ACTA UNIVERSITATIS AGRICULTURAE ET SILVICULTURAE MENDELIANAE BRUNENSIS
SBORNÍK MENDELOVY ZEMĚDĚLSKÉ A LESNICKÉ UNIVERZITY V BRNĚ

Ročník LV 19 Číslo 6, 2007

VLIV FORMÁTU UKLÁDANÝCH DOKUMENTŮ NA PRODUKTIVITU INFORMAČNÍCH SYSTÉMŮ

O. Trenz

Došlo: 12. července 2007

Abstract

TRENZ, O.: *The impact of document format on productivity of information systems*. Acta univ. agric. et silvic. Mendel. Brun., 2007, LV, No. 6, pp. 177–186

Document processing is significant and limiting factor of the efficiency of every information system that creates, stores, and revises documents during its operation. Selection of proper document format together with the data that describe the document can substantially influence the speed of processing the documents.

We should consider the document format already in the phase of systems design. If documents that are processed are of one type it is useful to follow one descriptive style during their creation. This can be achieved through using sample document templates.

Not less important is also the method of storing the documents. This aspect is crucial mainly in cases when the presentation and processing of the documents is fundamental. The design of the structure and methods of storing the documents are influenced by the operations that will be carried out with the documents. Big impact on the structure has e.g. implementation of efficient searching algorithm in full text or semantic variant.

This paper doesn't provide complex solution of document systems in information systems but uses suitable approaches and comparisons to show effective way how to implement such system with the compliance with new technologies.

document, document system, document structure, document scheme

MATERIÁL A METODY

Problematika tvorby a správy dokumentů ve vztahu k informačnímu systému je velmi často diskutované téma. Vhodná implementace ve spojení s vhodnou politikou práce s dokumenty může ve výsledku značně zefektivnit práci s informačním systémem na dokumentové úrovni. Bezesporu vhodná volba struktury dokumentu spolu s dokumentovým formátem umožní provádět dokumentové operace efektivněji a ve výsledku tedy levněji; toto je velmi podstatný prvek ovlivňující návrh celého systému (Malo a Raszková, 2006). Pokusme se podrobněji zabývat jednotlivými aspekty v návrhu struktury dokumentů, jejich vhodným ukládáním ve spojení s nutnými režijními prvky a optimalizací pro vyhledávání v rámci dokumentů. Proveďme taktéž srovnání se současnými přístupy při využití posledních poznatků v této oblasti.

Pro přiblížení problematiky práce s dokumenty se dále zabývejme poradenským systémem, nebo přesněji řečeno poradenským informačním systémem. Je evidentní, že takovýto informační systém musí disponovat dokumentovým modulem, jenž je podporou při vytváření, revidování a zneplatňování dokumentů. Dokumentový modul mimo správu dokumentů ve většině případů umožňuje taktéž vytvářet popisné záznamy o vkládaných zdrojích a v neposlední řadě má integrováno i vyhledávání napříč uloženými dokumenty.

Dle specializace poradenského systému mohou mít dokumenty povahu pracovních dokumentů, vyhlášek, vládních nařízení, výrobních postupů, výkresové dokumentace, nabídky produktů, faktury, vlastních dokumentů zákazníků, nebo například formu případové studie.

Vhodný prvotní návrh práce s dokumenty může ve výsledku značně zjednodušit práci s daným poradenským systémem. Již v momentě návrhu je nutno zohlednit specifické rysy práce s dokumenty v daném poradenském systému a snížit tak časovou obslužnou režii. Zejména pak eliminací té časové složky, jež reprezentuje přístup k souborům na disku.

Je evidentní, že námi modelovaný poradenský informační systém, jakožto obecně jakýkoliv obecný informační systém, bude v dnešní době již provozován na bázi internetových technologií. Je to nutný předpoklad jeho dostupnosti, konzistence a integrity jím poskytovaných dat. Avšak ani tento předpoklad nevylučuje v odůvodněných případech možnost provozovat tento systém na lokální stanici.

Důležitým prvkem, který nám umožní zvýšit efektivitu práce s dokumenty, jsou vlastní popisná data dokumentů. Může se jednat o jednoduchý metadatový popis vypovídající o obsahu dokumentu a usnadňující nám činnost při hledání relevantních dokumentů napříč všemi zdroji. Důležitá jsou bezesporu i pracovní data zachycující druh dokumentu, kdy byl uveřejněn a kým a další informace usnadňující práci s dokumenty. Pokusme se dále provést návrh vhodného navázání těchto dat na dokument spolu s výběrem vhodného dokumentového formátu, za jehož pomoci budou styčná data v konečné formě zaznamenána.

Výběr vhodného formátu dokumentu

Klíčový pro dokumentový systém a potažmo poradenský systém je formát, v němž budou ve výsledku dokumenty ukládány. Formát dokumentu má vliv nejen na možnou podporu strukturovanosti dokumentu, rychlost v něm vyhledávaných informací a v neposlední řadě taktéž i rychlost přístupu k samotnému dokumentu uloženého na disku počítače. Pokusme se v krátkosti shrnout, v jakých formách a jakými způsoby může být dokument uložen a vystihnout jejich nejvýznačnější charakteristické rysy.

I: *Formáty určené pro ukládání dat a informací*

Druh formátu	Klady/Zápory
TXT	formát vhodný jen pro popis a tvorbu zdrojů
DOC	licencovaný formát, paměťově náročný, netextový
PDF	vhodné jen pro dokumenty jenž se nebudou dále upravovat
HTML, xHTML	umožňuje oddělit data od zobrazené informace, snadnější přístup k dílčím částem
XML	podpora strukturovanosti uložených dat, rychlejší přístup, podpora dodatečné funkcionality

U každého informačního systému se setkáváme s velkým množstvím různorodých dat, která musíme předem určitým způsobem zpracovat. Některá data máme uložená ve formě dokumentů, nejinak tomu je i u systému poradenského. Zde se setkáváme s daty a informacemi, které na základě požadavku klienta (uživatele) musíme být schopni ve vhodné formě zveřejnit. Některá data jsou svou povahou velmi specifická a ve své aktuální hodnotě doplněna až v momentě zveřejnění dokumentu, tj. dokument je sestaven na základě zákaznického dotazu a kritická data[1] jsou doplněna až v momentě uveřejnění dokumentu.

Pokud v jednoduchosti vyhodnotíme formáty dokumentů uvedené v tabulce I můžeme dojít k následujícím závěrům. Formáty DOC a PDF jsou formáty netextových souborů, přičemž jejich formát je licencovaný, ostatní formáty jsou čistě textové. Je evidentní, že pro uložení dat a informací nám stačí použít formát textový, avšak pokud chceme zdroje i vhodnou formou prezentovat (myšleno v jiné formě než jednolitý neformátovaný text), jeví se jako vhodnou volbou formát HTML. Pokud však vyžadujeme od zdroje i určitou strukturovanost přesahující možnosti HTML, je jedinou volbou formát XML. XML umožňuje použít vhodně definované značky, jež nám usnadní vytváření strukturovanosti dokumentu a implementaci dalších technologií. Jedná se o formát čistě textový, pro výsledné zobrazení se bude muset použít vhodný zobrazovací styl.

Rozeberme si možnosti, jakou efektivní formou můžeme data ukládat a zveřejňovat, avšak již nyní zaveďme si úmluvu, že výslednou finální verzi dat a informací určenou ke zveřejnění budeme nazývat dokumentem. Povětšinou se bude jednat o vhodně prezentovanou textovou informaci získanou ze zdroje, doplněnou o aktuální informace, na níž je dále aplikován vhodný zobrazovací styl. Naproti tomu dokument, jež je tvořen zdrojovými daty, popřípadě vhodně strukturálně členěn, při zachování podmínky oddělení

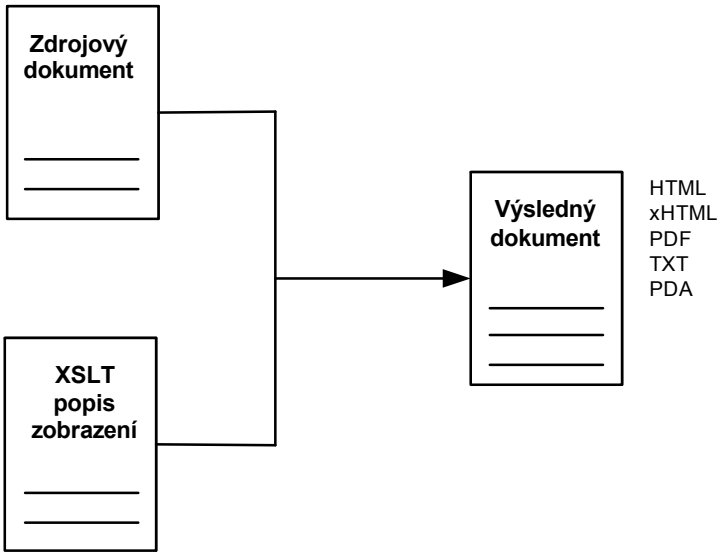
1 kritická data jsou data proměnná v čase, proto je vhodné jejich aktuální stav uveřejňovat až v momentě sestavení dokumentu

obsahu od výsledného zobrazení, nazvěme dokumentem zdrojovým. Pokusme se tyto pojmy zpřesnit.

Zdrojovým dokumentem *v informačním systému chápejme strukturovaný soubor obsahující vhodně členěná data, jež jsou zde uložena bez ohledu na výsledné zobrazení.*

Výstupním dokumentem, dále jen dokumentem, *v poradenském systému* ***chápejme*** *vhodně strukturovanou textovou informaci doplněnou o další objekty, na něž byl aplikován konkrétní zobrazovací styl a jež jsou určeny ke zveřejnění uživateli systému.*

Pokud použijeme výše popsaného přístupu oddělení dat od výsledné struktury prezentovaného dokumentu, je evidentní, že výsledný dokument se sestaví až v momentě prezentace na základě dílčího požadavku (požadavků), a to ve formě předepsané konkrétní výstupní šablonou. Je to velmi progresivní přístup umožňující vytisknout výslednou formu dokumentu až dle konkrétního aplikačního použití (prohlížení, prezentace, tisk, zobrazení na PDA). Tento přístup umožňuje zachování konzistence dat a předchází utváření duplicitních dokumentů, tedy stejných souborů rozdílných jen dle výsledného použití.



1: *Postup sestavení výsledného dokumentu*

Hledáme tedy formát souborů, jenž nám umožní uložit výchozí data, definovat (popsat) strukturu dokumentu v ryzí formě a to tak, abychom mohli členit dokument do dílčích částí, podle výsledného použití. Může se jednat např. o části jako úvod, rozbor problematiky, metodická východiska, vlastní řešení, závěr, doporučení, materiálové zdroje a to bez vlivu na výsledné zobrazení, které bude definovat zobrazovací styl.

XML dokumenty

Vezmeme-li v potaz zhodnocení tabulku I, je evidentní, že formáty souborů, jež těmto podmínkám vyhovují, jsou HTML (xHTML) a XML. Výhodnější však pro potřeby dokumentů je textový formát XML, v jehož rámci můžeme používat svoje vytvořené popisné značky umožňující nám definovat požadovanou strukturu dokumentu, popřípadě uložit další doplňující informace, jež ve výsledku nebudou zobrazeny a mají čistě popisnou úlohu.

Pokud budeme chtít, aby formát dokumentu umožňoval oddělení struktury dokumentu od jeho výsledného vzhledu, při podpoře přenositelnosti těchto dokumentů, jeví se jako nejpříhodnější právě formát XML, na něhož lze aplikovat speciální jazyk XSL podporující strukturovaný zápis dokumentů (Bradley, 2003). Takto vytvořený textový formát umožňuje po aplikaci vhodného stylu vytvořit adekvátní HTML výstup, který je klíčový, pokud chceme poradenský systém budovat na základě využití internetových technologií.

Je evidentní, že pro tvorbu tištěného výstupu není HTML výstup příliš vhodný, proto se jednoduše při požadavku tisku aplikuje na XML zdrojový dokument styl optimalizovaný pro tisk. Zde se jeví jako nejpříhodnější výstupní tiskový formát PDF, u něhož je zaručena velmi dobrá tisková přenositelnost. Výhodou pro volbu XSLT jako převodního filtru je také to, že podporuje převod do PDF (XSLT, 2007).

Tedy nutnou podmínkou, která musí být v dokumentovém systému splněna, je uveřejňovat jen takové dokumenty, jež splňují XML zápis. Počáteční režie při jejich vytváření je následně vykoupena úsporou času

při jejich dalším využívání. Tímto způsobem zvýšíme značně efektivitu prezentované informace (pro web i tisk) a podpoříme tak další přidružené služby na dokument navázané, za důležité lze považovat např. vyhledávání dokumentů na základě full textu, popřípadě sémantických spojitostí.

Pro převod dokumentů do XML podoby lze bezesporu, dle povahy dokumentu, použít i vhodně sestavené algoritmy, které ve spojení s adekvátním aplikačním prostředím usnadňují tuto konverzi. Tuto možnost blíže uvádíme v části s názvem Dokumentové šablony.

Na druhou stranu se v rámci informačního systému mohou vyskytovat i takové dokumentové zdroje, jež jsou jen doplňující přílohou dokumentů stávajících. Ty mohou zůstat v původním, standardizovaném formátu, nejlépe v PDF. V tomto případě se může jednat např. o různá nařízení vlády, zákony, popřípadě další dokumenty obdobné povahy, jež dle zaměření poradenského systému jsou také součástí zdrojů. I zde je však vhodné ohodnotit takový dokument popisnými daty a to tak, aby mohl být zařazen do množiny dokumentů, které umožňují dohledávání informací za pomoci klíčových slov zadaných uživateli.

Pokusme se dále rozebrat možnosti, jež komukoli přináší použití formátu XML a na něj navázané technologie při sestavování XML dokumentů, jejich publikování při využití konkrétních publikačních šablon a jakým způsobem nejefektivněji na takto vytvořený dokument navázat pracovní a popisná metadata.

Metadata dokumentů

Pokud přistoupíme ke kroku opatřit zdrojové dokumenty vhodnými metadaty, jež nám mohou pomoci při implementaci některých technologií, stavíme se před problém, jak tato metadata elegantně na dokument navázat. Pokud by se jednalo o HTML dokument, není tak problematické uložit daná data rovnou do určité částí HTML dokumentu, a to jako popisující komentáře. Pokud se jedná o dokumenty PDF nebo DOC, není tato možnost použitelná. Proto je vhodné navrhnout koncepci ukládání metadat u těchto formátů a to mimo výsledný dokument. Sice je nutná další režie na jejich správu a revidování, naproti tomu se pro práci s nimi nemusí procházet celý dokument a i výsledná manipulace je o hodně efektivnější. Jak již bylo naznačeno, můžeme rozlišit následující koncepce navázání metadat (bez vysvětlení dalšího vlivu na výslednou funkcionalitu):

- metadata uložená uvnitř struktury dokumentu,
- metadata uchována externě v databázi a navázána na daný dokument.

Oba dva přístupy popisují koncepci, ne však výslednou realizaci. Pro zjednodušení nerozlišujeme pracovní a popisná metadata, jsou-li potřeba. Je evidentní, že pracovní metadata (data zaznamenávající práci s fyzickým dokumentem) budou editována a referencována mnohem častěji než metadata popisná (popisující obsah dokumentu). Tato skutečnost vychází již z jejich povahy; pokud někdo reviduje dokument o nové skutečnosti a nezmění se povaha dokumentu, je potřeba zaznamenat skutečnosti spjaté s touto editací. Povaha popisných dat však může zůstat nezměněná. Na základě tohoto rozboru můžeme upřesnit, že pokud to dokument umožňuje, je vhodné popisná metadata uložit jako jeho součást a pracovní metadata vně (např. u dokumentů typu HTML), v opačném případě mohou být pracovní a popisná metadata uložena vně dokumentu a to i na jednom společném místě (dokumenty typu DOC, PDF, …).

Pokud poradenský systém vytváříme na základech obdobných HTML zápisu, můžeme popisná metadata integrovat dovnitř dokumentu již při jejich tvorbě. Obdobně tomu může být u již zmiňovaných XML dokumentů (Michael, 2006). Pracovní metadata o dokumentech jsou pak vytvářena kdykoliv, kdy je to potřebné, přičemž jsou vždy uložena mimo zdrojový dokument.

Pracovní metadata

Pro zjednodušení neuvažujme dokumenty sestavované až na základě definování požadavku od uživatele. Pod životním cyklem dokumentu chápejme průběh etapy zdrojového dokumentu od jeho vytvoření, přes uveřejnění, revidování, zneplatnění a po periodické zálohování, jež také musí být součástí funkcionality dokumentového systému. Je to poměrně dlouhá cesta skládající se z jednotlivých dílčích etap, pro jejichž definování, popřípadě identifikaci, v jaké etapě se daný dokument nachází, je nutné zavést vhodná metadata, s jejichž pomocí jsme tyto skutečnosti schopni vyhodnocovat a evidovat, neboť se jedná o metadata zachycující administrativu a práci s dokumenty; pojmenujme je jako w–metadata (pracovní metadata).

***Data** dokumentů umožňující zavést evidovanou práci s dokumenty a to včetně jejich uveřejňování, zneplatnění a zálohování, spolu s řízením přístupu k dokumentům nazývejme **pracovními metadaty**, či zkráceně „**w–metadaty**“.*

Pracovní metadata by měla umožňovat ukládat a uchovávat takové údaje, jež je vhodné počas života dokumentu evidovat a vyhodnocovat. Výběrem se jedná zejména o:

- informace spjaté s vytvořením dokumentu (autor, datum vytvoření, evidenční-registrační číslo dokumentu, číslo revize, datum revize),

- schopnost prohlásit dokument za neplatný a před uživateli ho skrýt,
- kontrola přístupu k dokumentu (kontrola přístupových práv, záznam ID uživatele, kdo dokument použil a druh akce, kterou na něj aplikoval, omezení některých akcí dle vhodně zvolených přístupových práv),
- možnost smazat dokument (přístupné jen autorovi, popřípadě uživateli s vyšším oprávněním),
- umožnit práci s dokumenty v rámci sítě (sdílení více uživateli, pokud u uveřejněného dokumentu dochází k revidování obsahu, je uzamčen pro jiný zápis),
- třídění dokumentů do organizovaných, tematicky orientovaných celků (studie, studijní podklady, projekty, formuláře, korespondence, …) usnadňujících vyhledávání napříč dokumenty,
- třídění dokumentů na základě pracovních údajů o dokumentu (datum vytvoření, autora dokumentu, počtu revizí, četnosti přístupů k dokumentu),
- vyhledávání v dokumentech za pomoci klíčových slov (full text), popřípadě na základě logické spojitosti (sémantické vyhledávání),
- automatická archivace dokumentů (předcházení ztrátám),
- vymezení přístupových práv pro jednotlivé uživatele (některé dokumenty má možnost shlédnout jen expert, operátor poradenského systému, jiné libovolný zákazník),
- zaznamenávat četnost požadavků adresovaných na daný dokument od klientů, popřípadě od konkrétního klienta.

Pracovní data jsou vytvářena na základě práce se zdrojovým dokumentem. Při každé jeho editaci, počínaje jeho vytvořením, zálohováním, uveřejněním, zneplatněním, se vždy provede záznam o tom, kdo tuto činnost provedl, spolu s časem změny a ve spojení s dalšími nutnými informacemi. Tuto činnost lze zautomatizovat za pomoci prováděcího skriptu dokumentového systému. Poté vždy na základě iniciačního podnětu dojde k úpravě pracovního záznamu w-metadat navázaných na dokument.

Vytváření evidenčních záznamů zachycující změny je důležitý krok podporující politiku práce s dokumenty a bezesporu nutný počin nutný k odhalování havárií a chyb daného poradenského, respektive informačního systému.

Popisná data dokumentů

Metadaty v tomto smyslu rozumějme takové údaje, která nám daný dokument po obsahové stránce popíší a věcně ohodnotí. Je to nezbytný předpoklad zavedení sémantického vyhledávání nad dokumenty, popřípadě zefektivnění vyhledávání full textového. Je vždy snahou rozdělovat dokument do více logických částí, které lze popsat nezávisle na sobě (může se jednat např. o úvod, abstrakt, seznámení s problematikou, řešení problému, postup řešení, shrnutí, závěr, zdroje, odkazy, ...). Právě popis dílčích částí dokumentů vystihuje podstatu použití popisných dat. V závislosti na formě dokumentu, zvláště při oddělení struktury od vzhledu, jak je tomu možno u XML, nemusí být metadata pro uživatele ve výsledné podobě dokumentů vůbec vizuálně přístupná. To je hlavní přínos přístupu oddělení obsahu zdrojového dokumentu od zobrazených dat a informací v uveřejněném dokumentu. Tedy výsledný dokument je sestaven na „na míru“ dle požadavku zadavatele dotazu, tedy uživatele a to bez zbytečných dat jeho zadání neodpovídajících.

Popisná metadata *dokumentů jsou určena pro* ***ohodnocení*** *obsahové náplně jednotlivých částí zdrojových dokumentů a umožňují ve svém důsledku zavést efektivní vyhledávání, na základě této skutečnosti je zvěme „****s-metadaty****“.*

Pokud budeme vytvářet s-metadata uvnitř XML dokumentů, bude se jednat o ucelené popisné úseky tvořené z výrazů vystihující vypovídajícím způsobem obsahovou náplň dané části dokumentu. Pod pojmem část dokumentu chápejme dílčí oblast zdrojového dokumentu definovanou za pomoci strukturálních značek. Tento přístup umožní vytvořit dílčí popisy dokumentu ve spojení s možností omezit vyhledávání v dokumentech na pevně definované oblasti.

Aby bylo možné sjednotit popisné výrazy vystihující podstatu částí dokumentů, je vhodné, aby vycházely z oborově orientovaného slovníku zaměřeného na problematiku, jako zveřejněný a ve výsledku publikovaný text. Přestože poradenský systém je systémem uzavřeným, lze tímto přístupem značně zefektivnit práci s dokumenty a to tak, že se nebude pracovat s celým textem dané části dokumentu, ale jen s dílčími s-metadaty. Přístup samozřejmě lze zobecnit na jakýkoliv systém. V rámci XML budou takto tvořená popisná data oddělena vhodným značkováním vycházející z OWL slovníku (OWL, 2007).

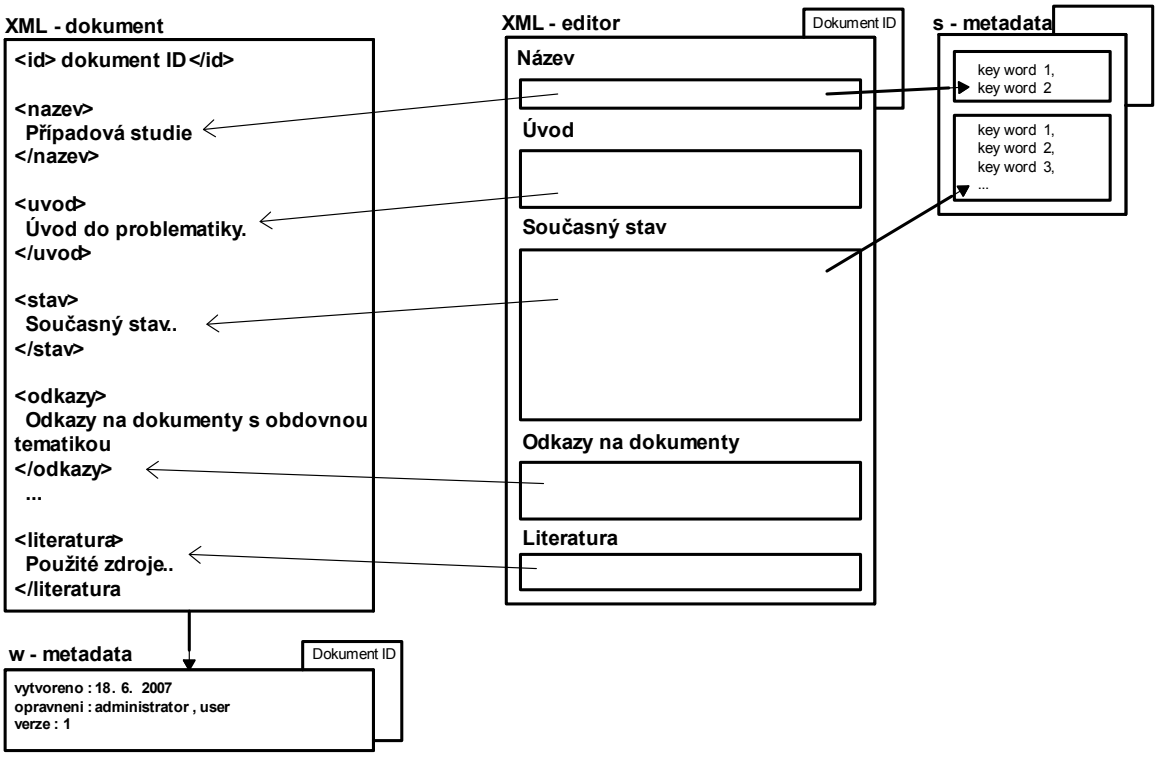
Dokumentové šablony

Jelikož se v dokumentovém systému a obecně v jakémkoliv informačním systému používá několik standardizovaných typových dokumentů, je vhodné zavést předem připravené šablony dokumentů, jež ve výsledku usnadňují práci při finálním sestavování dokumentů. Tento přístup též umožní sjednotit vzhled všech dokumentových sestav a usnadní jejich správu.

Pro podporu efektivního vytváření zdrojových dokumentů ukládaných v XML je vhodné vytvořit editor; bude umožňovat psaní částí dokumentů v roz-

hraní ne nepodobnému klasickým kancelářským produktům. Může se jednat o vyplňování formuláře s definovanými oblastmi odpovídající částem v dokumentu XML, kde po vhodné konverzi budou uloženy.

Je zřejmé, že textovou podobu zdrojového dokumentu XML lze vytvářet v jakémkoliv textovém editoru umožňujícím ukládat čistý text, bez definování vlastností textu, avšak v tomto případě je nutné takto vytvořený text doplnit o strukturální značky a dále o oddíly s popisnými s-metadaty. Neboť ne vždy lze zaručit vkládání či přepis dokumentů osobou strukturální problematiky dokumentů znalou, je volba uživatelsky přívětivého editoru dobrou volbou.



II: *Využití šablon dokumentů a vazba metadat na dokument*

Vlastní kapitolou je vytváření popisných metadat dokumentu. Je vhodné je tvořit pro každou část dokumentu zvlášť, neboť ne o všechny části má uživatel při vyhledávání zájem. Ve výsledku lze takto docílit zefektivnění celého procesu hledání a zlepšit tak jeho produktivitu. To se projeví zejména vzhledem k dohledaným informacím v jednotlivých částech dokumentu napříč všemi dokumenty dokumentového systému. Tato popisná metadata vytváří buď tvůrce dokumentu nebo administrátor této sekce poradenského systému a to z toho důvodu, aby byla zachována jednotnost zápisu a zároveň, aby zápis odpovídal realitě. Pokud by struktura XML dokumentu byla velmi mohutná[2] a docházelo by při procesu vyhledávání dokumentů k časovým průtahům, je možné uložit popisná data externě, obdobně jako data pracovní. Výsledkem procesu hledání je dokument, popřípadě část dokumentu odpovídající kritériím zadaným do vyhledavače. Dle relevantnosti k danému vyhledávacímu dotazu či výběru akce dojde k sestavení výsledného dokumentu určeného k zobrazení, popřípadě k setřídění dokumentů dle četnosti výskytu klíčových slov nebo klíčových spojitostí.

Vyhledávání informací v dokumentech

Integrace vyhledávacích funkcí do informačního potažmo poradenského systému je velmi podstatná a to zejména v souladu se zvyšováním uživatelské přívětivosti informačních systémů. Ve výsledku umožní uživateli vyhledat relevantní informace napříč všemi

2 zde mohutností dokumentu rozumějme jeho velikost danou počtem textových řádků

dokumenty a to ve zlomku času oproti manuelnímu dohledávání. Můžeme volit implementaci vyhledávání například na bázi full textu, kdy se dohledávají řetězce či slova obsahující podřetězce, které jsou shodné se zadaným výrazem nebo na bázi sémantického vyhledávání.

Na vhodné implementaci full textové technologie je přímo závislá i rychlost prohledávání dokumentů vedoucí ve svém důsledku k nalezení relevantní informace. Časové hledisko je v tomto případě klíčovým prvkem ovlivňujícím použitelnost.

Pokud máme uzavřený systém dokumentů (rozumějme v tomto pojetí s konečným počtem dokumentů), je výhodnější zavést sémantickou variantu vyhledávání. U tohoto typu vyhledávání nedochází k dohledávání čistě klíčových slov, ale spíše obsahové náplně a logických spojitostí se zadaným klíčovým slovem a slov po obsahové stránce podobných.

Nutnou podmínkou pro zavedení sémantického vyhledávání nad dokumenty je zavedení optimalizovaných dokumentů, tedy takových dokumentů, jež splňují alespoň následující kritéria:

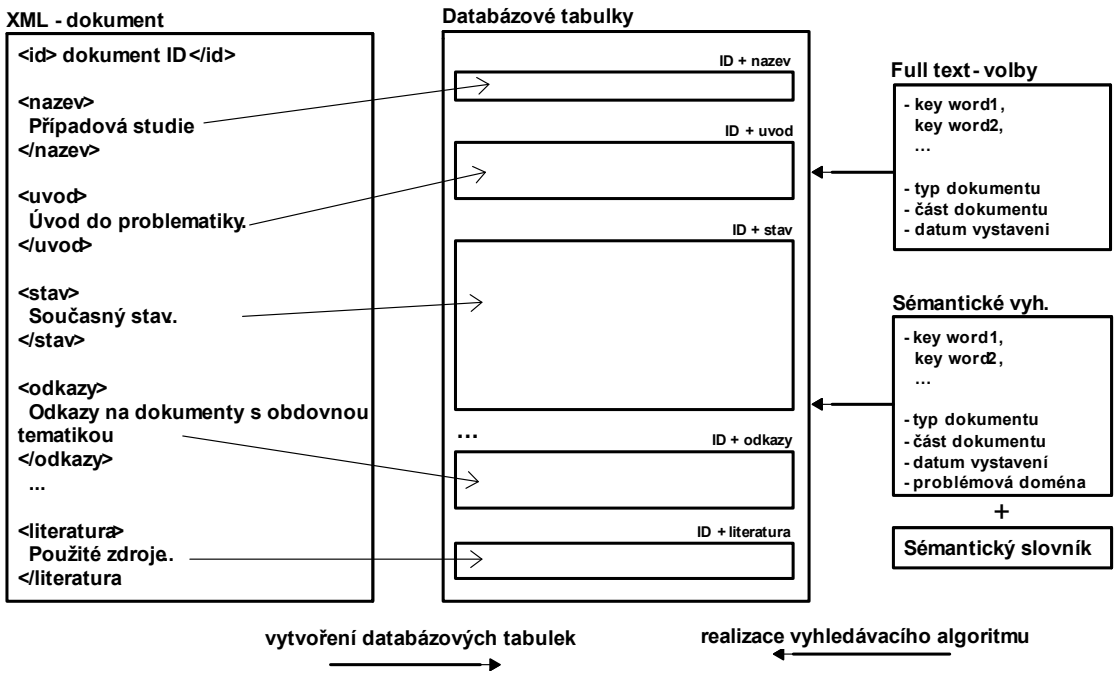
- vhodně členěná struktura
- bylo provedeno ohodnocení dokumentu, části dokumentu.

Těmito kritérii, pokud budou splněna, můžeme značně zefektivnit práci s dokumenty. Pokud si tyto požadavky podrobněji rozebereme, jako nejpříhodnější se bezesporu jeví formát XML pro vytváření zdrojových dokumentů. Uvažujme dokument vstupující do procesu dohledávání na základě full textu, popřípadě sémantických spojitostí, jen jako XML modelový dokument.

Velmi výhodné je v rámci popisných dat dokumentu uchovávat informaci o tom, o jaký typ dokumentu se jedná a to zejména ve spojitosti s problémovou doménou, jíž se poradenský systém zabývá.

Pokud tuto informaci použijeme při zadávání parametrů u vyhledávacího algoritmu ve spojení například s časovým údajem, kdy se platnost daného dokumentu vyskytla (nová vyhláška, studie, …) s upřesněním na jakou část zdrojového dokumentu má být brán zřetel, dojde k podstatné redukci relevantních dokumentů daná kritéria splňující a zvýšíme tak přesnost dohledávání.

Metodou vyhledávání rozumějme způsob, s jehož pomocí realizujeme vlastní hledání a následně vyhodnocení získaných údajů. Podstatou je procházení jednotlivých slov v oblastech dat zdrojového dokumentu a hledání spojitosti se slovy (slovními řetězci), která byla zadána do vstupů vyhledávacího enginu v rámci vyhledávání full textem. Další variantou je hledání na bázi sémantických spojitostí, kdy se prohledávají oblasti zdrojového dokumentu s s-metadaty. Výstupem pak jsou dokumenty (popřípadě jejich části), které zadaným podmínkám nejvíce vyhovují.



III: *Implementace vyhledávacího algoritmu*

Na základě sémantického vyhledávání získáme lepší orientaci v datech dokumentů a to za pomoci vyhodnocení jejich logické struktury a následně sémantických spojitostí. Tento přístup vychází z návrhu sémantického webu, kde však z důvodu nehomogenity jednotlivců v současné době nelze tuto myšlenku plně implementovat. Využijeme-li však tyto přístupy u uzavřeného dokumentového systému, jímž poradenský systém bezesporu je, potom bez obtíží tuto omezující podmínku překleneme. Takto můžeme překonat zejména některá slabá místa původní myšlenky, zejména v oblasti popisu dokumentových zdrojů a sestavit takto algoritmus umožňující efektivní vyhledávání v rámci dokumentů u poradenského systému.

Jak již bylo zmíněno, pro sémantické vyhledávání je klíčová problémová doména, jíž se v tomto případě poradenský systém zabývá. Stěžejním prvkem je zachycení sémantických spojitostí, tedy spárování dohledávaných klíčových slov s příbuznými popisnými výrazy. Toho můžeme docílit pomocí přidruženého Sémantického slovníku, jenž bude využit při vlastním vyhledávání.

Pokud chceme zefektivnit přístup k jednotlivým datovým položkám, popřípadě k s-metadatům a w-metadatům zdrojových dokumentů, je možnost uložit tento výchozí soubor zapsaný za pomoci XML značek do databáze. Je výhodné vytvořit databázové tabulky identifikované ID zdrojového dokumentu, popřípadě jeho částí. Následné hledání klíčových slov v jednotlivých tabulkách je o mnoho rychlejší než v původním dokumentu a to nejenom z důvodu procházení kratších bloků řetězců, ale zejména na základě podpory zpracování řetězců, kterou disponuje většina současných databází, jako jsou MySQL či Oracle. Jedná se o tzv. podporu full textové indexace, jež značně zrychlí vlastní vyhledávání klíčových skutečností a subřetězců v rámci záznamů.

Klíčové je za jakékoliv situace snižování výsledných nákladů celého systému, proto se jako vhodná volba při výběru databáze jeví výběr databáze MySQL, jež je volně dostupná a s plnou uživatelskou podporou. Je to vhodná volba zejména pro malé a střední firmy (Zajíc, 2005).

DISKUSE

Klíčovým prvkem každého poradenského systému je práce s dokumenty a právě na tuto část by měl být kladen velký důraz, neboť správná volba dokumentového formátu a členění dokumentu má vliv nejenom na rychlost odezvy při práci s dokumenty a i na implementaci dalších technologií souborový-dokumentový přístup využívající.

Volba formátu dokumentů nejenom ovlivní práci s vlastními daty, má také podstatný vliv na členění dokumentu a ovlivňuje i efektivitu vlastního vyhledávání aplikovaného nad dokumenty. Proto se výběr formátu XML pro uložení datové základny dokumentů se jeví bezesporu jako dobrá volba, zejména z jeho podstaty oddělení obsahové části od prezentované informace. Takto můžeme docílit použití jednoho dokumentového zdroje pro velké množství rozdílných výstupů, aniž bychom vytvářeli duplicitu dat, využitím vhodné výstupní šablony. To je klíčový rozdíl od ostatních formátů, jež touto možností ve formě podpory standardu formátu nedisponují. Podobnými úskalími publikování na webu ve spojení s podporou strukturovanosti dokumentu za pomoci XML (Fukala, 2003).

Byla zde popsána i možnost popisu dokumentů s pomocí s-metadat a podpora práce s dokumenty v součinnosti s w-metadaty, jenž nám uchovávají záznamy o režijní práci s dokumenty a ve výsledku umožňují zefektivnit práci a to nejenom ve spojení s implementací dalších technologií. V tomto případě jsou metadata dílčími prvky systému, které zlepší správu dokumentového systému a implementaci technologií s dokumenty souvisejícími. Je to další přístup usnadňující celkovou práci se systémem.

Jako nejefektivnější forma uložení bylo vyhodnoceno uložení XML strukturovaných dokumentů a jejich dílčích částí ve formě záznamů v databázi. Při využití volně dostupné databáze MySQL je to značně progresivní řešení zvyšující nejenom efektivitu přístupu k jednotlivým oblastem původně zavedeného XML zdrojového dokumentu, ale zejména zrychlení veškerých vyhledávacích algoritmů. Tento přístup značně zvyšuje přenositelnost a taktéž kompaktnost výsledného řešení. Pokud budeme provozovat systém čítající velké množství dokumentů, je to z hlediska délky obsluhy jednotlivých transakcí jediná schůdná cesta, při současném zachování použitelnosti tohoto informačního systému.

SOUHRN

V článku byla rozebrána problematika tvorby optimálních dokumentů pro komputerizovaný poradenský systém a možnosti, jež nám současné technologie k jeho tvorbě poskytují.
Jako nejvhodnější varianta se jeví tvorba dokumentu ve formě zápisu XML, jenž nám umožní oddělit obsahovou strukturu dokumentu od jeho výsledné uveřejněné formy a to za pomoci vhodného výstupního stylu.

Na základě těchto skutečností byla specifikována vhodná struktura zdrojového dokumentu s možnostmi převodu informací do podoby zdrojového dokumentu. Dále byly naznačeny možnosti jak vytvářet popisná a zdrojová data spolu jejich vazbou na dokument, jež nám umožní zvýšit produktivitu práce s dokumenty.

Byly také naznačeny možnosti v implementaci vyhledávání nad dokumenty v uzavřeném (poradenském) systému. Jsou zde ozřejměny v dnešní době používané metody ve spojení s možností specifikovat část dokumentu, jež bude procházen s ohledem na hledaný výraz a typ prohledávaného dokumentu. Jak již bylo zmíněno, je poradenský systém uzavřený systém a to je výhodné pro implementaci vyhledávacích technologií zavedených nad dokumenty.

Taktéž byla uvedena myšlenka zvýšení časové efektivity přístupu k jednotlivým částem dokumentu a podpora při vyhledávání napříč všemi dokumenty, popřípadě napříč konkrétními částmi dokumentu a to za pomoci uložení zdrojového dokumentu ne ve formě dílčího souboru, ale záznamů v databázi.

Práce s dokumenty je ustředním a limitním faktorem výsledné efektivity každého informačního systému, jenž vytváří, uchovává a reviduje dokumenty při jeho činnosti vytvořené. Volba vhodného formátu dokumentů, ve spojení s daty, jež dokument popisují, může značně ovlivnit rychlost výsledné práce s těmito dokumenty.

Již v době návrhu systému bychom se měli zabývat typy dokumentů, se kterými se bude pracovat. Pokud se zpracovávají dokumenty stejných typů, je vhodné při jejich vytváření dodržet jednotný popisný vzhled. Toho lze docílit například za pomoci využití vzorových šablon dokumentů.

Neméně důležitý je i způsob, jehož pomocí budou dokumenty uchovávány v rámci provozovaného systému. Toto hledisko je klíčové zejména pokud je prezentování a práce s dokumenty u tohoto systému stěžejní. Návrh členění struktury a způsob ukládání dokumentu je ovlivněno také tím, jaké operace v rámci systému na nich plánujeme provádět. Velký vliv na výslednou strukturu má například implementace efektivního vyhledávacího algoritmu, a to ať již ve své fulltextové či sémantické variantě.

Tento materiál si neklade za cíl provést komplexní řešení dokumentového systému v rámci informačního systému, ale na vhodných přístupech a srovnáních ukázat efektivní cestu v souladu s novými technologiemi, jak takovýto systém efektivně realizovat.

dokument, dokumentový systém, struktura dokumentu, šablony dokumentu

Článek vznikl v rámci řešení VZ MSM 6215648904/03/03.

LITERATURA

BRADLEY, N.: *XML – kompletní průvodce*, Grada Publishing Praha 2003, 536 stran, ISBN 80-7169-949-7.

FUKALA, M.*: Publikování dokumentů na webu*. Bulletin CVT : Občasník Centra výpočetní techniky [online]. 2003, roč. 4., č. 3. [cit. 2007-06-06]. VŠB Ostrava. Dostupný z WWW: <http://www1.vsb.cz/cvt/bulletin/b3/clanek/clanek.html>. ISSN 1213-9904.

MALO, R., RASZKOVÁ, M.: *Automatizace tvorby podnikových dokumentů*. In Firma a konkurenční prostředí 2006 – Sekce 7. IS/IT a konkurenceschopnost podniků. Brno: KONVOJ, spol. s r. o., 2006, s. 53–60. ISBN 80-7302-097-1.

Michael J. Young.: *XML – Krok za krokem*, Computer Press, 2006; ISBN: 80-251-1070-2.

OWL Web Ontology Language [online]. W3C, c1994–2007 , 2007/07/10 [cit. 2007-06-06]. HTML. Text v angličtině. Dostupný z WWW: <http://www.w3.org/TR/owl-features/>.

XSLT Tutorial [online]. 1999–2007 [cit. 2007-06-06]. Dostupný z WWW: <http://www.w3schools.com/xsl/>.

ZAJÍC, P.: *MySQL (38) – Fulltext a praxe : Příklady na použití fulltextu v MySQL*. [online]. 2005 [cit. 2007-06-06]. Dostupný z WWW: <http://www.linuxsoft.cz/article.php?id_article=960>.

Adresa

Ing. Oldřich Trenz, Ústav informatiky, Mendelova zemědělská a lesnická univerzita v Brně, Zemědělská 1, 613 00 Brno, Česká republika