Scalable and sustainable – OCR & document image analysis in the cloud

New Trends in Humanities Computing

Lotte Wilms – Koninklijke Bibliotheek, IMPACT Project
René van der Ark – Koninklijke Bibliotheek, Research programmer
Clemens Neudecker – Koninklijke Bibliotheek, Technical Project Manager IMPACT
Background: 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)
  - Early Dutch Books Online – [http://www.earlydutchbooksonline.nl](http://www.earlydutchbooksonline.nl)

- **Timeframe covered:** 1618 - 1995
Optical Character Recognition


DJgn i f paffato te S' aö'Jifeert mo?üen/bah .)etgi'uotbciraetail)i.r/JtmelchontDecht te / sbnbe bele btr felbrr geiufttceert baer bnber eeniglje jprant o^fen/bie ftcb .met beSpaenfcbeu enbeemgljen bifet Cbeiupcen berbonbru befe
Challenges in OCR
Answering the challenges – IMPACT

  Large-scale integrating research project, funded by the EC
  - Consortium of 26 partners
  - Coordinated by the National Library of the Netherlands (KB)
  - EU funding: € 12 100 000 (FP7 ICT Work Programme)
  - From 2012: sustainable Centre of Competence with alternative resources

- Main objectives:
  - Innovate OCR technology
  - Capacity building in mass-digitisation
IMPACT Solutions

- From a technical perspective:
  > 20 software toolkits for solving different problems

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

→ IMPACT Interoperability Framework (IIF)
Architecture

- IMPACT Interoperability Framework: Technologies
  - Java 6
  - Generic Web Service Wrapper
  - Apache Maven
  - Apache Tomcat
  - Apache Axis2
  - Apache Synapse
  - Taverna Workflow Engine

- IMPACT Interoperability Framework: Dataset
  - Hosted in the UK
  - PHP/mysql database, frontend for search
  - approx. 5 TB raw data (images, text files, metadata) and growing
How does it work?

1. Digitisation/OCR challenges registered and tagged in database
   → Warped text

2. Database contains 99.95% correct result: “ground truth”
How does it work?

1. Researcher develops new method to tackle a problem

2. Research prototype is wrapped to a SOAP web service
How does it work?

- A Web service is integrated as a workflow module.
- Workflow module can be evaluated, based on the ground truth.
Current setup

- Proxy receives requests from users and distributes the load to the available worker nodes (= server with all services installed).
- Main effect: Process parallelization, Load distribution, Fail over
- Drawback: Data is sent to worker nodes all around Europe = huge amount of data needs to be sent over the net!
Proposed setup

- Improve speed and availability for concurrent users
- Remove constraints for large-scale processing
Benefits

- Scalable platform
- Availability of resources to a large number of users
- Enable research into scalable computing for OCR & DIA
- Consolidation of support and maintenance
- Various interfaces (web/local)