Word Document

Jarosław Kuchta

Ten dokument opisuje strukturę dokumentu DOCX tworzoną przez program Microsoft Word. Jest to podsumowanie napisane na bazie standardu ECMA-376-1:2016 z rozszerzeniami zawartymi w standardzie ISO/IEC 29500-1:2016. Standard ECMA-376 opisuje kilka standardów obsługiwanych przez różne aplikacje Microsoft Office. Tutaj wybrano substandard WordprocessingML, w niektórych miejscach występują odwołania do innych substandardów, zwłaszcza DrawingML.

Literatura

1. ECMA-376-1:2016 Office Open XML File Formats
2. [MS-DOCX]: Word Extensions to the Office Open XML (.docx) File Format
3. [MS-OI29500]: Office Implementation Information for ISO/IEC 29500 Standards Support
4. Office Open XML Part 4: Markup Language Reference, December 2006

# Dokument jako pakiet ZIP

Plik DOCX jest w istocie pakietem folderów i plików spakowanych algorytmem ZIP. Można zmienić rozszerzenie pliku z „docx” na „zip” i podejrzeć strukturę pakietu. Podstawowa struktura jest pokazana na rys. 1.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | a) | b) | |  | Obraz zawierający tekst  Opis wygenerowany automatycznie |   Rys. 1. Podstawowa struktura pliku DOCX: a) katalog główny, b) katalog word |

Pakiet zawiera plik Content\_Types.xml oraz trzy katalogi:

* \_rels
* docProps
* word

Plik Content\_Types.xml zawiera definicje zawartości pakietu, w tym poszczególnych plików, zwanych tu „częściami” (parts).

<Types xmlns="http://schemas.openxmlformats.org/package/2006/content-types">

<Default Extension="rels" ContentType="application/vnd.openxmlformats-package.relationships+xml"/>

<Default Extension="xml" ContentType="application/xml"/>

<Override PartName="/word/document.xml" ContentType="application/vnd.openxmlformats-officedocument.wordprocessingml.document.main+xml"/>

<Override PartName="/word/numbering.xml" ContentType="application/vnd.openxmlformats-officedocument.wordprocessingml.numbering+xml"/>

<Override PartName="/word/styles.xml" ContentType="application/vnd.openxmlformats-officedocument.wordprocessingml.styles+xml"/>

…

</Types>

Pliki z rozszerzeniem „rels” to w istocie pliki XML zawierające definicje tzw. „relacji” (relationships) wiążących ze sobą poszczególne „części”. Folder „\_rels” w katalogu głównym dokumentu zawiera plik (bez nazwy, z samym rozszerzeniem) „.rels”, który opisuje relacje głównych „części” do całego dokumentu.

<Relationships xmlns="http://schemas.openxmlformats.org/package/2006/relationships">

<Relationship Id="rId3" Type="http://schemas.openxmlformats.org/officeDocument/2006/relationships/extended-properties" Target="docProps/app.xml"/>

<Relationship Id="rId2" Type="http://schemas.openxmlformats.org/package/2006/relationships/metadata/core-properties" Target="docProps/core.xml"/>

<Relationship Id="rId1" Type="http://schemas.openxmlformats.org/officeDocument/2006/relationships/officeDocument" Target="word/document.xml"/>

</Relationships>

Są w nim zdefiniowane trzy relacje o następujących identyfikatorach:

* rId1 – relacja do głównego pliku dokumentu; plik ten ma nazwę „document.xml” i znajduje się w folderze „word”,
* rId2 – relacja do głównych właściwości dokumentu zapisanych w pliku „core.xml” w folderze „docProps”; główne właściwości są wspólne dla różnych typów dokumentów Office,
* rId3 – relacja do rozszerzonych właściwości dokumentu zapisanych w pliku „app.xml” w folderze „docProps”; rozszerzone właściwości dotyczą dokumentów konkretnej aplikacji Office.

W każdym folderze podrzędnym może się znajdować podkatalog „\_rels” zawierający definicje relacji dla części umieszczonych w tym folderze. Dla przykładu katalog „/word” zawierający główny plik „document.xml” ma podkatalog „\_rels”, który zawiera plik „document.xml.rels”, który wiąże plik „document.xml” z innymi częściami:

<Relationships xmlns="http://schemas.openxmlformats.org/package/2006/relationships">

<Relationship Id="rId3" Type="…/settings" Target="settings.xml"/>

<Relationship Id="rId2" Type="…/styles" Target="styles.xml"/>

<Relationship Id="rId1" Type="…/numbering" Target="numbering.xml"/>

<Relationship Id="rId6" Type="…/theme" Target="theme/theme1.xml"/>

<Relationship Id="rId5" Type="…/fontTable" Target="fontTable.xml"/>

<Relationship Id="rId4" Type="…/webSettings" Target="webSettings.xml"/>

</Relationships>

Odwołania z jednej części dokumentu do innej mogą być jawne (z użyciem identyfikatorów relacji) lub niejawne (przez położenie w tym samym folderze). Dla przykładu część główna dokumentu odwołuje się niejawnie do tabeli czcionek i definicji stylów.

Firma Microsoft oferuje ramowe biblioteki do odczytywania pakietów, które zdecydowanie ułatwiają nawigację po poszczególnych częściach dokumentu, ich ładowanie i tworzenie. Są to:

* System.IO.Packaging – do odczytywania i zapisywania plików ZIP,
* DocumentFormat.OpenXml.Packaging – do odczytywania dokumentów Office,
* DocumentFormat.OpenXml.Wordprocessing – do odczytywania dokumentów programu Word.

Podobne biblioteki są też dla wyższych wersji programu Word oraz innych aplikacji Office.