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.

IMPACT: Centre of Competence in digitisation

Hildelies Balk – Pennington de Jongh, IMPACT Project director, KB National Library of the Netherlands

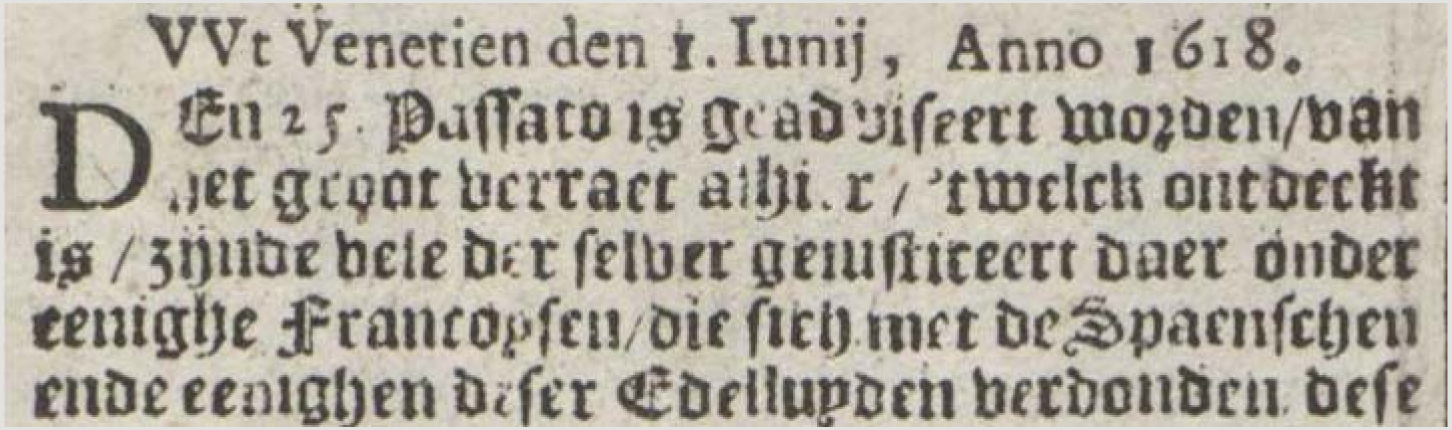
IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands.

A short introduction...

http://www.youtube.com/user/theimpactproject

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

Historical text: typical OCR results



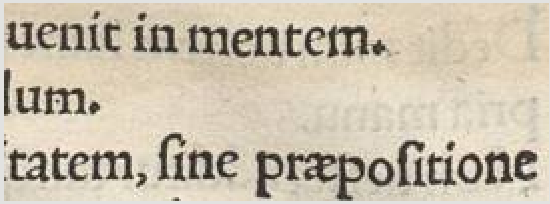
VVt Venetien den 1. Iunij, Anno 1618.

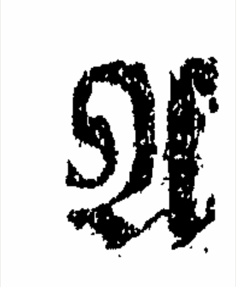
DEn 25. Paſſato is geadviſeert worden/van
het groot verraet alhier/ 'twelck ontdeckt
is/ zijnde vele der ſelver geiuſticeert daer onder
eenighe Francoyſen/die ſich met de Spaenſchen
ende eenighen deſer Edelluyden verbonden. deſe

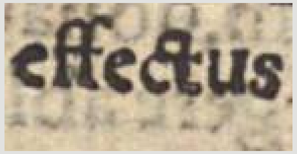
VVt Venetien den 1.Junij, Anno 1618.
DJgn i f paffato te S' aö'Jifeert mo?üen/bah .)
etgi'uotbciraetail)i.r/JtmelchontDecht
te /sbnbe bele btr felbrr geiufttceert baer bnber
eeniglje jprant o^fen/bie ftcb .met beSpaenfcbeu
enbeeemgljen bifet Cbeiiupcen berbonbru befe

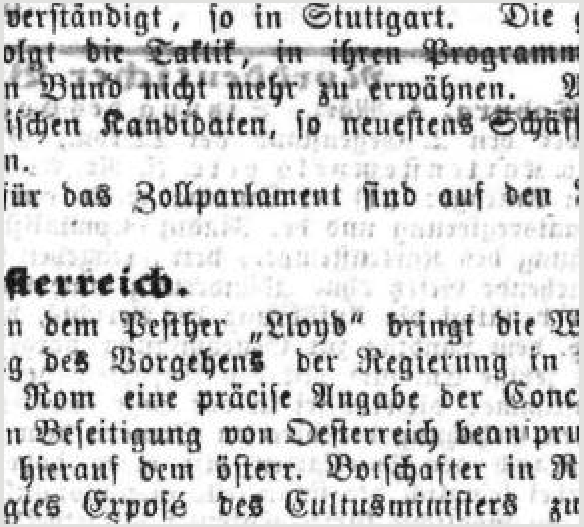
IMPACT is supported by the European Community under the FP7 ICT Work Programme. The project is coordinated by the National Library of the Netherlands.

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

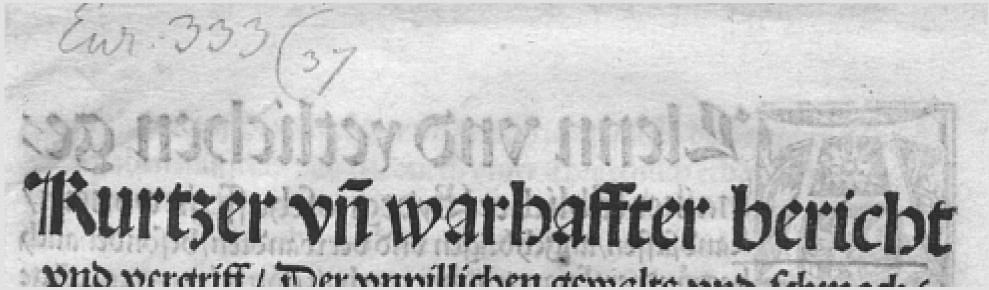

uenit in mentem.
lum.
tatem, ſine præpoſitione

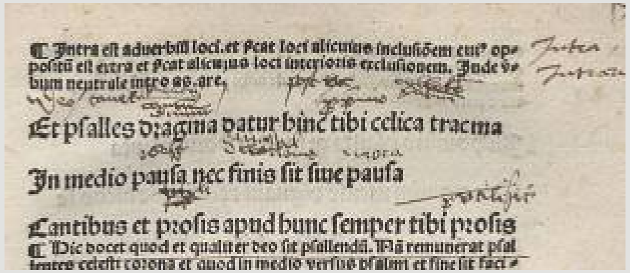



effectus

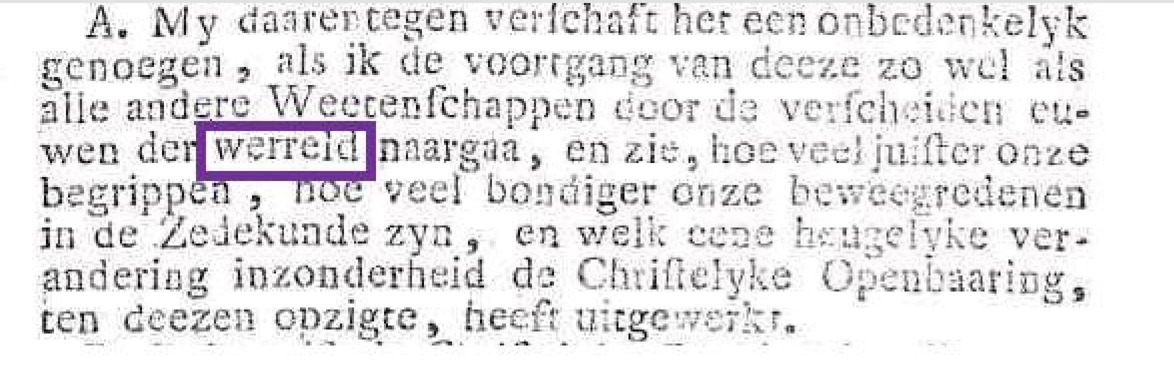

verſtändigt, ſo in Stuttgart. Die
olgt die Taktik, in ihren Programm
n Bund nicht mehr zu erwähnen.
iſchen Kandidaten, ſo neueſtens Schäff
n.
für das Zollparlament ſind auf den

terreich.
n dem Peſther „Lloyd“ bringt die W
g des Vorgehens der Regierung in
Rom eine präciſe Angabe der Conc
n Beſeitigung von Oeſterreich beanſpru
hierauf dem öſterr. Botſchafter in R
gtes Expoſé des Cultusminiſters zu


Ewr. 333 (37
Kurtzer vñ warhaffter bericht
vnd vergriff / Der vnwillichen gewalts vnnd ſchmach


Intra eſt aduerbiũ loci. et ſcat loci alicuius incluſiõem cui⁹ oppoſitũ eſt extra et ſcat alicuius loci interioris excluſionem. Unde verbum neutrale intro as. are.
Et pſalles dragma datur hinc tibi celica tracma
In medio pauſa nec finis ſit ſine pauſa
Cantibus et proſis apud hunc ſemper tibi proſis
Hic docet quod et qualiter deo ſit pſallendũ. Quã remunerat pſallentes celeſti corona et quod in medio verſus pſalmi et fine ſit faci

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 euwen der werreld naargaa, en zie, hoe veel juiſter onze begrippen, hoe veel bondiger onze beweegredenen in de Zedekunde zyn, en welk eene heugelyke verandering 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 werld vvereld weerelts werlde tswerels werreldts weereldt wereldje waereldje weurlt wald weëled

Institutional Challenge: lack of knowledge and expertise → inefficiency

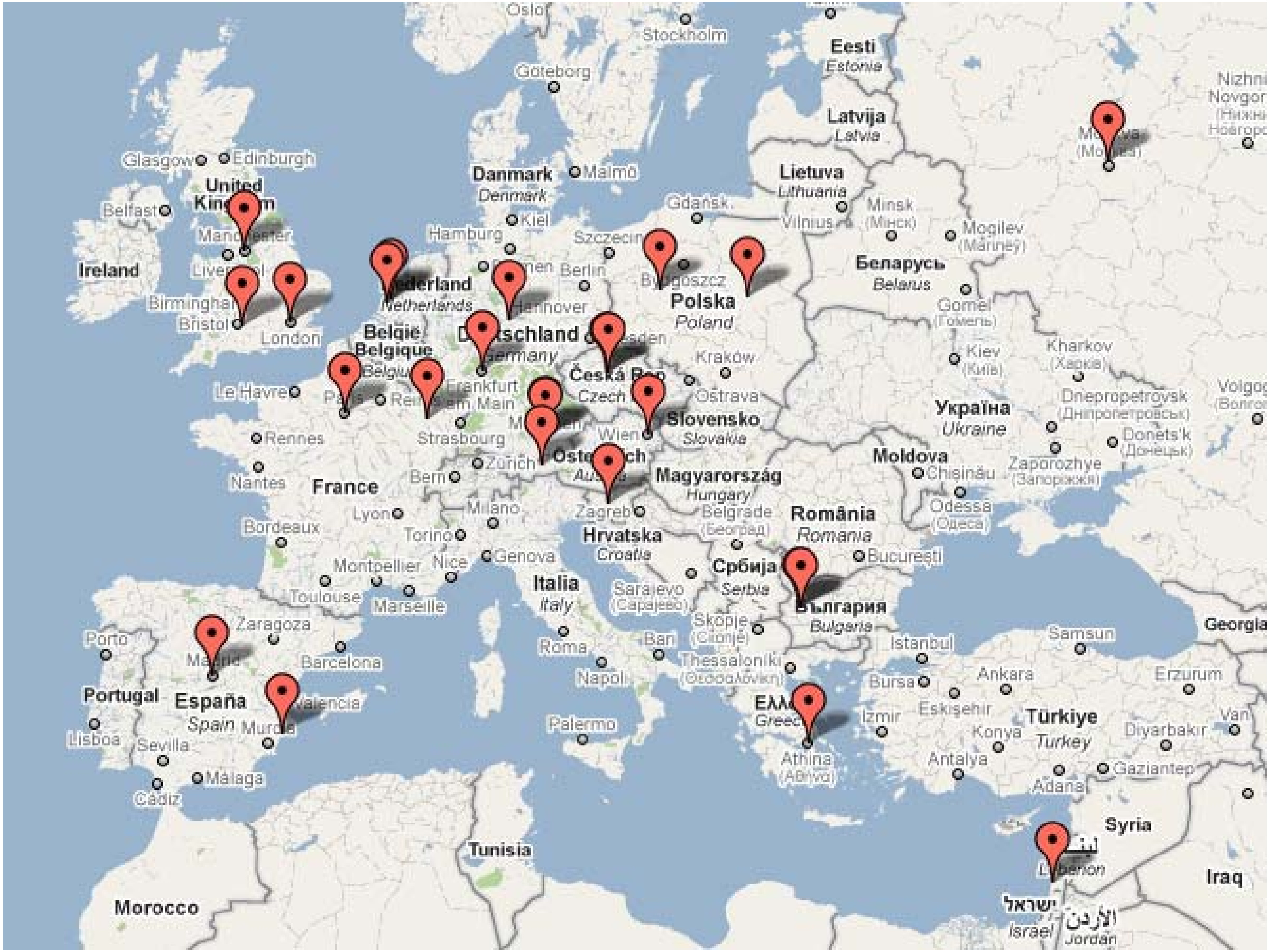
Answering the challenges – IMPACT

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

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

- Main objectives:

  - Innovate OCR & language technology
  - Build capacity in mass-digitisation


Oslo
Stockholm
Eesti
Estonia
Göteborg
Nizhni
Novgor
Latvija
Latvia
Glasgow
Edinburgh
Danmark
Denmark
Malmö
Lietuva
Lithuania
United
Kingdom
Belfast
Kiel
Gdańsk
Minsk
(Минск)
Vilnius
Mogilev
(Магілеў)
Manchester
Hamburg
Szczecin
Ireland
Liverpool
Bremen
Berlin
Bydgoszcz
Беларусь
Belarus
Nederland
Netherlands
Birmingham
Hannover
Polska
Poland
Gomel
(Гомель)
Bristol
London
België
Belgique
Deutschland
Germany
Dresden
Kiev
(Київ)
Kharkov
(Харків)
Le Havre
Paris
Reims
Frankfurt
am Main
Česká Rep.
Czech
Kraków
Ostrava
Україна
Ukraine
Dnepropetrovsk
(Дніпропетровськ)
Volgog
Rennes
Strasbourg
München
Wien
Slovensko
Slovakia
Donets'k
(Донецьк)
Nantes
France
Bern
Zürich
Österreich
Austria
Magyarország
Hungary
Moldova
Chişinău
Zaporozhye
(Запоріжжя)
Lyon
Milano
Zagreb
Belgrade
(Београд)
România
Romania
Odessa
(Одеса)
Bordeaux
Torino
Hrvatska
Croatia
Genova
Nice
Montpellier
Српбија
Serbia
Bucureşti
Toulouse
Marseille
Italia
Italy
Sarajevo
(Сарајево)
България
Bulgaria
Zaragoza
Porto
Skopje
(Скопје)
Georgia
Roma
Bari
Samsun
Istanbul
Madrid
Barcelona
Thessaloniki
(Θεσσαλονίκη)
Portugal
España
Spain
Napoli
Bursa
Ankara
Erzurum
Valencia
Ελλάδα
Greece
Eskişehir
Lisboa
Murcia
Palermo
Izmir
Türkiye
Turkey
Van
Sevilla
Athina
(Αθήνα)
Konya
Diyarbakır
Málaga
Antalya
Gaziantep
Cádiz
Adana
Syria
Tunisia
Lebanon
Iraq
Morocco
ישראל
Israel
الأردن
Jordan

IMPACT Main Achievements at this point

- Improved commercial OCR (ABBYY ‘IMPACT’ Finereader 10 on market)
- Effective tool for OCR correction with volunteer involvement (CONCERT) ready for implementation
- Novel Approaches to preprocessing, OCR and post correction available
- Computerlexica for nine languages close to delivery
- Digitisation Framework with evaluation tools available
- Facility to plug in other tools (Your tools!)
- Large Dataset with sophisticated Ground Truth close to final delivery
- Knowledge bank with guidelines and learning resources close to delivery
- A unique network bringing together experts from different communities
- New website under construction
- Centre of Competence to be launched at final conference 24-25 October 2011

A Centre of Competence in digitisation: Why

- Challenges in digitisation of historic material still there
- No lack of novel approaches to improve access: IMPACT and many others
- Challenge: implementation in real life
- Challenge: direction of research
- Need for real life datasets with Ground Truth
- Need for real life testing and evaluation
- Need for support in implementing the solutions

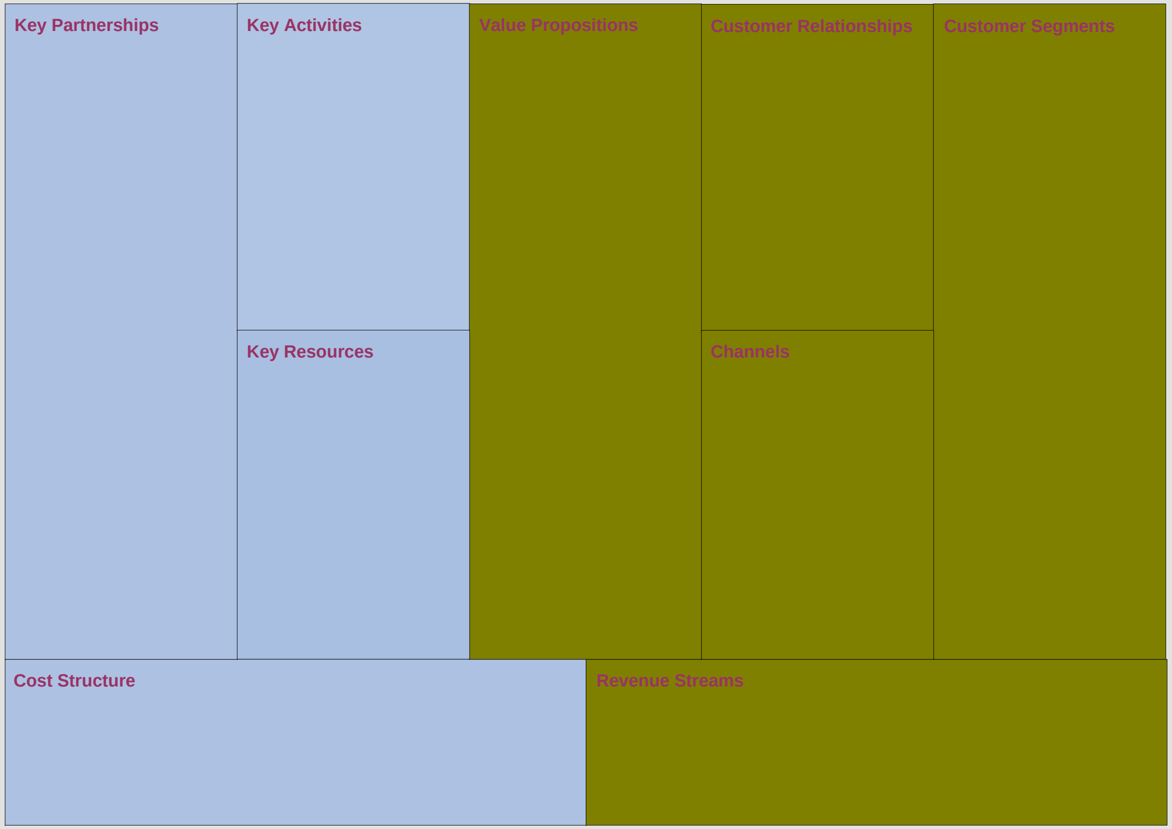
Improving Access to Text

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.

Building the Business Model: Methodology

Business Model Generation (http://www.businessmodelgeneration.com)



Key Partnerships

Key Activities

Key Resources

Value Propositions

Customer Relationships

Channels

Customer Segments

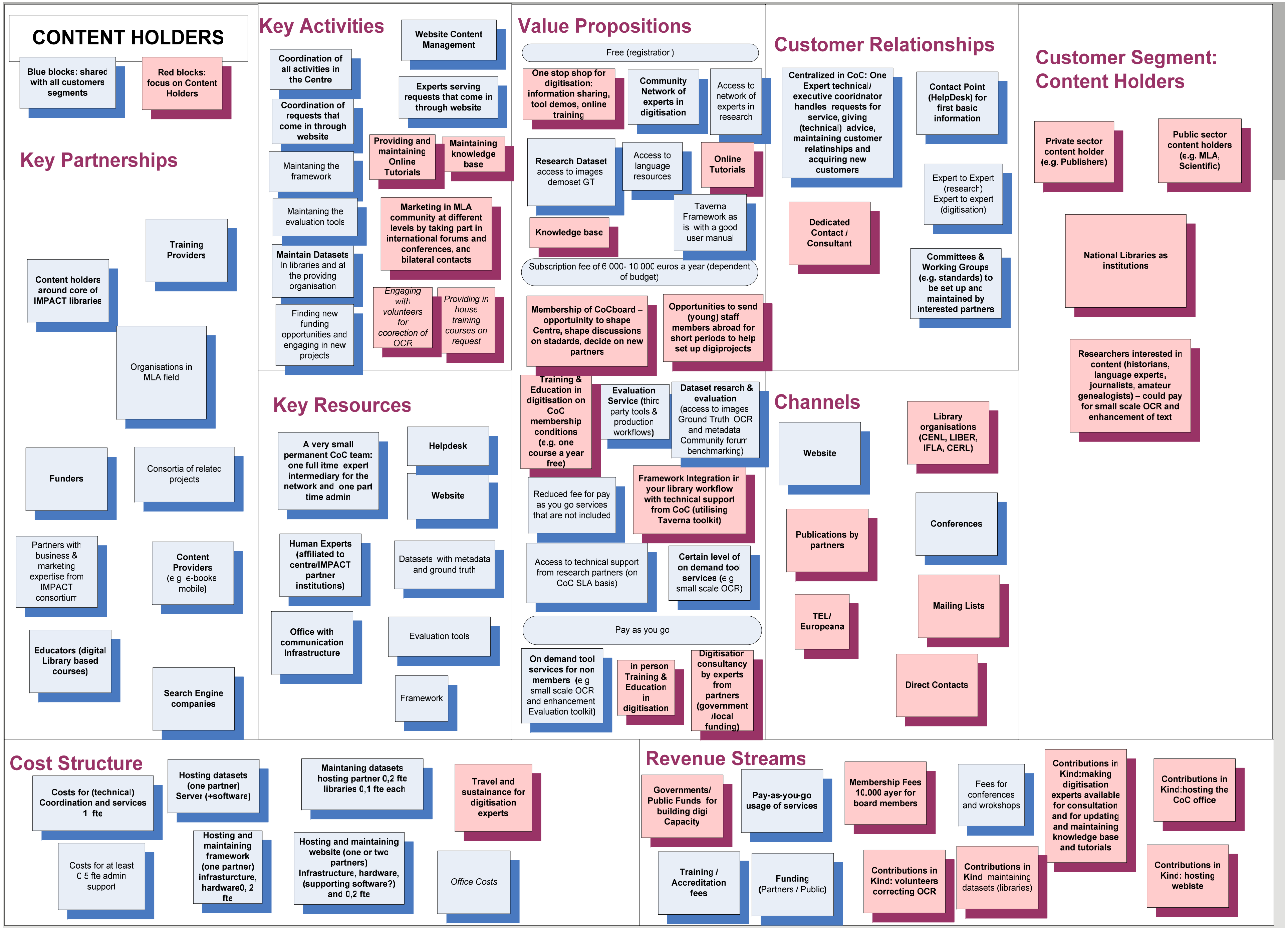
Cost Structure

Revenue Streams

Building the Business Model: sessions

Building the Business Model: Gathering ideas

Building the Business Model: Finding commitment



CONTENT HOLDERS

- Blue blocks: shared with all customers segments
- Red blocks: focus on Content Holders

Key Partnerships

- Training Providers
- Content holders around core of IMPACT libraries
- Organisations in MLA field
- Funders
- Consortia of relatec projects
- Partners with business & marketing expertise from IMPACT consortiurr
- Content Providers (e g e-books mobile)
- Educators (digital Library based courses)
- Search Engine companies

Key Activities

- Website Content Management
- Coordination of all activities in the Centre
- Experts serving requests that come in through website
- Coordination of requests that come in through website
- Providing and maintaining Online Tutorials
- Maintaining knowledge base
- Maintaning the framework
- Maintaning the evaluation tools
- Marketing in MLA community at different levels by taking part in international forums and conferences, and bilateral contacts
- Maintain Datasets In libraries and at the providng organisation
- Engaging with volunteers for coorection of OCR
- Providing in house training courses on request
- Finding new funding opportunities and engaging in new projects

Key Resources

- A very small permanent CoC team: one full itme expert intermediary for the network and one part time admin
- Helpdesk
- Website
- Human Experts (affiliated to centre/IMPACT partner institutions)
- Datasets with metadata and ground truth
- Office with communication Infrastructure
- Evaluation tools
- Framework

Value Propositions

Free (registratior)

- One stop shop for digitisation: information sharing, tool demos, online training
- Community Network of experts in digitisation
- Access to network of experts in research
- Research Dataset access to images demoset GT
- Access to language resources
- Online Tutorials
- Knowledge base
- Taverna Framework as is with a good user manual

Subscription fee of 6 000- 10 000 euros a year (dependent of budget)

- Membership of CoCboard – opportuinity to shape Centre, shape discussions on stadards, decide on new partners
- Opportunities to send (young) staff members abroad for short periods to help set up digiprojects
- Training & Education in digitisation on CoC membership conditions (e.g. one course a year free)
- Evaluation Service (third party tools & production workflows)
- Dataset resarch & evaluation (access to images Ground Truth OCR and metadata Community forum benchmarking)
- Reduced fee for pay as you go services that are not included
- Framework Integration in your library workflow with technical support from CoC (utilising Taverna toolkit)
- Access to technical support from research partners (on CoC SLA basis)
- Certain level of on demand tool services (e g small scale OCR)

Pay as you go

- On demand tool services for non members (e g small scale OCR and enhancement Evaluation toolkit)
- in person Training & Education in digitisation
- Digitisation consultancy by experts from partners (government /local funding)

Customer Relationships

- Centralized in CoC: One Expert technical/ executive cooridnator handles requests for service, giving (technical) advice, maintaining customer relatinships and acquiring new customers
- Contact Point (HelpDesk) for first basic information
- Expert to Expert (research) Expert to expert (digitisation)
- Dedicated Contact / Consultant
- Committees & Working Groups (e.g. standards) to be set up and maintained by interested partners

Channels

- Website
- Library organisations (CENL, LIBER, IFLA, CERL)
- Publications by partners
- Conferences
- TEL/ Europeana
- Mailing Lists
- Direct Contacts

Customer Segment: Content Holders

- Private sector content holder (e.g. Publishers)
- Public sector content holders (e.g. MLA, Scientific)
- National Libraries as institutions
- Researchers interested in content (historians, language experts, journalists, amateur genealogists) – could pay for small scale OCR and enhancement of text

Cost Structure

- Costs for (technical) Coordination and services 1 fte
- Hosting datasets (one partner) Server (+software)
- Maintaning datasets hosting partner 0,2 fte libraries 0,1 fte each
- Travel and sustainance for digitisation experts
- Costs for at least 0 5 fte admin support
- Hosting and maintaining framework (one partner) infrasturcture, hardware0, 2 fte
- Hosting and maintaining website (one or two partners) Infrastructure, hardware, (supporting software?) and 0,2 fte
- *Office Costs*

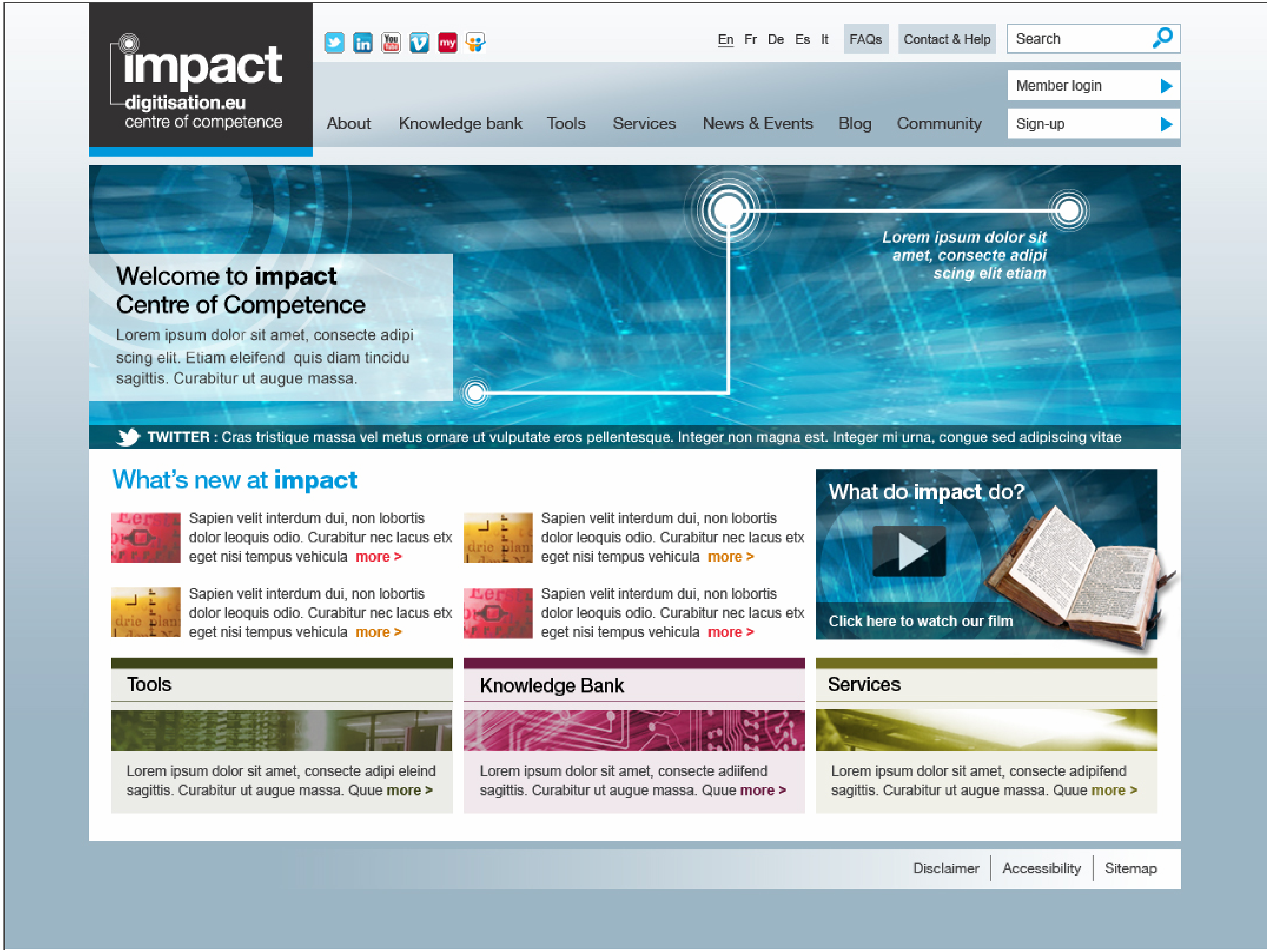
Revenue Streams

- Governments/ Public Funds for building digi Capacity
- Pay-as-you-go usage of services
- Membership Fees 10.000 ayer for board members
- Fees for conferences and wrokshops
- Contributions in Kind:making digitisation experts available for consultation and for updating and maintaining knowledge base and tutorials
- Contributions in Kind:hosting the CoC office
- Training / Accreditation fees
- Funding (Partners / Public)
- Contributions in Kind: volunteers correcting OCR
- Contributions in Kind maintaining datasets (libraries)
- Contributions in Kind: hosting webiste

Centre of Competence: What*

- Not for profit organisation
- Mission: to support the process of making Europe’s heritage accessible in digital form
- Web based international community with small core facility for support
- Curates IMPACT achievements and provides tools, services and facilities for further advancement of the State of the Art in this field
- Focuses on practical solutions
- Distributed effort by IMPACT partners spreads risk and ensures continuing engagement
- Income generation by offering number of resources and services at a fee (mix of subscription and pay as you go)

**pending approval of IMPACT consortium*



impact digitisation.eu centre of competence

En Fr De Es It FAQs Contact & Help Search

Member login

Sign-up

About Knowledge bank Tools Services News & Events Blog Community

Welcome to **impact** Centre of Competence

Lorem ipsum dolor sit amet, consecte adipi scing elit. Etiam eleifend quis diam tincidu sagittis. Curabitur ut augue massa.

Lorem ipsum dolor sit amet, consecte adipi scing elit etiam

TWITTER : Cras tristique massa vel metus ornare ut vulputate eros pellentesque. Integer non magna est. Integer mi urna, congue sed adipiscing vitae

What's new at **impact**

Sapien velit interdum dui, non lobortis dolor leoquis odio. Curabitur nec lacus etx eget nisi tempus vehicula more >

Sapien velit interdum dui, non lobortis dolor leoquis odio. Curabitur nec lacus etx eget nisi tempus vehicula more >

Sapien velit interdum dui, non lobortis dolor leoquis odio. Curabitur nec lacus etx eget nisi tempus vehicula more >

Sapien velit interdum dui, non lobortis dolor leoquis odio. Curabitur nec lacus etx eget nisi tempus vehicula more >

What do **impact** do?

Click here to watch our film

Tools

Lorem ipsum dolor sit amet, consecte adipi eleind sagittis. Curabitur ut augue massa. Quue **more >**

Knowledge Bank

Lorem ipsum dolor sit amet, consecte adiifend sagittis. Curabitur ut augue massa. Quue **more >**

Services

Lorem ipsum dolor sit amet, consecte adipifend sagittis. Curabitur ut augue massa. Quue **more >**

Disclaimer | Accessibility | Sitemap

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

Centre of Competence: Who

- Digitisation practitioners in content holding institutions
- Researchers in the field of historical document processing and language technology
- Service providers
- End users (Researchers in the humanities who need access to the content)

- IMPACT Centre of competence offers distinct value to each of these target groups

Centre of Competence*: benefits for content holders**

- Exchange of best practice in community of content holders with digitisation programmes
- KnowledgeBank with comprehensive and up to date information and technology watch reports
- Training on demand and on line tutorials
- Online support through a Helpdesk
- Support in the implementation of the innovative IMPACT solutions for improving access to text
- Access to the IMPACT Dataset with Ground Truth and tools for evaluation
- Digitisation Framework: Guidelines on using the open source workflow management system Taverna in a digitisation workflow
- Language Resources: a set of historical lexica that can be utilised within the digitisation workflow
- Conferences/workshops with focus on demonstration and implementation
- Working together on future practical solutions with scientific communities in the area of pattern recognition, language technology, image processing

**pending approval of IMPACT consortium **access depends on membership level*

Centre of Competence*: Benefits for researchers**

- New community: Bridging the gap between specialist research and real life needs
- Brings together scientific communities in the area of pattern recognition, language technology, image processing with a focus on large scale digitisation
- Access to content holding community
- Access to large real life datasets and ground truth
- Working on implementation of research prototypes/products into digitisation environment
- Facilities for testing and evaluating new tools and IMPACT tools
- Working groups and comittees
- Access to new projects and funding opportunities
- Conferences/workshops with focus on demonstration and implementation

**pending approval of IMPACT consortium **access depends on membership level*

Centre of Competence*: Benefits for service providers**

- Access to content holding community with large scale digitisation programmes
- Access to large real life datasets and ground truth
- Facilities for testing and evaluating new tools and IMPACT tools
- Attend conferences
- Working groups and comittees
- Make yourself known to your clients through yellow pages on website
- Sponsorsship and exhibition opportunities
- Working together with content holders and researchers on practical solutions

**pending approval of IMPACT consortium **access depends on membership level*

Join the Centre by becoming a member!

Three levels of membership (pending final decision IMPACT consortium) :

- Open (registration) access to forum, part of content
- Basic membership (fee): access to all facilities, reduced fee for conferences
- Premium membership (fee): member of the Board, privileges such as free entry to conferences

Want to sign up?

→ Mail to impact@kb.nl for information on membership

→ Join us now already on LinkedIn

→ Follow us on Twitter (@impactocr)

→ Access through www.impact-project.eu

→ Come to our conference on 24 and 25 October

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:

www.impact-project.eu

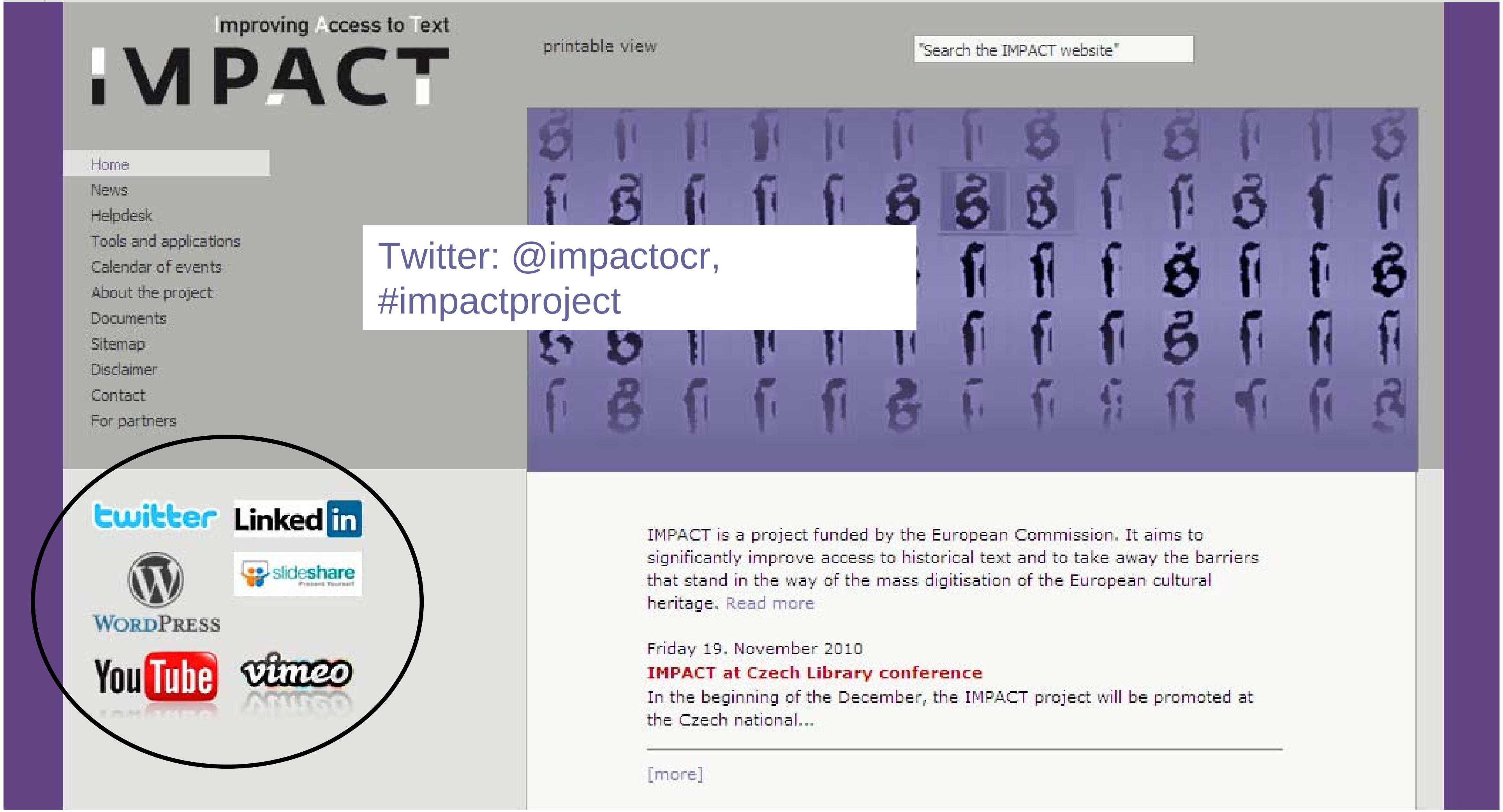
Improving Access to Text

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.

www.impact-project.eu

Twitter: @impactocr, #impactproject



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]