AMPHORA - NÁSTROJ PRO INDEXOVÁNÍ WEBOVÝCH STRÁNEK.

Václav Snášel, Jiří Dvorský, Petr Šaloun, Daniela Ďuráková

VŠB – Technická univerzita, 17. listopadu 15, 708 33 Ostrava

Abstract

Textová databáze AmphorA poskytuje informace s přidanou hodnotou týkající se informačních zdrojů na internetu. Propojení textové databáze virtuální knihovnou vytváří výkonný vyhledávač využívající spojení moderní technologie a tradičního popisu dokumentu.

1. Úvod

Při tvorbě textové databáze AmphorA jsme se původně zabývali indexováním dokumentům, které jejich autoři připravili v textovém editoru, byl použit MS Word. Tyto dokumenty byly dostupné na lokálním pevném disku. Pro zvýšení informační hodnoty dodávaného produktu, bylo rozhodnuto zařadit do systému i zdroje dostupné na internetu a do textové databáze přidat i texty z různých URL. V tomto okamžiku jsme museli řešit dvě otázky:

1. kde vzít seznam vhodných URL adres,
2. stažení a uložení již stažených WWW stránek.

Seznam smysluplných URL adres nám zcela logicky poskytla virtuální knihovna, dále zkráceně VK, jejímž cílem je shromažďovat hodnotné URL adresy. Stahování a ukládání WWW stránek bylo vyřešeno v rámci systému AmphorA. Textová databáze AmphorA obsahuje výkonný nástroj na stahování internetových stránek. Tyto stránky se stahují pouze v textovém formátu, protože ten je pro fulltextový vyhledávač zajímavý. Stažené stránky jsou ukládány ve formě XML dokumentů. Uložení stažených WWW stránek ve formě XML dokumentu umožňuje získat další informace pro indexování a prezentaci vyhledaného dokumentu.

Stažená stránka je uložena ve tvaru XML dokumentu, který umožňuje k textu WWW stránky připojit další informace viz následující příklad.

Příklad 1.

```
<?xml version="1.0" encoding="Windows-1250" ?>
- <DOCUMENT>
  - <HEADER>
      <HEADLINE>AGRITEC s. r. o., Šumperk</HEADLINE>
      <SOURCE>http://www.agritec.cz/</SOURCE>
      <DATE>11.1.2001</DATE>
      <TIME>2:5:55</TIME>
      <DESCRIPTION></DESCRIPTION>
      <STOPWORD>Ltd</STOPWORDS>
      </HEADER>
  - <DOC>
```

```
- <TEXT>
    <P>AGRITEC s. r. o., Šumperk</P>
    <P>http://www.vukrom.cz</P>
    <P>výzkum, šlechtění a služby s. r. o.,</P>
    <P>Šumperk</P>
    <P>AGRITEC</P>
    <P>Research, Breeding Services Ltd.,</P>
    <P>Šumperk, the Czech Republic</P>
    <P></P>
    <P>výzkum | šlechtění |</P>
    <P>služby | profil firmy |</P>
    <P>e-mail | mapa</P>
    <P>on-line databáze genových zdrojů luskovin</P>
  </TEXT>
 </DOC>
</DOCUMENT>
```

AmphorA obsahuje pouze jednoduchý nástroj pro editaci indexovaných www stránek, proto je vhodné propojit Amphoru s jiným nástrojem, umožňujícím pohodlnou správu URL odkazů, jakou je virtuální knihovna.

1.1 Virtuální knihovny

Pro usnadnění orientace v prostoru WWW (World Wide Web) byly vytvořeny vyhledávací služby, které jsou velice oblíbené a které jsou využívané většinou uživatelů Internetu. Virtuální knihovny spadají do kategorie předmětově orientovaných vyhledávacích služeb. V současnosti většina z nich nabízí i rozhraní umožňující hledání prostřednictvím klíčových slov. V tomto článku popisujeme spojení VK s plnotextovým systémem. Databáze VK jsou vytvářeny odborníky z řad profesionálních informačních pracovníků a knihovníků, kteří výběrem kvalitních zdrojů, připojením popisu a hodnocení podle přesně stanovených kritérií vytvářejí přidanou hodnotu poskytovanou virtuální knihovnou.

Informační zdroje ve virtuálních knihovnách jsou zpravidla organizovány logicky, obdobně jako je tomu u tradičních informačních zdrojů, na jejichž tvorbě se informační profesionálové podílejí. Kromě kvalitního pořádacího systému nabízejí tyto služby uživatelům Internetu také odkazy na kvalitní informační zdroje. Definici pojmu VK obsahuje např. [1].

2. Aplikace Virtuální knihovna

Námi vytvořená virtuální knihovna uchovává záznamy obsahující základní informace o každém ze zadaných zdrojů, tj. jeho název, autory, URL adresu a jazyk spolu se stručným popisem, deskriptory a klíčovými slovy. Každý zdroj je jednoznačně určen svojí URL adresou. Vyhledávací služba virtuální knihovny je schopna poskytnout informace o všech zdrojích, které odpovídají vyhledávací podmínce.

Pro usnadnění přidávání, vyhledávání, modifikace, rušení a pro kontrolu aktuálnosti záznamů jsou k dispozici podpůrné nástroje. Celý databázový systém je přístupný z kteréhokoliv počítače připojeného k Internetu.

Propojení virtuální knihovny s textovou databází AmphorA přináší uživateli další účinné možnosti při vyhledávání a zpracování elektronických informačních zdrojů v prostředí WWW. Propojení je definováno datově s využitím standardu XML.

3. Propojení VK a AmphorA

Plnotextový nástroj AmphorA je s VK provázán výstupem v XML formátu. Systém AmphorA jsme popsali v [4]. Výstup v hlavičce určuje vlastnosti záznamu VK, mezi značkami <TEXT> pak obsahuje text odkazovaného dokumentu. Dokument obsahuje informaci o datu a času, umístění, klíčových slovech či stop-slovech. Umístění rozhoduje o případném budoucím přístupu k dokumentu. Lokální soubory budou dostupné jen v rámci intranetu.

Textová část XML dokumentu může obsahovat celý text libovolného dokumentu. To nám při plnotextovém zpracování dává jistotu, že můžeme vyhledávat jak podle obsahu dokumentu, tak využít obsah záznamu položky VK a vyhledávat tedy s využitím přidané hodnoty záznamu ve VK.

Dokument zpracovaný VK a Amphorou může být:
1. plně k dispozici, umístění v Intranetu,
2. k dispozici na Internetu, propojení jen přes URL – po zaindexování zůstala jen hlavička, textový obsah je odstraněn;
3. odkaz je vyhledávacím nástrojem popsaným knihovníkem, další vyhledávání pak typicky využívá možností odkazovaného serveru (např. yahoo.com, www.springer.de apod.)

Obsah XML (viz Příklad 1.) dokumentu je v systému AmphorA využíván v následujících subsystémech:

- vyhledávací,
- prezentační.

Vyhledávací subsystém využívá informace ze sekcí DESCRIPTION a STOPWORDS. Tyto informace jsou v průběhu indexování připojeny k indexu dokumentu.

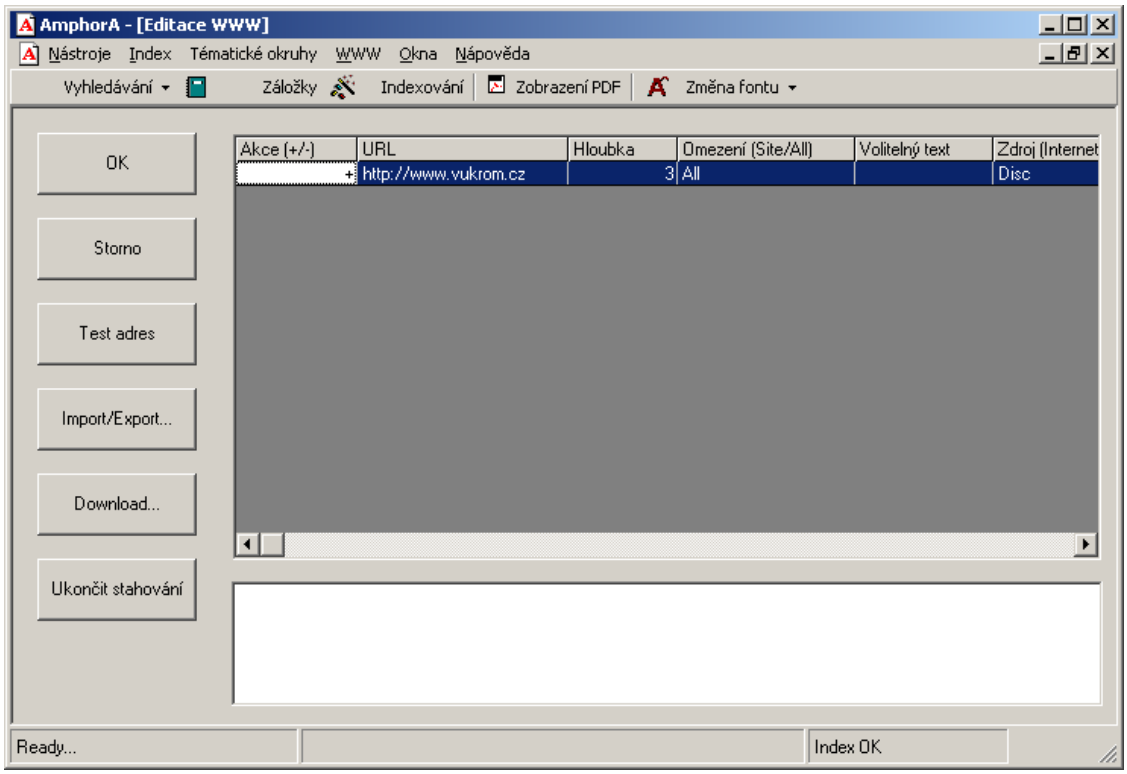
Prezentační subsystém využívá informace ze sekcí HEADLINE, SOURCE, TEXT.

- HEADLINE je vypsán do hlavičky zobrazeného dokumentu.
- SOURCE umožňuje zobrazit zdrojový dokument.
- TEXT obsahuje textový tvar dokumentu ve kterém jsou vyznačeny slova relevantní pro daný dotaz. Položka TEXT může být prázdná, v tomto případě je jako vyhledaný dokument prezentován dokument z položky SOURCE.

4. Indexování www stránek

Další možností propojení VK a systému AmphorA je možnost indexování www stránek. Indexování WWW stánek je možno provádět pomocí menu pro editaci WWW.

Vzhled aplikace ukazuje obrázek 1. Položka menu „Editace WWW adres“ slouží k editaci seznamu adres internetových serverů. Po zadání adres je možné automaticky stahovat stránky z uvedených serverů. V tabulce specifikujeme URL adresy, hloubku stromu stránek kam až se mají zkoumat odkazy ve stahovaných stránkách (hloubka nula znamená jen v tabulce specifikovanou stránku), omezení „Site“ tj. stahovat stránky jen v rámci tohoto serveru, „All“ stahovat i stránky z odkazů mimo daný server.



Obr. 1: Rozhraní programu AmphorA pro virtuální knihovnu.

Druhá položka v menu WWW je „Stoplist WWW adres“. V tomto okně je možno specifikovat URL adresy ze kterých se nemají stahovat žádané stránky, jako příklad si můžeme představit různé internetovské vyhledávače. Jejich stránky obsahují tisíce a tisíce odkazů prakticky kamkoliv a stahování takového množství stránek by zcela znehodnotilo ostatní data. Při stahování WWW stránek jsou jednotlivé odkazy na něž program ve stránkách narazí filtrovány přes tento seznam.

Položka „Volitelný text“ obsahuje seznam deskriptorů a stop slov jak bylo popsáno v předcházejících částech.

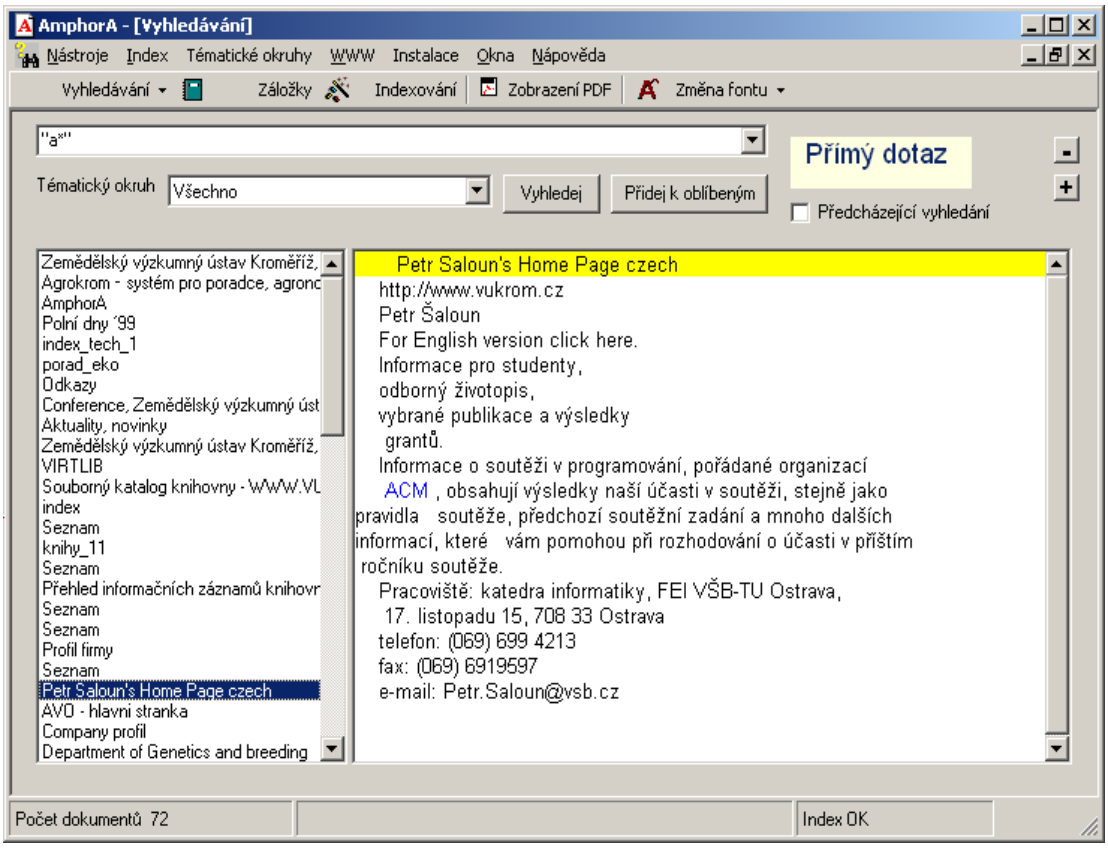
Tlačítko Import/Export umožňuje výměnu dat s virtuální knihovnou.
Výsledkem stahování dokumentů jsou XML dokumenty jejichž struktura byla popsána v předcházející části.

Na obrázku 2 uvádíme ukázku vyhledání v zaindexovaných www stránkách.

5. Závěr

V navrženém a realizovaném databázovém systému určeném ke zpracování elektronických dat byly spojeny výhody virtuálních knihoven (záznamy o kvalitních zdrojích informací, s odborným popisem dodaným knihovníkem) spolu s výhodami vyhledávacích strojů (automatické vyhledávání) a textových systémů. Systém dále obsahuje automatickou kontrolu aktuálnosti dat.

Popsaná VK je z větší části provozována Zemědělským výzkumným ústavem Kroměříž, s.r.o. Tato práce byla řešena na katedře informatiky FEI VŠB–TU Ostrava jako součást grantu MŠMT INFRA2 č. LB98227. Tento příspěvek vznikl za částečné podpory grantu číslo 201/00/1031 Grantové agentury ČR.



Obr. 2: Vyhledávání ve WWW stránkách.

Použitá literatura a www odkazy

1. Tkačíková, Daniela. Vyhledávací nástroje – klíč ke zdrojům Internetu [online]. [cit. 20. 03. 2001]. Dostupné na World Wide Web: <http://knihovna.vsb.cz/>.
2. Chudoba, Petr. Virtuální knihovna. Diplomová práce, FEI VŠB—TU Ostrava, 1999.
3. Pokorný Jaroslav, SNÁŠEL Václav, HÚSEK J. Dokumentografické informační systémy. Karolinum, skriptum MFF UK Praha, Praha 1998, ISBN 80-7184-764-X. 158 stran.
4. Snášel Václav, DVORSKÝ Jiří, ŠALOUN Petr, ĎURÁKOVÁ Daniela. Prostředky pro zpřístupnění a vyhledávání textových informací. Sborník z konference Tvorba softwaru 2000, ISBN 80-85988-49-6, s. 173-181.
5. http://www.vukrom.cz
6. Snášel Václav, DVORSKÝ Jiří, ŠALOUN Petr, ĎURÁKOVÁ Daniela. Propojení virtuální knihovny s textovou databází AmphorA. AKP 2001, v tisku.