



Tools for text digitisation and transcription

Tomasz Parkoła
Poznan Supercomputing and Networking Center

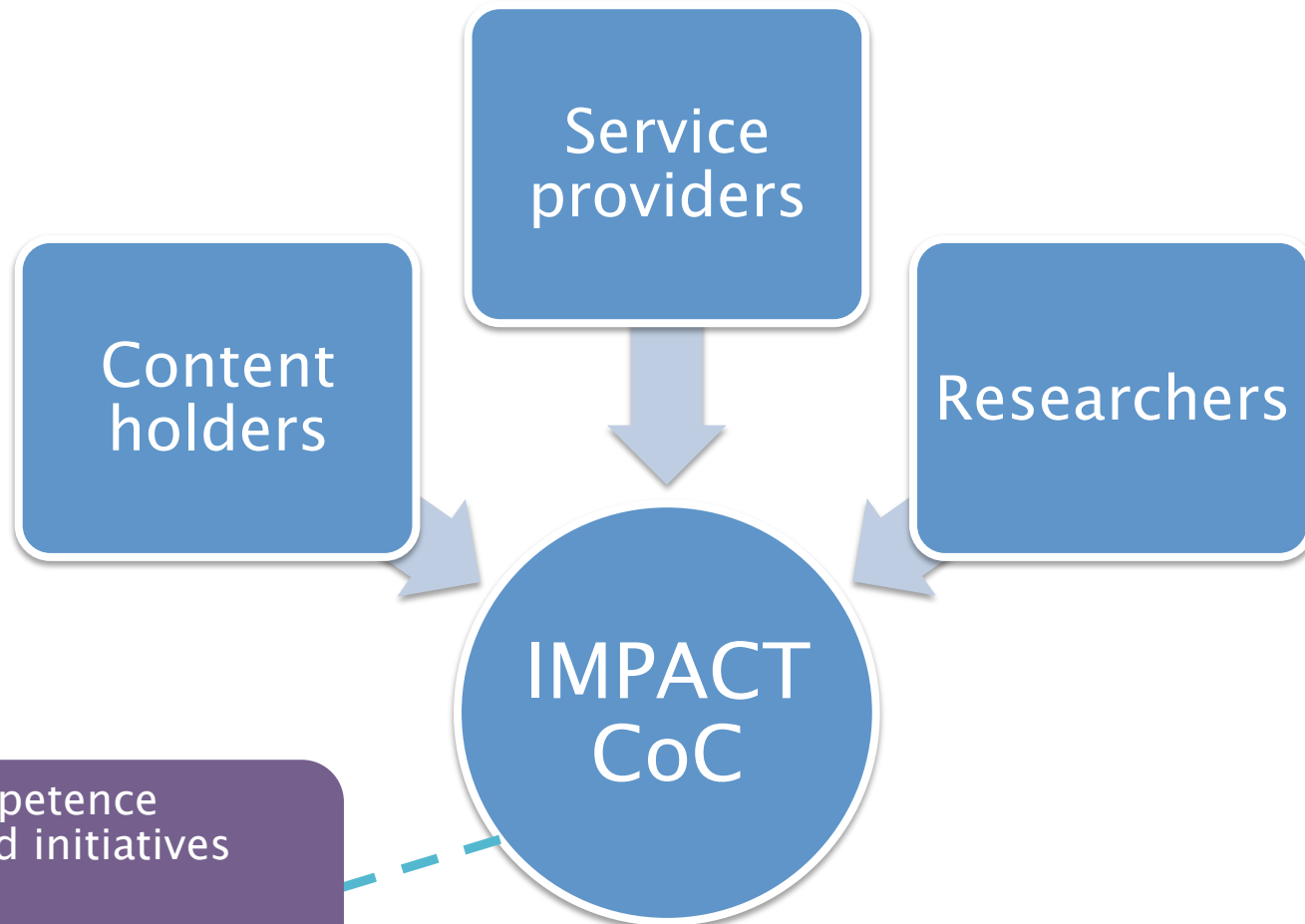


CERL annual seminar, 28.10.2014, Oslo, Norway

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

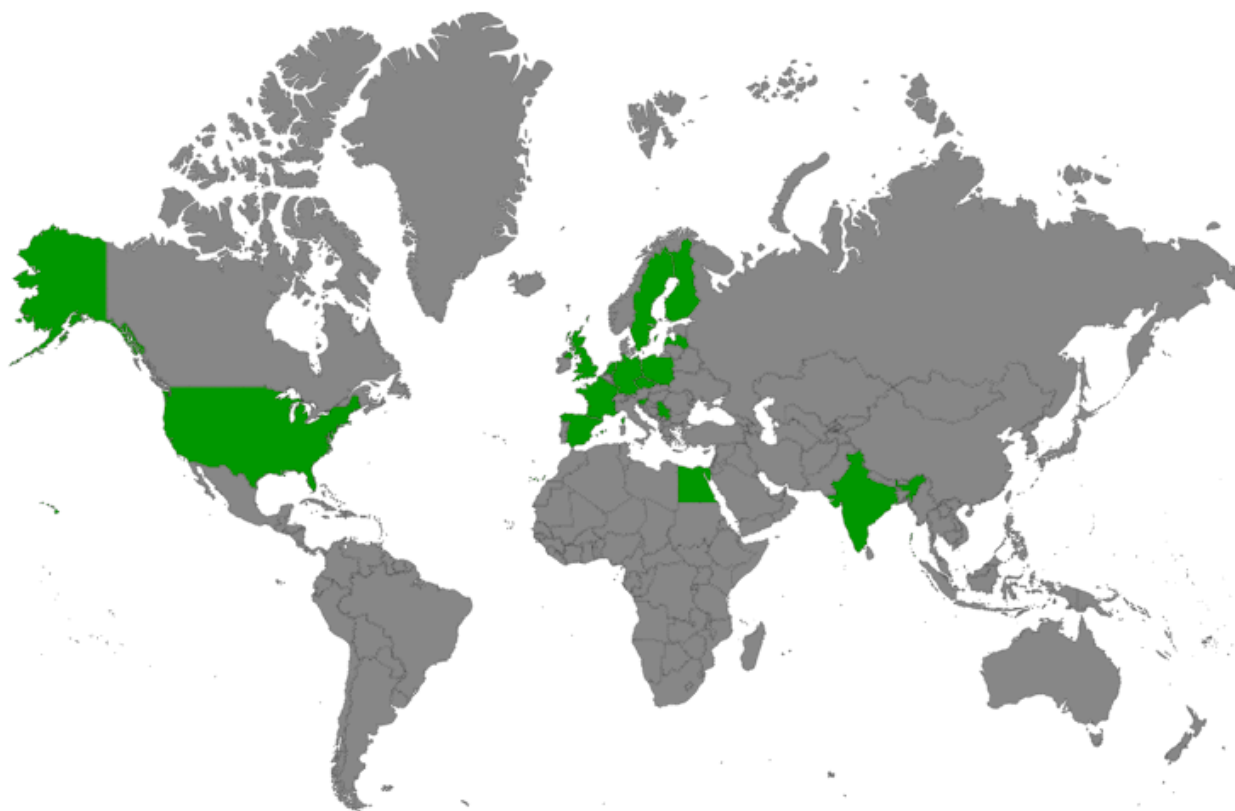
Standard members

- Biblioteka Uniwersytecka we Wrocławiu (Wrocław University Library)
- California Digital Library
- Centro de documentación teatral
- DIGIBIS
- Elzaburu
- Göteborgs Universitet
- Hochschulbibliothekszenrum 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
- Tecnológica
- Universitat de Barcelona
- Universidad Complutense de Madrid
- Universidad de Granada
- Universidad de Murcia
- Universidad de Salamanca
- Universidad de Valladolid
- University of Salford
- Vinfra

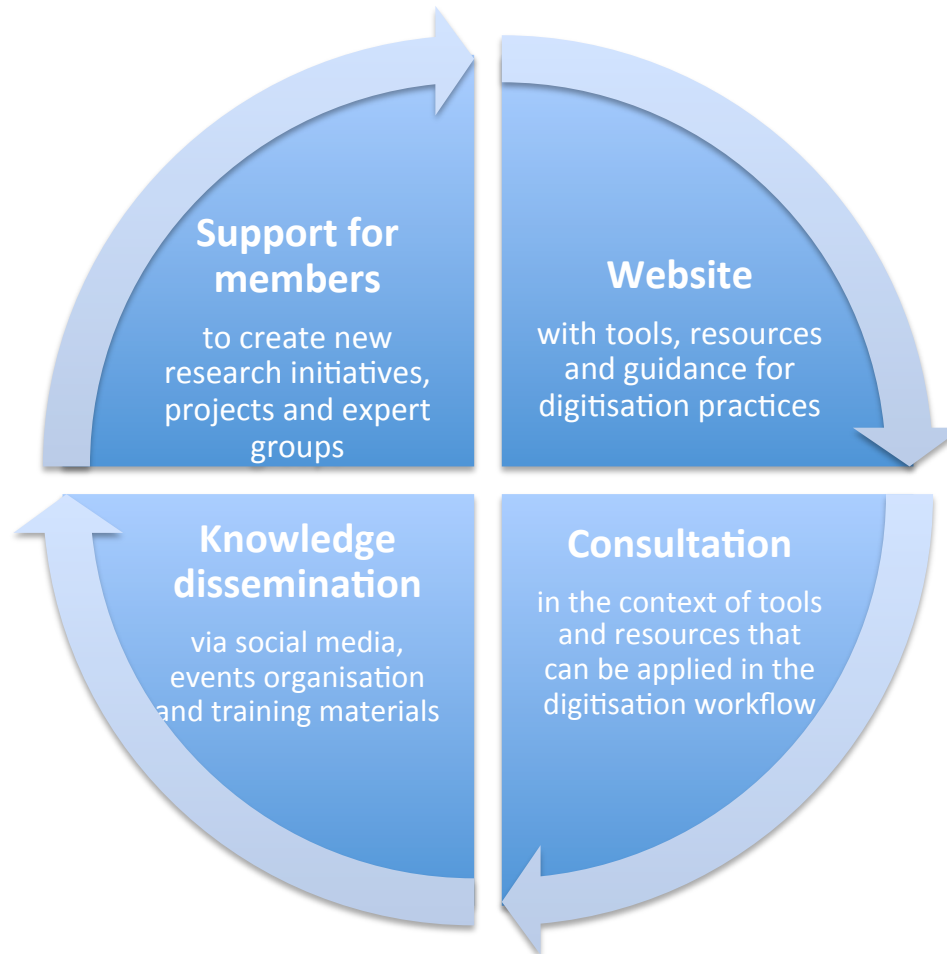
IMPACT CoC members



IMPACT CoC members



Main activities led by IMPACT CoC



Key benefits for members



<h3>Cultural heritage</h3>	<ul style="list-style-type: none">● Access, validate and identify best digitisation technologies● Meet the experts and define best practices● Share experience and guide innovation
<h3>Research centres</h3>	<ul style="list-style-type: none">● Learn about research challenges● Collaborate and provide solution<ul style="list-style-type: none">● Find project partners, sponsors or facilitators● Share knowledge and experience
<h3>Companies</h3>	<ul style="list-style-type: none">● Showcase your tools and services● Meet your target customers● Introduce innovation

Example: Geometric correction in the demonstrator platform

Demonstrator Platform

WORKFLOW CLIENT

Caution: Only execute workflows with strings as inputs and

Workflows containing services that require authentication ca (0.2.1).

Please upload your workflow file:

Nie wybrano pliku

Show input values, if available

Or login t

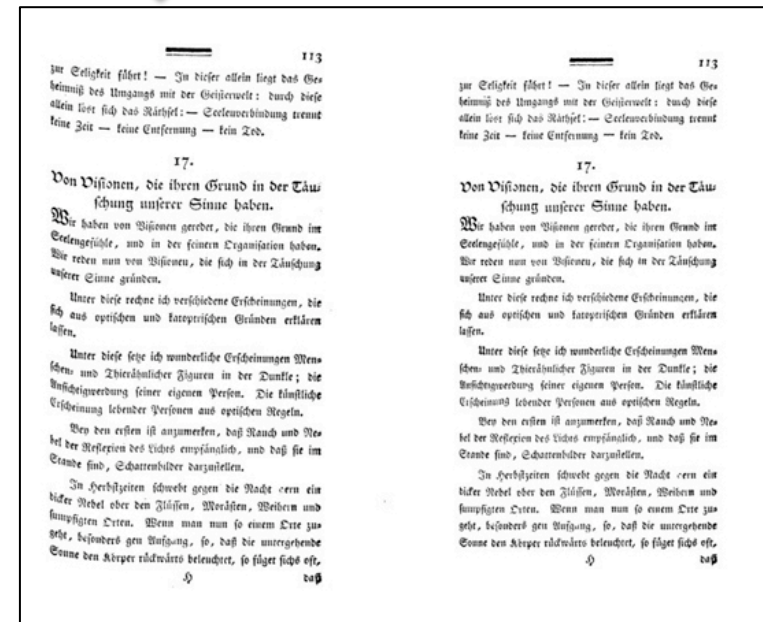
User:

Group: "1"

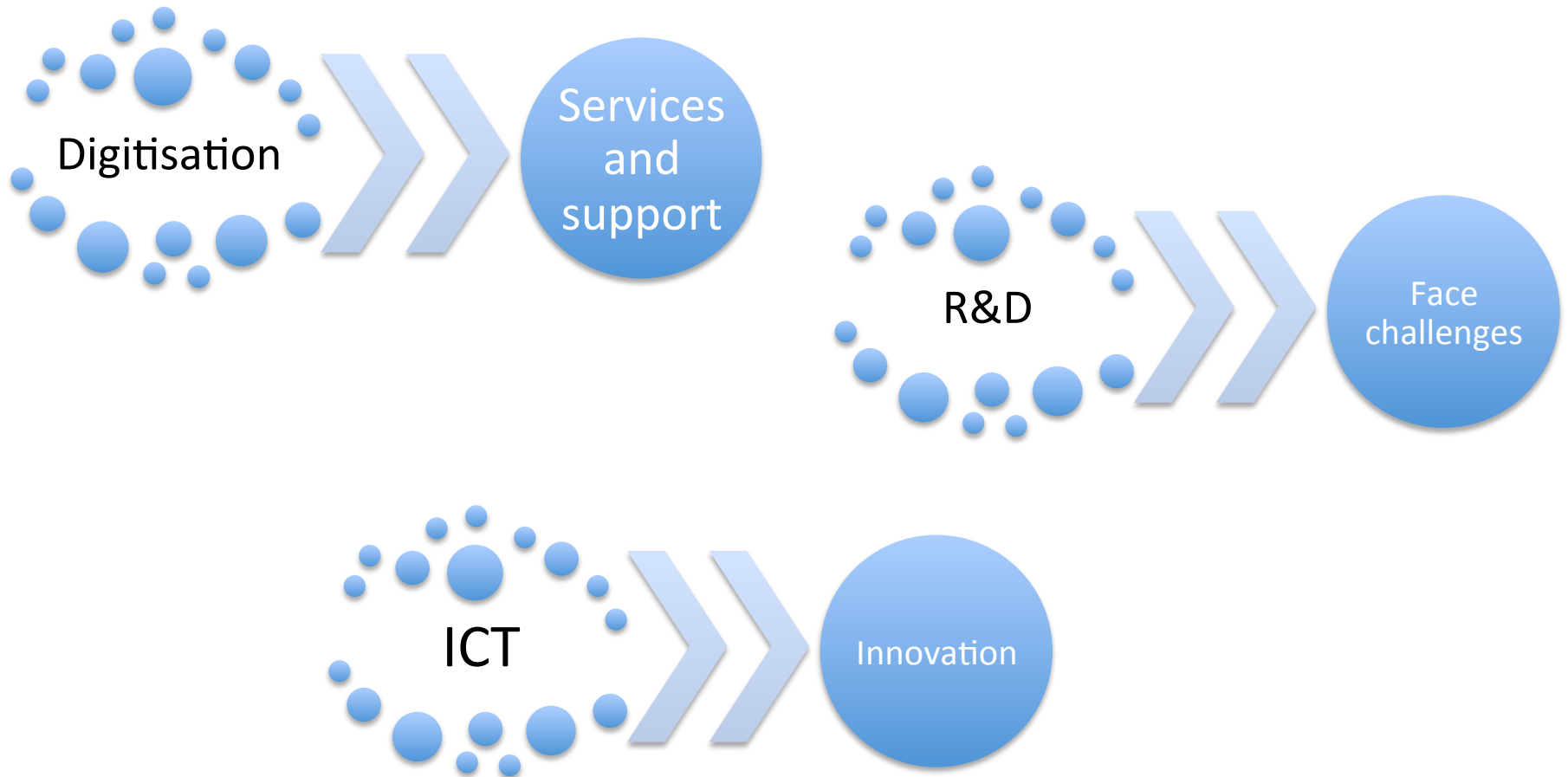
Workflow:

Show

- ✓ Complex Abbyy FRE 9 OCR and Evaluation with Border Removal and Page Curl
- Complex Experimental IMPACT Workflow
- Complex Full Evaluation
- Complex OCR Evaluation FRE9 FRE10 Tesseract2 (French) + Repository
- Complex OCR Evaluation FRE9 vs CONCERT
- Complex OCR Evaluation FRE9 vs FRE10 with IMPACT image enhancement
- Complex Segmentation Combination
- Helper Convert n PAGE-XML into TXT files
- Helper Count characters in PAGE XML batch
- Helper CSV to List Conversion
- Helper Retain PRImA ID
- Helper Timer
- Helper URL to List Conversion
- IMPACT ABBYY FineReader 10 Binarisation
- IMPACT ABBYY FineReader 10 OCR
- IMPACT ABBYY FineReader 9 OCR
- IMPACT ABBYY FineReader 9 PAGE V3
- IMPACT ALTO to Text Transformation
- IMPACT FineReader
- IMPACT Gimp PNG to TIF Conversion
- IMPACT Gimp TIF to JPEG Conversion
- IMPACT Gimp TIF to PNG Conversion
- IMPACT Iconv Text Encoding Conversion
- IMPACT ImageMagick Conversion**
- IMPACT INL Named Entities Recognizer
- IMPACT INL Word Evaluation
- IMPACT LMU OCR Profiler
- IMPACT NCSR Binarisation
- IMPACT NCSR Border Removal V4
- IMPACT NCSR Character Segmentation
- IMPACT NCSR Geometric Correction V4
- IMPACT NCSR OCR Evaluation
- IMPACT OCR Evaluation by Levenshtein distance
- IMPACT OCROPUS 0.3 Binarisation
- IMPACT OCROPUS 0.3 Deskew
- IMPACT OpenJPEG Conversion
- IMPACT PAGE to Text Transformation
- IMPACT Results Repository
- IMPACT Tesseract OCR V3
- IMPACT USAL Layout Evaluation



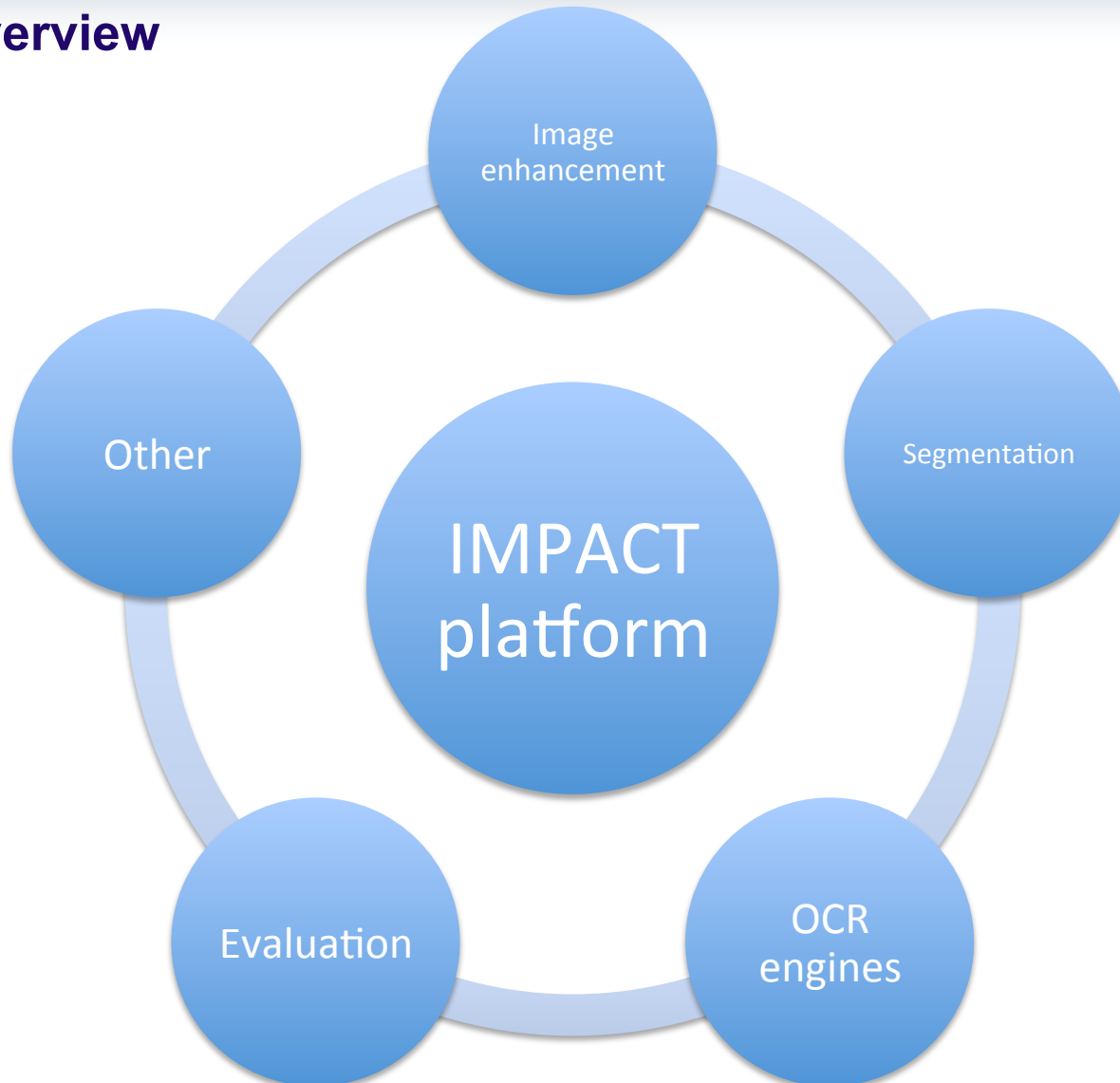
Expertise and experience: overview



Expertise and experience: examples



Tools: overview



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

Image
enhancement

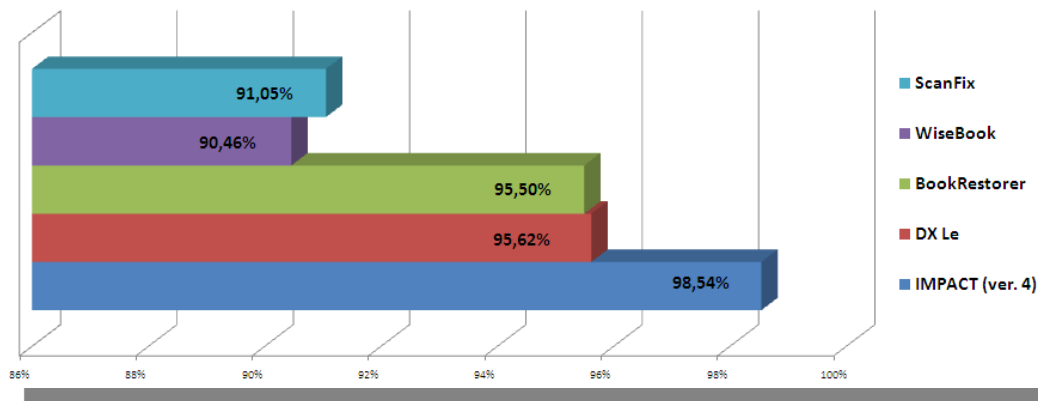
NCSR Border removal

NCSR Geometric Correction

NCSR Binarisation

Abby 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



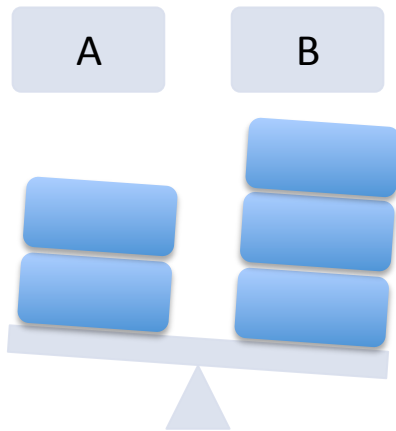
Tools: examples (<http://digitisation.eu/demonstrator-platform>)

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
 - TPDF 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: TPDF 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)



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

The screenshot displays the 'VIRTUALNE LABORATORIUM TRANSKRYPCJI' web portal. It features several key components:

- Cut Signs:** A tool for creating customised recognition profiles, showing a document image with a 'W' sign highlighted and a grid overlay.
- Identify the sign:** A tool for identifying the sign, showing a 'W' sign with a grid overlay and a 'The history of recognized signs' table.
- Main Interface:** A central window showing a document image with a 'W' sign highlighted and a grid overlay. The text below the image reads: "tysiąc barek leżących w wodzie uniemożliwiło żeglugę. Odrzańscy wodniacy tylko dziesiątą część taboru rzeczno-gebieli w posiadanie. Już w sierpniu 1945 roku jednak ruszył odrzańskim szlakiem pierwszy transport węgla. Odrzańscy wodniacy tego kanału aż do kopalń górnośląskich, aby u na wagony kolejowe i z nich na barki. W przewiozła prawie 2 miliony ton towaru, za 5 Odrzaliśmy notatki i nakreśliły na linii Odrzka, co piętnaście lat temu... Porównujmy dalej 598 zakładów przemysłowych, skupiających Osiem wyższych uczelni. Ponad 15 tys. stu Wrocław. Mówi się o nim, że jest stolicą Nadodrza i nie honorowe. Wrocław jest potężnym ośrodkiem mającym niemal udział w kształtowaniu obl i całego kraju. Ludność Wrocławia stanowi 1,4% w produkcji przemysłowej wynosi 2,6%. Zniszc widoczne. Miastu przybywa około 6.000 izb mie".
- Audyt (Audit):** A table for auditing the transcription results. The table has columns for 'Znak', 'Finalny', 'Oryginalny', 'Z kontekstem', 'Nieczytelny', 'Czcionka', 'Poprawny', 'Autor', and 'Decyzja'. The rows show the characters 'E', 'o', 'e', 'd', 'a', and 'u' with their corresponding original and contextual forms, and the decision made by the user.

The URL <http://wlt.synat.pcss.pl/> is displayed at the bottom left.

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



Thank you!

Tomasz Parkoła
tparkola@man.poznan.pl

CERL annual seminar, 28.10.2014, Oslo, Norway