|  |  |
| --- | --- |
| Symbol projektu:  DocxDocument. ModelGen | |
| Nazwa projektu:  **Generator modelu obiektowego dokumentu** | |
| Tytuł dokumentu: | **ModelDocumentation** |
| Nr wersji: | **2.0** |
| Data utworzenia: | **30.05.2023** |
| Data ostatniej aktualizacji: | **21.06.2024** |
| Osoba odpowiedzialna: | **Jarosław Kuchta (JK)** |
| Autorzy: | **Jarosław Kuchta (JK)** |

Historia dokumentu

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Wersja | Opis modyfikacji | Autor | Data otwarcia | Data zamknięcia |
| 1.0 | Pierwsza wersja | JK | 30.05.2023 | 27.10.2023 |
| 2.0 | Wersja z bazą danych | JK | 2.06.2024 |  |

# Wprowadzenie

Projekt ModelDocumentation służy do pobierania dokumentacji modelu obiektowego w standardzie OpenXml i przygotowania jej do dalszego wykorzystania. Są trzy źródła dokumentacji:

* deklaracje schematu OpenXml zapisane w plikach z rozszerzeniem XSD,
* specyfikacja standardu ECMA-376 opublikowana w kilku plikach PDF i przekonwertowana do formatu DOCX,
* pliki źródłowe pakietu bibliotek DocumentFormat.OpenXml.

Model obiektowy z bibliotek DocumentFormat.OpenXml jest słabo zdokumentowany. Większość komentarzy dokumentacyjnych zawartych w kodzie źródłowym jest trywialna. Najbardziej bogata dokumentacja jest w zawarta w standardzie ECMA, zwłaszcza w pliku „Ecma Office Open XML Part 1 - Fundamentals And Markup Language Reference.pdf”. Celem projektu jest powiązanie modelu obiektowego DocumentFormat.OpenXml z dokumentacją pobraną z pliku ECMA.

Projekt składa się z dwóch modułów:

* ModelDocApp – aplikacji konsolowej do parsowania plików i tworzenia bazy danych dokumentacji modelu OpenXml,
* ModelDoc – modułu dostępowego do bazy danych, który jest wykorzystywany przez aplikację ModelDocApp i może być wykorzystywany w innych aplikacjach.

Projekt działa w oparciu o bazę danych MS Access o nazwie Models.accdb położonej w ścieżce „d:\VS\Docs\DocumentModel”. Ten sposób magazynowania danych został wybrany ze względu na dużą łatwość edycji rekordów bazy danych. Do odwzorowania modelu klas encyjnych w model relacyjnej bazy danych jest wykorzystywany framework Microsoft.EntityFramework.Core.

# Model encji

Model encji dokumentacji składa się z następujących klas encyjnych:

* Namespace – przestrzeń nazw,
* SchemaElement – element ze schematu OpenXml,
* SchemaAttribute – atrybut ze schematu OpenXml,
* ComplexType – typ złożony ze schematu OpenXml,
* SimpleType – typ prosty ze schematu OpenXml,
* StandardElement – element ze standardu ECMA,
* StandardAttribute – atrybut ze standardu ECMA,
* Type – typ (klasa, typ wyliczeniowy) z biblioteki DocumentFormat.OpenXml,
* Property – właściwość klasy z biblioteki DocumentFormat.OpenXml.
* EnumValue – wartość typu wyliczeniowego z biblioteki DocumentFormat.OpenXml

Każda encja jest przechowywana w bazie danych w odpowiedniej tabeli, której nazwa jest nazwą encji w liczbie mnogiej (np. encja Namespace jest przechowywana w tabeli Namespaces).

## Przestrzenie nazw

*Przestrzeń nazw* (Namespace) organizuje nazwy typów. W jednej przestrzeni nazw wszystkie typy muszą mieć unikatowe nazwy. Przestrzenie zostały zaimportowane z kodu źródłowego biblioteki DocumentFormat.OpenXML. Każda przestrzeń nazw ma *nazwę złożoną* z kilku *nazw prostych* oddzielonych kropkami. Nazwa prosta jest *identyfikatorem* składającym się z ciągu liter (łacińskich) i cyfr i rozpoczynającym się od litery. Wszystkie przestrzenie nazw mają nazwy złożone rozpoczynające się od „DocumentFormat.OpenXml”.

Oprócz nazwy przestrzeń nazw ma *prefiks*, który jest krótkim (maks. 10 znaków) ciągiem liter i cyfr, pod którym przestrzeń nazw jest deklarowana w pliku XML dyrektywą xmlns. Ten prefiks jest dodawany przed nazwą elementu XML (oddzielony dwukropkiem). Nie wszystkie przestrzenie nazw mają prefiksy.

Przykład

<w:document   
...  
xmlns:w="http://schemas.openxmlformats.org/wordprocessingml/2006/main" xmlns:w14="http://schemas.microsoft.com/office/word/2010/wordml"   
...>

<w:body>

<w:p w14:paraId="1C1F454E" w14:textId="107F22F8" w:rsidR="007D6C6C" w:rsidRPr="00C0491F" w:rsidRDefault="00732D28" w:rsidP="00F043F7">

<w:bookmarkStart w:id="0" w:name="\_GoBack"/>

<w:bookmarkEnd w:id="0"/>

<w:r>

<w:rPr>

<w:noProof/>

<w14:ligatures w14:val="standardContextual"/>

</w:rPr>

Przestrzeni nazw jest aktualnie 131. Najważniejsze z nich (w kontekście dokumentów programu Word) to:

* XmlSchema – deklaruje podstawowe typy danych atrybutów,
* DocumentFormat.OpenXml – deklaruje podstawowe typy danych modelu OpenXml,
* DocumentFormat.OpenXml.Wordprocessing (w) – deklaruje elementy główne programu Word (w wersji 2007),
* DocumentFormat.OpenXml.Office2010.Word (w14) – deklaruje elementy programu Word dodane w wersji 2010,
* DocumentFormat.OpenXml.Math (m) – deklaruje elementy równań matematycznych,
* DocumentFormat.OpenXml.Drawing (a) – deklaruje elementy rysunkowe.

Encja Namespace ma następujące właściwości:

* ID: int – automatycznie nadawany identyfikator wewnętrzny,
* FullName: string(255) – złożona nazwa przestrzeni nazw,
* Prefix: string(10)? – opcjonalny prefiks,
* Description: text? – opcjonalny opis.

## Elementy schematu XSD

Dokumenty pakietu Office (a w szczególności dokumenty programu Word) są w swojej istocie pakietami ZIP składającymi się z plików XML.

Schemat standardu OpenXml zapisany w plikach z rozszerzeniem XSD zawiera definicje elementów i atrybutów plików XML. Elementy są definiowane głównie jako typy złożone (o nazwach zaczynających się od „CT\_”), a atrybuty jako typy proste (o nazwach zaczynających się od „ST\_”).

Przykład

<xsd:complexType name="CT\_Color">

<xsd:attribute name="val" type="ST\_HexColor" use="required"/>

<xsd:attribute name="themeColor" type="ST\_ThemeColor" use="optional"/>

<xsd:attribute name="themeTint" type="ST\_UcharHexNumber" use="optional"/>

<xsd:attribute name="themeShade" type="ST\_UcharHexNumber" use="optional"/>

</xsd:complexType>

W aplikacji ModelDocApp zebrano pliki XSD w jednym katalogu (tutaj: „D:\VS\Docs\OpenXml\Schemas”) i je sparsowano do bazy danych programu Access (plik Models.accdb). Zapisano je w tabelach o nazwach zaczynających się od „Schema”.

### Pliki i przestrzenie nazw

Schematy XSD zapisane są w kilkudziesięciu plikach z rozszerzeniem XSD. Każdy z tych plików definiuje elementy w określonej przestrzeni nazw schematu. Pliki schematu są przechowywane w tabeli SchemaFiles. Każdą encję reprezentuje klasa SchemaFile o właściwościach:

* Id: int {autonum, PK} – identyfikator encji,
* FileName: string – nazwa pliku (bez ścieżki i rozszerzenia),
* TargetNamespaceId: int {FK „SchemaNamespaces”} – identyfikator przestrzeni nazw definiowanej w tym pliku,
* SchemaNamespace? TargetNamespace – właściwość nawigacyjna do przestrzeni nazw definiowanej w tym pliku.

Przestrzenie nazw są przechowywane w tabeli SchemaNamespaces. Każdą encję reprezentuje klasa SchemaNamespace o właściwościach:

* Id: int {autonum, PK} – identyfikator encji,
* Url: string – URL przestrzeni nazw (długa nazwa),
* Prefix: string[10] – prefiks przestrzeni nazw powszechnie używany w schematach.

Przestrzeni nazw jest więcej niż plików XSD, bowiem wiele plików odwołuje się do standardowych przestrzeni nazw XML i XSD. Są trzy takie przestrzenie:

* xml http://www.w3.org/XML/1998/namespace
* xsd http://www.w3.org/2001/XMLSchema
* inkml http://www.w3.org/2003/InkML

Długie nazwy przestrzeni nazw (URL) zaczynają się od:

* http://purl.oclc.org/ooxml,
* http://schemas.microsoft.com,
* http://schemas.openxmlformats.org,
* urn:schemas-microsoft-com:office.

Nie ma jednego, ustalonego standardu dla przestrzeni nazw OpenXml. Przy wczytywaniu plików XSD z różnych źródeł trzeba czasami zmieniać URL przestrzeni nazw dla dopasowania się do innych plików. W przeciwnym wypadku mogą się pojawić podwójne definicje tych samych elementów.

W każdym pliku schematu mogą być wykorzystywane inne przestrzenie nazw. Do rejestracji używanych przestrzeni nazw służy tabela SchemaUsedNamespaces. Encja SchemaUsedNamespace reprezentuje relację wiele-wiele pomiędzy plikami schematów i przestrzeniami nazw i ma następujące właściwości:

* Id: int {autonum, PK} – identyfikator encji,
* SchemaFileId: int {FK „SchemaFiles”} – identyfikator pliku schematu,
* SchemaNamespaceId: int {FK „SchemaNamespaces”} – identyfikator przestrzeni nazw,
* Prefix: string[10] – prefiks przestrzeni nazw używany w danym pliku.

Prefiksy używane do odwołania się do tej samej przestrzeni nazw mogą być różne w różnych plikach. Zostały one ujednolicone w zebranych plikach XSD tak, aby mogły zostać jednoznacznie wpisane do tabeli SchemaNamespaces. Posłużyła to tego celu klasa pomocnicza SchemaNamespacePrefix, która została wykorzystana do tymczasowego przechowywania wyników kwerendy na tabelach SchemaFiles i SchemaNamespaces.

Ponieważ nie wszystkie przestrzenie nazw zdefiniowane w plikach XSD są wykorzystywane jawnie w innych plikach, więc niektóre prefiksy zostały arbitralnie ustalone.

### Model obiektowy XmlSchema

Wszystkie klasy w modelu XmlSchema opierają się na jednej, podstawowej klasie XmlSchemaObject (rys. 1). W hierarchii dziedziczenia najbardziej rozbudowana jest gałąź XmlSchemaAnnotated. Ta klasa abstrakcyjna reprezentuje wszystkie obiekty XmlSchema, które mogą mieć adnotacje. Jednocześnie obiekty tych klas tworzą główną strukturę schematu.

Rys. 2.Hierarchia dziedziczenia klas XmlSchema

Obiekt XmlSchema może zawierać:

* XmlSchemaSimpleType – typy proste,
* XmlSchemaComplexType – typy złożone,
* XmlSchemaGroup – grupy elementów, które mogą być wykorzystywane w całości przy definicji typów złożonych,
* XmlSchemaAttributeGroup – grupy atrybutów, które mogą być wykorzystywane w całości przy definicji typów złożonych,
* XmlSchemaElement – elementy globalne, które przypisują typy złożone do elementów XML,
* XmlSchemaAttribute – atrybuty globalne, które przypisują typy proste do atrybutów XML.

### Typy proste

Typy proste określają sposób interpretacji wartości atrybutów. Reprezentowane są przez obiekty klasy XmlSimpleType. Ich znaczenie precyzuje obiekt typu XsdSchemaSimpleTypeRestriction. Obiekt taki zazwyczaj odwołuje się do typu bazowego i może na niego nakładać ograniczenia, np. co do długości:

<xsd:simpleType name="ST\_LongHexNumber">

<xsd:restriction base="xsd:hexBinary">

<xsd:length value="4"/>

</xsd:restriction>

</xsd:simpleType>

Ograniczenia typu prostego mają fasetki (ang. Facets), które mogą być różnych typów (zależnie od typu bazowego):

* Dla typu xsd:hexBinary:
* XmlSchemaLengthFacet – określa długość ciągu w bajtach,
* Dla typu xsd:string i xsd:NCName:
* XmlSchemaMaxLengthFacet – określa maksymalną długość w znakach,
* XmlSchemaMinLengthFacet – określa minimalną długość w znakach,
* Dla typu xsd:string:
* XmlSchemaPatternFacet – określa wzorzec (wyrażenie regularne) łańcucha,
* XmlSchemaEnumerationFacet – zazwyczaj jest ich kilka, określają dopuszczalne wartości łańcucha,
* XmlSchemaWhiteSpaceFacet – określa sposób traktowania białych znaków (spacji). Dopuszczalne wartości to preserve (zachowuj), replace (zamieniaj na spacje), collapse (znaki nowego wiersza, tabulatory, spacje, powroty karetki są zastępowane spacjami, spacje wiodące i końcowe są usuwane, a wiele spacji jest redukowanych do jednej spacji). Nie używane w przetwarzanych plikach XSD,
* Dla typów liczbowych:
* XmlSchemaMaxInclusiveFacet – określa maksymalną dozwoloną wartość,
* XmlSchemaMaxExclusiveFacet -– określa maksymalną niedozwoloną wartość
* XmlSchemaMinInclusiveFacet – określa minimalną dozwoloną wartość,
* XmlSchemaMinExclusiveFacet – określa minimalną niedozwoloną wartość,
* XmlSchemaFractionDigitsFacet – określa liczbę cyfr dziesiętnych (nie używane),
* XmlSchemaTotalDigitsFacet – określa całkowitą liczbę cyfr (nie używane).

Przykład ograniczenia długości

<xsd:simpleType name="ST\_XmlName">

<xsd:restriction base="xsd:NCName">

<xsd:minLength value="1"/>

<xsd:maxLength value="255"/>

</xsd:restriction>

</xsd:simpleType>

Przykład wzorca

<xsd:simpleType name="ST\_TextScalePercent">

<xsd:restriction base="xsd:string">

<xsd:pattern value="0\*(600|([0-5]?[0-9]?[0-9]))%"/>

</xsd:restriction>

</xsd:simpleType>

Przykład wyliczenia wartości

<xsd:simpleType name="ST\_Em">

<xsd:restriction base="xsd:string">

<xsd:enumeration value="none"/>

<xsd:enumeration value="dot"/>

<xsd:enumeration value="comma"/>

<xsd:enumeration value="circle"/>

<xsd:enumeration value="underDot"/>

</xsd:restriction>

</xsd:simpleType>

Przykład wartości minimalnej i maksymalnej

<xsd:simpleType name="ST\_Integer255">

<xsd:restriction base="xsd:integer">

<xsd:minInclusive value="1"/>

<xsd:maxInclusive value="255"/>

</xsd:restriction>

</xsd:simpleType>

Przykład wartości wykluczonych

<xsd:simpleType name="ST\_FixedAngle">

<xsd:restriction base="ST\_Angle">

<xsd:minExclusive value="-5400000"/>

<xsd:maxExclusive value="5400000"/>

</xsd:restriction>

</xsd:simpleType>

#### Rejestrowanie typów prostych

Typy proste są rejestrowane w bazie danych w tabeli SchemaSimpleTypes. Encja SchemaSimpleType ma następujące właściwości:

* Id: int {autonum, PK} – identyfikator encji,
* SchemaNamespaceId: int {FK „SchemaNamespaces”} – identyfikator przestrzeni nazw, w której jest zdefiniowany dany typ,
* TypeName: string – nazwa typu,
* BaseNamespaceID: int {FK „SchemaNamespacess”} – identyfikator przestrzeni nazw typu bazowego (pomijany, gdy typ bazowy jest zdefiniowany w tej samej przestrzeni nazw),
* BaseTypeName: string– nazwa typu bazowego,
* Length: int – określenie długości dla typu bazowego hexBinary (w bajtach),
* MinLength: int – ograniczenie minimalnej długości łańcucha lub nazwy (w znakach),
* MaxLength: int – ograniczenie maksymalnej długości łańcucha lub nazwy (w znakach),
* MinInclusive: decimal – minimalna dozwolona wartość dla typu liczbowego,
* MaxInclusive: decimal – maksymalna dozwolona wartość dla typu liczbowego,
* MaxExclusive: decimal – maksymalna niedozwolona wartość dla typu liczbowego,
* MinExclusive: decimal – minimalna niedozwolona wartość dla typu liczbowego,

Wzorce są zapisywane w osobnej tabeli SchemaPatterns. Chociaż w analizowanych plikach XSD nie stwierdzono różnych wzorców przypisanych do tego samego typu prostego, to teoretycznie istnieje taka możliwość. Encja SchemaPattern ma następujące właściwości.

* Id: int {autonum, PK} – identyfikator encji,
* SimpleTypeID: int {FK „SchemaSimpleTypes”} – identyfikator typu prostego, do którego jest przypisany wzorzec,
* Pattern: string – tekst wzorca (wyrażenie regularne).

#### Typy wyliczeniowe

Typy wyliczeniowe w schemacie to są typy proste, które mają określone wartości do wyboru. Wartości te są zapisywane w tabeli SchemaEnumValues. Encja SchemaEnumValue ma następujące właściwości:

* Id: int {autonum, PK} – identyfikator encji,
* SimpleTypeId: int {FK „SimpleTypes”} – identyfikator typu prostego, dla którego podano tę wartość,
* EnumValueStr: string – łańcuch znaków, który jest rozpoznawalny jako dana wartość,
* EnumValueNum: int? – numer kolejny wartości (od zera).

Właściwość EnumValueStr nie musi być nazwą. Może być łańcuchem pustym. Przykładowe rozpoznawalne łańcuchy są następujące:

* -+
* 00
* 1
* 1024,768
* 1024x768
* 1pic
* 35mm
* Aa

Typ wyliczeniowy najczęściej jako bazowy ma typ string, ale też często występuje token, a w pojedynczych przypadkach ST\_String i byte.

#### Typy unijne

Typy unijne to są typy proste, które mogą być interpretowane alternatywnie, a jako zawartość w schemacie mają obiekty typu XsdSchemaSimpleTypeUnion, np.:

<xsd:simpleType name="ST\_Coordinate">

<xsd:union memberTypes="ST\_CoordinateUnqualified s:ST\_UniversalMeasure"/>

</xsd:simpleType>

Typy proste będące typami unijnymi mają jako właściwość BaseTypeName wpisane słowo kluczowe „union”.

Składowe unii są zapisywane w tabeli SchemaUnionMembers. Encja SchemaUnionMember ma następujące właściwości:

* Id: int {autonum, PK} – identyfikator encji,
* SimpleTypeId: int {FK „SimpleTypes”} – identyfikator typu prostego, do którego należy ta składowa,
* MemberNamespaceId: int? {FK „SchemaNamespaces”} – identyfikator przestrzeni nazw, do której należy typ składowej (pomijany, gdy typ składowej jest zdefiniowany w tej samej przestrzeni nazw),
* MemberTypeName: string – nazwa typu składowej.

#### Anonimowe składowe unii

W niektórych uniach jako składowe występują typy proste bez nazwy. Przykład:

<xsd:simpleType name="ST\_DoubleOrAutomatic">

<xsd:union memberTypes="xsd:double">

<xsd:simpleType>

<xsd:restriction base="xsd:string">

<xsd:enumeration value="auto"/>

</xsd:restriction>

</xsd:simpleType>

</xsd:union>

</xsd:simpleType>

Ponieważ odwołanie do składowych w bazie danych następuje przez nazwę, to takim typom w czasie zapisu do bazy danych nadawana jest nazwa „\_anon\_*N*”, gdzie *N* jest kolejnym numerem typu anonimowego.

#### Typy listowe

Typy listowe to typy proste, które określają taki sposób interpretacji łańcucha znaków jakby to miały być listy wartości. W schemacie te typy mają zawartość typu XsdSchemaSimpleTypeList, która określa typ elementów, np.:

<xsd:simpleType name="ST\_Sqref">

<xsd:list itemType="ST\_Ref"/>

</xsd:simpleType>

Typy proste będące typami listowymi mają jako właściwość BaseTypeName wpisane słowo kluczowe „list”.

Elementy składowe listy są zapisywane w tabeli SchemaListItems. Encja SchemaListItem ma następujące właściwości:

* Id: int {autonum, PK} – identyfikator encji,
* SimpleTypeId: int {FK „SimpleTypes”} – identyfikator typu prostego, do którego należy ta składowa,
* ItemNamespaceId: int? {FK „SchemaNamespaces”} – identyfikator przestrzeni nazw, do której należy typ składowej (pomijany, gdy typ składowej jest zdefiniowany w tej samej przestrzeni nazw),
* ItemTypeName: string – nazwa typu składowej.

### Typy złożone

Typy złożone określają sposób interpretacji elementów Xml. Reprezentowane są przez obiekty klasy XmlComplexType. Szczegóły są zapisane w atrybutach i elementach definicji.

Deklaracje typów złożonych, które nie mogą zawierać w sobie elementów, mają tylko

### Grupy elementów

### Grupy atrybutów

### Elementy globalne

### Atrybuty globalne

W kolejnych rozdziałach i podrozdziałach głównej części specyfikacji opisano elementy OpenXml. Dla każdego z nich podano jaki znacznik XML reprezentuje ten element, jakie inne elementy ten element może zawierać i jakie może mieć atrybuty XML. Atrybuty mają określone typy wartości. Mogą to być typy podstawowe XML Schema (jak np. xsd:string) lub typy proste OpenXml. Każdy z typów prostych jest opisany w osobnym podrozdziale specyfikacji ECMA.

Encja Element ma następujące właściwości:

* ID: int – automatycznie nadawany identyfikator wewnętrzny,
* SpecNum: varchar(255) – numer podrozdziału ze specyfikacji ECMA,
* Ns: int – identyfikator przestrzeni nazw,
* Tag: varchar(255) – znacznik elementu OpenXml,
* FullName: varchar(255) – złożona nazwa przestrzeni nazw,
* Description: varchar(MAX)? – opcjonalny opis elementu.
* Attributes: Collection<Attribute> – kolekcja atrybutów OpenXml,
* ChildElement: Collection<ChildElement> – kolekcja relacji wiążących element OpenXml z podrzędnymi elementami OpenXml.

Encja Attribute ma następujące właściwości:

* ID: int – automatycznie nadawany identyfikator wewnętrzny,
* ElementID: int – identyfikator encji Element, do której należy
* Ns: int – identyfikator przestrzeni nazw,
* Tag: varchar(255) – znacznik atrybutu OpenXml,
* FullName: varchar(255) – złożona nazwa przestrzeni nazw,
* Description: varchar(MAX)? – opcjonalny opis elementu.
* SimpleTypeID: int – identyfikator typu prostego wartości atrybutu.

Encja SimpleType ma następujące właściwości:

* ID: int – automatycznie nadawany identyfikator wewnętrzny,
* SpecNum: varchar(255)? – opcjonalny numer podrozdziału ze specyfikacji ECMA,
* Ns: int – identyfikator przestrzeni nazw,
* TypeName: varchar(255) – nazwa typu prostego OpenXml,
* FullName: varchar(255) – złożona nazwa przestrzeni nazw,
* Description: varchar(MAX)? – opcjonalny opis elementu.
* Values: Collection<SimpleTypeValue> – kolekcja wartości typu prostego.