impact
digitisation.eu
centre of competence

Tools for text digitisation and transcription

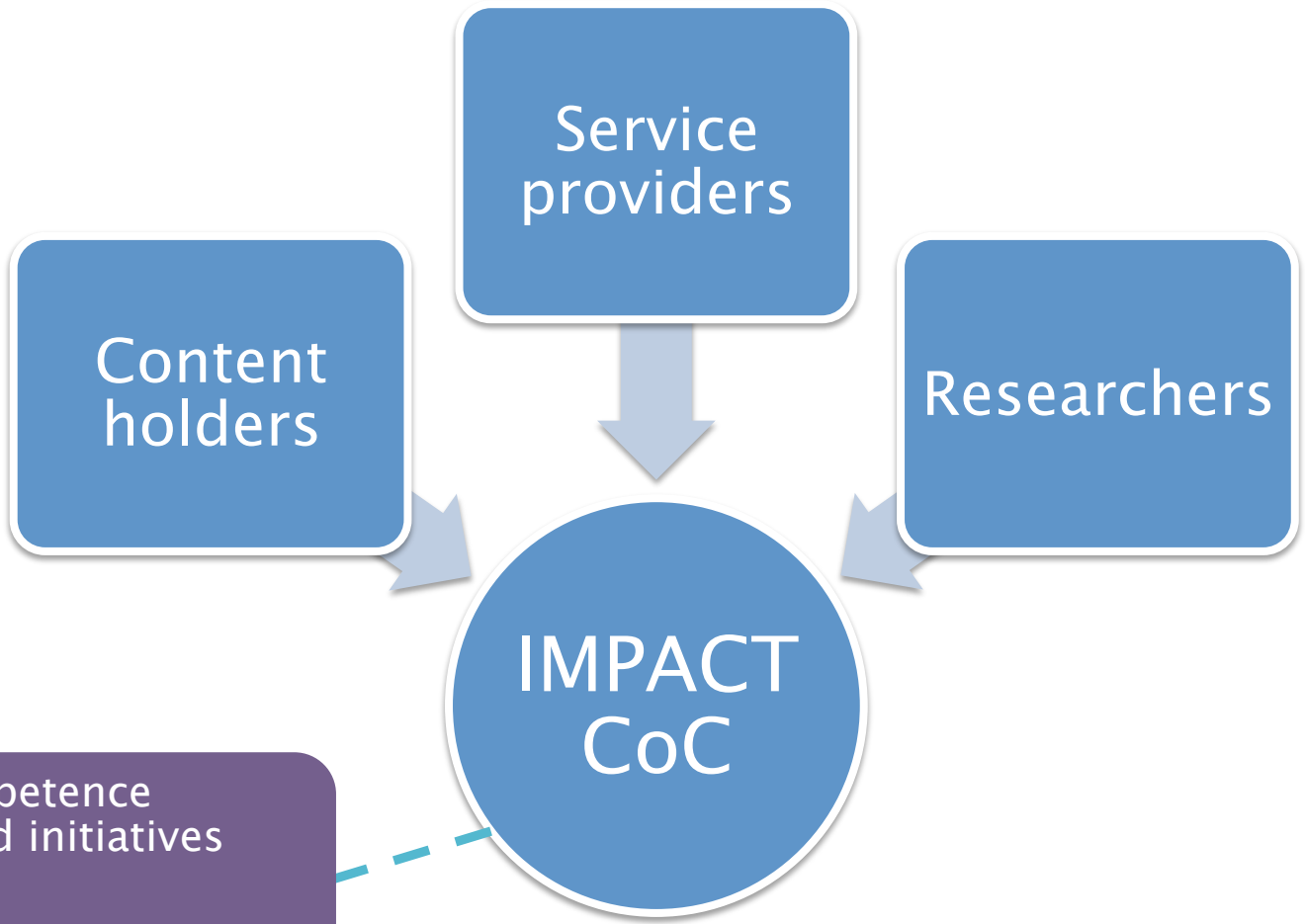
Tomasz Parkoła
Poznan Supercomputing and Networking Center

CERL annual seminar, 28.10.2014, Oslo, Norway

Agenda

- IMPACT Center of Competence
- Expertise, tools & events
- PSNC example
- Summary

IMPACT Centre of Competence in digitisation



- Other competence centres and initiatives
- Europeana
- Research infrastructures

IMPACT CoC members

Premium members

- Biblioteca Nacional de España
- Bibliothèque nationale de France
- British Library
- Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung
- Fundación Biblioteca Virtual Miguel de Cervantes (Management and headquarters)
- Instituut voor Nederlandse Lexicologie
- Koninklijke Bibliotheek
- Contentra Technologies (formerly Planman Technologies)
- Poznań Supercomputing and Networking Center
- Universidad de Alicante

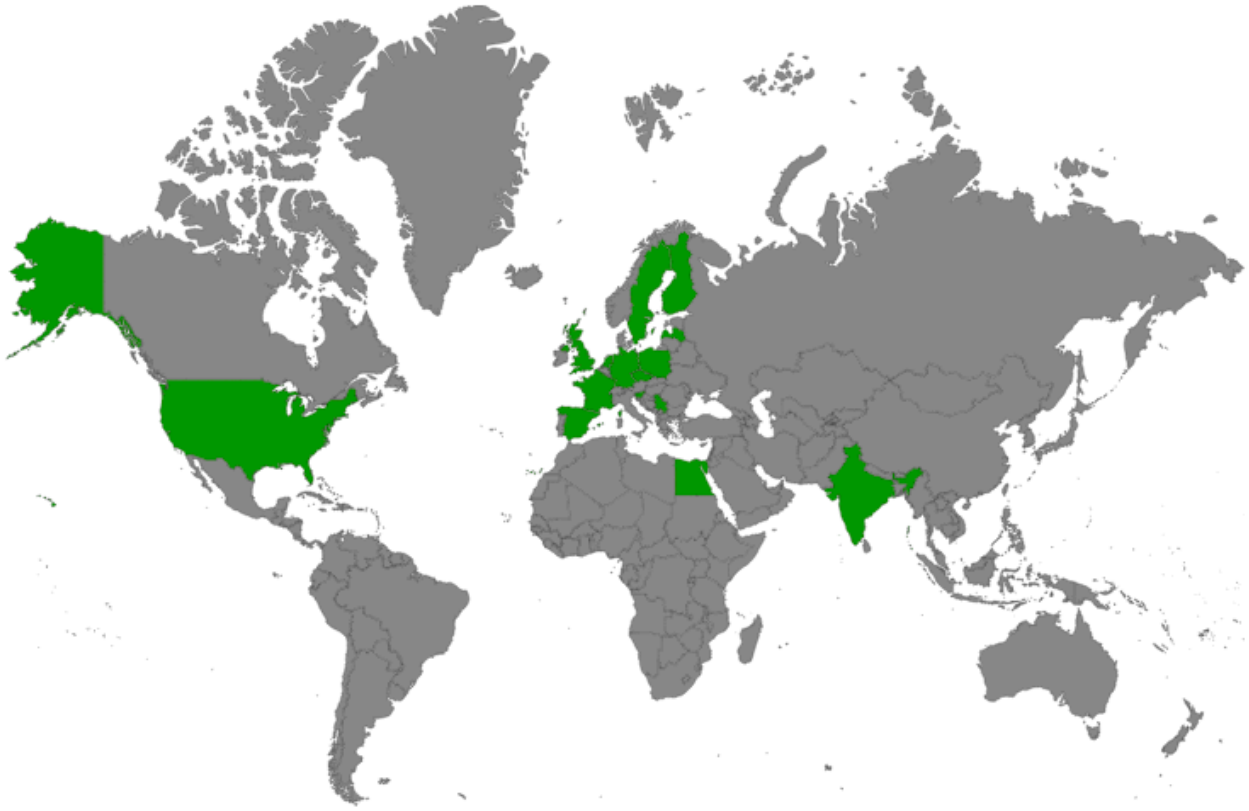
Standard members

- Biblioteka Uniwersytecka we Wrocławiu (Wroclaw University Library)
- California Digital Library
- Centro de documentación teatral
- DIGIBIS
- Elzaburu
- Göteborgs Universitet
- Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen (University Library Centre of North Rhine-Westphalia)
- i2s Digibook
- KU Leuven
- Kungliga Biblioteket (National Library of Sweden)
- LIBNOVA
- Ludwig-Maximilians-Universität, Centrum für Informations- und Sprachverarbeitung

Standard members (cont.)

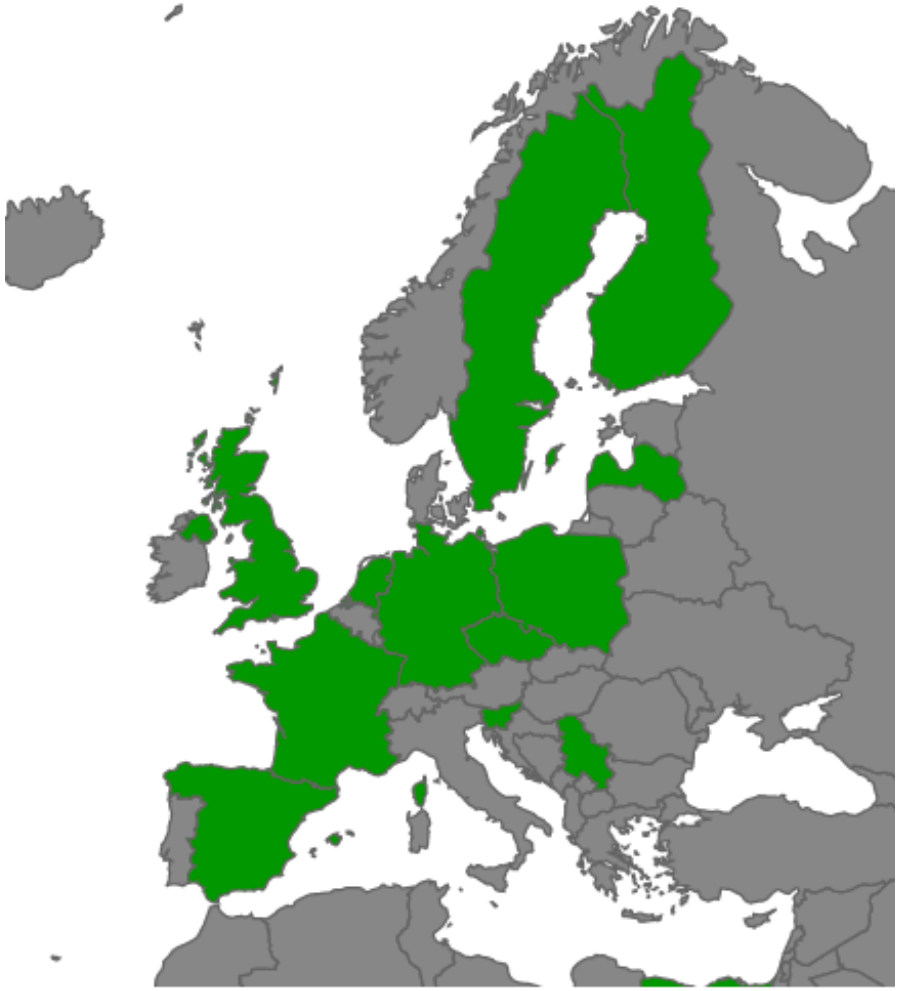
- Narodna in univerzitetna knjižnica (National Library of Slovenia)
- National Library of Czech Republic
- National Library of Egypt
- National Library of Finland
- National Library of Latvia
- National Library of Serbia
- Staats- und Universitätsbibliothek Bremen
- Tecnilógica
- Universitat de Barcelona
- Universidad Complutense de Madrid
- Universidad de Granada
- Universidad de Murcia
- Universidad de Salamanca
- Universidad de Valladolid
- University of Salford
- Vinfra

IMPACT CoC members



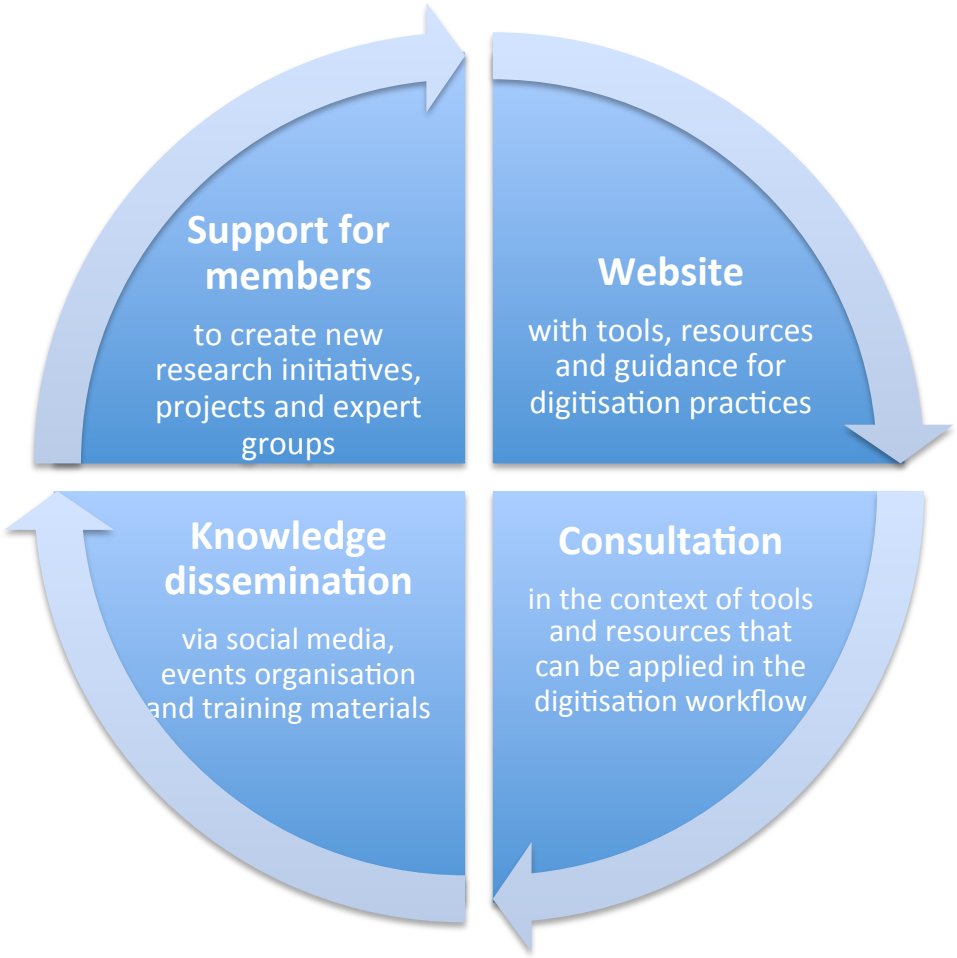
source: http://www.amcharts.com/visited_countries/

IMPACT CoC members



source: http://www.amcharts.com/visited_countries/

Main activities led by IMPACT CoC



Key benefits for members

Cultural heritage

- Access, validate and identify best digitisation technologies
- Meet the experts and define best practices
- Share experience and guide innovation

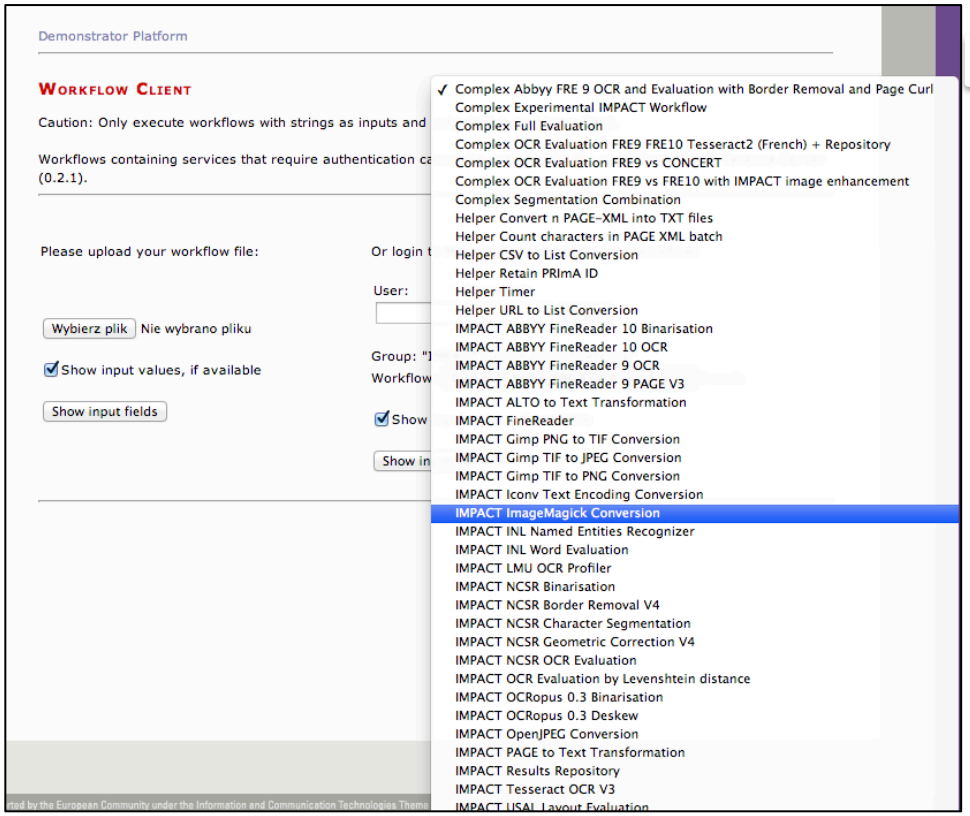
Research centres

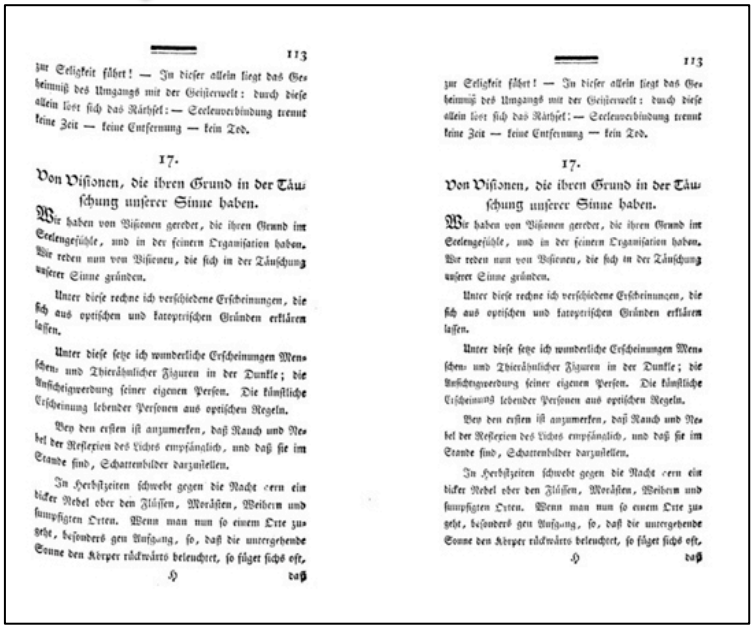
- Learn about research challenges
- Collaborate and provide solution
  - Find project partners, sponsors or facilitators
- Share knowledge and experience

Companies

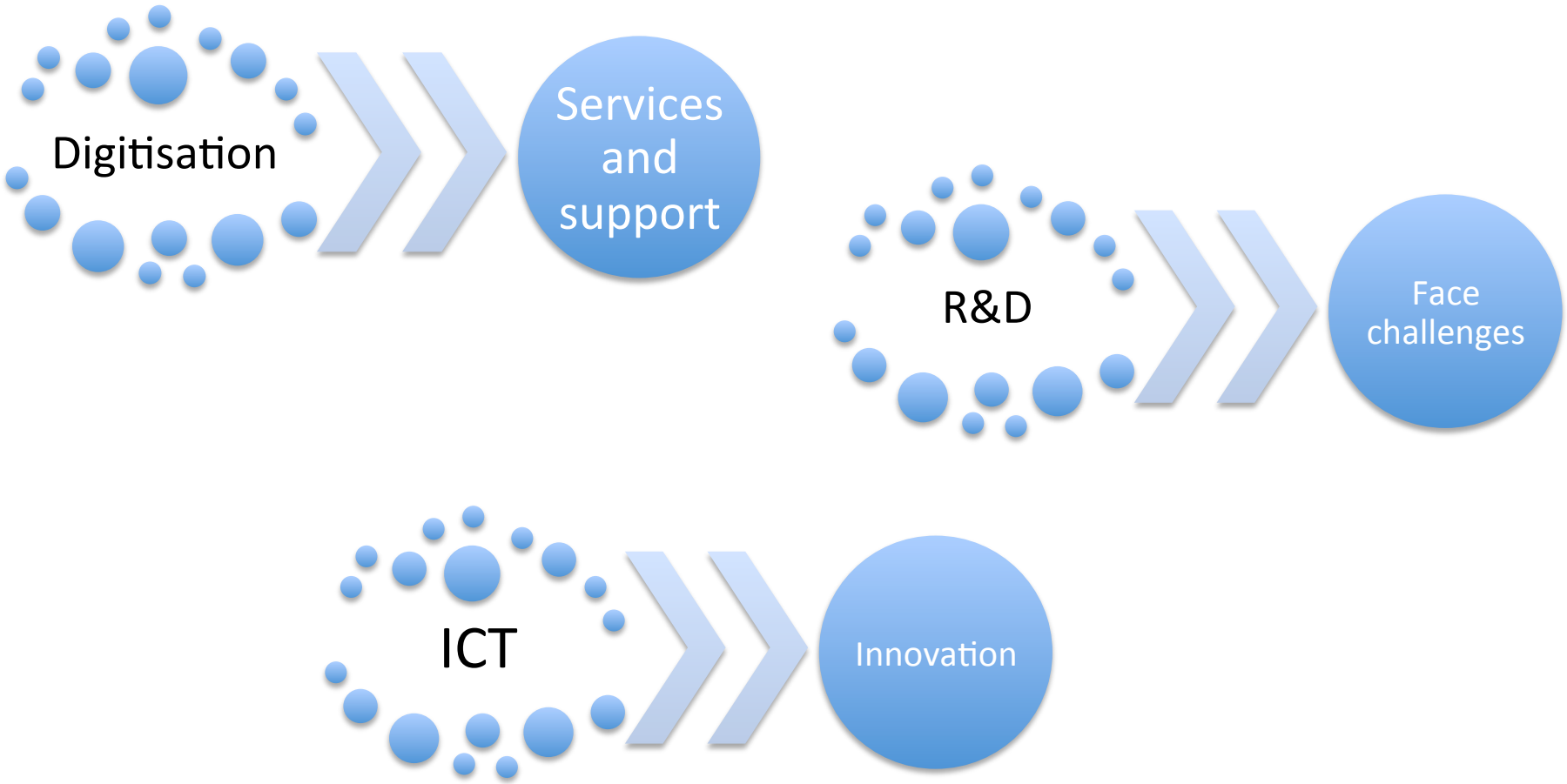
- Showcase your tools and services
- Meet your target customers
- Introduce innovation

Example:
Geometric correction in the demonstrator platform



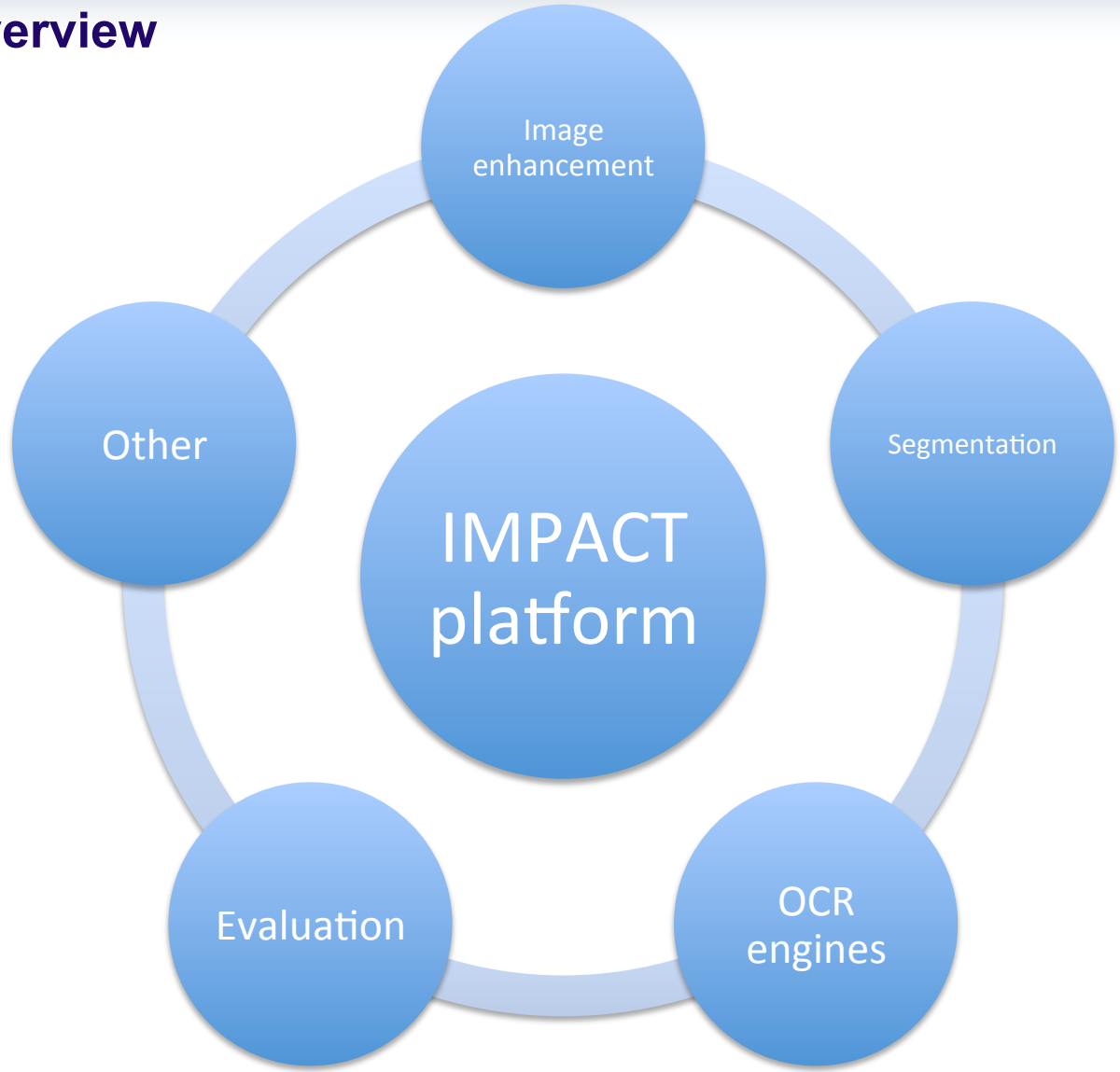


Expertise and experience: overview



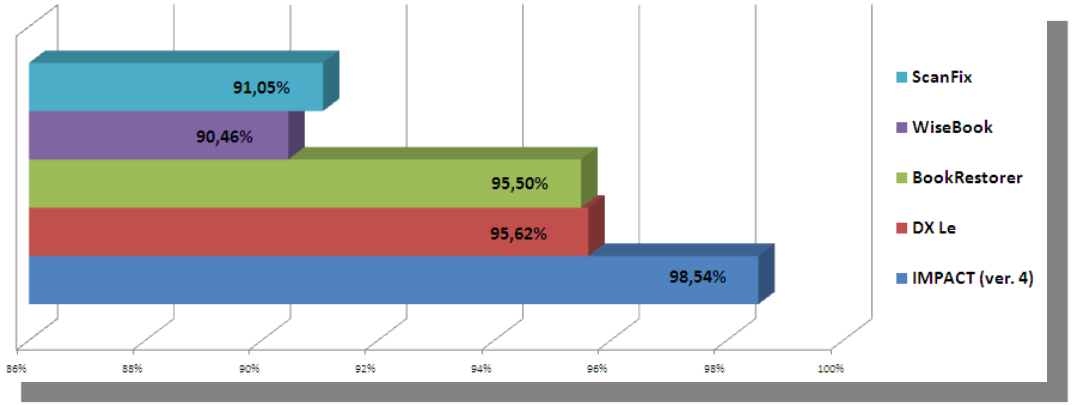
Expertise and experience: examples

Tools: overview



Tools: examples (http://digitisation.eu/demonstrator-platform)

Image enhancement	NCSR Border removal
	NCSR Geometric Correction
	NCSR Binarisation
	Abbyy FineReader 10 Binarisation
	Unpaper



Tools: examples (http://digitisation.eu/demonstrator-platform)

Segmentation	Abbyy FineReader 10 Segmentation
	Uni. Salford region, line, word Segmentation Service
	NCSR character segmentation.
	Uni. Innsbruck

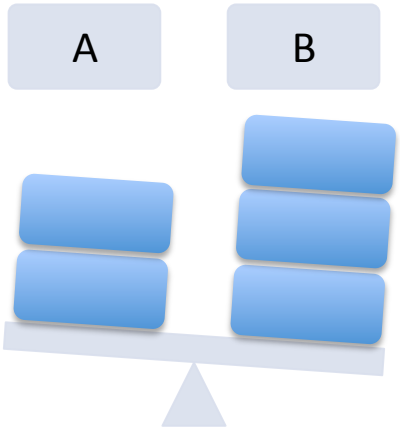
Tools: examples (http://digitisation.eu/demonstrator-platform)

OCR engines	Abbyy FineReader 10 OCR
	Abbyy FineReader 10 with external dictionary
	Uni. Salford Typewritten OCR
	Tesseract 3.00
	Gocr
	Ocropus
	Cuneiform

Tools: examples (http://digitisation.eu/demonstrator-platform)

Evaluation

- NCSR OCR Evaluation service
- Uni. Salford layout evaluation service
- INL word evaluation service



Tools: examples (http://digitisation.eu/demonstrator-platform)

Other	ALTO and PAGE XML transformation
	Uni. Salford ground-truth normalisation
	Uni. Salford PAGE XML to svg
	JP2
	Exif

Resources

- Linguistic data
  - OCR/IR lexica: Slovene, German, Spanish, Czech, Polish, Dutch, English, French and other coming soon.
- Images and ground truth
  - Czech, Spanish, Polish, Bulgarian, Slovene, Biodiversity Heritage Library and other coming soon.
- Annotated corpora
  - Spanish.

Recent activities

- Past events (2013-2014)
  - Developer workshop for tools integration
  - TPDL tutorial (state of the art tools for text digitisation)
  - Digitisation Days, DATecH and Succeed awards
  - Workshop to investigate interoperability issues in digitisation
- Upcoming events (2014-2015)
  - **November 28th, 2014: Succeed in digitisation. Spreading Excellence. http://www.succeed-project.eu/succeed-digitisation**
  - **September, 2015: TPDL 2015 organised by PSNC (premium member of IMPACT CoC) http://tpdl2015.info/**
  - **2016: Digitisation Days and DATecH**
- Supporting take-up of tools and resources
  - A dozen of cultural heritage institutions validating and integrating state of the art digitisation tools (Tesseract, ScanTailor, ImageMagick, JHOVE, NER, Korrektor, Gimp, Alchemy API, COBaLT, Omnipage, Abbyy FR)
- Cooperation with other initiatives and centres of competence (e.g. Open Preservation Foundation)

PSNC example: who are we?

PSNC is a R&D centre located in Poznan, Poland, focused on ICT in the context of :

- Cloud technologies (archiving & computing) and HPC
- Network technologies (protocols, tools, management)
- **Innovative applications & services**

PSNC example: who are we?

In the context of cultural heritage we provide

- Polish National aggregator for Europeana and others
- **DInGO toolset for digitisation projects with over 100 production-mode deployments**
- **Virtual Transcription Laboratory tool with OCR training module and OCR execution support**
- Expertise (over 20 R&D projects, Digital Libraries Conference, training, workshops, consultancy)

dLab
DIGITISATION WORKFLOW MANAGEMENT

dArceo
LONG-TERM PRESERVATION SERVICES

dLibra
DIGITAL LIBRARY FRAMEWORK

dMuseion
DIGITAL MUSEUM FRAMEWORK

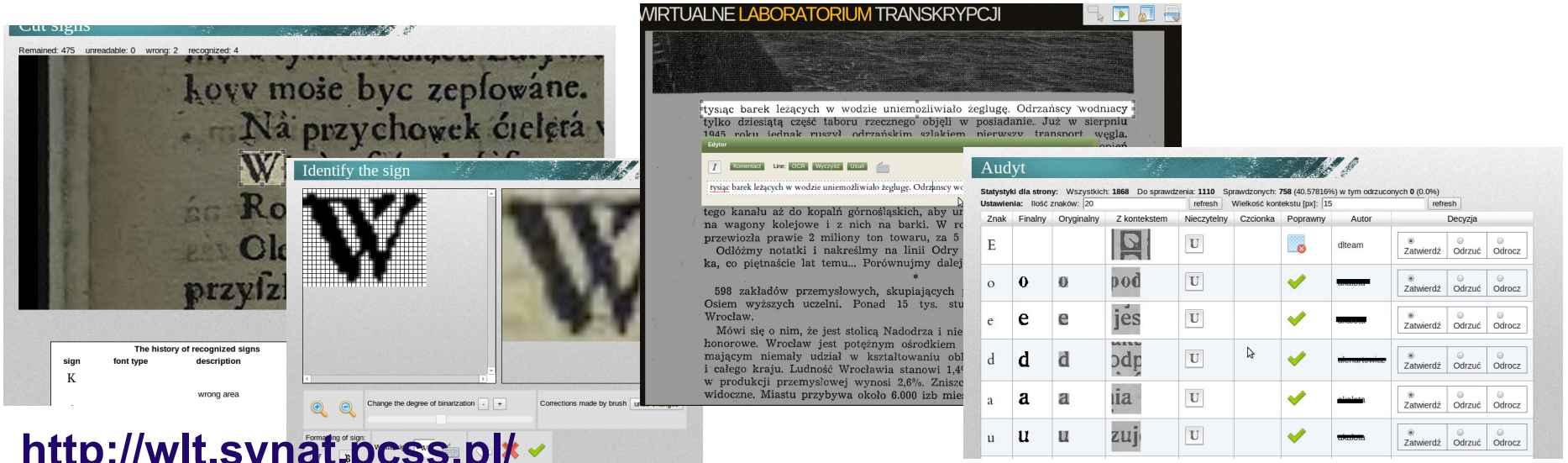
DIGITAL LIBRARIES FEDERATION

VIRTUAL TRANSCRIPTION LABORATORY

PSNC example: Virtual Transcription Laboratory

Web portal which provides access to:

- **Cutouts** tool (creation of customised recognition profiles)
- **OCR engine** with multiple recognition profiles
- **Transcription editor** with QA interface and group work support
- **TXT, hOCR and ePUB** output formats



http://wlt.synat.pcss.pl/

PSNC example: why?

1. High quality, efficient mass digitisation is currently one of the crucial challenges for cultural heritage institutions.
2. PSNC is involved in the IMPACT CoC to help these institutions to overcome existing barriers and to face new challenges in this context.
3. We believe that expertise of IMPACT CoC members can significantly contribute to successful R&D activities and cutting-edge information technologies for digitisation.

Summary: Join us!

Become standard member

- get support in the digitisation programmes
- access part of the IMPACT CoC resources
- meet the experts advice

Become premium member

- steer activities led by IMPACT CoC
- get full access to IMPACT CoC resources
- actively investigate and innovate digitisation

Contact: info@digitisation.eu

impact
digitisation.eu
centre of competence

Thank you!

Tomasz Parkoła
tparkola@man.poznan.pl

CERL annual seminar, 28.10.2014, Oslo, Norway