

XML i nowoczesne technologie

zarządzania treścią 2009/10 – Zadanie 3 Poprawkowe

Należy napisać program w Javie (lub po ustaleniu ze sprawdzającym w innym języku programowania) czytający dokument tekstowy Open Document i zapisujący w dokumencie XHTML informacje o zawartości dokumentu tekstowego zgodnie z poniższym opisem.

Parametry wywołania

Program będzie wywoływany z następującymi parametrami:

- ścieżka do wejściowego pliku ODT,
- ścieżka do wynikowego pliku XHTML.

Pliki i działanie

Plik wejściowy plik w formacie Open Document Text (.odt), do tworzenia danych testowych będzie używany Open Office 3.1.

Program powinien czytać plik wejściowy nie modyfikując go. Pliki .odt są archiwami ZIP i należy „strumieniowo” dekompresować plik źródłowy, a wewnętrznym plikiem, którego zawartość ma zostać odczytana, jest `content.xml`. Parsowanie dokumentu XML ma także odbywać się strumieniowo, zgodnie z SAX lub StAX. Tworzenie dokumentu wynikowego może odbywać się za pomocą dowolnego API spośród omawianych na zajęciach (DOM, JAXB, StAX, SAX).

Zarówno podczas odczytu jak i zapisu należy poprawnie (w sposób właściwy dla danego API) obsługiwać przestrzeń nazw. W szczególności nie można zakładać, że prefiksy i/lub części lokalne nazw kwalifikowanych są wystarczające do ich identyfikacji.

Statystyki

Dokument wynikowy ma być dokumentem XHTML zawierającym listę numerowaną, której elementy opisują poszczególne rozdziały dokumentu. Lista może zawierać zagnieżdżone listy itd., odzwierciedlając strukturę rozdziałów w dokumencie.

Opis rozdziału na każdym poziomie powinien zawierać:

1. tytuł rozdziału odczytany z nagłówka,
2. liczbę akapitów w tym rozdziale (nie uwzględniając rozdziałów niższego poziomu),
3. rozdziały niższego poziomu zawarte wewnątrz tego rozdziału jako analogicznie ("rekurencyjnie") utworzone listy zagnieżdżone.

Za rozdział poziomu N uznajemy fragment dokumentu od nagłówka stopnia N do następnego nagłówka poziomu N lub mniejszego, lub do końca dokumentu, jeśli brak takiego nagłówka. Jeśli w rozdziale poziomu N pojawia się bezpośrednio nagłówek stopnia N+2 lub większego, to rozumiemy to jako niejawne otwarcie rozdziałów o pośrednich poziomach, z pustymi tytułami. Można przyjąć, że nie będzie nagłówków stopnia > 5.

Można przyjąć, że jako nagłówki traktujemy elementy `text:h` (stopień podany w atrybucie `text:outline-level`), a jako akapity liczymy (tylko) elementy `text:p` (prefiks `text` mógłby być inny, ale ma oznaczać przestrzeń nazw `urn:oasis:names:tc:opendocument:xmlns:text:1.0`).

Jako tytuł traktujemy cały tekst zawarty w nagłówku, wraz z zawartością elementów potomnych.

Informacje organizacyjne

W razie wątpliwości można pytać mailowo autora zadania: *czarnik@mimuw.edu.pl*

Rozwiązania (w postaci archiwum ZIP o nazwie równej loginowi studenckiemu) należy wysłać do **1 marca 2010** włącznie na adres: *czarnik@mimuw.edu.pl* z konta mailowego na students.

Uzupełnienia i odpowiedzi

1.