Institution: University of Salford

Unit of Assessment: B11 Computer Science and Informatics

Title of case study: The Pattern Recognition and Image Analysis (PRImA) Research Laboratory: Digitising Europe's printed cultural heritage

1. Summary of the impact

From image capture to online access, the Pattern Recognition and Image Analysis (PRImA) Research Laboratory at the University of Salford is instrumental in improving access to printed works in Europe's libraries and archives. Through developing document analysis, digitisation software tools and evaluation frameworks, and supporting their deployment, PRImA research demonstrates the following impact:

- Improving access to personal, collective and community histories through increasing the availability of previously inaccessible, now digitised objects;
- Maximising production efficiency, quality and volume, whilst significantly reducing the costs of digitisation;
- Providing libraries and archives with access to new generation digitisation technologies, contributing to their sustainability and improving their skills base;
- Bringing benefit to the wider European economy by supporting continued investment in and dissemination of the technologies;
- Supporting exponential improvements in digitisation technologies through partnership.

2. Underpinning research

The key researchers and positions they held at the institution at the time of the research are as follows: Dr Apostolos Antonacopoulos, Senior Lecturer, School of Computing, Science and Engineering, (from 2005).

2005 onwards: The PRImA Research Laboratory relocated to the University of Salford in 2005, expanded considerably and is now one of the most active research laboratories in the field, the UK and internationally. Driven by the need for increased access to information; the transformation of the heritage of books, newspapers and other documents into modern digital media has led to new challenges in the area of digitisation. Document analysis and digitisation software tools and evaluation frameworks developed by PRImA address this demand, leading to the development of new approaches to the mass digitisation of printed text. *Digitising Europe's printed cultural heritage* is underpinned by the following research:

- **2008:** A geometric approach for accurate and efficient performance evaluation of layout analysis methods: A major component of performance evaluation of document layout analysis methods is the comparison of ground truth regions with regions resulting from segmentation methods. Previous approaches favour either accuracy or efficiency, resulting in an impractical compromise:
  - Antonacopoulos led the development of an improved approach that uses polygons to accurately describe both segmentation and ground truth regions, which has been validated using data from the ICDAR page segmentation competitions. **[1]**
- **2009:** A Realistic Dataset for Performance Evaluation of Document Layout Analysis:
  - Antonacopoulos led the compilation of a new dataset on which to evaluate layout analysis methods and a methodology for its creation based on a wide range of contemporary documents. Rendering comprehensive and detailed representation of both complex and simple layouts, and on colour originals, in-depth information is recorded both at the page and region level. **[2]**
- **2010:** The PAGE (Page Analysis and Ground-Truth Elements) Format Framework: No document representation formats adequately supported individual stages within an entire sequence of document image analysis methods (from document image enhancement to layout analysis to OCR) and their evaluation:
  - Antonacopoulos 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]**
- **2011:** Scenario Driven In-depth Performance Evaluation of Document Layout Analysis Methods:

- Antonacopoulos led the development of an advanced framework for evaluating the performance of layout analysis methods combining efficiency and accuracy by using a special interval based geometric representation of regions, including a wide range of sophisticated evaluation measures provides the means for a deep insight into the analysed systems. **[4]**

- **2011:** Aletheia - An Advanced Document Layout and Text Ground-Truthing System for Production Environments: Large-scale digitisation has led to a number of new possibilities for adaptive and learning based methods in the field of Document Image Analysis and Optical Character Recognition (OCR). For ground truth production of large corpora, however, there was still a gap in terms of productivity. Ground truth is crucial for evaluation at the development stage of tools and for quality assurance in the scope of production workflows for digital libraries:
  - Antonacopoulos led the development of Aletheia, a production-quality system for accurate and yet cost-effective ground truthing of large amounts of documents. Aletheia aids the user with a number of automated and semi-automated tools developed and improved in association with major libraries across Europe and commercial service providers, which are now using the tool in a production environment. **[5]**
- **2012:** A robust hybrid approach for text line segmentation in historical documents: Large-scale digitisation of historical documents demands robust methods that cope with the presence of frequent distortions and noisy artefacts:
  - Antonacopoulos led the development of a hybrid text line segmentation method that uses a novel data structure and a rule base to combine the strengths of top-down and bottom-up approaches while minimising their weaknesses. The effectiveness of the approach has been methodically evaluated in the context of large-scale digitisation. **[6]**

3. References to the research

Key outputs

1. D. Bridson, A. Antonacopoulos, "A Geometric Approach for Accurate and Efficient Performance Evaluation of Layout Analysis Methods", *Proceedings of the 19th International Conference on Pattern Recognition (ICPR2008)*, Tampa, Florida, USA, December 7-11, 2008, IEEE-CS Press. DOI
2. A. Antonacopoulos, D. Bridson, C. Papadopoulos, S. Pletschacher, "A Realistic Dataset for Performance Evaluation of Document Layout Analysis", *Proceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR2009)*, Barcelona, Spain, July 2009, pp. 296-300. DOI
3. S. Pletschacher, A. Antonacopoulos, "The PAGE (Page Analysis and Ground-Truth Elements) Format Framework", *Proceedings of the 20th International Conference on Pattern Recognition (ICPR2010)*, Istanbul, Turkey, August 23-26, 2010, IEEE-CS Press, pp. 257-260. DOI
4. C. Clausner, S. Pletschacher, A. Antonacopoulos, "Scenario Driven In-Depth Performance Evaluation of Document Layout Analysis Methods", *Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR2011)*, Beijing, China, September 2011, pp. 1404-1408. DOI
5. C. Clausner, S. Pletschacher, A. Antonacopoulos, "Aletheia - An Advanced Document Layout and Text Ground-Truthing System for Production Environments", *Proceedings of the 11th International Conference on Document Analysis and Recognition (ICDAR2011)*, Beijing, China, September 2011, pp. 48-52. DOI
6. C. Clausner, A. Antonacopoulos, S. Pletschacher, "A Robust Hybrid Approach for Text Line Segmentation in Historical Documents", *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*, Japan, November 11-15, 2012, IEEE-CS Press. ISBN 978-1-4673-2216-4

Key Grants

7. IMPACT: Improving Access to Text (1/1/2008 – 30/6/2012) €1.3M European Commission FP7.

8. Europeana Newspapers: European Newspapers Online (1/2/2012 – 31/1/2015) €426K European Commission ICT-PSP.
9. EMOP: Early Modern OCR project (1/10/2012 – 30/9/2014) US$86K Andrew W. Mellon Foundation, USA.
10. SUCCEED: Support Action Centre of Competence in Digitisation (1/1/2013 – 31/12/2015) €192K European Commission FP7.

4. Details of the impact

Techniques developed by PRImA and their partners are of international standing and used by a number of European libraries (and beyond) and commercial service providers, they influence international digitisation policy such as the Digital Agenda for Europe in the area of digitisation, and are funded through a range of incremental funding streams:

- **2008-2012:** EU 7th Framework Programme: The Improving Access to Text (IMPACT) **project:** Supporting the EU's i2010 vision to significantly improve access to Europe's cultural heritage, the British Library, the National Library of the Netherlands and the University of Salford led a group of 15 institutions from across Europe to remove the barriers that stand in the way of the mass digitisation of the European cultural heritage.
- Libraries and archives around the world had relied on digitisation service providers whose best technologies were designed primarily for modern business documents (the service providers' largest commercial market) and were not able to take fully into account the significant challenges posed by ageing books and newspapers:
  - Antonacopoulos established the common baseline for evaluating different approaches to mass digitisation through the development of a comprehensive, large scale reference dataset with ground truth at various levels, compiled in partnership with the content-holding partners and representative of their collections. Antonacopoulos defined evaluation metrics and scenarios, and the tools to implement them. **[7]**
- The dataset, hosted by the IMPACT Centre of Competence in Digitisation, is a unique resource; for researchers to create new methods of digitisation; libraries who want to evaluate their holdings for digitisation and service providers who put together workflows of different methods to identify what works best in given scenarios to enable the objective evaluation of printed holdings. Covering texts from as early as 1500, containing material from newspapers, books, pamphlets and typewritten notes, and created, maintained and expanded by University of Salford (PRImA) researchers, the dataset forms a repository of document images reflecting the holdings and priorities of major European libraries, running to over 600,000 document images, ground truthed in 50,000 cases.
- The British Library, the National Library of the Netherlands and the PRImA research lab explored methods of improving Optical Character Recognition (OCR) for use in the digitisation of less standardised material, making a significant impact on the digitisation of historical documents, by focusing extensive research expertise to exceptional material in both breadth and volume, such as the collections in the British Library. This collaboration increased resource discovery success for historic mass digitisation, maximised production efficiency, quality and volume, whilst significantly reducing the costs of digitisation.
  - Aly Conteh, e-Strategy & Information Systems, Programme Manager, British Library said: *"It is absolutely vital institutions like the British Library, the National Library of the Netherlands and technical experts like the University of Salford work together, sharing our experiences and resolving the challenges we face in digitising historic texts to ensure that we deliver digital resources, which are sustainable."*
- **2011-onwards:** The National Library of the Netherlands utilised Aletheia to determine the expected quality in results in large scale digitisation of their holdings. Based on these findings they used Aletheia to create a specification for large digitisation contracts to service providers, ensuring high quality and cost-effectiveness (more content digitised).
- Novel features of Aletheia include the support of top-down ground truthing with sophisticated split and shrink tools as well as bottom-up ground truthing, supporting the aggregation of lower-level elements to more complex structures. Special features have been developed to support working with the complexities of historical documents.

International commercial service providers have taken up the Aletheia methodology using the technology to implement institutions' digitisation plans, improve their effectiveness and lower their digitisation costs. The technology is openly available for reuse via the PRImA website.

- **2012:** The Wellcome Library commissioned Antonacopoulos to evaluate what current OCR methods could achieve in the digitisation of the their archives, including papers of Francis Crick, who discovered (with James Watson) the double helix of DNA, and helped to crack its code. In question was a combination of typewritten documents, notes, and, for example, versions of Francis Crick's seminal paper demonstrating the construction of DNA. Antonacopoulos evaluated which types of document would yield high quality text after digitisation and utilised Aletheia to ground truth the documents to test the required level of accuracy. During the pilot phase, the Library planned to add over 1 million images of archives to the Wellcome Library website. The evaluation results helped Wellcome assess its digitisation strategy as well as the prioritisation for digitisation of certain document types over others.
  - "*Regarding our archival collections, there is a wide range of content that is theoretically OCR'able. To find out we commissioned the University of Salford PRImA (Pattern Recognition and Image Analysis Research) to do a benchmarking exercise from which we could determine whether we could rely on raw OCR outputs, should not OCR this type of material at all, or to test various methods to improve OCR'ability.*" Dr. Christy Henshaw**,** Digitisation Programme Manager, Wellcome Library, Wellcome Trust, Wellcome Library
- **2012 onwards:** EU 7th Framework Programme: The Europeana Newspapers: A Gateway to European Newspapers Online project, with the objective of the provision of more than 18 million newspaper pages online, is undertaking the aggregation and refinement of newspapers through the European Library.
  - Antonacopoulos leads on quality assessment of the refinement performed on the digitised newspaper pages, developing state-of-the-art methods that have drastically improved the search and retrieval possibilities in digitised historical newspaper content. **[8]**
- **2012 onwards:** The EMOP: Early Modern Optical Character Recognition (OCR) project aims to preserve and improve access to literary cultural heritage by using innovative applications of OCR technology and crowd-sourced corrections.
  - Antonacopoulos leads on the development of a web-based system for crowd-sourced editing of textline segmentation and entity labelling, and further development of Aletheia for use in font-training for improved OCR systems. **[9]**
- **2013:** The National Library of Slovenia is trialling Aletheia in the creation of a dictionary of old Slovenian and the Dutch Institute of Lexicology has commissioned Antonacopoulos to create an historical dictionary utilising Aletheia to extract terms from relevant texts for the dictionary. The National Library of Turkey has commissioned a case study focused on how OCR might perform on Ottoman scripts, utilising Aletheia and the evaluation tools to produce the report whether the digitisation of this material can proceed.
- **2013 onwards:** The SUCCEED: Support Action Centre of Competence in Digitisation (FP7) project, supports the European Commission in executing the Digital Agenda for Europe in the area of digitization:
  - Antonacopoulos leads on evaluation, to assess on behalf of, and inform the European Commission of the effectiveness of, individual software methods as well as complete digitisation workflows in a variety of digitisation application scenarios. **[10]**

5. Sources to corroborate the impact

a) Corroboration of resolving challenges in digitising historic texts, Digitisation Programme Manager, Wellcome Library, Wellcome Trust

b) Corroboration of OCR project, e-Strategy & Information Systems, Programme Manager, British Library