# Digital resources in Nasjonalbiblioteket

Kristin Bakken

# NB’s digitization program

- The total of NB’s collection is being digitized
- Our digitization program includes alle media types (books, newspapers, periodicals, manuscripts, photo, prints, posters, maps, music, film, radio, television and web content)

- We have established a digital long-term repository in the mountain in Mo i Rana
- We are also a repository for other institutions, i.e. NRK, Aftenposten, Adresseavisen, publishing companies

# The digitization program - status

- We have digitized
  - ca. 80.000 books
  - ca. 300.000 hours of radio broadcast
  - ca. 400.000 photos
- The digitization capacity is growing
- So far 17 different digitization routes/production lines have been established for equally many media types and formats

# Digitized text

- NB photoscan and produce ocr-text
- We store both image and text
- We reuse metadata from our catalogues (f.ex. BIBSYS)
- We display facsimili pages on internet, but base the searches on the ocr-version
- We don’t proof-read the ocr-version
- We receive digital text directly from such text producers, i.e. Aftenposten, Adresseavisen + publishing houses

# The digital repository

- A physical storage infrastructure (computer room, storage medium, storage and computing network, servers)
- Software that hides the physical infrastructure from the applications
- All documents have a unic id (URN)
- All data are stored in three copies that are physically isolated from each other – one copy on disc and two copies on tape in two different tape robots
- Integrity is secured by a system of check sums

- The system supports migration between technology generations and conversion to different formats
- – but the processes still need manual initiation and surveillance
- Our storage capacity today is one petabyte, that is 1000 terrabyte (multiplied by three)
- NB participates in the research project LongRec

# NB and Språkbanken/Clarin

- NB has large amounts of digitized data and a growing capacity for converting data
- Digital sources forward research
- Our data could be used for development regarding
  - Tools for retrieval
  - Methodology for automatic classification and cataloging
  - Tools for ocr-processing and document structure analysis
  - Tools and methodology for speech processing

- *And* for development of linguistic research bases, for instance corpora
- NB can offer digital long-term storage facilities for language data