

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 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

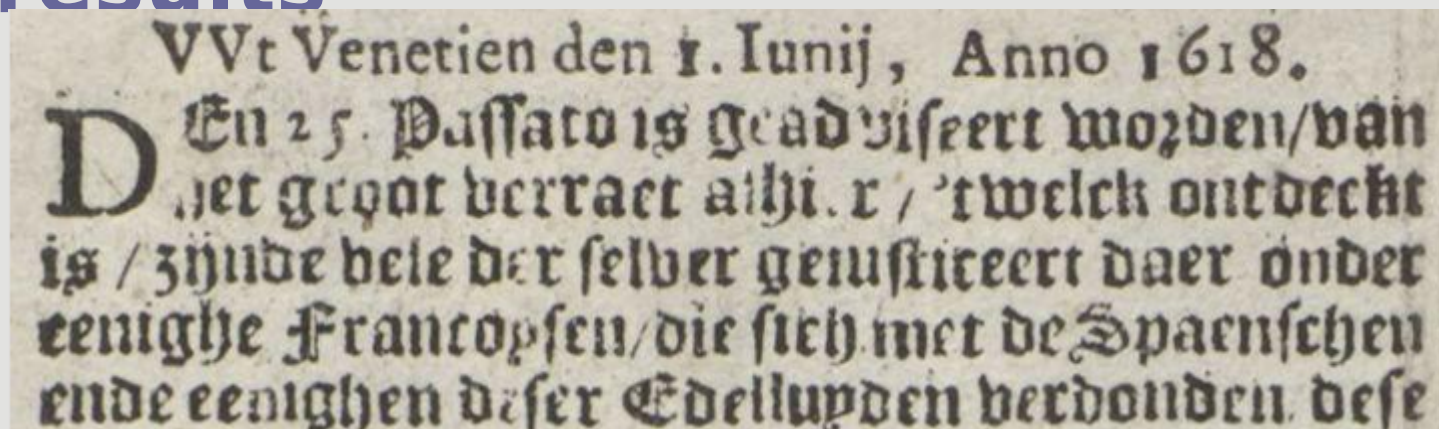
- Challenges to be addressed
- The IMPACT Project and Outcomes
- The IMPACT Centre of Competence
- Get involved now!
- Twitter: @impactocr, #impactproject



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



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 geiufstceert baer bnber eeniglje jprant o^fen/bie ftcb
.met

beSpaenfcben enbeeemgljen bifet Cbeiupcen berbonbru befe

OCR Challenges: damaged pages, bleed through, difficult layout, historic fonts ... and many more

uenit in mentem.
lum.
tatem, sine præpositione

21

effectus



verständigt, so in Stuttgart. Die
folgt die Taktik, in ihren Programm
n Band nicht mehr zu erwähnen. V
ischen Kandidaten, so neuesten Schaff
n.
für das Zollparlament sind auf den
sterreich.
n dem Vester „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 R
gtes Exposé des Cultusministers zu

Eur. 333 (37)
Kurtzer vñ warbaffter bericht
vnd vergriff / Der vnwilligen gemelten

Intra est adverbium loci, et sciat loci alicuius inclusionem cuius op
positum est extra et sciat alicuius loci interioris exclusionem. Inde de
bim neutrale in tro ss are.
Et psalles Diagma datur hinc tibi edica tracma
In medio pausa nec finis sit siue pausa
Lantibus et prosis apud hunc semper tibi prosis
Die docet quod et qualiter deo sit psallendi. Ita remunerat psal
tentes celesti corona et quod in medio versus psalmi et sine sit fact

Language Challenges: Spelling variants, orthographical variants, inflected forms...and more

A. My daarentegen verſchaft het een onbedenkelyk
genoegen, als ik de voortgang van deeze zo wel als
alle andere Weetenſchappen door de verſcheiden eu-
wen der **wereld** naargaa, en zie, hoe veel juifter onze
begrippen, hoe veel bondiger onze beweegredenen
in de Zedekunde zyn, en welk een hangelyke ver-
andering inzonderheid de Chriſtelyke Openbaaring,
ten deezen opzigte, heeft uitgewerkt.

Historical variants of the Dutch word 'wereld' (world):

werelt weerelt wereld weerelds wereldt werelden weereld werrelts waerelds weerlyt
wereldts vveerelts waereld weerelden waerelden weerlt werlt werelds sweerels
zwerlys swarels swerelts werelts swerrels weirelts tsweerelds werret vverelt werlts
werrelt worreld werlden wareld weirelt weireld waerelt werreld wereld vvereld weerelts
werlde tswerels werreldts weereldt wereldje waereldje weurlt wald weëled

Institutional Challenge: lack of knowledge and expertise → inefficiency





Addressing 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



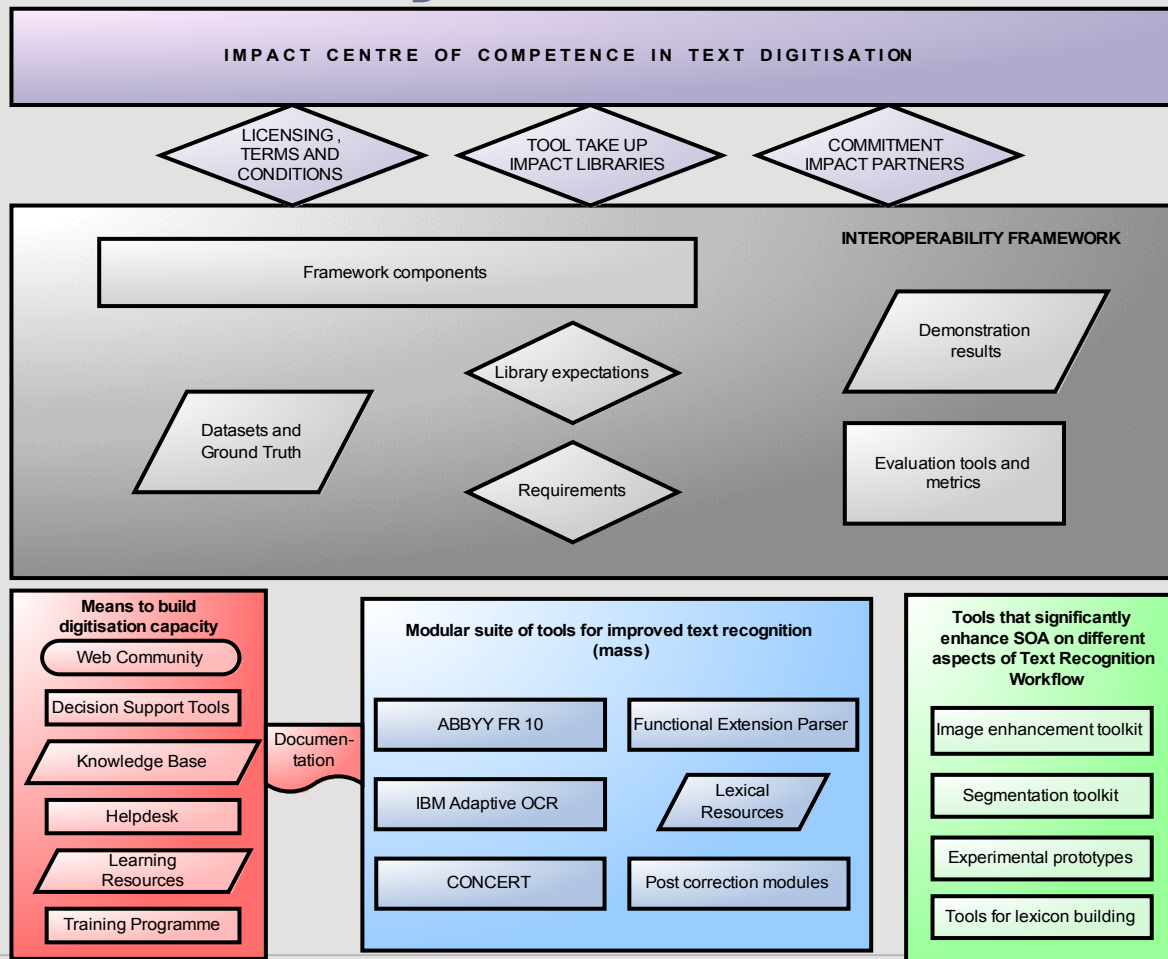
The IMPACT Vision

- 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



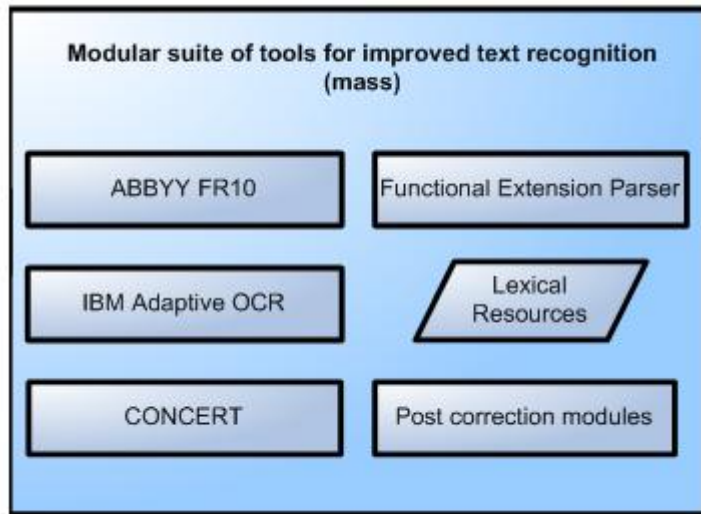
Five key outcomes of the project will lead to fulfilling this vision →

IMPACT Key outcomes



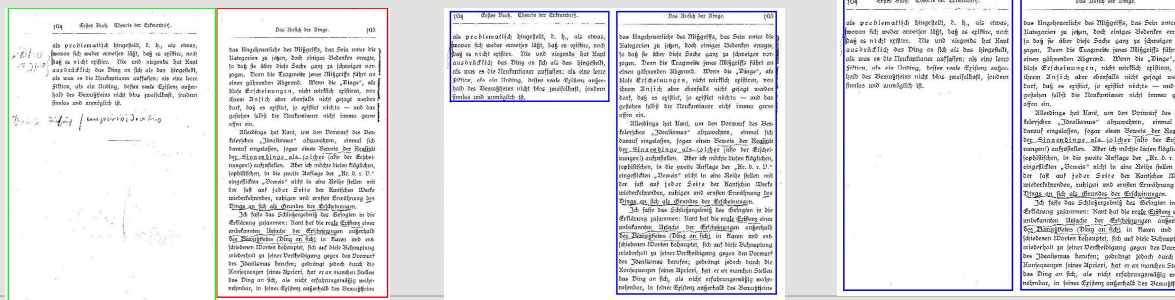
1. A modular suite of tools and resources to improve text recognition, ready for implementation in a mass digitisation workflow

ABBYY® FineReader Engine 10

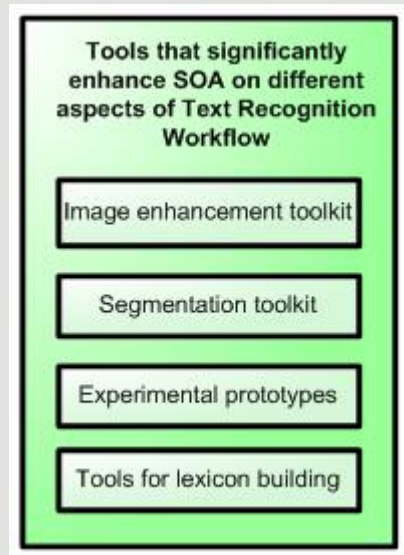


CONCERT: OCR Correction with volunteer involvement

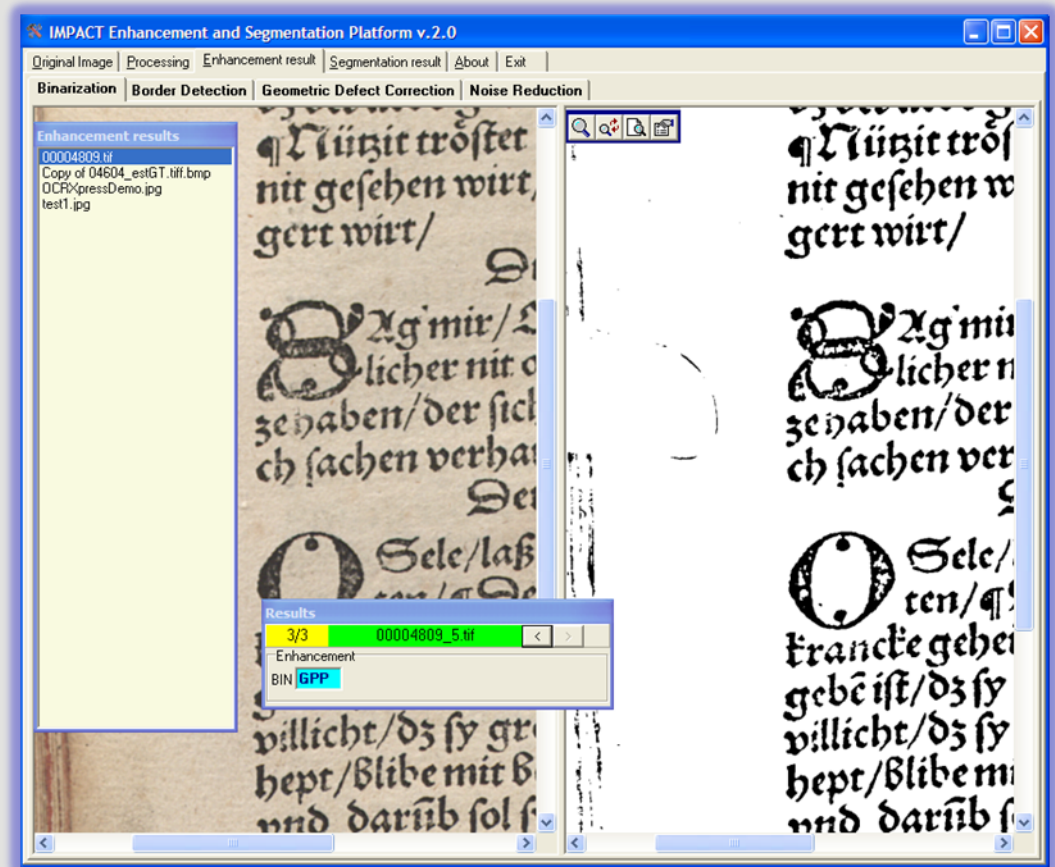
Functional Extension Parser: structural metadata



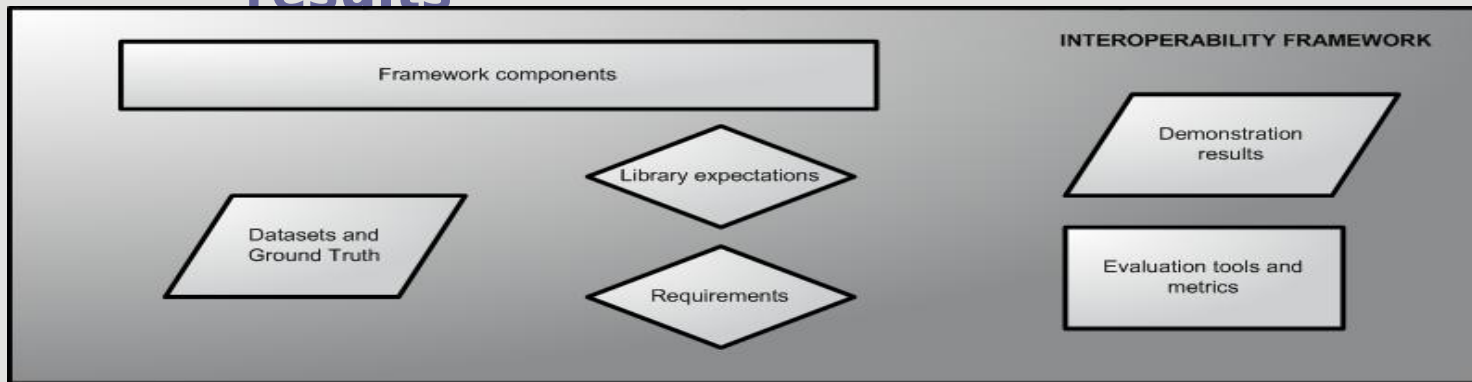
2. Research prototypes that significantly advance the state of the art of research in text recognition



Platform
Enhancement
and
segmentation



3. A free and Open Source Interoperability Framework with tools and resources for evaluating and demonstrating results



Facilities for

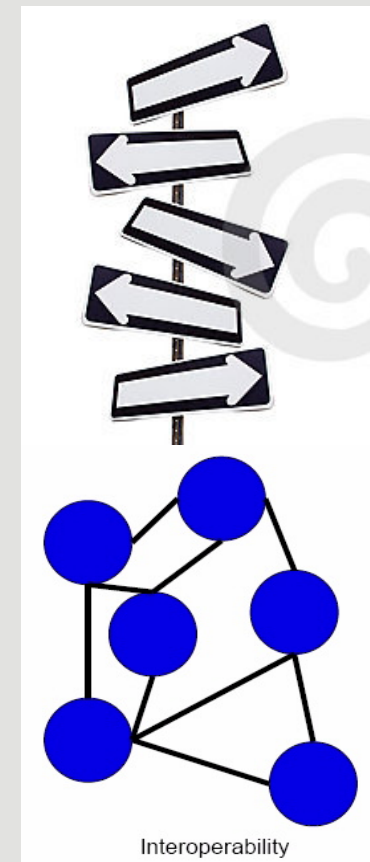
- wrapping all tools as web services,
- creating workflows with both IMPACT- and external tools
- instruments and resources for demonstrating and evaluating results

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...”

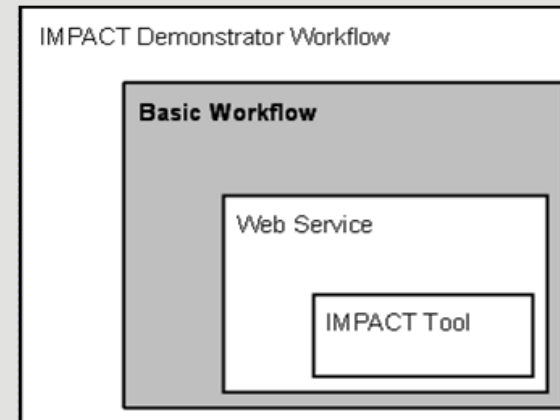
→ IMPACT Interoperability Framework (IIF)



Technical Framework

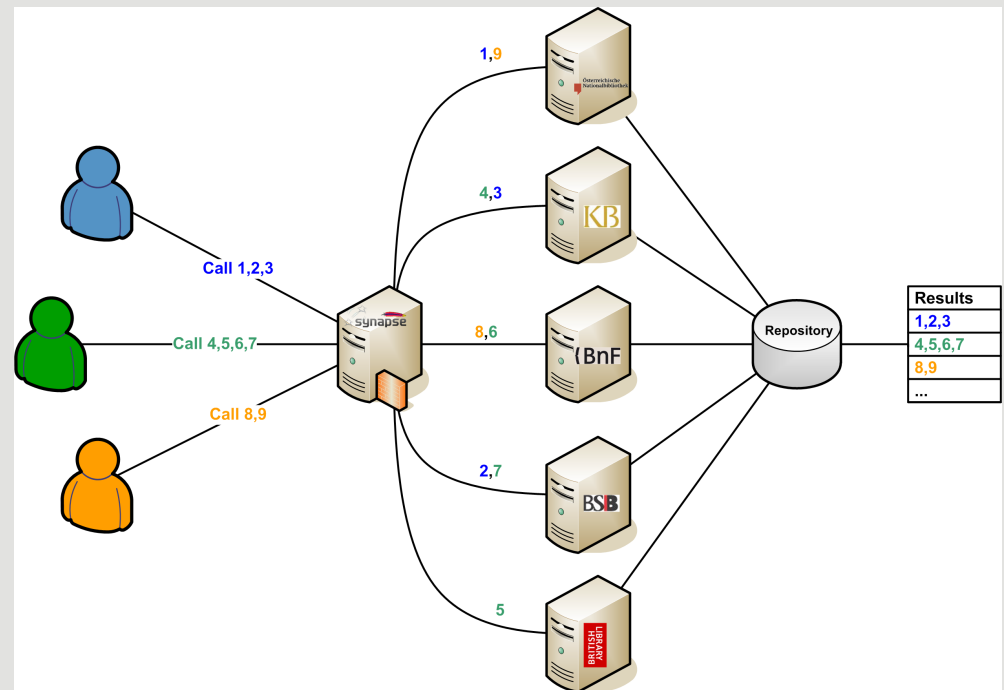
- Java 6
- Apache Axis2
- Apache Tomcat
- Apache httpd (optional)

- 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

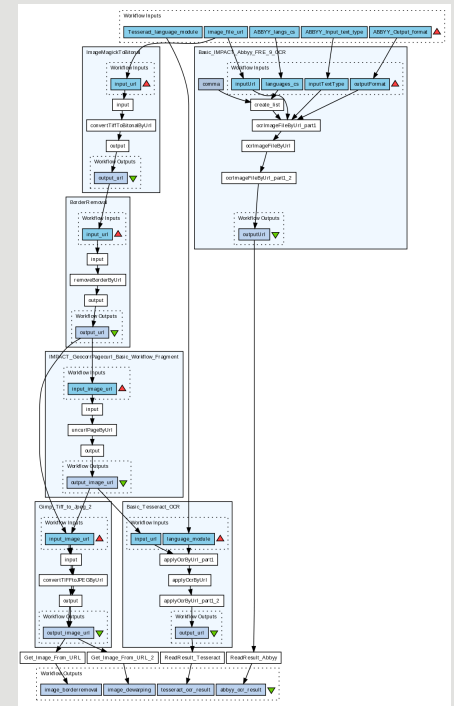
- Clustering
(load balancing & fail over)
- Monitoring (soapUI)
- Central Results Repository (WebDAV)
- HTTPS encryption & authentication



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



Taverna 2 **IMPACT IBM Concert UploadPagesUri BasicWorkflow (v1)** [View](#)
Created: 26/02/10 @ 09:30:10 | Last updated: 26/02/10 @ 09:30:12 [Download \(v1\)](#)

Credits: [Sven](#)

License: Creative Commons Attribution-Share Alike 3.0 Unported License

Updated pages to the IBM Concert tool. It is required that you login to Concert and get a session key before applying this basic workflow (see note on the side). The result files should be returned as word, xls, or character level. The Concert Web tool is available at <http://doc-proc.hack4lib.com/S000ConcertPagesUri/>. Under construction, the update does still not work correctly!

Rating: 0.0 / 5 (0 ratings) | Versions: 1 | Reviews: 0 | Comments: 0 | Citations: 0

Viewed: 0 times | Downloaded: 0 times

This Workflow has no tags!

Workflow

Taverna 2 **Helper Transform XML using XSLT Basic Workflow (v1)** [View](#)
Created: 03/03/10 @ 14:11:23 | Last updated: 03/03/10 @ 14:30:14 [Download \(v1\)](#)

Credits: [Sven](#)

License: Creative Commons Attribution-Share Alike 3.0 Unported License

The input XML document is converted using the input XSLT document to an output document which can be another XML, the HTML, or plain text, depending on the XSLT used in this workflow.

Taverna Workbench 2.2.0

File Edit View Workflow Advanced Help

Tools and applications: [Tools](#) [Tools and applications](#) [Workflow](#) [Workflow](#) [Workflow](#)

Workflow 1: [Workflow 1](#) [Workflow 1](#) [Workflow 1](#)

Workflow 1 input fields

image_file_uri: [image_file_uri](#) [Browse...](#)

groundtruth_file_uri: [groundtruth_file_uri](#) [Browse...](#)

[Execute workflow](#)

Aus dem Gerichtsakte.

Diebstahl. (Orig.-Ver.) Ferdinand Schenz, fälschlich Karl gebürtig, 31 Jahre alt, seines Zeichens ein Mühljunge, welcher wegen Diebstahls und ein Mal wegen Raubes abgeurteilt ist am 1. November v. J. mittels Nachschlüssel Einlaß auf das Zidamend, ließ sich mittelft eines bei den Rattermühlen gefohl- s Kirchengesicht herab, wo er von den Seitenallären Boten gegen- in 94 fl. 10 kr. entwendete. Unter den letzteren befand sich auch u geipenbete silberne Taschengeld, deren Feiger auf 3 lbr ger-

Improving Access to Text
IMPACT

printable view

Search the IMPACT website

Home About the project News Calendar of events Tools and applications Documents Startup Database Contact For partners

Login: [Create New User](#) [Log In](#)

Workflow Client

Please upload your workflow file:

Workflow 1: [Browse...](#)

[Show input fields](#)

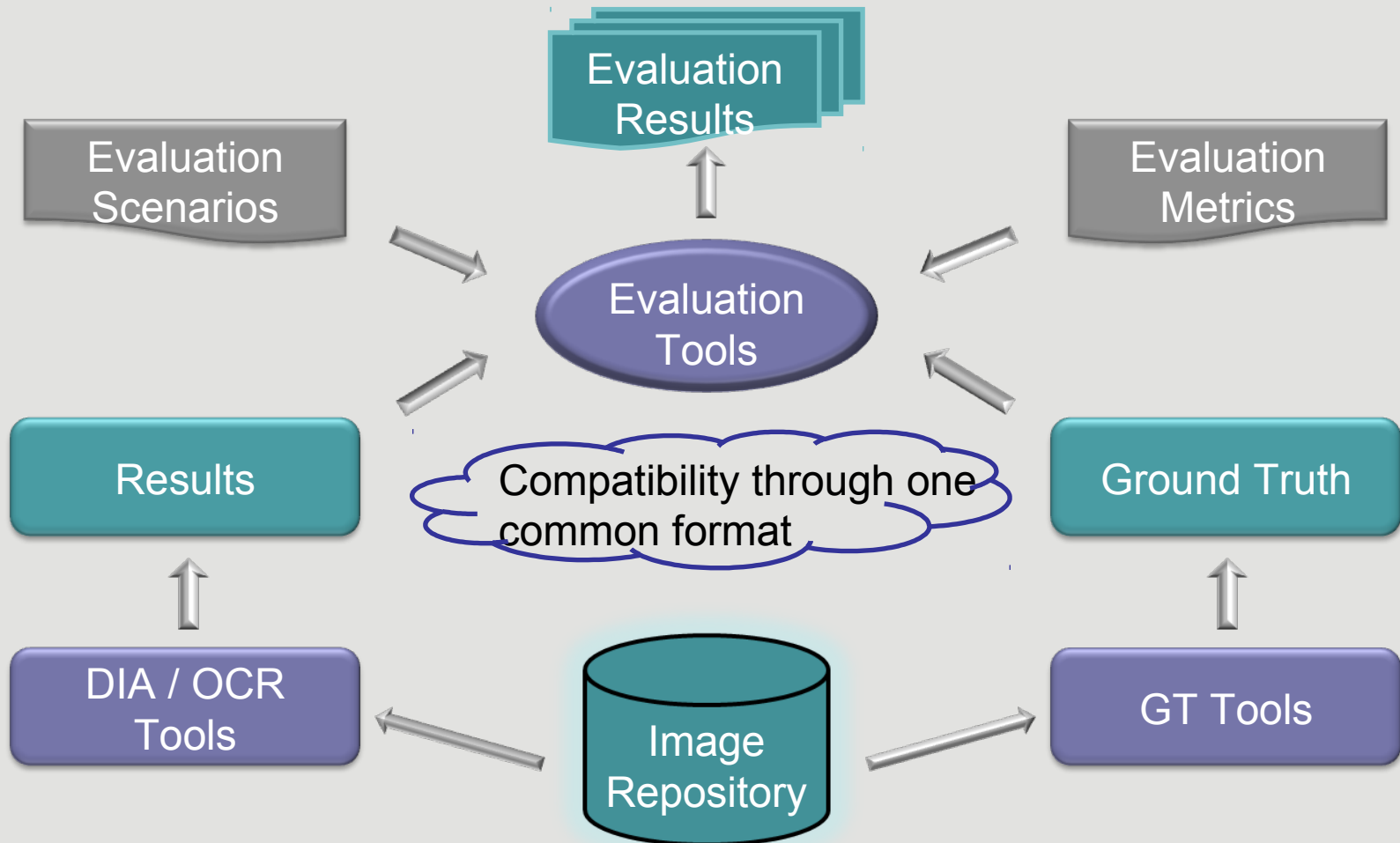
Workflow 1 input fields

image_file_uri: [image_file_uri](#) [Browse...](#)

groundtruth_file_uri: [groundtruth_file_uri](#) [Browse...](#)

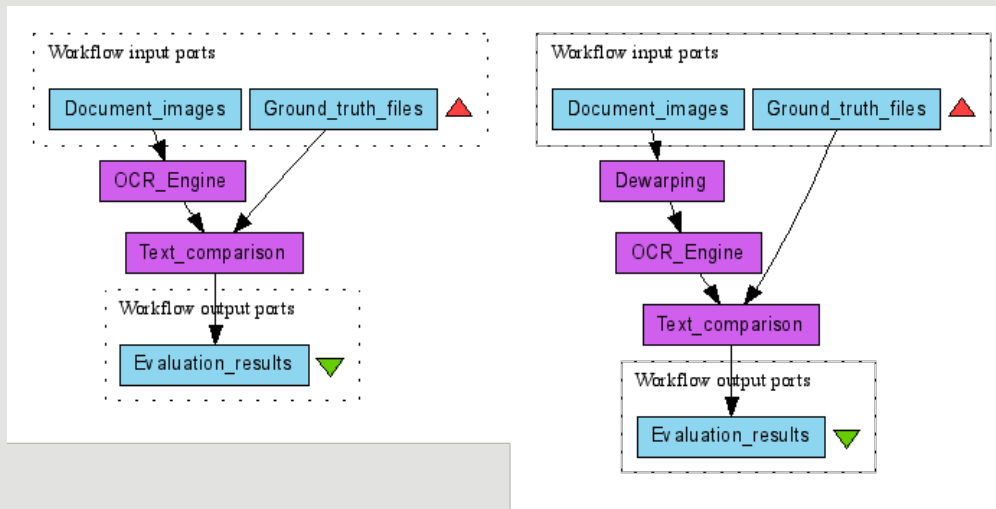
[Execute workflow](#)

Evaluation Framework



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



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

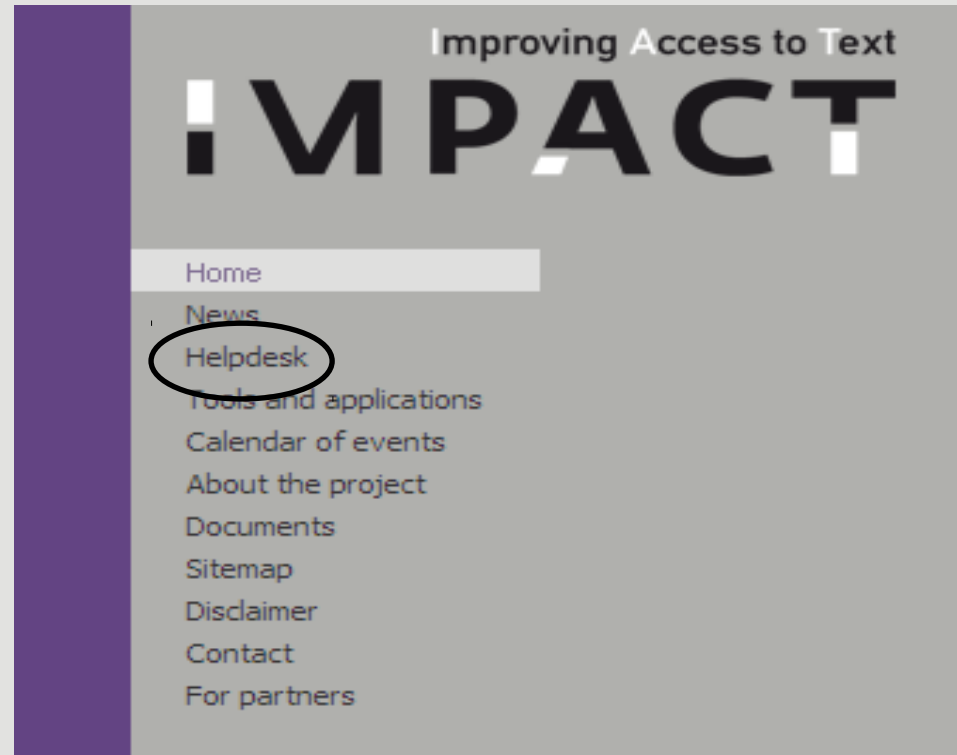
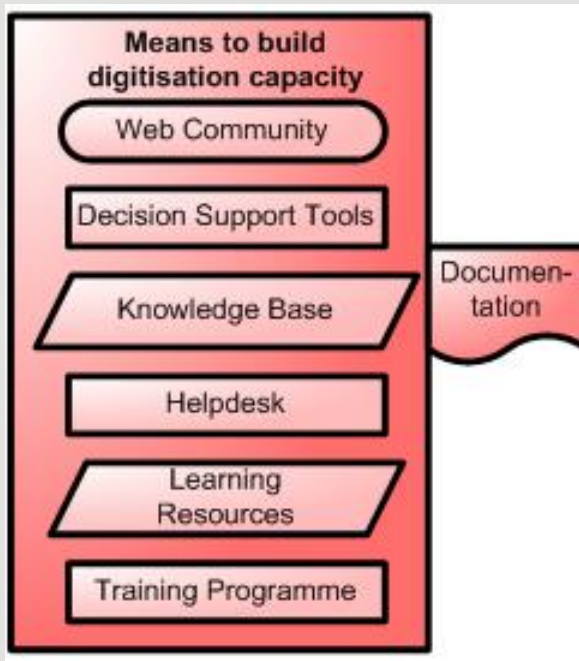
- Modular
- Flexible
- Transparent
- Expandable



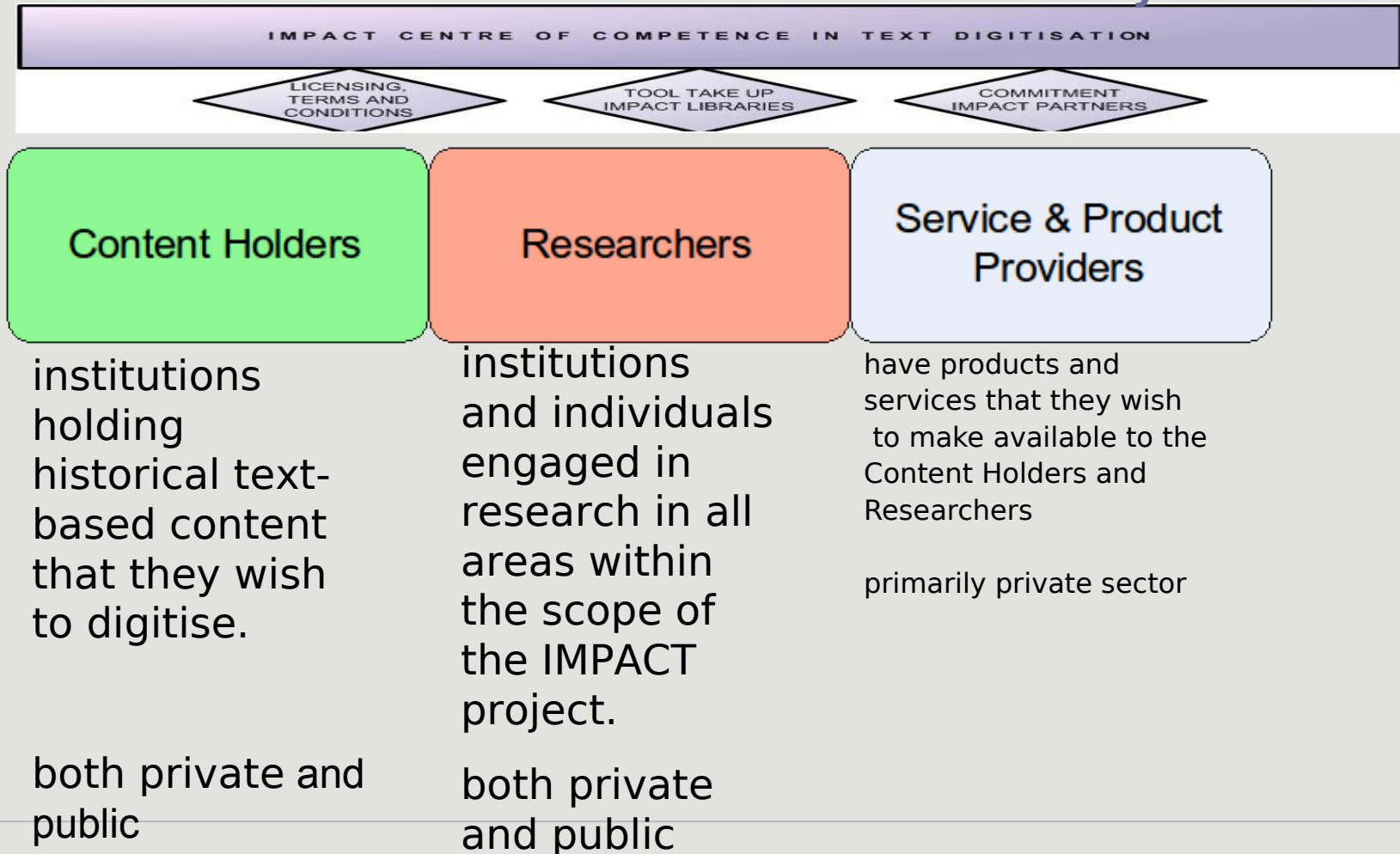
→ Potential use: Production platform, evaluation framework, long-term preservation system - and many more!



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





- 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

IMPACT Improving Access to Text

Discussions Members Promotions Jobs Search Manage More... Invite others

Start a discussion or share something with the group...
Maximum length is 200 characters.
Attach a link Share

My Activity

Most Popular Discussions

IMPACT Briefing Paper: Optical Character Recognition <Draft Pilot Release>
IMPACT has pleasure in releasing a draft pilot version of a Briefing Paper that provides an overview of Optical Character Recognition for libraries and other practitioners of mass digitisation.
This is the first 'draft' section available from our upcoming 'Digitisation ...
<http://www.impact-project.eu/uploads/media/IMPACT-ocr-bp-pilot-1b.pdf>
posted 1 month ago
David Bruchmann 2 days ago • @Richard, ... »
See all 15 comments »

IMPACT Best Practice Guide: Optical Character Recognition Sections 1 & 2 <Draft Pilot Release>
This week, as part of our gradual release of pilot materials in a 'draft' form, we have pleasure in releasing the first two sections of our Best Practice Guide on Optical Character Recognition, which can be downloaded from:
[http://www.impact-project.eu/uploads/media/IMPACT-o ...](http://www.impact-project.eu/uploads/media/IMPACT-o...)
Impact | Improving access to text: Pilot Tools
<http://www.impact-project.eu/index.php?id=138>
posted 1 month ago
Lotte Wilms 1 month ago • Lotte likes this.

Manager's Choice


IMPACT Storage Estimator - Released for Feedback <STOP PRESS>
Ed Bremner See all »

Updates: Last 7 Days

Elisabeth Freyre has joined the group.
1 day ago • Send message

David Bruchmann and 1 more commented on: IMPACT Briefing Paper: Optical Character Recognition <Draft Pilot Release>
2 days ago • 15 comments

Lieke Ploeger commented on: IMPACT Tools on the Technology Watch List of Europeana
4 days ago • Like • 1 comment
See all updates »



- 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

Improving Access to Text

IMPACT



KB

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

www.impact-project.eu

Improving Access to Text

IMPACT

[printable view](#)

[Home](#)

[News](#)

[Helpdesk](#)

[Tools and applications](#)

[Calendar of events](#)

[About the project](#)

[Documents](#)

[Sitemap](#)

[Disclaimer](#)

[Contact](#)

[For partners](#)

Twitter: @impactocr,
#impactproject

[twitter](#)

[Linked in](#)



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](#)

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...

[\[more\]](#)