Digital Single Market

Projects news and results29/06/2012

From the printed page to bits: new tools for mass digitisation

EU-funded research has developed a suite of automated text recognition and processing tools that improve the fidelity and searchability of digitised texts from museum and library archives. 'These days anything that is not digital is not visible,' states Hildelies Balk, head of European Projects at Koninklijke Bibliotheek in The Hague, Netherlands. 'For libraries and national archives this problem is even more pronounced today than before because most people only look at the internet now. If something is not online then they presume it is not available. So national libraries, archives and museums now have an obligation to make everything available electronically. We need to scan and digitise books, documents and printed materials en masse as quickly and as accurately as we can.'

 [1]

The digitisation process is relatively straightforward. First you scan a document to create an image of the page - and this is where the process stopped in the earliest days of digitisation. Today, however, the scanned image is then processed, typically using 'Optical character recognition' (OCR) software to extract the text into a digital format. Once the text is digitised this way it makes the entire document available for indexing and accessible to search engines.

The searchability of historical texts suddenly transforms collections into a powerful cultural resource. Previously you had to go to a specific institution and look for a particular document. Today a quick keyword search, for example, can pull back thousands of documents; you can identify a huge volume of important sources with no prior knowledge at all.

Get the picture?

But is this conversion from printed words to machine readable text sufficiently accurate to be able to trust search results? 'We wanted to improve or create new tools downstream of the actually scanning which would reduce the errors created by OCR,' explains Dr Balk. 'This mass digitisation is generating an immense resource and in the near future I think will see a proliferation of applications which exploit and even monetise the resource. But we have to be confident that the digital version of a historical text is a true copy of the original.'

For the last four and a half years Dr Balk has coordinated the FP7 'Improving access to text' (Impact [2]) project. One of the main aims of the project has been to improve the accuracy and reliability of output text by developing a suite of software tools and processing modules which could be applied (sometimes in sequence) to scanned images.

Before any OCR can be applied to a scanned image it must first be 'cleaned up. The University of Salford in the UK, the National Centre for Scientific Research 'Demokritos' in Athens and the OCR technology specialist ABBYY, based in Moscow, worked on a variety of image processing algorithms which could analyse and adjust the scanned image. One tool [3] looks at the alignment of characters on the page and straightens out lines of text which have become skewed, perhaps because they were near the spine of a book. Another algorithm [4] can remove the random appearance of black and white pixels (known as 'salt and pepper' noise) which frequently occur in scanned images.

A likely character

The project looked at various options for improving OCR results. One important area of collaboration was a close partnership with the OCR software developer and vendor ABBYY. 'We chose to work with this company because its OCR software is so widely used across Europe by libraries for digitisation,' says Dr Balk. 'ABBYY opened up its software development kit to us and worked in close partnership to integrate our research into its software. It has been great to see our research go into the improvement of a product which is already in use.'

'We weren’t interested in improving the OCR per se,' explains Dr Balk, 'as this is reasonably well developed, but the nature of historical texts can sometimes make the OCR less accurate. We wanted to develop tools that would take this historical context into account.'

For example, historical documents often have complicated layouts, with multiple columns and drop capitals. They also often use different typefaces that are not encountered in modern materials. The Impact project generated a set (known as a corpus) of 50 000 digital transcripts drawn from a set of more than half a million scanned pages from several European national libraries. These so-called 'ground truths' which are confirmed to be nearly perfect transcriptions can be used to 'train' OCR software to recognise new typefaces or cope with unusual page layouts, and also to test applications for their results.

The project also produces historical dictionaries which OCR software can use to improve its transcriptions. As the OCR works through a scanned image it puts the characters it recognised together to form 'words', then checks that the words actually exist; if they don’t then the software will typically second-guess the words by finding those with closely matching spellings.

But most OCR software will use modern dictionaries with modern words. 'Researchers want to read the actual content of documents, with the original spellings,' says Dr Balk, 'but for searching the document you will not want to look for 10, in some cases over 50, different spellings of one word. We

have compiled dictionaries [5] of arcane words for nine languages and spellings and mapped them to modern synonyms and spellings. This way the OCR will be able to transcribe a document word for word, but it will also be possible to use the dictionary to convert to modern spellings too. The dictionary helps to make digitisation more accurate, but also more flexible and useable too.'

The human touch

With mass digitisation it is important that these tools work automatically - the millions of pages requiring digitisation make it impossible for people to check transcriptions for accuracy. Nevertheless, the project has developed novel technologies that will allow users to verify OCR output quickly and easily.

Computational linguists at the University of Munich worked on an algorithm [6] that is able to identify the likelihood that words in the OCR transcription are correct or not. The algorithm takes account of the time period and original language of the document and information on established patterns for spelling and historical linguistics. From this it can identify whether misspelled words, for example, are likely to be OCR errors (which will be highlighted) or valid historical spelling variants.

Scientists from IBM Israel Science and Technology developed another system that combines a novel approach to OCR. This 'adaptive OCR' [7] called CONCERT [8] adds a clever collaborative correction system that encourages volunteer involvement to improve the accuracy of the automated OCR output through human error correction.

'Impact has produced a suite of tools and partners are currently testing them to evaluate their impact on the accuracy and fidelity of transcription,' notes Clemens Neudecker, technical manager of European projects at Koninklijke Bibliotheek. 'We want to assess how much they individually improve output, but also their impact when they are combined in a chain of post-scanning processing. We are also making sure that all these tools are interoperable by publishing a technology architectural framework [9] so that libraries can use the tools and process digitised documents without having to worry about formats and file conversions.'

The project is due to end in June 2012, but the collective expertise of the partners and their experience of using and developing digitisation tools is now being opened up to the mass digitisation community through the Impact Centre of Competence [10].

The IMPACT project received EUR 12.1 million (of total EUR 17.1 project budget) in research funding from the EU's Seventh Framework Programme (FP7) under the ICT theme.

Useful Links:

- 'Improving access to text' project website [11]
- IMPACT project factsheet on CORDIS [12]
- Impact Centre of Competence [10]
- ICT Challenge 4: Digital libraries and content [13]
- Europeana [14]

Related Articles:

- Feature Stories - Digitising our cultural heritage [15]

Country: NETHERLANDS

Information Source: Hildelies Balk, Koninklijke Bibliotheek, The Hague, the Netherlands

Share this page

Source URL: https://ec.europa.eu/digital-single-market/en/news/printed-page-bits-new-tools-mass-digitisation

Links
[1] https://ec.europa.eu/digital-single-market/sites/digital-agenda/files/newsroom/offer_id_85383_3000.jpg
[2] http://www.impact-project.eu/home/
[3] http://www.digitisation.eu/tools/image-enhancement-toolkit/geometric-correction-arbitrary-warping/
[4] http://www.digitisation.eu/tools/image-enhancement-toolkit/border-detection-and-removal/
[5] http://www.digitisation.eu/tools/language-resources/
[6] http://www.digitisation.eu/tools/ocr-post-correction-and-enrichment/text-and-error-profiler/
[7] http://www.digitisation.eu/tools/ocr-engines/ibm-adaptive-ocr-engine/
[8] http://www.digitisation.eu/tools/ocr-post-correction-and-enrichment/collaborative-correction-platform/
[9] http://www.digitisation.eu/tools/interoperability-framework/
[10] http://www.digitisation.eu/
[11] http://www.impact-project.eu/
[12] http://cordis.europa.eu/projects/rcn/85383_en.html
[13] http://cordis.europa.eu/fp7/ict/programme/challenge4_en.html
[14] http://www.europeana.eu/portal/
[15] http://cordis.europa.eu/fetch?CALLER=OFFR_TM_EN&ACTION=D&DOC=2&CAT=OFFR&QUERY=0137ffcea2f0:a4a4:20842f59&RCN=8696