Projekt zintegrowanego systemu informatycznego do studiowania polszczyzny XVII i XVIII wieku

Włodzimierz Gruszczyński
Instytut Języka Polskiego PAN

wlodzimierz.gruszczynski@ijp.pan.pl

Maciej Ogrodniczuk
Instytut Podstaw Informatyki PAN

maciej.ogrodniczuk@ipipan.waw.pl

Finansowanie projektu

Prezentacja powstała w ramach projektu pt.

Rozbudowa Elektronicznego Korpusu Tekstów Polskich XVII i XVIII w. i jego integracja z *Elektronicznym słownikiem języka polskiego XVII i XVIII w.*

finansowanego z funduszy Narodowego Programu Rozwoju Humanistyki MNiSW i realizowanego w latach 2019–2023 pod kierownictwem W. Gruszczyńskiego w IJP PAN we współpracy z Zespołem Inżynierii Lingwistycznej IPI PAN.

Założenia projektowanego systemu

1. Cel: stworzenie platformy informatycznej ułatwiającej studiowanie polszczyzny XVII i XVIII wieku.
2. Zasoby udostępniane obecnie odrębnie:
   a. *Elektroniczny słownik języka polskiego XVII i XVIII wieku* (*e-SXVII*),
   b. Elektroniczny Korpus Tekstów Polskich z XVII i XVIII Wieku (KorBa),
   c. Cyfrowa Biblioteka Druków Polskich i Polski Dotyczących z XVI, XVII i XVIII Wieku (CBDU),
   d. Kartoteka *Słownika języka polskiego XVII i 1. poł. XVIII w.*
3. Narzędzia:
   a. wyszukiwarki korpusowe,
   b. wyszukiwarki tekstowe,
   c. Chronofleks.

Dwa kierunki integracji zasobów historycznojęzykowych

- zasoby tego samego typu łączone na osi czasu (np. korpus diachroniczny),
- zasoby różnego typu łączone w obrębie tego samego okresu.

	przed XVI w.	XVI w.	XVII-XVIII w.	XIX w.	XX-XXI w.
słowniki			*e-SXVII*		
korpusy			KorBa		
biblioteki cyfrowe			CBDU		
kartoteki i in. zasoby			Kartoteka SXVII		

Zasoby podlegające integracji (1)

Elektroniczny słownik języka polskiego XVII i XVIII wieku

- dostęp: https://sxvii.pl/ lub https://xvii-wiek.ijp.pan.pl/pan_klient/;
- zawartość: ok. 39 tys. artykułów hasłowych na różnym stopniu opracowania (w tym ok. 25 tys. zalążków artykułów);
- formy hasłowe i in. formy fleksyjne (jedynie poświadczone) podawane w transkrypcji;
- cytaty dokumentacyjne podawane w kolejności chronologicznej w transliteracji;
- brak informacji statystycznej (ze względu na zmieniającą się podstawę materiałową).

Zasoby podlegające integracji (2)

Elektroniczny Korpus Tekstów Polskich z XVII i XVIII Wieku

- dostęp: http://korba.edu.pl/;
- zawartość:
  - obecnie 13,5 mln segmentów (ok. 700 tekstów z lat 1601–1772),
  - docelowo 25 mln segmentów (w tym ok. 4 mln z lat 1773–1800);
- teksty dostępne zarówno w transkrypcji, jak i transliteracji;
- teksty oznakowane morfosyntaktycznie przez tagery Concraft i Toygger;
- podkorpus o wielkości 0,5 mln segmentów oznakowany ręcznie;
- bardzo dokładne metadane źródeł;
- możliwości wyszukiwania wg różnych kryteriów.

Zasoby podlegające integracji (3)

Cyfrowa Biblioteka Druków Polskich i Polski Dotyczących z XVI, XVII i XVIII Wieku

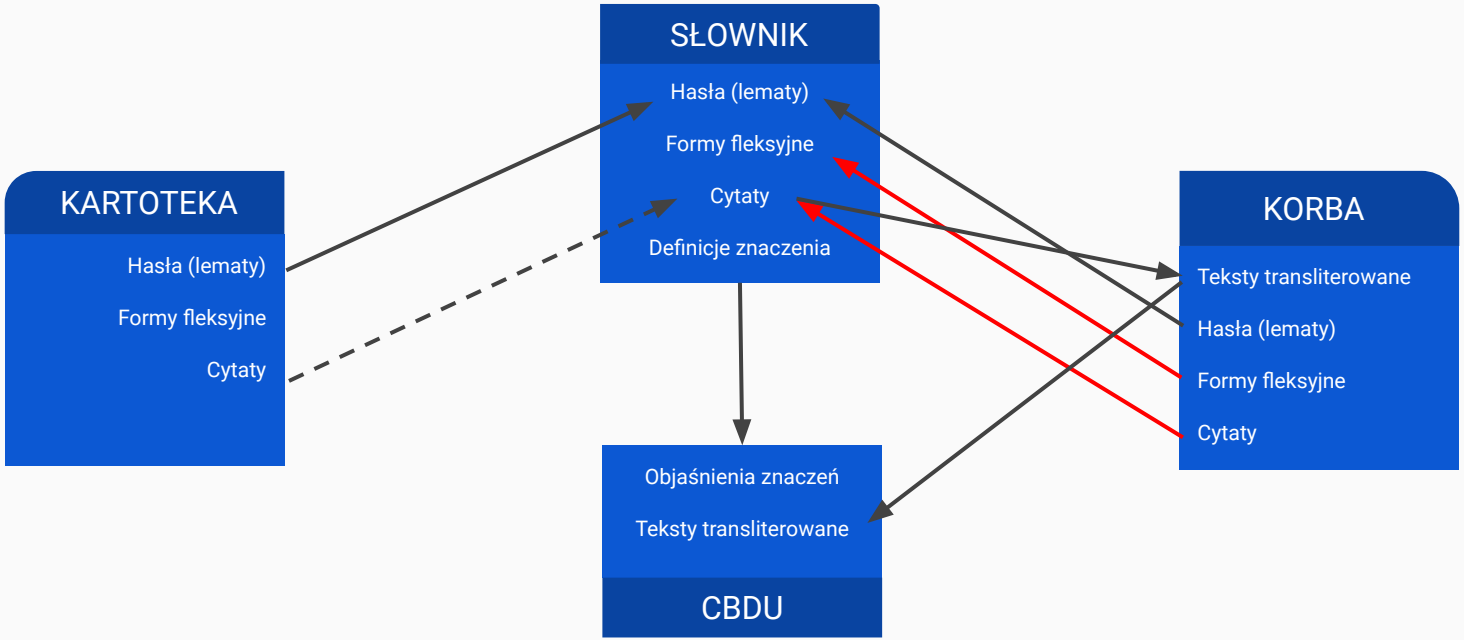
- dostęp: https://cbdu.ijp.pan.pl/;
- zawartość:
  - ok. 1500 skanów starodruków w 11 językach, w tym 326 w języku polskim,
  - bardzo bogate metadane, w tym informacje o relacjach łączących dany tekst z innymi (przeróbki, tłumaczenia);
  - słowniczki trudniejszych słów i zwrotów (przy tekstach polskich),
  - objaśnienia niektórych słów i zwrotów łacińskich;
  - komentarze historyczne dotyczące faktów opisywanych w tekstach.

Zasoby podlegające integracji (4)

Kartoteka *Słownika języka polskiego XVII i 1. poł. XVIII w.*

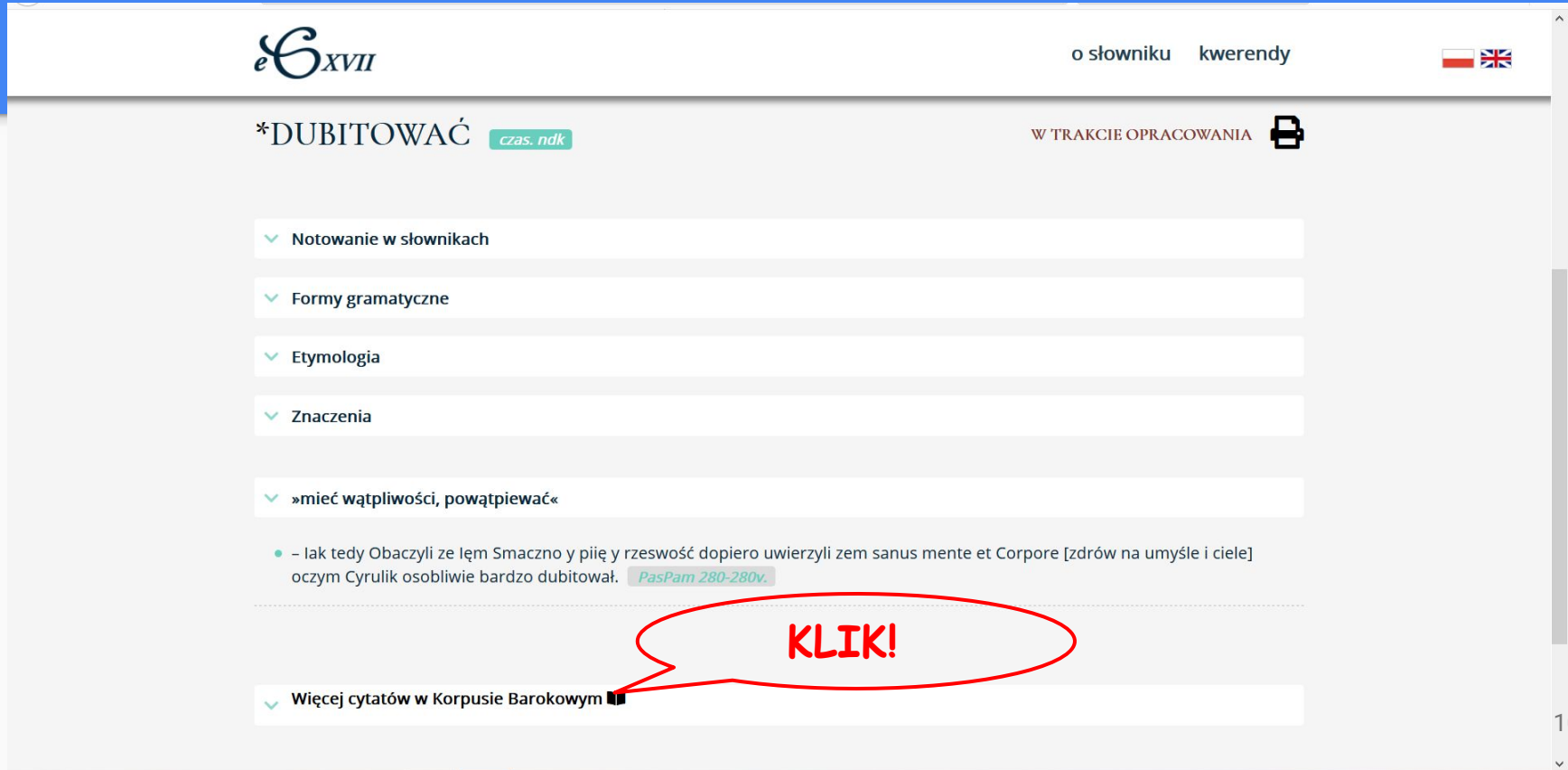
- dostęp: http://rcin.org.pl/dlibra/publication?id=20029&tab=3;
- zawartość: skany ok. 1,8 mln rękopiśmiennych kart cytatowych powstałych w latach 1955–2011 w wyniku ręcznej ekscerpcji źródeł z XVII i połowy XVIII w.;
- karty ułożone alfabetycznie wg form hasłowych wyróżnionych w cytacie wyrazów;
- część cytatów jest niekompletna;
- niektóre lematyzacje dyskusyjne.

Planowane i wprowadzone powiązania między zasobami



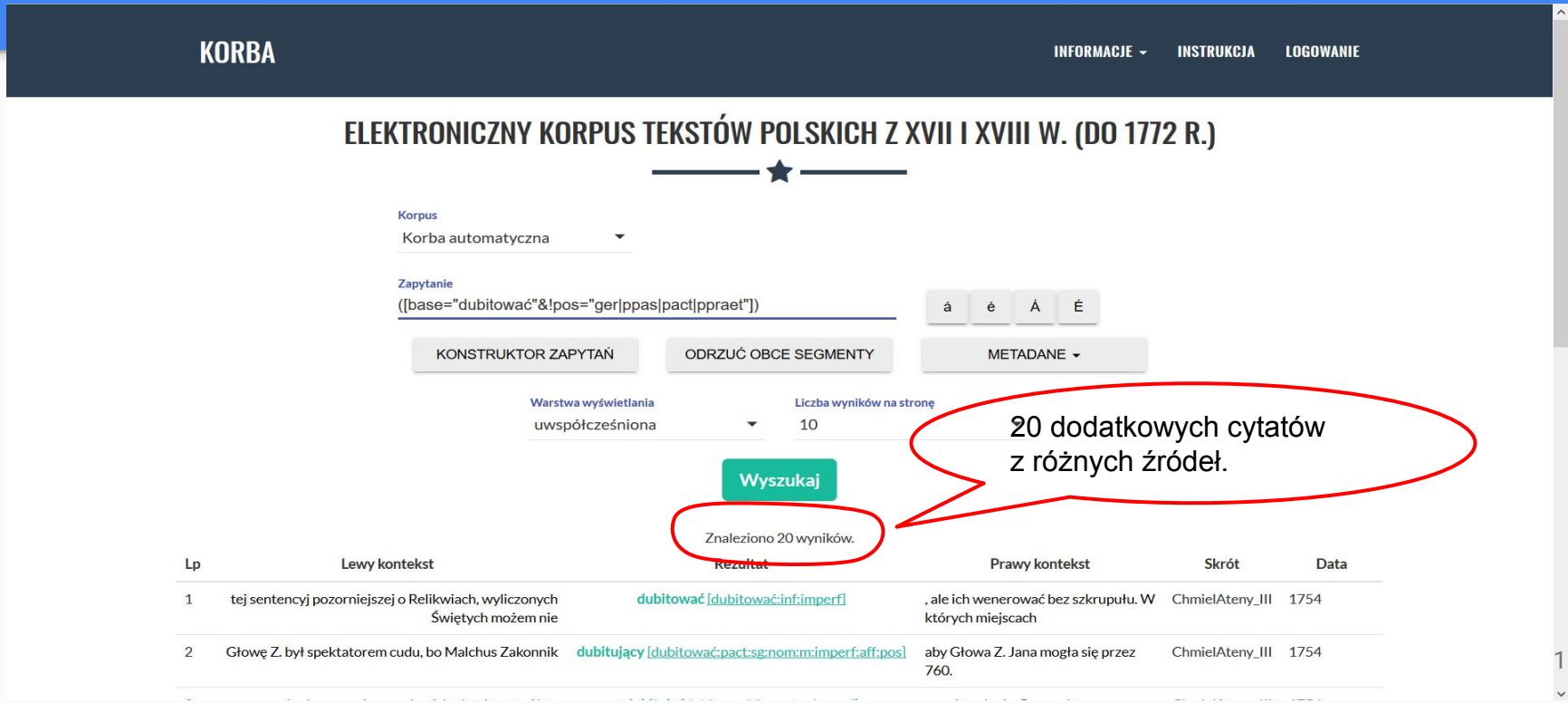
Wdrożone powiązania e-SXVII ← KorBa

1. Automatyczne wyszukiwanie wystąpień form wyrazu hasłowego w korpusie



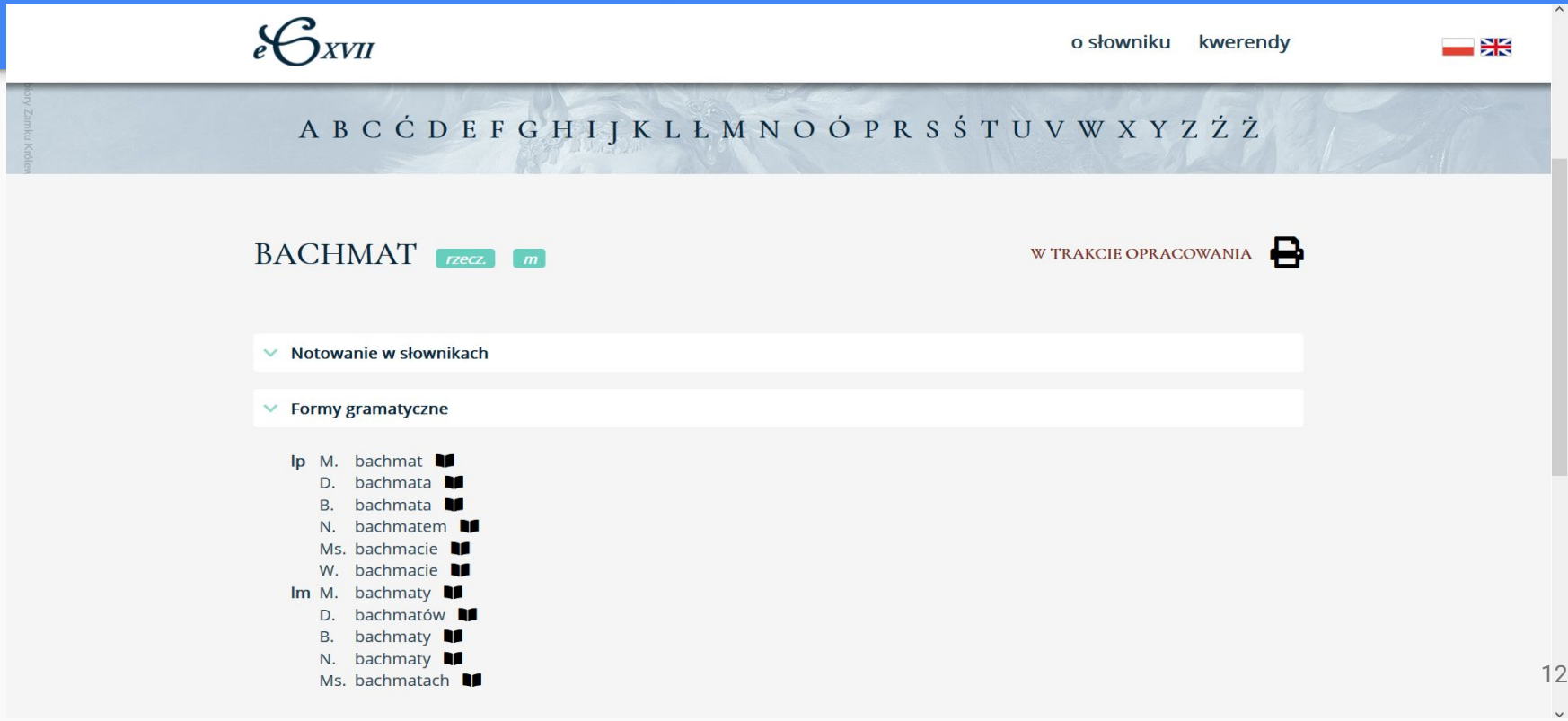
Wdrożone powiązania e-SXVII ← KorBa

1. Automatyczne wyszukiwanie wystąpień form wyrazu hasłowego w korpusie



Wdrożone powiązania e-SXVII ← KorBa

2. Automatyczne wyszukiwanie dodatkowych poświadczeń form fleksyjnych w korpusie



Braki we wdrożonych powiązaniach

1) Brak możliwości wyszukiwania form fleksyjnych niewykazanych w artykule hasłowym e-SXVII.
2) Problemy z wyszukiwaniem form fleksyjnych wariantów fonetycznych hasła.
3) Brak interfejsu dla autorów artykułów hasłowych w e-SXVII.

Eksperymentalne powiązania e-SXVII → CBDU

Cel:

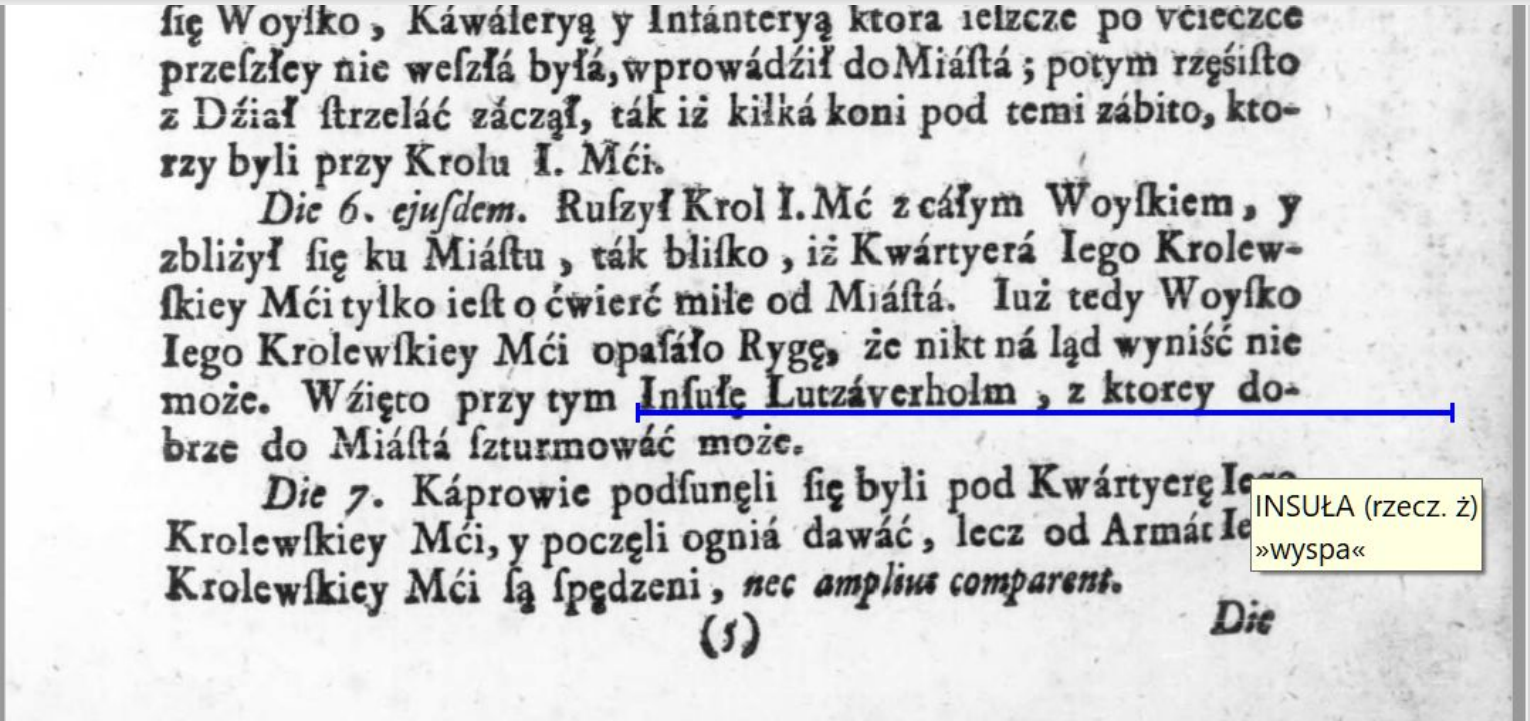
wykorzystanie definicji znaczeń zawartych w słowniku historycznym do objaśniania znaczeń wyrazów (archaizmów) wskazanych przez czytelnika w tekstach udostępnianych w bibliotece cyfrowej.

Eksperymentalne powiązania e-SXVII → CBDU

Problemy:

1. Konieczność utworzenia tekstowej wersji dokumentu dostępnego w postaci graficznej (pdf) i “podłożenia” jej pod wersję graficzną (skanowanie? przepisywanie?).
2. Konieczność lematyzacji każdej z form tekstowych, aby możliwe było odnalezienie odpowiedniego artykułu hasłowego w e-SXVII.
3. Wybór haseł (lematów), które są potencjalnie trudne dla czytelników (nieuwzględnianie haseł występujących w słowniku współczesnego języka polskiego? zmiana znaczenia hasła?)
4. Wybór definicji w wypadku haseł wieloznacznych (czy to w ogóle możliwe?)

Eksperymentalne powiązania e-SXVII → CBDU



ſię Woyſko, Káwáleryą y Intánteryą ktora ieizcze po vcieczce przeſzłey nie weſzłá byłá, wprowádźił do Miáſtá; potym rzęśiſto z Dźiał ſtrzeláć záczął, ták iż kiłká koni pod temi zábito, ktorzy byli przy Krolu I. Mći.

Die 6. ejuſdem. Ruſzył Krol I. Mć z całym Woyſkiem, y zbliżył ſię ku Miáſtu, ták bliſko, iż Kwártyerá Iego Krolewſkiey Mći tylko ieſt o ćwierć mile od Miáſtá. Iuż tedy Woyſko Iego Krolewſkiey Mći opaſáło Rygę, że nikt ná ląd wyniść nie może. Wźięto przy tym Inſuſę Lutzáverholm, z ktorey dobrze do Miáſtá ſzturmowáć może.

Die 7. Káprowie podſunęli ſię byli pod Kwártyerę Iego Krolewſkiey Mći, y poczęli ogniá dawáć, lecz od Armát Ie Krolewſkiey Mći ſą ſpędzeni, *nec amplius comparent.*

(5)

Die

INSUŁA (rzecz. ż)
»wyspa«

eSXVII

INSUŁA (rzecz. ż)
»wyspa«

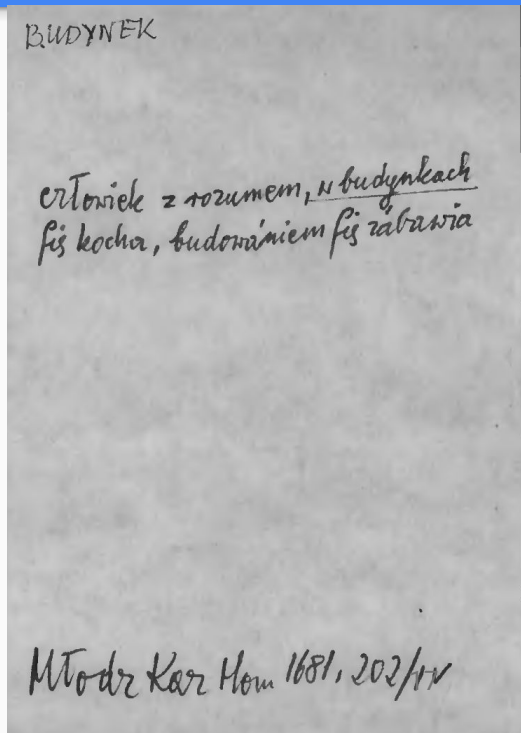
Próby powiązania kartoteki SXVII z e-SXVII

Cel:

Stworzenie narzędzia ułatwiającego autorom artykułów hasłowych w e-SXVII odnajdowania fiszek odnoszących się do opracowywanego artykułu hasłowego.

Problem:

Niemożliwe jest zastosowanie programów do OCR, ponieważ fiszki są w większości rękopiśmienne, robione przez różne osoby.



Próby powiązania kartoteki SXVII z e-SXVII

Eksperyment:

Podjęto próbę mającą na celu zindeksowanie (fragmentu) kartoteki, tzn. ustalenia dla każdego reprezentowanego w niej hasła adresu pierwszej fiszki odnoszącej się do danego hasła. Jeśli eksperyment się powiedzie, zindeksowana zostanie cała kartoteka i dzięki temu redaktor słownika (być może także jego czytelnik) będzie mógł uzyskać dostęp do odpowiedniego fragmentu kartoteki za pomocą jednego kliknięcia.

Dziękujemy za uwagę.