Národního úložiště šedé literatury – digitální repozitář a centrální vyhledávací rozhraní

Pro časopis ITlib. Informačné technológie a knižnice

Petra Pejšová, NTK, petra.pejsova@techlib.cz, 232 002 485

Odevzdáno: 30.4.2010

Projekt Národního úložiště šedé literatury (NUŠL) se začal realizovat díky podpoře Ministerstva kultury České republiky v rámci programových projektů výzkumu a vývoje jako výzkumný záměr pod názvem „Digitální knihovna šedé literatury – funkční model a pilotní realizace“. Tento projekt je rozdělen do tří etap, které probíhají od roku 2008 do roku 2011, a řeší ho Národní technická knihovna společně s Vysokou školou ekonomickou v Praze.

Co je šedá literatura? Existuje několik definic z nichž asi nejznámější je definice vniklá v Luxemburgu v roce 1997 a rozšířená v New Yorku v roce 2004[1]. Volně přeloženo: *Šedá literatura jsou informace produkované na všech úrovních vládních, akademických, obchodních a průmyslových institucí jak v elektronické tak v tištěné podobě, které neprošly standardním vydavatelským procesem či nejsou distribuovány do standardní prodejní sítě, tj. jsou vydávány institucemi, jejichž hlavní činností není vydavatelská činnost.* Co vše si tedy lze pod šedou literaturou představit? Jsou to různé druhy zpráv (př. výzkumné, výroční, závěrečné z projektů), konferenční materiály (př. články, prezentace, postery), vysokoškolské kvalifikační práce (např. diplomové, disertační), firemní literaturu (např. katalogy, návody), studijní materiály (např. učební texty, osnovy, záznamy z přednášek), formální i neformální komunikace (např. telefonické rozhovory, e-amily, blogy), sociální sítě (např. Twiter, Facebook), diskuse na wiki aj.[2]

Hlavním cílem projektu je systematicky shromažďovat, dlouhodobě archivovat a zpřístupňovat odbornou šedou literaturu především z oblasti výzkumu a vývoje, státní správy a školství, ale i z komerčního sektoru a sféry „open access“, na národní úrovni a zvýšit tak její dohledatelnost a přístupnost. Souvisejícím cílem je informovat odbornou veřejnost a podporovat diskusi s oblasti šedé literatury. Za tímto účelem byly v roce 2008 zřízeny první české webové stránky plně se věnující šedé literatuře a od téhož roku se každoročně pořádá „Seminář ke zpřístupňování šedé literatury“. Pomocí obou těchto informačních kanálů poskytujeme informace jak o projektu samotném a jeho řešení tak všeobecné informace o šedé literatuře z tuzemska i ze světa. Témata reflektují aktuálně řešenou problematiku jako je stav systémů pro šedou literaturu, formáty standardy pro dlouhodobé archivování, standardizace pro popis a výměnu zdrojů, autorská práva, perzistentní identifikátory a další. Webové stránky naleznete na adrese **http://nusl.techlib.cz/**, kde je k dispozici i on-line sborník ze seminářů.

Obrázek č. 1: Náhled úvodní webové stránky projektu NUŠL

Klíčové pro zajištění hlavního cíle projektu je softwarové řešení. V případě NUŠL byly vybrány softwarové nástroje CDS Invenio pro digitální repozitář a indexační a vyhledávací systém FAST ESP pro uživatelské rozhraní. Tyto nástroje společně zabezpečují vytvoření efektivního automatizovaného sběru dat (metadat i digitálních dokumentů), jejich dlouhodobou archivaci a uživatelsky příjemné a snadné zpřístupnění.

Nástroj CDS Invenio (Cern Document Server), vybraný pro digitální repozitář, je vysoce modulární systém, který vznikl a dále se vyvíjí ve švýcarském CERNu ve spolupráci s École Polytechnique Fédérale de Lausanne (EPFL). CDS Invenio představuje vyzrálé řešení pro všeobecnou správu dokumentů, institucionální repozitář nebo rozsáhlý knihovní systém. Výhodami systému CDS Invenio jsou jeho propracovanost a flexibilita, kterou zajišťuje jeho modulární skladba, díky níž lze systém nastavit mnoha různými způsoby pro nejrůznější použití, dále pak jeho lokalizace do 18 jazyků včetně češtiny a slovenštiny. Vývojový tým CDS Invenia v CERNu má ve svých řadách též slovensky

[1] "Information produced on all levels of government, academics, business and industry in electronic and print formats not controlled by commercial publishing i.e. where publishing is not the primary activity of the producing body." (Luxembourg, 1997 - Expanded in New York, 2004, dostupné na WWW <http://www.greynet.org/index.html>)

[2] PEJŠOVÁ, Petra. Národní úložiště šedé literatury. Čtenář. 2010, roč. 62, č. 5, s. 176 - 180. ISSN 0011-2321

mluvícího člena. Naskýtá se tedy i možnost komunikace v nám blízkém jazyce. Systém umožňuje vlastní definici metadatového schématu, přičemž interní jmenná konvence je podle MARC21. CDS Invenio bylo vybráno pro digitální repozitář také proto, že patří mezi Open Source software, tzn. že je možné ho volně instalovat, používat i upravovat. To nám umožňuje jeho nastavení pro ukládání šedé literatury a následné šíření mezi spolupracující organizace.

Systém CDS Invenio jsme v první fázi v roce 2009 instalovali v jeho základní verzi, tak jak je šířen mimo CERN. Při instalaci jsme využili aplikaci Virtual Box, která lze nainstalovat na všechny víceméně známé platformy. Pak jsme v rámci Virtual Boxu instalovali CDS Invenio nad operačním systémem Linux-Debian. Toto řešení jsme zvolili, protože umožňuje snadnou distribuci nastavené verze CDS Invenia pro šedou literaturu spolupracujícím institucím. Nastavení pro šedou literaturu obsahuje především definovaný metadatový formát NUŠL, šablony pro vkládání dokumentů vysokoškolských kvalifikačních prací, konferenčních materiálů, zpráv, autorských prací, studijních materiálů a firemní literatury. Dále zahrnuje nastavení sbírek a podsbírek podle druhů dokumentů a vyhledávací indexy a rejstříky. Například ve sbírce zprávy jsou podsbírky výroční zprávy, průběžné zprávy z projektů, závěrečné zprávy z projektů, výzkumné zprávy, technické zprávy a statistické zprávy.

Obrázek č. 2: Digitální repozitář NUŠL

Záměrem projektu NUŠL je též vytvoření integrující platformy repozitářů, které obsahují šedou literaturu. Tuto integrační funkci zajišťuje indexační a vyhledávací systém ESP FAST, který poskytuje zabezpečené, relevantní a škálovatelné vyhledávání nad připojenými repozitáři. Díky funkcionalitě systému ESP FAST je možné určit kontext a účel dotazu, vyhledat odpovídající termíny jak v metadatech, tak v dokumentech, a obdržet odpovědi vyskytující se v souvislostech. Uživatelé tak získají přesné výsledky společně s kontextovou dynamickou navigací pro další hledání souvisejících informací. V současné době jsou do centrálního vyhledávacího rozhraní připojeny čtyři hlavní zdroje[3] s více jak 31 tisíci záznamy. Jedná se o šedou literaturu z katalogu NTK, kterou představují především vysokoškolské kvalifikační práce, konferenční materiály, firemní literatura a zprávy. Dále je připojen repozitář vysokoškolských kvalifikačních prací VŠE a bibliografická databáze Evidence publikační činnosti ASEP, která obsahuje z šedé literatury především konferenční materiály a výzkumné zprávy. Do centrálního vyhledávacího rozhraní je zapojen digitální repozitář NUŠL v CDS Inveniu, který nyní přechází z testovacího na běžný provoz. Řešení vychází ze záměru vytvořit v pilotní fázi projektu komfortní aplikační prostředí, zaměřené zejména na zprovoznění centrálního vyhledávacího rozhraní, které dokáže zpřístupnit uživatelům jak data z digitálního repozitáře, tak data z vybraných úložišť šedé literatury v jednom interaktivním prostředí. Cílem je převést co nejvíce zdrojů přímo do digitálního repozitáře NUŠL, který zajistí dlouhodobou archivaci a přístupnost dat. V centrálním vyhledávacím rozhraní by pak měly zůstat jako samostatné zdroje šedé literatury pouze repozitáře, které budou samostatně zajišťovat dlouhodobou archivaci a přístup k datům. Vyhledávání v centrálním rozhraní si můžete vyzkoušet na **www.nusl.cz** .

Obrázek č. 3: Centrální vyhledávací rozhraní NUŠL

Ideální formou spolupráce je sklizení dat z lokálního repozitáře dodavatelské instituce přes protokol OAI-PMH. Dodavatelská instituce připraví OAI-PMH set pro šedou literaturu a při importu dat do digitálního repozitáře NUŠL se provede konverze do metadatového formátu NUŠL.

Na podporu spolupráce s institucemi, které ještě nemají vlastní repozitář, nabízíme dvě možnosti využití systému CDS Invenio nastaveného pro šedou literaturu. První možností je, že spolupracující organizace bude vkládat dokumenty přímo do digitálního repozitáře NUŠL v systému CDS Invenio. Bude používat metadatový formát NUŠL a přednastavené šablony pro vkládání dokumentů šedé literatury. Pro organizaci vytvoříme v rámci digitálního repozitáře NUŠL její vlastní sbírku, přístupnou přes vstupní stránku - viz http://invenio.ntkcz.cz. Organizaci budou přidělena příslušná administrátorská práva na správu její sbírky, tj. vkládání, editace a přidělování práv k zveřejňování dokumentů z její sbírky. K administraci sbírky v rámci NUŠLu bude organizaci k dispozici stručná dokumentace s odkazy na úplnou dokumentaci od CERNu. Případně může správu sbírky (kromě vkládání dokumentů a vyplňování metadat) převzít NTK, což bude ošetřeno v příslušné smlouvě mezi NTK a organizací.

Druhou možnost představuje instalace nastaveného CDS Invenia pro šedou literaturu, který si spolupracující organizace nainstaluje na vlastním HW jako lokální repozitář. Instituce nebude

[3] Zdroj představuje databázi, která obsahuje šedou literaturu.

zasahovat do nastavení pro šedou literaturu, tzn. bude používat metadatový formát NUŠL a připravené šablony pro vkládání dat. Dále si může lokální instalaci CDS Invenia upravit dle vlastních požadavků a použít i pro jiné druhy literatury. Následná spolupráce bude probíhat tak, že data budou z lokálního repozitáře importována do digitálního repozitáře NUŠL přes protokol OAI-PMH přímo bez nutnosti konverze dat. Tato forma spolupráce bude ošetřena v příslušné smlouvě mezi NTK a organizací. Lokální instalace CDS Invenia je dostupná ke stažení a vyzkoušení na stránkách projektu NUŠL http://nusl.techlib.cz/ v oddílu Software společně s návodem k instalaci a popisem, jak vytvářet a upravovat sbírky a šablony.

Dlouhodobá archivace, možnost snadného vyhledání a rychlého přístupu k odborné šedé literatuře poskytuje výhody jak pro vědecké pracovníky, tak pro instituce. Pro vědeckého pracovníka poskytuje přínos v podobě centrálního archivu jeho prací, zlepšuje dostupnost jeho prací a zvyšuje šíření výsledků z jeho výzkumů. Instituci přináší výhody ve zvýšení viditelnosti a prestiže před veřejností a financujícími orgány. V neposlední řadě je významnou výhodou též možnost přístupu veřejnosti k výsledkům výzkumů. Tomuto tématu se na druhém ročníku Semináře pro zpřístupňování šedé literatury podrobně věnovala Sophia Jonesová z Univerzity of Notthingham.[4] NUŠL je právě takovým místem v digitálním prostoru, který trvale uchovává a zpřístupňuje intelektuální produkci vědců i institucí.

Více o šedé literatuře a spolupráci organizací v této oblasti se můžete dozvědět v letošním roce na následujících akcích:

Prezentace Národního úložiště šedé literatury na konferenci INFORUM 2010 - 26. května 2010 na VŠE v sekci Informační zdroje 3 krát jinak

Twelfth International Conference on Grey Literature - 6. - 7. prosince 2010 v NTK – celosvětové setkání mezinárodní komunity pro šedou literaturu

3. ročník Semináře ke zpřístupňování šedé literatury - 8. prosince 2010 v NTK

[4] JONES, Sophia. Open Access and digital repositories: the role of the DRIVER project. 22. 10. 2009 [online]. Praha: Národní technická knihovna, 2009. Dostupný z WWW: http://nusl.techlib.cz/sbornik/2009/ ISSN 1803-6015.