

Application National Programme Investments in Large-Scale Research Facilities

Administrative data

Title

Digital Databank for Newspapers

Applicant

Koninklijke Bibliotheek
P.O. Box 90407
2509 LK The Hague
The Netherlands

Contact person

H.J. Jansen
Head Research & Development
Koninklijke Bibliotheek
P.O. Box 90407
2509 LK The Hague
The Netherlands

tel. +31 (0)70 314 0413

E-mail: hans.jansen@kb.nl

Executive summary

This application describes the development and exploitation of a digital databank for newspapers as a large text corpus for scientific research, which further builds upon the existing technical and organisational infrastructure at Koninklijke Bibliotheek (KB). KB now wants to specifically use this infrastructure for the benefit of the scientific community by developing and exploiting a digital databank for newspapers as a large text corpus for scientific research.

Newspapers are an ideal candidate for mass digitisation. They contain a particularly large amount of material and fall within a clearly defined area. Newspapers form an important source for political and economic research and for research into opinions about art, literature and science. Newspapers are particularly suitable for diachronic research as they allow changes to be followed over time. Further this entire corpus is particularly suitable for language technology research.

Firstly a selection will be made of a number of major national newspapers, supplemented by several influential, longstanding regional and local newspapers. This will also take the scientific significance of a newspaper into account. A number of specific search functionalities will be developed for the service. On the basis of the aforementioned selection, the digital databank for newspapers will be filled with about 8 million pages from newspaper titles. The total costs for developing and exploiting the databank during the project period are M€ 13. The project will run for four years.

Science case

The digital databank for newspapers will form an essential new building block within the humanities knowledge infrastructure, and will offer new possibilities for innovative research. The databank will optimise the accessibility of material and increase the efficiency of the research, thereby making it possible to tackle previously unanswerable research questions. Due to its size (about 25 billion words) the databank will form a unique source for groundbreaking research in text mining and language use.

The databank is particularly important for two types of subject disciplines. Firstly subject disciplines that have always been based on investigating sources and which will benefit from scaling up and automation. Scaling up through making vast amounts of digital information accessible, might lead to breakthroughs in these areas and an increased rate of progress.

Further the digital databank for newspapers will play a major role within subjects where the methods and techniques can be further developed by applying them to a large-scale text corpus. The possibility of handling so much data could lead an enormous step forwards in the techniques for extracting knowledge and information from texts.

Talent case

Digital text corpora can easily be consulted from a different physical location. The unlimited accessibility of the digital databank for newspapers signifies an important stimulus for the research of Dutch source material by foreign scientists.

Innovation case

The digital databank for newspapers will also fulfil an important role in disseminating the cultural-scientific heritage within the wider societal context. Newspapers have a high information value and form highly attractive material for a broad group of interested parties.

Partnership case

The newspaper pilots already completed by KB have demonstrated the feasibility and potential of a digital databank for newspapers. The databank is perfectly in keeping with similar initiatives in Scandinavia and the UK and will immediately bring the Netherlands to the forefront of developments in this area. The importance of digital content is also recognised in Brussels.

A number of parties will be worked with to realise the digital databank for newspapers. KNAW and the separate research institutes will be consulted during the selection of the material and the development of specific search functionalities. KB will turn to fellow institutes who possess additional material to complete titles. Consultations with the publishing sector are another important element within this project. The inclusion of material in the databank will be discussed and negotiated with the publishers involved. Copyright will be a key issue in these talks.

Business Case

With the digital databank for newspapers, KB is further building on its existing technical and organisational infrastructure. The management and maintenance of the digital databank for newspapers will therefore become a part of the existing KB organisation. This construction forms an important guarantee for the long-term status of the service.

Technical case

A number of technical issues are associated with the mass digitisation of large text corpora. These are related to the digitisation of the material, the development of a generic and scalable infrastructure, the metadata and the rendering of the material. KB has already acquired experience with the technical aspects stated in a large number of projects.

Digital Databank for Newspapers

Introduction

In its report *Kennisambitie en Researchinfrastructuur* the Innovation Platform has indicated that additional funding measures are needed for the large-scale research infrastructure in the Netherlands. Such research facilities are of considerable societal and economic importance, are essential for the quality of research and function as a magnet for talent.¹ The report distinguishes three large scientific areas (the natural and engineering sciences, life sciences and medical sciences and the humanities and social sciences), each with its own characteristics and requirements with respect to infrastructure.

The arts and humanities are typified by the requirement for a large-scale *data* infrastructure. During the workgroup's discussions, mass digitisation of large text corpora (such as newspapers and journals) was referred to on several occasions in this context, due to the broad utility of such data collections, not only for scientific research, but also for the general public.

This application describes the development and exploitation of a digital databank for newspapers as a large text corpus for scientific research, which is a further extension of the existing technical and organisational infrastructure at the Koninklijke Bibliotheek (KB).

Description of the facility

Since the mid-1990s, KB has gained experience with digitisation in a large number of projects. As initiator and coordinator of the long-term national programmes *Metamorfoze* (national programme for the conservation of paper-based heritage) and *Het Geheugen van Nederland* (national digitisation programme) KB has built up a successful technical and organisational infrastructure for mass digitisation. The House of Representatives of the Dutch Parliament has recently commissioned KB to start on the project *Staten-Generaal Digitaal*. This project, the digitisation of all parliamentary documents since 1814 (duration 6 years, size 2.5 million pages, budget M€ 10.5), also makes use of the knowledge KB has acquired.

As a result of these programmes and this infrastructure, some 90 scientific and cultural institutions in the Netherlands and abroad (Library of Congress, The British Library) have been served in recent years. KB now wants to specifically use this infrastructure for the benefit of the scientific community by developing and exploiting a digital databank for newspapers as a large text corpus for scientific research.

The digitisation of analogue, scientific sources requires a wise selection of material. A responsible choice can be made on the basis of the following criteria:

- *Size and composition of the collection*: the sources should form a substantial collection within a clearly defined area. A large collection can also be created by virtually linking physically-separated sources;
- *Expected use*: The sources should be interesting for many potential users; not only for the scientific researchers, but also for a broad group of other interested parties.

¹ *Kennisambitie en Researchinfrastructuur. Investeren in Grootschalige Kennisinfrastructuur*. Report issued by the Innovation Platform (July 2005).

Newspapers are an ideal candidate for mass digitisation. They contain a particularly large amount of material and fall within a clearly defined area. Newspapers form an important source for political and economic research and for opinions about art, literature and science. Newspapers are particularly suitable for diachronic research, as they allow changes to be followed over the course of time. And this entire corpus is particularly suitable for linguistic research.

Newspapers are printed for use on a single day but the information value is timeless. Unfortunately newsprint is a material that is only intended for one-off use and not for conservation purposes. That is why newspapers from the depots of archives and libraries can only be made available for a limited number of years. KB owns the largest collection of newspapers in the Netherlands. Digitising these newspapers clearly falls within KB's profile as the national library of the Netherlands with a special interest in Dutch history, language and culture in a broad international context.

Over the past few centuries more than 5000 national, regional and local newspaper titles have been published in the Netherlands. Dutch archives and libraries contain tens of kilometres of material. As previously stated, newsprint cannot be preserved for a long time. The paper browns and acidifies rapidly, and eventually the majority crumbles if used. Therefore since the 1970s, KB and other institutes have been microfilming newspaper material in order to preserve its contents. KB now possesses a large number of national and regional newspapers and Dutch Indies newspapers on microfilm. However this does not lead to much improvement in the accessibility of the material, as microfilm readers are not particularly user-friendly.

The digital databank for newspapers will be developed as follows.

Firstly a number of major national newspapers will be selected, supplemented by several influential, longstanding regional and local newspapers. This will also take the scientific significance of a newspaper into account. KNAW and the separate research institutes will be consulted during this selection process.

An indispensable part of this project will be the national inventory of the available material to establish which titles are already available on microfilm or in digital form.

KB has recently launched a particularly successful newspaper pilot on the Internet, *Historische dagbladen in beeld* (see page 8), which contains 350,000 newspaper pages that can be searched in an integral manner. The specific technical environment developed for this newspaper pilot will be converted into a generic and scalable environment.

Digitisation of the newspapers will preferably be done from microfilm. Although much of the material from KB and other institutes has already been filmed, this project will need to investigate which microfilms satisfy the current quality requirements for a high quality digital conversion.

The project will be completed in phases, starting with the content of the current KB newspaper pilot. After this as much digitised material as possible will be added from elsewhere. A next phase concerns the mass digitisation of usable microfilms from KB and from other parties. Finally an assessment of the missing material will be made so that this can be collected, (filmed) and digitised.

A number of specific search functionalities will be developed for the service. The expertise of the KNAW will be used to make an inventory of the existing research questions and preferences of the scientific target group.

On the basis of the aforementioned selection, the digital databank for newspapers will be filled with about 8 million pages from national, regional and local newspaper titles. The total costs for

developing and exploiting the databank during the project period are M€ 13. The project will run for four years.

Science case

The importance of digitising sources for scientific research is evident. Research in the humanities without source materials is unthinkable. In the report *Een digitale bibliotheek voor de geesteswetenschappen* NWO argues that the knowledge contained in the sources, forms the basis for new descriptions, new explanations and new theories.²

As far back as 1997, the KNAW advisory report *De computer en het alfaonderzoek* stated that the use of digital source material would provide various benefits for research. Since then various documents have referred to the importance of digitising our scientific and cultural sources.³ Digital disclosure contributes to the efficiency, efficacy and reliability of research. Moreover it enables previously unanswerable research questions to be investigated in a systematic manner. Last year the report *E-based Humanities and E-humanities on a SURF platform* was published in which the development of large digital text corpora was highlighted as a top priority: “*E-humanities can only take off if digital research corpora exist in a sufficient number and in a sufficiently large size*”.⁴

The digital databank for newspapers could serve a large number of scientific disciplines, including political science, literature studies, history of art, sociology, economics, historical research and diachronic linguistics. Multidisciplinary research, such as text mining (interdisciplinary between informatics and linguistics) and history computing (on the boundary between history and informatics) would also greatly benefit from a mass text corpus.

The databank will optimise the accessibility of material and increase the efficiency of the research thereby making it possible to tackle previously unanswerable research questions. Due to its size (about 25 billion words) the databank will form a unique source for groundbreaking research in text mining and language use. However the databank is unlikely to lead to a major scientific breakthrough. In the humanities there is more of an emphasis on the cumulative process of knowledge acquisition. Moreover scientific breakthroughs can neither be planned nor predicted. So the focus is on creating optimal conditions for research. The digital databank for newspapers will form an essential new building block within the humanities knowledge infrastructure, and will offer new possibilities for innovative research.

The databank is particularly important for two types of subject disciplines. Firstly subject disciplines that have always been based on investigating sources and which will benefit from scaling up and automation. History, sociology, economics and cross-disciplinary areas such as economic psychology, already make extensive use of archives (including newspaper archives) in order to validate and measure trends over time, for example changing attitudes in the description of regimes and political parties. Scaling up through making vast amounts of digital information

² *Een digitale bibliotheek voor de geesteswetenschappen. Aanzet to een programme voor investering in een landelijke knowledge infrastructuur voor humanities and cultuur*. Beleidsnota ICT van het Gebiedsbestuur Geesteswetenschappen NWO (1999).

³ E.g.: *Alles uit de kast. Op weg naar een nationaal investeringsprogramma digitale infrastructuur cultureel erfgoed* (Wetenschappelijk Technische Raad SURF, 1998); *Het Instituut Nederlandse Geschiedenis en elektronische bronontsluiting*. (ING, 1998); *Digitalisering van het cultureel erfgoed*, brief van de Staatssecretaris Van der Ploeg aan de Tweede Kamer (27 mei 2002); *Wetenschapsbudget 2004. Focus op excellentie en meer waarde*.

⁴ Joost Kircz, *E-based Humanities and E-humanities on a SURF platform*. Report commissioned by SURF-DARE, p. 25 (Amsterdam 2004).

accessible, might lead to breakthroughs in these areas and an increased rate of progress. It will be possible to perform investigations and case studies faster and in greater numbers.

A large diachronic text corpus such as the digital databank for newspapers is an important precondition for answering innovative questions posed by linguists and literary researchers. Up until now these researchers have only had access to relatively small text corpora. Mainly as a result of its size, the digital databank for newspapers will provide important new research challenges. Independent of the contribution to improved accessibility, large text corpora can contribute to the development of language models for language technology. Representative text corpora are important for models in which language needs to be quantified (e.g. language models for recognition purposes) due to the need for statistical support.

The emergence of digital text files was vitally important for the development of speech recognition. Speech recognition research goes through a phase in which not only audio material is needed but also a vast amount of text material for the building up of a vocabulary.

Also for diachronic research (rates of change in language use, the adoption of loan words, etc.) a text corpus such as the digital databank for newspapers is of major importance.

The digital data bank for newspapers provides new possibilities for research into literary canon forming and author recognition. In the case of literary canon forming this mainly concerns the change of the canon over time and the media specificity of the canon forming. In the case of author recognition research, anonymous reviews and other contributions are mainly attributed to authors on the base of comparing styles.

Further the digital databank for newspapers will play a major role within subjects where the methods and techniques can be further developed by applying them to a large-scale text corpus. For example in the new application areas of text mining and information extraction, it is already known that increasing the amount of data available facilitates the knowledge and information mining process. However, in practice there are very few cases where a considerable amount of data is available from the same type of source over a considerable period of time. Smaller periods have however been widely available for some time (since news services and newspapers have existed in digital form), but the challenge is to acquire information over much longer periods, in which languages and entities (countries, people, political movements, conflicts) completely change. The need to cope with so much data (the digital databank for newspapers will contain 25 billion words) could lead to enormous progress being made in the technologies for extracting knowledge and information from texts. The digital databank for newspapers will therefore become a unique source for groundbreaking research in text mining.

This science case is supported by:

- Dr K. van Dalen-Oskam, Huygens Institute;
- Dr H. Wals, International Institute of Social History;
- Prof. J.C.H. Blom, Netherlands Institute for War Documentation;
- Prof. E.O. Postma, Institute for Knowledge and Agent Technology, Maastricht University;
- Prof. F.M.G. de Jong, Language Technology and Computer Linguistics, University of Twente;
- Dr P. Wittenburg, Max Planck Institute for Psycholinguistics, Nijmegen;
- Dr A. van den Bosch, Induction of Linguistic Knowledge (ILK) / Computational Linguistics, Tilburg University.

Talent case

Digital text corpora can easily be consulted from a different physical location. The unlimited accessibility of the digital databank for newspapers signifies an important stimulus for the research of Dutch source material by foreign scientists.

In recent years foreign researchers, for example Simon Schama and Jonathan Israel, have written a number of seminal overviews in the area of Dutch history and culture. Further institutes such as the Centre for Dutch Studies in Münster (which carries out a lot of research on Dutch history in the 19th and 20th centuries) and the Society for Netherlandic History in the United States (unites researchers and students in the area of Dutch history) will greatly benefit from a databank that can be remotely consulted.

Innovation case

Up until the mid-1990s, ICT had scarcely played a role in the dissemination and accessibility of the cultural-scientific heritage. However, since the widespread use of Internet that is no longer the case. Thanks to the Internet, heritage documentation has now gained an important new distribution medium.

Digitisation offers spectacular new user possibilities for the collections of libraries, museums and archives. With ICT, the richness and breadth of the cultural-scientific heritage can be significantly expanded.

In “*Meer dan de som. Beleidsbrief cultuur, 2004-2007*” the State Secretary for Education, Culture and Science indicated that the digitisation of heritage collections is of major importance and that it potentially has strong links with education, science, the arts industry and the knowledge economy.⁵

The digital databank for newspapers will also fulfil an important role in this broader societal context. Newspapers have a high information value and form highly attractive material for a broad group of interested parties.

It is expected that the use of newspapers will increase explosively, once the material can be simply searched via the Web. The user figures for the KB website *Historische kranten in beeld* (www.kb.nl/kranten) support this assumption (1.5 million pages requested in the first month).

Partnership case

The importance of digitising newspapers is also recognised in our neighbouring countries.

In Scandinavia four national libraries are cooperating in *TIDEN, The Nordic Digital Newspaper Library*, which contains Danish, Swedish, Norwegian and Finnish newspapers from the period 1640–1860. This databank has been available via the Web for a number of years.

In the UK the British Library started the *National Newspaper Project* last year. In this project, 2 million pages of national, regional and local newspapers from the period 1800–1900 are being digitised. With this project the British Library intends to serve the research and higher education communities as well as the general public. The databank will be made available via the Web. The project is being financed by the Joint Information Systems Committee (JISC).

⁵ *Meer dan de som. Beleidsbrief Cultuur 2004-2007* van de Staatssecretaris van OCW aan de Tweede Kamer (3 november 2003).

KB has also acquired the experience necessary for the digitisation of newspapers. In 1999 it started the project *The Roaring Twenties*, in which three national newspapers from the period 1920–1930 were digitised and disclosed to a limited extent. In the subsequent project *Oorlog en Revolutie* that was started in 2002, three national newspapers once again took centre stage, but now in the period 1910–1919. The objective of this project was more ambitious: digitisation and disclosure on the basis of full text retrieval. The results of this project have become available this year via the successful website *Historische kranten in beeld*.

Finally in 2004, KB in partnership with the Permuseum started the pilot project *Dutch newspapers 1618–1700*. This intends to make an electronic, searchable catalogue of all newspapers printed in the Netherlands between 1618 and 1700, and will include digital images of these.

The successful newspaper pilots have demonstrated the feasibility and potential of a digital databank for newspapers. The databank fits perfectly within the stated European initiatives and will immediately bring the Netherlands to the forefront of developments in this area.

The importance of digital content is also recognised in Brussels:

“The demand for quality digital content in Europe, with balanced access and user rights, by a broad community, be they citizens in society, students, researchers, SMEs and other business users, or people with special needs wishing to augment their knowledge, or “re-users”, wishing to exploit digital content resources to create services, is increasingly apparent.”⁶

Realising the digital databank for newspapers will require cooperation with a number of parties. KNAW and the separate research institutes will be consulted during the selection of the newspaper titles. KNAW’s expertise will also be used during the inventory of research questions from the scientific target group. Specific search functionalities will be developed on the basis of this.

KB will turn to fellow institutes who possess additional material to complete titles.

Consultations with the publishing sector are another important element within this project. The inclusion of material in the databank will be discussed and negotiated with the publishers involved. Copyright will be a key issue in these talks. Other aspects include the possible exclusion of recent volumes and how publishers can benefit from making digital material available.

Business Case

The setting up of a digital databank for newspapers is far beyond the means of the usual granting sources. The total development costs are M€ 12.1. The exploitation costs have been claimed for the project period of four years (2006-2009) and amount to M€ 0.9. After this period the exploitation costs will be borne by KB.

With the digital databank for newspapers, KB is further building on its existing technical and organisational infrastructure. The management and maintenance of the digital databank for newspapers will therefore become a part of the existing KB organisation. This construction forms an important guarantee for the long-term status of the service.

⁶ Decision No. 456/2005/EC of the European Parliament and of the Council of 9 March 2005, establishing a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable, preamble point 5.

Technical case

A number of technical issues are associated with the mass digitisation of large text corpora. Digitising from microfilm requires a different type of scanning than digitising from the original and also makes specific demands on the microfilms concerned. A generic and scalable infrastructure is required for the storage and accessibility of large quantities of digitised material, and it must also be possible to add to this in the future. Different search functionalities will need to be provided, including searching in the full-text version of material. In order to serve as many researchers and other potential users as possible, the option of searching in different digitised collections and catalogues simultaneously, from both KB and other institutes, will also need to be provided.

KB has already acquired experience with the technical aspects stated in a large number of projects.

- *Digitisation*

For the programme *Het Geheugen van Nederland*, guidelines were drawn up for compiling the specifications needed to ensure digitisation of a qualitatively high value. In the *Metamorfoze* programme technical requirements were also compiled that microfilms had to satisfy in order to be digitised. Experience in scanning from microfilm has been acquired in the newspaper pilot *Historische kranten in beeld*. During this pilot experience was also acquired with the conversion to full-text by means of optical character recognition, the different software packages which can be used for this and the requirements that can be made with respect to the accuracy of the character recognition. It is expected that the expertise in the area of digitisation will further increase as a result of the mass digitisation of large text corpora.

KB has established quality procedures and trained quality managers to check the quality of – mostly external – digitised images. The quality procedures and the management of large quantities of digitised material necessitate the use of software for supporting the workflow in the process; from supply to storage and making the files available. In the project *Staten-Generaal Digitaal*, in which 2.5 million pages are being digitised, such software is already being used.

- *Generic and scalable infrastructure*

The storage of 8 million digitised newspaper pages requires about 250 terabyte of storage space. Although the infrastructure of KB is suitable for the storage of large digitised collections, such a storage size requires an extension of KB's current – scalable – storage system.

The newspaper pilot has been set up in a stand-alone environment in which various search functions could be tested. For the mass digitisation of newspapers, use will be made of the generic infrastructure present at KB for the storage of digital files, indexing, rendering, search facilities and workflow. The material from the newspaper pilot can easily be transferred to this infrastructure.

- *Metadata*

Metadata will be added to make the digital databank for newspapers searchable. The metadata are stored in XML (Extensible Markup Language), a standard of the World Wide Web Consortium for the structuring of data. By making use of XML, the metadata from the newspaper pilot and from newspapers already digitised elsewhere can easily be combined with the collection yet to be digitised. It will also be simple to simultaneously search in the digital databank for newspapers and other collections and/or catalogues. The infrastructure necessary for integrated searching in different files is already present at KB.

As open standards are being used – such as XML for structuring of the full-text and the metadata and MPEG21 DIDL for structuring of the digital files – it will also be easy to exchange data with the parties.

- *Rendering*

The digital databank for newspapers will be made available on a freely accessible website with various search functions. The present search and browse options from the website *Historische kranten in beeld* are being evaluated for this purpose. This website has been built with off-the-shelf software. As such software offers few possibilities for adjustments and/or expansion with new functions, the new website will be developed in-house by KB. The technologies needed to realise that are already available.

Desired search functions that currently have to be provided through standard software, such as the highlighting of search terms in pages found, will also be developed in-house by KB. For this use can be made of the tools that were built in the project *Staten-Generaal Digitaal*.