Stav implementace perzistentních identifikátorů v NK ČR a výhled do budoucnosti

Ladislav Cubr
Marek Melichar
Jan Hutař

Národní knihovna Praha
jmeno.prijmeni@nkp.cz

PID obecně

V příspěvku se budeme zabývat problematikou perzistentních identifikátorů (PID) v digitálním světě. Začněme definicí: identifikátor je řetězec znaků, který reprezentuje určitý objekt. Jako takový má význam jen v určitém kontextu, přesněji řečeno systému. Tomuto systému slouží obvykle pro jednoznačnou identifikaci objektu, pořádání nebo klasifikace. Např. „IN-3115b“ je řetězec znaků, který nám neřekne nic. Teprve pokud známe systém (kontext), ve kterém je použit (např. konkrétní knihovna), zjistíme, že jde o signaturu pro konkrétní (a žádnou jinou) knihovní jednotku. Samotný identifikátor bez znalosti systému, který ho vytvořil, je tedy k ničemu. Systém mu dodává význam a udržuje jeho platnost.

Aby měl identifikátor nějakou hodnotu a plnil svůj účel, musí systém a instituce, ve kterém vznikl, zajistit následující:

1. Perzistenci (trvání) vztahu mezi identifikátorem a objektem – identifikátor bude mít smysl tak dlouho, jak dlouho vydrží vztah mezi ním a objektem
2. Jedinečnost PID - jeden identifikátor musí odkazovat jen na jeden objekt v určitém kontextu[1]
3. Perzistenci objektu samotného - tu zajišťuje daná digitální knihovna či instituce, která digitální objekt ukládá. Zanikne-li takový objekt, význam PID se přirozeně snižuje, ale stále si drží jistý význam jako informace z metadat.

Díky povaze digitálního světa může systém spravující PID objekt nejen identifikovat (tj. poskytnout metadata), ale může zajistit přímo získání objektu (poskytnout data).

PID mají dlouhou tradici i v oblasti knihovnictví. Již v roce 1898 zavedla Kongresová knihovna systém LCCN (The Library of Congress Card Number) k identifikaci a kontrole katalogizačních lístků[2]. Od roku 1970 funguje celosvětově systém PID pro knihy pod označením ISBN a další. Bohužel využití těchto tradičních systémů v digitálním prostředí není z mnoha důvodů jednoduché a mnohdy ani možné.

Systém PID pro digitální objekty v praxi

Co je třeba pro zavedení systému PID pro digitální objekty? Nejprve musíme rozhodnout, jaké objekty nebo části objektů chceme identifikovat (granularita - tj. např. bude identifikátor označovat celou knihu nebo její jednotlivé kapitoly nebo dokonce stránky?). Poté vybraným

[1] je ovšem možné, že jeden objekt bude mít dva identifikátory, každý ovšem přidělený v jiném systému
[2] http://www.loc.gov/marc/lccn.html

objektům přidělíme PID[3]. Z PID vytvoříme registr, který bude obsahovat informace o aktuálním umístění označovaných objektů (URL). Registr je nutné udržovat stále aktuální. Dále je třeba zajistit službu, která na základě zadání PID digitální objekt vyhledá, a buď ho přímo dodá, nebo dodá metadata, tj. informace o místě uložení, případně další informace (resolver). Nejnáročnějším úkolem je zajistit dlouhodobé uchovávání a správu registru. Toto vše musí být zajištěno dlouhodobě.

V současnosti existuje několik systémů PID určených primárně pro digitální objekty. Ačkoliv mají rozdílné architektury, všechny jsou založeny na tomtéž principu – centrální autorita spravuje registr PID a zajišťuje získání objektu na základě zadání PID. Nejznámější jsou: DOI (správcem je DOI Foundation) – např. pro články v agregátorech nebo databázových centrech (např. EBSCO), ARK (California Digital Library), PURL (OCLC), Handle (CNRI) nebo URN:NBN (IANA a jednotlivé národní knihovny).

Situace v ČR

V roce 2007 byla vytvořena pracovní skupina pro PID, která měla za úkol zjistit, jaká jsou očekávání, požadavky a zkušenosti zainteresovaných institucí v této oblasti. Byl vytvořen web pro sdílení informací a společnou práci[4]. Velmi záhy se ukázalo, že původní představa, že se na základě kritérií a požadavků vybere jeden systém, který by vyhovoval všem, není reálná. Není možné vyhovět všem požadavkům v rámci jednoho systému identifikátorů. Každá knihovna i projekt má svoje specifika, která se musejí odrážet i v případné implementaci. Jak ukazují zkušenosti ze zahraničí, dokonce i v rámci jedné instituce, např. knihovny, je běžné, že se provozuje více systémů identifikátorů. Dnes je již zřejmé, že tomu tak bude i v Národní knihovně v Praze (dále NK ČR), kde se počítá s URN:NBN a zároveň v projektu Kramerius je již využíván systém Handle.
Od příspěvku „Perspektivy trvalých identifikátorů v ČR“ z konference Knihovny současnosti v Seči 2007[5] se situace v ČR posunula o krok dál. Zatímco v roce 2007 ještě žádná česká instituce systémy PID plnohodnotně nevyužívala, tedy tak že by přidělovala identifikátory a zároveň je resolvovala, o rok později nastaly první změny. Do open source SW Kramerius byla implementována možnost využití systému Handle. Na ÚVT Karlovy univerzity se chystají do konce roku 2008 nasadit systém Handle pro DigiTool.

Handle je od března 2008 využíván v NK ČR v projektu Kramerius. Tamní Kramerius je napojen na server systému handle.net (verze 3.1.0). Pro používání systému Handle je nutná registrace prefixu u CNRI pro každou instanci Krameria a použití systému je zpoplatněno. Prostřednictvím identifikátorů systému Handle jsou v Krameriu označovány monografie i periodika. Identifikace jde až na nejnižší úroveň popisu (titul> ročník> výtisk> strana> obrázek).

Národní knihovna a URN:NBN

NBN je jmenný prostor v rámci celosvětového identifikačního systému URN, který umožňuje globálně identifikovat jakýkoliv objekt jakéhokoliv typu na světě. Podsystém NBN zaregistrovala

[3] Volba znaků pro řetězec identifikátoru je oblastí, kterou se zde nebudeme podrobněji zabývat. Existují různá doporučení, která je dobré zvážit:
http://wiki.dlib.indiana.edu/confluence/display/INF/Filename+Requirements+for+Digital+Objects
[4] http://pid.ndk.cz/
[5] Hutař, J., Coufal, L. Perspektivy trvalých identifikátorů v ČR. Příspěvek na konferenci Knihovny současnosti v Seči 2007 (http://pid.ndk.cz/dokumenty/prezentace-z-konferenci/perspektivy-trvalych-identifikatoru-v-cr)

Finská národní knihovna. NBN systém jako takový je určený národním knihovnám. V pravomoci všech národních knihoven je jmenný prostor „URN:NBN:XX“, kde XX je kódem státu dané národní knihovny dle normy ISO 3166[6]. NK ČR se rozhodla použít systém URN:NBN z důvodu jeho značného rozšíření v zemích Evropské unie. Několik významných evropských národních knihoven jej už úspěšně využívá (zejména ve Skandinávii, Německu a Itálii) a mají také praxi se zaváděním a provozováním resolveru.
NK ČR je vzhledem k svému statutu národní knihovny jediným tzv. registrátorem první úrovně v rámci jmenného prostoru „cz“. Politika NK při udělování PID bude taková, že za tento jmenný prostor bude přidělovat registrátorům tzv. druhého stupně jejich individuální jmenný prostor. NK si mj. vymezí jmenný prostor NKP pro svoje digitální objekty. Registrátor druhé úrovně, který dostane svůj jmenný prostor přidělen, si pak bude sám rozhodovat o poslední části řetězce URN:NBN, následujícím za pomlčkou za jeho jmenným prostorem. Při výběru znaků pro PID je třeba dodržovat doporučení RFC 2141[7]. Např. Moravská zemská knihovna dostane přidělen jmenný prostor (např. MZK) a přidělí sama poslední část identifikátoru (např. ABC123). Výsledný identifikátor bude vypadat následovně URN:NBN:CZ:MZK-ABC123.

Technologická implementace – Resolver

Pojem resolver je obtížně přeložitelný (možné varianty lok/aliz/átor, analyzátor nebo vyhledávač váží různé konotace), proto jej necháváme v původní podobě. V kontextu systému URN:NBN, který se Národní knihovna chystá implementovat, má resolver řadu funkcí:

- Přiděluje nová jedinečná URN:NBN.
- Udržuje registr vztahů PID-URL-digitální objekt.
- Pomáhá na základě zadání PID vyhledat označované objekty.
- Spravuje PID, tj. sbírá a kontroluje PID od registrátorů druhého řádu.
- Může také fungovat jako záložního archiv digitálních objektů registrátorů druhého řádu.

Architektura a implementace NBN resolveru v NK

NK se rozhodla využít decentralizovaný systém vyvinutý v Itálii, založený na vzájemně sdílených (peer-to-peer) sítích. Národní knihovna bude fungovat jako centrální uzel sítě a bude automaticky sbírat nová nebo aktualizovaná NBN z registrů registrátorů druhé úrovně. Data zkontroluje, vyřeší případy duplicit (stejné NBN pro různé dokumenty, stejná MD5 pro různá NBN), a nakonec bezchybná data zařadí do centrální databáze. Tuto databázi bude průběžně distribuovat všem členům sítě.
Každý registrátor druhé úrovně bude provozovat software, který ho připojí do sítě, umožní mu vytvářet nová NBN, spravovat lokální databázi, podílet se na peer-to-peer lokalizaci NBN. Může také provozovat webové rozhraní pro lokalizaci URN:NBN. Pokud bude hledané NBN spadat do podřízeného jmenného prostoru tohoto registrátora, jeho resolver provede vyhledávání přímo ve svém registru, pokud ne, dotáže se centrálního uzlu nebo jiných registrátorů v síti. Centrální uzel

[6] Všechny ostatní dvoupísmenné kombinace, neobsazené současnými státy jsou vyhrazeny pro případné budoucí státy.
[7] RFC 2141 – viz http://www.ietf.org/rfc/rfc2141.txt

bude synchronizován s podřízenými uzly. Systém funguje i při výpadku centrálního nebo jakéhokoli lokálního registru, vyhledávání pak probíhá přes vzájemnou síť.
Výše popsaný systém je výstupem Italského výzkumného projektu řízeného nadací Fondazione Rinascimento Digitale[8]. Systém vystavěn na základu převzatého ze systému DSpace využívá Javu, databázi PostgreSQL a server Tomcat. V současnosti je k dispozici prototyp aplikace, další vývojová fáze by měla přinést lepší instalační balíček, dokumentaci, doplňky pro komunikaci s jinými národními systémy URN:NBN a pro vyhledávání jiných PID (ARK, DOI).

Systém je vyvíjen jako open source. Národní knihovna ho testuje a hledá partnery ochotné spolupracovat při testování spolupracovat. Tento aplikační systém by měl na celonárodní úrovni zajistit využití identifikátoru URN:NBN, jehož garantem bude Národní knihovna ČR.

[8] http://www.rinascimento-digitale.it/indexEN.php