Institution: University of Salford

Unit of Assessment: 11

Title of case study: Enabling digital transformation through effective digitisation

Period when the underpinning research was undertaken: January 2005 – December 2020

Details of staff conducting the underpinning research from the submitting unit:

Name(s):	**Role(s) (e.g. job title):**	**Period(s) employed by submitting HEI:**
Prof. Apostolos Antonacopoulos	Head of the PRImA Research Lab; Professor in Pattern Recognition	January 2005 – Present
Dr Christian Clausner	Research Fellow in Computer Science	October 2009 – Present
Stefan Pletschacher	Lecturer in Computer Science	July 2008 – September 2017; October 2017 – Present

Period when the claimed impact occurred: January 2014 – December 2020

Is this case study continued from a case study submitted in 2014? N

1. Summary of the impact

The sheer volume and variability of information contained in hard-copy documents makes digitisation too costly and complex. As a result, only a fraction of those billions of documents exist in an effective digital fully searchable and editable text form with semantic annotations to allow further analysis and use – the basis of digital transformation. Research by the Pattern Recognition and Image Analysis (PRImA) Laboratory at the University of Salford, working with major libraries and organisations internationally such as the Office for National Statistics, the British Library and FamilySearch, has created new methodologies and corresponding software that:

1. Fundamentally improved the accessibility and utility of critical but previously inaccessible organisational data, as well as of personal, community and cultural heritage
2. Directly influenced digitisation strategy in major national and other large content-holding organisations worldwide by enabling them to evaluate possibilities and create specifications for cost-efficient and effective large-scale digitisation programmes
3. Enabled the creation of new large-scale automated document recognition systems by large organisations, using machine learning approaches.

2. Underpinning research

Converting the vast legacy of books, newspapers and other information-rich documents into modern digital media (a critical part of digital transformation in organisations) has led to new challenges in the area of digitisation. Document analysis and digitisation software tools and evaluation frameworks developed by Salford's Pattern Recognition and Image Analysis (PRImA) Laboratory helped address this demand, leading to new approaches in the mass digitisation of printed text and numerical data.

2.1. The *Aletheia* system: Methods and software for accurate and efficient digitisation

Since 2009, Salford's researchers have led the development of *Aletheia* **[3.1]**, a production-quality software system for very accurate and yet efficient (cost-effective) analysis, recognition and annotation of large amounts of scanned documents. In contrast to existing systems which simply apply the same processes to large amounts of documents, *Aletheia* aids the user to achieve very high precision with an ever-expanding number of automated and semi-automated tools developed and improved over the years. This is based on PRImA research and feedback from stakeholders across the world (major content-holding institutions and commercial service providers), several of which have been using the tool in production environments. *Aletheia* has also been used frequently by organisations globally to create training data for building new mass digitisation systems.

2.2. The *PAGE XML* format

No existing document representation formats adequately supported individual stages within an entire digitisation workflow (from document image enhancement to layout analysis to text recognition) and its evaluation. Since 2009, Pletschacher and Antonacopoulos have led the development of *PAGE*, a new XML-based page image representation framework that records information on image characteristics in addition to layout structure and page content **[3.2]**. *PAGE XML* has been widely adopted by organisations around the world as the de facto standard format to represent highly detailed data within digitisation workflows as well as to store data used to train machine learning algorithms. Supporting tools and a viewer have been implemented by PRImA and used by organisations deploying this format.

2.3. Performance evaluation methodology, tools and resources

Salford's researchers pioneered in 2011 the concept of evaluating the performance of digitisation methods as well as complete workflows according to the use-scenario of the resulting data **[3.3]**. Prior to this, only generic error-rate metrics were widely used, which did not allow content-holding institutions to assess whether existing or about to be commissioned digitisation outputs were fit for purpose. Software tools have been implemented in order to support this realistic performance evaluation methodology and to provide in-depth information at several levels (high-level benchmarking for institutions and detailed low-level reports for developers). New evaluation metrics have been progressively added (e.g. **[3.4]** in 2019) enabling realistic and accurate evaluation in even more complex document situations where other metrics fail, such as in newspaper pages where the wrong grouping of text lines into coherent articles affects the meaning of stories. Several test datasets on which systems can be evaluated have also been created by PRImA and used in international challenges and by organisations deploying this evaluation methodology.

2.4. Optimised digitisation workflows and system training/adaptation

Complex information-rich documents have not been digitised yet, despite their important content, due to the difficult challenges they pose. In particular, extracting validated numerical information from printed tabular documents is a significant and yet very common problem, demanding data modelling and analysis as well as character recognition. In 2017, Salford's researchers led the development of an optimised workflow and novel methodology **[3.5]** which was used in the world's first large-scale digitisation project involving detailed numerical information, funded by ONS (1961 England and Wales Census – Small Area Statistics data). The adaptation (training) of document recognition systems to different document types and application scenarios is essential when creating and improving specialised workflows such as this. In 2018, the team led the development of the first comprehensive training infrastructure **[3.6]**, significantly improving previous very time-consuming and laborious manual processes. Another advantage of this approach is that it does not need significant amounts of data available for training, a common problem with historical documents, which also prevents the use of machine learning approaches.

3. References to the research

3.1. C. Clausner, S. Pletschacher, A. Antonacopoulos, Aletheia - An advanced document layout and text ground-truthing system for production environments, *11th International Conference on Document Analysis and Recognition (ICDAR2011), Beijing, China, September 2011, IEEE-CS Press,* pp. 48-52. https://doi.org/10.1109/ICDAR.2011.19

3.2. S. Pletschacher, A. Antonacopoulos, The PAGE (Page Analysis and Ground-truth Elements) format framework, *20th International Conference on Pattern Recognition (ICPR2010), Istanbul, Turkey, August 2010, IEEE-CS Press,* pp. 257-260. https://doi.org/10.1109/ICPR.2010.72

3.3. C. Clausner, S. Pletschacher, A. Antonacopoulos, Scenario driven in-depth performance evaluation of document layout analysis methods, *11th International Conference on Document Analysis and Recognition (ICDAR2011), Beijing, China, September 2011, IEEE-CS Press,* pp. 1404-1408. https://doi.org/10.1109/ICDAR.2011.282

3.4. C. Clausner, S. Pletschacher, A. Antonacopoulos, Flexible character accuracy measure for reading-order-independent evaluation, *Pattern Recognition Letters*, Volume 131, March 2020, pp. 390-397. https://doi.org/10.1016/j.patrec.2020.02.003 **(REF2)**
3.5. C. Clausner, J. Hayes, **A. Antonacopoulos, S. Pletschacher**, Creating a complete workflow for digitising historical census documents: Considerations and evaluation, *Proceedings of the 4th International Workshop on Historical Document Imaging and Processing (HIP2017), Kyoto, Japan, November 2017, ACM,* pp. 83-88. https://doi.org/10.1145/3151509.3151525
3.6. C. Clausner, A. Antonacopoulos, S. Pletschacher, Efficient and effective OCR engine training, *International Journal on Document Analysis and Recognition (IJDAR)*, 23(1), 2020, pp. 73-88. https://doi.org/10.1007/s10032-019-00347-8 **(REF2)**

The underpinning research has been supported by international grants and contracts: European Commission funded projects (IMPACT 2008 – 2012, Europeana Newspapers Online 2012 – 2015, SUCCEED 2013 – 2015) EUR18,000,000 in total – EUR1,900,000 for PRImA; Office for National Statistics (Digitisation of 1961 Census) GBP251,000; Andrew W. Mellon Foundation, USA (EMOP 2012 – 2014) USD736,000 total – USD86,000 for PRImA; Australian Department of Defence (South Asian language analysis and tools 2016 – 2018) GBP18,000; and Lumex AS, Norway (Evaluation of new ZoneMaster system 2018 – 2020) GBP50,000.

4. Details of the impact

Research by PRImA has enabled and accelerated digital transformation through effective digitisation in several organisations around the world, impacting those organisations and the public in three main ways:

4.1. Enabled effective use of previously inaccessible information

Extended over the years, the capabilities of the methods included in the *Aletheia* document analysis system and the developed *PRImA evaluation methodology* and associated infrastructure have enabled PRImA to create and validate new specialised *recognition workflows* **[2.4]** targeted at difficult to digitise – and hence virtually inaccessible (digitally) – document collections. As a result, information that was previously simply displayed but not able to be utilised (e.g. numerical information in several tabular forms) has been successfully extracted, semantically labelled and is used in complex queries and calculations.
The **Office for National Statistics (ONS)** had an irreplaceable set of very detailed statistical information from the 1961 Census of England and Wales, in the form of a large number of microfilmed computer printouts from that era, some degraded in quality. While unusable in its existing image form, using the *PRImA numerical information digitisation methodology and workflow*, this information was 'recovered' between 2018 and 2019 and now exists in a semantically labelled and structured form, on a par with current digitally-born information and usable for statistical analyses across years/geographies. The ONS's Head of Historical Census Support confirmed: '*The impact on ONS has been significant and wide in reach. First, valuable but previously inaccessible data has been made accessible to the public in a highly operable and quality assured way that would not have been realistically feasible otherwise. Second, based on the success of this first project, in 2020 ONS made the significant strategic investment to digitise, in the same manner, the printed census volumes from 1921 to 1961'* **[5.1]**.
FamilySearch, the world's largest genealogy organisation with 1.3 billion name records online, accessed by millions of users, has licensed *Aletheia* and used it since 2014 (by training some of their hundreds of thousands of volunteers) to efficiently extract, semantically annotate and make available millions more historical family records freely searchable online. This information was previously inaccessible, being part of FamilySearch's unindexed collection of 4.6 billion images of different documents from archives around the world (tabular registry records, newspaper obituaries etc.) spanning the 16th to 18th Centuries in 12 different languages **[5.2]**.

4.2. Influenced digitisation strategy internationally

The combination of the *Aletheia* document analysis system **[2.1]** (to efficiently and accurately create sample data, in *PAGE XML format* **[2.2]**) and the *PRImA evaluation methodology* **[2.3]** has allowed institutions to carry out targeted feasibility studies for different types of material and different use scenarios (possibilities ranging from simple keyword search to full natural language processing of text content). The results enabled strategic decisions to be made about the selection of material to prioritise for digitisation and what the expected quality and utility can be for that material. The latter directly affects the extent to which a given use scenario can be fulfilled and indicates the associated cost for manual intervention/correction during digitisation.

The British Library (BL), the world's largest library, in its search to identify ways to digitise challenging rare material (where standard commercial systems fail), initiated two international challenges: one for early Indian printed books (run in 2017 and 2019) and one for Arabic scientific manuscripts (run in 2017 and 2018) **[5.3]**. *Aletheia* was used to manually digitise and annotate representative samples from the target collections. Organisations and individuals around the world submitted document recognition methods whose results were compared, using the *PRImA evaluation methodology*, against the specified samples. These challenges enabled the BL to gain an understanding of the viability of the state-of-the-art and open problems in order to identify suitable organisations to work further with, and accordingly to make decisions on the direction of its digitisation strategy.

The Wellcome Collection, the library and archive of the Wellcome Trust – one of the largest non-governmental funders of scientific research internationally, had already digitised a relatively small part of its holdings of the Medical Officer of Health reports, which provide medical statistics for all hospitals in England and Wales over a period of 150 years. However, this tabulated numerical information could not be accessed nor queried in a useful way (e.g. what was the deadliest disease in different cities in the 19th Century?). Using *Aletheia* and the *PRImA numerical information digitisation methodology* and workflow **[2.4]**, a scoping study was conducted in early 2018 demonstrating the potential use scenarios for the remaining not yet digitised collection and proposing specialised workflows, enabling Wellcome to make informed decisions on further digitisation efforts **[5.4]**.

The Berlin State Library (SBB), one of the largest libraries in the world and one of the most important research libraries in the German-speaking world, has been using *Aletheia* and the *PRImA evaluation methodology* since 2015 to decide which titles in its collections to include in its large-scale strategic digitisation projects. As a leading partner in the OCR-D project, a major German government-funded initiative to implement and deliver full-text digitisation of all German texts published between 16th and 18th Centuries, the SBB has specified that its workflow uses *PAGE XML* as the main functional format to process and store all digitised material **[5.5]**. Through its leading role in the high-profile OCR-D initiative (multi-year multi-million project funded by the German government to digitise and transcribe all early-modern printed works in the German language), the SBB also used the *PAGE XML format* to store and employ datasets in the training and evaluation of new deep-learning based document recognition systems created by the project **[5.5]**.

The National Library of The Netherlands (KB) used *Aletheia* in early 2020 to prepare comprehensive samples from its already digitised newspaper holdings in a major effort to evaluate whether the outputs of its past digitisation projects satisfied its intended use scenarios. Correspondingly, the KB was able to decide how best to use its digitisation budget: whether the collections involved needed to be re-processed or whether it could focus solely on new digitisation projects **[5.6]**.

The National Library of Finland faced challenges when commercially available software was used to extract individual newspaper articles as part of its national digitisation effort. It used the *PRImA evaluation methodology* in 2018 to compare alternative systems and determine whether a newly created approach would deliver objectively better results **[5.7]**.

4.3. Equipped organisations across the world to create new digitisation systems

The creation of new large-scale document recognition systems based on machine learning requires very significant amounts of accurately labelled data (ground truth) used to train such systems. The *Aletheia* document analysis system **[2.1]** has been used extensively by organisations to efficiently create high quality training data (in *PAGE XML* **[2.2]**) for such systems. In addition, the ability to perform objective fine-grained performance evaluation (to assess the effectiveness of training) is a key requirement for improving document recognition systems. The *PRImA evaluation methodology* and framework **[2.3]** has been instrumental in guiding the development of systems by being the only comprehensive framework that can be flexibly adapted to report on different use scenarios.

In addition to using *Aletheia* to transcribe and make genealogical information accessible to the public (see 4.1 above), thousands of **FamilySearch** subcontractors and volunteers used *Aletheia* since 2017 and produced detailed annotated data in the *PAGE XML format.* This data was used to develop new machine-learning based large-scale text recognition systems used in the processing of millions of scanned historical genealogical records **[5.2]**.

Google's deep-learning based document recognition web service (Cloud Vision) was evaluated among other state-of-the-art methods in the British Library challenge to recognise Arabic scientific manuscripts. Subsequently, using the training data produced for the challenge with *Aletheia* and the *PRImA evaluation methodology*, Google trained a new prototype Arabic text recognition system, substantially improving the performance of the web service (from practically unusable to useful). Writing about the *PRImA evaluation methodology*, the Director in the Perception Team at Google Research stated it to be '*among the most useful tools we have drawn on to benchmark the performance of our recognition systems'* **[5.8]**.

Still in the commercial sector, **Lumex AS (Norway)** and **Skilja GmbH (Germany)** used the *PRImA evaluation methodology* to improve their document analysis systems between 2018 and 2020. The evaluation results were also used both for marketing purposes and to officially report to the Norwegian and German Funding Councils funding their research and development **[5.9]**.

The **Australian Department of Defence** funded the development of extensions to *Aletheia* to enable the analysis and recognition of documents in south east Asian scripts, where a single character of text is composed as a hierarchy of graphemes. It subsequently used *Aletheia* to create labelled data in the *PAGE XML format* and trained an open-source system to recognise south east Asian documents, thus increasing its automated document analysis capabilities **[5.10]**.

English Literature researchers led by the **Center for Digital Humanities at Texas A&M University** used *Aletheia* to create training data for a new document recognition system to convert large and significant collections (EEBO and ECCO) of previously poorly scanned 15th to 18th Century books to full-text for use by humanities scholars worldwide **[5.11]**.

5. Sources to corroborate the impact

5.1. Testimonial: Office for National Statistics (March 2021), on enabling use of previously inaccessible information (4.1)

5.2. Testimonial: FamilySearch (March 2021), on use of *Aletheia* to make records accessible in new formats (4.1) and to recognise historical family records (4.3)

5.3. Research Publication: *'Cross-disciplinary Collaborations to Enrich Access to Non-Western Language Material in the Cultural Heritage Sector'* (2019), The British Library, on influence of PRImA methodology on digitisation strategy (4.2)

5.4. Blog Post: Wellcome Collection (27 June 2018), on influence of *PRImA methodology* on digitisation strategy (4.2)

5.5. Testimonial: Berlin State Library (March 2021), on influence of PRImA methodology on digitisation strategy and on training and evaluation of document recognition systems (4.2)

5.6. Blog Post: National Library of The Netherlands (KB) (17 July 2020), on influence of *Aletheia* on digitisation strategy (4.2)

5.7. Research Publication: *'Detecting Articles in a Digitized Finnish Historical Newspaper Collection 1771–1929: Early Results Using the PIVAJ Software'* (2019), the National Library of Finland, on influence of *PRImA methodology* on digitisation strategy (4.2)

5.8. Testimonial: Google (February 2021), on use of *PRImA methodology* to train a new prototype Arabic text recognition system (4.3)

5.9. Testimonial: Lumex AS, Norway (March 2021), on use of *PRImA methodology* to improve document analysis systems (4.3)

5.10. Research Publication: *'Comparison of Visual and Logical Character Segmentation in Tesseract OCR Language Data for Indic Writing Scripts'* (2015), Defence Science & Technology Group, Australia, on use of *Aletheia* to improve document analysis systems (4.3)

5.11. Research Publication: *'Mass Digitization of Early Modern Texts With Optical Character Recognition'* (2017), Center for Digital Humanities Research, Texas A&M University, on use of *Aletheia* to create training data for a new document recognition system (4.3)