AiP Beroun s.r.o., http://www.aipberoun.cz, tel.: +420 311 611 237, fax: +420 311 611 238
Talichova 807, Beroun, 266 01; IČO: 25778943, DIČ: 25778943

ZDOKONALENÍ VIRTUÁLNÍHO BADATELSKÉHO PROSTŘEDÍ MANUSCRIPTORIA - VYUŽITÍ SROVNÁVÁNÍ FULLTEXTŮ PRO VYHLEDÁVÁNÍ V MANUSCRIPTORIU

Poskytnutí Konzultací

Zpráva ke smlouvě Smlouva o spolupráci ve výzkumu a vývoji

verze 1.0

AiP Beroun, autor Ing. Tomáš Psohlavec

Obsah

1 Úvod o dokumentu

AiP Beroun uzavřela s Národní knihovnou České republiky Smlouvu o spolupráci ve výzkumu a vývoji: Zdokonalení virtuálního badatelského prostředí Manuscriptoria - využití srovnávání FullTEXTů pro vyhledávání v Manuscriptoriu.

1.1 Účel

Tento dokument zpravuje čtenáře o konzultacích poskytnutých ze strany AIP Zadavateli.

1.2 Termíny a konvence

Termíny a konvence použité v tomto dokumentu, pokud zde nejsou přímo vysvětleny, jsou popsány a definovány v dokumentu [1].

1.3 Reference

V dokumentu se odkazujeme na následující literaturu:

[1] Manuscriptorium v.2.0 – analýza systému, AiP Beroun 2004

[2] SRU – Search/Retrieval via URL
http://www.loc.gov/standards/sru/

2 Zpráva o poskytnutých konzultacích

NK ČR v souladu se smlouvou požádala celkem o 3 konzultace směřující k formulaci uživatelských požadavků na nové funkce MnS související s porovnáváním fulltextů. Dvě čtyřhodinové konzultace byly pracovníkům NK ČR poskytnuty na pracovišti AiP Beroun, další čtyřhodinová konzultace proběhla na půdě NK ČR. Ze strany AiP Beroun se konzultací zúčastnil vždy nejméně jeden konzultant a jeden programátor.

Před zahájením konzultací pracovníci NKČR informovali zaměstnance AiP Beroun o aktuálním rozsahu funkčnosti aplikace porovnávání fulltextů.

Konzultace technického rázu byla následně poskytnuta i subdodavateli NK ČR, který vytváří vlastní aplikaci pro rozpoznávání fulltextů. S ohledem na výsledky předchozích konzultací byla jako forma API pro aplikaci porovnávání fulltextů doporučena technologie web services. Funkčnost poskytovaná přes API by měla být identická s funkčností nabízenou uživatelským interface dané aplikace.

Během konzultací byly posuzovány možnosti využití aplikace porovnávání fulltextů v MnS. Konzultovány byly jednak otázky uživatelské funkčnosti a dále také problematika změny work-flow systému MnS, které zavedení nové služby případně vyvolá v souvislosti se zpracováním dat fulltextů a jejich správou.

Konkrétní zvažované možnosti implementace se lišily v závislosti na aktuálních požadavcích postupně vznesených ze strany Zadavetele. Během konzultací tak byly popsány tři odlišné možné varianty řešení. Odlišnosti spočívají zejména ve změně základních předpokladů týkajících se správy dat - fulltextů:

- Varianta A: MnS i databáze aplikace pro porovnávání fulltextů pracují s totožným obsahem; databáze aplikace pro porovnávání fulltextů je plněna fulltexty pocházejícími z exportů Správního Systému MnS.
- Variana B: MnS a databáze aplikace pro porovnávání fulltextů nepracují s totožným obsahem. Oba zdroje jsou plněny fulltexty a spravovány nezávisle na sobě, lze předpokládat určitý průnik.
- Varianta C: MnS a databáze aplikace pro porovnávání fulltextů nepracují s totožným obsahem; MnS umožní uživateli vkládat fulltexty ve formátu TEI P5 do obou systémů prostřednictvím rozhraní Manuscriptoria. Databáze aplikace pro porovnávání fulltextů obsahuje všechny fulltexty MnS + fulltexty z dalších zdrojů.

Zejména varianty A a B se výrazně liší dopadem na work-flow zpracování dat fulltextů v systému MnS a tedy na implementaci do provozu skutečného Manuscriptoria. Ze strany NKČR bude teprve upřesněno, která z variant bude v budoucnu preferována.

Varianta C by v zásadě mohla být z hlediska MnS rozšířenou variantou A, protože nepředpokládáme, že by v databázi aplikace pro porovnávání fulltextů obsažený

doplněk k množině fulltextů v MnS byl pro uživatele MnS zajímavý - pokud by byl, je v zájmu NKČR, aby byl i součástí MnS, tj. jde nakonec o řešení podle varianty A.

2.1 Koncovo-uživatelské funkce v interface MnS

Okruh koncovo-uživatelských funkcí společných pro všechny varianty byl během konzultací stanoven tak, aby těžil jednak z funkcí nové, subdodavatelem NKČR dodávané aplikace pro porovnávání fulltextů a dále aby žádoucím způsobem využil a rozvinul množinu funkcí osobní digitální knihovny.

2.1.1 Funkce porovnání fulltextu

Výběr fulltextů k porovnávání

Uživatel bude mít k dispozici interface MnS, aby mohl

a) vybrat v MnS existující fulltext či jeho část, nebo

b) vložit část fulltextu, který v MnS obsažen není

a požádat o nalezení podobností v jiných plných textech.

Podle vybraného či zadaného textu budou hledány podobnosti v podmnožině fulltextů obsažených v databázi aplikace pro porovnávání plných textů. Tato podmnožina určená k porovnání bude podle informací pracovníků NKČR též vytvářena uživatelem, tj. uživatel vybere plné texty, ve kterých se bude hledat podobnost se zdrojovým textem. K tomu bude potřeba doplnit do MnS interface.

Způsob získání dat pro seznam, ze kterého lze takto vybírat se výrazně liší podle toho, zda se realizuje varianta A či B. V případě varianty A bude MnS načítat fulltexty zevnitř systému. V případě varianty B bude MnS žádat o seznam externí aplikaci.

Poznámka: Vzhledem k předpokládanému rozvoji MnS lze očekávat nárůst množství fulltextů v systému. Pokud očekáváme, že uživatel z nich bude vybírat množinu k porovnání, bude nutné umožnit mu vyhledávání nad množinou všech fulltextů. Pokud bude platit varianta A, lze využít technologií, se kterými MnS v současnosti pracuje. Pokud bude platit varianta B, bude nutno aplikaci pro porovnávání fulltextů opatřit vyhledáváním a příslušné API bude muset být vhodným způsobem rozšířeno (například o podporu protokolu SRU [2]).

Zobrazení výsledků porovnání

Po odeslání požadavku uživatelem vrátí aplikace pro porovnávání fulltextů výsledky porovnání. Ty jsou následně uživateli zobrazeny jako seznam, jehož položky tvoří informace o výsledku porovnání vybraného či vloženého textu s každým fulltextem ze seznamu fulltextů vybraných k porovnání.

Součástí informací takto získaných je informace o míře shody, ID fulltextu v databázi aplikace pro porovnávání fulltextů a buď přímo celé XML každého fulltextu či pravděpodobněji jen výtah identifikačních a obsahových informací (titul, uložení a podobné relevantní informace).

Seznam bude řazen dle informace o míře shody. Budou se zobrazovat pouze relevantní položky, u kterých bude dosaženo stanovené míry shody. Uživatel bude moci tuto hodnotu upravit.

Zobrazení porovnání pro konkrétní dvojici fulltextů

Ze seznamu výsledků porovnání bude možné vybrat k zobrazení konkrétní dvojici fulltextů. Předpokládá se zobrazení obou textů s barevným rozlišením podobných oblastí – v závislosti na funkcionalitě aplikace pro porovnávání fulltextů.

Uživatel tedy kliknutím vybírá danou dvojici, MnS do aplikace pro porovnávání fulltextů odesílá identifikátor dané dvojice a získá zpět domluveným způsobem otagované (kvůli např. barvení apod.) původní plné texty.

2.1.2 Návaznost na osobní digitální knihovnu

Je zřejmé, že porovnávání fulltextů – pokud bude v budoucnu nasazeno v rutinním provozu – bude funkce zajímavá především pro skupinu odborníků a bude relativně náročná na použití.

Bude tedy vhodné umožnit využití funkcí osobní digitální knihovny k urychlení práce či odstranění opakovaných vyhledávání.

Mělo by být možné uložit výsledky práce, tj. umožnit

- ukládat a načítat seznamy vybraných fulltextů,
- ukládat a načítat seznamy porovnaných fulltextů,
- umožnit dodatečné ruční úpravy těchto seznamů,
- umožnit porovnávání nových textů s již uloženými seznamy,
- uložit porovnání konkrétní dvojice fulltextů.

Navíc by mělo být možné opatřit takto uložené informace dodatečnými informacemi (popisky, odkazy apod.) na úrovni celého seznamu, dokumentů i jejich částí.

3 Závěr

Během konzultací a při následném shrnutí těchto základních uživatelských požadavků, jež byly při diskuzích se Zadavatelem formulovány, vyplývá, že jde o poměrně komplexní problematiku s výraznými přesahy za hranice aktuálně řešeného úkolu.

Pokud na základě zde shrnutých uživatelských požadavků Zadavatele vzniknou požadavky směřující k úpravě aplikace pro porovnávání fulltextů, bude nutné, aby její vývoj dále pokračoval.

V souvislosti s dalším plánováním kroků směřujících k implementaci technologie do ostrého provozu MnS považujeme za nutné zaměřit se na následující:

- Provedení zátěžových testů aplikace pro porovnávání fulltextů při simulaci reálného provozu MnS a využití databáze fulltextů takového rozsahu, jaký lze očekávat v nejbližších obdobích (nejméně 2-3 roky).
- Návrh řešení podle jedné z vybraných variant A či B; dosavadní výsledky diskuzí ukazují spíše na realizaci ve variantě A, ale obě řešení jsou možná.
- Správu fulltextů.
- Vyhledávání nad fulltexty.

Aby nasazení nové technologie mělo smysl v ostrém provozu, bude nutno podpořit vznik relevantního množství fulltextů, se kterými bude uživatel moci pracovat. (Již nyní počet fulltextů narůstá a tempo s jakým fulltexty budou přibývat, vzrůstá.)

Z hlediska AiP Beroun, jakožto technického garanta MnS, je proto podstatné zejména zajištění efektivní správy většího množství fulltextů.

Dále je nutné při zpřístupnění splnit specifická očekávání uživatelů ohledně zobrazení (je v MnS řešeno v minulých VaV) a především vyhledávání, které není dosud v MnS řešeno vůbec (nelze vyhledávat v obsahu fulltextů!). Tyto funkce zřejmě očekává výrazně širší obec uživatelů, nežli vlastní porovnávání. Zároveň je zřejmé, že samotné popsané porovnávání se bez vyhledávání nad obsahem fulltextů (pro výběr smysluplné množiny dokumentů ke srovnání) neobejde.

Bez vyřešení správy fulltextů, jejich prezentace a bez zajištění vyhledávání bude efektivní nasazení i kvalitní technologie porovnávání obsahu fulltextů do skutečného provozu problematické.

U vědomí těchto souvislostí Zadavatel společně s Řešitelem již v roce 2010 hledali zdroje financování těchto úkolů: výsledkem byla formulace projektu *TEXTORIUM – virtuální badatelská studovna elektronických edic* a jeho přihláška do programu do NAKI, bohužel neúspěšná.

Mají-li být výsledky aktuálního výzkumu využity v běžné praxi, je nutné najít způsob, jak splnit tyto ryze praktické (nevýzkumné) úkoly.