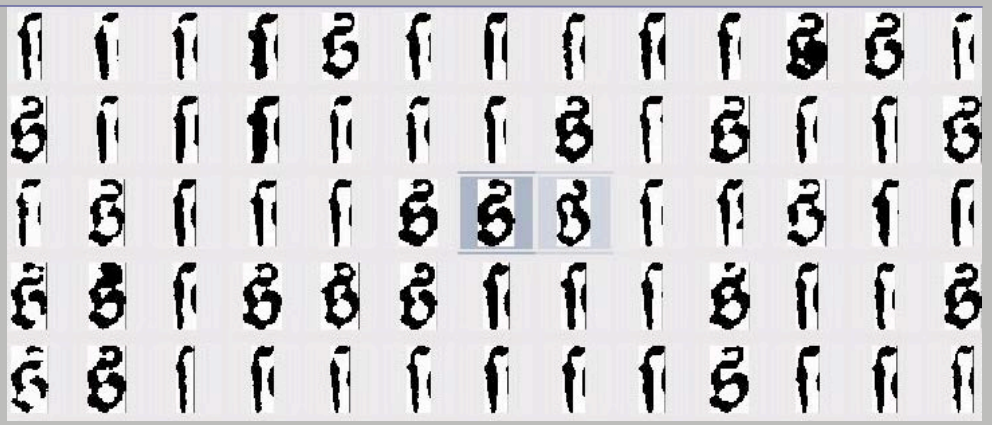
Improving Access to Text

IMPACT



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

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

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

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

  Dutch Parliamentary Papers – http://www.statengeneraaldigitaal.nl/

  Early Dutch Books Online – http://www.earlydutchbooksonline.nl

- Timeframe covered: 1618 - 1995

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

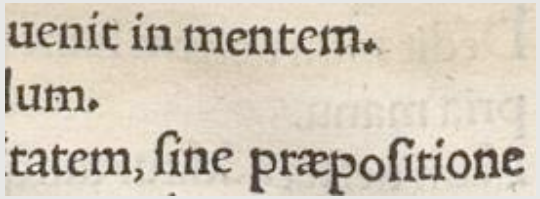
Optical Character Recognition

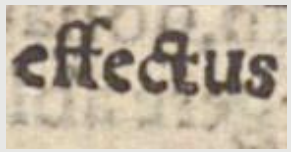
VVt Venetien den 1.Junij, Anno 1618.

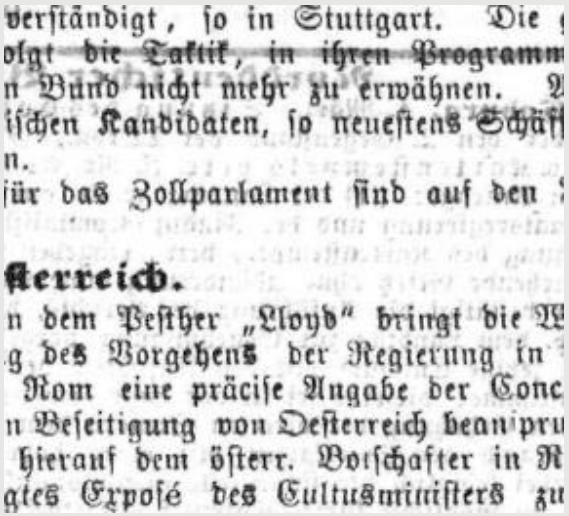
DJgn i f paffato te S' aö'Jifeert mo?üen/bah .)etgi'uotbciraetail)i.r/JtmelchontDecht te / sbnbe bele btr felbrr geiufttceert baer bnber eeniglje jprant o^fen/bie ftcb .met beSpaenfcbeu enbeeemgljen bifet Cbeiiupcen berbonbru befe

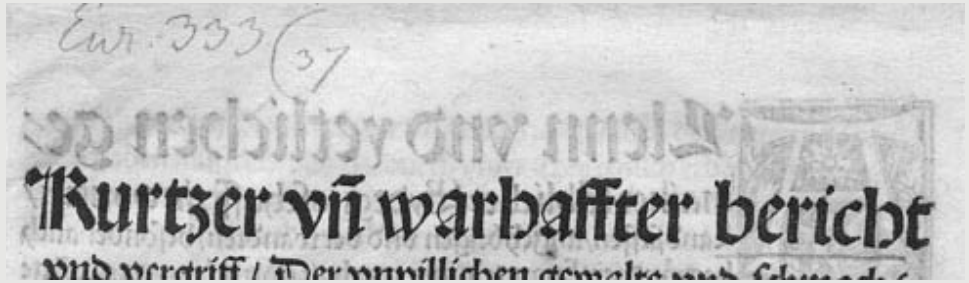
IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands.

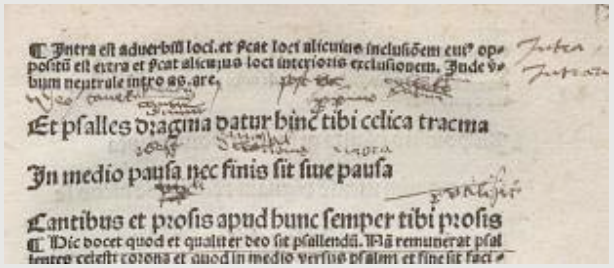
Challenges in OCR











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

Answering the challenges – IMPACT

- IMPACT – Improving Access to Text (2008 – 2011)

  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

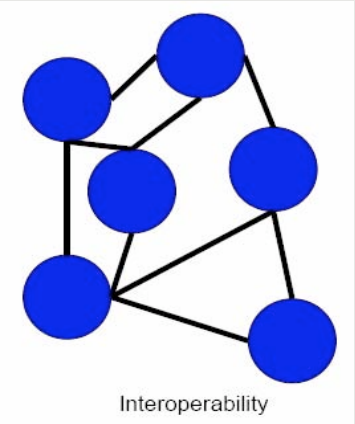
IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands.

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)



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

Architecture

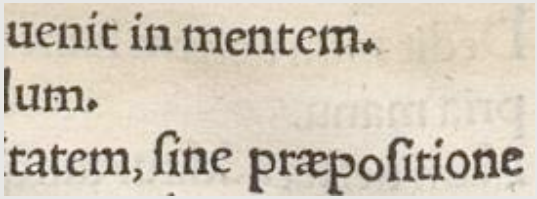
- IMPACT Interoperability Framework: Technologies
  - Java 6
  - Generic Web Service Wrapper
  - Apache Maven
  - Apache Tomcat
  - Apache Axis2
  - Apache Synapse
  - Taverna Workflow Engine

- 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

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

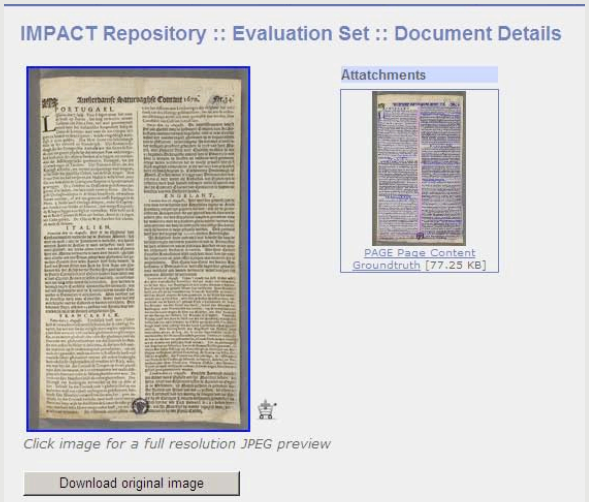
How does it work?

- Digitisation/OCR challenges registered and tagged in database



→ Warped text

- Database contains 99,95% correct result: “ground truth”

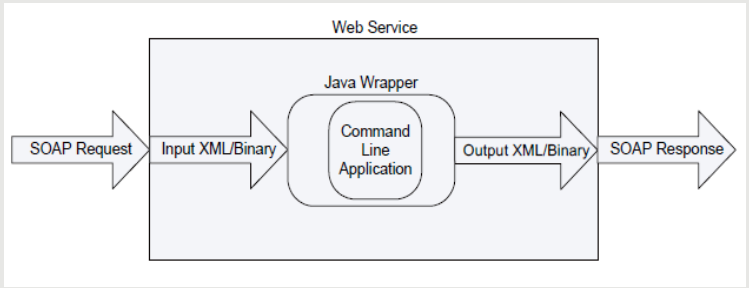


IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands.

How does it work?

- Researcher develops new method to tackle a problem

- Research prototype is wrapped to a SOAP web service



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

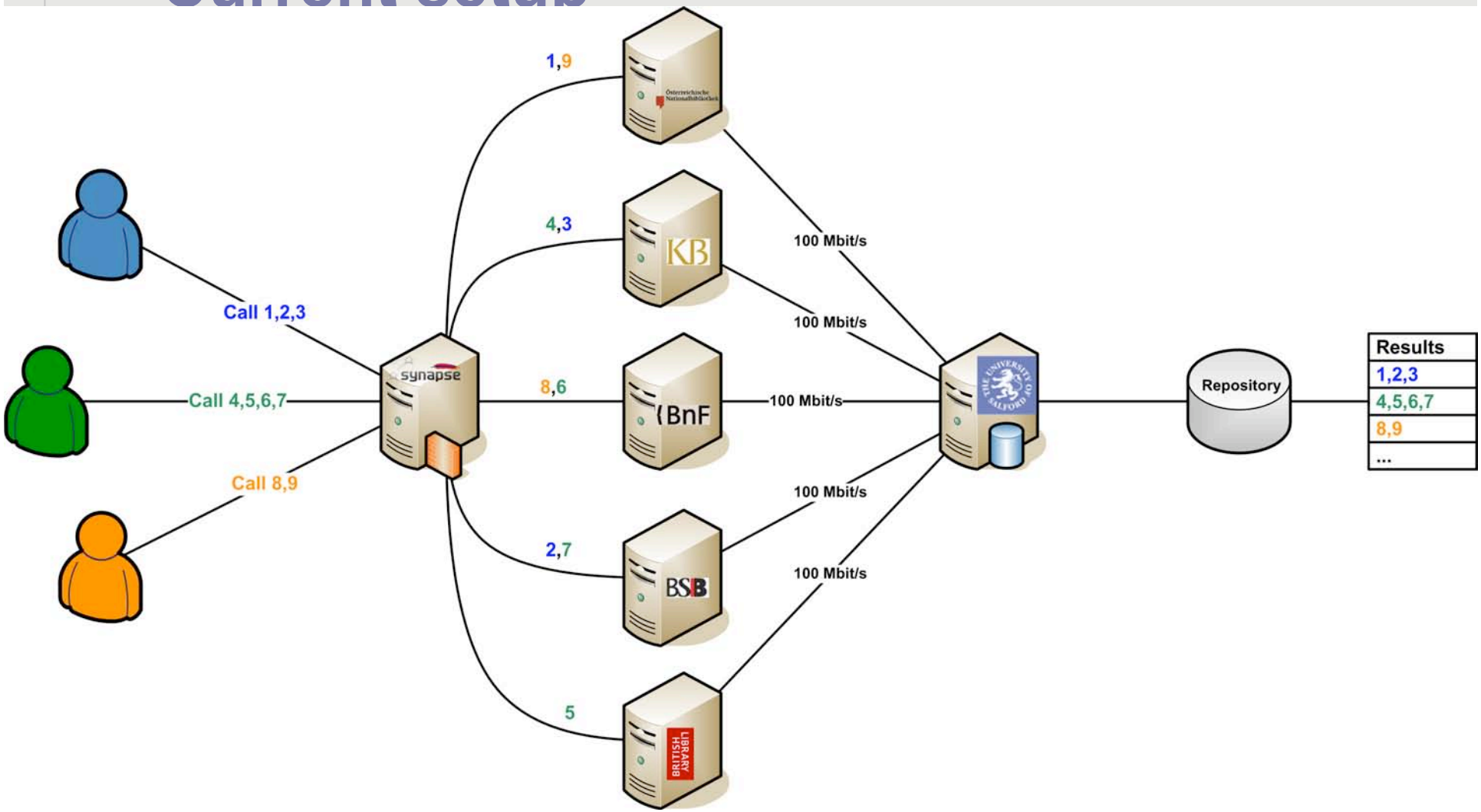
How does it work?

- Web service is integrated as a workflow module

- Workflow module can be evaluated, based on the ground truth

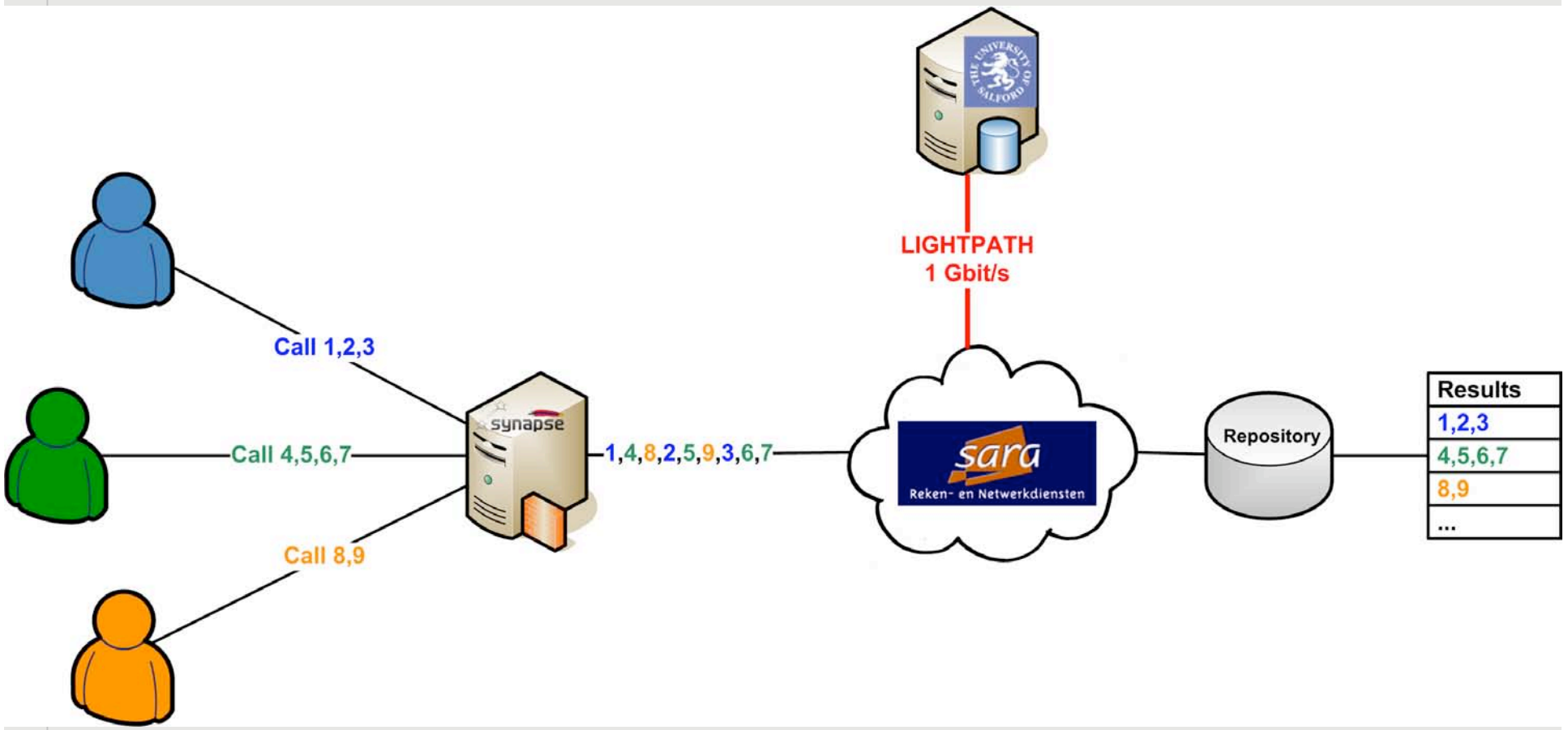
IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands.

Current setup



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

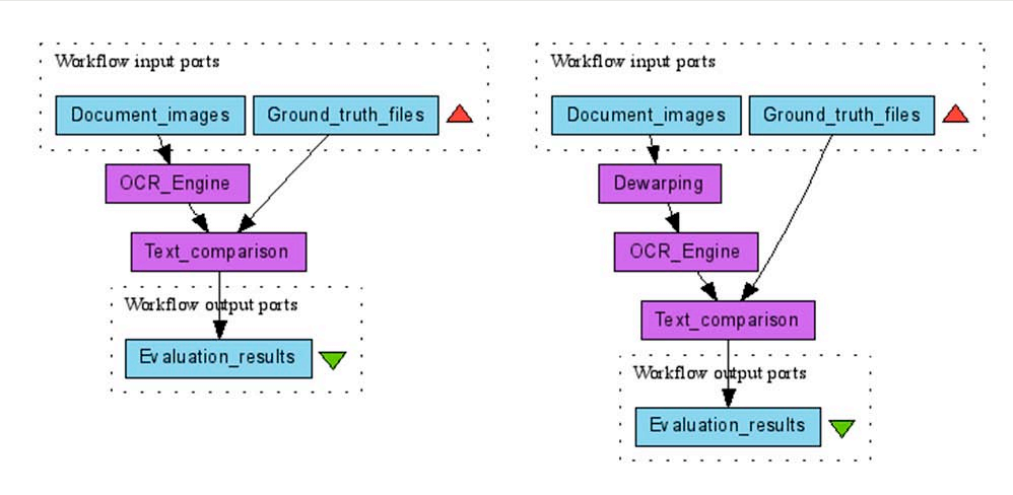
Proposed setup



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

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)



Improving Access to Text

IMPACT

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

Improving Access to Text

IMPACT

www.impact-project.eu

- Home
- News
- Helpdesk
- Tools and applications
- Calendar of events
- About the project
- FAQs
- Documents
- Sitemap
- Disclaimer
- Contact
- For partners

twitter LinkedIn slideshare WordPress YouTube vimeo

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

Monday 26. September 2011

Upcoming IMPACT Demo Days

In the next weeks, several IMPACT Demo Days will be held, in different languages of the IMPACT...

[more]

Friday 23. September 2011

Final IMPACT Conference, 24-25 Oct 2011: Speaker information now available