Przetwarzanie dokumentów XML i zaawansowane techniki WWW

"Wprowadzenie do walidacji dokumentów XML przez DTD (Document Type Definition)"

(Zajęcia 03 - 14.03.2016 r.)

Poza umiejętnością przygotowania optymalnej struktury plików XML, ważne jest aby poznać techniki walidacji tych plików, tzn. czy dany plik XML jest zgodny z założonym lub określonym wzorcem lub spełnia określone warunki.

1) Walidacja przez DTD (Document Type Definition)

Jedną z metod jest tzw. walidacja przez DTD - Document Type Definition (Definicja Typu Dokumentu) dzięki której można określić strukturę pliku XML oraz jakie elementy w danym pliku XML mogą wystąpić. W definicji DTD występują deklaracje elementów oraz adrybutów tych elementów wraz z zdefiniowaniem hierarchi wystąpień oraz ich ilości. Deklaracja elementu wygląda następująco:

```
<!ELEMENT nazwa ( element_zagniezdzony_1, ... , element_zagniezdzony_n) >
```

Dodatkowo dla każdego elementu zagnieżdzonego można określić ilość wystąpień tych elementów w następujący sposób:

- ***** -- **występuje zero lub wiecej razy,**
- **+** -- **występuje jeden lub wiecej razy (jedno wystąpienie obowiązkowe),**
- **?** -- **występuje zero lub jeden raz.**

przykład:

```
<ELEMENT biblioteka (ksiazka*) >
```

w elemencie **biblioteka** może wystąpić dowolną liczbę razu (od 0 - n) element zagnieżdzony **ksiazka**. Idąc dalej możemy zdefiniować jakie elementy zagnieżdzone wystąpią w elemencie **ksiazka**:

```
<!ELEMENT ksiazka (autor+, tytul, wydawca, rok) >
```

w tym przypadku element **autor** musi wsytapic 1-raz lub więcej.

Stwórzmy teraz definicje dla tych elementów zagnieżdzonych:

```
<!ELEMENT autor (#PCDATA) >
<!ELEMENT tytul (#PCDATA) >
<!ELEMENT wydawca (#PCDATA) >
<!ELEMENT rok (#PCDATA) >
```

w tym przypadku definiujemy już tylko elementy tekstowe nieposiadające dalszych zagnieżdzeń.
W przypadku definicji typu dokumentu nie możemy określić jakiego rodzaju dane mogą wystapić w danym elemencie. Jedyna co można podać to że będą to dane typu "**PCDATA**" czyli dane komputerowe.

Poza określeniem typów elementów można zcharakteryzować również występowanie atrybutów przez deklarację:

```
<!ATTLIST nazwa_elementu atrybut_1 type atrybut_2 type …. atybut_n type >
```

Zmodyfikujmy zatem strukturę lementu książka tak aby zawierała atrybuty:

```
<ksiazka id="A01" regal="F" >
```

zatem deklaracja atryputów powinna wyglądać następująco:

```
<!ATTLIST ksiazka
                id          ID #REQUIRED
                regal    (A|F|K|R|X|Z) "X" >
```

gdzie słowo kluczowe `ID` oznacza niepowtarzalność wartości atrybutu natomiast `#REQUIRED` wymaga jego wsystąpienia. Natomiast określenie w nazwiasie okrągłym `(A|F|K|R|X|Z) "X"` oznacza możliwe do wsytąpienie wartości atrybutu `regal` lub gdy nie jest podany to domyslnie uzyty zostanie atrybut `"X"`.

Definicję typu dokumentu możemy określić bezpośrednio za prologiem lub w osobnym pliku. W pierwszym przypadku należy bezpośrednio za prologiem dopisać:

```
<!DOCTYPE  nazwa_elementu _glownego [
    //// deklaracja DTD
]>
```

lub w przypadku deklaracji w osobnym pliku:

```
<!DOCTYPE nazwa_elementu _glownego SYSTEM "plik.dtd" >
```

gdzie `"plik.dtd`", zawiera deklaracje DTD. W tym przypadku do prolog powinien zawierać atrybut określający istnienie dodatkowego pliku:

```
<?xml version="1.0" encoding="utf-8" standalone="no" ?>
```

Walidacji pliku xml można dokonać wykorzystując parsery. Natomiast do testowania można użyć polecenia w systemie Linux:

```
> xmllint --noout --valid file.xml
```

Kompletny przykład:

Plik: biblioteka.xml

```
<?xml version="1.0" encoding="utf-8" standalone="no" ?>
<!DOCTYPE nazwa_elementu _glownego SYSTEM "biblioteka.dtd" >
<biblioteka>
     <ksiazka>
          <autor>D. Perkins</autor>
          <tytul>Wstep do fizyki wysokich energii</tytul>
          <wydawca>PWN</wydawca>
          <rok>1973</rok>
     </ksiazka>
</biblioteka>
```

Plik: biblioteka.dtd

```
<ELEMENT biblioteka (ksiazka*) >
<!ELEMENT ksiazka (autor+, tytul, wydawca, rok) >
<!ELEMENT autor (#PCDATA) >
<!ELEMENT tytul (#PCDATA) >
<!ELEMENT wydawca (#PCDATA) >
<!ELEMENT rok (#PCDATA) >
```

Zadanie 1:

Proszę dla przygotowanego na poprzednich zajęciach pliku XML utworzyć plik DTD z definicjalmi typu elementów.