Wyniki prac PCSS w etapach A9, A10, A11, A12 i A25 projektu SYNAT

Cezary Mazurek, Tomasz Parkoła, Juliusz Pukacki, Maciej Stroiński, Marcin Werla, Jan Węglarz
Poznańskie Centrum Superkomputerowo-Sieciowe

Założenia zadania badawczego PCSS w projekcie SYNAT

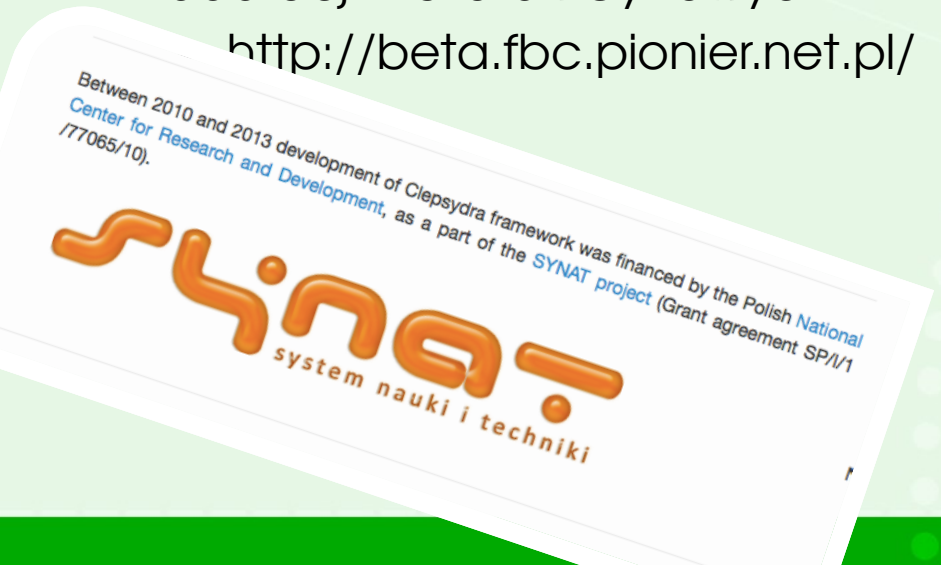
- Etapy A9 i A10 – Nowa architektura agregacji i wzbogacania danych
  - Rozproszone heterogeniczne źródła – w szczególności biblioteki, archiwa i muzea cyfrowe
  - Technologia oparta na Federacji Bibliotek Cyfrowych
  - Cel: Nowej generacji usługi agregacji danych

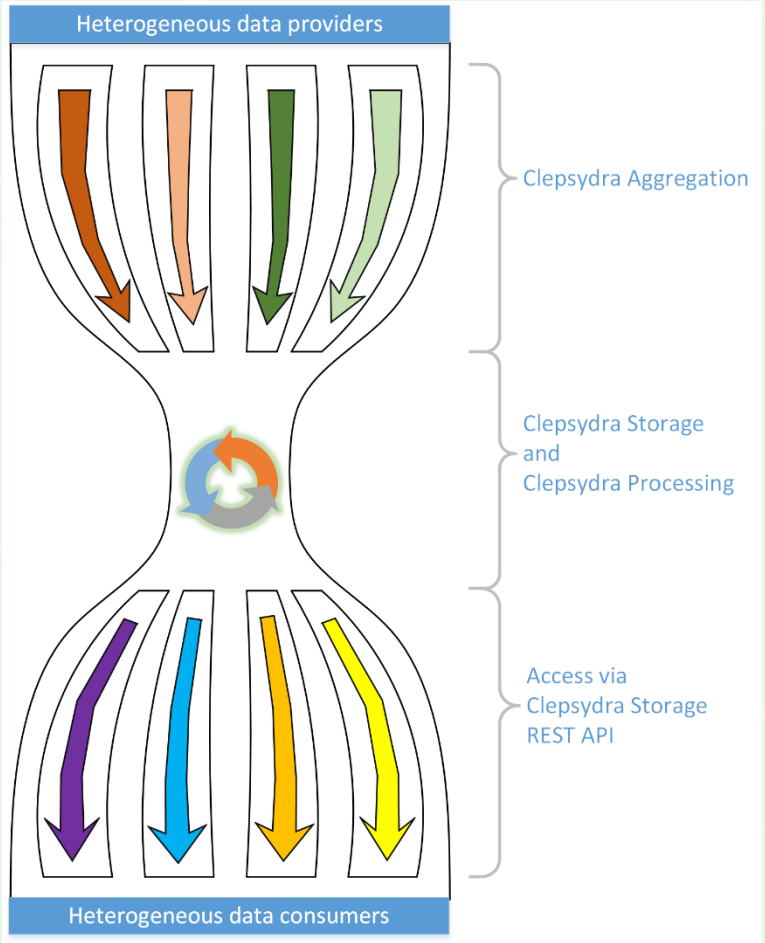
- Etap A11 – opracowanie systemu Wielofunkcyjnego Repozytorium Danych Źródłowych
  - Zunifikowany dostęp do wszelkich repozytoriów i usług magazynowania danych
  - Technologia oparta na oprogramowaniu dLibra
  - Cel: Usługi e-infrastruktury do bezpiecznego deponowania danych źródłowych

- Etap A12 – opracowanie Wirtualnego Laboratorium Transkrypcji
  - Przetwarzanie danych źródłowych na potrzeby badań cyfrowej humanistyki
  - Technologia opracowana z humanistami korzystającymi ze zbiorów polskich bibliotek cyfrowych
  - Cel: Rozwój usług cyfrowej humanistyki

Etap A9

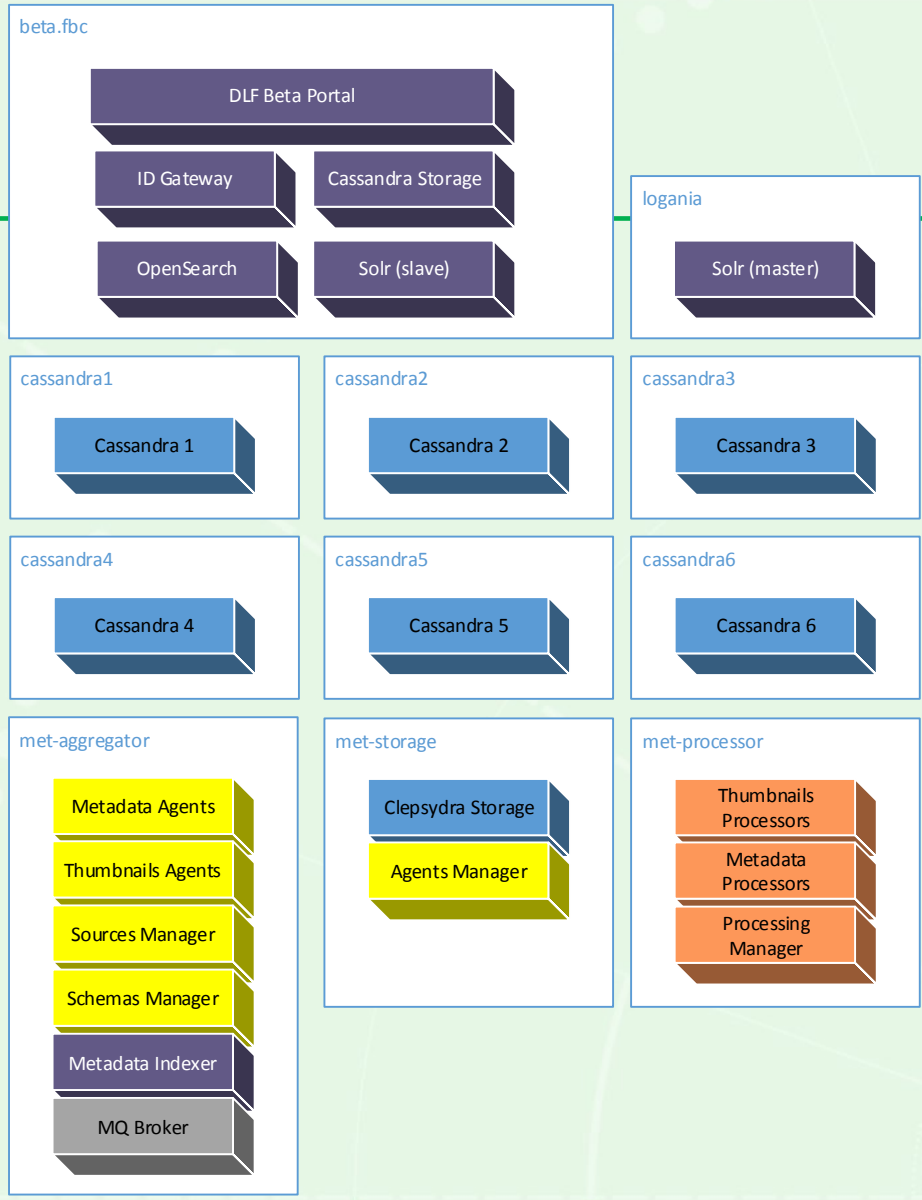
- Główny wynik:
  - Prototyp systemu agregacji i wzbogacania danych Clepsydra
  - http://fbc.pionier.net.pl/pro/clepsydra/
- Pilotażowe wdrożenie na potrzeby Federacji Bibliotek Cyfrowych
  - http://beta.fbc.pionier.net.pl/





Etap A9

- Pilotażowe wdrożenie (połowa lipca 2013)
  - 15.6M rekordów metadanych
  - 2.1M miniatur
  - Dane rozłożone na 6 węzłów bazy danych Cassandra:
    - Node1: 123.05 GB
    - Node2: 204.15 GB
    - Node3: 152.34 GB
    - Node4: 220.55 GB
    - Node5: 191.81 GB
    - Node6: 148.4 GB



Źródła danych

Źródła danych
Typy źródeł danych

#	Identyfikator	URL	Typ źródła	Opis		
550	http://dlibra.biblioteka.tarnow.pl/dlibra/oai-pmh-repository.xml	http://dlibra.biblioteka.tarnow.pl/dlibra/oai-pmh-	pmh	Zaimportowane	Edytuj	Usuń
600	http://dlibra.wbpg.org.pl/dlibra/oai-pmh-repository.xml			ane	Edytuj	Usuń
650	http://mbc.malopolska.pl/dlibra/oai-pmh-repository.xm			ane	Edytuj	Usuń
700	http://czytelnia.cnbop.pl/oai			ane	Edytuj	Usuń
750	http://pbc.gda.pl/dlibra/oai-pmh-repository.xml			ane	Edytuj	Usuń
800	http://jbc.jelenia-gora.pl/dlibra/oai-pmh-repository.xml			ane	Edytuj	Usuń
850	http://bbc.mbp.org.pl/oai-pmh-repository.xml			ane	Edytuj	Usuń
900	http://biblioteka.wejherowo.pl/dlibra/dlibra/oai-pmh-repository.xml			ane	Edytuj	Usuń
1000	http://bc.dominikanie.pl/dlibra/oai-pmh-repository.xml			ane	Edytuj	Usuń
1050	http://bc.pollub.pl/dlibra/oai-pmh-repository.xml			ane	Edytuj	Usuń
1100	http://kpbc.umk.pl/dlibra/oai-pmh-repository.xml			ane	Edytuj	Usuń
1150	http://bcul.lib.uni.lodz.pl/dlibra/oai-pmh-repository.xml			ane	Edytuj	Usuń
1200	http://dlibra.bg.uwm.edu.pl/dlibra/oai-pmh-repository.x			ane	Edytuj	Usuń
1250	http://rcin.org.pl/dlibra/oai-pmh-repository.xml			ane	Edytuj	Usuń
1350	http://jbc.bj.uj.edu.pl/dlibra/dlibra/oai-pmh-repository.x			ane	Edytuj	Usuń
1400	http://digital.fides.org.pl/dlibra/oai-pmh-repository.xml			ane	Edytuj	Usuń
1450	http://dlibra.itl.waw.pl/dlibra-webapp/dlibra/oai-pmh-repository.xml					
1500	http://delibra.bg.polsl.pl/dlibra/oai-pmh-repository.xml					
1600	http://bc.bdsandomierz.pl/dlibra/oai-pmh-repository.xn					
1750	http://bibliotekaormianska.pl/dlibra/oai-pmh-repository					
1800	http://mbc.cyfrowemazowsze.pl/dlibra/oai-pmh-repository.xml					
1850	http://www.bibliotekacyfrowa.pl/dlibra/oai-pmh-repository.xml					
1900	http://www.dlibra.karta.org.pl/dlibra/oai-pmh-repository.xml					

Edycja źródła danych

ID 550
Identyfikator http://dlibra.biblioteka.tarnow.pl/dlibra/oai-pmh-repository.xml
URL http://dlibra.biblioteka.tarno
Opis Zaimportowane
Typ źródła pmh

Właściwości

Klucz	Wartość	
harvestSetSpecs	true	Usuń
repositoryName	Tarnowska Biblioteka Cyfro	Usuń
repositoryURL	http://dlibra.biblioteka.tarno	Usuń
isPlanned	false	Usuń
defaultLang	pl	Usuń

Dodaj właściwość

Obsługiwane schematy

Schemat	Nazwa schematu w źródle	
mets	mets	Usuń
http://dlibra.biblioteka.tarnow.pl/dlibra	dlibra_avs	Usuń
set		
oai		

Za

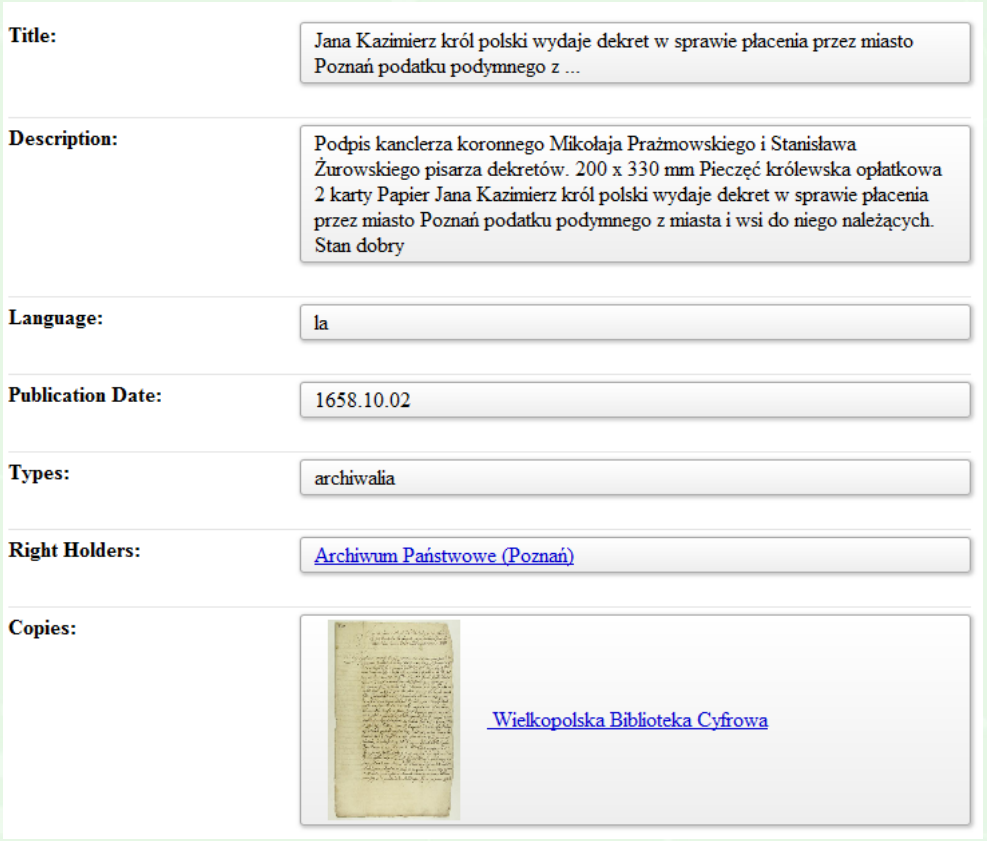
Zarejestrowani agenci

Agenci
Statystyki

#	Obsługiwany typ źródła	URL	Liczba obsługiwanych źródeł	
51	pmh	http://150.254.155.225:8080/oai-pmh-agent/rest/	226	Edytuj
251	th-abckrakow	http://met-aggregator.synat.pcss.pl:8080/th-abckrakow-agent/rest/	1	Edytuj
253	th-dlibra	http://met-aggregator.synat.pcss.pl:8080/th-dlibra-agent/rest/	70	Edytuj
101	nukat	http://150.254.155.225:8080/nukat-agent/rest/	2	Edytuj
201	mona	http://met-aggregator.synat.pcss.pl:8080/mona-agent/rest/	1	Edytuj
254	th-eprints	http://met-aggregator.synat.pcss.pl:8080/th-eprints-agent/rest/	5	Edytuj
301	csv	http://met-aggregator.synat.pcss.pl:8080/csv-agent/rest/	2	Edytuj
252	th-csv	http://met-aggregator.synat.pcss.pl:8080/th-csv-agent/rest/	2	Edytuj

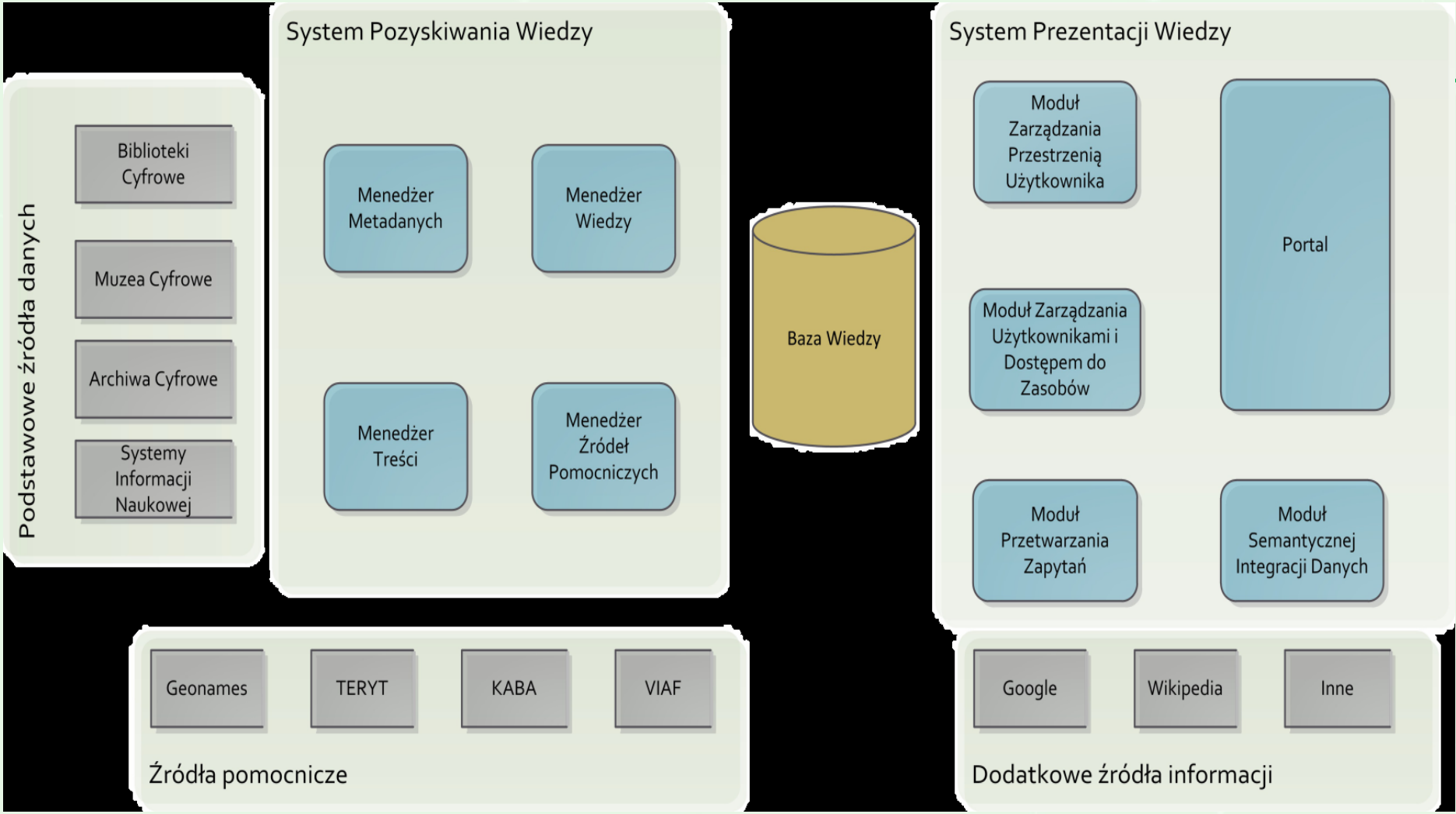
Etap A10

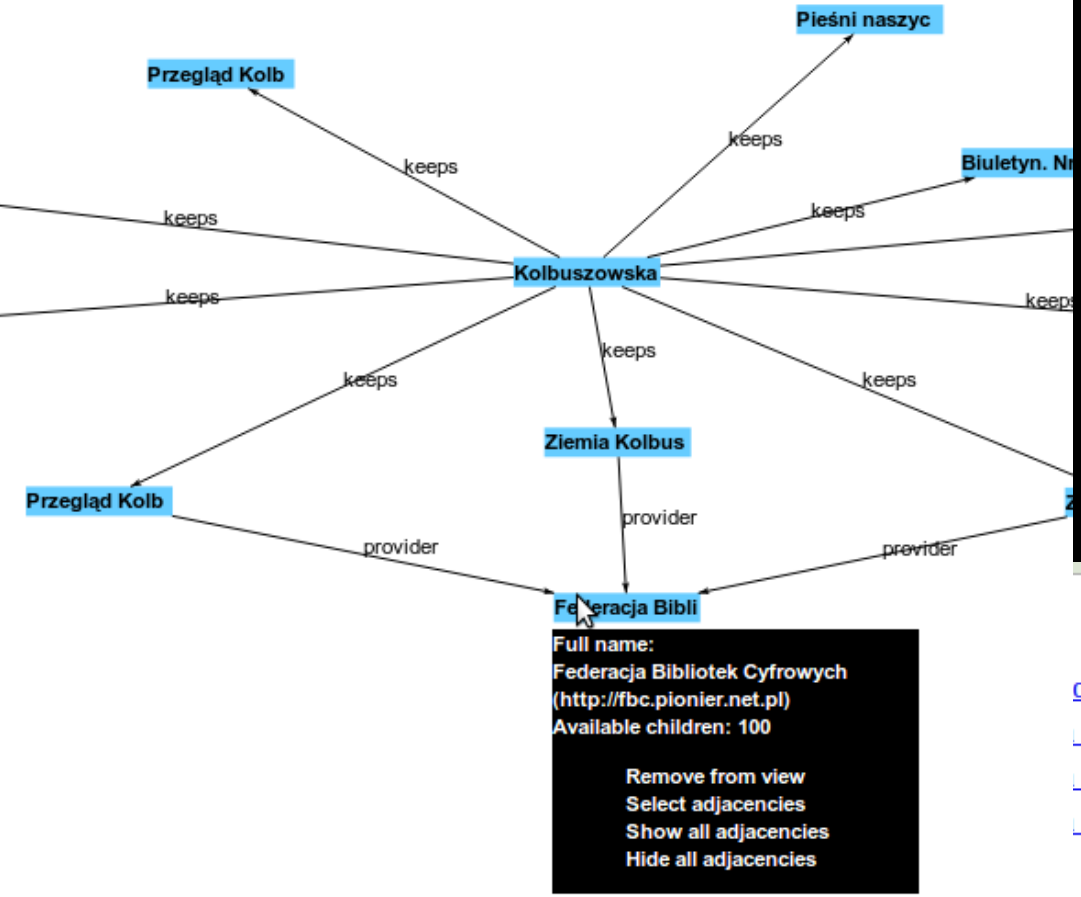
- Główny wynik:
  - Prototyp Zintegrowanego Systemu Wiedzy
    - Zintegrowany dostęp do heterogenicznych źródeł danych
    - Budowa centralnego magazynu danych z wykorzystaniem technologii semantycznych - Bazy Wiedzy (BW)
    - Budowa aplikacji dla użytkownika końcowego – portal z elementami społecznościowymi
- Prototypowa baza wiedzy
  - Rekordy metadanych
    - FBC (PLMET): 876 887
    - NUKAT (MarcXML): 2 191 825
  - Trójki RDF
    - Trójki jawne: 297 145 812
    - Wywnioskowane: 349 757 179
    - Łącznie: 646 902 991



Title:	Jana Kazimierz król polski wydaje dekret w sprawie płacenia przez miasto Poznań podatku podymnego z ...
Description:	Podpis kanclerza koronnego Mikołaja Prażmowskiego i Stanisława Żurowskiego pisarza dekretów. 200 x 330 mm Pieczęć królewska opłatkowa 2 karty Papier Jana Kazimierz król polski wydaje dekret w sprawie płacenia przez miasto Poznań podatku podymnego z miasta i wsi do niego należących. Stan dobry
Language:	la
Publication Date:	1658.10.02
Types:	archiwalia
Right Holders:	Archiwum Państwowe (Poznań)
Copies:	Wielkopolska Biblioteka Cyfrowa

Etap A10


Podstawowe źródła danych
Biblioteki Cyfrowe
Muzea Cyfrowe
Archiwa Cyfrowe
Systemy Informacji Naukowej
System Pozyskiwania Wiedzy
Menedżer Metadanych
Menedżer Wiedzy
Menedżer Treści
Menedżer Źródeł Pomocniczych
Baza Wiedzy
System Prezentacji Wiedzy
Moduł Zarządzania Przestrzenią Użytkownika
Portal
Moduł Zarządzania Użytkownikami i Dostępem do Zasobów
Moduł Przetwarzania Zapytań
Moduł Semantycznej Integracji Danych
Geonames
TERYT
KABA
VIAF
Źródła pomocnicze
Google
Wikipedia
Inne
Dodatkowe źródła informacji

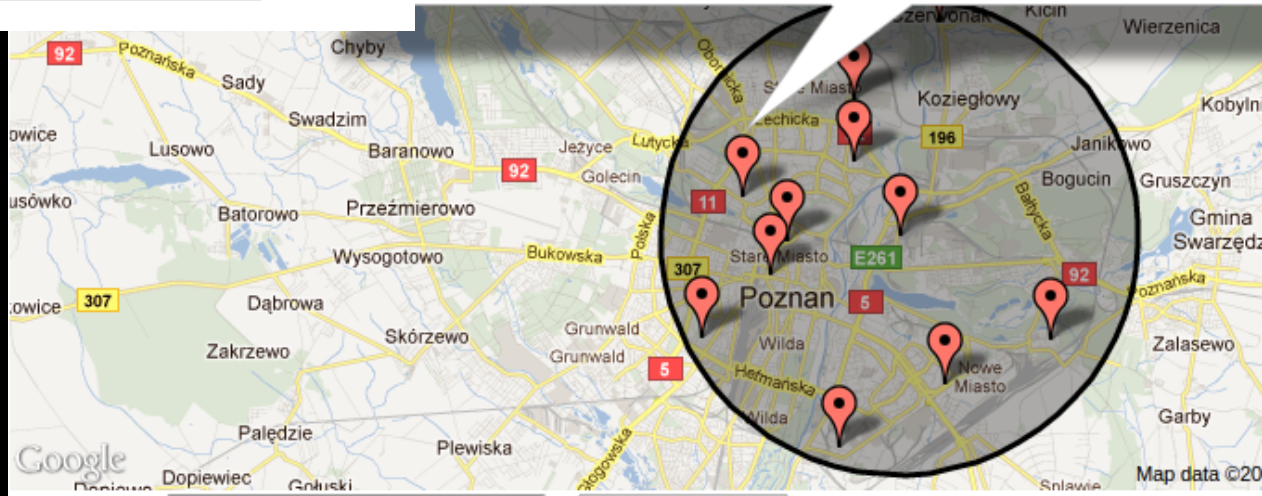


oznański pod względem dziejowym z zastosowaniem do topografii wspó

Poznańska i Nowiny Sportowe 1931.06.09 Nr23

Poznańska i Nowiny Sportowe 1930.03.11 Nr10

Poznańska i Nowiny Sportowe 1931.05.27 Nr21



Poznań

Poznań

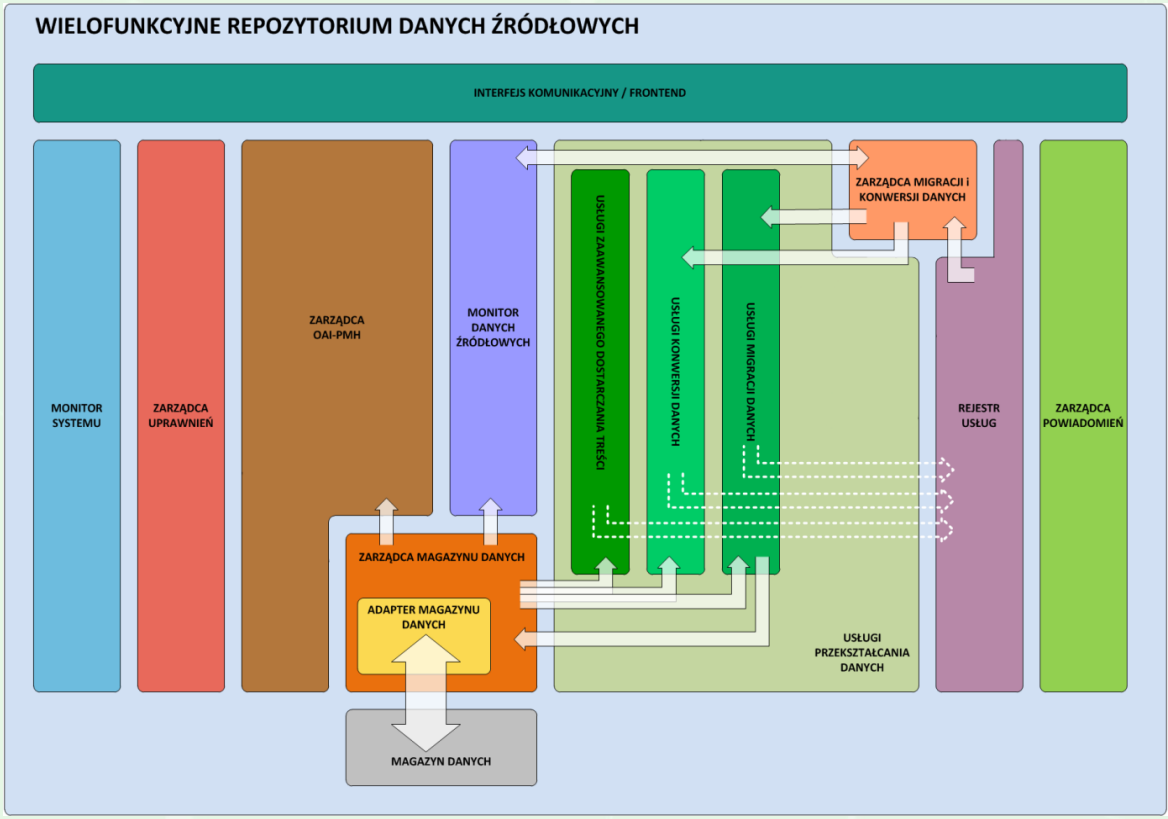
Change Location

Etap A11

- Główny wynik:
  - Prototyp Wielofunkcyjnego Repozytorium Danych Źródłowych
- Współpraca z Repozytorium Cyfrowym Instytutów Naukowych (RCIN)
  - WRDZ (dArceo) systemem długoterminowego przechowywania danych źródłowych
  - Integracja z systemem do zarządzania procesem digitalizacji dLab
- Dalsze wdrożenia dArceo (produktu powstałego na bazie prototypu WRDZ):
  - Książnica Karkonoska
  - Politechnika Śląska
  - Lubelska Biblioteka Wirtualna (prace w toku)

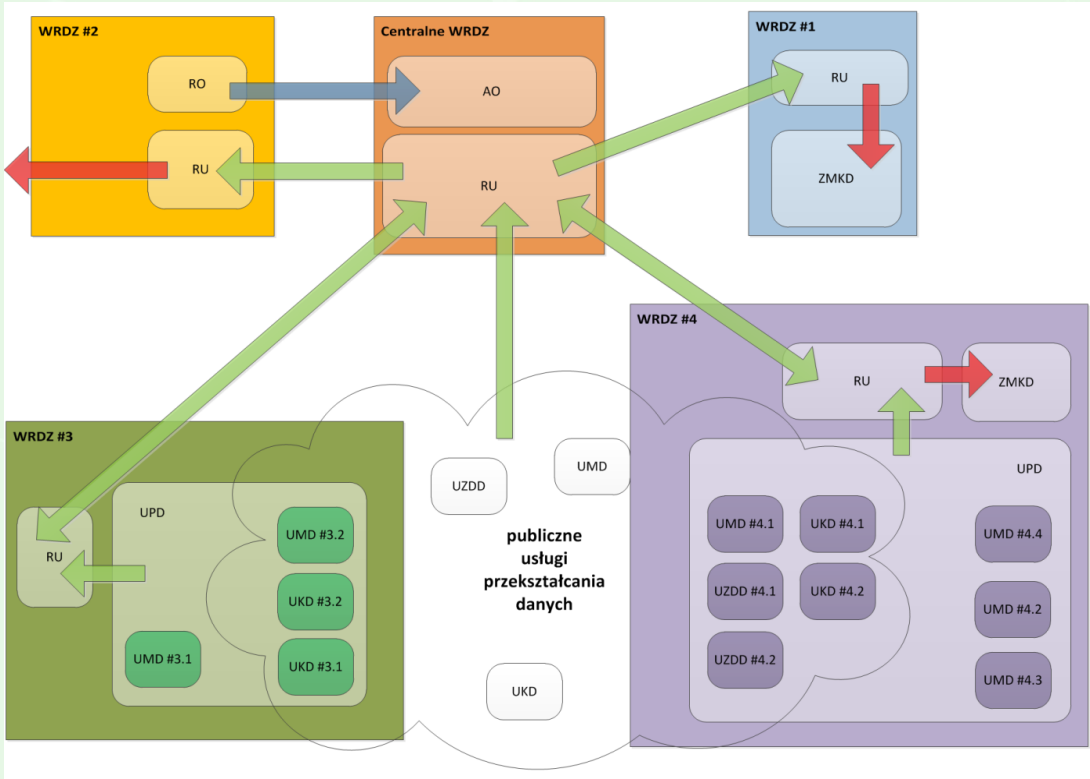
Etap A11

- Podstawowe funkcje WRDZ:
  - Wprowadzanie oraz odczyt danych źródłowych
  - Zarządzanie obiektami w magazynie danych
    - Wersjonowanie
    - Usuwanie
  - Wspierane magazyny danych
    - Serwer SFTP (PLATON U4, Krajowy Magazyn Danych)
    - Przestrzeń dyskowa (dysk sieciowy, macierz, itp.)



Etap A11

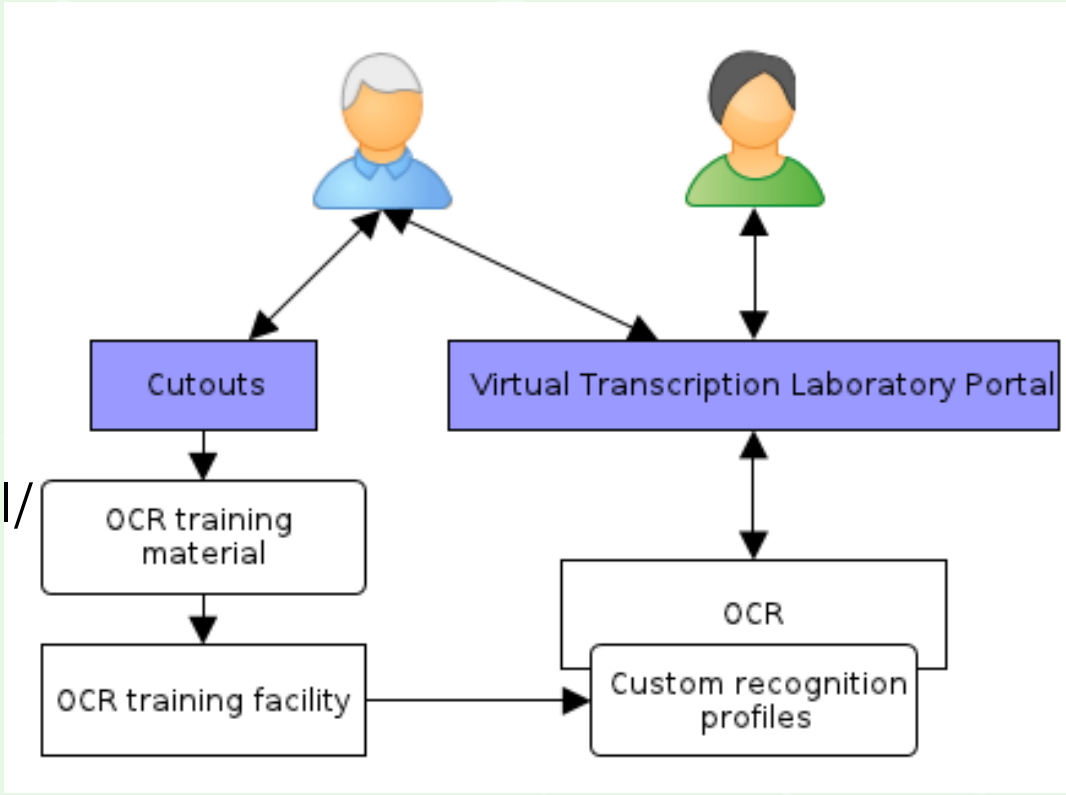
- Zaawansowane funkcje WRDZ:
  - Przechowywanie metadanych (kontener metadanych METS)
  - Ekstrakcja metadanych z dostarczonych informacji o obiekcie źródłowym
  - Migracja i konwersja danych źródłowych
  - Monitorowanie systemu
  - Komunikacja w ramach infrastruktury WRDZ
  - **Współdzielenie usług konwersji i migracji**
  - **Współdzielenie informacji o dostępnych zasobach**



Etap A12

- Główny wynik:
  - Prototyp Wirtualnego Laboratorium Transkrypcji
  - Prototyp systemu do przygotowywania materiału treningowego dla silników OCR („Wycinanki")
- Serwisy dostępne obecnie dla wszystkich pod adresem:

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



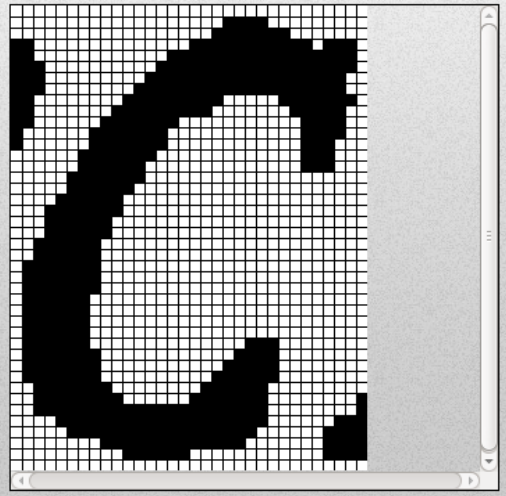
Wycinanie znaków

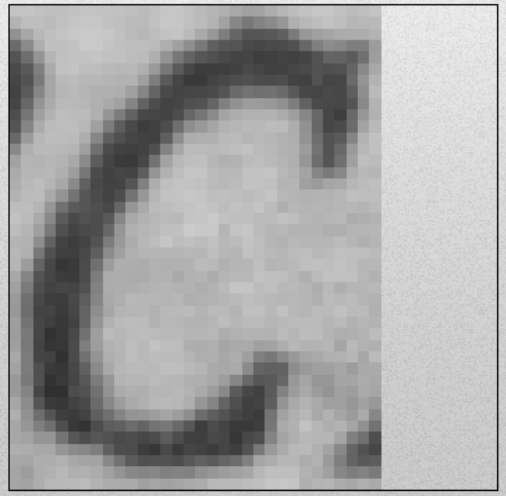
Pozostało: 2488 nieczytelnych: 1 błędnych: 7 opracowanych: 22

Historia opracowanych znaków

znak	typ czcionki	opis
z		
w		
a		
R		
R		
H	kursywa	

Identyfikacja znaku





Zmień stopień binaryzacji - +

Korekty wykonane pędzlem cofnij zmiany

Formatowanie znaku: *I* **B**

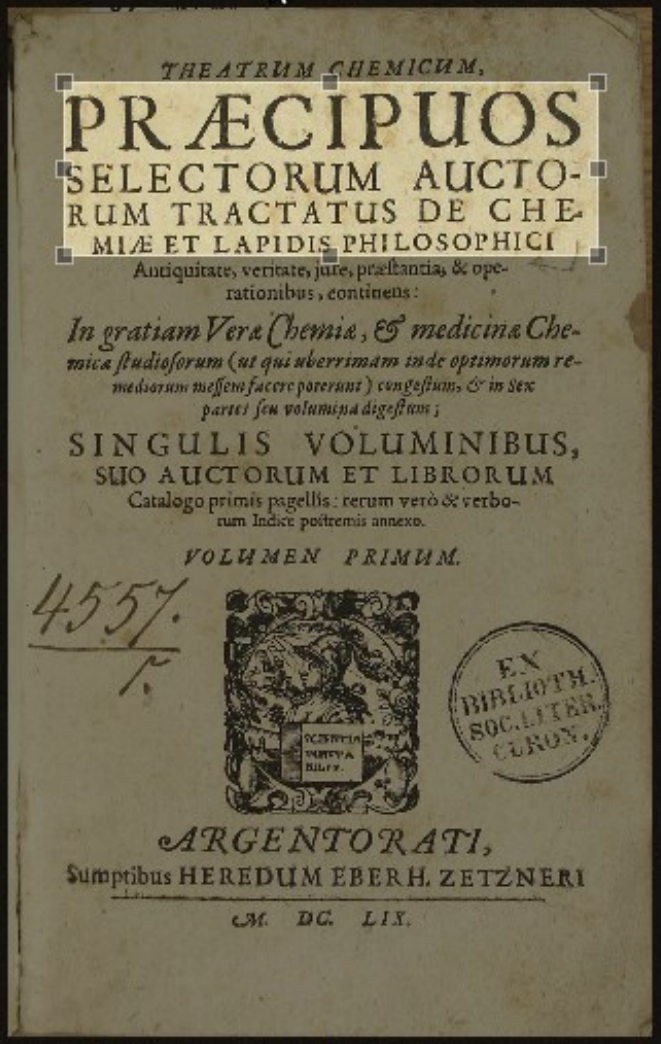
Wpisz znak: c

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 possent) congestum, & in sex partes seu volumina digestum;

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

VOLUMEN PRIMUM.

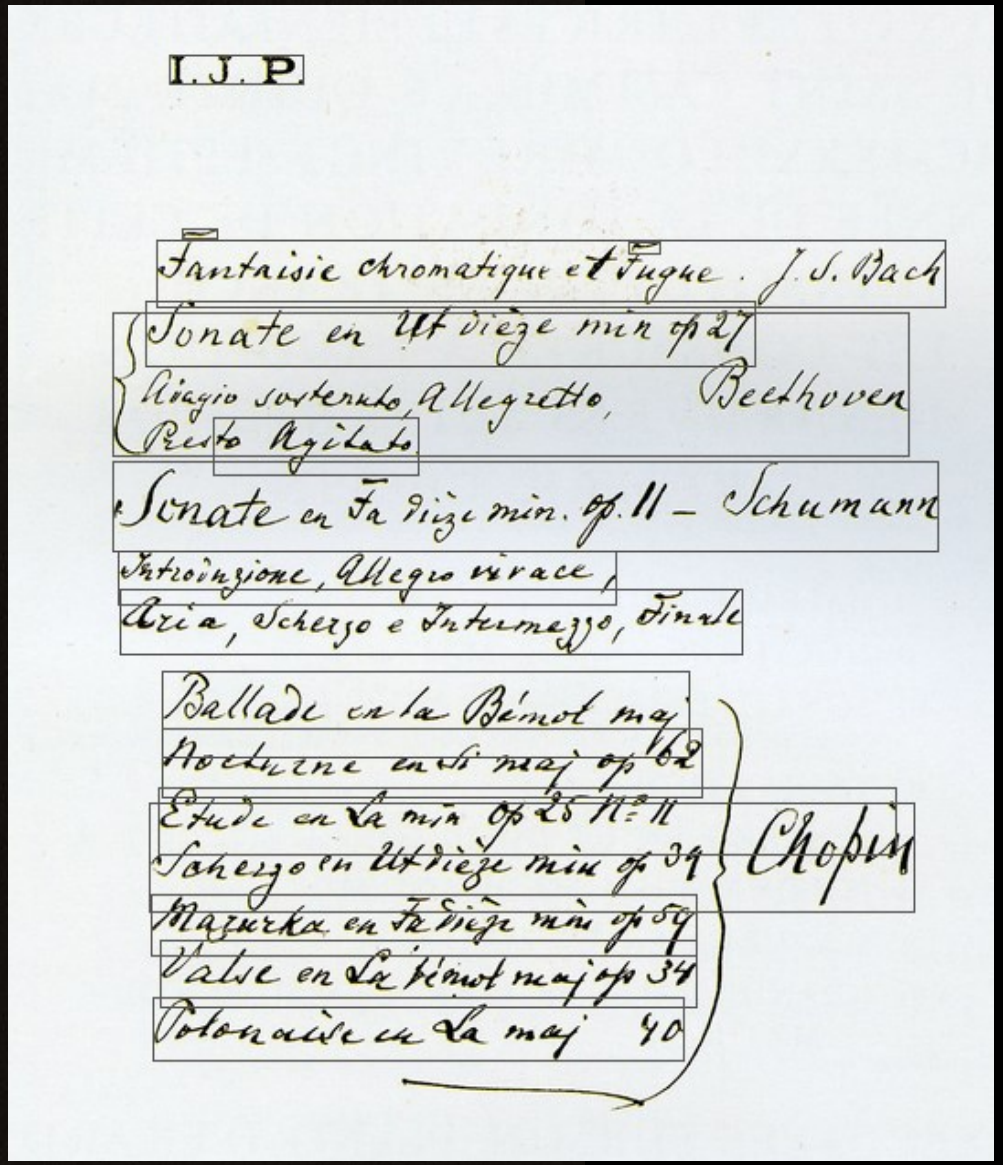
4557.
f.

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, Beethoven
Presto Agitato.

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.o 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

Interfejs webowy usługi OCR

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

topień ponad wodną 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.

*

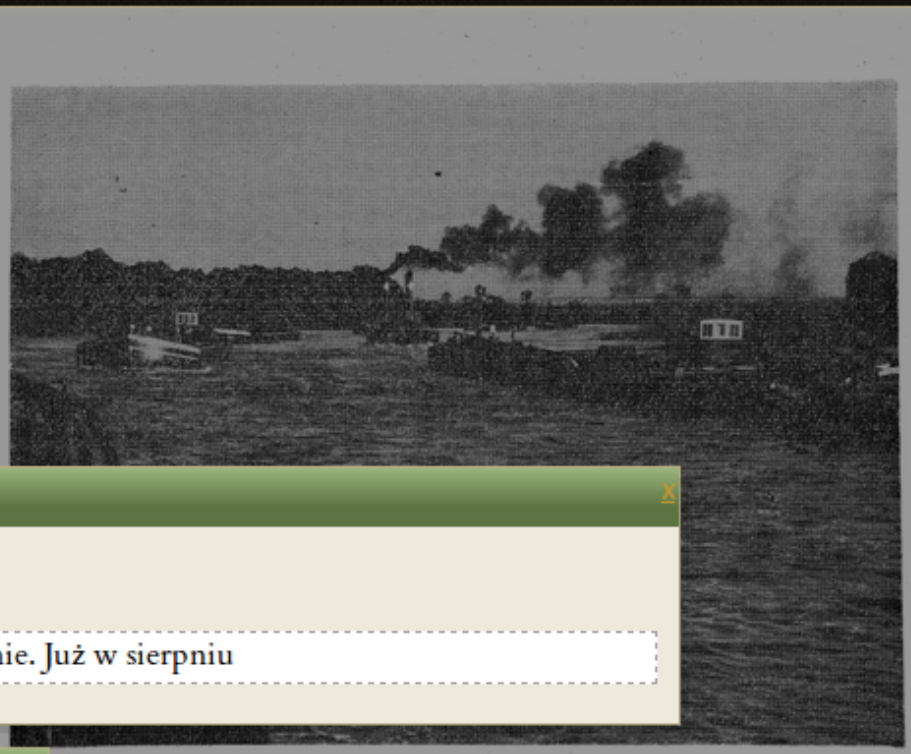
598 zakładów przemysłowych, skupiających ponad 63 tys. pracowników. Osiem wyższych uczelni. Ponad 15 tys. studentów. Dwa wydawnictwa... Wrocław.

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 ji przemysłowej wynosi 2,6%. Zniszczenia wojenne są coraz mniej Miastu przybywa około 6.000 izb mieszkalnych rocznie. Rosną nowe

Edytor transkrypcji

MIRTUALNE 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.

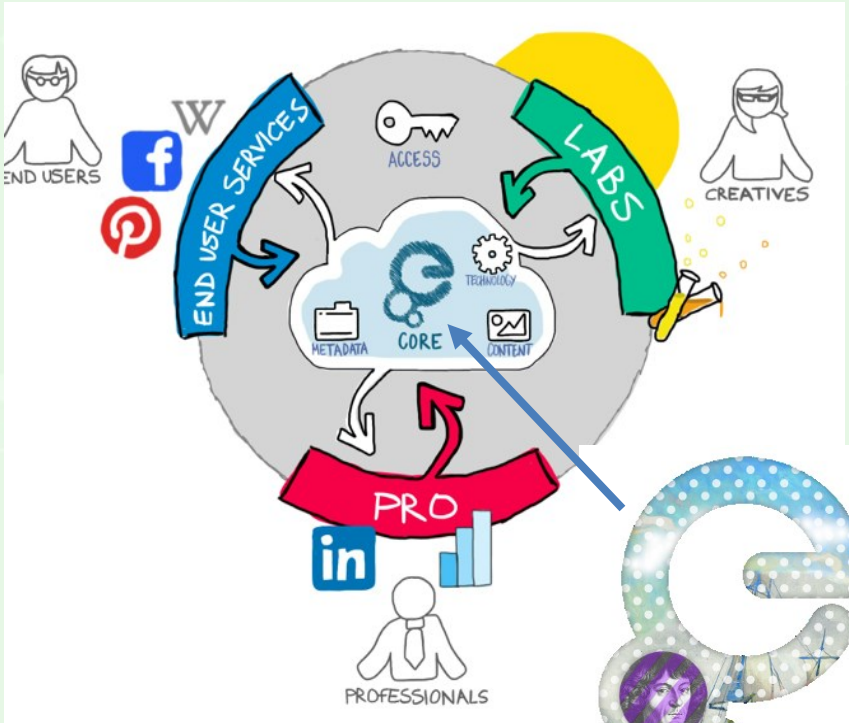
Mówi się o nim, że jest stolicą Nadodrza i nie jest to tylko jakieś określenie honorowe. Wrocław jest potężnym ośrodkiem pr
mającym niemały udział w kształtowaniu oblic
i całego kraju. Ludność Wrocławia stanowi 1,4%
w produkcji przemysłowej wynosi 2,6%. Zniszczenia wojenne są coraz mniej widoczne. Miastu przybywa około 6.000 izb mieszkalnych rocznie. Rosną nowe

Edytor transkrypcji

Europejski kontekst wykorzystania wyników prac PCSS

- Główne elementy koncepcji architektury systemu agregacji i wzbogacania danych Clepsydra wypracowanego przez PCSS w etapie A9 zostały wykorzystane przy projektowaniu chmurowego systemu agregacji i udostępniania danych **Europeana Cloud** (http://pro.europeana.eu/web/europeana-cloud)

- System ten będzie wdrożony do końca 2015 roku i stanie się technologiczną podstawą transformacji Europeany z portalu dostępowego do platformy na której każdy może budować swoje aplikacje
  - Europeana Cloud będzie kluczowym komponentem rdzenia tej platformy
  - Federacja Bibliotek Cyfrowych będzie jednym z trzech pierwszych użytkowników tego systemu (obok samej Europeany i The European Library)



europeana
cloud

Project Home | eCloud Blog | Partners & Work Packages | Project Documents | Newsletter | Communication Tools | Calend

LATEST NEWS

The first half of the Europeana Cloud project has seen plenty of discussion about the shape of the project. Those discussions are now feeding into the creation of some key publications, which will

Read More »

The Europeana Cloud project is currently developing a shared infrastructure with 3 specific aggregators (The European Library, The Polish Library Digital Federation and Europeana itself). But the

Read More »

Europeana Cloud will change the way that data is sent to Europeana, and will give researchers new tools to enrich and use that data.

Europejski kontekst wykorzystania wyników prac PCSS

- Wyniki prac etapu A11 zostały wykorzystane w projekcie europejskim SCAPE
  - Oprogramowanie dArceo oparte na prototypie WRDZ zostało zintegrowane z platformą SCAPE służącą do masowego przetwarzania danych na potrzeby długoterminowej archiwizacji
  - Scenariusz integracji został opracowany pod kątem składowania danych medycznych

SCAPE
SCAlable Preservation Environments

- Doświadczenie zyskane podczas prac realizowanych w etapie A11 pozwoliło uzyskać PCSS status członka w Fundacji Open Planets zrzeszającej instytucje z całego świata, zajmujące się problematyką długoterminowej archiwizacji
  - W ramach uczestnictwa w fundacji PCSS prowadzi dalej prace badawcze i rozwojowe związane z tematyką długoterminowej archiwizacji danych i aktywnie bierze udział w rozwoju narzędzi software'owych z tym związanych

Open Planets Foundation

Poznan Supercomputing and Networking Center Joins the OPF

We are delighted to welcome the Poznan Supercomputing and Networking Center (PSNC) as our latest affiliate member.

'PSNC brings new expertise and tools to the OPF membership', explained Ed Fay, Executive Director of OPF. 'They have developed dArceo, a long-term preservation system (http://dingo.psnc.pl/darceo/) which is already used by a number of institutions in Poland to preserve their digital content. As part of their membership contribution they will publicise the service as an open source package. In addition, PSNC will submit improvements to the FITS (http://projects.iq.harvard.edu/fits) code base, a tool which is widely used by both OPF members and the wider community'.

'As we speak about long-term preservation, our long-term goal is to improve PSNC's excellence in the field and share our expertise and tools with the preservation community' said Tomasz Parkoła, long-term preservation specialist, and member of PSNC's Digital Libraries Team. 'We believe that OPF can help us bring this to reality, by providing an excellent collaboration and networking environment. We are especially looking forward to sharing knowledge, creating new initiatives, investigating new ideas and running new projects. Our main focus is obviously on research and development activities."

The Poznan Supercomputing and Networking Center is a public ICT research and development institution working on broad range of topics, including network, storage, computing, applications and network services. It has been active in the long-term preservation and archiving domain for several years.

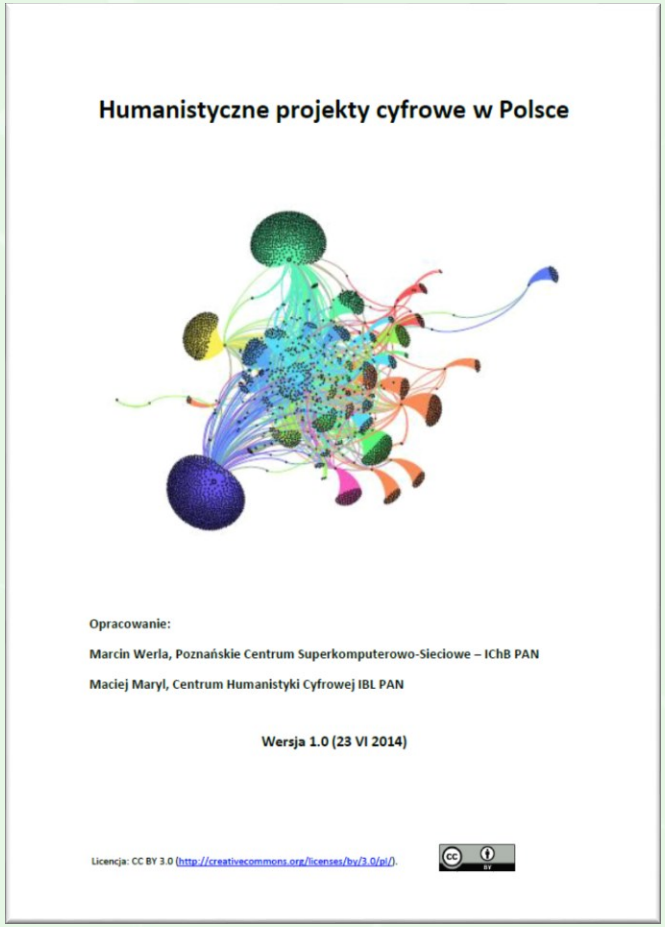
PSNC becomes the 19th member of Open Planets joining libraries, archives, research institutions, universities, and service providers collaborating on shared approaches to digital preservation.

Preservation Topics:
Open Planets Foundation
Submitted by Rebecca McGuinness on 29 April 2014 - 10:02am
Rebecca McGuinness's blog Log in or register to post comments

Europejski kontekst wykorzystania wyników prac PCSS

- Zaangażowanie się PCSS w ramach projektu SYNAT w środowisko polskiej humanistyki cyfrowej doprowadziło m.in. do nawiązania współpracy z Centrum Humanistyki Cyfrowej IBL PAN
- Obecnie CHC IBL PAN wspólnie z PCSS podejmują działania mające na celu powstanie polskiego konsorcjum instytucji zajmujących się badaniami w zakresie humanistyki cyfrowej i przyłączenie się tego konsorcjum do DARIAH - Digital Research Infrastructure for the Arts and Humanities

DARIAH-EU



Integracja w etapie A25 projektu SYNAT

- Wykorzystanie wyników prac PCSS w portalu INFONA
  - Etap A9: System Clepsydra
    - Konwersja otwartych metadanych z polskich bibliotek cyfrowych do formatu BWMETA
    - Udostępnienie metadanych w formacie BWMETA na potrzeby portalu INFONA
  - Etap A12: Wirtualne Laboratorium Transkrypcji
    - Możliwość logowania się do WLT przy użyciu konta portalu INFONA
    - Możliwość eksportu własnych obiektów z portalu INFONA do WLT w celu grupowej realizacji transkrypcji
    - Możliwość eksportu wyników transkrypcji z WLT do portalu INFONA m.in. w celu przeszukiwania pełnotekstowego

Dziękuję za uwagę!

I zapraszam do śledzenia naszych dalszych działań na stronie dl.psnc.pl

Poznańskie Centrum Superkomputerowo - Sieciowe

afiliowane przy Instytucie Chemii Bioorganicznej PAN,

ul. Noskowskiego 12/14, 61-704 Poznań,
tel : (+48 61) 858-20-00, fax: (+48 61) 852-59-54,
e-mail: office@man.poznan.pl, http://www.pcss.pl