User collaboration for improving access to historical texts

Clemens Neudecker, Koninklijke Bibliotheek
Asaf Tzadok, IBM Israel

30/06/2010 - Liber 2010, Aarhus
Background

- Text that is not digital is virtually invisible
- Digitised material is becoming available too slowly, in too small quantities and from too few sources
- OCR (Optical Character Recognition) does not produce satisfactory results for historical documents
- There is a lack of institutional knowledge and expertise which causes inefficiency and “re-inventing the wheel”
IMPACT – Improving access to text

- Funded by the EC as part of the 7th Framework Programme
- Coordinated by KB – National Library of the Netherlands
- EU funding: €12 100 000
- 26 partners: Libraries, Universities, Industry Partners
- Start date: 1 January 2008
- Duration: 48 Months
IMPACT – Aims and objectives

- Significantly improve mass digitisation of historical printed text by:
  - Innovating OCR software and language technology
  - Sharing expertise and building capacity across Europe
  - Ensuring that tools and services will be sustained after the end of the project
OCR Challenges in historical documents
...bad OCR results
KB Digital Library Programme

- **Goal:** Offer everyone access to everything published in and about the Netherlands through the internet

- **2013:** 10% of the publications published in and about the Netherlands available in digital form

- **Example projects:**
  - Historical Newspapers – [http://kranten.kb.nl](http://kranten.kb.nl)
  - Dutch Print Online – [http://www.dutchprintonline.nl/](http://www.dutchprintonline.nl/)

- **Timeframe covered:** 1618 - 1995
Crowd sourcing OCR correction

- Libraries cannot afford correction of OCR results on a scale of millions of pages
- Users expect highly accurate text for search
- Involve volunteers in correction of bad OCR
- Promising examples:
  - Distributed Proofreaders
  - Australian Newspaper Digitisation Project
  - Wikisource
  and many more!
Crowd sourcing example: PGDP

1. Sally Ann’s Experience
2. The New Organ
3. Aunt Jane’s Album
4. “Sweet Day of Rest”
5. Milly Baker’s Boy
6. The Baptising at Kittle Creek
7. How Sam Amos Rode in the Tournament
8. Mary Andrews’ Dinner Party
9. The Gardens of Memory
Crowd sourcing example: ANDP
IMPACT Collaborative Correction

- Web-based application to improve overall efficiency of collaborative correction by facilitating manual work for several tasks:
  1. Correction by means of “carpets”
  2. Font training of Adaptive OCR engine
  3. Enrichment of dictionaries
  → Higher productivity, higher motivation

- System flow:
Character session
Word session
Page session
Summary

- Mass digitisation of historical documents constitutes an enormous challenge not to be completely solved even by sophisticated OCR software

- Promising potential can be discovered in involving volunteers for OCR correction, as several examples show

- Current tools are not efficient enough for mass digitisation

- IMPACT Collaborative Correction provides a highly productive and motivating approach
Thank you!

http://www.kb.nl/

http://www.impact-project.eu/

@impactocr

http://impactocr.wordpress.com/