XML dla humanistów, i to cyfrowych... (pogadanka...)

Piotr Bański

Institut für Deutsche Sprache (IDS)
Mannheim

bansp@o2.pl

Warszawa, DELab UW, październik 2014

Czego potrzebujemy do pracy

Tekstu, głowy, tradycji badawczej, aparatu pojęciowego...

Co najmniej:

- słowników elektronicznych,
- archiwów tekstów, korpusów tekstów (zbiorów o bardziej konkretnym przeznaczeniu),
- (ontologii/terminologii – do bardziej zaawansowanych zastosowań),
- łatwości odnalezienia informacji/tekstów w Sieci [metadane!],
- sposobu na odpytanie słownika/archiwum (a więc języka zapytań),
- sposobu na „podwiązanie” dodatkowych danych o tekście do samego tekstu (a więc anotacji [uwaga: termin nie do końca zinstytucjonalizowany...]),
- sposobu na mówienie o tym samym (standardy, schematy)
- łatwości publikowania w Sieci (i nie tylko),
- oswojenia się z komputerem i komputerowym wspomaganiem badań.

Ćwiczenie:

- otwórz http://nkjp.pl/
- sprawdź... coś :-)
- Czy to pokazuje zalety XML-a? Nie, jedynie zalety zasobu językowego, który został stworzony przy zastosowaniu XML-a. Niezły, prawda?

Problemy typowe dla językoznawcy

- brak danych (czy aby na pewno?)
- brak dostępu do danych (licencje, przyzwyczajenia, nieuzasadniony “academic/non-commercial use”, itp.)
- brak chęci podzielenia się danymi (b. krótkowzroczne, mało etyczne)
- dane mają “złą” formę – trudno się do nich dobrać:
  - zamknięty format (Word! no ludzie...),
  - nietrwały format (Word 5.0, WordStar, itd. – kto to dziś otworzy...),
  - sztywny, niewygodny format (np. kolumnowe CSV, TSV; COCOA),
  - wymagający czasem (nie)szczególnych narzędzi,
  - format nieczytelny “gołym okiem” (binarny, np. relacyjne bazy danych).

Odpowiedź na dziś: XML

- nie, XML to nie jest panaceum, to jedynie narzędzie:
  - nieco wadliwe (problemy m.in. z nakładaniem się obszarów danych)
  - ale proste i 'giętkie' (
    - SGML – potwornie i zwykle niepotrzebnie skomplikowany,
    - HTML – zbyt szczegółowy, nastawiony na prezentację)
  - także... modne – a to istotne przy poszukiwaniu narzędzi, a także grup użytkowników (fora, listy dyskusyjne).
- „(bardzo?) stara szkoła” nie znosi XML, woli relacyjne bazy danych, ale
  - relacyjne bazy danych nie są dobrze dostosowane do opisu zasobów językowych (sztywne rekordy – por. zagnieżdżone sensy),
  - istnieją coraz to lepsze i szybsze XML-owe bazy danych (np. eXist, BaseX, etc.), które pozwalają 'odpytywać' XML językiem zbliżonym do SQL.

*ML w skrócie

SGML i XML to metajęzyki, służące do budowy bardziej wyspecjalizowanych (meta) języków znaczników. Np. HTML to aplikacja SGML na potrzeby WWW.

SGML
idea: lata 60-te
standard: lata 80-te

podzbiór →

XML
standard: 1998

→ XML 2.0? MicroXML? Zobaczymy, ale na razie jest stabilnie...

aplikacje

***ML x 100**
m.in. TEI P5

XHTML
„HTML (4.0) jako XML”

aplikacje

***ML x 10**
m.in. TEI P3

HTML
wiele wariantów, 'browser wars'; standaryzacja i wersja 4.x

→ **HTML5**
rozwinięcie HTML-a, ale nie SGML-a; połączenie języka opisu strony z językiem aplikacji. Istnieje też XHTML5.

Zalety XML...:

- uwypuklenie struktury danych,
- oddzielenie opisu danych od ich prezentacji,
- łatwość obróbki i związana z tym
- dostępność narzędzi.

… dla językoznawcy:

- to co powyżej, oraz
- możliwość ukrycia części danych lingwistycznych (np. opisania własności słów) w sposób niewpływający na przepływ tekstu,
- możliwość standaryzacji formatu opisu, a więc:
  - łatwość dzielenia się danymi,
  - łatwość powtarzania pomiarów (przy spełnieniu dodatkowych warunków).

[czas na demonstrację użyteczności XML-a, ale najpierw parę słów o nazwie]

XML (**eXtensible Markup Language**) to „rozszerzalny język znaczników”.

Czy używają Państwo na co dzień znaczników, znakowania przy pracy z tekstem?

Ależ tak – bez przerwy!

- odstępy między słowami, znaki interpunkcyjne, interlinie, zmiany kształtu liter (pogrubienie, kursywa) – to wszystko są rodzaje znakowania...

  to wszystko jest niedoskonałe (radzimy sobie czasem bez powyższych, ponieważ jesteśmy zdolnymi zwierzętami – ale jak biedna maszyna ma sobie z tym wszystkim poradzić?)

- konwencje: tylko ogólne (ale: akapit oddzielony linią czy wcięciem? wcięcie składa się ze spacji, znaku \t, czy wewnętrznego formatowania procesora tekstu?)

[czas na demonstrację użyteczności XML-a, ale najpierw parę słów o nazwie]

XML (**eXtensible Markup Language**) to „rozszerzalny język znaczników”.

Czy używają Państwo na co dzień znaczników, znakowania przy pracy z tekstem?

Ależ tak – bez przerwy!

- odstępy między słowami, znaki interpunkcyjne, interlinie, zmiany kształtu liter (pogrubienie, kursywa) – to wszystko są rodzaje znakowania...

  to wszystko jest niedoskonałe (radzimy sobie czasem bez powyższych, ponieważ jesteśmy zdolnymi zwierzętami – ale jak biedna maszyna ma sobie z tym wszystkim poradzić?)

- konwencje: tylko ogólne (ale: akapit oddzielony linią czy wcięciem? wcięcie składa się ze spacji, znaku \t, czy wewnętrznego formatowania procesora tekstu?)

[czas na demonstrację użyteczności XML-a, ale najpierw parę słów o nazwie]

XML (**eXtensible Markup Language**) to „rozszerzalny język znaczników”.

Czy używają Państwo na co dzień znaczników, znakowania przy pracy z tekstem?

Ależ tak – bez przerwy!

- odstępy między słowami, znaki interpunkcyjne, interlinie, zmiany kształtu liter (pogrubienie, kursywa) – to wszystko są rodzaje znakowania...

  to wszystko jest niedoskonałe (radzimy sobie czasem bez powyższych, ponieważ jesteśmy zdolnymi zwierzętami – ale jak biedna maszyna ma sobie z tym wszystkim poradzić?)

- konwencje: tylko ogólne (ale: akapit oddzielony linią czy wcięciem? wcięcie składa się ze spacji, znaku \t, czy wewnętrznego formatowania procesora tekstu?)

[czas na demonstrację użyteczności XML-a, ale najpierw parę słów o nazwie]

XML (**eXtensible Markup Language**) to „rozszerzalny język znaczników”.

Czy używają Państwo na co dzień znaczników, znakowania przy pracy z tekstem?

Ależ tak – bez przerwy!

- odstępy między słowami, znaki interpunkcyjne, interlinie, zmiany kształtu liter (pogrubienie, kursywa) – to wszystko są rodzaje znakowania...

  to wszystko jest niedoskonałe (radzimy sobie czasem bez powyższych, ponieważ jesteśmy zdolnymi zwierzętami – ale jak biedna maszyna ma sobie z tym wszystkim poradzić?)

- konwencje: tylko ogólne (ale: akapit oddzielony linią czy wcięciem? wcięcie składa się ze spacji, znaku \t, czy wewnętrznego formatowania procesora tekstu?)

(dalej...)

- Czy nowe terminy oznaczane są kursywą, czy wytłuszczeniem?
- A wtrącenia z innych języków?
- Czy nagłówki używają konkretnego stylu, czy robione są ręcznie (bold, small caps, etc.)?
- Czy daty albo nazwy własne są jakoś wyróżnione?
- Czy wiemy, która kropka kończy zdanie, a która należy do skrótu?

  (*np. Warszawa)*

A jeśli chcemy zmienić oznakowanie nagłówków? Albo rodzaj odstępów pomiędzy akapitami?

- w Wordzie też to mogę zrobić!
- no tak, takie rozdzielenie opisu i prezentacji to nie jest wyłączna zaleta XML-a, ale dodajmy to do wszystkiego innego, o czym tu mówiłem.

(dalej...)

- Czy nowe terminy oznaczane są kursywą, czy wytłuszczeniem?
- A wtrącenia z innych języków?
- Czy nagłówki używają konkretnego stylu, czy robione są ręcznie (bold, small caps, etc.)?
- Czy daty albo nazwy własne są jakoś wyróżnione?
- Czy wiemy, która kropka kończy zdanie, a która należy do skrótu?

  (*np. Warszawa)*

A jeśli chcemy zmienić oznakowanie nagłówków? Albo rodzaj odstępów pomiędzy akapitami?

- w Wordzie też to mogę zrobić!
- no tak, takie rozdzielenie opisu i prezentacji to nie jest wyłączna zaleta XML-a, ale dodajmy to do wszystkiego innego, o czym tu mówiłem.

Językoznawstwo...

A co jeśli chcemy wykonać jakieś pomiary językoznawcze?

Gdzie są granice zdań?

(to problem ogólny, ale czy wystarczy umieścić każde zdanie w osobnej linii? a co jeśli potem zajmiemy się słowami?)

Jak możemy oznakować własności słów? [atrybuty!]

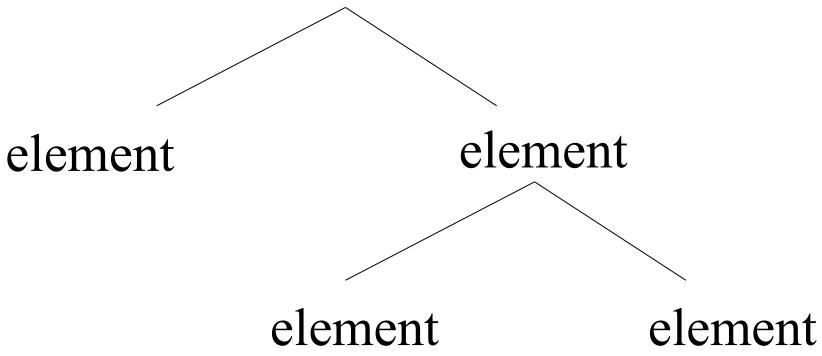
- Jak oznaczą Państwo własności gramatyczne słów w korpusie w Wordzie?
- A jeśli zechcą Państwo dołożyć lematy (= formy „słownikowe” wyrazów)?
- A informację o strukturach składniowych?
- Jak zaznaczyć, że niektóre spacje są nieistotne? (*po prostu, po ciemku*)
- A brak spacji? (*robiliście*)
- I jak się całość będzie czytało?

Rozszerzalny język znaczników: repertuar znaczników jest praktycznie nieograniczony. Możemy je dodatkowo opisywać atrybutami.

No to się teraz przestraszmy...

XML – podstawy składni

Struktura: drzewo, jeden “korzeń” (root)



Element: znacznik otwierający + zawartość + znacznik zamykający

```
<word>dziecko</word>
```

Atrybuty: wewnątrz znaczników otwierających:

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

XML – krótki tekst

Tekst: artykuł z Wikipedii o Powszechnej Deklaracji Praw Człowieka.

Wymyślmy jakieś znaczniki: czego potrzebujemy?

XML – krótki tekst

Tekst: artykuł z Wikipedii o Powszechnej Deklaracji Praw Człowieka.

Wymyślmy jakieś znaczniki: czego potrzebujemy?

<dokument>

<rozdzial> (umówmy się: bez diakrytyków)

<naglowek> (jak rozróżnić nagłówek rozdziału od nagłówka podrozdziału?)

<akapit>

<z>(danie) albo może <zdanie>? (tego w HTML-u nie ma...)

<w>yraz albo może <wyraz>?

<data>?

<nazwisko>?

… ?

A teraz: <?xml-stylesheet type="text/css" href="UDHR-Wikipedia.css"?>

XML – krótki tekst

Tekst: artykuł z Wikipedii o Powszechnej Deklaracji Praw Człowieka.

Wymyślmy jakieś znaczniki: czego potrzebujemy?

<dokument>

<rozdzial> (umówmy się: bez diakrytyków)

<naglowek> (jak rozróżnić nagłówek rozdziału od nagłówka podrozdziału?)

<akapit>

<z>(danie) albo może <zdanie>? (tego w HTML-u nie ma...)

<w>yraz albo może <wyraz>?

<data>?

<nazwisko>?

… ?

A teraz: <?xml-stylesheet type="text/css" href="UDHR-Wikipedia.css"?>

CSS: Kaskadowy arkusz stylów

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

(Możemy wyszukać „CSS2”, pojawi się specyfikacja W3C: http://www.w3.org/TR/CSS2/)

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

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

font-weight: bold; font-size: 1em
font-style: italic
border:solid black 1px; display: inline

Rozszerzalność, schematy

Repertuar znaczników jest nieograniczony,

- i bardzo dobrze, bo dzięki temu możemy opisać, co tylko nam się zachce,
- ale jak sprawić, abyśmy mogli swobodnie zamienić się tekstami i wciąż mieć pewność, że rozumiemy, co poprzednia osoba wykonała, co zaznaczyła?
- Pochodny problem: jak sprawić, żeby mój arkusz CSS działał ze znakowaniem kolegi?

Schematy i walidacja (rozróżnienie well-formedness od validity): proste DTD.

(Są też bardziej zaawansowane języki schematów: XML Schema i RelaxNG.)

Well-formedness – „poprawność podstawowa”: jeden korzeń, domknięte elementy, wartości atrybutów w cudzysłowie, poprawne zagnieżdżenie.

Validity – poprawność w odniesieniu do konkretnego schematu.

xmllint --noout yourfile.xml vs. xmllint --noout --valid yourfile.xml

Standardy

Mamy dwa różne znakowania i dwa różne DTD – co dalej?

→ Standardy (jak wszędzie; w tym kontekście znana jest organizacja ISO, i prowadzi ona także standaryzację opisu zasobów językowych – ale nie będziemy się dziś w to zagłębiać)

Standard dla humanistów: **TEI** (Text Encoding Initiative)

Czasem pełna standaryzacja nie jest potrzebna – czasem wystarcza **regularny** format, a więc **dopracowana struktura danych**. Z regularnej struktury da się wiele uzyskać.

Roma – narzędzie do tworzenia schematów

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

Wyderywujmy jakiś ogólny schemat...

(nie trzeba znać się na pisaniu DTD lub innych językach schematów, aby ich używać...)

oXygen automatycznie rozpoznaje schemat TEI. Robi to dzięki przestrzeni nazw TEI:

```
<TEI xmlns="http://www.tei-c.org/ns/1.0">
```

Przestrzenie nazw pozwalają na dalszą rozszerzalność zasobu znaczników. Ale to dłuższa historia i nie będziemy jej dziś zapewne poznawać :-)

Żeby nie było zbyt różowo...

Wady XML:

- aż nazbyt hierarchiczny (jeden korzeń drzewa, nakładanie się opisywanych obszarów)
- pewne błędy w dodatkowych standardach (np. problemy z Namespaces, XML Schema), niektóre standardy niemal nieużywane (XLink, XPointer)
- wielkość plików
  - ojej, mamy coraz większe dyski i RAM
  - ale i tak jest to istotny problem (wiem z doświadczenia)
- szybkość przetwarzania względem innych formatów danych
  - ale to się zmienia:
    - optymalizacja natywnych baz danych
    - streaming XSLT
  - poza tym format przechowywania danych nie musi być formatem serwowania danych (XML łatwo skompilować do zadanej postaci).

Tworzenie zasobów językowych

Standaryzacja – daje nam m.in.:

- możliwość użycia danych w wielu narzędziach oraz łączenia danych (np. łączenia kawałków słownika w jeden)
- ułatwia przechowywanie danych, nie pozwala im się zestarzeć (przypomnijmy WordStar...)

Ale standaryzacja nie dotyczy jedynie formatu samych danych.

Ważne:

- metadane opisowe (typ tekstu, dane twórcy, czas powstania, język/dialekt, itp.)
- treść metadanych opisowych i analitycznych:
  - czy kategoria „proz.hist” to to samo co „UDC 820+900”?
  - czy kategorie „A”, „adj”, „przym.” albo „f”, „fem”, „żeń” mają może coś wspólnego?

Metadane, nagłówki, ontologie

Metadane opisowe zawarte są zwykle w nagłówkach dokumentów. Nagłówek TEI (TEI header) jest narzędziem stosowanym nawet w systemach, które nie używają TEI do znakowania tekstów.

Do uzgadniania zawartości kategorii analitycznych stosuje się tzw. ontologie językoznawcze (wyszukajmy „GOLD ontology”) bądź repozytoria kategorii (wyszukajmy „ISO DCR” lub „ISOCat”).

Jak ontologie takie jak GOLD mają się do XML-a?

Do czystego XML-a niemal nijak. Ale do XML-a używanego przez językoznawców – jak najbardziej. Ponieważ zależy nam m.in. na tym, aby inni mogli porównać swoje wyniki z naszymi.

Dodatkowe materiały

„Learn the TEI”: http://www.tei-c.org/Support/Learn/

(szczególną uwagę warto zwrócić na „Gentle Introduction...”)

Intensive introduction to the TEI:
http://dev.stg.brown.edu/staff/Julia_Flanders/tei/brown2006/

Kurs technologii XML Tomaža Erjaveca:

http://nl.ijs.si/et/teach/esslli05/

Moje slajdy ze szkoły letniej TEI@Oxford-2010:

http://tei.oucs.ox.ac.uk/Oxford/2010-07-oxford/materials/workshops/languageResources/languageResources.pdf

TEI Wiki: http://wiki.tei-c.org/