Przetwarzanie dokumentów XML i zaawansowane techniki WWW

"Wprowadzenie do języka ścieżek XPath"

(Zajęcia 05 - 04.04.2016 r.)

1. Wprowadzenia do języka scieżek XPath

Wyrażenia XPath pozwalają na poruszanie się po strukturze dokumentów XML i jednocześnie pozwalają na wyłuskiwanie z nich danych szukanych. Do testowania wyrażeń XPath służy znane narzędzie **`xmllint`**.

Poruszanie się po pliku XML i wyłuskiwanie z niego żądanych danych polega na określeniu "ścieżki dostępu" podobnie jak to jest w strukturze kartotek w systemie Linux. W ogólności każdy plik XML składa się z węzłów (node). Aby wybrać ze struktury dokumentu XML określone dane musimy na nie wskazać przez wybranie odpowiedniego węzła. Do wybieranie węzłów służą:

`/` - wybranie pasujących danych węzła głównego
`//` - wybranie pasujących danych z dowolnego węzła w dokumencie XML
`.` - wskazanie na bieżacy węzeł w którym się znajdujemy
`..` - wybranie węzła rodzica (nadrzędnego).
`@` - wybranie atrybutów

Przykładowy plik XML na którym możemy testować wyrażenia znajduje się na stronie i ma postać:

```
<?xml version="1.0" encoding="UTF-8"?>
<book lang="pl">
  <title>Filozofia kosmologii</title>
  <bookinfo>
     <author>
     <firstname>Michał</firstname>
     <surname>Heller</surname>
     </author>
  </bookinfo>
  <chapter id="A01">
     <title>Kosmologia przez Einsteinem</title>
     <abstract>
     <para>Bardzo dawno temu...</para>
     </abstract>
     <section1>
     <title lang="en">Paradoks Olbersa</title>
     <para>Jeszcze daniej ….</para>
     </section1>
     <!-- A small comment -->
     <section2 id="B02">
     <title lang="pl">Paradoks Seeligera</title>
     <para>Tutaj Seelinger sie narodzil….</para>
     </section2>
  </chapter>
</book>
```

Narzędzia xmllint do testowania siceżek można używać na dwa sposoby:

1) przez bezpośrednie wywoływanie ścieżek i wyrażań z lini komend:

```
xmllint --xpath '//title' book.xml
```

wynikiem powinno być:

```
<title>Filozofia kosmologii</title><title>Kosmologia przez
Einsteinem</title><title lang="en">Paradoks Olbersa</title><title
lang="pl">Paradoks Seeligera</title>
```

2) lub przez wywołanie wbudowanego interpretera XPatha i testowanie ścieżek:

```
xmllint --shell book.xml
```

polecenie udostępnie konsole która pozwala na interpretację poleceń XPatha:

```
> cat //title
```

wynik:

```
 -------
<title>Filozofia kosmologii</title>
 -------
<title>Kosmologia przez Einsteinem</title>
 -------
<title lang="en">Paradoks Olbersa</title>
 -------
<title lang="pl">Paradoks Seeligera</title>
```

W obu przypadkach zwrócone dane są takie same.

1) dalsze przykłady w trybie shella:

a) polecenie `cat` - zwraca wyniki wyłuskania danych z żądanych węzłów jak było to pokazane poprzednio.

b) `grep` : pozwala na określenie ścieżki do danego elementu / wyszukanie ścieżki w struktrurze dokumentu:

```
> grep Paradoks
/book/chapter/section1[1]/title : t--       16 Paradoks Olbersa
/book/chapter/section1[2]/title : t--       18 Paradoks Seeligera
```

c) `xpath` : pozwala na zwrócenie informacji o testowanych węzłach:

```
> xpath /book/chapter
Object is a Node Set :
Set contains 1 nodes:
1  ELEMENT chapter
     ATTRIBUTE id
     TEXT
     content=A01
```

dostajemy informację o liczbie wystąpień danego elementu oraz czy zawiera atrybuty:

```
> xpath /book/chapter/section1
Object is a Node Set :
Set contains 2 nodes:
1  ELEMENT section1
2  ELEMENT section1
     ATTRIBUTE id
     TEXT
     content=B02

> xpath //@id
Object is a Node Set :
Set contains 2 nodes:
1  ATTRIBUTE id
     TEXT
     content=A01
2  ATTRIBUTE id
     TEXT
     content=B02
```

d) funkcja `string()` : zwracająca wartość łańcuch dla danej ścieżki:

```
> xpath string(/book/chapter/section1[1]/title)
  Object is a string : Paradoks Olbersa
```

warto zwrócić uwagę, że wybieranie konkretnego elementu wśród wielu tego samego typu jest realizowane przez podanie numeru w nawiasach kwadratowych podobnie jak odwołanie do numeru pozycje w tablicy.

d) funkcja `last()` zwracający ostatni element danego typu według ścieżki:

```
> xpath string(/book/chapter/section1[last()]/title)
Object is a string : Paradoks Seeligera
```

lub

```
> xpath string(/book/chapter/section1[last()-1]/title)
Object is a string : Paradoks Olbersa
```

e) wyszukiwanie po wartości atrybutu:

```
 > xpath string(//title[@lang='en'])
Object is a string : Paradoks Olbersa

 > xpath string(//title[@lang='pl'])
Object is a string : Paradoks Seeligera
```

f) funkcja `text()` : zwraca wartość węzła tekstowego jeśli istnieje

```
> xpath //title[@lang='pl']/text()
```

```
Object is a Node Set :
Set contains 1 nodes:
1  TEXT
      content=Paradoks Seeligera
```

ZADANIE:

Proszę przygotowac plik XML zawierający bazę danych o mieszkaniach na sprzedaż. W pojedyńczym rekordzie (o nazwie np. <mieszkanie>) powinien znaleźć się atrybut określający identyfikator mieszkania. W elementach potomnych należy umieścić następujące informacje:

- *lokalizacja mieszkania (nazwa miejscowości),*
- *powierzchnia,*
- *liczba pokoi,*
- *rok oddania do użytkowania,*
- *cena.*

W pliku XML powinny znaleźć się przynajmniej 4 pojedyńcze rekordy. Następnie proszę przygotować ścieżki XPath które:

- wyciągną wszystkie informacje o ostatnim mieszkaniu,

- cenę przedostatniego mieszkania

- pozwolą na zwrócenie danych o drugim mieszkaniu na podstawie podania identyfikatora.