The Current State-of-art in Newspaper Digitization

A Market Perspective

"*Four hostile newspapers are more to be feared than a thousand bayonets*."
Napoleon Bonaparte (1769-1821)

Market parties

The KB received 14 survey responses, and the respondents ranged in size from small private businesses (annual turnover approximately 250,000 Euros) to large multinational companies (annual turnover approximately 1,200,000,000 Euros). All of the large and mid-size companies have offshore facilities in low-wage countries such as India, the Philippines and Angola. These offshore production units are often involved in the digitization of microfilm, OCR enhancement (manual correction of automatically generated texts), rekeying and segmentation (identifying individual articles on newspaper pages and classifying the articles into genres such as news items, editorials, family announcements, advertisements, etc.). Some respondents have backgrounds as suppliers of ICT-services; others specialize in the digitization of cultural heritage collections or printed matter in general (newspapers, magazines, books, documents).

Many of the companies we surveyed, especially the larger ones, have dedicated R&D departments to keep track of current technology, aiming to improve their workflow and products. Most respondents indicate that they have strict in-house quality control procedures; some are ISO 9001:2000 certified, and others comply with Quality Assurance Plans or other standardized quality control procedures. A number of the companies offer web-based quality control and tracking systems that allow their customers live access to production reports. For logistics some companies use specially equipped vehicles for transportation, with GPS-traceability of the digitization objects as well as fire-and-waterproof boxes.

Although nearly all respondents have experience with large scale digitization projects, only 9 of them have been particularly involved in digitizing historical newspapers. The projects in which they were involved mainly relate to 19th and 20th century newspapers, and the size of those newspaper digitization projects varied from several thousands to 16 million pages. The scope of most of the projects was limited to several volumes of one specific newspaper title; only in few cases have respondents created databanks holding several different historical newspapers.

Almost all of the surveyed companies have extensive experience with digitizing from microfilm. Only a few of them also digitized large quantities from paper originals. Scanning from microfilm is cheaper than scanning from paper because the processing speed is much higher. The average capacity per month is estimated at a maximum of 120,000 pages in greyscale from paper originals, against one million pages in greyscale from microfilm. Other factors that may determine the processing speed are whether the source materials are scanned

in colour or greyscale and whether or not the newspapers may be removed from their binders prior to digitization.

Digital imaging

Most companies use specialized equipment for scanning from microfilm and paper originals. Sometimes this is commercially available hardware such as standard A0 or A1 flatbed scanners. Some companies use custom-made large-format scanners purposely built to digitize newspapers. To create master images the consensus approach is to scan at 300ppi. The preferred format is uncompressed lossless TIFF, although some respondents also suggest using JPEG (quality 10) or JPEG2000. Scanning from the originals is generally acknowledged to produce higher quality master images. There is some disagreement amongst the survey respondents as to whether one should scan in colour or greyscale. Scanning in colour produces a master that is closer to the original newspaper (more 'authentic') than greyscale. Also, according to some respondents colour images may lead to better OCR results, or at least provide better 'raw materials' to improve the OCR in due course. Choosing the appropriate format is also closely related to the issue of storage. A master image in TIFF format requires approximately twice as much storage space as a JPEG2000 (lossless) image and ten times as much as a JPEG (quality 10) image requires.

Frequently applied image enhancement technologies include tools for deskewing, despeckling, rotation, cropping, noise removal, balancing white backgrounds and image splitting. These tools are often used in semi-automated processes, with manual correction performed at the end. Some companies optimize images in order to improve OCR results. In their workflow they clearly distinguish between images produced for viewing and images that are specifically prepared for OCR processing. In this context the alternative of so-called hybrid PDFs is suggested. These PDFs embed different quality levels within a single file, e.g. one image optimized for the plain text and delivered as a bitonal image, and another image for the illustrations on the page, delivered in greyscale.

As the derivative for web delivery, most respondents recommend JPEG, mainly because of its efficient compression rate and zooming potential. Three respondents mention the JPEG2000 format as a suitable derivative. ISO-standard JPEG2000 is considered to be an efficient compression format because it produces relatively small files. One large digitization company strongly advises against using JPEG and – to a lesser degree – JPEG2000. It argues that in the case of bitonal and greyscale images, such as those with line-art drawings, JPEG compression can lead to low-quality images. According to this respondent, PNG is preferable to JPEG because it is presently more widely supported than the promising – but not yet generally accepted – JPEG2000. This view is supported by another respondent who believes that PNG provides the optimum compression for B&W and text 'images'. Two other respondents suggest PDF as an alternative format for derivatives. Since the majority of all users are familiar with PDF files, delivering newspaper pages or articles in PDF is a common feature of most newspaper web delivery systems.

Optical Character Recognition (OCR)

In order to make the newspaper pages searchable, the digital images of the pages must be transformed into machine-readable texts. Generating OCR text from the images is done with standard commercial software packages, custom-made software modules, or - to get the best of both worlds - a combination of the two. Some of the companies included in the market

research offer post-processing technology to improve the raw OCR, for instance by integrating existing lexicons, terminology lists or dictionaries.

The quality of OCR largely depends on the condition of the original newspaper. If a newspaper is damaged by tears, speckles or stains, if it is badly printed, if bleed-through of print from recto to verso of pages occurs, if it consists of many different fonts or typesets, the output of the OCR is likely to be affected. Also, if the individual volumes of a newspaper are bundled up in a binder, gutter shadow can lead to very disappointing OCR results (see Figure 2 below).


TWEEDE BLAD

BINNENLAND

Uit het Buitenland

Fransche Partij-verhoudingen. – Wat Oorlogsschepen kosten. – De militaire diensttijd in België

VOETBAL

R. K. Federatie.

TIJDSCHRIFTEN

Katholiek Sociaal Weekblad.

Voor den Middenstand.

Figure 2: An example of 'gutter shadow' caused by newspapers that were too tightly bound.

Accuracy rates, on either word or character level, should not be considered as watertight performance indicators for OCR software. Usually the quality of the OCR texts says more about the condition of the original materials than it does about the performance of the OCR software. For what it is worth, the rates respondents gave for newspaper digitization projects vary from 99.8% for 700,000 newspaper pages (word accuracy, manually corrected) to 68% (character accuracy, no correction) for 350,000 pages of early 20th century newspapers. Accuracy rates for historical texts from books or documents are generally higher than for texts from newspapers; this is not surprising when one considers the specific characteristics of newspapers. OCR results for 17th and 18th century newspapers (various fonts, different typesets, Old Dutch, etc.) can be expected to be far below the averages of more recent materials.

Because of the diversity of source materials, the quality of the uncorrected OCR can fluctuate from case to case. Some suppliers suggest that it may be desirable to manually correct or re-key those parts on a newspaper page that are most likely to contain the most important information such as headlines or the first few lines of paragraphs.

Zoning and segmentation

The quality of the machine-readable text of a newspaper page can be improved if individual text blocks are identified as such before the OCR is done. One respondent, a specialist in 'zoning' and 'segmentation' technology, distinguishes between three sequential phases: zoning, OCR and segmentation. In the 'zoning' phase the page is analyzed in order to identify all elements on a page, such as horizontal and vertical lines, text blocks and illustrations. For each element the group characteristics are defined and recorded. The final result of the zoning phase is a rough structural definition of the original page composition. The next step in the process is performing the OCR process of the text regions. During this process the position of every word and character is recorded in what is called 'bounding box' coordinates. In the final 'segmentation' phase the results of the layout analysis and the OCR are merged in order to distinguish between page objects such as articles, illustrations or advertisements.

The zoning, OCR (with word/character coordinates) and segmentation processes provide the 'raw materials' for searching at the article level, classifying 'blocks' on the newspaper page and highlighting hit-terms (see Figure 3 below). Many of the surveyed companies are involved in developing zoning and segmentation techniques. Some offer the whole process from digitization to segmentation and presentation as a package deal. Other companies have a modular approach; they deliver XML-based, segmented newspaper pages and offer the use of their presentation and search systems as options.

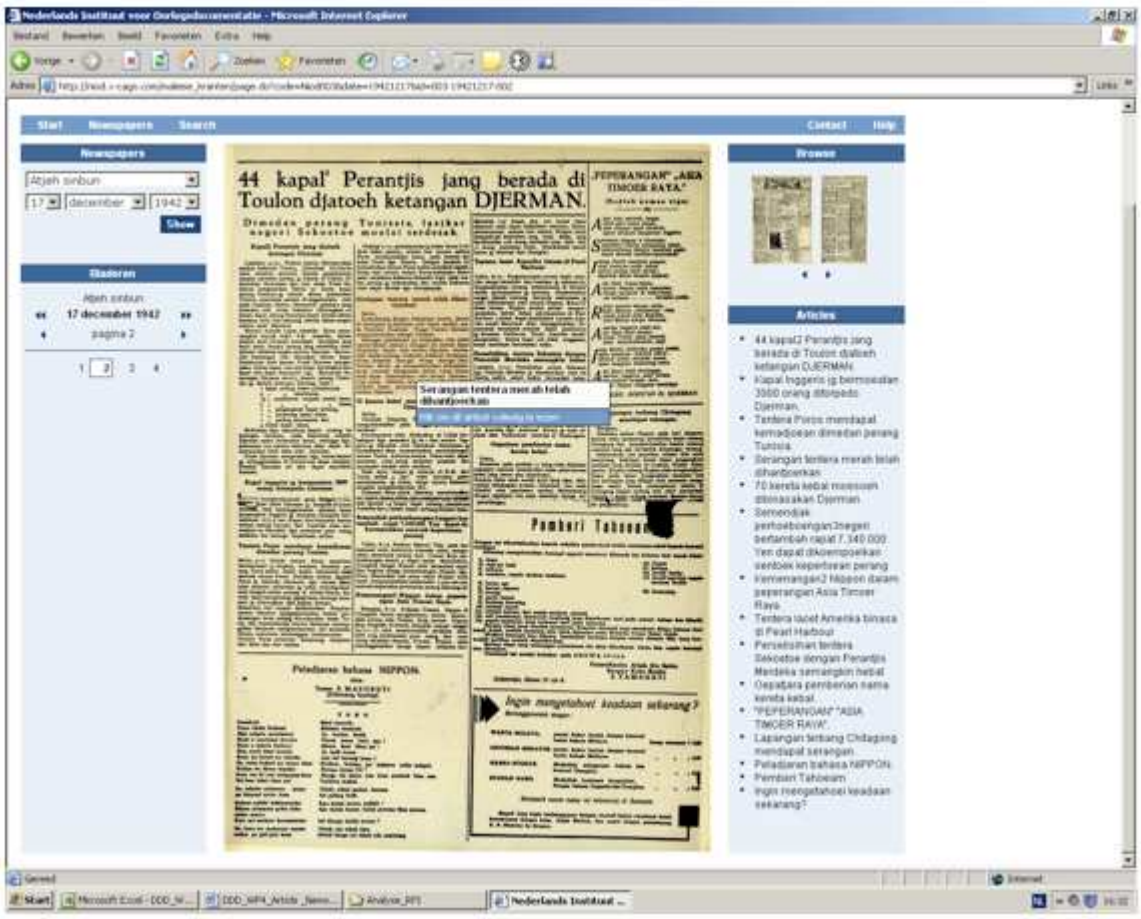

Figure 3: An example of a segmented newspaper page. The orange glow indicates a

highlighted article. On the right there is a list of article headlines (source: http://niod.x-cago.com/maleise_kranten).

Most respondents that apply segmentation techniques indicate that this is a semi-automated process. To a large extent, results of automatic segmentation depend on the consistency of the layout of the specific newspaper. The more predictable the layout, the better the automated segmentation. A manual check of the results is often necessary. Articles that have text appearing on different pages of the newspaper (so-called 'continuation' articles) need to be tailed manually. Segmenting a newspaper at article-level can be time consuming; the processing speed is estimated by one respondent at approximately 100 pages per hour. At that rate, to segment all 8 million newspaper pages for the DDD project to page-level would take one employee about 50 years!

For zoning and segmentation about half of all survey respondents use the ALTO-format. ALTO (Analyzed Layout and Text Object) is a standardized XML format for storing layout and content information.[5] It is currently used in large newspaper digitization projects such as the ones at the US Library of Congress,[6] the National Library of Australia[7] and the Bibliothèque nacional de France.

All companies in our market research have engaged in projects that aim to make newspapers accessible at the article level. In order to refine the search results at that level, manually correcting and/or rekeying the article headlines and/or the first lines of a paragraph is recommended.

Metadata

Some advanced segmentation techniques can automatically recognize and capture article headlines, page numbers and publication dates. The initial results after automated segmentation are largely determined by the level of irregularity in the layout. Nearly all respondents are able to provide basic metadata such as newspaper title, issue, page, article headline, etc. They support export of these elements to Dublin Core, METS, NEWSML[8] and custom-made schemas. Extraction of metadata from newspaper pages is often a semi-automated process. The regularities of a layout structure are recorded in a 'profile' that provides input to the software used for retrieval. Although the results from automated extraction can sometimes be quite satisfactory, to reach an acceptable level of accuracy manual correction and fine-tuning of the profile is often required.

Searchability

Many companies offer their own search engine modules. Some respondents provide hosting services to enable customers to remotely use these search engines for the purpose of their own projects. By creating different 'skins' every customer can get his own 'look and feel'. The search engines offered by the surveyed respondents for the most part are integrated within general document management systems.

Access to a data collection can be improved by increasing the 'intelligence' of the search engines. This can be done, for instance, by adding classification terms to the articles and/or pages (e.g. era, geographical location and named entities). In addition, articles can be divided into genres, such as news items, classified ads and editorials. However, if the OCR produces low-quality output, ways to improve access at an acceptable level are usually quite limited. In

the case of the KB 'Historical Newspapers pilot project'[9] applying 'fuzzy search' techniques led to an increase of word accuracy by a mere 11%. This means that 20 out of every 100 words (80% word accuracy) still remain incorrect and – as a consequence – irretrievable.

Apart from common operator searches (AND, OR, NEAR, etc.), current search technology offers a wide range of other possibilities. Respondents mentioned among others:

- user query-generated intelligence (a relevance indicator that takes into account previous searches),
- conceptual matching (relevance in relation to other similar patterns),
- data mining (retrieving logical contextual relationships and analogies in existing data and using this information to analyze other data),
- spelling variants and word correction (by input of historical lexicons),
- phonetic searches ('Soundex algorithms'),
- linguistic modules for analyzing grammatical variants and synonyms, e.g. by way of stemming (reducing words to their stem, which enables terms to be recognized regardless of grammatical variants) and
- auto-summarization.

Presentation

Search engines are often integrated into delivery modules. Many companies offer custom-made solutions for the presentation of newspapers. These are often proprietary and unavailable as separate modules. The University of California[10] and the National Library of Australia are currently developing article-level newspaper delivery systems based on open source software.

Basic features of currently available newspaper delivery systems include:

- zooming functionality
- hit-term highlighting
- segment highlighting, with a click option to display individual blocks separately

(Users may switch to different modes (image/ocr-text/pdf) according to their own preference.)

- browse-navigation from page to page

Some systems also facilitate:

- image-rotation by users
- a 'shopping basket' to store images from previous searches
- tool tips that display the headlines and the first lines of an article
- a table of contents with the article headlines per page

The user interfaces of current newspaper delivery systems sometimes embed standard viewers like Flashpaper or Adobe Acrobat Reader. Users need to have the right plug-ins to view the pages in these applications. Some survey respondents have developed their own interfaces, usually programmed in Java or Javascript/XHTML. These web applications are usually based on client-server interaction, which - especially with large files - can sometimes

limit the speed of delivery considerably. One respondent recommends using AJAX, a server-side scripting language that does not require a full page-reload when a user performs a specific action.[11] AJAX is intended to increase speed and interactivity and can be used, e.g., for fast navigation or zooming.

The performance of the server(s) can be one of the bottlenecks of newspaper delivery systems. The huge number of images and server-intensive actions, like zooming and – sometimes – retrieval of images at article-level, demands a powerful server environment, especially for sites that are heavily used. For the 8 million pages to be digitized by the DDD project, it is estimated that approximately 4 terabyte of server space will be required. One survey respondent advises building a prototype of the presentation system and performing stress tests at a very early phase of the DDD project in order to avoid unpleasant surprises.

Conclusions

The market research completed by the KB DDD project provides a bird's eye overview of current state-of-the-art technology in the field of newspaper digitization. Market trends reflect shifting goals on the side of those that produce the requirements, i.e. the cultural heritage institutions. There is a growing tendency to open up newspaper collections at the article level. Zoning and segmentation technologies provide article-level access. Also, in an increasing number of newspaper digitization projects, the digitized articles are classified into specific genres such as news items, classified ads, editorials, etc. These technologies can be very labour-intensive, as they are usually semi-automated processes. The accuracy of OCR technology is improving, but – especially for historical texts – there is still a lot that needs to be done.

The challenge for the DDD project – and in general all newspaper digitization projects – is to find a balance between accessibility and feasibility within the limitations of available resources. Our market research provided useful input for the 'Invitation to Tender' that the KB published in November 2007. At the start of 2008, the first newspapers in the project are scheduled to be digitized. As the project moves from theory into practice, quantity and quality levels will be put to the test in a real-life situation. American golfer Sam Snead once said, 'Practice puts brains in your muscles.' With the digitization of 8 million pages in store for the DDD project, we are bound to collect a lot of brainpower. We will probably need it too.

Some examples of newspaper digitization projects

- Australian Newspapers Digitization project, URL: http://www.nla.gov.au/ndp/.
- Chronicling America, URL: http://www.loc.gov/chroniclingamerica/.
- Colorado's Historic Newspaper Collection, URL: http://www.coloradohistoricnewspapers.org.
- Indonesian Newspaper Project, URL: http://niod.x-cago.com/maleise_kranten/.
- Leeuwarder Courant archive (in Dutch), URL: http://www.archiefleeuwardercourant.nl/.
- Newspaper Digitization Project British Library 1800-1900, URL: http://www.bl.uk/collections/britishnewspapers1800to1900.html.
- The Scotsman, URL: http://archive.scotsman.com/.

Notes

1. S. Puglia and E. Rhodes, 'Digital Imaging - How Far Have We Come and What Still Needs to be Done?' in: *RLG Diginews* 15 April 2007, vol. 11, issue 1, URL: <http://www.rlg.org/en/page.php?Page_ID=21033#article2>.

2. IFLA newspaper section, URL: <http://www.ifla.org/VII/s39/>.

3. Project website at URL: <http://www.kb.nl/projectnewspapers>.

4. See original Request for Information at URL: <http://www.kb.nl/hrd/digi/ddd/Request for Information.pdf>.

5. ALTO (Analyzed Layout and Text Object) homepage, see URL: <http://www.ccs-gmbh.com/alto/general.html>.

6. Chronicling America, see URL: <http://www.loc.gov/chroniclingamerica/>.

7. Australian Newspapers Digitization project, see URL: <http://www.nla.gov.au/ndp/.>

8. News Markup Language, a mark-up language for global news exchange, see URL: <http://www.newsml.org/pages/index.php>.

9. Pilot project 'Historische Kranten in Beeld', see URL: <http://kranten.kb.nl/> (in Dutch).

10. California Newspaper project, see URL: <http://www.cbsr.ucr.edu/cnp/>.

11. See for example URL: <http://atalasoft.com/ajaxviewer/>.