briefing paper

digital preservation europe

Interoperabilita identifikátorů

Objekty v digitálních sítích pocházejí z různých zdrojů, a bývají označovány různými typy identifikátorů, které jsou jim přiděleny na základě různých oficiálních nebo neoficiálních standardů nebo speciálních katalogizačních systémů. K usnadnění dlouhodobé ochrany digitálních dokumentů a pro potřeby trvalého využívání a výměny informací je velmi důležité, aby uživatelé mohli používat identifikátory (a dokumenty s nimi spojené) napříč různými aplikacemi. Takováto interoperabilita identifikátorů se netýká pouze technologie, ale také dané komunity a účelu, ke kterému tato komunita identifikátory potřebuje.

Interoperabilita

Interoperabilita je schopnost nezávislých systémů vyměňovat si srozumitelné informace a spouštět navzájem mezi sebou určité akce, přičemž toto společné fungování jednotlivých systémů přináší přidanou hodnotu. Jedná se tedy o schopnost volně spojených nezávislých systémů navzájem spolupracovat a komunikovat. Identifikátory jsou lexikální jednotky, které označují objekty v těchto systémech. Objekt je to, co je označeno identifikátorem.

Informační zdroj (digitální dokument) může být součástí několika systémů najednou nebo být identifikován různými systémy různě, a tak je nutné zajistit interoperabilitu jak v rámci různých identifikačních systémů, tak v rámci různých implementací téhož systému trvalé identifikace. Interoperabilita trvalých identifikátorů slouží například pro:

- interoperabilitu metadat (metadata vyjadřují vztah mezi dvěma objekty)
- vytvoření standardních mechanismů pro vyjádření vztahů mezi stejnými objekty označenými různými identifikačními standardy
- vytvoření služby, která stojí nad více než jedním systémem, např. vyhledávání „souvisejícího obsahu“, spojování multimediálních objektů atd.

Několik takových případů interoperability identifikátorů bylo prozkoumáno v projektech ISO TC46SC9 a RIDIR.

Identifikátory, které byly přiděleny v jednom kontextu, mohou být trvale využívány i v kontextu jiném, aniž by bylo třeba obracet se na správce identifikačního systému v původním kontextu. I v případě, že by informační zdroj byl součástí jen jednoho systému, platí, že jakmile je identifikovatelný, může ho identifikovat (a skrytě modifikovat) i jiný systém. Kontext a výchozí podmínky, v kterých byl identifikátor původně přidělen, nemusí uživatel identifikátoru v jiném systému vůbec znát. Určitý systém může například obsahovat pouze abstraktní reprezentace díla, jiný naopak pouze konkrétní vyjádření těchto abstraktních děl. Pokud o sobě oba systémy vědí, mohou si navzájem vyměňovat dodatečné informace o svém zaměření, pokud o sobě však nevědí, je třeba zajistit jejich interoperabilitu jinak.

Můžeme rozlišovat tři typy interoperability:

- Syntaktická interoperabilita. Schopnost systému zpracovat syntaxi znakového řetězce a rozpoznat, že jde o identifikátory (případně zahájit nějakou činnost), a to i v případě, že se v systému vyskytuje více takových syntaxí najednou.
- Sémantická interoperabilita. Schopnost systému zjistit, zda dva různé identifikátory skutečně odkazují k úplně stejnému objektu, a pokud neodkazují, pak zjistit, jaký je vztah mezi těmito dvěma odlišnými objekty.
- Uživatelská interoperabilita. Schopnost systému spolupracovat a komunikovat s využitím identifikátorů, a přitom brát ohled na právní otázky a omezení použitelnosti dokumentů označovaných těmito identifikátory v systémech.

Jde tedy o tři související oblasti: uživatelské interoperability je možné dosáhnout pouze tehdy, když je zajištěna sémantická interoperabilita, a sémantické interoperability lze dosáhnout jen za předpokladu, že je zajištěna interoperabilita syntaktická.

Syntaktická interoperabilita

Syntaktická interoperabilita znamená, že dva systémy přesně dodržují technické specifikace pro zpracování řetězce znaků identifikátorů a tyto specifikace umožňují odhadnout pravděpodobný rozsah možných identifikátorů. V některých případech se mohou vyskytovat pravidla, která určí, jak lze určitý identifikátor vložit do syntaxe jiného identifikačního schématu.

Interoperabilita identifikátoru však může být velmi široká, a pak je těžké odhadnout pravděpodobný rozsah identifikátoru: existují také jiné než „webové“ identifikátory, například v telekomunikačních nebo vysílacích sítích, a jiné

globálně jedinečné identifikátory, např. identifikátor ISBN, který nebyl původně zamýšlen pro použití v digitálním světě. Schémata registrů a výchozí podmínky identifikátoru, např. závislost na nějakém protokolu, se musí definovat nebo musí být možné je dohledat, jinak by nešlo použít identifikátor mimo kontext jednoho katalogu (např. v Dublin Core pole „identifikátor zdroje“, DC element syntax: DC.Identifier). Neexistuje jeden centrální registr, který by obsahoval všechny identifikátory. Některé identifikátory lze registrovat ve schématech URI, jiná schémata identifikátorů však mohou být soukromá nebo obtížně zjistitelná. Jednotný registr jmenných prostorů je – podobně jako domény DNS – součástí specifikace URN (ale moc se nevyužívá). Informační schéma URI vzniklo v knihovnické a vydavatelské oblasti (konkrétně při spolupráci na rozvoji standardů OpenURL), protože existovala potřeba společného identifikačního systému jmenných prostorů typu URI (jako systému prostých identifikátorů: tj. takových, které jsou určeny čistě pro účely identifikace, nikoli vyhledávání, lokalizace nebo získávání objektu). Cílem bylo definovat taková schémata URI, která by odkazovala k informačním zdrojům, které sice mají své identifikátory v rámci veřejných jmenných prostorů, ale nemají implementace v rámci URI, například identifikátory LCCN.
Některé identifikátory jsou čistě abstraktními označeními (jen jména), jiné vznikly s předpokladem dalšího využití, např. vyhledávání, zpřístupňování nebo lokalizace. Tyto výchozí předpoklady mohou mít dalekosáhlý dopad: například u specifikace URI se předpokládá, že identifikátory URI budou využity k zpřístupnění objektu, a síťový odkaz identifikátoru URI je implicitně postaven na systému DNS: specifikace URI neobsahuje žádná ustanovení pro jiné fungování než na základě DNS. V případě, že identifikátory explicitně obsahují nebo implicitně předpokládají nějaký konkrétní protokol, lze využít mechanismus „proxy“ (který překládá jeden protokol do jiného), který zajistí syntaktickou interoperabilitu.

Sémantická interoperabilita

Sémantická interoperabilita řeší běžný, ale obtížný problém. Znakové řetězce identifikátoru jednoho systému sice může jiný systém zpracovat po syntaktické stránce, ale jak jiný systém zjistí, co tyto znaky vlastně znamenají? Pokud systém A řekne „majitel“ a B řekne „majitel“, mluví skutečně o tomtéž? A pokud A řekne „vydáno“ a B „zpřístupněno“, znamená každý výraz něco jiného? Pro účinné zajišťování interoperability entit je potřeba těchto věcí:

- k jednotnému identifikátoru musí být připojen popis objektu, na který odkazuje, a to s použitím strukturovaného souboru prvků, které poskytují informace o tomto objektu (tzn. aby byl identifikátor interoperabilní, musí být spojen s nějak strukturovanými metadaty)
- jediným způsobem, jak jednoznačně rozhodnout, že jeden identifikátor odkazuje k témuž objektu jako jiný identifikátor, jakkoliv jsou oba identifikátory odlišné, je, že oba identifikátory musí sdílet stejný referenční rámec. K tomu je zapotřebí strukturovaná ontologie (explicitní formální specifikace toho, jak reprezentovat objekty, které jsou součástí daného systému, a vztahů mezi nimi), obecný model, který umožňuje generovat konzistentní nové vztahy, a metoda, jak zaznamenat dohody mezi systémy, jejichž identifikátory jsou použity.

Dvě hlavní iniciativy pracující na ontologiích, které umožňují srovnávání identifikátorů ve společném referenčních rámci, jsou konceptuální referenční model CIDOC a rodina aplikací odvozených z projektů sémantické interoperability založených na <indecs> (jako je například ONIX). Obě iniciativy mají mnoho společného a existují snahy prozkoumat jejich podobnosti s knihovnickými aktivitami typu RDA. Byla zahájena činnost společné iniciativy směřující k vytvoření sdíleného rámce pro kategorizaci zdrojů. Ontologie se sice ještě příliš nevyužívají pro plně automatické operace (které se očekávají v sémantickém webu), ale již se využívají v oblasti multimediálních metadat a schémat pro messaging, kde zajišťují jednoznačné, rozšiřitelné a přesné definice výrazů.

Uživatelská interoperabilita

Schémata identifikátorů mohou reprezentovat různá uživatelská oprávnění, která omezují nakládání s dokumenty, jež označují. Autorita registrující identifikátory musí zvážit, do jaké míry smí spolupracovat s jinými systémy nebo zveřejňovat dokumenty. Přestože v syntaktické a sémantické rovině nemusí být žádný problém, mohou existovat ještě další překážky bránící plné interoperabilitě. Přidělení a použití daného identifikátoru může být vázáno na uživatelská práva, kvalitu nebo správu dat, řízení a účastnické požadavky. Podobná omezení najdeme v komerčním i nekomerčním prostředí.
Sémantická interoperabilita, která využívá možnosti přiřadit systém k sdílenému ontologickému rámci, bude vyžadovat bilaterální smlouvu mezi dvěma schématy, které potvrdí přesný záměr obou systémů identifikace (nebo poznámku, že přiřazení není autorizováno jednou ze zúčastněných stran). Při používání identifikátorů je prostě třeba zohlednit povinnosti komunity uživatelů.
Každý registr identifikátorů má povinnost zajistit uživatelům patřičná data. Nemůže proto spoléhat na metadata někoho jiného, protože nad nimi nemá plnou kontrolu. Každý registr identifikátorů také musí zajistit, aby se jeho standardy implementovaly na základě obchodního modelu: metadata mají komerční hodnotu, což vede registry k zavádění vlastních standardů. Nelze proto očekávat, že registr poskytne metadata někomu, s kým nemá uzavřenou smlouvu.

Pokud obě strany souhlasí, může být uzavřena bilaterální smlouva, která specifikuje typ spolupráce (například: registrační autority mohou souhlasit se sdílením nebo porovnáváním svých objektů a aktualizací metadat). Pokud strany nesouhlasí, nelze vytvořit závazek interoperability. Systémy identifikátorů by měly takovou spolupráci usnadňovat tím, že poskytnou jasné informace týkající se práv a povinností. Jde často o požadavky na formální standardizační procesy.

Persistence a interoperabilita

Persistence je úzce svázána s interoperabilitou. Persistence je „interoperabilita s budoucností“, tj. nezávislé systémy, které si vyměňují srozumitelné informace a navzájem iniciují nějaké akce, jsou odděleny v čase.

Text DPE „Persistentní identifikátory pro kulturní dědictví“ se zabývá požadavky a předpoklady trvalých identifikátorů podrobněji. Některé systémy identifikace mohou plnit konkrétní, ovšem relativně krátkodobé potřeby (například sdílení videa v sociálních sítích), jiné se zaměřují na trvalost a ochranu a zavazují se udržovat registry identifikátorů a metadata.

Programový architekt musí zvážit výhody jednotlivých systémů identifikace, a vyhnout se těm, jež nejsou v souladu s jeho cíli. URL se často považují za nejběžnější „identifikátory“ (Viz. definice identifikátoru od DC: „Řetězec používaný k jedinečné identifikaci zdroje. Výchozí identifikátorem pro zdroje je URL.“), přestože ve skutečnosti je URL identifikátorem lokace. V důsledku nejobvyklejšího modelu jednoho přesměrování na jedno URL se snadno zaměňuje zdroj a lokace, vztah mezi identifikátorem a objektem není přímý, a tak se snadno naruší. Používání URL je nenáročné, ale také nepříliš trvalé. Systémy identifikace, které URL nějak vylepšují (PURL, ARK, N2T) nebo se mu zcela vyhýbají (Handle, DOI), jsou z hlediska dlouhodobé ochrany vhodnější.

Byl vytvořen mechanismus, který má podpořit interoperabilitu identifikátorů: v *Architektuře digitálních objektů Kahn/Wilenski* jsou digitální objekty spojeny s metadaty a odkazy na umístění v repozitářích. Tyto digitální objekty nenahrazují existující formáty a datové struktury, pouze poskytují společný rámec pro zabalení těchto formátů a struktur tak, aby byly jednotně interpretovány, a tedy také aby byly přenositelné do různých i z různých heterogenních informačních systémů a bez ohledu na změny způsobené časem. Existuje řada implementací této architektury, ovšem dosud není široce přijímána (přestože Handle identifikátor, součást této architektury, je sám o sobě používán poměrně často).

Poučení

- Snažte se nevynalézat to, co už bylo vynalezeno. Pokud se vám zdá, že potřebujete vytvořit nový systém identifikace, prozkoumejte již existující systémy a ujistěte se, zda váš problém nemůže vyřešit nějaký již existující identifikační systém.
- Pokud potřebujete nový systém identifikace, zjistěte si, zda můžete využít nějaký již existující protokol nebo registr identifikátorů pro implementaci svého systému.
- Při registraci svého systému si zajistěte adekvátní veřejný jmenný prostor.
- Dobře specifikujte strukturované schéma metadat a zveřejněte jej, aby bylo snadné zajistit sémantické přirazení k vašemu systému
- Zvažte uživatelské a komerční potřeby dalších možných uživatelů systému.
- Poskytněte jasné informace týkající se práv a závazků pro uživatele vašeho systému.
- Používejte takové systémy identifikace, které budou mít mechanismus pro zajištění persistence identifikace

Odkazy a zdroje

DPE Briefing paper "Trvalé identifikátory pro kulturní dědictví“
http://www.digitalpreservationeurope.eu/publications/briefs/cz_trvale_identifikatory.pdf

RIDIR project (Resourcing IDentifier Interoperability for Repositories)
http://www.hull.ac.uk/ridir/

<indecs> project (Interoperability of data in ecommerce systems). Metadata Framework: Principles, model and data dictionary.
http://www.doi.org/topics/indecs/indecs_framework_2000.pdf

Identifier Interoperability: A Report on Two Recent ISO Activities.
D-Lib magazine, April 2006
http://www.dlib.org/dlib/april06/paskin/04paskin.html

The RDA/ONIX Framework for Resource Categorization.
D-Lib magazine, Jan/Feb 2007
http://www.dlib.org/dlib/january07/dunsire/01dunsire.html

CIDOC Conceptual Reference Model
http://cidoc.ics.forth.gr/index.html

“Info URI” registration scheme
http://info-uri.info/registry/docs/misc/faq.html

Digital Object Architecture
Kahn R.E. & Wilensky R. "A Framework for Distributed Digital Object Services".
http://www.cnri.reston.va.us/cstr/arch/k-w.html

Norman Paskin, *Tertius Ltd*
n.paskin@tertius.ltd.uk

Překlad NKP