Kartografické listy / Cartographic Letters, 2013, 21 (2), 50-59

NÁSTROJE NA HARMONIZÁCIU GEOGRAFICKÝCH NÁZVOV

Juraj VALIŠ, Juraj STRAKA

Tools for the harmonization of geographical names

Abstract: Data harmonization is one of several requirements that need to be met to support functional data sharing – interoperability. Geographical names related data is used as an example for demonstration of data harmonization capabilities of spatial ETL (Extract Transform Load) tools. The paper presents a reusable sequence of steps leading to creation of schema-valid GML (Geography Markup Language) file. These steps deal mostly with GML instance generation, reformatting of input data according to implementation rules, merging this data into a GML template and finally validating the output for correct syntax and schema. The entire harmonization is carried out on the fly by a single spatial ETL tool.

Keywords: data harmonization, geographical names, INSPIRE directive

Úvod

Geografické informačné zdroje – geografické informácie, geografické informačné systémy (GIS) a geografické informačné služby tvoria spolu s pravidlami prístupu, využitia, zdieľania a s koordinačnými mechanizmami základné prvky priestorových informačných infraštruktúr.

Budovanie alebo aktualizácia takýchto infraštruktúr prebieha v súčasnosti na národnej úrovni členských štátov Európskeho spoločenstva. Tieto aktivity sú výsledkom implementácie smernice Európskeho Parlamentu a Rady 2007/2/ES, ktorou sa zriaďuje Infraštruktúra pre priestorové informácie v Európskom spoločenstve (INSPIRE). Pre budúcich používateľov tejto infraštruktúry to znamená najmä vytvorenie prístupu k priestorovým údajom prevažne zameraných na oblasť životného prostredia. Politika prístupu k týmto harmonizovaným údajom presne definuje podmienky, obsah, formu ako aj ďalšie aspekty ich využívania. Forma údajov sa týka predovšetkým ich dohovoreného referenčného systému, formátu a štruktúry (INSPIRE TWG GN, 2010).

Cieľom príspevku je prezentácia možného praktického postupu transformácie vzorky Geografického názvoslovia ZBGIS® (Základná báza GIS – národný „dataset“) do harmonizovanej formy v súlade s implementačnými pravidlami smernice INSPIRE (európsky „dataset“) na príklade témy Zemepisné názvy.

1. Zemepisné názvy ako priestorové údaje

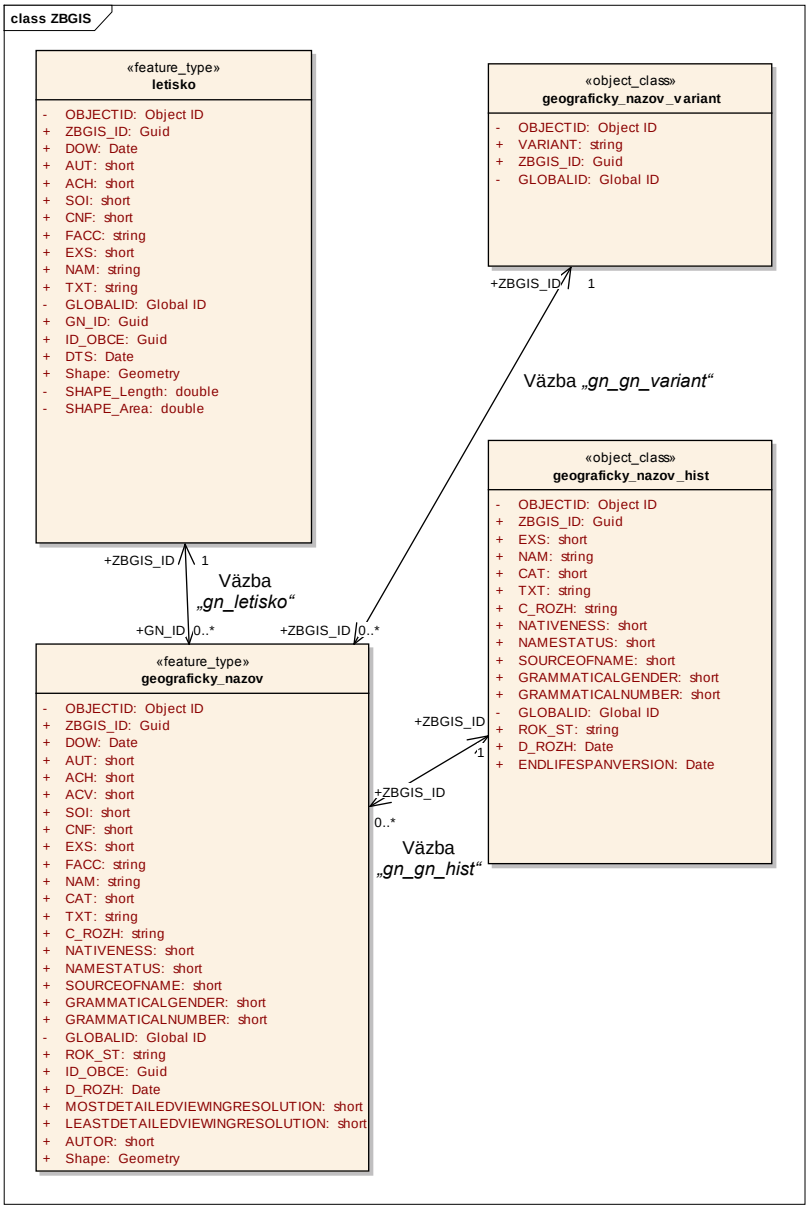
1.1 Zdrojový údajový model

Geografické názvy sú v rámci ZBGIS® reprezentované vo forme jednej hlavnej triedy objektov „geografických názov“ s bodovou geometriou, v súborovej geodatabáze, ktorá je ďalej relačne prepojená na dve tabuľky s historickým a variantným názvoslovím, ako aj na ďalšie triedy objektov. Triedou objektov sa rozumie špeciálny prípad tabuľky, ktorá obsahuje aj stĺpec, kde je uložená geometrická časť priestorového údaju (používa sa špeciálny údajový typ „Geometry“). Tým pádom je možné údaje uložené v triede objektov zobraziť v mapovom okne. Popri triedach objektov môžu v súborovej geodatabáze existovať aj obyčajné tabuľky, ktoré nemôžu obsahovať stĺpec s geometriou, a teda nie je možné ich zobraziť vo forme mapy, je možné ich prehliadať iba vo forme bežného tabuľkového náhľadu. V príklade na obr. 1 je uvedená trieda objektov letisko. Prehľad väzieb uvádza tab. 1 upravená podľa zdroja z Úradu geodézia kartografie a katastra Slovenskej republiky (ÚGKK SR, 2013).

Ing. Juraj VALIŠ, PhD., Mgr. Juraj STRAKA, Prírodovedecká fakulta UK, Mlynská dolina, 842 15 Bratislava, e-mail: valis@fns.uniba.sk, straka@fns.uniba.sk

Tab. 1 Prehľad relačných väzieb pre geografické názvoslovie

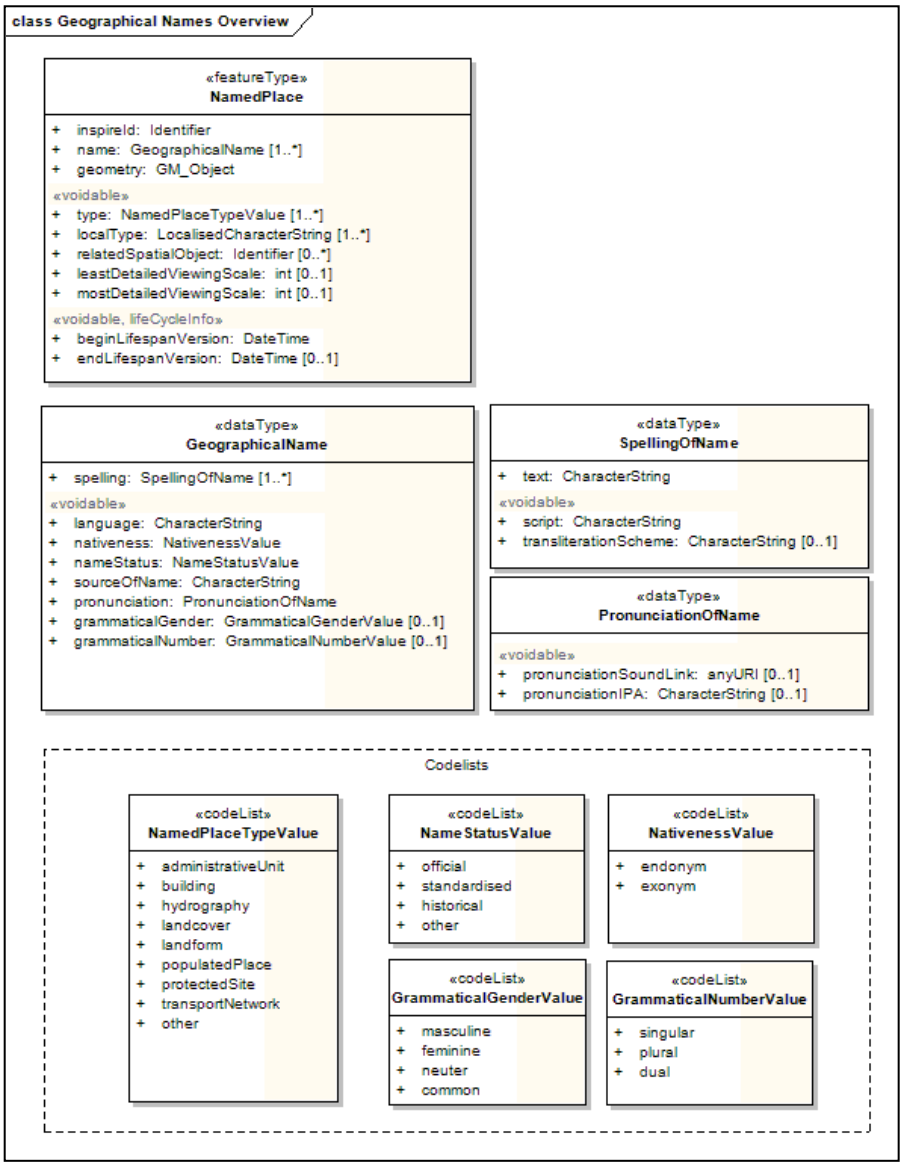
Názov väzby	Kardinalita	Zdrojová trieda objektov	Primárny kľúč	Cieľová trieda objektov	Cudzí kľúč
gn_letisko	1:N	geograficky_nazov	ZBGIS_ID	letisko	GN_ID
gn_gn_hist	1:N	geograficky_nazov	ZBGIS_ID	geograficky_nazov_hist	ZBGIS_ID
gn_gn_ variant	1:N	geograficky_nazov	ZBGIS_ID	geograficky_nazov_variant	ZBGIS_ID



Obr. 1 Zdrojová schéma – Geografické názvy ZBGIS® (upravené podľa ÚGKK SR, 2013)

1.2 Cieľový údajový model

Schéma INSPIRE GeographicalNames predstavuje cieľový konceptuálny údajový model. Tento model je možné vyjadriť s použitím formálneho modelovacieho jazyka UML ako tzv. UML diagram (obr. 2). Pod skratkou UML sa rozumie „Unified Modeling Language“ (čo možno voľne preložiť ako zjednotený jazyk na modelovanie). Jeho vývoj zastrešuje konzorcium Object Management Group (známe aj pod skratkou OMG®). Momentálne poslednou vydanou verziou je 2.4.1. V súčasnosti je prijatá aj v podobe ISO normy 19505 a jednotlivé súčasti špecifikácie OMG je možné prevziať z webovej stránky: *http://www.omg.org/spec/UML/2.4.1/*. Vyjadrovanie sa pomocou diagramov je typické pre grafické jazyky na tvorbu konceptuálnych schém, čo výrazne zjednodušuje čitateľnosť návrhu (INSPIRE DRAFTING TEAM, 2013).



Obr. 2 Cieľová schéma – INSPIRE Geographical Names (upravené podľa INSPIRE TWG GN, 2010)

Ako uvádza Sagris et al. (2013) na základe vytvoreného konceptuálneho modelu sa dá vytvoriť XML/GML schéma, nazývaná aj ako konceptuálna schéma, resp. iný zahraničný termín ju označuje ako „Geospatial Community Data Specification“. Skratkou XML sa rozumie „eXtensible Markup Language“, čiže rozšíriteľný značkovací jazyk, ktorý slúži najmä ako jednoduchý textový formát pre štruktúrované údaje. Technológia XML je štandardom W3C konzorcia a viac o nej je možné dozvedieť sa na webovej stránke: *http://www.w3.org/standards/xml/*. Skratkou GML sa rozumie „Geography Markup Language“, čo je špeciálny prípad formátu XML pre kódovanie priestorových údajov. Ide o štandard konzorcia OGC (Open GIS Consortium), ktorý je prijatý vo forme štandardu tohto konzorcia (*http://portal.opengeospatial.org/files/?artifact_id=20509*), a tiež vo forme ISO normy (ISO 19136). Vo všeobecnosti je možné GML schému (špeciálny prípad schémy XML) vytvoriť viacerými spôsobmi. Môže ju vytvoriť programátor priamou editáciou ako postupnosť rezervovaných slov povolených pre definíciu schémy XML. Iný prístup, nemenej programátorsky náročný, umožňuje metodikami softvérového inžinierstva uskutočniť konverziu diagramu UML tried do schémy XML. Uvedeným prístupom sa zaoberajú napr. (Gherabi a Bahaj, 2011, Routledge et al., 2002). Štruktúra skutočnej geografickej databázy môže byť opísaná aplikačnými modelmi a aplikačnými schémami. Pre účely vyhodnotenia súladu implementovanej databázy by mala byť aplikačná schéma „namapovaná“ na konceptuálnu schému. V prípadoch, keď je treba integrovať rôzne „datasety“ od rôznych organizácií/inštitúcií, môžu byť údaje transformované do štruktúry konceptuálnej schémy využitím „mapovacích“ parametrov. Konceptuálne modely sú predmetom dohody medzi členmi geoinformačných komunít, ktorými môžu byť používatelia, správcovia, resp. poskytovatelia údajov.

Konceptuálne schémy komunity regulovanej smernicou INSPIRE, odvodené z konceptuálnych modelov, sa dajú získať priamym prevzatím vo forme Aplikačných schém GML na webovej stránke: *http://inspire.jrc.ec.europa.eu/index.cfm/pageid/541/downloadid/1698*.

Pre tému INSPIRE GeographicalNames je takáto schéma prezentovaná jedným súborom s príponou XSD. Schému sme použili na generovanie prázdnej inštancie GML pre typ priestorového objektu NamedPlace. Následným napĺňaním prázdnych inštancií GML (hodnotami atribútov a geometrie z objektov zdrojového údajového modelu ZBGIS®) sa dá vytvoriť konečný súbor GML, ktorý obsahuje transformované údaje v harmonizovanej forme. Okrem uvedenej štruktúry typov priestorových objektov, atribútov a zoznamov kódov, zahŕňajú implementačné pravidlá aj požiadavky na súradnicový referenčný systém a formát údajov v harmonizovanej forme. Súradnicový systém, použitý na sprístupňovanie harmonizovaných údajov, má byť založený na referenčnom systéme ETRS89 pre všetky lokality definované jeho geografickým rozsahom. Zároveň majú byť údaje kódované vo formáte GML verzie 3.2.1.

2. Kódovanie priestorových údajov v jazyku GML

2.1 Schémy jazyka GML

Jazyk GML definuje rôzne prvky schém XML, ako napr. objekty, geometriu a topológiu prostredníctvom hierarchie objektov GML. Štandard GML poskytuje niekoľko schém pre opis geografických údajov v jazyku XML. Schémy opisujú viaceré aspekty, ako napr. geometriu, topológiu, hodnoty, súvislé povrchy, časové a priestorové referenčné systémy apod. Avšak táto základná skupina schém poskytuje len všeobecnú oporu pre prvotné štruktúry, ktoré konkrétna „aplikačná schéma“ môže používať, rozširovať alebo vytvárať zložené objekty apod. Vlastná aplikačná schéma opisuje konkrétne typy objektov pre konkrétnu aplikačnú doménu (v našom prípade pre oblasť Zemepisných názvov). Hlavným prínosom uvedeného prístupu je značná miera flexibility, resp. univerzálnosť použitia pri prezentácii širokého spektra priestorových objektov (Lu et al., 2007).

2.2 Ukážka prevodu zo súboru XSD do súboru GML

Na krátkom príklade prezentujeme princíp generovania prázdnej inštancie GML podľa pravidiel aplikačnej schémy (súbor XSD). Príklad bol vytvorený na základe podkladového postupu publikovaného Konečným a Kubíčkom (2012). Ako autori uvádzajú „*V dokumente XSD sú definované okrem iného nasledujúce pravidlá: jednotlivé typy elementov, ich obsah (ďalšie podelementy), vzájomné poradie elementov, max. a min. počet jednotlivých elementov, atď.*“ (Konečný a Kubíček, 2012, s. 36).

Uvádzame ukážku zápisu pravidiel pre inštanciu obsahujúcu informácie o výslovnosti zemepisného názvu:

```
<element name="PronunciationOfName" substitutionGroup="gml:AbstractObject" ty-
pe="gn:PronunciationOfNameType">
  <complexType name="PronunciationOfNameType">
    <sequence>
      <element minOccurs="0" name="pronunciationSoundLink" nillable="true">
        <complexType>
          <simpleContent>
            <extension base="anyURI">
              <attribute name="nilReason" type="gml:NilReasonType"/>
            </extension>
          </simpleContent>
        </complexType>
      </element>
      <element minOccurs="0" name="pronunciationIPA" nillable="true">
        <complexType>
          <simpleContent>
            <extension base="string">
              <attribute name="nilReason" type="gml:NilReasonType"/>
            </extension>
          </simpleContent>
        </complexType>
      </element>
    </sequence>
  </complexType>
```

Konkrétna inštancia GML „PronunciationOfName“ bude teda obsahovať dva subelementy „pronunciationSoundLink“ a „pronunciationIPA“, ktoré by mali obsahovať prepojenie na zvukový záznam náležitej výslovnosti zemepisného názvu a taktiež zápis výslovnosti s použitím znakov medzinárodnej fonetickej abecedy. Subelementy inštancie GML môžu byť vnorenými subinštanciami alebo vlastnosťami danej inštancie. Odlíšenie inštancie od vlastnosti je realizované odlišným spôsobom zápisu – zatiaľ čo názvy vlastností začínajú malým písmenom, názvy elementov korešpondujúcich s inštanciami tried GML začínajú veľkým písmenom (Lu et al., 2007). Vzhľadom na pozitívnu hodnotu atribútu XML „nillable“ pre obidva subelementy nie je ich vyplnenie povinné, avšak v tomto prípade by mal v atribúte „nilReason“ uvedený dôvod, prečo je tomu tak.

Dohovorené hodnoty dôvodov nevyplnenia elementov v súčasnosti zahŕňajú tieto prípady (INSPIRE DRAFTING TEAM DATA SPECIFICATION, 2013):

– správca údajov prirodzene nezbiera hodnotu tejto vlastnosti (vyplní sa „Unpopulated“),
– správca údajov nepozná hodnotu vlastnosti pre konkrétny objekt (vyplní sa „Unknown“),
– správca údajov si neželá hodnotu vlastnosti zverejniť (vyplní sa „Withheld“).

V nasledujúcej ukážke je znázornené, ako by mohla byť vyplnená inštancia GML. Vyplnený je len element s fonetickým zápisom výslovnosti, keďže správca údajov neeviduje odkaz na zvukový záznam výslovnosti názvu:

```
<gn:PronunciationOfName>
       <gn:pronunciationSoundLink xsi:nil="true" ilReason="Unpopulated">
       </gn:pronunciationSoundLink>
       <gn:pronunciationIPA xsi:nil="false">
              viena
       </gn:pronunciationIPA>
</gn:PronunciationOfName>
```

Výsledné štruktúry GML môžu vytvárať veľké súbory (Kubíček a Konečný, 2012), ale dajú sa účinne komprimovať napr. konverziou do formátu Gzip (Lu et al., 2007) a dosiahnuť veľkosť súborov s pôvodným binárnym kódovaním, ktoré boli transformované do formátu GML. Východiská k migrácii údajov medzi zdrojovou a cieľovou schémou opisuje nasledujúca kapitola.

3. Transformácia údajov

Vyššie uvedené rozdielne typy údajových modelov sťažujú migráciu údajov najmä preto, že tieto modely nie sú vzájomne dobre prepojiteľné. Riešením je porovnanie spoločných atribútov medzi vstupnou aj výstupnou schémou (napr. pomocou porovnávacích tabuliek) a následná transformácia údajov, napr. pomocou nástroja ETL (Safe Software, 2012). Skratkou ETL sa rozumie postupnosť operácií Extract, Transform a Load, teda získanie požadovanej časti alebo celku spracovávaných údajov, ich transformácia podľa potreby a uloženie do cieľového úložiska s definovaným formátom a súradnicovým systémom.

3.1 Porovnávacie tabuľky

V tejto kapitole uvádzame stručný prehľad vzájomných spoločných atribútov medzi výstupným a vstupným údajovým modelom (tab. 2). Porovnávacia tabuľka znázorňuje prepojenie, resp. vzájomné „namapovanie“ údajových štruktúr na úrovni tried objektov a atribútov predstavujúce v hrubých rysoch transformáciu údajov.

Pre detailnejšiu špecifikáciu transformácie údajov do cieľového údajového modelu je treba opísať aj prepojenie hodnôt atribútových domén. V našom experimente sme realizovali párovanie číselníka kategórie zemepisného názvu (doména hodnôt pre atribút CAT v ZBGIS®) na príslušnú hodnotu číselníka *NamedPlaceTypeValue* v schéme INSPIRE vzťahom 1:1 (každej hodnote zo zdrojového číselníka bola priradená nová hodnota z cieľového číselníka). Vzhľadom na veľký rozsah položiek pôvodnej atribútovej domény uvádzame príklad párovania niektorých hodnôt v tab. 3.

Na základe uvedeného prehľadu sa dá prispôsobiť priestorový nástroj ETL a hodnoty spoločných atribútov preniesť zo zdrojového do cieľového údajového modelu.

3.2 Nástroj ETL

Vlastný prenos, resp. migráciu údajov sme uskutočnili v prostredí Safe Software Feature Manipulation Engine (FME). Ide o prostredie prispôsobené pre návrh nástrojov ETL s cieľom hromadného spracovania údajov. V procese ich spracovania sa dajú reťaziť viaceré operácie, ktoré vedú k vytvoreniu výsledného súboru GML v takej podobe, v ktorej by ho poskytla Ukladacia služba INSPIRE.

Samotná štruktúra budúceho súboru GML je daná aplikačnou schémou GML (uvedenou v kapitole 2), pomocou ktorej sa dá vytvoriť prázdna inštancia elementu ľubovoľnej úrovne. Keďže hlavným typom priestorového objektu pre tému INSPIRE GeographicalNames je *NamedPlace*, bola generovaná prázdna inštancia tejto triedy GML, ktorá obsahovala všetky subelementy a atribúty predpísané v konceptuálnej schéme. Výrez z nej je znázornený nižšie:

```
   <gn:NamedPlace gml:id="">
     <gn:beginLifespanVersion></gn:beginLifespanVersion>
     <gn:endLifespanVersion xsi:nil="" nilReason=""></gn:endLifespanVersion>
     <gn:geometry>
       <gml:Point gml:id="Point_" srsName="urn:ogc:def:crs:EPSG::4258"
srsDimension="2">
         <gml:pos></gml:pos>
       </gml:Point>
     </gn:geometry>
     <gn:inspireId>
       <base:Identifier>
         <base:localId></base:localId>
         <base:namespace></base:namespace>
         <base:versionId></base:versionId>
       </base:Identifier>
     </gn:inspireId>
   ...
```

Tab. 2 Porovnanie spoločných atribútov cieľovej a zdrojovej schémy

INSPIRE Geographical Names			Geografické názvy ZBGIS		
tabuľka	*atribút*	*údajový typ*	*tabuľka*	*atribút*	*údajový typ*
NamedPlace	geometry	GM_Object	geograficky_nazov	SHAPE	Geometry
	inspireId	Identifier		čiastočne ZBGIS_ID	GUID
	name	Geographical Name		NAM	Text
	Begin Lifespan Version	DateTime		DTS	Date
	End Lifespan Version	DateTime			
	leastDetailed Viewing Resolution	MD _Resolution	geograficky_nazov	LEAST DETAILED VIEWING RESOLUTION	Short Integer
	mostDetailed Viewing Resolution	MD _Resolution		MOST DETAILED VIEWING RESOLUTION	Short Integer
	type	NamedPlace TypeValue		CAT	
Geographical Name	spelling	Spelling OfName	geograficky_nazov	NAM	Text
	language	Character String			
	nativeness	Nativeness Value			
	nameStatus	NameStatus Value		NAME STATUS	Short Integer
	sourceOf Name	Character String		SOURCE OF NAME	Short Integer
	pronunciation	Pronunciation OfName			
	grammatical Gender	Grammatical GenderValue		GRAMMATICAL GENDER	Short Integer
	grammatical Number	Grammatical NumberValue		GRAMMATICAL NUMBER	Short Integer
Spelling OfName	text	Character String	geograficky_nazov	NAM	
	script	Character String			
	transliteration Scheme	Character String			
Pronunciation OfName	pronunciation SoundLink	URI			
	pronunciation IPA	Character String			

Tab. 3 Ukážka párovania číselníka „geograficky_nazov_CAT“ na číselník „NamedPlaceTypeValue“

Geograficky_nazov_CAT	NamedPlaceTypeValue
aleja	landcover
autokemping	other
cesta	transportNetwork
gejzír	hydrography

Následne sme mohli pristúpiť k napĺňaniu jednotlivých subelementov a atribútov hodnotami zo zdrojových údajov tak, ako je to znázornené v tab. 2. Ak neboli požadované hodnoty obsiahnuté v zdrojových údajoch v potrebnej podobe, modifikovali sme ich pomocou vstavaných transformačných funkcií v rámci FME. Príkladom môže byť transformácia geometrickej zložky údajov do geografického súradnicového systému založeného na referenčnom systéme ETRS89 (EPSG:4258), ďalej reklasifikácia číselníkov 1 : 1, zmena formátu dátumu (z údajového typu ESRI DateTime na ISO formát požadovaný Implementačnými pravidlami – teda YYYY-MM-DD“T“HH:MM:SS) apod.

V prípadoch, keď malo dôjsť k naplneniu elementu, ktorý mal v konceptuálnom modeli *stereotyp* <<*voidable*>>, resp. v konceptuálnej schéme definíciu atribútu *nillable*=“*true*“ a zároveň v zdrojových údajov neexistovala vhodná hodnota, resp. ju nebolo možné odvodiť, bol daný element nechaný prázdny, hodnota jeho atribútu *xsi:nil* bola nastavená na „*true*“ a ako dôvod nevyplnenia sa uviedla hodnota „*nezaznamenávaný*“ – „*Unpopulated*“.

Špeciálnym prípadom sú atribúty s kardinalitou vyššou ako 1, čo v prípade súborov GML znamená, že v rámci jednej inštancie sa určitý sub-element bude opakovať viackrát. Pre oblasť geografického názvoslovia je napríklad bežné, že názov konkrétnej lokality sa z historických dôvodov menil, no poloha lokality zostala nezmenená. V prípade inštancie triedy GML NamedPlace sa táto skutočnosť prejaví tak, že koreňový element <*NamedPlace*> bude mať viac sub-elementov <*GeographicalName*>. Ukážka výrazov *XPath*, s využitím funkcie „*fme:get-attribute*“ pre napĺňanie prázdnej inštancie GML, je znázornená nižšie:

```
   ...
   <gn:geometry>
   <gml:Point gml:id="Point_{fme:get-attribute("ZBGIS_ID")}"
srsName="urn:ogc:def:crs:EPSG::4258" srsDimension="2">
     <gml:pos>{fme:get-attribute("_y")} {fme:get-attribute("_x")}</gml:pos>
   </gml:Point>
   </gn:geometry>
   <gn:inspireId>
     <base:Identifier>
       <base:localId>{fme:get-attribute("ZBGIS_ID")}</base:localId>
       <base:namespace>SK.ZBGIS.GN</base:namespace>
       <base:versionId>{fme:get-attribute("_timestamp")}</base:versionId>
   ...
```

Vo výslednom súbore GML sa potom nachádzalo toľko naplnených inštancií GML <*NamedPlace*>, koľko bodov so zemepisnými názvami obsahovala vstupná geodatabáza ZBGIS®. Okrem samotných inštancií GML bolo treba zahrnúť do hlavičky súboru GML aj súradnice priestorového rozsahu (*bounding box*), automaticky odvodené vstavanou funkciou FME *„BoundingBoxAccumulator“*, pre ktoré je vyhradený subelement <*gml:Envelope*>. Pre overenie správnosti vnútornej štruktúry súboru GML sme v hlavičke doplnili aj platné odkazy na samotnú aplikačnú schému INSPIRE Geographical Names a schému pre štandard GML vo verzii 3.2.1 (prevzaté z webovej stránky INSPIRE: *http://schemas.opengis.net/gml/3.2.1/gml.xsd*).

Záver

Vzhľadom na skutočnosť, že jazyk GML je rozšírenou verziou jazyka XML (zameranou na modelovanie priestorových objektov), dá sa syntaktická a schematická správnosť dokumentov GML overiť validáciou. V našom prípade sme uskutočnili validáciu v režime offline s využitím desktopového nástroja Altova XML Spy a v režime online s využitím XSD validátora na webovej adrese: *http://www.freeformatter.com/xml-validator-xsd.html*. V oboch prípadoch bola validita potvrdená.

Výsledný súbor GML je k dispozícii na webovej adrese: *http://158.195.40.237/Geographical Names/GN_Brat_kraj.xml*.

Použitý spôsob transformovania údajov do štruktúry INSPIRE sa dá použiť pri poskytovaní priestorových údajov INSPIRE vo forme tzv. preddefinovaných „datasetov". To znamená, že poskytovateľ údajov ich môže v takejto forme sprístupniť priamo na prevzatie, príp. sa dajú súbory GML zhromaždiť v databáze s podporou údajového typu BLOB (Binary Large Object) a následne sprístupniť údaje vo forme INSPIRE Ukladacej služby s priamym prístupom.

V kapitole 3.2 uvádzame problém s kardinalitou vyššieho stupňa, keď sa viacero podradených záznamov (relačne prepojených na nadradený objekt vo väzbe 1:N) transformuje na viacnásobne opakovaný subelement nejakej inštancie GML. V našom experimente sa nám nepodarilo tento aspekt transformácie úplne vyriešiť, pri migrácii údajov bol prenesený vždy iba prvý asociovaný záznam z väzby 1 : N. Riešením by mohli byť sofistikovanejšie konštrukcie XPath. Napriek tomu sa metóda transformácie s využitím princípov priestorových nástrojov ETL javí ako vhodná, čo dokazujú aj výsledky validácie. Samozrejme existujú aj iné možnosti transformácie údajov (modelom riadená architektúra, ontologické mapovanie, ...), ale preferencia priestorových ETL je podmienená ich aspektom flexibility a opätovnej použiteľnosti. Aj preto je možné prepojenie ETL z tohto experimentu použiť ďalej, ako vzor pre transformáciu údajov iných tém priestorových údajov asociovaných s iniciatívou INSPIRE.

Príspevok vznikol s podporou projektu APVV-0326-11.

Literatúra

GHERABI, N., BAHAJ, M. (2011). Robust Representation for Conversion UML Class into XML Document using DOM. *International Journal of Computer Applications (0975-8887)*, 33, 9, November 2011, pp. 22-28. [online] [cit. 2013-12-06]. Dostupné na:
<http://arxiv.org/ftp/arxiv/papers/1205/1205.5921.pdf>.

INSPIRE DRAFTING TEAM DATA SPECIFICATION (2013). *INSPIRE Generic Conceptual Model ver. 3.4rc3* [online]. [cit. 2013-12-06]. Dostupné na:
<http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/D2.5_v3.4rc3.pdf>.

INSPIRE TWG GN (2010). INSPIRE THEMATIC WORKING GROUP GEOGRAPHICAL NAMES. *INSPIRE Data Specification on Geographical Names – Guidelines* [online]. [cit. 2013-12-06]. Dostupné na: <http://inspire.jrc.ec.europa.eu/documents/Data_Specifications/INSPIRE_DataSpecification_GN_v3.0.1.pdf>.

KONEČNÝ, M., KUBÍČEK, P. (2012). *Datové infrastruktury pro prostorově informační společnost*. Brno (Masarykova Univerzita), 97 s.

LU, C. T., DOS SANTOS R.F, SRIPADA, L.N., KOU, Y. (2007). Advances in GML for Geospatial Applications. *Geoinformatica*, pp. 131-157.

ROUTLEDGE, N., GOODCHILD, A., BIRD, L. (2002). UML and XML Schema. In Xiaofang Zhou (ed.), *Conferences in Research and Practice in Information Technology*. Vol. 5, Melbourne, Australia, pp. 157-166. [online] [cit. 2013-12-06]. Dostupné na:
<http://crpit.com/confpapers/CRPITV5Routledge.pdf>.

SAFE SOFTWARE (2012). *Harmonise Your Spatial Data for INSPIRE with FME*. [online]. [cit. 2013-12-06]. Dostupné na:
<http://www.safe.com/videos/?video=http://cdn.safe.com/videos/Webinar-Harmonise-Your-Spatial-Data-for-INSPIRE-with-FME.mp4&category=webinar>.

SAGRIS, V., WOJDA, P., MILENOV, P., DEVOS, W. (2013). The harmonized data model for assessing Land Parcel Identification Systems compliance with requirements of direct aid and agri-environmental schemes of the CAP. *Journal of Environmental Management*, 118, pp. 40-48.

ÚGKK SR (2013). Úrad geodézie, kartografie a katastra Slovenskej republiky *Katalóg tried objektov ZBGIS.* [online] [cit. 2013-12-06]. Dostupné na: <http://www.skgeodesy.sk/files/slovensky/ugkk/geodezia-kartografia/zb-gis/kto_zbgis_2013_4.pdf>.

S u m m a r y

Tools for the harmonization of geographical names

In order to make EU member states' national spatial datasets INSPIRE compliant, certain data processing steps have to be carried out. This is generally referred to as "data harmonization" and involves changes to structure, format and/or spatial reference system of spatial data. In this case, we present one of the applicable practices to harmonize national Geographical names dataset (from Primary Base of GIS) with the INSPIRE Implementation Rules concerning the INSPIRE Geographical Names Application Schema. Data harmonization in principle means to compare source schema against the destination schema and propose (pair) the matching feature types, their attributes and attribute domains accordingly (create the matching tables). We have looked up these aspects within the national Geographical names dataset (1 point feature class with 2 related tables, Tab. 1) and paired them with features and attributes defined in INSPIRE Geographical Names Data Specification. The resulting pairs are shown in Tab. 2 and Tab. 3. As the INSPIRE Implementation Rules require the harmonized data to be stored in OGC GML format (version 3.2.1), we had to derive an empty GML instance from an XSD schema. This schema is machine- and human readable at the same time, so it was possible to automatically generate the empty GML instance and check it for correctness (whether it contains all mandatory XML elements and XML attributes) afterwards. This procedure is described in section 2.2. Once the empty GML instance and the matching tables are ready, the data itself can be transformed, migrated to the final GML structure, which is similar to filling-in a preprinted form. The empty GML instance plays the role of the form and the actual values being filled-in are represented by the source data from the national dataset. The Extract-Transform-Load Tools (ETL tools in short) can be used to meet this objective. We have designed our own ETL tools, which uses XPath expressions to extract the Geographical names spellings, status information, historical Geographical names etc. from the source data and load to the target GML instance document. More about this could be found in section 3.2. However, this workflow is unable to handle one-to-many relationships in source data. This would result in multiple nested sub-elements in the final GML document. Possibly, some more-sophisticated XPath expressions could help to resolve this issue. Apart from this issue, the validation (check, whether the resulting document meets the requirements of the INSPIRE Implementation Rules) finished with correct results. In this way, the harmonized data is ready to be made publicly available (for example, download from a web server as a "INSPIRE pre-defined dataset"). Naturally, the workflow presented in this paper is not the only possible solution. One can decide for a different approach (model-driven architecture, ontology mapping…), but the preference of ETL tools lies in their flexibility and re-usability. From this point of view, this paper could serve as a guide for transforming spatial data related to other themes from INSPIRE Annexes.