

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

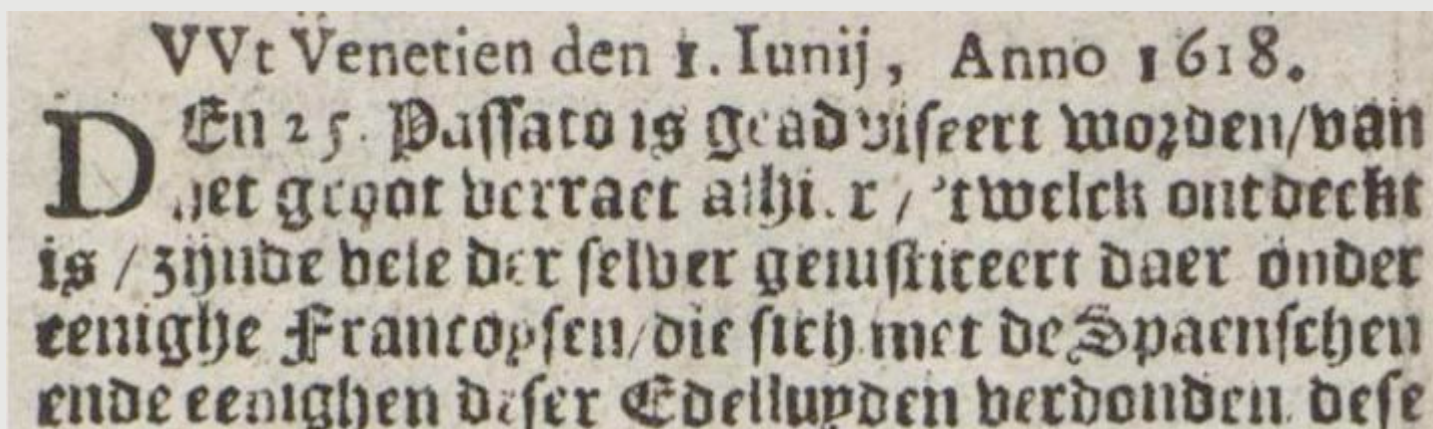
René van der Ark – Koninklijke Bibliotheek, Research programmer

Clemens Neudecker – Koninklijke Bibliotheek, Technical Project Manager IMPACT

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

Optical Character Recognition



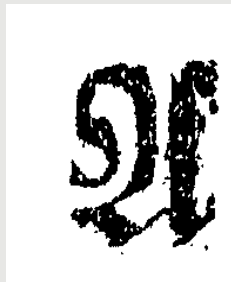
VVt Venetien den 1. Junij, Anno 1618.

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



Challenges in OCR

uenit in mentem.
lum.
tatem, sine præpositione



effectus



verständnis, so in Stuttgart. Die
folgt die Taktik, in ihren Programm
n Bund nicht mehr zu erwähnen. V
ischen Kandidaten, so neuestens Schäff
n.
für das Zollparlament sind auf den
Kerreich.
n dem Besther „Lloyd“ bringt die W
g des Vorgehens der Regierung in
Rom eine präzise Angabe der Conc
n Beseitigung von Oesterreich beanpru
hierauf dem österr. Vorschafter in M
gtes Exposé des Cultusministers zu

Eur. 333 (37)
Kurtzer vn̄ warbaffter bericht
vnd verdriff / Der vn̄willichen demelte vnd

¶ Intra est aduerbii loci, et p̄ae loci alicuius inclusiōem eius op̄
positū est extra et p̄at alicuius loci interiōis exclusiōem. Inde de
bim neutrale in tro so. are.
¶ Et p̄alles dyagma datur hinc tibi cœlica tracma
¶ In medio pausa nec finis sit siue pausa
¶ Antibus et profis apud hunc semper tibi profis
¶ Dic docet quod et qualiter deo sit p̄allendi. P̄ā remunerat p̄al
tentes cœlesti corōis et quod in medio versū d̄salini et sine sit facti

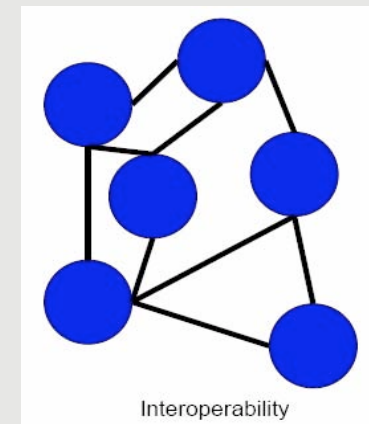
Answering the challenges – IMPACT

- 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 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

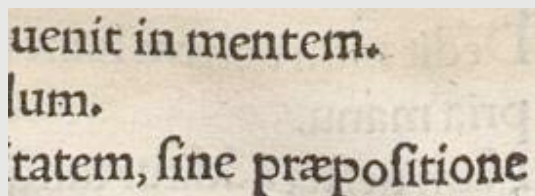
- IMPACT Interoperability Framework: Technologies
 - Java 6
 - Generic Web Service Wrapper
 - Apache Maven
 - Apache Tomcat
 - Apache Axis2
 - Apache Synapse
 - Taverna Workflow Engine

- 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?

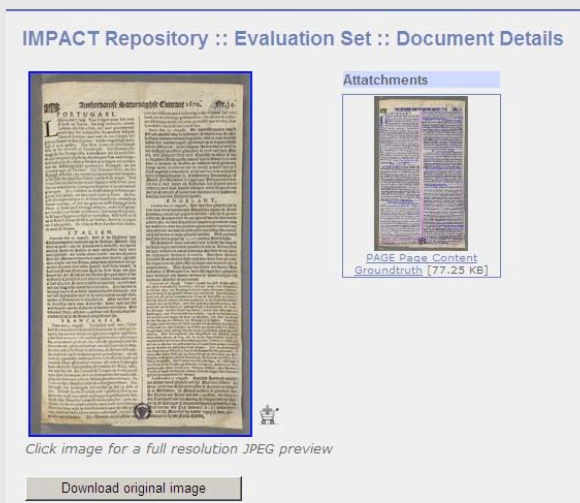
 Digitisation/OCR challenges registered and tagged in database



→ Warped text

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

IMPACT Repository :: Evaluation Set :: Document Details



Click image for a full resolution JPEG preview

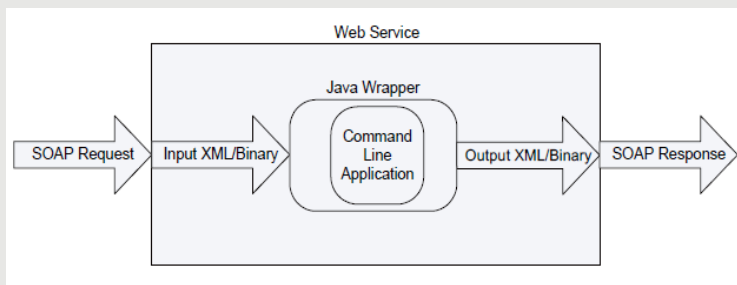
Download original image

How does it work?

 Researcher develops new method to tackle a problem



 Research prototype is wrapped to a SOAP web service



How does it work?

 Web service is integrated as a workflow module

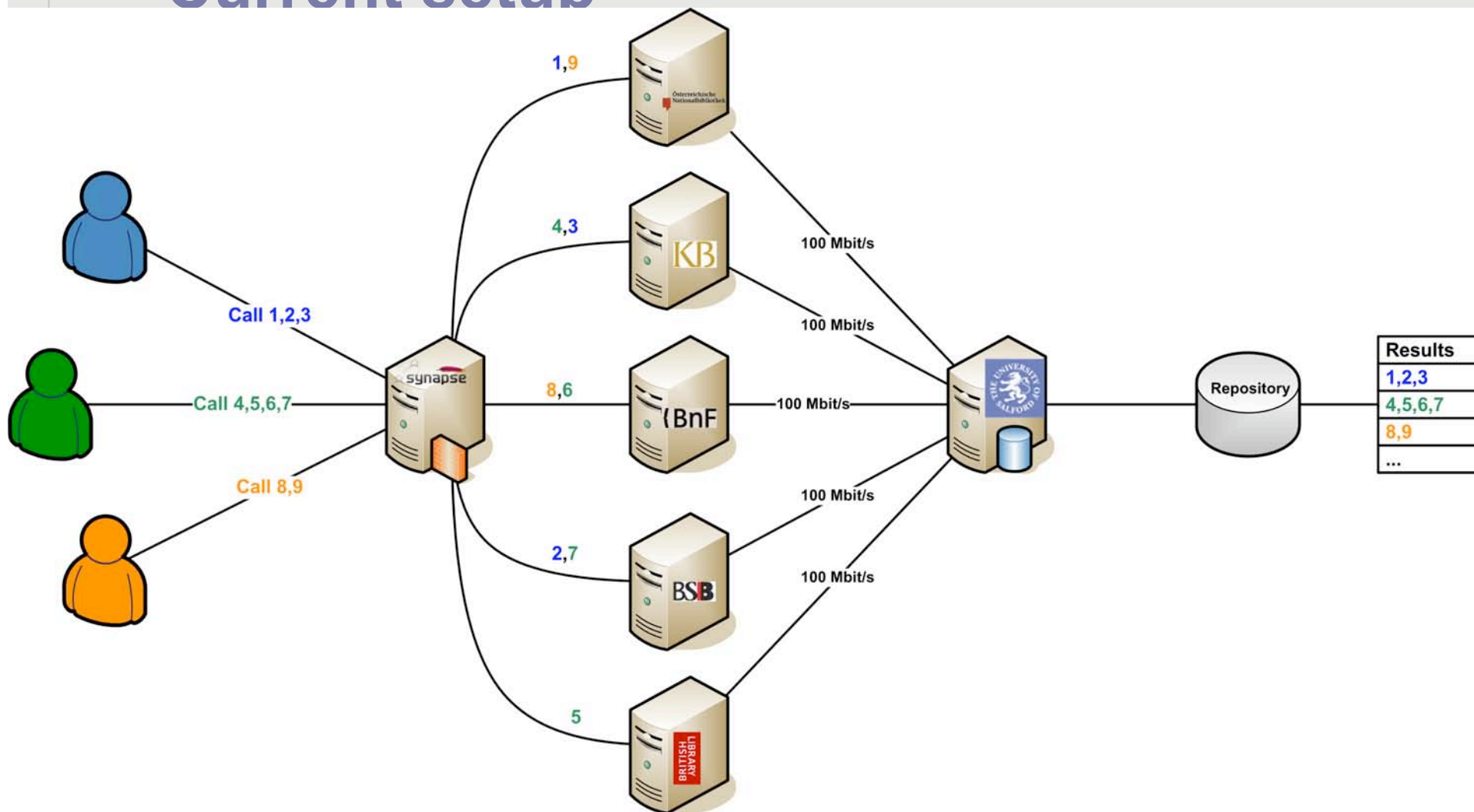


 Workflow module can be evaluated, based on the ground truth



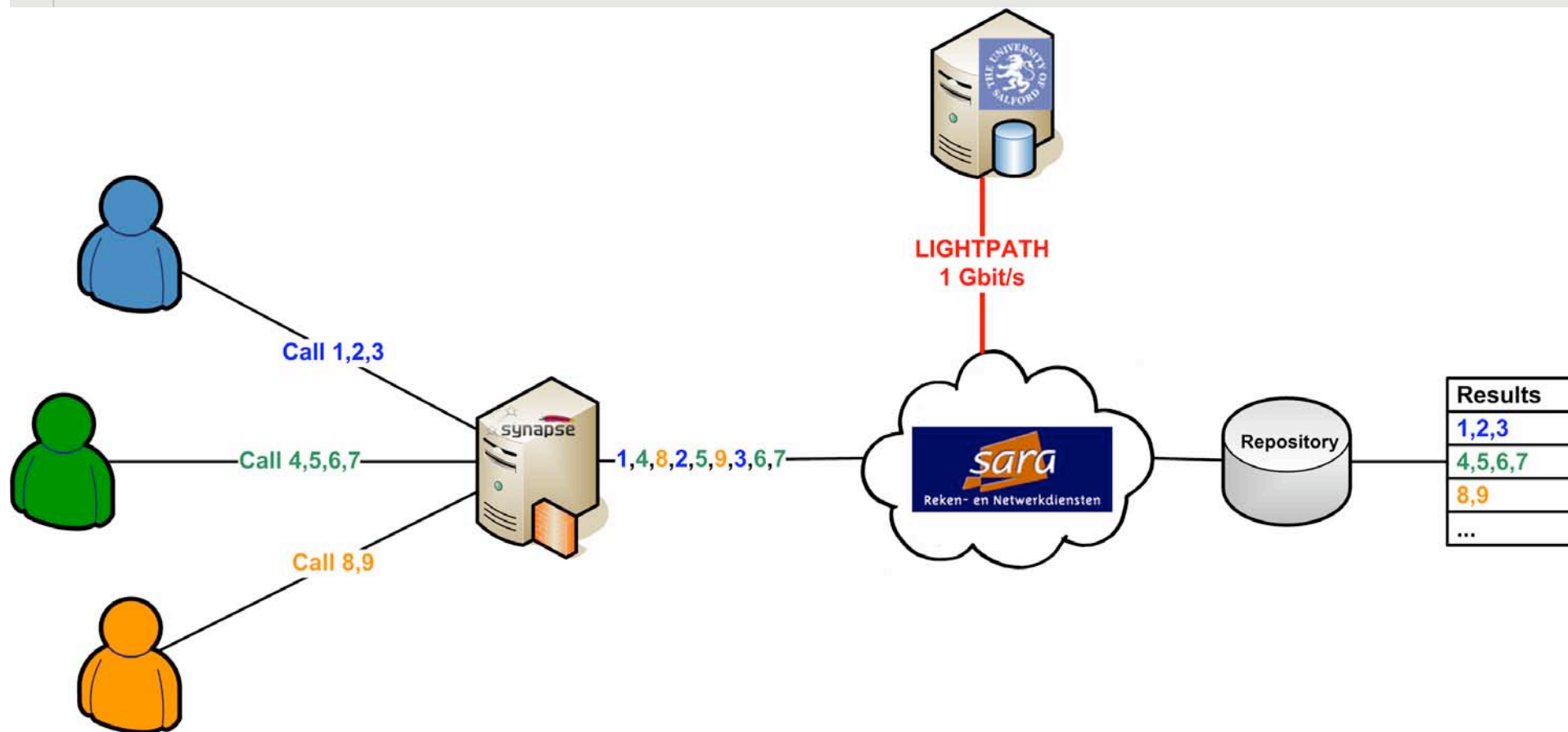


Current setup



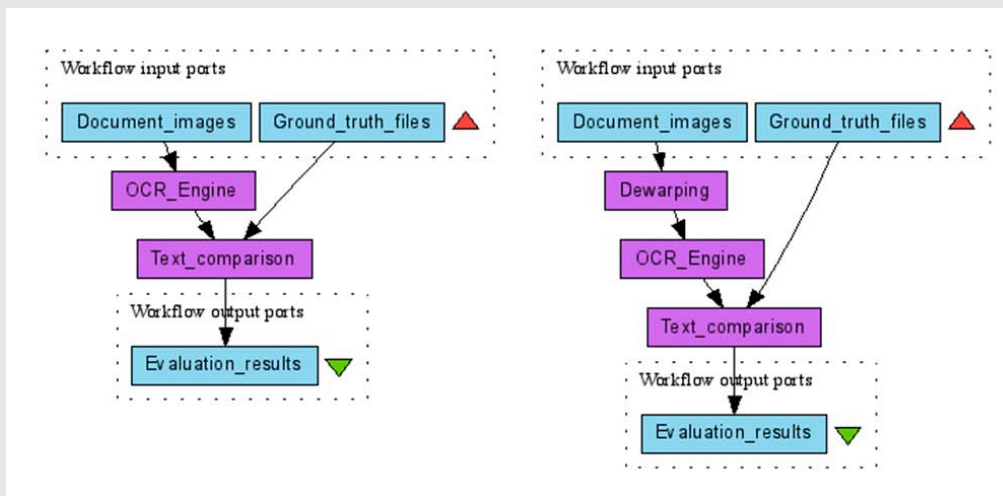


Proposed setup



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

Home

News

Helpdesk

Tools and applications

Calendar of events

About the project

FAQs

Documents

Sitemap

Disclaimer

Contact

For partners

www.impact-project.eu



twitter

LinkedIn



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

