Specifikace software pro NUŠL

Úvod

Hlavní funkce digitálních knihoven pro ukládání jakýchkoliv dokumentů, tedy i pro úložiště šedé literatury, jsou získávání, zpracovávání, uchovávání, zpřístupňování a ochrana dat. Specifika šedé literatury spočívají zejména v organizační oblasti získávání dokumentů a řešení autorských práv. Tato specifika nijak významně neovlivňují výběr software pro úložiště šedé literatury, výběrová hodnotící kritéria jsou v zásadě totožná s kritérii pro výběr software pro provoz jakékoliv digitální knihovny. Rozhodujícím kritériem pro výběr vítězného návrhu bude, jak SW podporuje hlavní funkce digitální knihovny. Podpůrná kritéria pak mají za úkol zjistit další žádoucí vlastnosti, jako jsou uživatelský a administrátorský komfort, rozšiřující funkce, hardwarové nároky, provozní softwarová platforma, dostupnost služeb, možnost nasazování SW ve více etapách apod. Výběru software pro úložiště šedé literatury či jakékoliv jiné digitální knihovny je třeba věnovat velkou pozornost, protože zejména z důvodu její digitální povahy není žádné rozhodnutí tak důležité, jako právě výběr vhodné softwarového řešení.

Z tohoto pohledu lze požadavky na software pro STK NUŠL rozdělit do pěti oblastí:

- **Řízení obsahu digitální knihovny** – toto kritérium se vztahuje k hlavním funkcím digitální knihovny, zahrnuje tvorbu, ukládání, kontrolu a verzování obsahu (např. při migraci souborů) digitální knihovny, jeho prohlížení lineární (rejstříky) i hierarchické, správu metadat, různé mechanismy pro získávání dat (např. harvesting - OAI-PHM, web services), interoperabilitu pro spolupráci s jinými digitálními knihovnami (např. Z39.50, OAI-PMH, SRU), podporu pro velké objemy dat, vícejazyčné vyhledávání nad metadaty a soubory (fulltextové), logické a proximitní vyhledávání, podporu různých formátů ukládání souborů - textových (např. DOC, RTF, PDF), obrázkových (např. TIFF, JPEG, GIF, PNG), prezentačních (např. MS PPT, Adobe PDF a PostScript), strukturovaných (př. HTML a XML), audio a video (např. WAV, MP3, AVI, MPEG4, Real audio a video), podpora kódování (př. ASCII, UNICODE, UTF-8).

- **Uživatelské rozhraní** – přístup přes www s podporou hlavních internetových prohlížečů, možnost přizpůsobení uživatelského rozhraní pro různé role uživatelů zvenku i zevnitř knihovny, vícejazyčné uživatelské rozhraní, vícejazyčná prezentace výsledků vyhledávání, funkce web2.0

- **Správa a řízení přístupu uživatelů** – klasifikace a profily uživatelů na základě jejich rolí (např. RBAC), autentifikace uživatelů na základě uživatelského jména a hesla, řízení přístupu IP adres, přístup přes proxy, monitorování a reporting pro vyhodnocování chování uživatelů pro pozdější využití ke zlepšení služeb digitální knihovny, autorizace uživatelů (např. Shibolet, Edurom, OpenID), zaznamenávání aktivit uživatelů pro účely účtování

- **Administrace systému** – tento požadavek se vztahuje ke správě provozního prostředí digitální knihovny, která může být zejména u digitálních knihoven velkého rozsahu velice pracná a náročná. Spadají sem nástroje pro správu metadat, nastavování automatických mechanismů sběru dat, indexování, automatické generování klíčových slov, formátování výstupů apod.

- **Jiné požadavky** – uživatelská a administrátorská dokumentace, helpdesk (podpora ze strany výrobce/implementátora systému), rozšiřující funkce pro uživatele jako jsou diskusní a komunikační skupiny, automatické upozorňování na různé události (výskyt nových dokumentů splňující určitá kritéria), RSS výstupy, uživatelské schránky a komentáře k dokumentům, požadavky na provozní hardware a software apod.

Hlavní proces NUŠL

Získávání obsahu - poskytovatelé obsahu zpřístupní data různými způsoby, prostřednictvím web services, přes standard OAI-PMH, přes e-mail, předáním dat přes ftp, na nosiči, přímým vložením. V případě, že metadata získaná přes OAI-PMH obsahují URL na soubor(y) a dohoda s poskytovatelem to umožňuje, stahují se i samotné soubory.

Zpracování obsahu – takto poskytnutá data je nutno upravit, metadata převést do formátu NUŠL pomocí konverzních šablon, soubory zkonvertovat do prezentačního formátu (PDF/A), připravit dávku pro import do databáze s možností její kontroly, uložit do databáze metadata a prezentační, případně i originální podobu souborů v různých formátech, vygenerovat indexy a klíčová slova, zformátovat vyhledávací výstupy.

Zpřístupňování obsahu – uživatelé přistupují přes www rozhraní anonymně jako veřejnost, pokud přistupují z určitých institucionálních adres, systém je přihlásí jako anonymní institucionální uživatele, nebo se hlásí jménem a heslem, které jim umožní získat knihovní systém, případně jejich autorizace probíhá přes systémy jako jsou Shibolet, Edurom, OpenID. Uživatelé mají možnost prohlížet obsah podle různých kritérií lineárně v seznamech i hierarchicky v stromové struktuře, případně i formou Web 2.0, vyhledávat nad metadaty i obsahem souborů pomocí booleovských výrazů, fulltextově, proximitně. Mají možnost ukládat výsledky vyhledávání do trvalých uživatelských schránek, (nákupní košík), psát komentáře k obsahu i formou diskusních skupin, vyžádat si zprávy e-mailem nebo so své pracovní plochy o událostech ve sbírkách (výskyt nového záznamu atd.). Systém umožňuje zaznamenávat aktivity uživatelů pro účely účtování, případně zpřístupňovat placené služby prostřednictvím rozhraní na jiný systém.

Uchování a ochrana obsahu – kontrola kvality obsahu z hlediska správce obsahu, duplicity dat, kvalita metadat, možnost formulovat dotazy nad databází, kontrola primárních souborů na udržitelnost obsahu atd.

Administrace systému - adminitrátorské nástroje pro řízení, konfiguraci a administraci celého systému včetně řízení operací a jejich časování nad centrální databází i s možnosti ručních zásahů (např. posun spuštěné časové náročné úlohy na pozdější dobu), správa uživatelských účtů, rolí a přístupových práv, automatické upozorňování na různé události, možnost formulovat administrátorské dotazy nad databází, systém zálohování dat.

Obecná funkční specifikace

1.1. systém a databáze pro ukládání velkého počtu knihovnických záznamů na základě definovaných metadat z různých interních i externích zdrojů a pro jejich rychlé a efektivní vyhledávání a poskytování koncovým uživatelům přes webové uživatelské rozhraní, podporující interoperabilitu s jinými digitálními úložišti dokumentů

1.2. automatické testování kvality a integrity metadat (např. délka polí, numerický/alfanumerický obsah, povinná/nepovinná/závislá pole) a vytváření zpráv o případných nedostatcích pro administrátory a katalogizátory

1.3. automatická extrakce klíčových slov z textových dokumentů na základě frekvence výskytu výrazů z definovaného slovníku

1.4. konverze metadat z proprietárních formátů do standardních formátů i vlastního formátu NUŠL, plná konfigurovatelnost formátu vstupů i výstupů, (systém popisu struktury vstupních a výstupních formátů, podobně jako XSLT pro XML), součástí systému by měla být sada nastavení konverzí z/do obvyklých a známých formátů, možnost editace těchto popisů přes uživatelské webové rozhraní

1.5. formátování bibliografických metadat pro různé účely, např pro různé výstupy při vyhledávání a perezentaci dokumentů, oddělení administrace obsahu dat a administrace jejich výstupního vzhledu, možnost editace těchto popisů přes uživatelské webové rozhraní

1.6. automatický mechanismus pro sběr dat (harvesting), podporující běžné standardy, založený na OAI-PHM (Open Archive Initiative Protocol for Metadata Harvesting) + OAI-ORE? nebo OAI-x?, a podporující obousměrnou interoperabilitu s jinými digitálními knihovnami, web harvesting (stahování dat prostřednictvím webu), web submission (poskytování dat přes webové rozhraní od autorizovaných subjektů včetně konverzí fulltextových dokumentů z různých jiných textových i obrázkových formátů), e-mail upload (možnost nahrávání dat poslaných e-mailem), podpora vyhledávání v jiných digitálních knihovnách

1.7. systém indexování metadat, odkazů a fulltextových souborů a mapování těchto indexů na tagy metadat pro rychlé vyhledávání v databázi, správa definice indexů

1.8. systém klasifikace záznamů v databázi podle různých kritérií pro pozdější vyhledávání (např. frekvence výskytu slov, hodnota tagu, počet zobrazení záznamu, počet vyžádání dokumetu apod.)

1.9. kontrola vstupních dat proti existující databázi s nastavitelnými kriterii, např. pro zamezení duplicitního uložení dat

1.10. adminitrátorské nástroje pro řízení, konfiguraci a administraci celého systému včetně řízení operací a jejich časování nad centrální databází i s možnosti ručních zásahů (např. posun spuštěné časové náročné úlohy na pozdější dobu), řízení přístupu k systému na základě rolí (RBAC – Role Based Access Control)

1.11. hromadný vstup předem naformátovaných dat (XML) do centrální databáze včetně kontroly konzistence dat (bulk upload)

1.12. uživatelské webové rozhraní pro vyhledávání, prezentaci a poskytování dokumentů s podporou různých nadstavbových funkcionalit, jako jsou osobní uživatelské schránky pro ukládání dokumentů, (obdoba nákupních košíků), nastavování různých upozornění na výskyt dokumentů s definovanými klíčovými slovy, diskusní uživatelské skupiny, komunikační uživatelské nástroje (nástěnky, uživatelské komentáře a hodnocení k dokumentům apod.)

1.13. vyhledávací nástroj s webovým rozhraním, umožňující vyhledávání podle slov i frází včetně složitějších dotazů s booleovskými kriterii, strukturovaná prezentace výsledků vyhledávání (třídění podle typu dokumentů), v případě nenalezení podle zadaných kritérií návrh alternativ

Na základě výše uvedených požadavků stanovujeme pro účely vyhodnocení tyto skupiny kritérií:

- **Řízení obsahu** – nástroje a postupy podporující ukládání obsahu do digitální knihovny a řízení procesu tohoto ukládání, verzování

- **Získávání obsahu** – import a export obsahu, podpora různých formátů dokumentů

- **Metadata** – podpora různých metadatových formátů je důležitá pro indexování, ukládání, zpřístupňování a ochranu obsahu

- **Podpora vyhledávání** – týká se celé řady vyhledávacích a prohlížecích funkcí, vyhledávání v metadatech, fulltextové vyhledávání, hierarchické prohlížení apod.

- **Řízení přístupu a ochrana soukromí** – řízení přístupu a ochrana soukromí zahrnuje administraci hesel, uživatelských účtů s přístupovými právy včetně možnosti získat zapomenuté heslo atd.

- **Podpora výstupů a dotazování** – toto kritérium souvisí s možností vyhodnocovat využití digitální knihovny a odhalovat vzorce chování uživatelů pro vylepšování poskytovaných služeb, zaznamenávání aktivit uživatelů pro účely účtování

- **Udržitelnost, ochrana dat** – ochrana metadat, konzistence a integrity datové základny, zálohování, podpora případné migrace dat

- **Interoperabilita –** možnost obousměrné integrace s jinými distribuovanými systémy na úrovni metadat, prohledávání i získávání a poskytování dokumentů, podpora OAI-PHM, Z39.50

- **Uživatelské rozhraní** – tato kategorie se týká podpory více jazyků a schopnosti přizpůsobit uživatelské rozhraní různým potřebám různých uživatelů či rozdílných implementací

- **Podpora standardů** – standardy jsou důležité pro sdílení a dlouhodobé uchování digitálního obsahu. Jedná se o celé spektrum oblastí standardů od metadat přes interoperabilitu až po formáty uložených dokumentů

- **Nástroje pro automatizaci** – tato kategorie se týká nástrojů pro automatizované získávání obsahu, harvesting, generování metadat, údržbových činností atd.

- **Podpora, služby** – důležitý aspekt u všech softwarových systémů. Mnoho dobrých systémů digitálních knihoven se mj. nachází v oblasti Open Source, kde je třeba na toto hledisko dbát zejména. Důležitá je dokumentace, helpesk, sběr požadavků na vylepšení, případně diskusní fóra atd.

- **Hardware a provozní software** – hardwarové nároky na úložiště dat, provoz a zálohování systému, zabezpečení jeho dostupnosti, podporované operační a databázové systémy atd.

2. Jednotlivá dílčí kritéria

2.1. Řízení obsahu

2.1.1. Podpora systému pro vícenásobné sbírky v rámci jedné instalace

2.1.2. Nástroje administrátora pro parametrizaci ukládání obsahu (nastavit průběh importu, vstupní formáty průběh konverze)

2.1.3. Lze nastavit šablonu stránky zvlášť pro každou sbírku?

2.1.4. Definice oprávění pro ukládání obsahu na základě rolí?

2.1.5. Konfigurovatelné oprávnění pro vkládání obsahu v rámci různých sbírek (role pro sbírky)?

2.1.6. E-mailové informace uživatelům o událostech ve sbírkách, jaké informace lze nastavit

2.1.7. E-mailové informace administrátorům o událostech ve sbírkách, jaké informace lze nastavit

2.1.8. Uživatelské (individuální dodavatel dat) revize uložených dat

2.1.9. Umožňuje správci obsahu revize poskytnutých dat

2.2. Získávání obsahu

2.2.1. Nahrávání komprimovaných dat

2.2.2. Nahrávání dat ze známé adresy – URL

2.2.3. Hromadné nahrávání dat (bulk upload)

2.2.4. Hromadné nahrávání metadat k existujícím sbírkám

2.2.5. Hromadný export/přenositelnost obsahu do jiného systému

2.2.6. Možnost administrátorsky omezit povolené formáty obsahu

2.2.7. Entita k nahrání může obsahovat více souborů a/nebo typů souborů

2.2.8. Textové soubory

2.2.8.1. ASCII

2.2.8.2. Unicode

2.2.8.3. RTF

2.2.8.4. Jiné

2.2.9. Obrázky

2.2.9.1. JPEG

2.2.9.2. TIFF

2.2.9.3. GIF

2.2.9.4. Jiné

2.2.10. Prezentace

2.2.10.1. MS PPT

2.2.10.2. Adobe PDF

2.2.10.3. Adobe PostScript

2.2.10.4. Jiné

2.2.11. Strukturované formáty

2.2.11.1. HTML

2.2.11.2. XML

2.2.11.3. SGML

2.2.11.4. Jiné

2.2.12. Multimédia

2.2.12.1. Wave

2.2.12.2. MP3
2.2.12.3. Real Audio
2.2.12.4. AVI
2.2.12.5. MPEG
2.2.12.6. Real Video
2.2.12.7. Jiné
2.2.13. Možnost získat minulé verze souboru/záznamu
2.2.14. Možnost identifikovat změny
2.2.15. Možnost porovnat změny

2.3. Metadata

2.3.1. Možnost nahrávání, změn a indexování akceptovaného obsahu v reálném čase
2.3.2. Podporované formáty/standardy metadat
2.3.2.1. MARC 21
2.3.2.2. Dublin Core
2.3.2.3. EAD
2.3.2.4. LOM
2.3.2.5. METS
2.3.2.6. MODS
2.3.2.7. VRA CORE Categories
2.3.2.8. Jiné
2.3.3. Možnost přidat/smazat položky metadat
2.3.4. Možnost nastavit počáteční (defaultní) hodnoty položek metada
2.3.5. Podpora Unicode pro metadata

2.4. Podpora vyhledávání

2.4.1. Fulltextové vyhledávání
2.4.1.1. booleovská logika
2.4.1.2. neúplné výrazy / náhražkové znaky
2.4.1.3. vyhledávání frází
2.4.1.4. proximitní vyhledávání
2.4.1.5. přibližné vyhledávání podobných výrazů
2.4.2. Prohledávání popisných metadat
2.4.2.1. booleovská logika
2.4.2.2. neúplné výrazy / náhražkové znaky
2.4.3. Prohledávání vybraných položek metadat
2.4.4. Prohlížení záznamů podle
2.4.4.1. autora
2.4.4.2. názvu
2.4.4.3. data vydání
2.4.4.4. předmětu
2.4.4.5. sbírek
2.4.4.6. přidaných/upravených polí
2.4.4.7. více položek najednou
2.4.5. Třídění výsledku prohledávání podle
2.4.5.1. autora

2.4.5.2. názvu
2.4.5.3. data vydání
2.4.5.4. významu (relevance)
2.4.5.5. jiného hlediska

2.5. Řízení přístupu a ochrana soukromí

2.5.1. Uživatelská hesla přiděluje systém
2.5.2. Hesla si volí sami uživatelé
2.5.3. Funkce pro získání zapomenutého hesla
2.5.4. Založení účtu uživatele
2.5.5. Editace účtu uživatele
2.5.6. Smazání účtu uživatele
2.5.7. Omezení přístupu na úroveň pole/objektu
2.5.8. Omezení přístupu na úroveň sbírek
2.5.9. Lze definovat skupiny/role uživatelů
2.5.10. Omezení přístupu podle rolí
2.5.11. Sbírky lze přizpůsobit jednotlivým rolím
2.5.12. Fitrování zdrojových IP adres
2.5.13. Filtrování proxy
2.5.14. Přístup založený na kreditech
2.5.15. Podpora kryptování při zadávání citlivých dat
2.5.16. Podpora digitálních podpisů

2.6. Podpora výstupů a dotazování

2.6.1. Systém generuje statistiky o užívání
2.6.2. Pokud ano, jaké?
2.6.3. Lze nastavit časy pro spouštění výstupních zpráv
2.6.4. Lze výstupní zprávy upravovat?
2.6.5. Existují uzpůsobitelné šablony výstupních zpráv?
2.6.6. Zaznamenávání aktivit uživatelů pro účely účtování

2.7. Bezpečnost, ochrana dat

2.7.1. Je zabezpečena trvalá identifikace dokumentů?
2.7.2. Přiděluje identifikátory systém?
2.7.3. Existuje podpora CNRI Handles?
2.7.4. Podporuje systém kontrolu kvality?
2.7.5. Pokud ano, jak?
2.7.6. Existuje popsaná strategie ochrany digitálních dat?
2.7.7. Pokud ano, krátce popište

2.8. Interoperabilita

2.8.1. Podpora OAI-PHM – metadata harvesting
2.8.2. Podpora protokolu Z39.50
2.8.3. E-mail upload
2.8.4. Web upload
2.8.5. Web harvesting (OAI-ORE, OA-X …)

2.8.6. Rešeršní protokol Dienst

2.8.7. Rešeršní protokol SDLIP

2.9. Uživatelské rozhraní

2.9.1. Lze přizpůsobovat vzhled uživatelského rozhraní?

2.9.2. Lze ke statickým/dynamickým stránkách připojit upravené záhlaví/zápatí?

2.9.3. Podpora vícejazyčného uživatelského rozhraní v rámci jedné instalace systému?

2.10. Podpora standardů – sumarizace

2.10.1. Strukturovaná data - HTML, XML, SGML …

2.10.2. Metadata - Dublin Core …

2.10.3. Text - ASII, Unicode, RTF …

2.10.4. Obrázky - JPG, TIFF, GIF …

2.10.5. Prezentace – MS PowerPoint, Adobe PDF, Adobe PostScript …

2.10.6. Multimédia - wav, mp3, avi, mpeg, real audio, real video …

2.11. Nástroje pro automatizaci

2.11.1. Systém pro vstup metadat

2.11.2. Generování indexů

2.11.3. Generování HTML stránek

2.11.4. Generování výstupních sestav

2.12. Podpora, služby

2.12.1. Dokumentace, příručky

2.12.2. Diskusní fórum, e-mailové skupiny

2.12.3. Hlášení chyb

2.12.4. Sběr požadavků na rozvoj

2.12.5. Podpora – helpdesk

2.13. Hardware, provozní software

2.13.1. Hardwarové nároky na server

2.13.2. Hardwarové nároky na klientské stanice

2.13.3. Podporované serverové operační systémy

2.13.4. Podporované klientské operační systémy

2.13.5. Podporované databáze (MySQL, Oracle …)