

Posudek vedoucího diplomové práce

Jméno a příjmení: *Bc. Michal Madleňák*
Název tématu: *Správce sbírek založený na technologiích sémantického webu*
Vedoucí: *Ing. Petr Včelák, KIV*

Informace k zadání

Cílem diplomové práce bylo vytvoření aplikace pro organizaci a procházení sbírek různých typů heterogenních dat s použitím technologií sémantického webu. Motivací bylo umožnit uživateli zvýšit přehled o jeho dokumentech a přinést možnosti moderní organizace a pohledu na data.

Součástí má být seznámení se se sémantickými technologiemi, analýza typů dokumentů, nalezení vhodných knihoven pro získání metadat a nalezení vhodných ontologií pro strukturování metadat. Na základě nastudované teorie a analýzy má být navržen a implementován grafický uživatelský nástroj pro organizaci a procházení sbírek dle automaticky získaných nebo od uživatele doplněných metadat.

Aktivita během řešení, konzultace, komunikace

Diplomant pracoval samostatně, aktivně a komunikace byla rychlá. Pravidelně se účastnil konzultací, které byly převážně orientovány na řešení nejasností nebo problémů. Dosažené výsledky a postup prací zmiňoval.

Formální stránka

Diplomová práce obsahuje 66 stran čistého textu v jednoduchém řádkování. Dále je v práci 81 zdrojů literatury a 20 příloh. Minimálně 6 příloh obsahuje nic neříkající náčrty GUI. Dokumentu chybí seznam obrázků, tabulek a výpisů kódu. Realizovaný softwarový nástroj je na přiloženém CD-ROM ve zdrojové i binární podobě. K dispozici je i uživatelská příručka. Na přiloženém CD-ROM je několik testovacích souborů, i tři používané ontologie, což považuji za přínosné. Ovšem, v případě souboru *2MP3.ZIP* se může jednat o autorsky chráněná hudební díla, což nikde není licenčně ošetřeno ani zmíněno.

Po formální stránce je dokument spíše průměrný. Práce obsahuje řadu překlepů včetně překlepu přímo v názvu zadání u titulního listu. Místy se setkáme s nejednotností formátování textu, ve větším množství jsou přítomny typografické chyby. Celkový dojem kazí černobíle a nekvalitně vytištěné obrázky. Při čtení textu je až obtěžující použití nadměrného množství křížových odkazů z nichž některé jsou chybně. V textu se opakují již řečené části, shodné odstavce (kapitola 6.5 a 7) a stejné názvy kapitol (6.4.1 a 6.4.2).

Základní kostra práce je logicky rozvržena na informace o sémantickém webu a následně dle fází vývoje. Bylo by možné zvolit jiné pořadí podkapitol, aby nebylo potřeba tolik dopředných odkazů (od kapitoly 2.2.2 jsou příklady v RDF/XML, ale způsoby serializace RDF dat uvádí až kapitola 2.7). V teoretické části autor velmi často *zabíhá do nepodstatných detailů* (např. detailní podoba URI v kapitole 2.1), které nejsou dále v práci použity ani se tématu přímo netýkají (kapitoly 2.2.2, 2.2.3, 2.2.4, 2.4, 2.10.1, 2.10.2, 2.10.3 a 2.7.5; Ilustrace 3. a 5.). Z pohledu zadání jsou vybrané *významné informace utopeny* až na 4. úrovni nadpisů (4.2.2.1–4.2.2.4) a tedy i těžko dohledatelné. Kapitola 4.2.3.2 se týká analýzy, ale popisuje provedené implementační chyby. Kapitoly 5.3 až 5.8 (str. 38–40) obsahují pouze jednořádkové věty s odkazy na přílohy, což působí nadbytečně, obzvláště v případě náčrtů GUI.

V teoretické části *chybí ucelená analýza existujících formátů*, způsobu získání jejich metadat a chybí i výčet existujících pro ně vhodných ontologií. Významné kapitoly věnované metadatům (2.5) a ontologiím (2.6) jsou psány pouze populární formou. Pro omezenou, implementovanou skupinu formátů jsou informace o existujících knihovnách a ontologiích až v analýze (4.2.2.4).

Text obsahuje řadu *nepřesností, chybných, zavádějících* nebo *odporujících si* definic a informací: nativní/perzistentní; reifikace; RDF Schema je slovník, nikoliv ontologie; v diagramu užití (Ilustrace 10, str. 37) *chybí případ užití „procházení sbírek“*, které je hlavním cílem nástroje; diplomant si ve věci volby NetBeans při analýze 4.1.1.4 (str. 26) volí z důvodu jejich znalosti, ale při implementaci

6.2.3 (strana 43) znalost už popírá; v práci jsou popisována RDF úložiště Jena a Sesame, přitom v Tabulce 1 (str. 28) uvádí i úložiště Virtuoso TS, Virtuoso SQL a blíže nedefinovaný řádek relační databáze MySQL; o existenci *xdg-open*, byl student informován, ale informaci ignoroval (4.2.4.1).

Kladně a jako užitečné hodnotím vložené ukázky kódu čtení metadat z různých typů souborů.

Realizační část

Z pohledu uživatele nesouhlasím s tvrzením diplomanta – aplikace není uživatelsky příjemným nástrojem. Práce s méně jak 10 soubory je nepohodlná až obtížná. V případě rozsáhlé sbírky obrázků nebo hudebních nahrávek už bude aplikace nepoužitelná. Je pravdou, že zobrazování sbírek souborů bylo domluveno řešit v externím programu, ale považuji za chybu zobrazovat pouze vypočítaný hash v seznamu souborů nebo výsledků vyhledávání namísto náhledu nebo alespoň jiných vybraných metadat. Seskupování souborů pro docílení lepší organizace a pohledu na data není možné, ale je možné použít základní formu vyhledávání.

Diplomant řeší extrakci pouze velmi omezené množiny metadat z podporovaných formátů. O automatizaci získání metadat se ani nepokouší a získává je pouze „ručně“.

Zásadní nedostatek vidím v chybném použití sémantických technologií, které vede ke ztrátě interoperability získaných metadat. Chybí strukturování metadat s použitím tříd z ontologií. Všechna ze souboru získaná metadadata jsou vždy uložena v jediné instanci bez typu! Vznikají duplicity v datech, protože není možné docílit vazeb typu „sbírka–soubory“, „interpret–hudební alba–stopy“, apod.

Importovat lze pouze jedinou ontologii ke zvolené skupině dat (hudba, obrázky, dokumenty). Při opakovaném importu vždy dojde k odstranění předchozí. Lze editovat identifikátor, anglický nebo český název vlastnosti. Pro třídy lze měnit pouze názvy. Ontologie ze všech skupin dat nelze z aplikace exportovat a tedy ani přenést jinam.

Zdrojové kódy jsou stručně komentovány, ale v některých místech by bylo vhodné detailnějších komentářů. Je přítomna vlastní nefungující implementace logování chyb. Výjimky jsou ošetřovány použitím *java.util.logging.Logger*. Místa jsou přímé výpisy do konzole. Lokalizace by byla problematická, neboť textové řetězce jsou i přímo ve zdrojových kódech. Vytvářený soubor *errors.txt*, datové úložiště *.storeage/* a *.store_direction/* jsou nevhodně zapisovány v místě aplikace a spuštění bez zápisových oprávnění končí pádem aplikace (GNU/Linux). Soubor *errors.txt* je i po chybách vždy prázdný. Konfigurace preferovaných externích prohlížečů je interně uvnitř aplikace.

S tvrzením diplomanta, které uvádí ve dvou kapitolách citaci ze „života“ ve znění: „*V dnešní době zákazníka nezajímá funkčnost aplikace, ale hlavně vzhled.*“, se neztotožňuji. Z jeho pohledu však předkládaný nástroj uvedenou citaci doslovně splňuje – úvodní obrazovka (navržená grafikem) se povedla, ale použitelnost a nezbytné funkce jsou nedostatečně propracované nebo byly zcela zanedbány.

Splnění požadavků zadání

Požadavky zadání považuji za splněné značně minimalistickým způsobem. Ve svém důsledku realizace degraduje sémantické technologie na pouhé úložiště dat bez dalších výhod. Řada zásadních otázek nebyla vůbec uvažována a některé problémy jsou bohužel velmi zjednodušeny nebo i zanedbány.

Dosažené výsledky v teoretické a především v realizační části nepovažuji za použitelné ani přínosné pro další projekty nebo pro možnost na ně navázat.

Závěrečné hodnocení

Po zvážení všech dosažených výsledků a především po přihlédnutí k připomínkám v realizační části, jsem nucen celkově uvést – hodnocení známkou **nevyhověl** a diplomovou práci **nedoporučuji k obhajobě**.

Otázky

Na základě textu diplomové práce a praktického využití v praxi pokládám následující otázky k obhajobě.

1. Jaký počet metadat je automaticky získáván a uchováván v RDF úložišti ze vstupních souborů? Uveďte konkrétní čísla pro všechny podporované formáty v přehledné formě.
2. Kolik jste jako zákazník za aplikaci správce sbírek v této podobě ochoten zaplatit? Proč?

V Plzni 1. 6. 2012



Ing. Petr Včelák
KIV, FAV, ZČU