KB Koninklijke Bibliotheek
Nationale bibliotheek van Nederland

IFLA

BnF

IFLA International Newspaper Conference

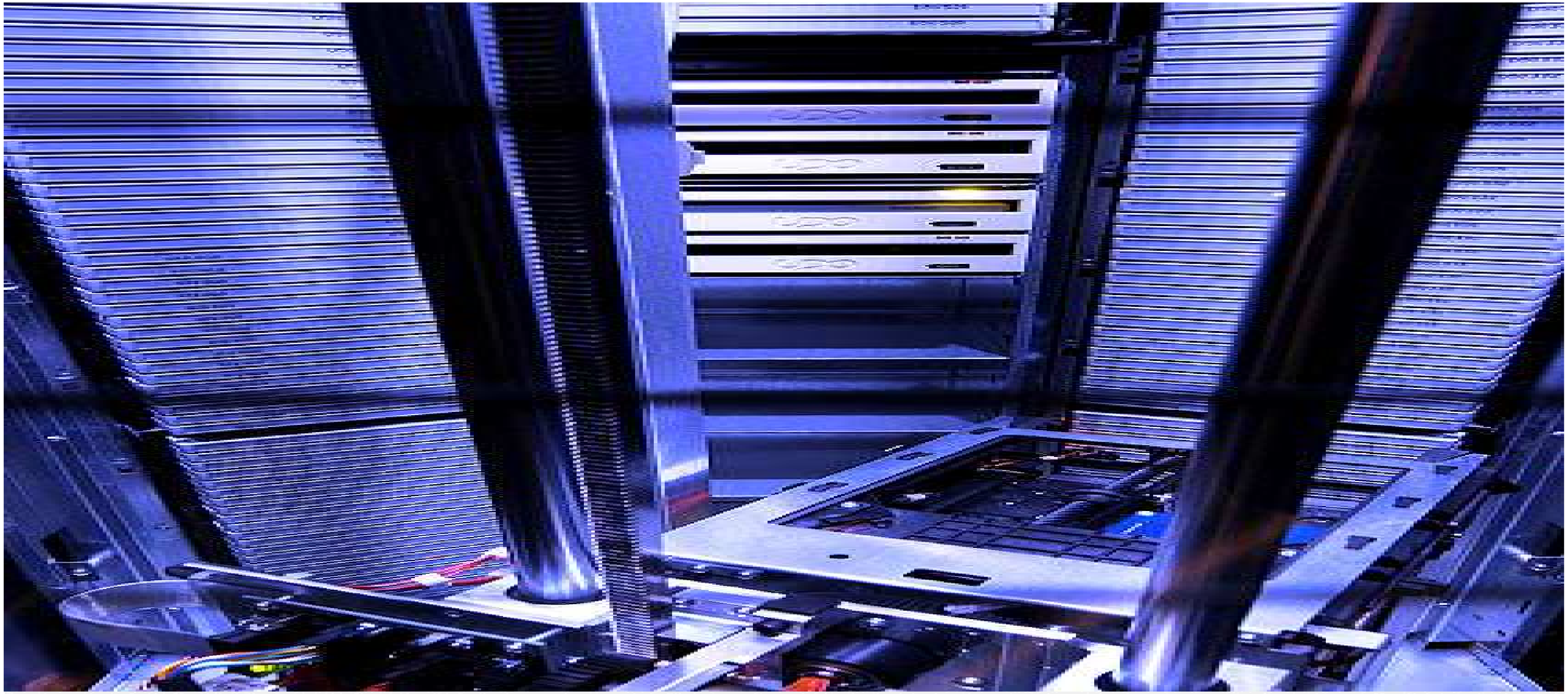
“Newspaper Digitization and Preservation. New prospects. Stakeholders, Practices, Users and Business Models”

11-13 April 2012
BnF, Paris

With the support of:

ZEUTSCHEL The Future of the Past · CCS · Isako · Bookkeeper · EUROPRESSE.COM · PLANMAN TECHNOLOGIES · i2S DigiBook · Kirtas · diadeis groupe numen

Koninklijke Bibliotheek
Nationale bibliotheek van Nederland



Post Digitization: Challenges in Managing a Dynamic Dataset

Jasper Faase, 12 April 2012

Mission

- The Koninklijke Bibliotheek is the national library of the Netherlands: we bring people and information together.
- Our core values are: accessibility, sustainability, innovation and cooperation.

Vision

We

- Offer everyone everywhere access to everything published in the Netherlands
- Promote permanent access to digital information nationally and internationally
- Play a central role in the (scientific) information infrastructure of the Netherlands

From vision to practice

- Mass-digitisation: both with public funding and in public-private partnerships (Metamorfoze, Google, Proquest)
- Between 2009 and 2013 10% of all Dutch publications (60 million pages) will be digitized. This will be sped up by designating digitization part of the library's primary process
- Target for newspaper digitization: 9 million pages by the end of 2012

Results so far

- Digitized: 7.4 million pages of Dutch national, colonial, regional and local newspapers (1618-1995)
- Contracts with 19 publishers and representative bodies of freelancers (photographers, journalists) to clear 105 titles for free online access
- 5 million pages published online via www.kranten.kb.nl; this figure will grow to 9 million pages by November 2012
- Around 50,000 unique visitors and 80,000 visits a month

Post Digitization challenges

1. Improving results of Optical Character Recognition
2. Preserving large amounts of digital objects in order to guarantee access in the long term
3. Setting up a viable model for cooperation at national level to improve accessibility and completeness of digital collections

Challenge 1: Improving OCR

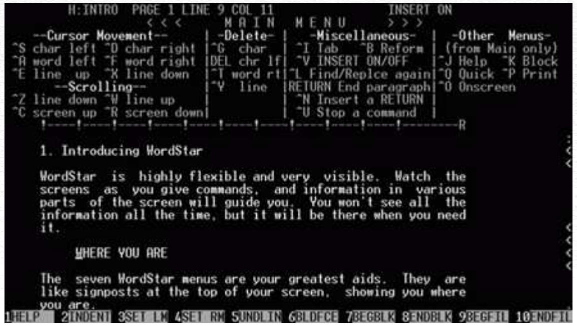
- Target groups (especially linguists) value high quality OCR
- Poor recognition rates of software means OCR of pre-1850 historical texts can often not be used for research
- Current situation at KB: OCR is a static dataset. No method in place to improve this data structurally
- Language technology is evolving quickly

Challenge 1: Improving OCR

- Develop a method for management of OCR as a dynamic dataset
- Use results of the European Project IMPACT (http://www.impact-project.eu/) within KB’s digitization workflow (2012)
- Introduce crowdsourcing to improve OCR of individual texts to a high degree of accuracy (2013)
- Implement a toolset so language technology can be used to structurally improve OCR of existing datasets (2013-2014)

Challenge 2: Digital preservation

- Digital preservation in order to guarantee permanent access
- Use of digital copies is replacing use of originals
- Permanent access is not self evident and is not the main priority for many of our partners
- Risks:



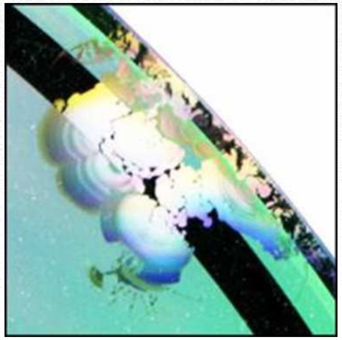
```
      H:INTRO  PAGE 1 LINE 9 COL 11                 INSERT ON
              < < <      M A I N   M E N U      > > >
  --Cursor Movement--   | -Delete- |   -Miscellaneous-   |  -Other  Menus-
^S char left ^D char right |^G  char  | ^I Tab    ^B Reform | (from Main only)
^A word left ^F word right |DEL chr lf| ^V INSERT ON/OFF    |^J Help   ^K Block
^E line  up  ^X line down  |^T word rt|^L Find/Replce again|^Q Quick ^P Print
     --Scrolling--         |^Y  line  |RETURN End paragraph|^O Onscreen
^Z line down ^W line up    |          | ^N Insert a RETURN |
^C screen up ^R screen down|          | ^U Stop a command  |
    !----!----!----!----!----!----!----!----!----!----!----!--------R

    1. Introducing WordStar

    WordStar  is  highly flexible and very  visible.  Watch  the
    screens  as  you give commands,  and information in  various
    parts  of the screen will guide you.  You won't see all  the
    information all the time, but it will be there when you need
    it.

         WHERE YOU ARE

    The  seven WordStar menus are your greatest aids.  They  are
    like signposts at the top of your screen,  showing you where
    you are.
1HELP   2INDENT 3SET LM 4SET RM 5UNDLIN 6BLDFCE 7BEGBLK 8ENDBLK 9BEGFIL 10ENDFIL
```



Challenge 2: Digital preservation

- At the moment KB is only preserving born digital data in its E-Depot
- Extending scope to digitized collections will mean building a new scalable E-Depot for long-term & safe storage for over 500 TB of data
- This new depot has to provide the library with one infrastructure for processing, access and storage

Safe storage

Click to enlarge

+

Preservation metadata

+

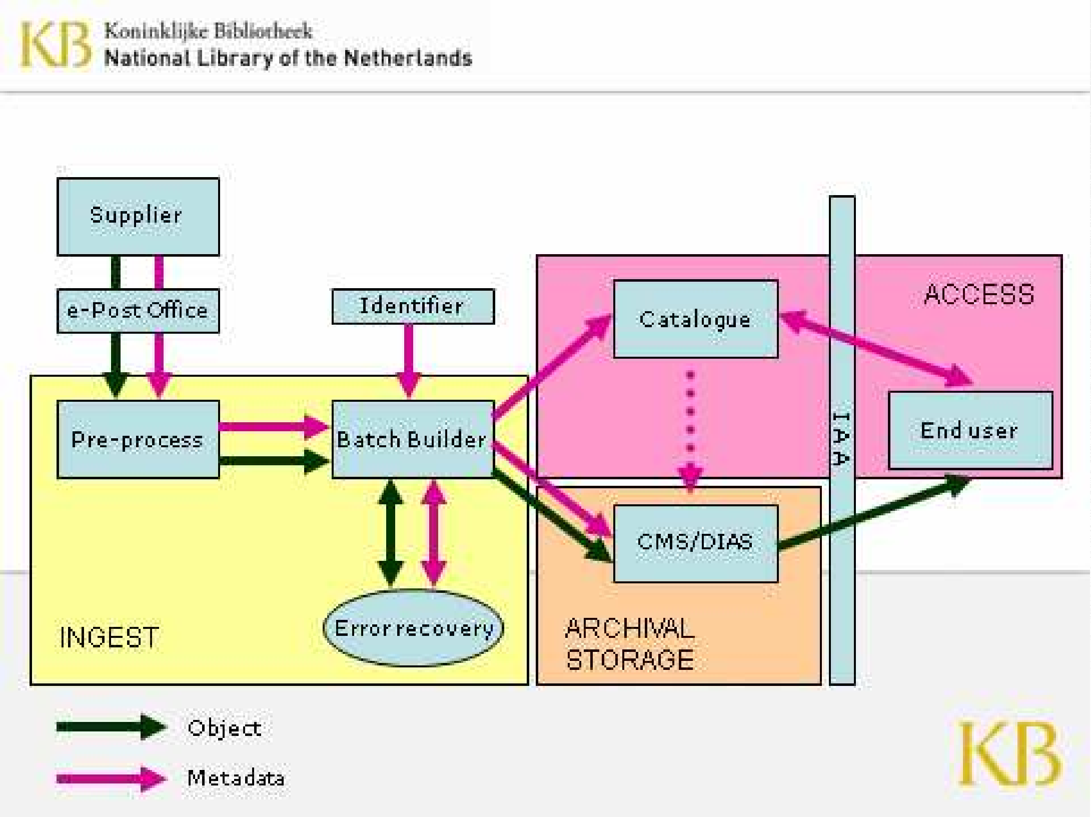
Permanent access

Challenge 2: Digital preservation

- Requirements for the new system:
  - Can be scaled up to deal with the enormous growth of digital collections
  - Able to deal with a growing diversity of digital collections
  - New functionality: tools for characterisation, format-conversion and other preservation functionality now available

Challenge 2: Digital preservation

- Requirement for the library: Preservation and data management at the core of library processes



Challenge 3: Cooperation

- Fragmented access to digital publications; especially digitized newspapers
- Lack of standardization and limited use of metedatastandards (such as ALTO and protocols for sharing data OAI and SRU)
- Lack of coordination of digitization efforts

Challenge 3: Cooperation

- Centralizing mass-digitisation of books, magazines and newspapers - financed by Metamorfoze (National Programme for Preservation of Cultural Heritage)
- Starting a joint project between academic libraries and KB to develop a platform for digitized publications, including newspapers
- Developing a model for cooperating to provide permanent access to digitized heritage from libraries and newspaper archives
- And the biggest internal challenge: merging the National Library and the National Archive

Thanks for your attention

Jasper.Faase@kb.nl