Štěpán Šrubař
ročník 2.
14.11.2006

Uživatelské formáty metadat

1. Úvod

Objem digitálních dat stále roste a to především díky digitálním multimédiím. Nejen k jejich používání ale také k jejich organizaci potřebujeme metadata. Jde o informace k danému digitálnímu obsahu. Aby se dala tato metadata snadno vytvářet, číst a uchovávat, vznikla spousta formátů. Velké společnosti vytvářely formáty metadat, které byly schopny uchovat všechny možné informace. Aby v této volnosti nevznikl chaos, bylo definováno, co, kam a jaké informace ukládat a co vlastně znamenají. Pro běžného uživatele to však bylo příliš složité. Dokud totiž měl několik digitálních multimédií, nepotřeboval žádná metadata, neboť si vše pamatoval. Jakmile množství těchto nezávislých souborů multimédií vyrostlo do řádově stovek, pak bylo třeba jednoduché organizace, která umožňovala zadat a uchovat základní informace. V případě hudby to jsou například název, autor, album a číslo stopy. V případě digitální fotografie čas pořízení, parametry pořízení a popis. V případě videa hlavně jazyk zvuku a titulků a seznam kapitol. Z tohoto důvodu vznikaly různé formáty metadat podle typu dat a jejich formátu.

Jednoduché formáty metadat lze snadno vytvářet stejně jako programy, které toto ještě více usnadňují. S rostoucí poptávkou po větším množství typů uchovávaných metadat vznikají novější formáty, které jsou odvozeny ze starších a stále jsou dostatečně jednoduché. Já jsem se zaměřil na široce rozšířené formáty především pro hudbu a obrázky. Dále budu používat pojem identifikátor, což je označení kategorie, která je společná pro digitální data. Jedná se například o pojem autor, název, žánr apod. Samotná metadata odpovídající konkrétnímu obsahu budu nazývat popis. Jde například o Jim Carrey, Flying, rock apod. Dvojici identifikátor,

popis budu nazývat značka (tag). Pod tímto označením budu také uvažovat dodatečné informace jako jeho délka, případně speciální ohraničující symboly.

2. Formáty metadat pro zvukové formáty

2.1 APE

APE tag [1] je formát metadat vytvořený pro formát ztrátové komprese zvuku MPC. Dnes je však používán i pro bezztrátové formáty komprese jako Monkey's audio, WavPack a OptimFROG a lze jej použít i pro formát MP3. Tato vlastnost je třeba implementována v programu TagScanner. Využití pro jiné formáty záleží jen na podpoře ze strany přehrávačů. Například přehrávač Foobar2000 toto umožňuje. APE tag je dnes ve verzi 2 – APEv2. Metadata lze umístnit na konec souboru a narozdíl od APEv1 i na začátek. Identifikátor musí být v kódování tisknutelných znaků ASCII délky 2 až 255. Vhodnější je použít definované identifikátory. Na druhou stranu není povoleno použít označení ID3, TAG, OggS and MP+. Identifikátory jsou závislé na velikosti písmen (case sensitive), avšak doporučuje se toho nevyužívat. Popis může být textový v UTF-8, celočíselný, s pohyblivou čárkou, datum, čas, rozsah časů, časovaná slova, odkazy, binární data a seznam údajů. Jiný obsah se povoluje, ale není doporučen. Seznam lze například využít k zápisu více autorů jednoho díla. Je zde také definovaný tag dummy, který lze využít jako tzv. vata (padding). Pokud je hlavička na začátku souboru a nemáme systém, který umožňuje rozšiřovat soubor uprostřed, pak vytvoříme značku, která pouze zabírá místo a nemá informační charakter. V případě, že chceme do nějakého popisu něco připsat, tak posuneme značky mezi tímto popisem a prázdnou značkou a tuto prázdnou značku zmenšíme, abychom mohli zanechat data za ní beze změny. V případě zmenšení nebo smazání nějaké značky pouze přesuneme mezilehlé značky a tuto prázdnou zvětšíme. Takto nemusíme vůbec zpracovávat či přesouvat zbytek souboru.

Příklad: "Artist=Melanie C", "Album=Northern Star", "Date=1999", "Title=Go!", "Track=1".

2.2 ID3

Asi nejznámější formát metadat pro zvuky a hudbu je ID3 [2]. V první verzi označované ID3v1 (1996) se začal používat s formátem komprese MP3. Jako jediný zde jmenovaný formát zabírá fixní velikost 128 B, tudíž má pevně dány identifikátory a délka popisu je pak shora omezena na dnes nedostatečnou velikost. V roce 1997 byla proto vytvořena verze ID3v1.1, která se dnes běžně označuje stále ID3v1, neboť přináší pouze minoritní rozšíření a je zpětně kompatibilní. Dalším problémem byla nemožnost zapsat názvy v různých abecedách, neboť se používalo kódování ASCII. Proto přišla verze ID3v2. Ta už umožňuje vytvářet popisy libovolné délky. Od verze ID3v2.4 (1.11.2000) lze ukládat záznamy ve znakové sadě Unicode v kódování UTF-8 nebo UTF-16. Identifikátory jsou definovány specifikací a jsou ukládány jako čtveřice písmen. To znepříjemňuje přidávání nových identifikátorů do specifikace. Například přehrávač Winamp ve verzi 5.3 podporuje přehrávání s hlasitostí uloženou v metadatech skladby. Tato informace je tedy ukládána do značky uživatelského textu. Až se používání této možnosti rozšíří a bude nová revize ID3, pak budou pro toto nejspíš stanoveny nové identifikátory. ID3v2.4 umožňuje vatu (padding) za poslední značkou, kompresi a šifrování popisů.

Příklad: "TPE1=Melanie C", "TALB=Northern Star", "TYER=1999", "TIT2=Go!", "TRCK=1".

2.3 Vorbis comment

Pro formát Vorbis komprese zvuku a hudby kvality CD byla navržena obálka Ogg, se kterou přišel jednoduchý formát metadat Vorbis comment [3]. Sama specifikace uvádí nemožnost použití binárních dat, které by měly být přímo zakódovány do obálky Ogg. Struktura je velmi jednoduchá a složitější metadata by podle doporučení měla být uložena v nějakém XML formátu. Identifikátory mohou být libovolně dlouhé, avšak kódovány pouze v ASCII. Jsou nezávislé na velikosti písmen (case insensitive) a mohou se vyskytovat několikrát. Samotný popis je kódován do UTF-8. Značek může být až 4 Gi (Gibi = 2^{30}) [4] a délka

identifikátoru a popisu může být až 4 GiB. Specifikace uvádí 15 doporučených identifikátorů s jejich sémantikou (významem).

Příklad: "ARTIST=Melanie C", "ALBUM=Northern Star", "DATE=1999", "TITLE=Go!", "TRACK=1".

3. Formáty metadat pro obrazové formáty

3.1 EXIF

Jeden z nejpoužívanějších formátů metadat pro obrázky je EXIF [5]. Byl navržen společností JEITA (Japan Electronic Industry Development Association), která v září 2003 představila verzi 2.21, která je pouze drobným pozměněním verze 2.2 z dubna 2002. Kompletní specifikace je placená. Každá značka je složena z identifikátoru, formátu popisu, velikosti a hodnoty. Pokud je popis do velikosti 4 B, pak je vepsán přímo do hodnoty, pokud je větší, tak je v hodnotě uložena pozice popisu v souboru. Z tohoto vyplývá, že popis může být velikosti maximálně 4 GiB – 1 a jeho začátek může být až na pozici 4 Gi – 1. Identifikátory jsou zde pevně definovány a reprezentovány dvěma bajty. Jsou například pro rozložení barev v obrázku, rozměry, informace o snímání a mnoho dalších. Specifikace počítá se zakomponováním zvuku, ovšem pouze ve třech formátech (PCM, μ-law, AD-PCM). Popis může být kódován v ASCII 7b, číselný nebo nedefinovaného formátu, tudíž zde není podpora jiných abeced, přestože sama společnost je japonská. EXIF vychází z TIFF Rev. 6.0 a rozšiřuje jej. Revize 6.0 formátu TIFF vyšla již 3.6.1992 a definuje uložení dat včetně popisu jejich komprese a rozložení. Značky mají stejný formát jako v EXIF, avšak jejich identifikátory jsou odlišné.

Příklad: "256=1024", "257=768", "274=8", "532=[0, 255, 0, 255, 0, 255]"

3.2 IPCT core

IPCT core [6] je standard vytvořený společností International Press Telecommunications Council. Jedná se o převzatý formát Information Interchange Model (IIM) vytvořený Adobe. Celé je to samozřejmě kompatibilní s dnešní Extensible Metadata Platform (XMP) vytvořenou Adobe v roce 2001. Jelikož XMP je založeno na XML, pak by popis v IPCT core mohl být kódován do unicode, avšak ve specifikaci to není nikde explicitně vyjádřeno. Identifikátory jsou pevně definovány a každý odpovídající popis může být libovolně dlouhý, avšak kvůli zpětné kompatibilitě je lepší nepřekračovat maximální délku z dřívějšího IIM. Identifikátory creator, rights, description, subject a title dokonce odpovídají identifikátorům dublin core. Urgency, category a supplemental categories jsou specifikací zavrhovány.

Příklad: "Creator=Julie Doe", "Creator's Jobtitle=Mugwum contract photographer", "Headline=Shore Temple, Malibalipuram, India"

4. Schémata formátů

Formát APEv2
APE hlavička
APE značka 1
...
APE značka n
APE patička

APE hlavička/patička
'APETAGEX'
číslo verze
celková velikost
počet značek (n+1)
příznaky

Příznaky	
bit 0	pouze pro čtení
bit 1, 2	kódování popisu
bit 29	hlavička, ne patička
bit 30	ne patička
bit 31	hlavička

Značka
velikost popisu
příznaky
identifikátor (ASCII)
0x00
popis (UTF-8)

Formát ID3v2
hlavička
rozšířená hlavička
značky
vata (padding)
patička

Hlavička (patička)
'ID3' ('3DI')
verze (0x0400)
příznaky hl.
celková velikost

Značka
identifikátor (4 B)
velikost popisu
příznaky zn.

Příznaky zn.	
bit 2	zrušení značky
bit 3	zrušení souboru
bit 4	jen pro čtení
bit 10	skupinová značka
bit 13	komprese zlib
bit 14	šifrování
bit 15	nesynchronní
bit 16	indikátor velikosti dat

Formát Vorbis comment
délka názvu kodéru
název kodéru
počet značek
délka značky 1
značka 1
...
délka značky n
značka n
bit hodnoty 1

Značka
délka značky
identifikátor (ASCII)
0x3D (=)
popis (UTF-8)

Formát EXIF 2.2 pro JPEG
začátek soboru obrázku
aplikační segment značek 1
aplikační segment značek 2

Aplikační segment značek 1
ASZ 1 označení
ASZ 1 délka
EXIF identifikační kód
TIFF hlavička
0. IFD
hodnota 0. IFD
1. IFD
hodnota 1. IFD
1. IFD data obrázku (náhled)

IFD
identifikátor (2 B)
typ (číslo, zlomek, ASCII)
velikost popisu
odkaz na popis

5. Shrnutí a návrh budoucího formátu

Nejjednodušším a zároveň velmi flexibilním formátem metadat je zde Vorbis comment. Typovou definici zavádí některé formáty přímo ve specifikaci podle identifikátorů, některé umožňují definovat typ přímo ve značce (EXIF). Identifikátory jsou buď textové (APEv2, Vorbis comment), písmenné (ID3v2) nebo číselné (EXIF). Žádný formát neumožňuje vnořování značek ani plnou podporu unicode. V případě vytvoření nového formátu metadat bych použil XML formát s možnostmi: seznam doporučených identifikátorů se sémantikou (profily), možnost vnořovat značky, identifikátory i popisy v unicode se specifikací v jakém jsou jazyce, opakování stejných identifikátorů nebo jejich shlukování, podpora typů celé a racionální číslo, text, datum, časové i číselné rozmezí a seznam. Stromové struktury bych zavedl pomocí vnořování identifikátorů. Takovýto formát by měl být schopen popsat všechny typy multimédií (text, zvuk, obraz, video). Při jejich spojování a rozdělování by se pak jednoduše automaticky spojovala, rozdělovala a případně duplikovala metadata.

6. Odkazy

[1] http://wiki.hydrogenaudio.org/index.php?title=APEv2
[2] http://en.wikipedia.org/wiki/Id3
[3] http://www.xiph.org/vorbis/doc/v-comment.html
[4] http://en.wikipedia.org/wiki/Binary_prefix,
http://cs.wikipedia.org/wiki/Bin%C3%A1rn%C3%AD_p%C5%99edpona
[5] http://en.wikipedia.org/wiki/EXIF, http://cs.wikipedia.org/wiki/Exif
[6] http://www.iptc.org/IPTC4XMP/

7. Zápis Dublin core

TITLE = Uživatelské formáty metadat
LANGUAGE = cs : ISO 639-2
CREATOR = Bc. Štěpán Šrubař
SUBJECT = metadata, Vorbis comment, ID3v2, APEv2, EXIF, IPTC core
DESCRIPTION = přehled několika formátů pro ukládání metadat používaných běžnými uživateli
PUBLISHER = FI MU
DATE = 2006-11-24
FORMAT = vnd.oasis.opendocument