POZNAŃ SUPERCOMPUTING AND NETWORKING CENTER

POZNAŃ SUPERCOMPUTING AND NETWORKING CENTER

Country-scale infrastructure for creation of full text versions of historical documents from Polish Digital Libraries

Adam Dudczak, Miłosz Kmieciak, Marcin Werla
{name.surname}@man.poznan.pl

Interedition Symposium
Scholarly Digital Editions, Tools and Infrastructure
19-20 March 2012

Polish Digital Libraries Federation

- The Polish Digital Libraries Federation aggregates information from digital libraries distributed across entire Poland
- Operates since 2007
- Over 70 digital libraries
- Almost 920 000 metadata records

DIGITAL LIBRARIES FEDERATION

PIONIER POLISH OPTICAL INTERNET

Role of Polish DLF

- Common gateway to resources from Polish digital libraries
- Works as an intermediary between Polish digital libraries and Europeana
- Offers services for both regular users and digital librarians
  - Basic statistics
  - Map of Polish Digital Libraries
  - Duplicates detection
  - Digitisation plans
  - E-learning courses

Resources in Polish DLF

- Almost 920 000 of digital objects, most of them are described as containing text
- Mostly newspapers, journals and books (72%)
- Oldest objects from 11th century,
- Most objects from 19th and beginning of 20th century
- DjVu is the most widely used format for representation of the text (79%)

Description of the problem

- Having one common access point with metadata search is great but full text is better then most detailed metadata!
- Libraries contains a lot of materials which might be useful for researchers
- Lack of full text search results in poor visibility of those resources
- This issue is address by PSNC's work in the frame of SYNAT (http://www.synat.pl) project.

How OCR is used?

- Survey was held in September-October of 2010
- We received responses from 26 major institutions
- Survey covered 70% of resources gathered by Polish DLF
- We were asking about all sort of things related to creation of full text search
  - Type and number of documents
  - Digitisation practices
  - Usage of OCR software

Results of the survey

- Scanning resolution between 300-600 PPI, colour depth depends on type of resource
- FineReader and Document Express are most widely used OCR software packages
- Only 3 institutions use training capabilities of OCR software

- 40% of objects were a subject of OCR
- No one does the correction of OCR results

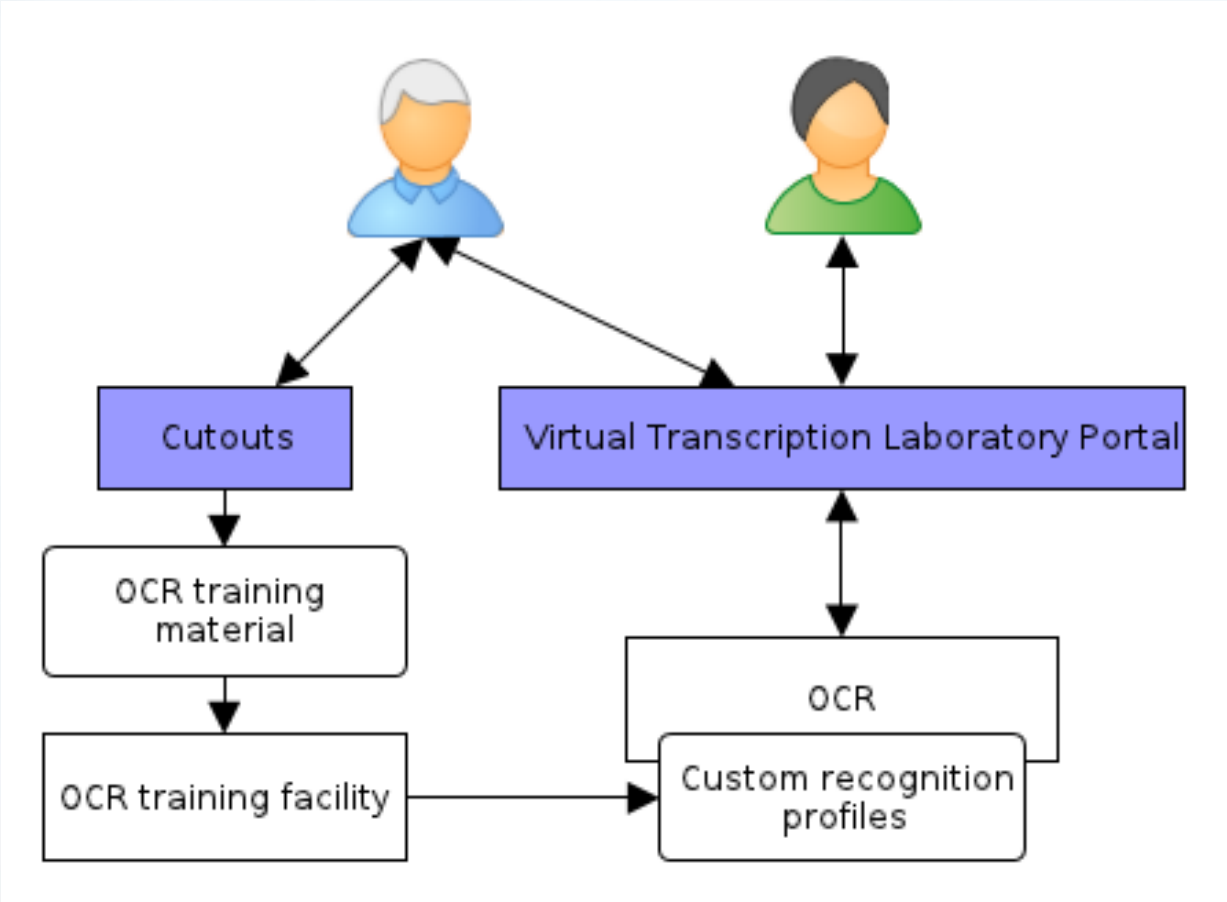
Conclusions

- No correction of OCR results
  - Librarians are not interested in 100% correct text. OCR results are used as a search aid.
  - Lack of human resources to perform correction
- Limited usage of training capabilities
  - Training can improve OCR quality for historical documents
- Lack of tools which would integrate training and correction into digitisation workflow

General assumption

- Integration with national aggregation infrastructure (Polish DLF)
- Well suited for digitisation workflow
  - Creation of custom recognition profiles
  - Crowdsourced correction of both new and existing resources
- Useful for researchers willing to work in distributed environment on historical projects

Most important components

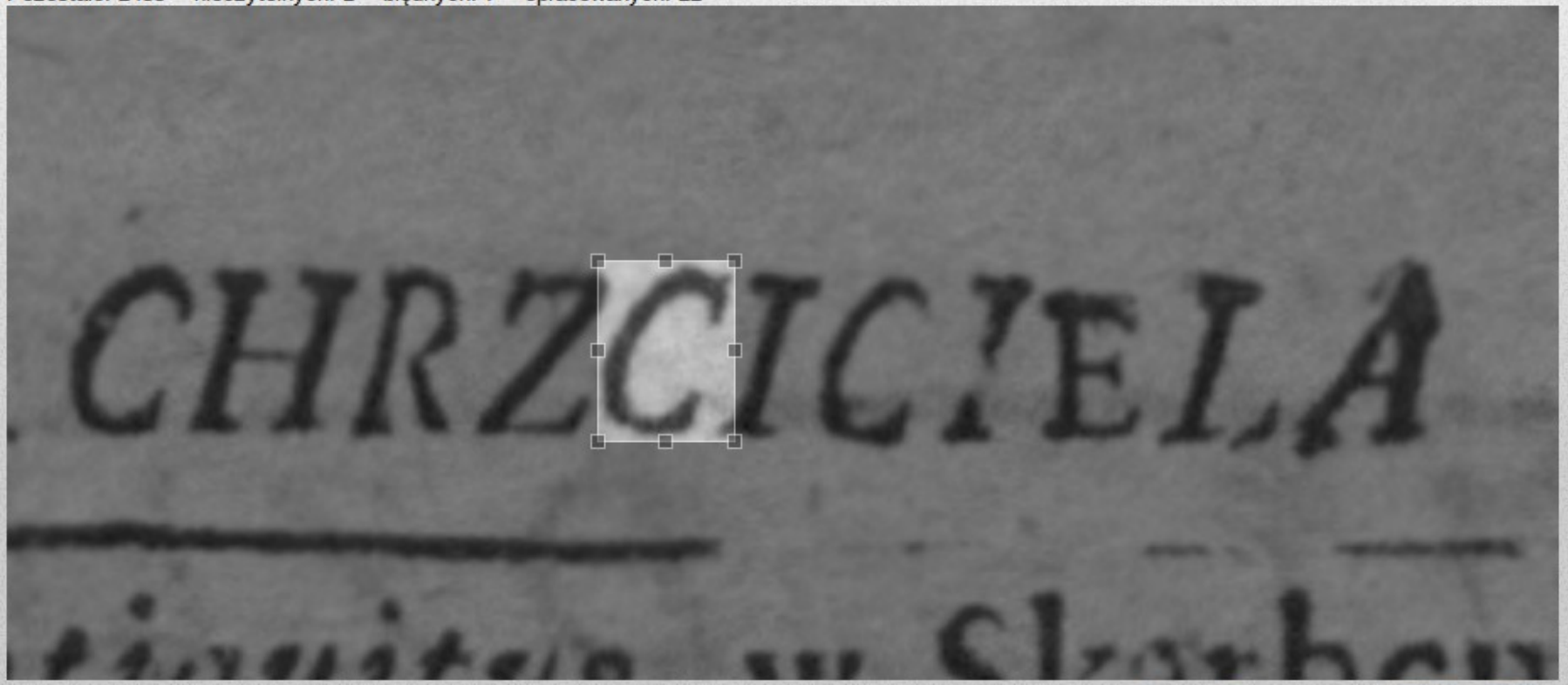


OCR and supporting services

- OCR Service implemented on top of Tesseract 3.x
  - well-known, free, open source solution
- Support for recognition of multiple modern languages (including Polish)
- OCR supporting services
  - Preparation of training data (Cutouts, ...)
  - Training itself (OCR training facilities)

Wycinanie znaków

Pozostało: 2488 nieczytelnych: 1 błędnych: 7 opracowanych: 22

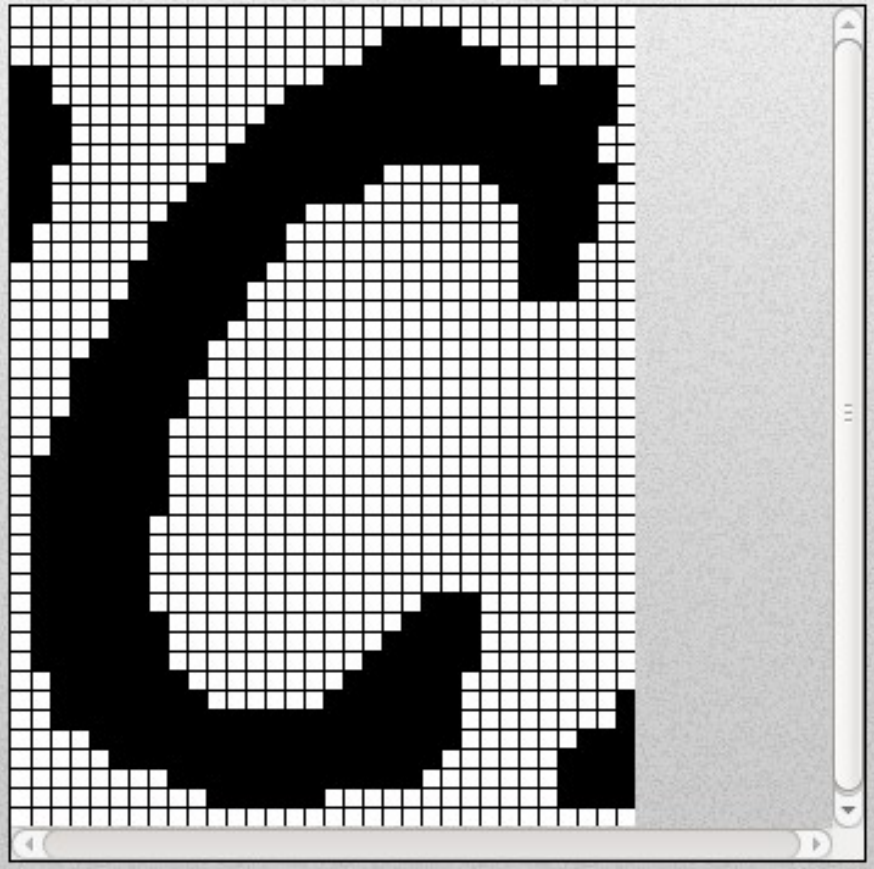


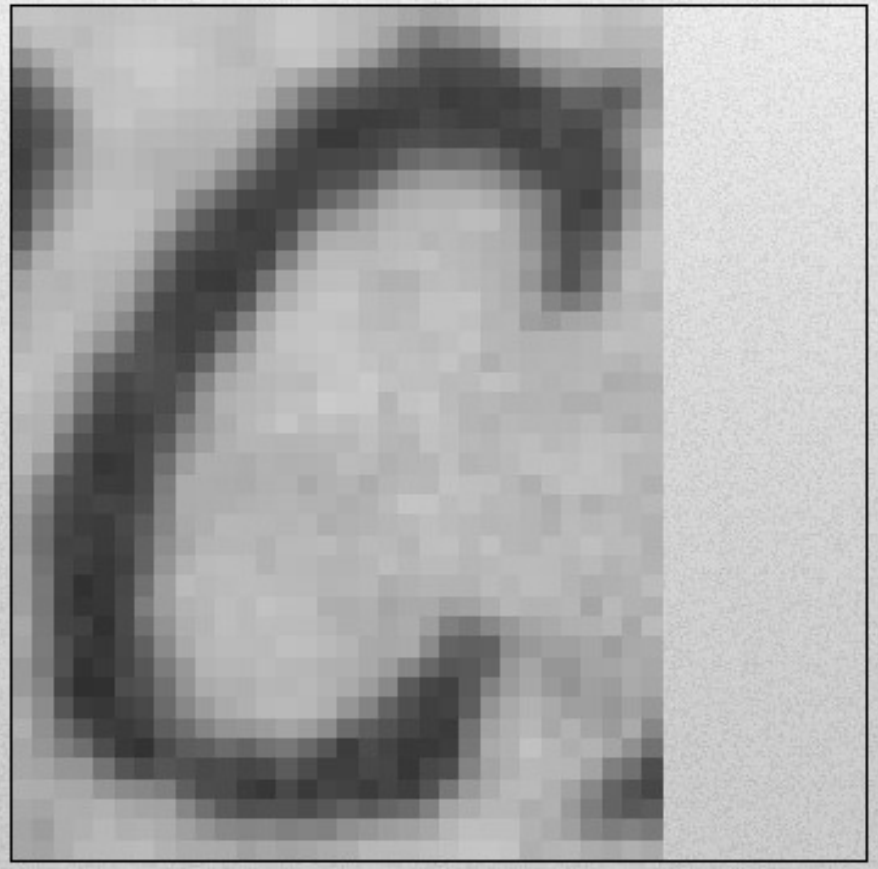
Historia opracowanych znaków

znak	typ czcionki	opis	
z			wycofaj
w			wycofaj
a			wycofaj
R			wycofaj
R			wycofaj
H	kursywa		wycofaj

Screenshot from “Cutouts”

Identyfikacja znaku





Zmień stopień binaryzacji - +

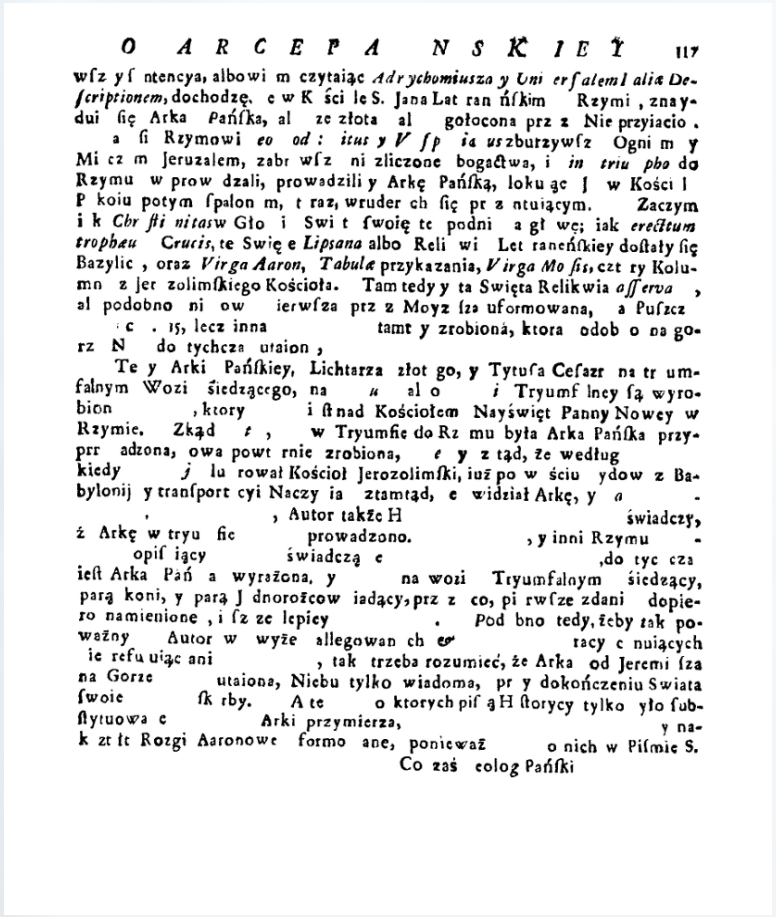
Korekty wykonane pędzlem cofnij zmiany

Formatowanie znaku:

Wpisz znak: c

Screenshot from “Cutouts”

Tesseract training



O ARCEPANSKIEY 117

wſzyſtentencya, albowiem czytaiąc *Adrychomiusza y Universalem aliæ Deſcriptionem*, dochodzę, że w Kościele S. Jana Lateraneńſkim Rzymi, znayduie ſię Arka Pańſka, ale ze złota ale gołocona przez Nie przyiacioł. a ſi Rzymowi *eo od: itus y Veſpaſianus* zburzywſzy Ogniem y Mieczem Jeruzalem, zabrawſzy niezliczone bogactwa, ie *in triumpho* do Rzymu w prow dzali, prowadzili y Arkę Pańſką, loku ąc J w Kościele Pokoiu potym ſpalonym, teraz, wruderach ſię prezentuiącym. Zaczym iak *Chriſtianitas* w Głowie Swiata ſwoię też podniefła głowę; iak *ereƈtum trophæum Crucis*, te Swięte *Lipſana* albo Relikwie Leteraneńſkiey doſtały ſię Bazylice, oraz *Virga Aaron*, *Tabulæ* przykazania, *Virga Moſis*, cztery Kolumny z Jerozolimſkiego Kościoła. Tam tedy y ta Swięta Relikwia *aſſervatur*, ale podobno nie owa pierwſza przez Moyzeſza uformowana, na Puſzczy *Exod: cap.* 15, lecz inna *ad formam* tamtey zrobioną, ktora podobno na gorze *Nebo* do tychczas utaiona,

Tedy Arki Pańſkiey, Lichtarza złotego, y Tytuſa Ceſazra na tryumfalnym Wozie ſiedzącego, na *Arkuſie* albo *Bramie* Tryumfalney ſą wyrobione *veſtigia*, ktory *Arkus* ieſt nad Kościołem Nayświęt: Panny Nowey w Rzymie. Zkąd *patet*, że w Tryumfie do Rzymu była Arka Pańſka przyprowadzona, owa powtornie zrobiona, *Patet* y z tąd, że według *Egeſippa* kiedy *Pompejus* luſtrował Kościoł Jerozolimſki, iuż po wyściu Żydow z Babylonij y tranſportacyi Naczynia ztamtąd, że widział Arkę, y *Tabulas Teſtimonij*, *item Cherubinos*, Autor także *Hiſtoriæ Scholaſticæ Judith* 3. świadczy, że Arkę w tryumfie *Tytuſa* prowadzono. *Adrichomius*, y inni Rzymu *ſplendorem* opiſuiący *Authores* świadczą że *in Arcu triumphali Tytuſa*, do tychczas ieſt Arka Pańſka wyrażona, y *Tytus* na wozie Tryumfalnym ſiedzący, parą koni, y parą Jednorożcow iadący, przez co, pierwſze zdanie dopiero namienione, ieſzcze lepiey *confirmatur*. Podobno tedy, żeby tak poważnych Autorow wyżey allegowanych *& probabiliter* racyocynuiących nie refutuiąc ani *infringendo*, tak trzeba rozumieć, że Arka od Jeremiaſza na Gorze *Nebo* utaiona, Niebu tylko wiadoma, przy dokończeniu Swiata ſwoie *effunder* ſkarby. A te *Vaſa* o ktorych piſzą Hiſtorycy tylko było ſubſtytuowane *adinſtar* Arki przymierza, *adinſtar Tabularum Teſtimonij*, y na kształt Rozgi Aaronowey formowane, ponieważ *nulla* o nich w Piſmie S. *poſt Reſtaurationem Templi mentio* Co zaś Teolog Pańſki *Apoc: cap:* 11. o Arce temi wſpomina ſłowy: *Et apertum eſt Templum DEI in Cælo & viſa eſt Arca*; nie ma ſię rozumieć, żeby Arka przymierza miała bydź do Nieba wzięta y przeniesiona, lecz przez nią rozumieią *Humanam Naturam Chriſti*. *Richardus*, y *Beda*, rozumieią drudzy Matkę Nayświętſzą, iako to *Bernardus* y *Suarez*.

O Loto-

Virtual Transcription Laboratory

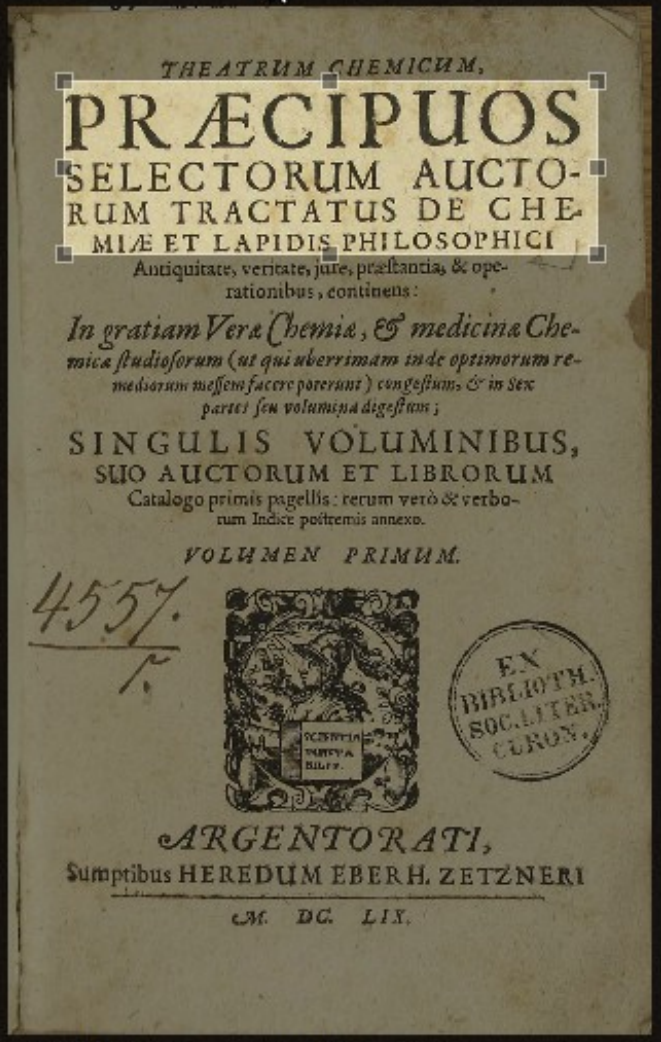
- Virtual Transcription Laboratory will allow to integrate text recognition and correction into digitisation workflows
- Users of VTL can upload scanned images and create textual version on this basis
- VTL gives access to OCR service
- It allows to correct existing text
- At any time user can export results of work in hOCR format

WIRTUALNE LABORATORIUM TRANSKRYPCJI

POWRÓT DO STRONY PROJEKTU

Zaznacz obszar, który chcesz poddać przetwarzaniu OCR

-- default -- Start OCR


THEATRUM CHEMICUM,

PRÆCIPUOS

SELECTORUM AUCTORUM TRACTATUS DE CHEMIÆ ET LAPIDIS PHILOSOPHICI

Antiquitate, veritate, jure, præstantia, & operationibus, continens:

In gratiam Veræ Chemiæ, & medicinæ Chemicæ studiosorum (ut qui uberrimam inde optimorum remediorum messem facere poterunt) congestum, & in sex partes seu volumina digestus;

SINGULIS VOLUMINIBUS, SUO AUCTORUM ET LIBRORUM Catalogo primis pagellis: rerum verò & verborum Indice postremis annexo.

VOLUMEN PRIMUM.

4557. f.

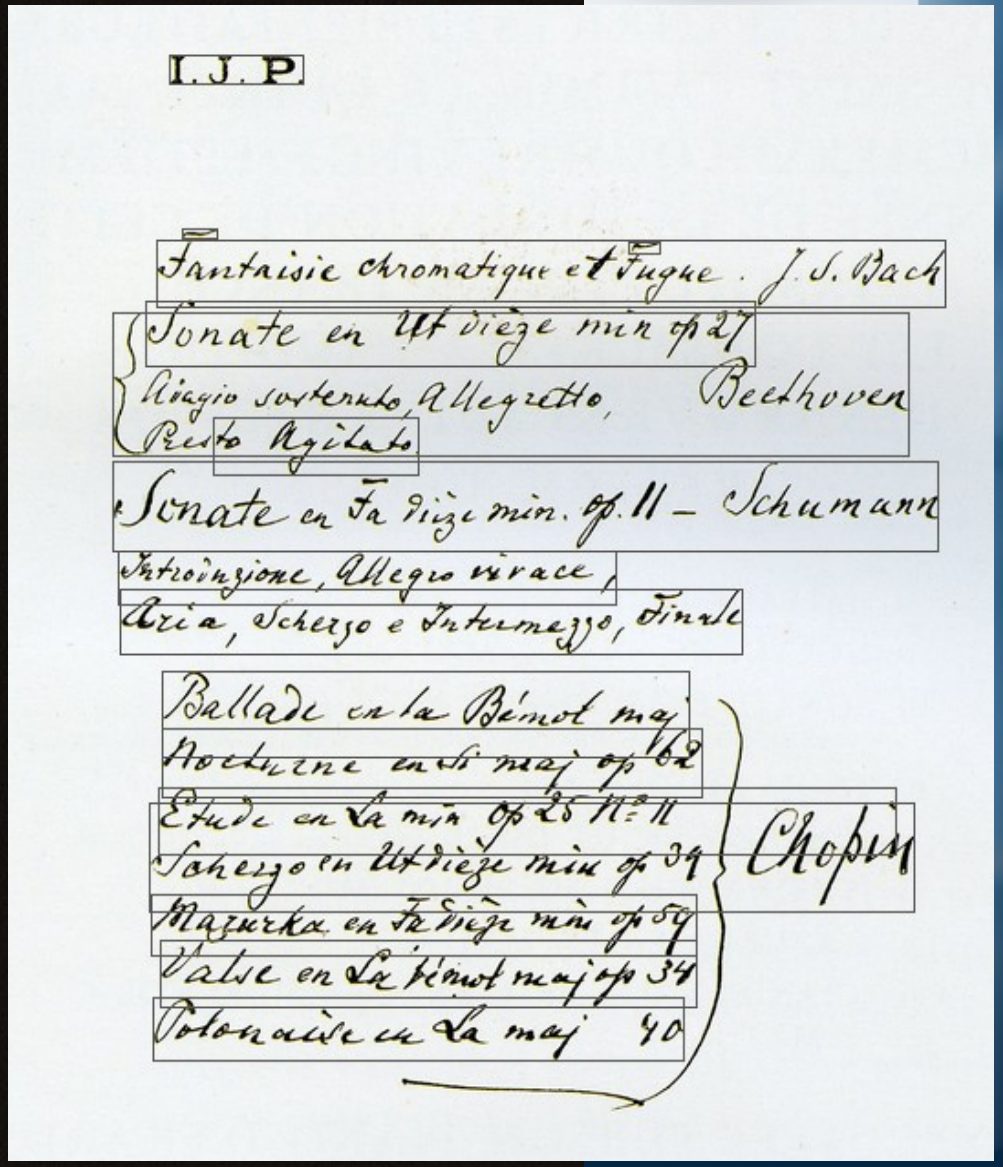
EX BIBLIOTH. SOC. LITER. CURON.

ARGENTORATI,

Sumptibus HEREDUM EBERH. ZETZNERI

M. DC. LIX.

Create Transcript | Create Boundaries


I. J. P.

Fantaisie chromatique et Fugue . J. S. Bach

Sonate en Ut dièze min op 27
Adagio sostenuto, Allegretto, Presto Agitato. Beethoven

Sonate en Fa dièze min. op. 11 – Schumann
Introduzione, Allegro vivace,
Aria, Scherzo e Intermezzo, Finale

Ballade en la Bémol maj
Nocturne en si maj op 62
Etude en La min Op 25 N° 11
Scherzo en Ut dièze min op 39 } Chopin
Mazurka en Fa dièze min op 59
Valse en La Bémol maj op 34
Polonaise en La maj 40

Screenshots from webinterface of OCR service

tysiąc barek leżących w wodzie uniemożliwiało żeglugę. Odrzańscy wodniacy tylko dziesiątą część taboru rzecznego objęli w posiadanie. Już w sierpniu 1945 roku jednak ruszył odrzańskim szlakiem pierwszy transport węgla.

Edytor

I Komentarz Line: OCR Wyczyść Usuń

tysiąc barek leżących w wodzie uniemożliwiało żeglugę. Odrzanscy wodniacy

opień

ponad

vodną

eniem

tego kanału aż do kopalń górnośląskich, aby uniknąć przeładowywania węgla na wagony kolejowe i z nich na barki. W roku 1960 „Żegluga na Odrze" przewiozła prawie 2 miliony ton towaru, za 5 lat przewiezie dwa razy tyle.

Odłóżmy notatki i nakreślmy na linii Odry te poprawki. Już nie ta rzeka, co piętnaście lat temu... Porównujmy dalej.

*

598 zakładów przemysłowych, skupiających ponad 63 tys. pracowników. Osiem wyższych uczelni. Ponad 15 tys. studentów. Dwa wydawnictwa... Wrocław.

Mówi się o nim, że jest stolicą Nadodrza i nie jest to tylko jakieś określenie honorowe. Wrocław jest potężnym ośrodkiem przemysłowym i kulturalnym, mającym niemały udział w kształtowaniu oblicza nie tylko Nadodrza, ale i całego kraju. Ludność Wrocławia stanowi 1,4% ludności Polski. Jego udział

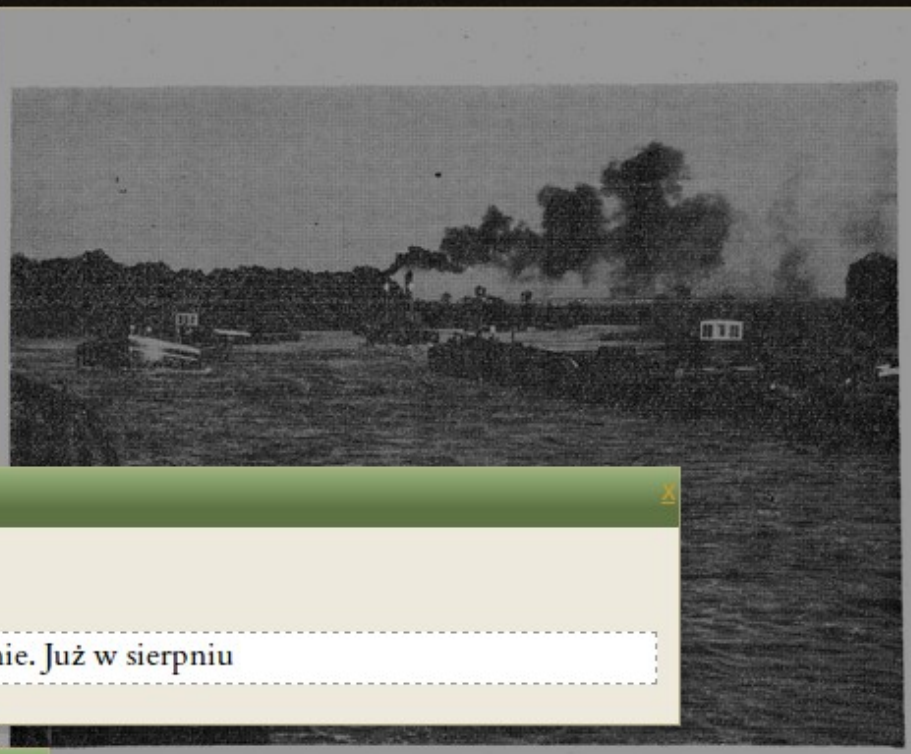
nosi 2,6%. Zniszczenia wojenne są coraz mniej

widoczne. Miasto przybywa około 6.000 izb mieszkalnych rocznie. Rosną nowe

Screenshots from transcription editor

WIRTUALNE LABORATORIUM TRANSKRYPCJI

tysiąc barek leżących w wodzie uniemożliwiało żeglugę. Odrzanscy wodniacy
tylko dziesiątą cześć taboru rzecznego objęli w posiadanie. Już w sierpniu
1945 roku jednak ruszył odrzańskim szlakiem pierwszy transport węgla.
Oczyszczono baseny portowe, odbudowano nabrzeża, zbudowano nowy stopień
wodny, buduje się nowe zbiorniki retencyjne.



Edytor

I Komentarz Line: OCR Wyczyść Usuń

tylko dziesiątą cześć taboHru rzecznego objęli w posiadanie. Już w sierpniu

Klawiatura

Poprzednie Nastepne

ċ	Č	č	Ď	ď	Đ	đ	Ē	ē	Ĕ
ĕ	Ė	ė	Ę	ę	Ě	ě	Ĝ	ĝ	Ğ
ğ	Ġ	ġ	Ģ	ģ	Ĥ	ĥ	Ħ	ħ	Ĩ
ĩ	Ī	ī	Ĭ	ĭ	Į	į	İ	ı	IJ
ij	Ĵ	ĵ	Ķ	ķ	ĸ	Ĺ	Í	Ļ	ļ
Ľ	ľ	Ŀ	ŀ	Ł	ł	Ń	ń	Ņ	ņ

tysiąc barek leżących w wodzie uniemożliwiało żeglugę. Odrzańscy wodniacy tylko dziesiątą część taboru rzecznego objęli w posiadanie. Już w sierpniu 1945 roku jednak ruszył odrzańskim szlakiem pierwszy transport węgla. Oczyszczono baseny portowe, odbudowano nabrzeża, zbudowano nowy stopień wodny, buduje się nowe zbiorniki retencyjne.

Przedsiębiorstwo „Żegluga na Odrze" otrzymało kredyty na budowę ponad 200 barek motorowych. Kanał Gliwicki łączy się obecnie magistralą wodną z Zakładami Azotowymi w Kędzierzynie. Trwają studia nad przedłużeniem tego kanału aż do kopalń górnośląskich, aby uniknąć przeładowywania węgla na wagony kolejowe i z nich na barki. W roku 1960 „Żegluga na Odrze" przewiozła prawie 2 miliony ton towaru, za 5 lat przewiezie dwa razy tyle.

Odłóżmy notatki i nakreślmy na linii Odry te poprawki. Już nie ta rzeka, co piętnaście lat temu... Porównujmy dalej.

•

598 zakładów przemysłowych, skupiających ponad 63 tys. pracowników. Osiem wyższych uczelni. Ponad 15 tys. studentów. Dwa wydawnictwa... Wrocław.

Mówi się o nim, że jest stolicą Nadodrza i nie jest to tylko jakieś określenie honorowe. W
mającym n
i całego kra
w produkcji przemysłowej wynosi 2,6%. Zniszczenia wojenne są coraz mniej widoczne. Miastu przybywa około 6.000 izb mieszkalnych rocznie. Rosną nowe

Screenshots from transcription editor

Ongoing work

- Experimenting with Tesseract training on top of data released in IMPACT project
  - http://dl.psnc.pl/activities/projekty/impact/results/
- Release of custom recognition profile dedicated to Polish documents written in Gothic script
- Public release of Virtual Transcription Laboratory

Future plans

- Direct import of content from digital library
- Inclusion of various crowdsourcing techniques e.g. games

POZNAŃ SUPERCOMPUTING AND NETWORKING CENTER

QUESTIONS?

Poznań Supercomputing and Networking Center
affiliated to the Institute of Bioorganic Chemistry of the Polish Academy of Sciences,
ul. Noskowskiego 12/14, 61-704 Poznań, POLAND,
Office: phone center: (+48 61) 858-20-00,
fax: (+48 61) 852-59-54,
e-mail: office@man.poznan.pl, http://www.man.poznan.pl