IMPACT Centre of Competence

Text digitisation Faster, Better, Cheaper

Hildelies Balk and Clemens Neudecker, KB

Workshop Recent Developments in OCR for Digital Libraries, Rouen 31 March 2011
Overview of this presentation

- Challenges to be addressed
- The IMPACT Project and Outcomes
- The IMPACT Centre of Competence
- Get involved now!

Twitter: @impactocr, #impactproject
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
  - Dutch Parliamentary Papers – http://www.statengeneraaldigitaal.nl/
  - Dutch Print Online – http://www.dutchprintonline.nl/

- Timeframe covered: 1618 - 1995
Historical text: typical OCR results

DJgn i f paffato te S' aö'Jifeert mo?
uen/bah )etgi'uoetbciraetail)i.r/JtmelchontDecht te /
sbnbe bele btr felbrr geiufttceert baer bnber eeniglje jprant o^fen/bie ftcb .met
beSpaenfcbeu enbeeemgljen bifet Cbeiiupcen berbonbru befe
IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands.

OCR Challenges: damaged pages, bleed through, difficult layout, historic fonts ... and many more
Language Challenges: Spelling variants, orthographical variants, inflected forms...and more

Historical variants of the Dutch word ‘wereld’ (world):

werelt weereelt wereld weerelden weereeld werrelts waerelds weerlyt wereldts vveerelts waereld weerelden waerelden weerlt werlt werelds sweerels zwerlyys swarels sweerelts werelts weeerelts sweerels weirelts tsweerelds werret vverelt werlts werrelt worreld werlden wareld weirelt weireld waerelt werreld wereld vvereld weereelts werlde tswerels werredt tsweeeldt wereldje waereldje weurlt wald weëled
Institutional Challenge: lack of knowledge and expertise → inefficiency
Adressing these challenges: The IMPACT project

- Consortium of 26 partners
  - Good mix of public and private partners
  - Users, researchers and industry work together to find solutions
  - Each established in a large international network
- Coordinated by the National Library of the Netherlands (KB)

- Large-scale Integrating Project
- EU funding: €12 100 000 (FP7 ICT Work Programme)
- Start date: 1 January 2008 (extended feb-april 2010)
- Duration: 4 years
- From 2012: sustainable Centre of Competence with alternative resources

- Currently, over 125 people across Europe, Israel and Russia involved in the project
IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands.
The IMPACT Vision

- We make digitisation of historical printed text in Europe **better, faster, cheaper**

- We provide tools, services and facilities for further **advancement of the State of the Art** in this field

Five key outcomes of the project will lead to fulfilling this vision ➔
IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands.

**IMPACT Key outcomes**

---

**IMPACT CENTRE OF COMPETENCE IN TEXT DIGITISATION**

- Licensing, Terms and Conditions
- Tool Take Up
- Impact Libraries
- Commitment

**Framework components**

- Datasets and Ground Truth
- Library expectations
- Requirements

**INTEROPERABILITY FRAMEWORK**

- Demonstration results
- Evaluation tools and metrics

---

**Means to build digitisation capacity**

- Web Community
- Decision Support Tools
- Knowledge Base
- Helpdesk
- Learning Resources
- Training Programme

- Documentation

**Modular suite of tools for improved text recognition (mass)**

- ABBYY FR 10
- IBM Adaptive OCR
- CONCERT

- Functional Extension Parser
- Lexical Resources
- Post correction modules

**Tools that significantly enhance SOA on different aspects of Text Recognition Workflow**

- Image enhancement toolkit
- Segmentation toolkit
- Experimental prototypes
- Tools for lexicon building
1. A modular suite of tools and resources to improve text recognition, ready for implementation in a mass digitisation workflow

**CONCERT:**
- OCR
- Correction with volunteer involvement
- Functional Extension Parser: structural metadata
2. Research prototypes that significantly advance the state of the art of research in text recognition
3. A free and Open Source Interoperability Framework with tools and resources for evaluating and demonstrating results

Facilities for

- wrapping all tools as web services,
- creating workflows with both IMPACT- and external tools
- instruments and resources for demonstrating and evaluating results
Tools & Applications

- OCR (C++, C#),
- Image Processing & Lexica (DLL),
- Command Line Tools (Win/Linux),
- Java, PHP, Perl, etc.

+ 3rd party software!

“One ring to rule them all...”

→ IMPACT Interoperability Framework (IIF)
**Technical Framework**

- Java 6
- Apache Axis2
- Apache Tomcat
- Apache httpd (optional)

- Focus on interoperability
- Web based, platform independent
- Highly standardised (SOAP/WSDL)
- Easy deployment (e.g. hot deployment & update)
- Open source (Apache License 2.0, LGPL)
Infrastructure

- Clustering (load balancing & fail over)
- Monitoring (soapUI)
- Central Results Repository (WebDAV)
- HTTPS encryption & authentication
Integration: Workflows

- OCR workflow = data pipeline
- Building blocks = processing steps (nodes)
- Integration = interaction between nodes

⇒ Collaboration with myGrid (paper@TPDL2011)
Workflow management

- Web 2.0 style registry: myExperiment
- Local client: Taverna Workbench
- Web client: project website
Evaluation Framework

Evaluation Results

Evaluation Tools

Compatibility through one common format

Evaluation Scenarios

Results

DIA / OCR Tools

Image Repository

Evaluation Metrics

Ground Truth

GT Tools
**Evaluation**

- Tool A vs Tool B
- Tool A(v1) vs Tool A(v2)
- Workflow X (Tool A + Tool B) vs Workflow Y (Tool A + Tool C)
- Workflow X vs previously digitised material

Users can identify optimal workflow for source material
Benefits

- Modular
- Flexible
- Transparent
- Expandable

Potential use: Production platform, evaluation framework, long-term preservation system - and many more!
4. The means for building digitisation capacity in Europe
5. A Centre of Competence in text digitisation with a business model that can sustain itself for 3 years

- **Content Holders**: Institutions holding historical text-based content that they wish to digitise. Both private and public.
- **Researchers**: Institutions and individuals engaged in research in all areas within the scope of the IMPACT project. Both private and public.
- **Service & Product Providers**: Have products and services that they wish to make available to the Content Holders and Researchers. Primarily private sector.

IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands.
 IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands.

- To be launched Q3 of 2011
- One stop shop for all content holders in Europe
- Main objective: delivering faster, better and cheaper digitisation of text
- Multi-sided platform, delivering different product and services to different customer segments
- will employ a Freemium business model
  - offering basic products and services for free,
  - charging for premium or special features
  - facilitating income generation to aid sustainability

GET INVOLVED NOW
LinkedIn group: IMPACT Improving Access to Text

- Building an online community
- Collecting feedback on IMPACT deliverables (to be incorporated in later versions)
- Discussions on topics related to digitisation, OCR & language technology
IMPACT final conference: 24-25 October 2011

Digitisation & OCR: Better, faster, cheaper
*Solutions of the IMPACT Centre of Competence and future challenges*

- Presentation of final results of IMPACT & related research in the area of OCR, digitisation and language technology

- Location: The British Library, London, UK
- Registration and more news available through the IMPACT website
- Early bird fee until June 2011
IMPACT is a project funded by the European Commission. It aims to significantly improve access to historical text and to take away the barriers that stand in the way of the mass digitisation of the European cultural heritage. Read more

Friday 19. November 2010
IMPACT at Czech Library conference
In the beginning of the December, the IMPACT project will be promoted at the Czech national...

[more]