Získání metadat z obrazových souborů a jejich uchování v RDF

Martin Kryl[1]

1 Úvod

V současné době existuje velké množství formátů souborů, do kterých lze zapisovat data. Datový soubor může kromě samotných dat obsahovat i popis dat, tzv. metadata. Přečtením a pochopením metadat lze získat informace užitečné pro další zpracování souborů, jejich analýzu nebo vyhledávání.

Struktura metadat je většinou definována pouze pro konkrétní formát souboru nebo malou množinu formátů. Jeden formát souboru může navíc obsahovat několik typů metadatových struktur. Metadata z různých struktur se mohou významově překrývat.

Cílem této práce je vytvořit plugin pro program MetaMed, který je vyvíjen na Katedře informatiky a výpočetní techniky v rámci výzkumné skupiny Medicínské informační systémy. Plugin umožní číst metadata z obrazových souborů JPEG, TIFF a PNG. Extrahovaná metadata budou zapsána do RDF (Resource Description Framework) modelu.

2 Popis řešení

Otázka extrakce metadat z obrazových souborů není nová a existuje řada aplikací či knihoven, které problém řeší. Na základě srovnání množin čtených metadat a dalších vlastností nástrojů Exiv2, ExifTool a Metadata Extractor je vybrán Metadata Extractor ke čtení metadat v projektu. Nástroj neřeší otázku zápisu získané informace do RDF modelu. Výstup v podobě namapovaných metadat na RDF vlastnosti nabízí z nalezených nástrojů pouze framework Aperture, který nevyhovuje požadovanému rozsahu zpracovatelných metadat.

Byly hledány RDF slovníky a ontologie definující pojmy použitelné k zapsání získaných metadat. Největší množinu užitečných pojmů poskytla sada ontologií NEPOMUK Information Element. Dále jsou použity Dublin Core, Geo, Friend of a Friend. Slovníky, které jsou definovány pro Extensible Metadata Platform vycházející z konceptu RDF, nemohou být použity z důvodu nedostupnosti jejich serializace. Pro metadata, která nemohla být namapována na existující vlastnost, byla v ontologii *http://mre.kiv.zcu.cz/ontology/2015/03/image.owl* vytvořena patřičná vlastnost. Ontologie dále obsahuje definici vlastností, které mají charakter kontrolovaných slovníků. NEPOMUK Exif ontologie například požaduje ve vlastnosti *resolutionUnit* hodnotu *2*, pokud je jednotkou rozlišení DPI. Pro uživatele je užitečné zapsat i hodnotu, které kód odpovídá. Proto je získaná hodnota zapsána jak do této vlastnosti, tak do vlastnosti v nově vytvořené ontologii.

Za účelem definování způsobu mapování získaných metadat na RDF vlastnosti byla vytvořena ontologie *http://mre.kiv.zcu.cz/ontology/image-mapping.owl*. Pro každý metadatový tag, který Metadata Extractor zná, je definována instance RDF třídy *imageMetadata*. V instanci lze určit vlastnost, na kterou je mapováno, a typ uzlu, ke kterému je vlastnost přiřazena. Dále

[1] student navazujícího magisterského studijního programu Inženýrská informatika, obor Informační systémy, e-mail: kryl@students.zcu.cz

je možné nastavit transformaci čtené hodnoty, složení výsledné hodnoty z více čtených hodnot, rozdělení čtené hodnoty na více podřetězců, či zapsání hodnoty jako prvek kontrolovaného slovníku. Je zaveden mechanismus pro vytvoření více definic zapsání jednoho tagu.

Při implementaci bylo zavedeno 16 transformací extrahovaného řetězce. Některé z nich jsou obecné a bylo by je možné použít při vytváření nových mapovacích definic. Jde například o funkci k odstranění písmenných znaků z řetězce, převod racionálních čísel na desetinná, dělení celého čísla jiným nebo výběr řetězce ze zadané množiny, který je na indexu odpovídajícímu čtené hodnotě. Implementace dále řeší jednoznačné pojmenování vytvářených RDF zdrojů na základě jejich vlastností. Zdroj popisující fotoaparát použitý k vytvoření série snímku je tak ve výsledném RDF modelu pouze jednou a všechny zdroje odpovídající snímkům ze série na něj odkazují.

3 Výsledek

Řešení bylo testováno na sadě 277 obrazových souborů o celkové velikosti 680 MB. Konfigurace testovacího stroje byla Intel Core i3-2100 3,1 GHz, 8 GB RAM, OS Windows 7, 64 bitová verze. Zpracování bylo dokončeno za 16,4 vteřiny, z toho 10,7 vteřin trvalo čtení metadat knihovnou Metadata Extractor. V případě, že soubory již byly načteny v paměti po předešlém běhu programu, se doba zpracování zkrátila na 8,4 vteřiny. Jeden soubor byl zpracován průměrně za 59 ms při prvním čtení a 30 ms při opakovaném čtení.

V tabulce 1 uvádím souhrnně počty metadatových tagů, se kterými se práce zabývá. Makernote tagy jsou specifické tím, že jde o proprietární metadatové formáty výrobců zařízení, které nejsou zdokumentovány. Některé z nich byly popsány pomocí reverzního inženýrství. V tabulce jsou jako mapované uvedené takové tagy, kterým odpovídá vlastnost některého z existujících slovníků. Pro zbylé jsou použity vlastnosti vytvořené ontologie. V posledním sloupci je počet tagů, které se vyskytují alespoň v jednom ze souborů testovací sady.

Typ metadat	Známé	Mapované	Ve vzorku
Makernote	722	107	600
Ostatní	430	207	276

Tabulka 1: Tabulka s počtem tagů, které Metadata Extractor zná, které jsou v řešení mapovány na existující RDF vlastnosti, resp. které se vyskytovaly v testovacím vzorku.