Persistentní identifikátory pro NUŠL – rozhodovací kritéria

Dokument vznikl jako podpora výběru perzistentního identifikátoru pro ***Národní úložiště šedé literatury*** *v rámci řešení projektu "Digitální knihovna pro šedou literaturu - funkční model a pilotní realizace", který podporuje Ministerstvo kultury. Projekt řeší Státní technická knihovna společně s Vysokou školou ekonomickou v Praze. Viz* http://nusl.stk.cz.

Úvod

Webové technologie otevřely obrovské možnosti v oblasti dostupnosti elektronických informací a způsobily tak revoluční změny ve způsobech, jak knihovny, archivy a jiní poskytovatelé informací mohou zpřístupňovat elektronické dokumenty různého druhu svým uživatelům.
Digitální knihovny tak mohou poskytovat přístup k informacím v distribuovaném prostředí otevřené sítě, kde je v zásadě lhostejné, kde se informace nebo uživatel nacházejí, a to prostřednictvím relativně standardního software na straně poskytovatele i uživatele, kterému postačí webový prohlížeč.

Distribuované digitální knihovny mají v tomto prostředí dva úkoly:
- uchovávat a archivovat informace,
- udržovat a pečovat o dostupnost těchto informací tak, aby je uživatelé nalezli, aby ukazatele či identifikátory směrovaly k dokumentům i po delší době.

Úspěch digitální knihovny, zejména distribuované, je tak založen na dlouhodobě konzistentním propojení zdrojů a na trvanlivosti (persistenci) identifikátorů a směrníků, které tyto zdroje zpřístupňují.

Tradičním ukazatelem, identifikátorem, směrníkem, „linkem" k informacím ve webovém prostředí je hyperlink URL (Uniform Resource Locator). Nicméně s časem se mnoho těchto linků přeruší, přestanou platit, nezpřístupňují původní zdroj.

Důvody pro toto přerušené spojení mohou být:
- informace, kterou link zpřístupňoval, byla z nejrůznějších důvodů odstraněna (pro vlastníka již nemá význam, nereflektuje současný stav věcí, není aktuální ….)
- soubor byl opět z různých důvodů přemístěn (reorganizace webu nebo souborové struktury, změna webové domény, změna vlastníka organizace …)

Řešením těchto problémů jsou organizační opatření, minimalizující nebezpečí narušení ukazatelů, a zejména užití systému persistentních identifikátorů. Existuje několik takových systémů (Handle, URI, URN, DOI, OAI, PURL, NBN atd., viz dále).

Je třeba podotknout, že systém persistentních identifikátorů sám o sobě nic nevyřeší, může být efektivní jen v případě, že je udržován. Pokud je zdroj informací přemístěn, je třeba jeho nové umístění propojit s persistentním identifikátorem ve zvoleném systému, což obvykle zajišťuje databáze tzv. resolveru (úložiště persistentních identifikátorů, mapující je na aktuální úložiště informačních zdrojů).

Organizace, která se rozhodne používat systém persistentních identifikátorů, se musí sama vnitřně rozhodnout takový systém plně podporovat, vytvořit pro něj organizační podmínky a na praktické úrovni si zodpovědět řadu otázek a vyřešit řadu úkolů. Hlavním úkolem je navrhnout samotný identifikátor, a jednou z hlavních otázek je, zda se organizace připojí k existující implementaci systému persistentních identifikátorů, nebo zda bude implementovat vlastní systém včetně resolveru. K návrhu identifikátoru a zodpovězení této otázky vede cesta přes řadu dílčích otázek, které tento dokument definuje.

Existuje několik vlastností, které musí splňovat systém pro perzistentní identifikátory, bez ohledu na druh materiálu.

Jedinečnost

Potřeba jedinečnosti bude zajištěna komponentou centrální autority přijatého systému. Identifikátory musí být unikátní v rámci digitálních zdrojů organizace, ale mohou být unikátní i celosvětově. Globální jedinečnosti lze dosáhnout, když bude využíván systém, který je využíván po celém světě a je zaštítěn buď 1) v každé zemi národní autoritou nebo 2) jedinou celosvětovou autoritou centrální.

Zajištění trvalosti = perzistence

Organizace musí udržovat propojení aktuálního umístění zdroje na perzistentní identifikátor. Je důležité, aby zdroj identifikovaný perzistentním identifikátorem nebyl nikdy přesunut nebo odstraněn, aniž by došlo k aktualizaci informací o jeho umístění v registru perzistentních identifikátorů.

Rozšiřitelnost

Tento systém musí být rozšiřitelný a schopný zahrnout všechny zdroje, které požadují identifikátor.

Flexibilita

Identifikační systém bude mnohem efektivnější, pokud je schopen vyhovět speciálním požadavkům pro různé typy materiálů nebo sbírek. Přístup „jedna velikost pro všechny" není vždy nejpraktičtější. Přiměřenou úroveň inteligence pro podporu procesů a systémů pro různá použití lze začlenit tam, kde je to užitečné, a vynechat ji v případě, kde to není účelné.

Snadnost použití

Přestože není absolutně nejdůležitější a nezbytná pro strojově generované perzistentní identifikátory, systém bude obecně úspěšnější, pokud je snadno pochopitelný a použitelný a pokud umožňuje snadné použití citací. Je tedy žádoucí pokud možno zabránit tvorbě příliš dlouhých identifikátorů, které lze tolerovat tehdy, pokud jsou určeny pouze pro interní potřebu nebo strojové zpracování.

Rozhodovací kritéria pro výběr perzistentního identifikátoru (PID) pro NUŠL

Typ identifikátoru

Perzistentní identifikátor může představovat buď náhodný výběr znaků bez jakýchkoliv souvislostí, které neobsahují žádné informace o objektu a určuje je náhodný řetězec znaků (anglicky „dumb number"). Nebo se může jednat o nějaký systém, který je vytvořen za účelem identifikace. Takovýto systém inteligentních identifikátorů může mít různou mírou složitosti a konkrétnosti. Obecně lze říci, že lidem se lépe pamatují a používají inteligentní identifikátory s vestavěnou mnemotechnikou než bezvýznamná posloupnost znaků, ale pro strojní zpracování je toto hledisko irelevantní. U inteligentních identifikátorů lze také snáze dosáhnout jedinečnosti, a to nejen mezi různými organizačními jednotkami, ale zejména globálně. Generování náhodných řetězců znaků musí být naopak pečlivě kontrolováno a sledováno v celé organizaci s cílem zajistit jedinečnost, kterou neřeší mimo organizaci. Inteligentní identifikátor může obsahovat relační informace, kdy menší komponenty jsou určeny odkazem na větší subjekty, nebo sbírky, kterých jsou součástí.

Varianty:
- Náhodný řetězec znaků („dumb number")
- Inteligentní identifikátor

Hierarchie

Jedním z nejběžnějších použití inteligentních identifikátorů v knihovním světě je začlenění relačních informací, které zobrazují organizaci a hierarchie digitálních sbírek nebo agregátů. Menší složky, jako jsou digitální obrazy stránek nebo agregáty nižší úrovně, jsou identifikovány odkazem na větší sbírky nebo subjekty, ke kterým patří. Tato forma identifikace se hodí zejména pro digitální náhrady fyzických sbírek, i když se může také použít k zobrazení hierarchie digitální sbírky, jako jsou čísla časopisů a jejich články. Tento systém hierarchie ale není nutný, v případě že každá úroveň např. časopisu (titul, ročník, číslo, stránka) má vlastní identifikátor, není třeba, aby syntaxe identifikátoru poskytovala na první pohled informaci o hierarchii. Většina uživatelů takto neuvažuje a problémem se může stát to, když se hierarchie změní (u časopisů asi ne, ale např. u uměle vytvářených sbírek určitých dokumentů).

Varianty:
- Podporovat
- Nepodporovat

Granularita (úroveň odkazu)

Rozhodnutí o úrovni detailu odkazu, na které budou perzistentní identifikátory přidělovány, závisí na vnímání potřeb ukládaného materiálu. Granularita se bude lišit podle různých použití a materiálů. Pro mnohé potřeby vystačí citace prostřednictvím webové stránky nejvyšší úrovně, která slouží jako vstupní bod na sbírku webových souborů s vlastními odkazy, nebo odkazem na článek v časopise, či stránky nebo kapitoly v knize. Nicméně některá použití mohou vyžadovat jemnější úroveň detailů. Granularitu si musí určit každá instituce pro jednotlivé typy dokumentů sama.

Příklad monografie:
- Celek

- Kapitoly
- Stránky
- Obrázky

Verze

Verze na první pohled stejného dokumentu může být odlišná různými způsoby, může mít jiný obsah, jiný formát, nebo jiné rozlišení ve stejném formátu. Každá verze objektu, pokud sledujeme verze, vyžaduje samostatný perzistentní identifikátor. Vztah mezi verzemi může být vyjádřen v identifikátoru pomocí kódu verze nebo datováním verze nebo kódem typu verze nebo v metadatech. Je tedy nutné zvážit, zda a jak zaznamenávat vztahy mezi jednotlivými verzemi.

Varianty:
- Nesledovat verze
- Obsah
- Formát
- Rozlišení ve stejném formátu

Zajišťující autorita

Na zajišťující autoritě závisí nejdůležitější úkol pro úspěšný systém PID a to organizační zajištění. Zajišťující autorita garantuje perzistentnost identifikátoru, čímž na sebe bere zodpovědnost za jeho dlouhodobé přetrvání. Zajišťující autorita zajistí správu resolveru pro perzistentní identifikátory – přidělování jmenných prostorů různým institucím a hlavně údržbu registru vztahů mezi PID a digitálními objekty a aktuálnost tohoto registru. Přidělování samotných PID může zajišťovat centrální instituce, nebo pak instituce s vlastním jmenným prostorem sama (viz dále). Zajišťující autorita nemusí být specializovaná na šedou literaturu, ale musí být schopna zajistit vybrané schéma PID a jeho perzistentnost.

Varianty:
- Státní technická knihovna
- Národní knihovna České republiky
- Jiná

Doporučení identifikátorů pro lokální úložiště

Vzhledem k tomu, že cílem projektu jsou doporučení pro lokální úložiště šedé literatury, zahrnuli jsme tento cíl též do tohoto rozhodovacího dokumentu. Doporučení bude brát ohled na co nejjednodušší implementaci pro lokální úložiště ŠL a zajištění kompatibility s NUŠL.

Varianty:
- Sdílet identifikátory s NUŠL
- Vlastní jakékoliv identifikátory
- Vlastní identifikátory dle koncepce NUŠL

Generování identifikátorů pro lokální úložiště

Na doporučení identifikátorů pro lokální úložiště navazuje rozhodnutí o zajištění jejich generování, které bude vybráno v závislosti na zvoleném schématu a centrální autoritě.

Varianty:
- Dostane přidělen seznam
- Bude mít někde prostor pro generování
- Bude si muset nainstalovat SW pro generování

Dostupnost řešení

Toto kritérium je důležité z hlediska harmonogramu projektu. Vybrané schéma perzistentního identifikátoru musí být dostupné včetně revolveru v závislosti na plánu implementace SW nejpozději do konce roku 2009.

Varianty:

- Ihned
- V průběhu roku 2009
- Vyvíjené řešení

Služby resolveru

Hlavní funkcí resolveru je údržba registru vztahů mezi PID a digitálními objekty a přidělování nových jedinečných PID. Dále poskytuje resolver různé služby jako jsou vyhledání a dodání metadat nebo samotného dokumentu apod.

Varianty:

- Vyhledání platné url adresy
- Dodání metadat
- Dodání digitálního objektu
- Spolupráce s lokálními resolvery

Finanční hledisko

Z dlouhodobé perspektivy není možné ani opomenout finanční hledisko, které budeme zvažovat v rámci možností rozpočtu.

Varianty:

- Pořizovací náklady dále zdarma
- Pořizovací a udržovací náklady
- Pravidelné roční příspěvky
- Zdarma

Finanční hledisko lokálních úložišť

Pro lokální úložiště bude jistě finanční hledisko velmi důležité, proto jsme ho také zařadili do rozhodovacího dokumentu.

Varianty:

- Pořizovací náklady dále zdarma
- Pořizovací a udržovací náklady
- Pravidelné roční příspěvky
- Zdarma

Rešerše

BELINI, Emanuele; CIRINNA, Chiara; LUNGHI, Maurizio. *Persistent Identifiers for Cultural Heritage* [online]. Fondazione Rinascimento Digitale. [cit. 2008-11-09]. Dostupný z www: <http://www.digitalpreservationeurope.eu/publications/briefs/persistent_identifiers.pdf>.

CUBR, Ladislav.; MELICHAR, Marek; HUTAŘ, Jan. Stav implementace perzistentních identifikátorů v NK ČR a výhled do budoucnosti. In *Seminář ke zpřístupňování šedé literatury 2008 : 1. ročník semináře zaměřeného na problematiku uchovávání a zpřístupňování šedé literatury, 8. 10. 2008* [online]. Praha : Státní technická knihovna, 2008. Dostupný z WWW: <http://nusl.stk.cz/images/PID_text.pdf>. ISSN 1803-6015.

HILSE, Hans-Werner; KOTHE, Jochen. *Implementing Persistent Identifiers : Overview of concepts, guidelines and recommendations* [online]. London : Consortium of European Research Libraries; Amsterdam : European Commission on Preservation and Access, C2006. [cit. 2008-11-09].]. Dostupný z www: <http://webdoc.sub.gwdg.de/edoc/ah/2006/hilse_kothe/urn%3Anbn%3Ade%3Agbv%3A7-isbn-90-6984-508-3-8.pdf>.

NLA Guidelines for the Development and Application of a Persistent Identifier Scheme for Digital Resources : Appendix 1 [online]. Canberra : National Library of Australia, c2008. [cit. 2008-11-09]. Dostupný z www: <http://www.nla.gov.au/initiatives/persistence/PIappendix1.html>.

NLA Guidelines for the Development and Application of a Persistent Identifier Scheme for Digital Resources : Appendix 2 [online]. Canberra : National Library of Australia, c2008. [cit. 2008-11-09]. Dostupný z www: <http://www.nla.gov.au/initiatives/persistence/PIappendix2.html>.

Persistent identifiers [online]. Canberra : National Library of Australia, c2008. [cit. 2008-11-09]. Dostupný z www: <http://www.nla.gov.au/initiatives/persistence.html>.

Persistent Identification Systeme : Report on a consultancy [online]. Conducted by Diana Dack. Canberra : National Library of Australia, 2001. [cit. 2008-11-09]. Dostupný z www: <http://www.nla.gov.au/initiatives/persistence/PIcontents.html>.