XML i nowoczesne technologie zarządzania treścią

Wykład monograficzny
Semestr zimowy 2006/07

Szymon Zioło
sziolo@mimuw.edu.pl

Sprawy organizacyjne

- Strona internetowa wykładu:
  - http://www.mimuw.edu.pl/~sziolo
- Terminy zajęć:
  - wykłady: czwartki, 16:15 – 17:45,
  - pracownia:
    - Radek Bartosiak – czwartki, godz. 18:15 – 20:00,
    - Mikołaj Rybiński – czwartki, godz. 18:15 – 20:00,
    - Patryk Czarnik – piątki, godz. 8:30 - 10:00,
    - Patryk Czarnik – piątki, godz. 10:15 - 12:00.
- Konsultacje: w siedzibie firmy ABG Ster-Projekt po indywidualnym umówieniu.
- Kryteria zaliczenia:
  - pracownia: punktowane zadania zaliczeniowe,
  - wykład: egzamin pisemny.

Plan wykładu

- Wprowadzenie: historia rozwoju technik znakowania tekstu.
- Technologia i standardy:
  - podstawowe koncepcje XML-a,
  - definiowanie typów dokumentów, modelowanie informacji,
  - przekształcenia XSLT,
  - standardy związane z XML-em,
  - wykorzystanie XML-a we własnych aplikacjach.
- Zastosowania biznesowe:
  - technologie modelowania, kodowania i wymiany wiedzy,
  - kategoryzacja, klasyfikacja i wyszukiwanie informacji,
  - XML a bazy danych,
  - systemy i technologie zarządzania treścią i publikowania treści,
  - XML w elektronicznej wymianie danych i integracji aplikacji.

Zawartość zajęć

- Dodatkowo na wykładzie:
  - *case studies*:
    - struktura formularzy ubezpieczeniowych KEDU ZUS,
    - generatory przekształceń XSLT w systemie empolis Impera,
    - neutralna pula zasobów w wydawnictwie Planeta Actimedia.
  - pokazy oprogramowania.
- Pracownia:
  - modelowanie informacji,
  - korzystanie z parserów XML-a w Javie,
  - tworzenie przekształceń XSLT,
  - korzystanie z innych standardów związanych z XML-em.

O mnie

- Absolwent MIMUW-u.
- Główny analityk w firmie ABG Ster-Projekt:
  - kierowanie pracami analitycznymi w projektach,
  - kursy, szkolenia,
  - merytoryczne wsparcie sprzedaży,
  - zainteresowania: systemy obiegu dokumentów i zarządzania treścią.
- Inicjator powstania grupy newsowej pl.comp.xml.
- Redaktor prowadzący wydań 6'2001, 6'2003 i 6'2004 czasopisma Software 2.0 poświęconych XML-owi.
- Autor kursów komercyjnych „Podstawy XML-a" i „Modelowanie informacji w XML-u".

Moje projekty (wybór)

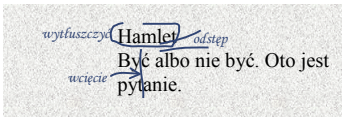
- Planeta Actimedia, Barcelona (wydawnictwo encyklopedyczne) – system zarządzania treścią oparty na koncepcji neutralnej puli zasobów.
- Prokom Software, Gdynia – projekt struktury formularzy ubezpieczeniowych ZUS i technologii ich wizualizacji.
- Schlumberger Oilfield Services, Paryż – system zarządzania dokumentacją techniczną.
- Wolters Kluwer, Mechelen, Belgia – projekt mechanizmu Publication Build dla grupy wydawnictw.
- Polska Telefonia Cyfrowa, Warszawa – system obiegu dokumentów strukturalnych Document Collection Office.
- Ubezpieczeniowy Fundusz Gwarancyjny, Warszawa – system obiegu dokumentów i spraw związanych z działalnością UFG.

Historia rozwoju technik znakowania tekstu

Markup – znakowanie

- Markup Languages (eMeLe):
  - SGML – Standard Generalized Markup Language,
  - HTML – Hypertext Markup Language,
  - XML – Extensible Markup Language.
- Markup:
  the process of marking manuscript copy for typesetting with directions for use of type fonts and sizes, spacing, indentation, etc.
  (The Chicago Manual Of Style).

Prehistoria: znakowanie tekstu



Znakowanie tekstu w epoce komputerów

Treść

Hamlet Być albo nie być. Oto jest pytanie

+

Formatowanie, adjustacja

{nowy_wiersz} {bold} {wyłącz_bold} {wcięcie}

=

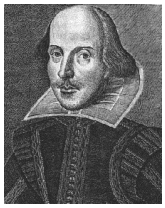
Dokument

Hamlet

Być albo nie być. Oto jest pytanie.

Przykłady języków znakowania

- **Frame (MIF)** <Font <FTag `B'>><br><String `Hamlet'>
- **QuarkXPress** <B>Hamlet<B>
- **RTF** {\b\f5\cf1 Hamlet}
- **Ventura** <B>Hamlet<D>
- **TeX/LaTeX** \textbf{Hamlet}
- **PostScript** /Times-BoldR 900 ff
(Hamlet)W
- **HTML** <B>Hamlet</B>



Korzenie

- Lata 60-te XX wieku:
  - 1967 – William Tunnicliffe, prezes Graphic Communications Association, podczas spotkania w Canadian Government Printing Office przedstawia ideę oddzielenia zawartości informacyjnej dokumentów od ich formatu,
  - Stanley Rice proponuje użycie uniwersalnych znaczników do znakowania struktury tekstu,
  - projekt GenCode definiuje sposób oznaczania tekstu ukierunkowany na jego strukturę.

Korzenie: INTIME

- INTIME – INteractive Textual Information Management Experiment:
  - projekt badawczy Charlesa Goldfarba (IBM Cambridge Scientific Center, koniec lat 60-tych XX wieku),
  - prototyp zintegrowanego systemu przetwarzania tekstu:
    - edycja tekstu,
    - repozytorium dokumentów,
    - wyszukiwanie;
  - wykorzystane technologie:
    - „maszyny wirtualne" na mainframie IBM 360,
    - *concurrent access to a disk file*,
    - *context editors*.

Edytor kontekstowy

```
LOCATE /researchers/
researchers. A system which integrates
CHANGE /researchers/analysts/
analysts. A system which integrates
CHANGE /edit/edit/ *
In online systems, text editing is
are known as "context" editors. They
NEXT
provide a retrieval capability: e.g.,
QUIT
```


Wnioski z projektu INTIME

The usefulness of a retrieval program can be affected by its ability to identify the structure and purpose of the parts of text (e.g., footnotes, abstracts, citations). [...] *A heuristic routine for identifying new paragraphs in normal text was developed for INTIME, but a more sophisticated facility is needed. A typesetting command language could convey such information, but present languages deal with the appearance of the text, not with the purpose which motivated it.*

C. Goldfarb, „SGML: The Reason Why and the First Published Hint", Journal of the American Society for Information Science, Volume 48, Number 7 (July 1997)

Programy i ich formaty

- Prawie każda aplikacja wprowadza swój wewnętrzny format.
- Nowe wersje tej samej aplikacji wprowadzają zmiany do używanego formatu:
  - wsteczna kompatybilność,
  - brak możliwości zapisu do formatu poprzednich wersji.
- Aplikacje dostarczają konwerterów:
  - tylko do najpopularniejszych formatów,
  - możliwość utraty danych podczas konwersji.

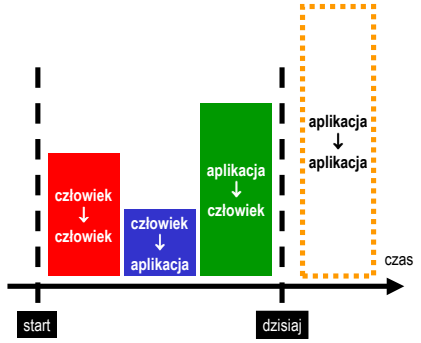
Standardy

- Nie istnieją uznane standardy.
- Istnieją substandardy w różnych dziedzinach:
  - dokumenty biurowe: Microsoft Word,
  - teksty naukowe: Postscript, TeX,
  - Internet: HTML, GIF, JPG,
  - elektroniczna wymiana danych: EDIFACT.
- Standard musi być:
  - własnością publiczną,
  - otwarty i jawny,
  - niezależny od konkretnego producenta oprogramowania.

Potrzeba struktury

- Masa informacji cyfrowej powoduje potrzebę struktury:
  - jeden format dokumentu nie wystarczy dla 5 miliardów ludzi,
  - ale nie możemy operować milionami niekompatybilnych formatów.

Ewolucja Internetu



Rozwój języków uogólnionego znakowania tekstu

- 1969: GML – Generalized Markup Language (IBM; Goldfarb, Mosher, Laurie).
- 1986: SGML – Standard Generalized Markup Language, ISO 8879:1986.
- 1991: powstaje World Wide Web.
- 1994: HTML 2.0 zdefiniowany jako zastosowanie SGML-a.
- 1998: XML – Extensible Markup Language, World Wide Web Consortium.

Wokół SGML-a

- Pierwsze szerzej znane zastosowania SGML-a:
  - Electronic Manuscript Project, Association of American Publishers, 1987,
  - CALS – Computer-Aided Acquisition and Logistic Support, US Department of Defense, MIL-M-28001, February 1988.
- Standardy pokrewne:
  - DSSSL – Document Style Semantics and Specification Language,
  - HyTime:
    - meta-notacja dla linków,
    - opis struktur multimedialnych, rozciągniętych w czasie.

World Wide Web Consortium (W3C)

- Kuźnia standardów internetowych, np.:
  - HTML – Hyper Text Markup Language,
  - HTTP – Hyper Text Transfer Protocol,
  - CSS – Cascading StyleSheets,
  - ...
- XML – Extensible Markup Language:
  - najważniejsza rekomendacja ostatnich lat,
  - twórcy: Tim Bray (Netscape), Jean Paoli (Microsoft), C.M. Sperberg-McQueen (University of Illinois).
- Obecne dominują prace nad standardami związanymi z XML-em.

W3C WORLD WIDE WEB consortium

Idea SGML/XML (1)

Oddzielenie znaczenia tekstu od sposobu prezentacji

<OSOBA MÓWIĄCA>Hamlet</OSOBA MÓWIĄCA>
<WYPOWIEDŹ>Być albo nie być.
Oto jest pytanie.</WYPOWIEDŹ>

Sposób prezentacji

- OSOBA MÓWIĄCA
  - nowy akapit
  - do lewej
  - wytłuszczenie
- WYPOWIEDŹ
  - nowy akapit
  - wcięcie na 2 cm
  - do lewej

Hamlet
Być albo nie być. Oto jest pytanie.

Inny sposób prezentacji

- OSOBA MÓWIĄCA
  - na marginesie
  - tekst pionowo
  - niebieski
  - hiperlink do opisu postaci na początku dramatu
- WYPOWIEDŹ
  - nowy akapit
  - kursywa
  - ew. użyj syntezatora mowy z ustawieniami dla OSOBY MÓWIĄCEJ

Hamlet *Być albo nie być. Oto jest pytanie.*

Idea SGML/XML (2)

Stworzenie najodpowiedniejszego modelu dla naszych własnych dokumentów.

<OSOBA MÓWIĄCA>Hamlet</OSOBA MÓWIĄCA>
<WYPOWIEDŹ> <NUDA> Być albo nie być.
Oto jest pytanie.</NUDA> </WYPOWIEDŹ>

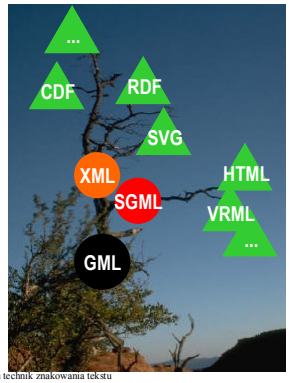
Najodpowiedniejszy model

- Przykłady:
  - encyklopedia: **<nazwisko>, <imie>, <ur>, <zm>, <wymowa>, <etymologia>, <liczba-mieszk>**
  - prawo: **<promulgator>, <rocznik>, <poz>, <art> <sąd>, <sygn-wyroku>, <teza>**
  - dokument techniczny: **<part-number>, <function-name>**
  - patenty: **<wynalazca>, <nr-zgloszenia>**
  - ubezpieczenia: **<data-polisy>, <wartość-polisy>**

Język – metajęzyk

- Stan wyjściowy:
  - Wieża Babel (brak wspólnego języka),
  - czy w ogóle możliwy jeden wspólny język?
- Wspólny metajęzyk:
  - znana gramatyka,
  - jednolita metodologia,
  - takie same narzędzia.
- Dowolnie wiele języków specyficznych dla zastosowań.

Genealogia XML-a



Co to jest XML?

- XML to nie język programowania.
- XML to sposób zapamiętywania danych wraz z ich strukturą w dokumencie tekstowym:
  - otwarty,
  - elastyczny,
  - bezpłatny,
  - niezależny od platformy sprzętowej.
- XML to rama składniowa do tworzenia języków specyficznych dla zastosowań.
- Użycie XML-a nie zwalnia od myślenia (analizy, projektowania, ...)

Jak wygląda XML?

```
<?xml version="1.0"?>
<zeznanie-sprawcy nr="1313/2001">
<autor>st. asp. Jan Łapówka</autor>
<miejsce>Dołowice Górne</miejsce>
<treść>Wypadek dnia
<data>13.10.2001r</data>
o godzinie <godzina>13:13</godzina>
(<dzien-tygodnia>piątek
</dzien-tygodnia>) miał  miejsce nie
z mojej winy. <poszkodowany>Alojzy
M.</poszkodowany> nie miał żadnego
pomysłu w którą stronę uciekać, więc
go przejechałem.</treść>
</zeznanie-sprawcy>
```

Deklaracja XML
Element główny
Atrybut
Element
Znacznik początkowy
Znacznik końcowy
Zawartość tekstowa

HTML ↔ XML

- Znaczenie elementów i ich atrybutów z góry określone.
- Interpretację elementów określa standard, a w praktyce przeglądarki internetowe.
- To, co jest poprawne również określają przeglądarki internetowe.
- Znaczenie elementów i ich atrybutów określa użytkownik lub aplikacja.
- <p> może w jednym dokumencie oznaczać paragraf, w drugim pomoc, a w trzecim pismo odręczne.
- Poprawność XML-a jest ściśle określona przez specyfikację.

SGML ↔ XML

- Filozofia: jeden duży system zarządzania treścią.
- Konieczność definiowania struktury.
- Skomplikowana składnia, wiele opcji.
- Trudność tworzenia parserów.
- Bardzo drogie narzędzia.
- Filozofia: wiele małych komunikujących się ze sobą modułów.
- Opcjonalne definiowanie struktury.
- Uproszczona składnia.
- Łatwość tworzenia parserów.
- Darmowe narzędzia.

Klasy zastosowań XML-a

Zarządzanie dokumentami, treścią, wiedzą:

- Pierwotne zastosowanie SGML-a.
- Dokumenty tworzone przez człowieka i przeznaczone dla człowieka.
- Długi czas życia dokumentów.
- Typowy model mieszany zawartości.

Elektroniczna wymiana danych, integracja aplikacji:

- Nowa klasa zastosowań XML-a.
- Dokumenty tworzone oraz przetwarzane automatycznie
- Dokumenty tworzone tylko na czas komunikacji.
- Konieczność dokładnego kontrolowania struktury i zawartości.

Dwie twarze XML-a

Dokument tekstowy:

```
<zeznanie-sprawcy>
Wypadek dnia <data>
13.01.2001 r.</data>
o godzinie <godzina>13.13
</godzina> (<dzien-
tygodnia>piątek
</dzien-tygodnia>) miał
miejsce nie z mojej winy.
<poszkodowany>Alojzy
M.</poszkodowany> nie miał
żadnego pomysłu w którą
stronę uciekać, więc go
przejechałem.
</zeznanie-sprawcy>
```

Baza danych:

```
<zamowienie>
  <pozycja>
    <nazwa>Papier</nazwa>
    <jednostka>ryza
    </jednostka>
    <ilosc>3</ilosc>
  </pozycja>
  <zamawiajacy id="123456">
    <imie>Szymon</imie>
    <nazwisko>Zioło
    </nazwisko>
    <firma>ABG Ster-Projekt
    </firma>
  </zamawiajacy>
</zamowienie>
```


Gdzie szukać dalej: historia XML-a

- Charles F. Goldfarb's SGML Source Home Page:
  www.sgmlsource.com
- Wypych, W., *Na początku był rękopis, czyli o historii XML-a*:
  Software 2.0, 6/2001

Gdzie szukać dalej

- W3C – The World Wide Web Consortium:
  - www.w3.org
- XML.com:
  - www.xml.com
- The XML Industry Portal, hosted by OASIS:
  - www.xml.org
- The XML Cover Pages:
  - www.oasis-open.org/cover/xml.html
- Paweł Stroiński, Kurs języka XML:
  - www.wckp.lodz.pl/~pabloware/xml
- Grupa dyskusyjna pl.comp.xml

Gdzie szukać dalej

- Megginson, D., „Structuring XML Documents”, Prentice Hall PTR, 1998
- Goldfarb, C., Prescod, P., „The XML Handbook, 5th Edition”, Prentice Hall PTR, 2003
- Marchal, B., „XML w przykładach”, Wydawnictwo Mikom, 2000
- Young, M. J., „XML krok po kroku”, Wydawnictwo RM, 2000
- McLaughlin, B., „Java i XML”, Wydawnictwo Helion, 2001
- Elliotte, R. H., Scott, W., „XML. Almanach”, Wydawnictwo Helion, 2002
- Arciniegas, F., „XML. Kompendium programisty”, Wydawnictwo Helion, 2002
- Ray, E. T., „XML. Wprowadzenie (wydanie 2)”, Wydawnictwo Helion, 2004
- ...