

# Wczytywacz i konwersja na XML

Maciej Ogrodniczuk  
Michał Lenart

Zespół Inżynierii Lingwistycznej  
Instytut Podstaw Informatyki  
Polskiej Akademii Nauk



INSTYTUT PODSTAW INFORMATYKI  
POLSKIEJ AKADEMII NAUK  
ul. Jana Kazimierza 5, 01-248 Warszawa

Sesja naukowa kończąca projekt Korpusu Barokowego  
Warszawa, 18 czerwca 2018

# Wczytywacz w pigułce

## Serwis internetowy:

- zarządzający procesem transliteracji tekstów (wgranie wstępnej wersji przez skryptora, przesyłanie poprawek, redakcja),
- wykrywający błędy w metaoznaczeniach (brakujące znaczniki, źle oznaczone języki obce itp.),
- dokonujący konwersji pliku w Wordzie do postaci XML-owej (DOC → TEI),
- wyświetlający wyniki konwersji w formie czytelnej dla użytkownika (co ułatwia znajdowanie błędów).

# Proces transliteracji

## Trzy role użytkowników:

- redaktor — inicjuje i zarządza procesem,
- skryptor — przepisuje tekst oraz oznacza go strukturalnie,
- korektor — weryfikuje anotację.

## Krok po kroku:

- 1 redaktor tworzy zapis w rejestrze tekstów (metadane),
- 2 redaktor przypisuje tekst skryptorowi,
- 3 skryptor opracowuje transliterację w edytorze Microsoft Word,
- 4 skryptor zapisuje dokument w formacie DOC,
- 5 skryptor wgrywa tekst na stronę aplikacji wczytywacza,
- 6 redaktor weryfikuje tekst i odsyła go do korekty lub do skryptora, jeśli błędów jest zbyt wiele.

# Widok listy tekstów

## KORBA — wczytywacz tekstów

renata (Redaktor)

[Strona główna](#)

[Nowy tekst](#)

[Wyszukiwarka](#)

Pracuj jako:

[skryptor](#)




[korektor](#)

[Zgłoś błąd](#)

[Zmień hasło](#)

[Wyloguj](#)

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18

Identyfikator	Metadane	Stan	Redaktor	Skryptor	Korektor	Słów w doc (w TEI)	Znaków	Tekst	Akcje
PetrSInst	<b>Sebastian Petrycy</b> <b>Instrukcja albo</b> <b>nauka jak się</b> <b>sprawować czasu</b> <b>moru</b> Mikołaj Lob Kraków 1613	Gotowy	renata			11077 (10942)	72526 (40,2str)	nagłówek   źródło   podgląd   XML	przywróć do redakcji 
BirkEgz	<b>Fabian Birkowski O</b> <b>Egzorbitancjach</b> <b>kazania dwoje</b> Andrzej Piotrkowczyk Kraków 1632	Gotowy	renata	renata		8657 (8340)	57010 (31,6str)	nagłówek   źródło   podgląd   XML	przywróć do redakcji 
NowZwyc	<b>Anonim Nowina o</b> <b>zwycięstwie cesarza</b> <b>z Wiednia do Lublina</b> <b>przyniesiona 17</b> <b>września 1634</b> 1634	Gotowy	renata	renata		206 (201)	1512 (0,8str)	nagłówek   źródło   podgląd   XML	przywróć do redakcji
TwarkPoch	<b>Kasper Twardowski</b> <b>Pochodnia miłości</b> <b>bożej</b> Walerian Piątkowski Kraków 1628	Gotowy	renata	renata	renata	10322 (10103)	67241 (37,3str)	nagłówek   źródło   podgląd   XML	przywróć do redakcji 

# Widok metadanych

## KORBA — wczytywacz tekstów

renata (Redaktor)

Strona główna

Nowy tekst

Wyszukiwarka

Raporty

Pracuj jako:

skryptor

korektor

Zgłoś błąd

Zmień hasło

Wyloguj

Wczytanych tekstów: ,  
zawierających w sumie ( wg  
TEI) słów.

Zakorzyszonych tekstów: ,  
zawierających w sumie ( wg  
TEI) słów.

Id: \*

Tytuł: \*

Autor: \*   Anonimowy

Tłumacz:   Anonimowy

Drukarnia:

Rok:

Rok wydania  
niepewny (wydano  
nie wcześniej, niż  
podany rok):

Typ mowy: \*

Rodzaj: \*

pieśni  fraszki i epigramaty  epitafia  satyry  sielanki  kazania  pisma polityczne  
 polemiki religijne  mowy okolicznościowe  traktaty  dialogi  pamiętniki  kroniki  relacje  
 opisy podróży  herbarze  akta sejmikowe  wilkierze  księgi sądowe  inventarze  rejestry  
Gatunek:  diariusze sejmowe  rozmówki do nauki języka  podręcznik  przysłowia  kalendarze  przewodniki  
 żywoty świętych  poematy epickie  przypowieści, specula (zwierciadła)  modlitwy  panegiryk  
 lamentsy  emblematy  przywileje  konstytucje sejmowe  bajki  księgi liturgiczne

alchemia  anatomia  architektura  astrologia  astronomia  biologia  botanika  budownictwo  
 chemia  egzotyka  ekonomia  filozofia  fizyka  geografia  gospodarstwo  gramatyka  
Tematyka:  górnictwo  historia  hutnictwo  języki  kulinaria  matematyka  medycyna  mitologia  
 miłość  muzyka  myślistwo  obyczajowość  poetyka  polityka  prawo  religia  retoryka  
 wojskowość  zielarstwo  zoologia  żeglarsstwo

Poetyka żartu:

Miejsce: \*

Region: \*

Wydanie

# XML-owy format reprezentacji danych

Bazujący na formacie Narodowego Korpusu Języka Polskiego:

- anotacja zewnętrzna,
- XML TEI P5.

`text_structure.xml`:

```
<front>
  <titlePage>
    <docTitle>
      <titlePart type="main">
        The DUNCIAD, VARIOURVM.</titlePart>
      <titlePart type="sub">
        WITH THE PROLEGOMENA of SCRIBLERUS.</titlePart>
      <docAuthor>Gal Anonim</docAuthor>
    </docTitle>
    ...
```

# Konwersja do XML-a i wykrywanie błędów

## DOC → TEI:

- zachowuje wszystkie teksty dostępne w źródle,
- wykrywa ewentualne błędy anotacji — w większości błędy strukturalne lub braki,

## Wykrywane typy błędów/ostrzeżeń:

- niekompletne znaczniki → brak możliwości zamiany tekstu na poprawny element XML-owy,
- błędnie zagnieżdżone znaczniki (np. znacznik nowej strony przed zakończeniem poprzedniej),
- tekst poza oznaczeniem stron (prawdopodobnie wynik pominięcia znaczników),
- niespójne oznaczenie języka obcego (częsty błąd: oznaczenie jedynie części słowa obcego).

Dziękuję!



**MERKVRIVSZ  
P O L S K I,**

Dzieie wſzytkiego ſwiátá w ſobie zámykáajúcy  
dla Informácyey poſpolitey.

*W Krákwie 3. Ianuarij 1661.*