



61-139 Poznan
ul. Jana Pawła II 10
phone: (+48 61) 858-20-01
fax: (+48 61) 852-59-54

kpcontact@man.poznan.pl
<https://ehum.psnc.pl>

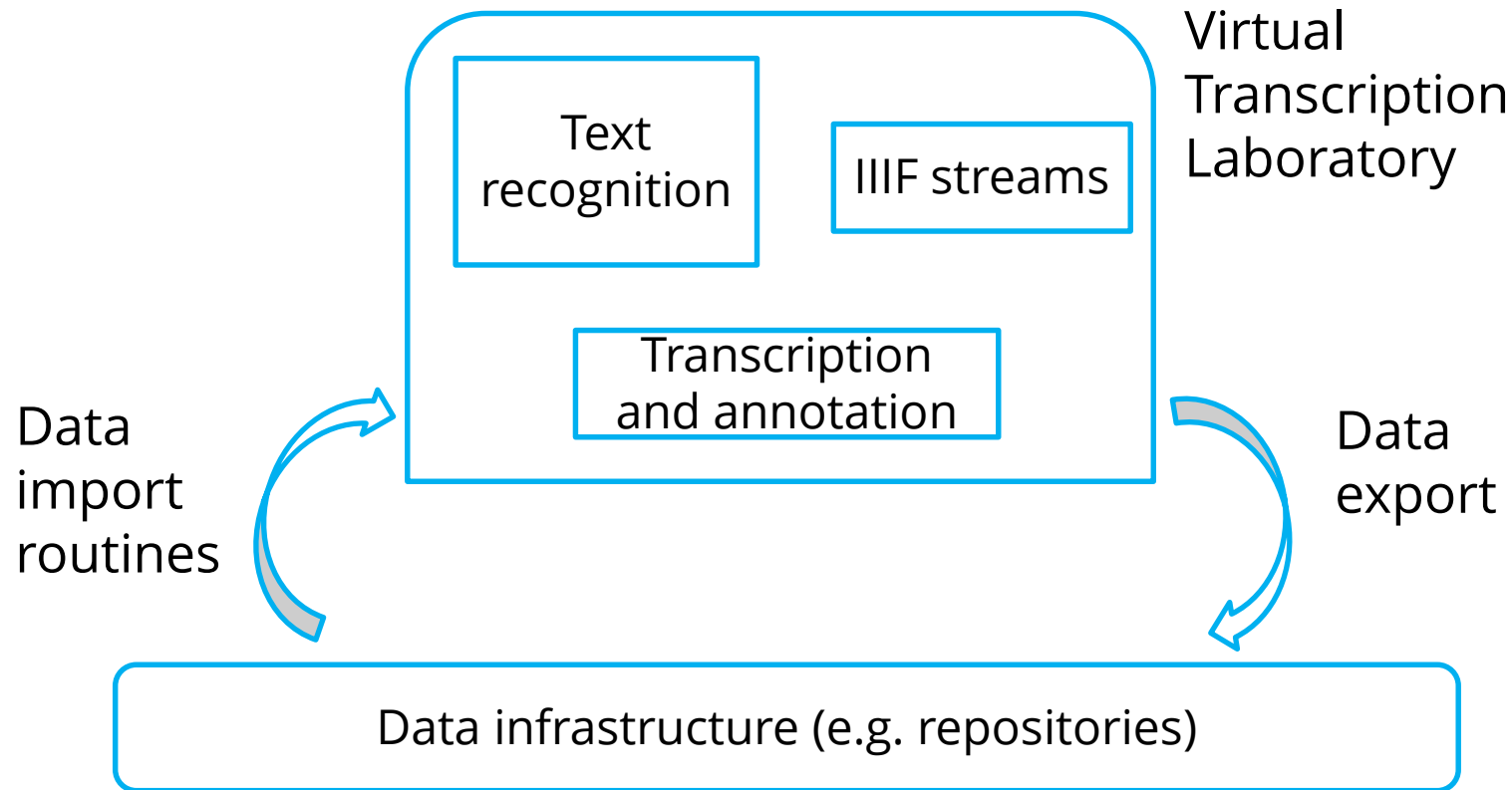
Virtual Transcription Laboratory: IIIF and OCR enabled transcription platform

Tomasz Parkoła, Poznan Supercomputing and
Networking Center

The challenge

- Thousands of digital assets are available online in various data repositories
- We need to address researcher's needs and provide good quality data
- Extend data infrastructure with semi-automated transcription system
 - The main challenge – increase availability of full-text historical documents
 - There is no solution that integrates transcription tool with character recognition engine, IIIF-based streams and automated training routines

Our approach



Virtual Transcription Laboratory

- Import module
 - TEI, IIIF manifest, OAI-PMH (METS profile),
 - Supported formats: PDF, images
- Export module
 - eBooks (ePub, MOBI, PDF)
 - plain text, hOCR, TEI
- Recognition module
 - Based on Tesseract 4.0 (LSTM)
 - Trained with validated input from users
- Transcription and annotation
 - Line level recognition by default, but new regions can be also added
 - Working groups supported with review/feedback proces
- Existing deployments: <https://translab.uni.wroc.pl/> (institutional) and <https://wlt.pcss.pl/> (general purpose)



Poznan Supercomputing and Networking Center

61-139 Poznan

ul. Jana Pawła II 10

phone: (+48 61) 858-20-01

fax: (+48 61) 852-59-54

kpcontact@man.poznan.pl

<https://ehum.psnc.pl>

