the Wikibase software be used to facilitate and rationalize the creation and maintenance of authority data?
3. How can Linked Open Data facilitate new forms of mutualisation that can reduce the volume of data to be maintained?

This first question aimed to study strategies for acting on the semantic debt rather than letting it accumulate. We have observed that automation is hardly possible before several pre-treatment steps, which are characterised by their time-consuming nature. Moreover, if algorithms can help to reconcile entities and save time, we saw a trend appear that was reminiscent of the Pareto principle, which states that 80% of the effects are the product of 20% of the causes. Indeed, we noted that a minority of data with uncertain reconciliation scores, would represent 80% of the processing time, as manual checks are extremely time-consuming. In such cases, if there are not enough resources to perform thorough manual audits or the volume is too large, alternative solutions exist. By adopting a “good enough” approach, it is possible to provide the user with as much information as possible, without ignoring the ambiguity or uncertainty surrounding certain data, or jeopardizing the credibility of the institution (the goal is to being able to transparently indicate that data are imperfect and that some uncertainty remains). Moreover, it must be kept in mind that if the semantisation of data wants to reconcile data with controlled vocabularies or URIs, it raises the question of what data-source will be used as referential: a question that requires cost-benefit analyses.

The second question is intended to explore how the functionalities offered by the Wikibase software can be used to facilitate and rationalize the creation and maintenance of authority data? As noted by Lovins and Hillmann (2017) in the context of bibliographic vocabularies, it is useful to have maintenance functions directly integrated into tools and workflows. The Wikibase software allows this, by providing tools such as ‘versioning’, which allows the user to consult the history of the operations that have been carried out at any time and to return to a previous version if necessary, or the monitoring lists that facilitate the quality control of the data. This represents a radical shift from current, more informal practices, which include a great deal of oral or personal exchanges, which cannot then be archived and accessed by others in the future. Furthermore, the systematic recording and display of operations traditionally performed anonymously, creates visibility to data creation and maintenance tasks. The possibility to quantify these operations represents an opportunity to add value to valorize tasks that are usually invisible and therefore to create arguments for a more efficient allocation of structural resources (although such tools might also be used less benevolently for productivity control purposes).

This third research question wanted to observe the concrete possibilities of data sharing or the reuse offered by Linked Open Data technologies, in the 2020s. It turns out that this remains far from obvious. Since the Wikibase instance federation project is still under development, efforts to synchronize data from different instances proved to be laborious and finally even created a new form of maintenance that needs to be done. For example, in the context of reuse of Wikidata items, a cost-benefit trade-off must be made in order to determine whether it is preferable to import this data to use them as new elements of the Wikibase or rather to use them as external identifiers. Furthermore, if external data can be retrieved using federated queries (see figure 22), one must keep in mind that this requires a pivot between the two sources concerned. This alignment step is very resource intensive, as we have pointed out. It also requires that the data is freely accessible through a SPARQL access point, a practice which is not yet widely used in the context of archives and cultural heritage. Given that the value of linked data increases with the number of shared datasets, this means that, depending on the domain, the benefits that can be derived from such from such queries may be small or non-existent. Last but not least, this leads to new challenges in the design of search interfaces, which must be able to clearly and transparently display the provenance of data from external sources, but also that this type of information will not be systematically available and that its quality greatly varies.

Finally, it has to be noted that once the data is structured, semantised and linked, additional efforts are required to make it readily available for users. These efforts include both the presentation of data in a more user-friendly format, and the development of tools to translate questions formulated in natural language into SPARQL queries – questions that would benefit from further investigation in the future.

4.3.3 Future work

In this project, we have highlighted the importance of working with structured machine-readable metadata. By experimenting with authority records and publishing them as knowledge graphs, we were able to show the possibilities and limitations and to establish recommendations that can be generalised by other institutions. The results of this work, but also the experience gained during the project, have been a crucial step to enable subsequent research and have already led to new initiatives.

In particular, concerning the Belgian National Library (KBR), the ADOCHS researcher was able to share her expertise and provide advice on issues such as the use of a Wikibase instance in new research projects, or the sharing of authority data via Wikidata. She was also able to actively participate in the reflection launched by KBR on the topic of a national cooperation for the management of Belgian authorities. In addition, it should be noted that the effort in publishing LOD at KBR continues through the development of the LOD-ISNI project (https://www.kbr.be/en/projects/lod-isni/).

Building upon the expertise gained through the project, Cegesoma has decided to pursue his efforts in the publication of machine-readable authority data as part of a new ambitious project: Wikibase Resistance (https://cegesoma.be/en/project/wikibase-resistance). Hosting structured, multilingual, human- and machine-readable data, this knowledge base will allow searching among several hundred thousand descriptions of about 150,000 resistance fighters, starting from a name, a date of birth, or a place of residence.

The ADOCHS project notably shed light on the need to adopt long-term visions and to pay attention to the issue of maintenance. Concerned about this, the promoters of the Wikibase Resistance project – including one of the ADOCHS researchers – do not wish to develop it in isolation and have initiated a global campaign at the level of the Belgian national state archives, in order to adopt permanent identifiers based on the domain name *data.arch.be*. Currently, under construction, the platform should be made public at the end of 2022.

Hopefully, this initiative will also be an opportunity to investigate further the aforementioned issue of user experience, be it through a more user-friendly presentation of the data or the possibility to convert natural language questions into the SPARQL query language.

4.4 Processes Quality

Overall, the purpose of digitisation projects is to generate the richest information in order to answer questions and satisfy target audiences while serving the development of the cultural institution offering the service. Based on the definition of quality as defined by ISO 9000 (Ref), heritage and documentary digitisation involves :

- Technical skills and in-depth knowledge of images and metadata, as well as a good understanding of digitisation equipment.
- An overview of the most frequent quality problems and their causes, so that a coherent methodology can be developed upstream of the project and unforeseen events during the digitisation process can be handled as effectively as possible.
- Management of the working environment, which includes the organisation of spaces and the control of light, as well as the measurement and analysis of atmospheric conditions, in order to make the work flow more smoothly, to guarantee the best conditions for image capture and to minimize the potential deterioration of heritage collections.

- The definition and use of organisational tools to coordinate the various positions in the digitisation chain and facilitate communication between the operators in the studio.
- More broadly, each digitisation project should be part of an overall policy for the preservation, digitisation and dissemination of its digitised content, as well as a data management strategy.

This planning stage is rarely fully carried out, due to a lack of budget, resources or working methods. Therefore, any initiative should start with a diagnosis of the project's objectives and its target audience before moving on to define the technical components. What is the purpose of digitisation? What service does it provide and how does it contribute to the short, medium and long term development of the scientific and cultural institution? This approach must be thought out and planned according to the technical, human and financial resources of the institution. To be efficient, the system must be adapted to the real resources of the organisation initiating the project. In concrete terms, the answers to these questions should be reflected in three documents:

1. *The Digital Data Strategy*, that defines, approves and communicates strategies, policies, standards, architectures and procedures surrounding data management. The better the data created, classified and processed by the organisation is, the more readable, relevant and reliable the information system will be.

2. *The Digitisation Policy*, that follows directly from the data management strategy, and brings together all the principles governing the digitisation of heritage, documentary and archival collections. It also includes all the procedures for developing and carrying out projects as well as the role of the teams during the process.

3. And finally the *Specifications* that outline in detail all the stages of the project. This document includes a summary of the services to be provided, the presentation of the documents to be digitised, their breakdown into batches, the execution deadlines, the parameters for calibrating the chain or the procedure to be followed in case of anomalies. In short, it is the concrete plan for conducting the digitisation project.

Once the *Specifications* have been established, it is then possible to create a map of the digitisation chain. This visual representation of the different stages offers a valuable overview to add to the specifications: it makes it possible for everyone to better understand their roles in the process, to understand the sequence and interactions between the stages and to visualize the key stages in the quality management of the project. This involves defining the operational chain of the digitisation project, the time frame for each task and the quality controls associated with these tasks. Those controls can be divided into three categories:

- Visual control, which checks the legibility and aesthetics of the image (margins, background colour, blurring, etc.).

- Technical control, which consists of checking the technical elements of the digitised content (e.g. resolution/definition, file formats, metadata (EXIF, IPTC XMP model or others), presence of a colour profile, test pattern analysis, etc.).

- Integrity or consistency control, which involves checking compliance with models that have an impact on the integration of images in the digital document management tool (examples: tree structure, file name, directory name, completeness, etc.).

Finally, each digitisation project should be evaluated to identify the strengths and weaknesses of the project and to determine what changes need to be made to improve the digitisation chain. This approach is based on the *PCDA* – Plan - Do - Change - Act – quality management method, which aims to create a virtuous circle that constantly strives to improve the quality of projects and institutions.

5. DISSEMINATION AND VALORISATION (BIBLIOGRAPHY)

5.1 Websites

- Creation and maintenance of the project website: www.adochs.be (long-term access via a capture made by Internet Archive available at: https://web.archive.org/web/20210815054248/http://adochs.be/)

- ADOCHS YouTube channel with the recordings of presentations made during the closing day of the ADOCHS project: https://www.youtube.com/channel/UCTn40lwGvET9lYub4MKr96Q

- Interactive appendices of Anne Chardonnens' PhD thesis: https://linkingthepast.org

- Wikibase prototype created in the context of ADOCHS: https://adochs.arch.be

5.2 Organisation of conferences, workshop, symposia

- 31/05/2018, "Machine Learning for Information Management", co-organised by Dr. Chardonnens at the ULB, Brussels, with Dr. van Hooland.
- 21/11/2019, "Linking the Past", International Conference organised by the ADOCHS project at KBR, Brussels (program and slides available online: http://adochs.be/linking)
- 14/09/2021, "Image & Data Processing in the Cultural Heritage Sector", International Conference organised by the ADOCHS project online (program, slides and recordings available online: http://adochs.be/idpchs/)

5.3 Participation to conferences, workshops, seminars and other valorisation activities

Brault, C., "Digitisation & Quality", plenary session of the Digit staff of the federal scientific establishments, 26 November 2021

Chardonnens, A., "Going beyond Google Analytics: Extending the possibilities for the monitoring and reporting of usage data for cultural heritage collections", the 9th Qualitative and Quantitative Methods in Libraries International Conference (QQML2017), Limerick (Ireland), 22-26 May 2017

Chardonnens, A., "Présentation de l'étude de cas « BelgicaPress » sur les requêtes d'utilisateurs", Cours d'Architecture des Systèmes d'Information du Professeur Seth van Hooland, Université libre de Bruxelles, Brussels, 9 November 2017

Chardonnens, A., "ADOCHS", Présentation du projet devant le personnel Digitp@t des Archives de l'État, CegeSoma, Brussels, 16 November 2017

Chardonnens, A., “Close-reading of linked data: a case study in regards to the quality of online authority files”, DARIAH-EU Workshop: Trust and Understanding: the value of metadata in a digitally joined-up world, Archives de l’État en Belgique (Bruxelles), 14-15 May 2018

Chardonnens, A., “Les entités Personne dans les collections patrimoniales”, (Linked) Data Archiving and Curation: TIC Collaborative and beyond (a joint internal DARIAH-BE and TIC Collaborative workshop), Amsab-ISG (Gand), 25 June 2018

Chardonnens, A., “Web Analytics & Requêtes utilisateurs”, Cours d’Architecture des Systèmes d’Information du Professeur Seth van Hooland, Université libre de Bruxelles, Brussels, 8 November 2018

Chardonnens, A., “Wikidata dans le paysage des données ouvertes et liées”, GLAM & Wikimedia : bilan des projets en Suisse et perspectives avec Wikidata (2019), Bibliothèque nationale de Suisse, Bern, 28 March 2019

Chardonnens, A., “Wikidata et Wikibase : des données structurées”, Cours d’Architecture des Systèmes d’Information du Professeur Seth van Hooland, Université libre de Bruxelles, Brussels, 7 November 2019

Chardonnens, A., “From CegeSoma spreadsheets to Linked Data : a Wikibase journey”, Linking the Past, KBR, Brussels, 22 November 2019

Chardonnens, A., “Linked (Open) Data & Data Mining in Archives and Libraries”, Into the Contents of Collection (journée d’étude internationale organisée par l’Association professionnelle des Archives et Bibliothèques de Belgique), KBR, Brussels, 13 December 2019

Chardonnens, A., “Wikibase @CegeSoma : méthodologie et retour d’expérience”, formation Linked Open Data donnée par Jean Delahousse, CegeSoma, Brussels, 27 January 2020

Chardonnens, A., PhD Public Defense to obtain the degree of Doctor in Information and Communication, Faculté de Lettres, traduction & communication, online, 15 December 2020

Chardonnens, A., “Le projet Wikibase Résistants aux Archives de l’État en Belgique”, Webinaire wiki, data et GLAM 2021, online, 10 June 2021

Chardonnens, A., “Le projet Wikibase Résistants”, Séminaire interuniversitaire UCL-UAntwerpen, CegeSoma, Brussels, 5 October 2021

Gillet, F., Rizza, E., & Chardonnens, A., “CegeSoma case study”, Workshop Wikidata/Wikibase, Ugent, Gent, 03-05 July 2019

Lu, T., “Towards Physical damage Identification and Removal in Document Images”, 7th European Workshop on Visual Information Processing (EUVIP), 26-28 November 2018

Lu, T., “Towards Content Independent No-reference Image Quality Assessment Using Deep Learning”, 4th IEEE International Conference on Image, Vision and Computing (ICIVC), Xiamen, China, July 5-7, 2019.

Lu, T., “A Deep Transfer Learning Approach to Document Image Quality Assessment”, poster presentation, 15th International Conference on Document Analysis and Recognition (ICDAR), Sydney, Australia, 20-25 September 2019

Lu, T., & Dooms, A., “Noise Characterization for Historical Documents with Physical damages”, SPIE International Conference on Optics, Photonics and Digital Technologies for Imaging Applications VI, 113530F, 4-6 April 2020.

Lu, T., PhD Public Defense to obtain the degree of Doctor of Sciences, Department of Mathematics and Data Science, online, 23 October 2020

6. PUBLICATIONS

6.1 Peer reviewed

Chardonnens, A., Rizza, E., Coeckelbergs, M., & Van Hooland, S., 2018, "Mining User Queries with Information Extraction Methods and Linked Data", Journal of Documentation, 74(5), pp. 936-950.

Lu, T., & Dooms, 2018, “Towards Physical distortion Identification and Removal in Document Images”, 7th European Workshop on Visual Information Processing (EUVIP)

Lu, T., & Dooms, 2019, “Towards Content Independent No-reference Image Quality Assessment Using Deep Learning”, 4th IEEE International Conference on Image, Vision and Computing (ICIVC), Xiamen, China, July 5-7 2019.

Lu, T., & Dooms, 2019, “A Deep Transfer Learning Approach to Document Image Quality Assessment”, 15th International Conference on Document Analysis and Recognition (ICDAR), Sydney, Australia, September 20-25 2019.

Lu, T., & Dooms, A., 2020, “Probabilistic Homogeneity for Document Image Processing”, Pattern Recognition, vol. 109, pp. 107591.

Lu, T., & Dooms, A., 2020, “Noise Characterization for Historical Documents with Physical damages”, Proc. SPIE 11353, Optics, Photonics and Digital Technologies for Imaging Applications VI, 113530F.

Lu, T., & Dooms, A., 2020, “Bayesian damage recognition in document images based on a joint global and local homogeneity model”, Pattern Recognition, vol. 118, pp. 108034.

Rizza, E., Chardonnens, A., & Van Hooland, S., 2019, “Close-reading of Linked Data: a case study in regards to the quality of online authority files”, ABB: Archives et Bibliothèques de Belgique -

Archiefen Bibliotheekwezen in België, Trust and Understanding: the value of metadata in a digitally joined-up world, ed. by R. Depoortere, T. Gheldof, D. Styven and J. Van Der Eycken, 106, pp. 37-46.

Lemmers, Frédéric. (2018) « Digitizing Sound Archives at Royal Library of Belgium. Challenges and Difficulties Encountered During a Major Digitization Project. Bibliothek, Forschung und Praxis 42.2:263-271; Article-DOI : 10.1515/bpf-2018-0035 ;

Vanweddingen, Vincent, Chris Vastenhoud, Marc Proesmans, Hendrik Hameeuw, Bruno Vandermeulen, Athena Van der Perre, Frédéric Lemmers, Lieve Watteeuw, Luc Van Gool. (2018). "A Status Quaestionis and Future Solutions for Using Multi-light Reflectance Imaging Approaches for Preserving Cultural Heritage Artefacts." In: Euromed 2018: Digital Heritage. Process in Cultural Heritage: Documentation, Preservation, and Protection. 204-2011.

6.2 PhD Dissertations

Chardonnens, A., 2020, "La gestion des données d'autorité archivistiques dans le cadre du Web de données", PhD Dissertation, 420 p., [online], http://hdl.handle.net/2013/ULB-DIPOT:oai:dipot.ulb.ac.be:2013/315804.

Lu, T., 2020, "Homogeneity Models for Image Processing in the Cultural Heritage Sector", PhD Dissertation, 245 p.

6.3 Other publications

Tan Lu, Ann Dooms, 2019, European Patent Application EP 19173367.4: Computer Implemented Method For Segmenting A Binarized Document.

Brault, C., 2021, "Digitisation & Quality. Guide to managing and controlling quality in a heritage and document digitisation project", 100 p., available online (in French, English and Dutch), on the Cegesoma website: https://www.cegesoma.be/en/publication/digitisation-quality-guide-cegesoma

7. ACKNOWLEDGEMENTS

A word of thanks goes out to:

- Staff members of the State Archives/KBR working in digitisation workshops who agreed to be interviewed in the context of the *Quality Control Procedures* work package
- Staff members of the State Archives/KBR working in IT department for their help and support
- Domain experts of the State Archives/KBR for sharing their knowledge with the project researchers
- Prof. Apostolos Antonacopoulos and Dr. Christian Clausner, respectively Director and Researcher of PRImA - University of Salford Manchester, for their support; especially for our work on document image processing
- The EHRI (European Holocaust Research Infrastructure) fellowship programme and the Foundation Jewish Contemporary Documentation Center (CDEC) which has welcomed one of the researchers in the context of her work on metadata improvement
- Florence Gillet, who coordinated the research project during its first 4 years
- Members of the follow-up committee from which researchers have received solid advices: Sally Chambers, Brecht Declercq, Hans van Dormolen, Frédéric Dufaux, Harco Gisjbers, Péter Király, Jean-Philippe Moreux, Maarten Savels, Julie Urbain, Erwin Verbruggen.

8. REFERENCES

A

Agrawal and Doermann, (2009): Mudit Agrawal and David Doermann, 2009, Clutter noise removal in binary document images, 2009 10th International Conference on Document Analysis and Recognition, pp. 556–560.

Abes, 2019, FNE - Preuve de Concept en cours, [online], https://fil.abes.fr/2019/09/04/fne-preuve-de-concept-en-cours

B

Sebastian et al. (2018): Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek, 2018, Deep neural networks for no-reference and full-reference image quality assessment, IEEE Transactions on Image Processing 27, no. 1, 206–219.

Brunnström et al. (2013): Kjell Brunnström, Sergio Ariel Beker, Katrien De Moor, Ann Dooms, Sebastian Egger, Marie-Neige Garcia, Tobias Hoss-feld, Satu Jumisko-Pyykkö, Christian Keimel, Mohamed-Chaker Larabi, Bob Lawlor, Patrick Le Callet, Sebastian Möller, Fernando Pereira, Manuela

Pereira, Andrew Perkis, Jesenka Pibernik, Antonio Pinheiro, Alexander Raake, Peter Reichl, Ulrich Reiter, Raimund Schatz, Peter Schelkens, Lea Skorin-Kapov, Dominik Strohmeier, Christian Timmerer, Martin Varela, Ina Wechsung, Junyong You, and Andrej Zgank, 2013, Qualinet White Paper on Definitions of Quality of Experience.

Batini, C., Cappiello, C., Francalanci, C., & Maurino, A., 2009, Methodologies for data quality assessment and improvement. ACM computing surveys (CSUR), 41(3), 1-52.

Berners-Lee, T., Fielding, R., & Masinter, L., 1998, RFC2396: Uniform resource identifiers (URI): generic syntax.

Boydens, I., 1999, Informatique, normes et temps, Bruylant, 570 p.

Boydens, I. and van Hooland, S., 2011, "Hermeneutics applied to the quality of empirical databases", Journal of Documentation, Vol. 67 No. 2, pp. 279-289.

Brazzo, L. et Mazzini, S., 2015, Open memory project, [online], https://www.bygle.net/wp-content/uploads/2015/04/Open-Memory-Project_3-1.pdf.

Bruce, T. R., & Hillmann, D. I., 2004, The continuum of metadata quality: defining, expressing, exploiting, [online], https://www.researchgate.net/publication/247818823_The_Continuum_of_Metadata_Quality_Defining_Expressing_Exploiting.

C

Cai et al. (2019): Hao Cai, Leida Li, Zili Yi, and Minglun Gong, 2019, Towards a blind image quality evaluator using multi-scale second-order statistics, Signal Processing: Image Communication 71, 88 – 99.

Chen et al. (2013): Kai Chen, Fei Yin, and Cheng-Lin Liu, 2013, Hybrid page segmentation with efficient whitespace rectangles extraction and grouping, 2013 12th International Conference on Document Analysis and Recognition, pp. 958–962.

Clausner et al. (2019): Christian Clausner, Apostolos Antonacopoulos, and Stefan Pletschacher, 2019, Icdar2019 competition on recognition of documents with complex layouts - RDCL 2019, 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 1521–1526.

Chardonnens, A., Rizza, E., Coeckelbergs, M., & Van Hooland, S., 2018, "Mining User Queries with Information Extraction Methods and Linked Data", Journal of Documentation, 74(5), pp. 936-950.

Chardonnens, A., 2020, “La gestion des données d’autorité archivistiques dans le cadre du Web de données”, PhD Dissertation, 420 p., [online], http://hdl.handle.net/2013/ULB-DIPOT:oai:dipot.ulb.ac.be:2013/315804.

Clair, K., 2016, "Technical debt as an indicator of library metadata quality", D-Lib Magazine, 22(11):3.

Claerr, T., Westeel, I., 2011, Manuel de la numérisation, Paris : Éd. du Cercle de la, 2011. 317 p.

Cunningham, W., 1992, "The WyCash portfolio management system", ACM SIGPLAN OOPS Messenger, 4(2):29–30.

D

Diefenbach, D., Wilde, M. D., & Alipio, S., 2021, October, Wikibase as an Infrastructure for Knowledge Graphs: The EU Knowledge Graph, In International Semantic Web Conference, pp. 631-647, Springer, Cham.

F

Flyvbjerg, B., 2006, "Five misunderstandings about case-study research", Qualitative Inquiry, 12(2). 219–245

G

Godby, J., Smith-Yoshimura, K., Washburn, B., Davis, K. K., Eslao, C. F., Folsom, S., & al., 2019, Creating library linked data with Wikibase: Lessons learned from project passage, [online], https://www.oclc.org/research/publications/2019/oclcresearch-creating-library-linked-data-with-wikibase-project-passage.html

H

Hungenaert, J. et Gillet, F., 2017, Studying user's digital practices and needs in Archives and Libraries. Final Report of the MADDLAIN project. Rapport, State Archives of Belgium, Centre for Historical Research and Documentation on War and Contemporary Society, Royal Library of Belgium, Département des Sciences et technologies de l'Information et de la Communication (ULB), Imec.

Hyvönen, E., 2020, "Sampo" Model and Semantic Portals for Digital Humanities on the Semantic Web, In DHN, pp. 373-378.

I

ICA, 2016, Records in Contexts Conceptual Model, [online], https://www.ica.org/en/records-in-contexts-conceptual-model.

ICOM/CIDOC Documentation Standards Group, 2021, The CIDOC Conceptual Reference Model (CRM), [online], http://www.cidoc-crm.org.

IFLA, 2017, IFLA Library Reference Model: A Conceptual Model for Bibliographic Information, [online], https://repository.ifla.org/handle/123456789/40

IFLA/FRANAR, 2009, Final Report - Functional Requirements and Numbering of Authority Records (FRANAR), [online], https://www.ifla.org/files/assets/cataloguing/frad/frad_2009-fr.pdf.

International Organization for Standardization, 2005, Quality management systems — Fundamentals and vocabulary (ISO 9000: 2005), Geneva : ISO.

International Organization for Standardization, 2015, Quality management systems — Fundamentals and vocabulary (ISO 9000: 2015), Geneva : ISO.

K

Kang et al. (2014a): Le Kang, Peng Ye, Yi Li, and David Doermann, 2014, Convolutional neural networks for no-reference image quality assessment, 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1733–1740.

Kang et al. (2014b): Le Kang, Peng Ye, Yi Li, and David Doermann, 2014, A deep learning approach to document image quality assessment, 2014 IEEE International Conference on Image Processing (ICIP), pp. 2570–2574.

Kim et al. (2017): Jongyoo Kim, Hui Zeng, Deepti Ghadiyaram, Sanghoon Lee, Lei Zhang, and Alan Conrad Bovik, 2017, Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment, IEEE Signal Processing Magazine 34, no. 6, 130–141.

Kim et al. (2019): Jongyoo Kim, Anh-Duc Nguyen, and Sanghoon Lee, 2019, Deep CNN-based blind image quality predictor, IEEE Transactions on Neural Networks and Learning Systems 30, no. 1, 11–24.

Kise et al. (1998): Koichi Kise, Akinori Sato, and Motoi Iwata, 1998, Segmentation of page images using the area voronoi diagram, Computer Vision and Image Understanding 70, no. 3, 370 – 382.

Király, P., 2019, Measuring metadata quality, PhD dissertation, [online], DOI: 10.13140/RG.2.2.33177.77920.

L

Lee et al. (2001): Seong-Whan Lee and Dae-Seok Ryu, 2001, Parameter-free geometric document layout analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 23, no. 11, 1240–1256.

Li et al. (2016): Yuming Li, Lai-Man Po, Litong Feng, and Fang Yuan, 2016, No-reference image quality assessment with deep convolutional neural networks, 2016 IEEE International Conference on Digital Signal Processing (DSP), pp. 685–689.

Li et al. (2017): Pengchao Li, Liangrui Peng, Junyang Cai, Xiaoqing Ding, and Shuangkui Ge, 2017, Attention based RNN model for document image quality assessment, 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), vol. 01, pp. 819–825.

Lu and Dooms, (2018): Tan Lu and Ann Dooms, 2018, Towards physical distortion identification and removal in document images, 2018 7th European Workshop on Visual Information Processing (EUVIP), pp. 1–6.

Lu and Dooms, (2019a): Tan Lu and Ann Dooms, 2019, Towards content independent no-reference image quality assessment using deep learning, 2019 IEEE 4th International Conference on Image, Vision and Computing (ICIVC), pp. 276–280.

Lu and Dooms, (2019b): Tan Lu and Ann Dooms, 2019, A deep transfer learning approach to document image quality assessment, 2019 International Conference on Document Analysis and Recognition (ICDAR), pp. 1372–1377.

Lu and Dooms, (2020a): Tan Lu, Dejan Ilic, and Ann Dooms, 2020, Noise characterization for historical documents with physical distortions, Optics, Photonics and Digital Technologies for Imaging Applications VI, vol. 11353, 2020, pp. 77 – 87.

Lu and Dooms, (2020b): Tan Lu and Ann Dooms, 2020, A novel contractive gan model for a unified approach towards blind quality assessment of images from heterogeneous sources, 15th International Symposium on Visual Computing (ISVC).

Lu and Dooms, (2021a): Tan Lu and Ann Dooms, 2021, Computer implemented method for segmenting a binarized document, patent PCT/EP2020/062909 (2021), filed.
Lu and Dooms, (2021b): Tan Lu and Ann Dooms, 2021, Probabilistic homogeneity for document image segmentation, Pattern Recognition 109 (2021), 107591.

Lu and Dooms, (2021c): Tan Lu and Ann Dooms, 2021, Bayesian damage recognition in document images based on a joint global and local homogeneity model, Pattern Recognition 118 (2021), 108034.

Larson, R. R., Pitti, D., & Turner, A., 2014, September, SNAC: The Social Networks and Archival Context project-Towards an archival authority cooperative, In IEEE/ACM Joint Conference on Digital Libraries, pp. 427-428, IEEE.

Lovins, D., & Hillmann, D., 2017, “Broken-world vocabularies”, D-Lib Magazine.

M

Mao and Kanungo, (2001): Song Mao and Tapas Kanungo, 2001, Empirical performance evaluation methodology and its application to page segmentation algorithms, IEEE Transactions on Pattern Analysis and Machine Intelligence 23, no. 3, 242–256.

Meng et al. (2007): Gaofeng Meng, Nanning Zheng, Yuanling Zhang, and Yonghong Song, 2007, Circular noises removal from scanned document images, Ninth International Conference on Document Analysis and Recognition (ICDAR 2007), vol. 1, pp. 183–187.

Moorthy and Bovik, (2010): Anush Krishna Moorthy and Alan Conrad Bovik, 2010, A two-step framework for constructing blind image quality indices, IEEE Signal Processing Letters 17, no. 5, 513–516.

Moorthy and Bovik, (2011): Anush Krishna Moorthy and Alan Conrad Bovik, 2011, Blind image quality assessment: From natural scene statistics to perceptual quality, IEEE Transactions on Image Processing 20, no. 12, 3350–3364.

N

National Information Standards Organization, 2007, A Framework of Guidance
for Building Good Digital Collections, [online], https://www.niso.org/sites/default/files/2017-08/framework3.pdf.

Neubert, J., 2017, Wikidata as a Linking Hub for Knowledge Organization Systems? Integrating an Authority Mapping into Wikidata and Learning Lessons for KOS Mappings, In Proceedings of the 17th European Networked Knowledge Organization Systems Workshop, pp. 14–25, [online], http://ceur-ws.org/Vol-1937/paper2.pdf.

O

O’Gorman (1993): Lawrence O'Gorman, 1993. The document spectrum for page layout analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 15, no. 11, 1162–1173.

P

Pižurica (2017): Aleksandra Pižurica, 2017, Image denoising algorithms: From wavelet shrinkage to nonlocal collaborative filtering, pp. 1–17, Wiley Encyclopedia of Electrical and Electronics Engineering.

Ohlig, J., 2018, Gemeinsam wieder Neuland betreten : Die Deutsche Nationalbibliothek und Wikimedia Deutschland. [online], https://blog.wikimedia.de/2018/11/02/gemeinsam-wieder-neuland-betreten-die-deutsche-nationalbibliothek-und-wikimedia-deutschland.

R

Rehman and Wang, (2012): Abdul Rehman and Zhou Wang, 2012, Reduced-reference image quality assessment by structural similarity estimation, IEEE Transactions on Image Processing 21, no. 8, 3378–3389.

S

Shafait and Breuel, (2009): Faisal Shafait and Thomas Breuel, 2009, A simple and effective approach for border noise removal from document images, 2009 IEEE 13th International Multitopic Conference, pp. 1–5.

Shah and Gandhi, (2018): Vatsal Shah and Vineet Gandhi, 2018, An iterative approach for
shadow removal in document images, 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1892–1896.

Sheikh et al. (2006): Hamid Sheikh, Muhammad Farooq Sabir, and Alan Conrad Bovik, 2006, A statistical evaluation of recent full reference image quality assessment algorithms, IEEE Transactions on Image Processing 15, no. 11, 3440–3451.

Souza et al. (2003): Andrea Souza, Mohamed Cheriet, Satoshi Naoi, and Ching Yee Suen, 2003, Automatic filter selection using image quality assessment, Seventh International Conference on Document Analysis and Recognition Proceedings, pp. 508–512 vol.1.

Stamatopoulos et al. (2011): Nikolaos Stamatopoulos, Basilis Gatos, Ioannis Pratikakis, and Stavros Perantonis, 2011, Goal-oriented rectification of camera-based document images, IEEE Transactions on Image Processing 20, no. 4, 910–920.

Stevenson, J., 2012, "Linking Lives : Creating an End-User Interface Using Linked Data", Information Standards Quarterly, 24(2/3):14–23, [online], https://www.niso.org/niso-io/2012/06/linking-lives

T

Talebi et al. (2018): Hossein Talebi and Peyman Milanfar, 2018, Nima: Neural image assessment, IEEE Transactions on Image Processing 27, no. 8, 3998–4011.

Tran et al. (2017): Tuan Anh Tran, Kanghan Oh, In-Seop Na, Guee-Sang Lee, Hyung-Jeong Yang, and Soo-Hyung Kim, 2017, A robust system for document layout analysis using multilevel homogeneity structure, Expert Systems with Applications 85, 99 – 113.

W

Wang et al. (2004): Zhou Wang, Alan Conrad Bovik, Hamid Sheikh, and Eero Simoncelli, 2004, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing 13, no. 4, 600–612.

Wang and Bovik, (2009): Zhou Wang and Alan Conrad Bovik, 2009, Mean squared error: Love it or leave it? a new look at signal fidelity measures, IEEE Signal Processing Magazine 26, no. 1, 98–117.

Wu and Hou, (2018): Jiafu Wu and Wenqi Hou, 2018, Mobile scanner : Document segmentation and object removal by exemplar-based image inpainting and spectral regularization.

W

Wikidata 2020, Wikidata : Notability, [online], https://www.wikidata.org/wiki/Wikidata:Notability.

W3C, 2014, Resource Description Framework, [online], https://www.w3.org/RDF

Y

Ye and Doermann, (2013): Peng Ye and David Doermann, 2013, Document image quality assessment: A brief survey, 2013 12th International Conference on Document Analysis and Recognition, 2013, pp. 723–727.

Z

Zhang et al. (2009): Li Zhang, Andy M. Yip, Michael Scott Brown, and Chew Lim Tan, 2009, A unified framework for document restoration using inpainting and shape-from-shading, Pattern Recognition 42, no. 11, 2961 – 2978.

Zhang et al. (2015): Lin Zhang, Lei Zhang, and Alan Conrad Bovik, 2015, A feature-enriched completely blind image quality evaluator, IEEE Transactions on Image Processing 24, no. 8, 2579–2591.