Format dokumentów w projekcie elektronicznego korpusu tekstów polskich z XVII i XVIII w.

Maciej Ogrodniczuk | Institute of Computer Science
Michał Lenart | Polish Academy of Sciences

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

Instytut Języka Polskiego PAN

21 listopada 2013

Struktura tekstu

Podobnie jak w NKJP, teksty w korpusie KORBA wraz z ich strukturą (podział na rozdziały, akapity, strony, przypisy, notki marginesowe itp.) znajdują się w plikach `text_structure.xml`.

```
<?xml version="1.0" encoding="UTF-8"?>
<teiCorpus xmlns="...">
  <xi:include href="KORBA_header.xml"/>
  <TEI>
    <xi:include href="header.xml"/>
    <text>
      <front>  <!-- strona tytułowa i potytułowa -->
      <body>   <!-- tekst właściwy - pozostałe strony -->
   </text>
 </TEI>
</teiCorpus>
```

Strona tytułowa i potytułowa

```
<front>
  <titlePage>
    ...
  </titlePage>
  <div type="postTitlePage">
    ...
  </div>
</front>
```

Użycie poszczególnych elementów TEI z przykładami:
`http://www.tei-c.org/release/doc/tei-p5-doc/en/html`.

Autor, tytuł i podtytuł

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

Drukarnia, miejsce, rok, numer wydania

```
<titlePage>
  ...
  <docImprint>
    <publisher>Jan Karcan</publisher>
    <pubPlace>Wilno</pubPlace>
    <docDate>1611</docDate>
  </docImprint>
  <docEdition>Wydanie pierwsze</docEdition>
  ...
</titlePage>
```

Części tekstu

Poszczególne części tekstu są reprezentowane przez elementy `<div>`; typ części znajduje się w atrybucie `@type`.

```
<body>
  <div type="volume">
    <p>...</p>
    <div type="part_level1">
      <!-- Rozdział 1 -->
      <p>...</p>
    </div>
    <div type="part_level1">
      <!-- Rozdział 2 -->
      ...
    </div>
  </div>
</body>
```

Tytuły, podtytuły i streszczenia części

```
<head type="title">O nieporządney chćiwośći
                   iedzenia,</head>

<head type="subtitle">to iest, o apetićie
                      przećiwnym ludzkiemu
                      przyrodzeniu.</head>

<head type="summary">POETA idąc zá zdániem dawnych
                     Pogan, trzy początki świátá,
                     y rzeczy będących ná nim,
                     bydz rozumiał: Bogá, Lepsze
                     przyrodzenie, y Chaos.</head>
```

Dopuszczalne atrybuty head

Wartość atrybutu `type`	**Znacznik strukturalny**
`volume`	[POCZĄTEK/KONIEC TOMU]
`part_levelN`	[POCZĄTEK/KONIEC POZIOMU N]
`motto`	[POCZĄTEK/KONIEC MOTTA]
`preface`	[POCZĄTEK/KONIEC PRE-TEKSTU]
`epilogue`	[POCZĄTEK/KONIEC POST-TEKSTU]
`index`	gdy zawiera indeks
`contents`	gdy zawiera spis treści
Wartość atrybutu `n`	wczytywane z tekstu znajdującego się wewnątrz znacznika [POCZĄTEK/KONIEC CZĘŚCI NUMEROWANEJ]

Podział na strony

Aktualnie w NKJP nie ma możliwości sprawdzenia, na której stronie tekstu źródłowego znajduje się znaleziony fragment.
Aby to umożliwić, do dokumentów `text_structure.xml` został włączony znacznik <pb> (ang. *page break*) oznaczający początek nowej strony.

Atrybut `n` zawiera numer strony lub inne oznaczenie identyfikujące.

Początek dokumentu zawiera oznaczenie strony pierwszej:

```
<front>
  <!-- początek strony tytułowej -->
  <pb n="1"/>
  <titlePage>
    ...
  </titlePage>
</front>
```

Śródtekstowy podział na strony

```
<body>
  <div type="volume">
    ...
    <p>Womit nic inszego nie iest/
       iedno wyrzucánie przez vstá tego/
       coby iedno w nim było. Cknienie záś
       iest początek niepewnego
       <pb n="2"/>
       poruszenia mocy wyganiáiącey/
       przez ktorą vśiełuie przez vstá wygnáć/
       to co iest żołądkowi przykro/
       ábo nieznośno.</p>
    <p>...</p>
  ...
</body>
```

Śródtekstowy podział na strony

Uwaga:

- Oznaczenie kustoszy jest przepisywane dla celów weryfikacyjnych, ale nie jest reprezentowane w formacie wynikowym.
- Jeśli słowo znajduje się na dwóch stronach jednocześnie (czyli zawiera przeniesienie), wówczas znacznik końca strony zostanie umieszczony za tym słowem.
- Każda strona zaczyna się znacznikiem <pb> — nawet strona tytułowa i potytułowa.

Śródczęściowy podział na strony

```
<body>
  ...
  <div type="volume">
    ...
    <div type="part">
      <head>Rozdział 1</head>
      ...
    </div>

    <!-- początek strony bezpośrednio w <div>,
         czyli np. pomiędzy częściami -->
    <pb n="12"/>

    <div type="part">
      <head>Rozdział 2</head>
      ...
```

Notki marginesowe

W notkach marginesowych zapisywane są różne fragmenty, które nie stanowią treści tekstu, ale które chcemy zachować.

```
<p>Wielkość zámyka w sobie/
   wielkie ábo nie wielkie iedzenie/
   y częste iedzenie. Co się tknie pierwszego.
   <note type="margin">Iák wiele ieść</note>
   Iák wiele mamy ieść/
   naznácza nam vśmierzenie głodu/
   to iest/ gdy się znieśie przez
   wźięty pokarm przyrodzona chćiwość iadłá.</p>
```

Uwaga: narzędzia automatycznie przetwarzające teksty muszą szczególnie uważać na ten znacznik, ponieważ jego zawartość może „zaśmiecać" właściwy tekst (notka marginesowa i przypis są jakby drugim wymiarem „płaskiego", jednowymiarowego tekstu).

Inne elementy

Przypisy:

```
<p>Y rzekł Mojżesz przed Pánem/
   mowiąc: Oto/
   Synowie Izráelscy nie usłuchali mię/
   á jákoż mię usłucha Fáráo/
   á jam nie obrzezánnych
   <note type="annotation">wolney wymowy nie mam.</note>
   warg?</p>
```

Podpis pod rysunkiem:

```
<note type="label">a. To iest tępy koniec kliniká,
ktory ma bydź zewnątrz w vlu.</note>
```

Żywa pagina:

```
<fw type="runningHead">O Przypadkách Brzemiennych.</fw>
```

Inne elementy

Znacznik strukturalny	Znacznik TEI
[ILUSTRACJA]	<figure/>
[WZÓR MATEMATYCZNY]	<formula/>
[ZAPIS NUTOWY]	<notatedMusic/>
[TEKST W JĘZYKU OBCYM]	<foreign xml:lang="kod"/>
[ALFABET NIEŁACIŃSKI]	<foreign xml:lang="x-nlws"/>
[INNA PRZERWA W TEKŚCIE]	
Znak specjalny	<g/>

Kody języków

Język	Kod
arabski	`ar`
czeski	`cs`
francuski	`fr`
hebrajski	`he`
hiszpański	`es`
litewski	`lt`
łacina	`la`
niemiecki	`de`
południowo-słowiański	`zls`
ruski	`x-ruski`
skandynawski	`x-skand`
turecko-tatarski	`x-turtat`
węgierski	`hu`
włoski	`it`

Kody języków

Oznaczenia rozpoczynające się literą `x` pochodzą spoza oficjalnej listy kodów językowych ISO; to tagi językowe do użytku prywatnego, wykorzystywane tam, gdzie standard okazuje się niewystarczający.

Z wewnętrznej korespondencji:

> *Zdecydowaliśmy się połączyć kilka języków w jedną grupę, bo w XVII w. niektóre języki albo jeszcze nie istniały, albo byłyby trudne do odróżnienia. Turecko-tatarski obejmuje turecki i tatarski, skandynawski — teoretycznie wszystkie języki skandynawskie, ale praktycznie szwedzki i duński. Zdecydowaliśmy się też nie używać określenia „ukraiński”.*

Kompatybilność

Opisany format zapisu jest oparty na NKJP, ale wprowadza wiele nowych znaczników. Narzędzia oparte na NKJP (np. Poliqarp, Pantera) obsługują tylko uproszczoną wersję formatu.

Wymagane jest dostosowanie wyżej wymienionych narzędzi lub stworzenie zubożonej wersji formatu.

Interfejs serwisu do konwersji

W ramach projektu powstaje serwis internetowy dla skryptorów i redaktorów.

Umożliwia on między innymi:

- konwersję plików `.doc` do formatu TEI za pomocą formularza na stronie,
- wykrywanie błędów w oznaczaniu tekstu (brakujące znaczniki, źle oznaczone języki obce itp.),
- podgląd wyników konwersji w formie czytelnej dla użytkownika (ułatwia znajdowanie błędów),
- zarządzanie procesem dostarczania tekstów — wgranie wstępnej wersji przez skryptora, przesyłanie poprawek, redakcja (jeszcze nie zaimplementowane).

Serwis do konwersji — strona główna

KORBA - wczytywacz tekstów

Strona główna Witaj **mlenart** | Administracja | Zmień hasło | Logout

---

Plik .doc: Choose File No file chosen

Wczytaj

Lista plików

ŻędzKom-poprawione-mlenart.doc	wynik konwersji	czytaj
próbka_KGO_1678-poprawione-mlenart.doc	wynik konwersji	czytaj
PetrSInst-mlenart.doc	wynik konwersji	czytaj

KORBA - wczytywacz tekstów

Strona główna Witaj **mlenart** | Administracja | Zmień hasło | Logout

---

- Błąd przy wczytywaniu pliku: Znacznik zamknięcia nieotwartej strony.

Strona 36

ze WLADYSLAWA IV spoiony dniá 13 Września rękomá Arcybiskupiemi w Krákowie zwielkim świátá wszytkiego weselem/ Państw/ Monárciy życzliwośćią Koronowána. záćmił/F2[KONIEC STRONY]
[KONIEC STRONY]
46
ORATOR
nie ostátnia/ nawiększe po Krolewsku do Niebá iść. Nászá Cecilia Renatá y po Krolewsku się vrodźiłá/ y po Krolewsku żyłá/ y po Krolewsku vmárłá/ y po Krolewsku do N

Plik .doc: Choose File No file chosen

Wczytaj

KORBA - wczytywacz tekstów

Strona główna Witaj **mlenart** | Administracja | Zmień hasło | Logout

---

- Dodano dokument PetrSInst-mlenart.doc, ale prawdopodobnie były błędy
- Prawdopodobnie źle zaznaczony język obcy

> R.Vnguenti sandalini ż 1.
> Pul Pimpinellae
> Scordii
> Mirrhae an -) 1
> Succi Citri żiiÿ
> Misce et f. linimen

- Prawdopodobnie źle zaznaczony język obcy

> R.Vnguenti sandalini ż 1.
> Pul Pimpinellae
> Scordii
> **Mirrhae an -)** 1
> Succi Citri żiiÿ
> Misce et f. linimentum pro corde.

Plik .doc: Choose File No file chosen

Serwis do konwersji — podgląd wyników

Strona tytułowa	
Tytuł	INSTRVCTIA Abo NAVKA/ IAK SIĘ SPRAWOWAC CZASV moru.
Podtytuł	W ktorey się zámyka: 1. Ochroná: Iáko się vchrániáć morowego powietrza. 2. Leczenie wszytkich niemal przypadków w nim/ gdzieby kogo opanowáło. Dla prostych nápisána/ krom discursow:
Miejsce wydania	W KRAKOWIE,
Drukarnia	W Drukárni Mikołáiá Lobá/
Rok wydania	Roku Pánskiego/ 1613.
Autor	Przez DOC. SEBASTYANA PETRYCEGO MEDYKA.

postTitlePage

motto

Trway w sławie/ poty/ nie<.>obyta Wieży/ Dokąd żołw wszytkiego świátá nie obieży: Abo dokąd mrowka morzá nie wypiie/ Lub dokąd vmárły znowu nieożyie.

preface

Szláchetnym y zacnym Pánom/ ICH MCIOM: P.

Serwis do konwersji — podgląd wyników

máterią ná wierzch ze wnętrza być wyrzuconą. Ták poznawszy powietrze teraz o tym będźiem mowić/ iák się go vchronić przez dietę/ y pewne lekárstwá. A potym iák go leczyć ma/ kto się go vchronić nie mogł.

Ochroná powietrza przez porządne życie.

Przestrogá generálna 1.

NIm o rzeczách do ochrony zarázy sczegulnie mowić pocznę/ napotrzebnieyszą powszechnią przestrogę naprzod położę. Do Páná Bogá się vćiecz z pokutą y prożbą.

O powietrzu.

Pan Bog przyczyna moru

Bo on sam táką plagę ná ludzi dopuscza: bez iego woli nie stáie się nic. Zadna ochroná nie pomoże/ gdy iego łaski y straży nad námi nie bedzie. Ptacy/ zá ie^o^ rządem y starániem swoię żywność máią/ á cożby ludźie nie mieli mieć: Ten ná ludzie przepuscza vporne choroby/ y prętkie zarazy: ten od nich broni. A to iescze dziwney táiemnicy Páńskiey przypisáć potrzebá/ iż czásem choc się kto nawięcey nátrąca ná nie/ minie y nie imie się go: a podczás tego/ co nie dáie przyczyny/ vchwyći. Co Doktorowie niektorzy doświádczyli náa sobie/ ktorzy choc chodźili do zarážonych/ niezárázili śiebie samych/ áni od śiebie drugich. Rychley ná tego padło co się nawięcey chronił.

Przestroga generalna 2.

Druga powszechnia przestrogá niech będźie. Ze złego powietrza/ na dobre vbiegáć. Co y nieme źwierzętá czynią: pod taki czás vćiekáią z iam swoich/ á inszych sobie szukaią zdrowych. A nie