

Korpus polszczyzny XVI wieku

Etap I

projekt finansowany przez



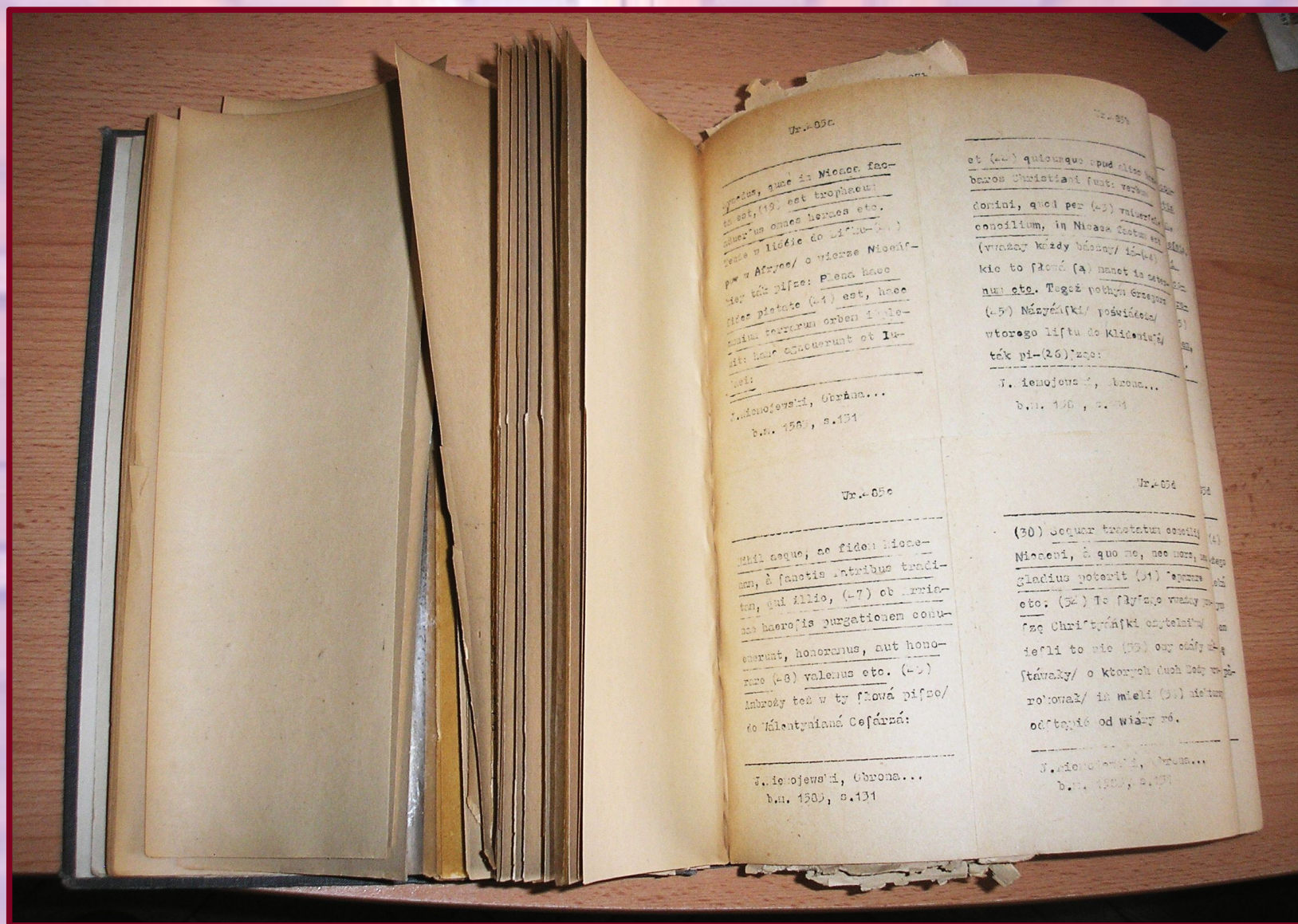
**NARODOWY PROGRAM
ROZWOJU HUMANISTYKI**

Korpus polszczyzny XVI wieku

Cele:

- Ocalenie bazy transliterowanych tekstów XVI-wiecznych – tzw. kanonu *Słownika polszczyzny XVI wieku*
- Udostępnienie kilku tekstów jako załączka korpusu
- Stworzenie narzędzi dla dalszego rozwoju korpusu

Korpus polszczyzny XVI wieku

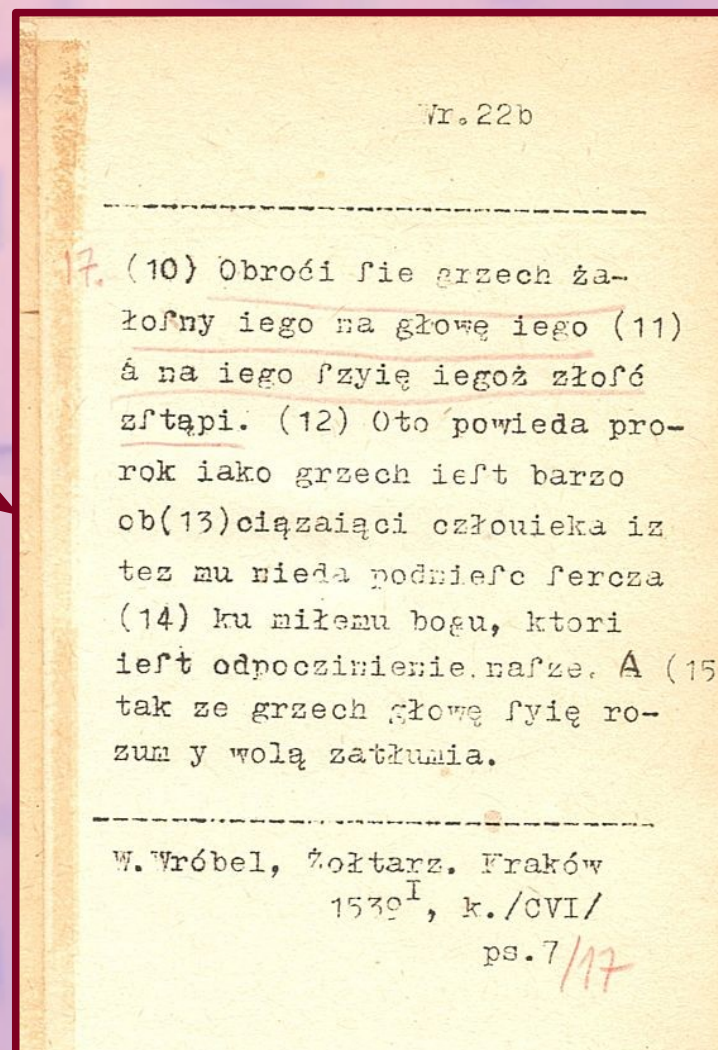
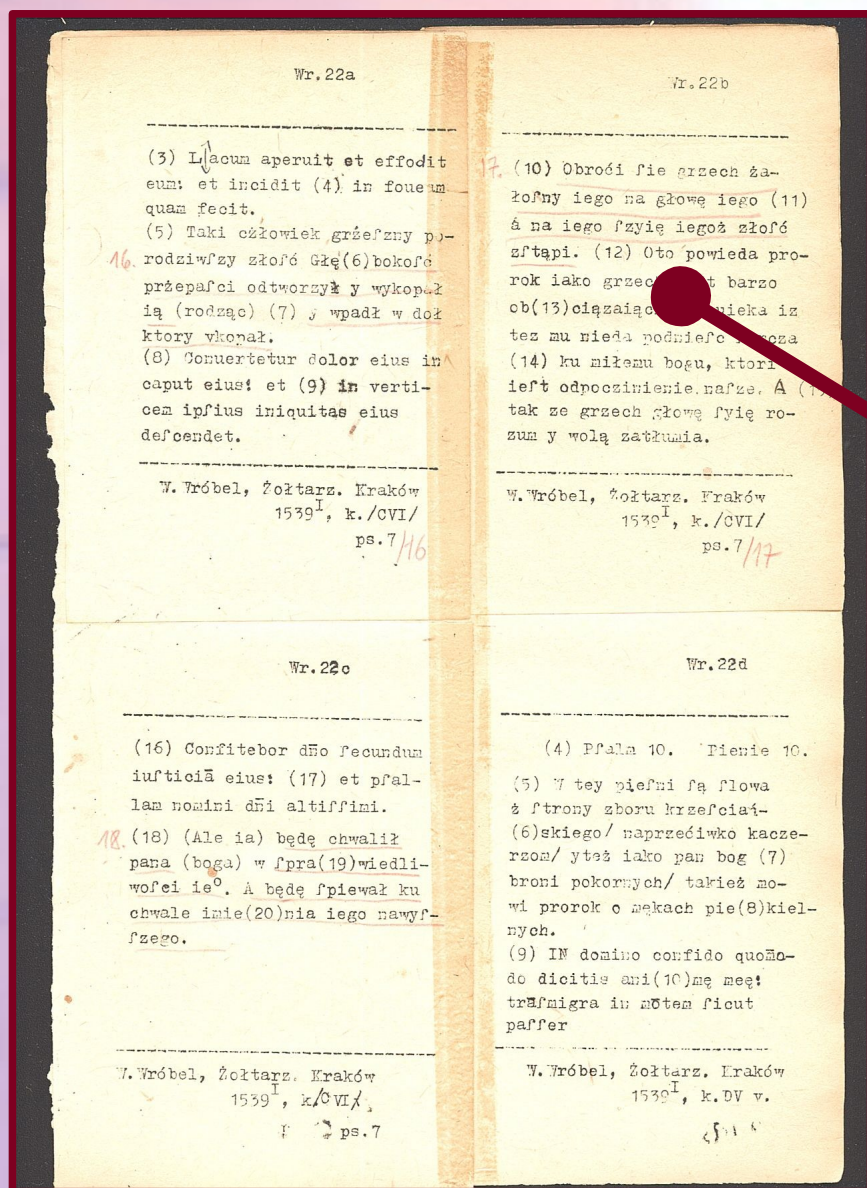


stan egzemplarzy archiwalnych

Korpus polszczyzny XVI wieku

matryca

ćwiartka - fiszka



Korpus polszczyzny XVI wieku

Założenia:

- format XML – TEI (P5) 
- transliteracja – postać oryginalna (nietypowe znaki, ligatury)
- pełna informacja lokalizacyjna (strona, wiersz, matryca)
- elementy typograficzne (marginesy, nagłówki, glosy, żywe paginy etc.)

Korpus polszczyzny XVI wieku

Kolejność prac:

- Opracowanie wyciągu z TEI (tagset) i ustalenie zestawu znaków (charset)
- Opracowanie edytora tekstów korpusowych
- Przepisanie tekstów
- Sprawdzenie z egzemplarzami archiwalnymi (1. korekta)
- Sprawdzenie z oryginałami (2. korekta – fotokopie, skany etc.)
- Opracowanie narzędzi do tagowania
- Tagowanie przykładowych tekstów
- Udostępnienie załączka korpusu

Korpus polszczyzny XVI wieku

Metody i narzędzia:

- TEI → Document Type Definition (DTD)
- **MUFI** Medieval Unicode Font Initiative ver. 3.0
(czcionka *Palemonas MUFI*)
- Python i Django
- Edytor tekstów XVI-wiecznych *KorpEd*
- Integracja z serwisem spxvi.edu.pl

Korpus polszczyzny XVI wieku

fragment DTD

```
129 <!ELEMENT head (#PCDATA | p | fw | sup | sub |  
choice | supplied | orig | reg | sic | damage | del  
| add | pb | cb | lb | milestone | decoNote |  
foreign | hi)*>  
130 = <!ATTLIST head  
131 —>rend CDATA "align(center)"  
132 >  
133 <!ELEMENT note (#PCDATA | p | sup | sub | choice |  
supplied | orig | reg | sic | damage | del | add |  
lb | milestone | decoNote | foreign | hi)*>  
134 = <!ATTLIST note  
135 —>type CDATA "gloss"  
136 —>place CDATA "margin"  
137 >
```

definicja elementu „margines”

Korpus polszczyzny XVI wieku

fragment tekstu XML

nagłówek XML (zwinięty)

nowy wiersz (pusty element)

kustosz

nowa strona (pusty element)

żywa pagina

```
1  <?xml version='1.0' encoding='UTF-8'?>
2  <!DOCTYPE TEI SYSTEM "korpus-xvi.dtd">
3  <TEI>
4  <teiHeader>
22 <text xml:lang="pl">
23 <body>
24 <div type="" n="">
25 ...
26 <lb n="16"/>dę niemąłą przez to. V bączné=
27 <lb n="17"/>go Páná tedy to nic. Ale v ó=
28 <lb n="18"/>wého Wilká/ który nigdy nie
29 <lb n="19"/>fyt/ thedy nic nie pómože/ nie=
30 <fw type="footer" place="bottom">
31 <fw type="kustosz" place="right">ma ón</fw>
32 </fw>
33 <pb n="R2"/>
34 <fw type="header" place="top">
35 <fw type="runHead" place="center">tych Książek.</fw>
36 </fw>
37 <lb n="1"/>ma ón przedsię ná cnotliwego
38 <lb n="2"/>bączności. Acz drugi baczy/ że
39 <lb n="3"/>iefth omyłká w Reieftrzech/
40 <lb n="4"/>przedsię thego nie będzie nic
41 ...
42 </div>
43 </body>
44 </text>
45 </TEI>
```

Korpus polszczyzny XVI wieku

KochOdpr KochOdpr.xml

Dokument | Edycja Narzędzia

Kustosz (fiv)

- Atrybuty (2)
- Wstaw
- Elementy puste
 - » Nowa strona
 - » Nowa kolumna
 - » Nowy wiersz nienumerowany
 - » Matryca
- Elementy i szablony
 - » Indeks górny [i]
 - » Indeks dolny [i]
 - » Uzupełnienie [i]
 - » Sic! [i]
 - » Usterka [i]
 - » Dopisek/Objaśnienie [i]
 - » Skrót
 - » Lege
 - » Skreślenie/Poprawka [i]
 - » Errata [i]
 - » Inicjał [i]
 - » Rycina/Ornament
 - » Wstawka obca [i]
 - » Wyróżnienie [i]
 - » p

[21] nálezch. **Inter caetera** trzy łą chory/ á trzeći iákoby

[22] Graeckim chorom przygánia: bo oni iuż oíobny

[23] chárakterem do thego máią: niewiem iáko tho w

[24] Polńkim ięzyku brzmieć będzie. Ale w tym niech

[25] będzie **arbitrium** W.M. álbo ráczey we wżytkim.

Aij

Bárzo-

strona: A2v

matryca: Wr2b

[1] Bárzobych tho był rad wczýnił/ żebych był łam

[2] praefens W.M. teraz łtużby łwe ofiárował/ ále

[3] mi złe zdrowie nie da. Nie radbych przedłię o-

[4] miełzkał przenośin W.M. iefli Salus ták będzie

[5] chćiáá. Zátym łie łáfce W.M. mego miłóści-

[6] wego Páná zálecám. Dat. w Czárnołefie dwu-

Znaki spoza klawiatury i encje

Grupa podstawowa

α	ε	ũ	p	ó	m	w	m
n	z	ś	ř	č	x	y	
ȳ	ÿ	α	β	γ			

Znaki modyfikujące

̂	̃	̄	̅	̆	̇	̈	̉
̊	̋						

Znaki specjalne

., ¶ § < > []

Š † ‡

Ligatury łacińskie

ō	ū	ē	ī	ā	œ	os	q
p	q	ſ	đ	đ	ŕ	z	z
9	'	3	f	9	o	3	R
æ	Æ	œ	CE	ā	ō		

Małe litery greckie

α	β	γ	δ	ε	ζ	η	θ
ι	κ	λ	μ	ν	ξ	ο	π
ρ	ς	σ	τ	υ	φ	χ	ψ
ω	'	'	'				

Duże litery greckie

A	B	Γ	Δ	E	Z	H	Θ
I	K	Λ	M	N	Ξ	Ο	Π

okno edytora KorpEd

Korpus polszczyzny XVI wieku

panel elementów

Kustosz (fw)

▣ Atrybuty (2)

Nazwa	Wartość	
type	kustosz	+
place	right	+

▣ Wstaw

▣ Elementy puste

- » Nowa strona
- » Nowa kolumna
- » Nowy wiersz nienumerowany
- » Matryca

▣ Elementy i szablony

- » Indeks górny [i]
- » Indeks dolny [i]
- » Uzupełnienie [i]
- » Sic! [i]
- » Usterka [i]
- » Dopisek/Objaśnienie [i]
- » Skrót
- » Lege
- » Skreślenie/Poprawka [i]
- » Errata [i]
- » Inicjał [i]
- » Rycina/Ornament

aktywny element

atrybuty

lista elementów

Korpus polszczyzny XVI wieku

panel znaków

znaki zapisane jako encje znakowe,
np. **´**

znaki zwykłe spoza klawiatury



Korpus polszczyzny XVI wieku

Stan realizacji na listopad 2013:

- większość z 271 tekstów tzw. kanonu jest już przepisana i sprawdzona pod względem zgodności z podstawą materiałową
- trwa opracowanie arkuszy przekształceń XSLT oraz XSL-FO

Korpus polszczyzny XVI wieku

przekształcenia XSLT

(przykładowy widok
w przeglądarce)

strona: A2

Wielmożne=
mu Pánu/ pánu Miko
láiiowi Firleiiowi z Dąbrowi=
ce

matryca: W1c

Káftellanowi Wiślickie=
mu. etc. Pánu mnie z dawná
wielce láfkáwému.

R

Ozumięć to
Miłościwy
pánie Wiślic=
ki/ iż Spráwá
káżda która ie=
dno ná pew=
nych á niepochybnych przycz=
nách/ vgruntowána/ álbo théz

matryca: W1d

iednostáyną Experiencyą ielt
doświadczoná/ może syc thák
ná Piśmie podać / iż sfośobem
iákoby iákiéy Náuki/ ludzióm

A ij

byśye

Korpus polszczyzny XVI wieku

tagowanie

Wynidź **roźlewáciu** krwie

BielKron 72

<w id="7840" form="1234">roźlewáciu</w>

id_Hasła	hasło	homonim	d
7838	rozleisty	0	
7839	rozlepić się	0	
7840	rozlewacz	0	
7841	rozlewaczka	0	

baza form gramatycznych

id formy

forma

1234

sb m V sg

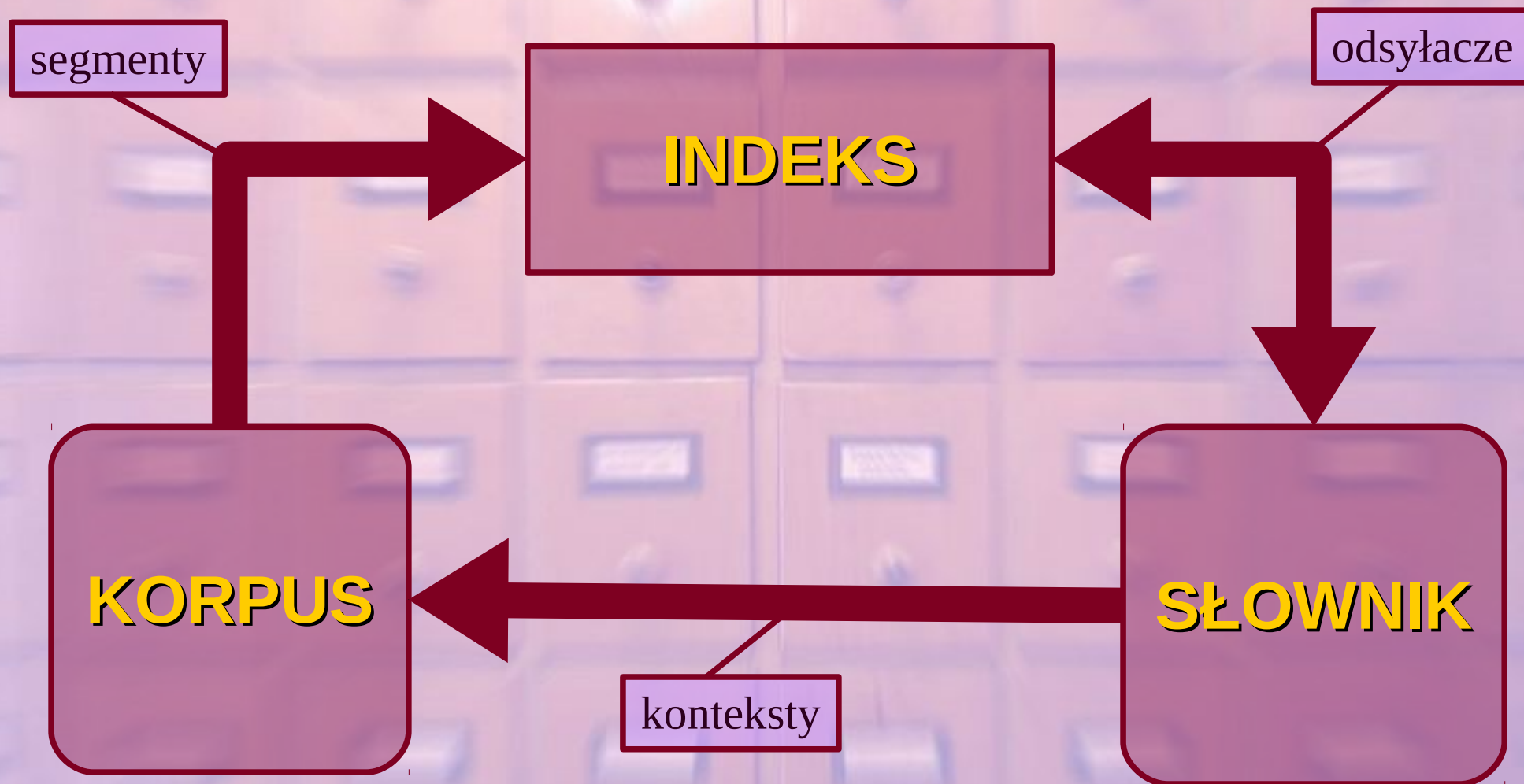
Korpus polszczyzny XVI wieku

Perspektywy:

- Integracja korpusu z elektronicznym indeksem *Słownika polszczyzny XVI wieku*
- Repozytorium tekstów polskich z XVI w.
- Możliwość połączenia z edycją transkrybowaną i normalizowaną
- Możliwość połączenia ze skanami oryginałów w bibliotekach cyfrowych

Korpus polszczyzny XVI wieku

integracja z e-słownikiem



Dziękuję za uwagę!