Standardy metadanych (TEI Header)

Piotr Bański

Institut für Deutsche Sprache
Digital Economy Lab, UW

bansp@o2.pl

Warszawa, DELab UW, listopad 2014

Metadane

Ogólnie: dane o obiektach modelowanego przez nas świata

Jeśli obiektami, które nas interesują, są telksty, to metadane można podzielić na dwa rodzaje:

- metadane formalne (czasem po prostu „metadane”): odnoszące się do tekstu jako całości, np. autor, wydawca, tytuł serii, data powstania tekstu, rok wydania, itd.
- metadane analityczne (lubimy mówić o nich „anotacje”, nawet jeśli słowniki nie zawierają tego znaczenia): dane odnoszące się do poszczególnych fragmentów tekstu, na różnym poziomie: morfemów, słów, zdań, wypowiedzi. Np. części mowy, podział na zdania i frazy, klasyfikacja nazw, zależności referencyjne.

Na dzisiejszych zajęciach będę używał terminu „metadane” w tym pierwszym znaczeniu.

Temat na dziś: TEI header

Niezwykle szeroki zakres zastosowań.

Oprócz podstawowych wartości metadanych (takich, jakie zapewnia np. Dublin Core) oferuje także własności strukturalne.

Nie trzeba używać znakowania TEI żeby użyć nagłówków TEI!

Szybka lektura na kiedyś:

http://blogs.it.ox.ac.uk/jamesc/2013/04/20/self-study-part-5-the-tei-header/

Odpowiedni rozdział TEI Guidelines (niekrótki...):

http://www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html

oXygen – ułatwienie pracy z TEI

Czy koniecznie ten edytor?

Nie, to jest mój sposób na dzisiejsze zajęcia. Twórcy oXygena współpracowali z TEI praktycznie od początku i jest on obecnie bardzo często używanym edytorem, ale...

… warto także sprawdzić propozycje zamieszczone pod adresem

http://wiki.tei-c.org/index.php/Category:Tools

A na dziś, zakładam że każdy już ściągnał pakiet z

http://www.oxygenxml.com/download.html

I być może nawet zainstalowali go Państwo? Jeśli nie, to ciach...

(Wykorzystamy możliwość użycia oXygena przez miesiąc)

TEI Framework

Instrukcje mamy tutaj:

http://blogs.it.ox.ac.uk/jamesc/2014/04/02/auto-update-your-tei-framework-in-oxygen/

Ale zróbmy to wspólnie.

Nowy dokument

File | New... | Framework Templates | TEI P5: TEI Simple

Porównajmy:

File | New... | Framework Templates | TEI P5: TEI Corpus

Trochę zabawy i odkrywania: zepsujmy coś i zobaczmy, co się stanie.

Superkrótko o XML

Element: znacznik otwierający + zawartość + znacznik zamykający

<word>dziecko</word>

Atrybuty: wewnątrz znaczników otwierających:

<word pos=”rzeczownik” rodz=”nijaki”>dziecko</word>

Infrastruktura TEI

ODD = „One document does it all”

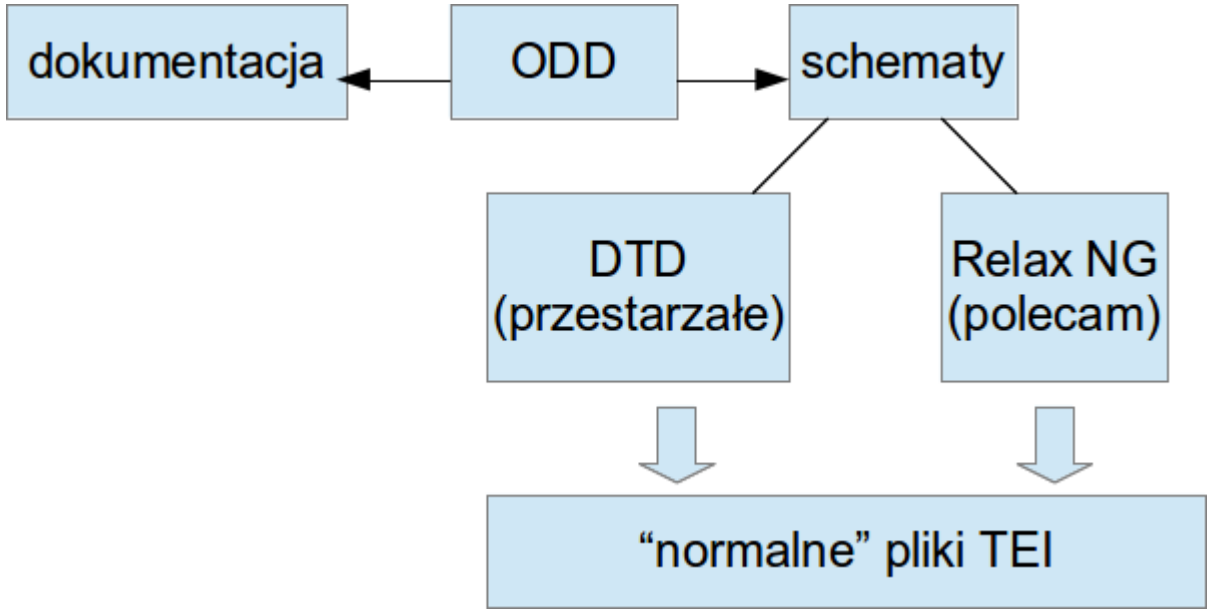
Plik ODD to plik TEI o specjalnych właściwościach:

- zawiera nagłówek (oczywiście), a jako zawartość znacznika <text> zawiera
- bardzo dokładną dokumentację (wraz przykładami) oraz
- opis modelu tekstu, który łatwo skonwertować do języka schematów.

http://wiki.tei-c.org/index.php/ODD

To co mam edytować? ODD czy...?

Dokumenty ODD to dokumenty XML „specjalnego przeznaczenia” – służą one do zmienienia (najczęściej zawężenia bądź modyfikacji) wybranych ogólnych schematów TEI. Ale proszę się nie bać: edytowanie TEI to niekoniecznie edytowanie ODD. Można się ODD nie przejmować i edytować te „normalne” pliki, których schematy derywowane są z ODD przez np. kierownika projektu.



Nawias: materiały TEI w zasięgu ręki

Warto poszperać na stronie TEI: http://www.tei-c.org/

Można tam, m.in., znaleźć odnośniki do Guidelines (oczywiście), projektów używających TEI (pod „Activities”), narzędzi („Tools”) – m.in. Roma i OxGarage, archiwów listy mailingowej i wiki (pod „Support”). Absolutnym klasykiem, używanym także poza TEI, jest tutorial „A gentle introduction to XML”.

Roma (rzut okiem)

http://www.tei-c.org/ (To się zawsze przyda.)

Wybierzmy „Tools” z menu. A spośród narzędzi, wybierzmy Romę.

Wybierzmy opcję „Build up”.

Możemy pobawić się ździebko pierwszym zbiorem danych – proszę np. wpisać siebie jako autora, wypełnić pole „Description”. Koniecznie proszę pamiętać o „Save” (czerwone, na dole).

Opcja „Save customization” u góry zachowuje plik ODD opisujący nasze zmiany (o ile je wprowadziliśmy). No to spróbujmy.

Roma z naszym własnym plikiem ODD

Wybierzmy opcję „browse” i załadujmy nasz dokument.

Możemy go poprawić i zapisać („save customization”), ale możemy wykorzystać Romę jedynie do stworzenia schematu, który pozwoli nam kontrolować poprawność nowych dokumentów XML.

Czyli: ładujemy nasz zmodyfikowany ręcznie ODD i prosimy Romę albo o wyderywowanie dokumentacji, albo o stworzenie schematu, którego potem użyjemy do sprawdzania poprawności naszych dokumentów XML.

OxGarage!

http://www.tei-c.org/oxgarage/

Możemy tu poszaleć.

(A jeśli są Państwo zbyt zmęczeni, to popatrzmy chociaż na opcje konwersji.)

Upiększenie ręczne: CSS

<?xml-stylesheet type="text/css" href="przykladowy_arkusz.css"?>

element { cecha-1: wartość-1;
cecha-2: wartość-2 }

dokument { display: block; }
naglowek {display: block; color: blue; }

akapit > zdanie (= wybierz <zdanie> wewnątrz elementu <akapit>)
naglowek[typ='glowny'] (= wybierz <naglowek typ=”glowny”>)

przydatne: font-weight: bold; font-size: 1em; font-style: italic

border:solid black 1px; display: inline

Przykłady

http://www.teibyexample.org/ (proszę spojrzeć w pola “Header”)

Przykładowy nagłówek z Narodowego Korpusu Języka Polskiego:

http://nlp.ipipan.waw.pl/TEI4NKJP/example_all_levels/header.xml

Wzór nagłówka słownika w projekcie FreeDict (lepiej go ściągnąć na dysk):

https://sourceforge.net/p/freedict/code/HEAD/tree/trunk/shared/lg1-lg2/lg1-lg2.tei

Dodatkowo wzór arkusza CSS z projektu FreeDict:

https://sourceforge.net/p/freedict/code/HEAD/tree/trunk/shared/freedict-dictionary.css