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` $\rightarrow$ `TEI`),
- wyświetlający wyniki konwersji w formie czytelnej dla użytkownika (co ułatwia znajdowanie błędów).

Proces transliteracji

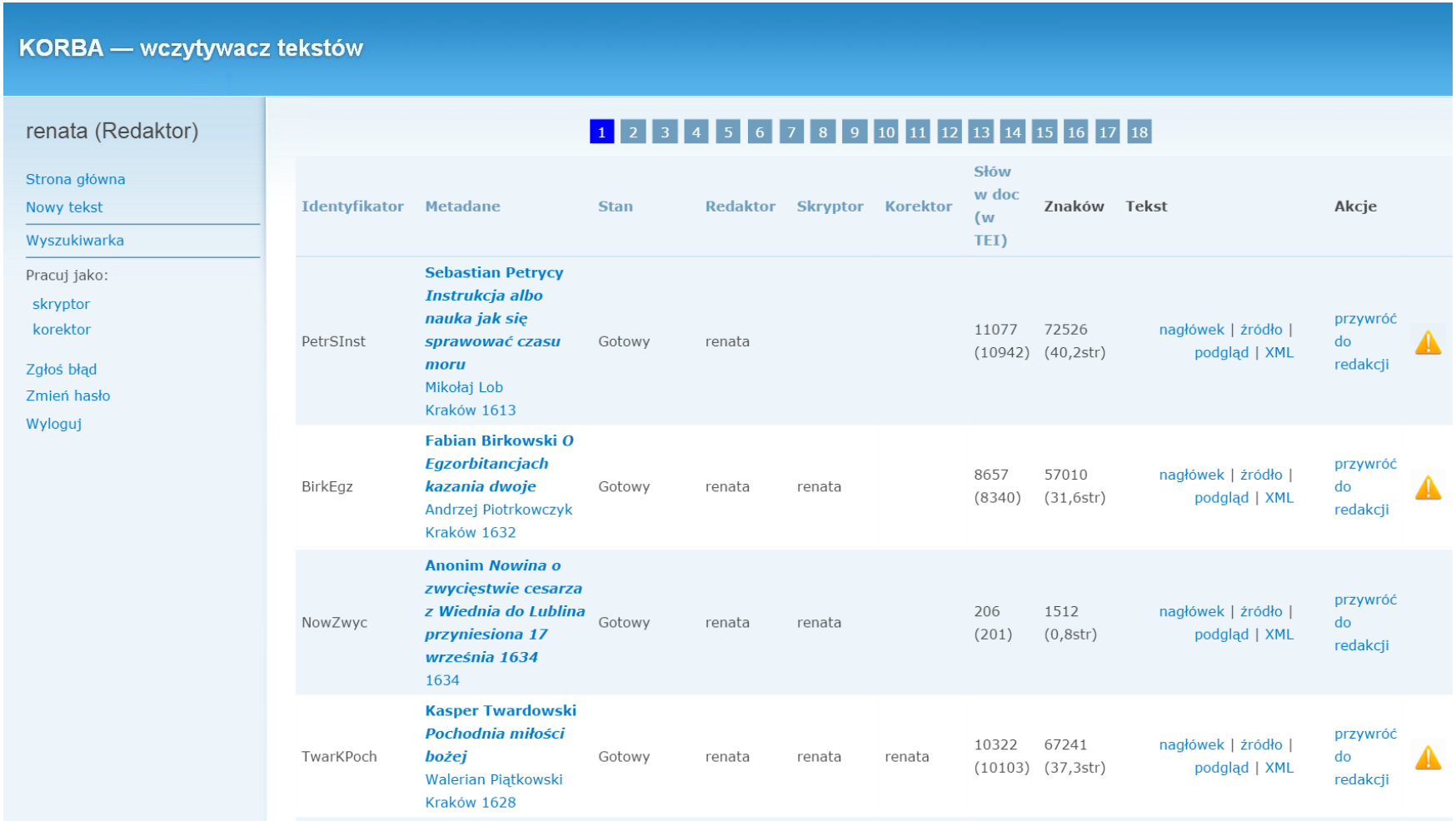
Trzy role użytkowników:

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

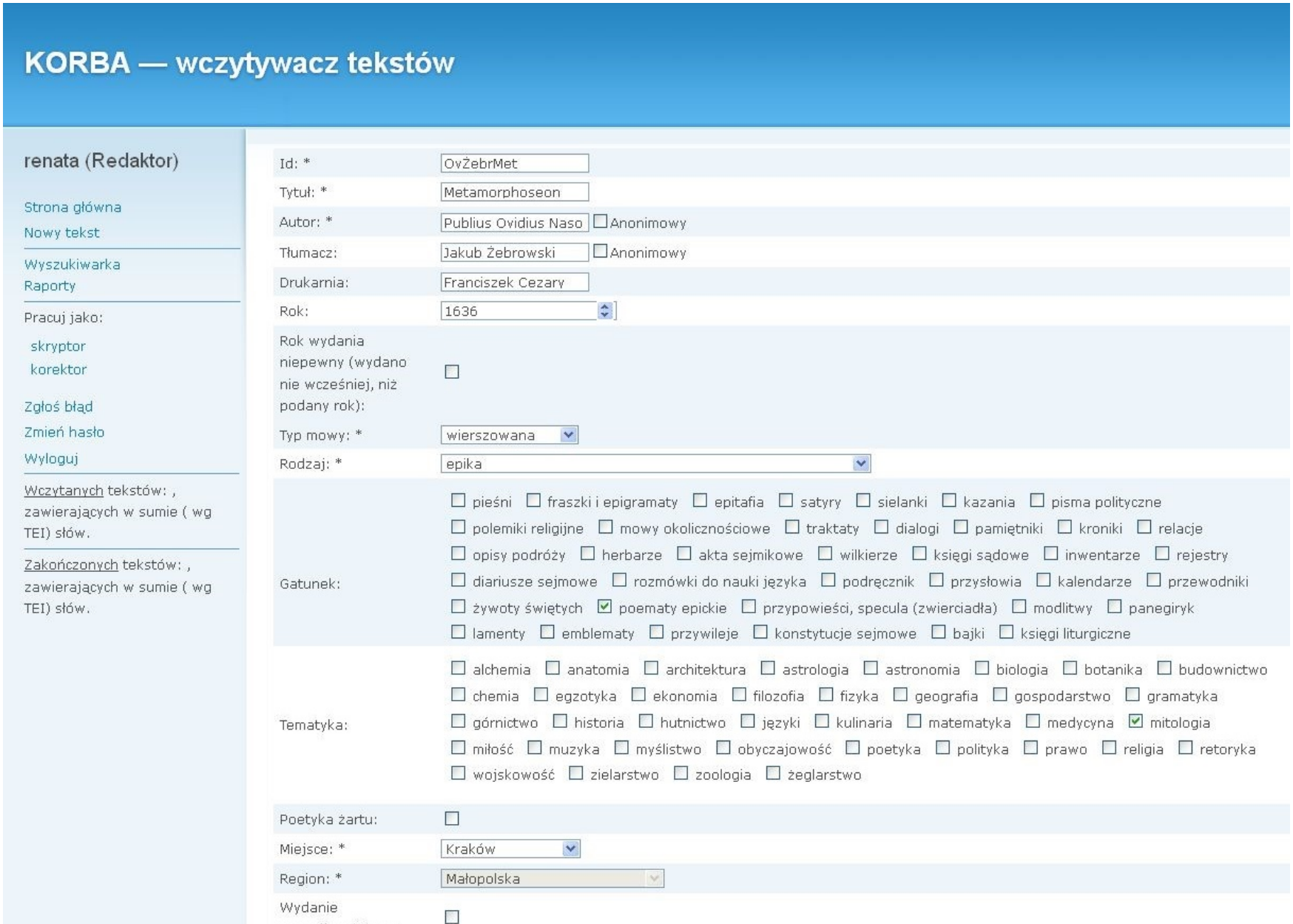
Krok po kroku:

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



Widok metadanych



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

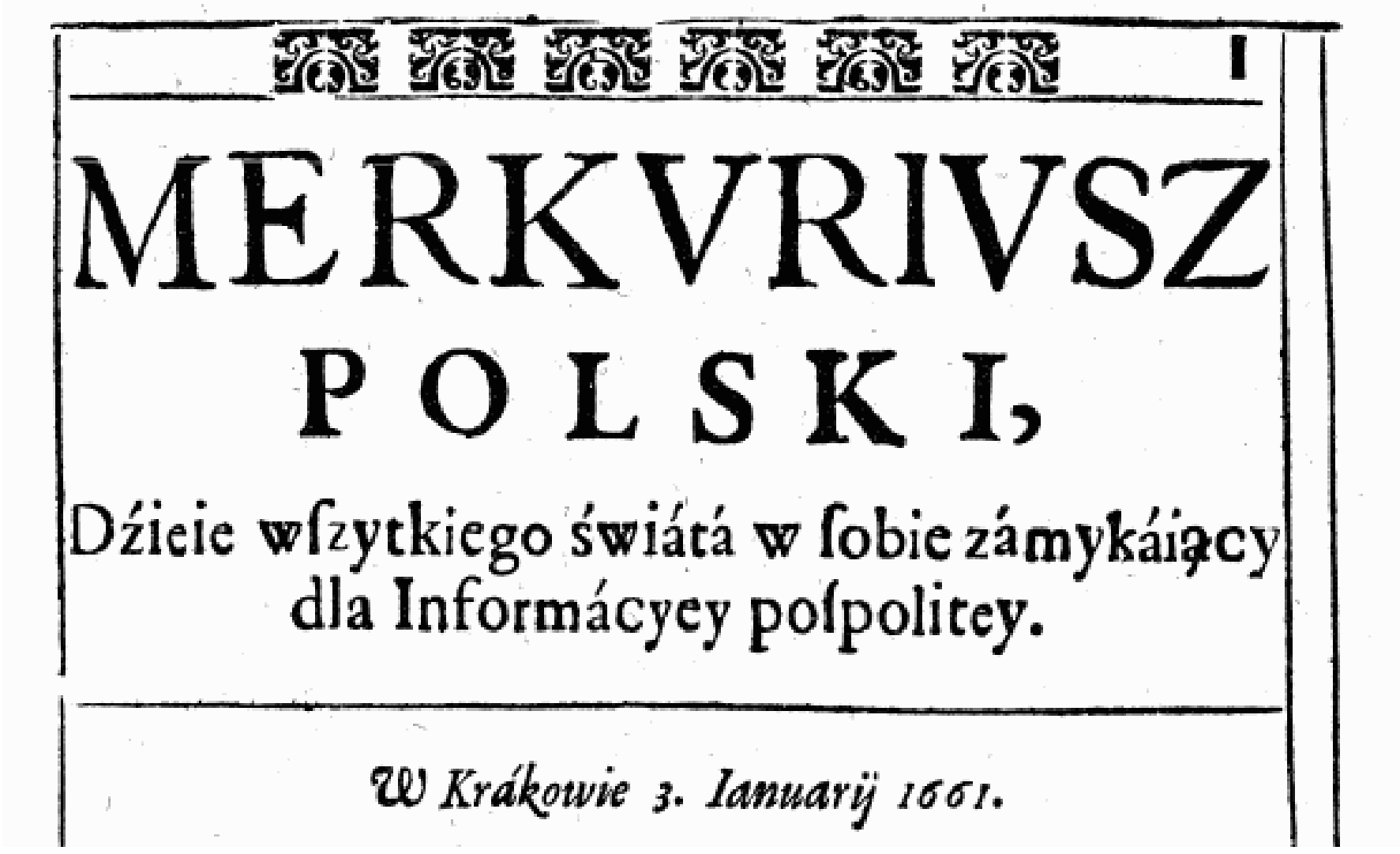
DOC → TEI:

- 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!



I

MERKVRIVSZ

POLSKI,

Dźieie wſzytkiego świátá w ſobie zámykáiący dla Informácyey poſpolitey.

W Krákowie 3. Ianuarij 1661.