DDI 3.0: Kvalitatívny posun v štandarde dokumentácie sociálnych dát

Juraj Švec, Slovenský archív sociálnych dát

V apríli 2008 bola zverejnená ostatná verzia štandardu dokumentácie sociálnych výskumov DDI 3.0[1]. Vychádza z predchádzajúcich verzií 1.* a 2.*, pričom sa od nich zásadne odlišuje v rozsahu možností využitia, čo umožňuje zmenená štruktúra dokumentácie. DDI (Data Documentation Initiative) je medzinárodný štandard zápisu dokumentácie sociálnych výskumov, ktorý využíva stále väčší počet inštitúcií, najmä sociálno-vedné archívy. Mnohé inštitúcie z USA, Kanady a Európy už nastúpili alebo plánujú nastúpiť cestu migrácie zo starších verzií na tretiu verziu. Informácie o štandarde a jeho starších verziách sú k dispozícii na oficiálnej webovej stránke; www.ddialliance.org, ako aj v časopise Data a výzkum – SDA Info, ktorý je prístupný na webstránke českého archívu SDA [Kalvas 2005; Vávra 2007].

Vzhľadom na to, že pri dokumentovaní výskumov v oblasti sociálnych vied je potrebné zozbierať a uchovať veľké množstvo informácií, a dokumentácia musí plniť viacero funkcií, DDI štandard sa už od uverejnenia prvej verzie vyvíja. Prvé dve verzie vychádzali z myšlienky elektronickej verzie klasického tlačeného codebooku, no to znamenalo, že zdedili aj mnohé nedostatky, ktoré vyplývali z obmedzení tlačeného textu, napríklad potreba tvorby celého nového dokumentu pre každý jednotlivý zber dát, či obmedzené možnosti zoskupovania výskumov. Tretia verzia DDI je veľkým krokom vpred nie len v oblasti využiteľnosti samotného štandardu. Zatiaľ čo sa tvorcovia snažili prispôsobiť špecifikáciu jej užívateľom, aj vďaka používaniu XML jazyka sa im darí zachovať určitú mieru kompatibility s predošlými verziami. To znamená, že je relatívne jednoduché migrovať aj s veľkými množstvami dokumentácie na novšiu verziu štandardu, bez čoho by bolo nezmyselné s takou frekvenciou vydávať nové verzie DDI špecifikácie (od uverejnenia prvej verzie uplynulo len osem rokov).

Tretia verzia DDI má nové vlastnosti, ktoré vznikli ako odpoveď na potreby tvorcov špecifikácie (ktorí väčšinou využívajú DDI aj v praxi), ako aj ďalších užívateľov z praxe, na vývoj v dokumentovaní a archivovaní dát, a na pokroky dosiahnuté v XML technológii. Tieto požiadavky si vyžiadali zásadnú úpravu štruktúry štandardu a niekoľko menších zmien. Medzi hlavné požiadavky kladené na novú verziu štandardu patria:[2]

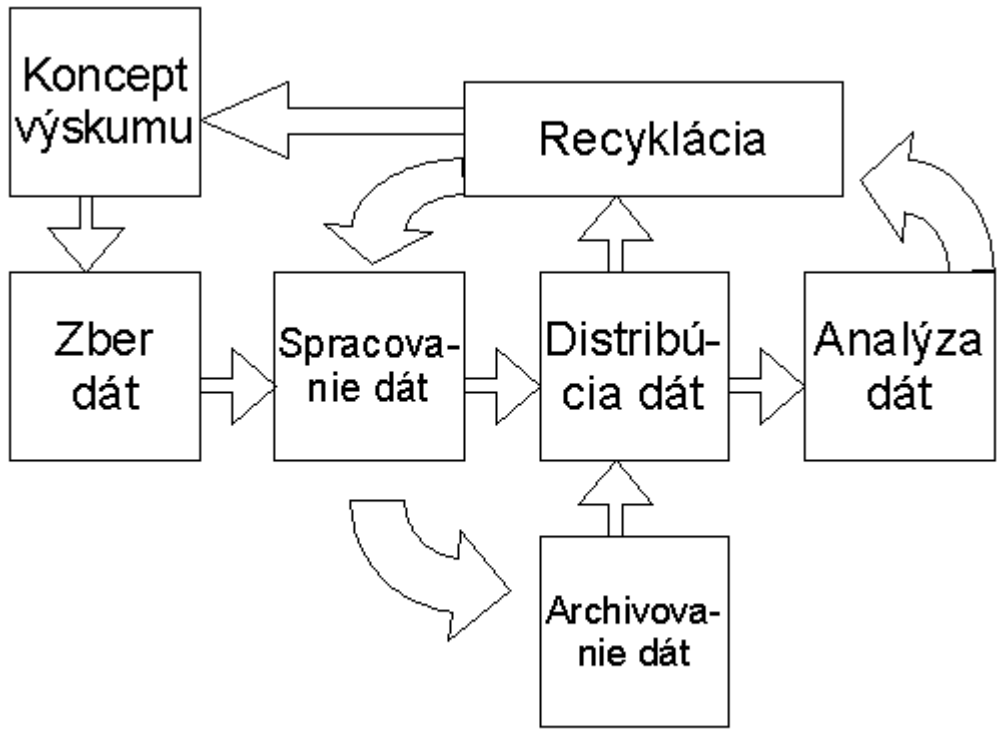
- zlepšiť a rozšíriť aspekty strojovej spracovateľnosti (machine-actionability) DDI na podporu programovania a softvérových systémov
- podporiť CAI nástroje pomocou rozšíreného popisu dotazníka (týka sa to obsahu a sledu/toku otázok)

1 O potrebnosti a užitočnosti štandardu dokumentácie metadát pojednáva článok [Blank, Rasmussen 2004].
2 Podľa [Gregory, Thomas 2008]

- podporiť popis dátových sérií (longitudinálnych výskumov, panelových štúdií, opakovaných výskumov, apod.)
- podporiť porovnateľnosť, konkrétne pri dizajnovaní výskumu (nástrojov), ako aj porovnateľnosť ex-post (harmonizácia výsledkov)
- zlepšiť podporu popisu komplexných dátových súborov (záznamové a súborové prepojenia)
- zabezpečiť zlepšenú podporu geografického obsahu na umožnenie prepájania ku geografickým súborom.

Podstatnou charakteristikou DDI 3.0, z ktorej vyplýva jeho zmenená štruktúra je orientácia na životný cyklus dát (Schéma 1). Prvé dve verzie DDI boli orientované na codebook, z čoho vyplýva, že zaznamenávali dokumentáciu k výskumu v jedinom časovom bode, teda až po vykonaní všetkých prác na samotnom výskume. Orientácia na životný cyklus dát umožňuje dokumentovanie už od počiatočných fáz výskumu, čo prináša mnohé výhody výskumníkom – producentom dát, ako aj užívateľom dát a archivárom.

Schéma 1:



Upravené podľa [Ionescu, Vardigan 2008]

Jednotlivé časti DDI pokrývajú všetky fázy tvorby dát; predprodukčné, produkčné a postprodukčné. Predprodukčné fázy sú konceptualizácia výskumu a zber dát, tieto zahŕňajú napríklad údaje o výskumnom zámere, autoroch, použitých konceptoch, financovaní, skúmanej populácii, metodológii zberu dát či o výskumnom nástroji. Produkčná fáza zahŕňa informácie o

spracovaní dát; napríklad tvorba dátového súboru, čistenie, alebo váženie. Postprodukčné fázy sú distribúcia, archivovanie a analýza dát. Prvé dve dokumentujú okrem iného premenné, ich kategórie (a ich kódy), formát archivovaných dát, dostupnosť dát a štatistiky. Fáza analýzy dát z pohľadu DDI označuje dokončenie konkrétneho prípadu (Instance). Prípad sa vytvorí spojením už vytvorených častí DDI dokumentácie, čím nadobúda samostatnú existenciu ako objekt a reprezentuje konkrétny výskum, resp. zber dát (reálny alebo virtuálny, napríklad účelným zoskupením viacerých dátových súborov). Recyklačná fáza označuje sekundárne využitie dát, čo implikuje nové vytvorenie niektorých komponentov DDI špecifikácie. Tvorenie dokumentácie počas jednotlivých fáz tvorby dát umožňuje modulárna štruktúra špecifikácie.

Tretia verzia DDI je komplexnejšia a štruktúrovanejšia ako predošlé verzie, no porozumenie celej koncepcie je dôležité najmä pre tých, ktorí sa profesionálne zaoberajú dokumentáciou/archiváciou, teda pre distribútorov. Jednotlivé súčasti špecifikácie sa môžu a nemusia použiť, v závislosti od charakteristík dokumentovaného výskumu a konkrétnych potrieb užívateľov. Bežní poskytovatelia a užívatelia dokumentácie ako výskumníci či študenti sa zrejme dostanú do styku len so špecializovanými programami určenými na poskytovanie dokumentácie a na prácu s dátami a dokumentáciou. Tieto programy budú pracovať len so súčasťami DDI, ktoré sú relevantné pre daný konkrétny výskum a účel. Napríklad dokumentácia výskumu s dátovým súborom s mikrodátami nemusí obsahovať modul popisujúci agregované dáta alebo viacdimenzionálne tabuľky.

Orientácia na životný cyklus dát umožňuje zaznamenávanie metadát tvorených rôznymi autormi v rôznych fázach tvorby dát. To znamená že dokumentácia môže byť zaznamenávaná najkompetentnejším človekom, takpovediac „pri zdroji.“ Tým sa zníži pravdepodobnosť chýb a predíde sa stratám informácií. Napríklad výskumník spolupracujúci na výskume zaznamená údaje o konceptoch či financovaní, agentúra zodpovedná za terénny zber dát zaznamená informácie o návratnosti dotazníkov či kódovaní odpovedí, a distribútor dát zdokumentuje procesy archivácie a zverejnenia, apod. Ďalšou výhodou je umožnenie všetkým zainteresovaným dokumentovať ich prácu priamo v DDI. Z orientácie na životný cyklus dát a z požiadavky využitia metadát na rôzne účely vyplynula potreba modularity špecifikácie.

Modulárna štruktúra DDI 3.0 umožňuje použitie len tých častí špecifikácie - modulov, ktoré sú v danom momente potrebné, a následné pridávanie modulov počas životného cyklu. Jednotlivé moduly zodpovedajú fázam životného cyklu dát. Pre špeciálne typy dát a formátov boli vytvorené špecializované moduly. Možnosť kombinovania rôznych modulov umožňuje výstižný a hutný popis rôznych typov dát, napr. samostatného dátového súboru, skupiny súvisiacich dátových súborov či skupiny súvisiacich štúdií.

Ďalšou výhodou modulárnej štruktúry je podpora znovupoužívania jednotlivých častí

dokumentácie. V záujme úspory prostriedkov, umožnenia priestorovej a časovej komparácie, či šetrenia serverového priestoru je v DDI 3.0 do značnej miery umožnené znovupoužívanie rôznych modulov a schém špecifikácie. Napríklad pokiaľ sa už zdokumentovaná séria otázok, schéma variantov odpovedí, kódovacia schéma alebo koncept použije v ďalšom výskume, stačí ak sa na príslušnom mieste v dokumentácii uvedie odkaz na túto dokumentáciu. To je záležitosť zápisu v XML špecifikácii, naopak v užívateľskom rozhraní sa môže zobraziť konkrétna informácia (metadáta), aj s poznámkou, kde už bola rovnaká schéma, koncept, apod. použitá.

V DDI špecifikácii je možné štruktúrovať a hierarchizovať dokumentáciu aj popisované výskumy. Napríklad medzinárodný sériový výskum s nepravidelne sa opakujúcimi modulmi (batériami otázok), naviac s nezávisle realizovanými národnými zbermi dát je v 3.0 možné veľmi presne popísať s jednoznačne určenými vzťahmi medzi dátami integrovanými za všetky krajiny, národnými dátami a dátami z rôznych časových období. V doterajších verziách špecifikácie bolo toto štruktúrovanie možné len s ťažkosťami, navyše takéto adaptácie (rozšírenie DDI špecifikácie) boli robené spravidla nejednotne, a teda medzi jednotlivými archívmi nekompatibilne.

S tým súvisí aj možnosť zoskupovania výskumov podľa rôznych kritérií, napr. podľa geografického pokrytia, skúmanej populácie, témy výskumov, druhu použitého nástroja a mnoho iných. Záleží od tvorcu dokumentácie, či a aké kritérium zoskupenia si zvolí pri tvorení skupiny metadát.

Vzhľadom na vývoj prebiehajúci v oblasti XML štandardov a požiadavky strojovej spracovateľnosti DDI dokumentácie, tretia verzia DDI už nie je vyjadrená v DTD jazyku, ale v XML schemach[3]. Schemy sú reálne súbory (s príponou .xsd), ktoré zodpovedajú modulom reprezentujúcim fázy životného cyklu dát. Schemy vyžadujú vyššiu mieru štandardizácie ako DTD, teda je pravdepodobnejšie, že dokumentácia bude kompatibilná medzi inštitúciami, čo spolu s modularitou vytvára ďalšiu nespornú výhodu DDI 3.0, a síce možnosť úzkej spolupráce národných archívov sociálnych dát pri dokumentovaní medzinárodných výskumov, najmä ich základných častí. Napríklad v prípade ISSP stačí ak jeden archív zdokumentuje základný (master) dotazník, a ostatné archívy môžu túto dokumentáciu v prípade potreby a dostatočnej vôle autorov jednoducho využiť pri dokumentácii svojej národnej verzie. Týmto by sa ušetrilo množstvo času a prostriedkov.

Okrem vyššie uvedených možností je samozrejme možné v novej verzii DDI aj ex-post dokumentovanie výskumu analogicky k predošlým verziám, teda ako nezávislej samostatnej štúdie s jedným zberom dát.

Tento článok je len skromnou rámcovou orientáciou v novej verzii DDI, so zámerom oboznámiť odbornú verejnosť s niektorými významnými zmenami a výhodami tohto štandardu. Na implementáciu je potrebné bližšie oboznámenie sa s celou špecifikáciou. DDI 3.0 aj predošlé verzie

3 Schema (anglický výraz) je zámerne napísané bez dĺžňa, aby sa dali odlíšiť od schém (anglicky: scheme).

budú koexistovať a budú podporované zo strany DDI Alliance. Pokiaľ to nie je užitočné, nie je nutné transformovať dokumentáciu v starších verziách do 3.0. Verzie sú medzi sebou do určitej miery kompatibilné, takže je možná aj koexistencia dokumentácie v rôznych verziách v rámci jednej inštitúcie, len na to treba prispôsobiť softvér. Ostatná verzia je už takmer rok zverejnená, no nakoľko DDI je len špecifikácia, na jej využitie sú potrebné nástroje, softvér. Nástroje na prácu s touto verziou sú zatiaľ obmedzené. Nasťastie aj v tejto oblasti sa Aliancia snaží plniť koordinačnú funkciu a k dispozícii už je niekoľko nástrojov na dokumentáciu v 3.0, ako aj na konverziu zo starších verzií.[4] Nesstar zatiaľ nie je prispôsobený na túto verziu. Ukážky využitia nových charakteristík DDI 3.0 sú na adrese: http://www.ddialliance.org/DDI/ddi3/use-cases.html.

Autor sa v novembri 2008 zúčastnil workshopu DDI 3.0 organizovaného nemeckým GESIS – Zentral archiv, kde získal základné informácie. 26. mája na konferencii IASSIST vo fínskom Tampere sa bude konať ďalší workshop DDI, a potom v júli na Cornell University, NY, USA. Na oficiálnej stránke DDI Alliance http://www.ddialliance.org/ je k dispozícii viacero textov popisujúcich DDI, ako aj návody ako ho používať.

Literatúra:

Blank, Grant, Karsten B. Rasmussen. 2004. „The Data Documentation Initiative: The Value and Significance of a Worldwide Standard.“ *Social Science Computer Review* 22: 307-318.

DDI 3.0 Part I Overview. 2008. DDI Alliance: http://www.ddialliance.org/ .

Gregory, Arofan, Wendy Thomas. 2008. „Putting DDI 3.0 to work for you! Part II.“ Príspevok prednesený v rámci workshopu na *34. výročnej konferencii IASSIST*. Palo Alto, California, 27. 5. 2008.

Ionescu, Sanda, Mary Vardigan. 2008. „Putting DDI 3.0 to work for you! Part I.“ Príspevok prednesený v rámci workshopu na *34. výročnej konferencii IASSIST*. Palo Alto, California, 27. 5. 2008.

Kalvas, František. 2005. „NESSTAR a DDI pro uživatele datových služeb." *SDA Info* VII. (1): 14-16.

Vávra, Martin. 2007. „Archivace sociologických dat." *Data a výzkum - SDA Info* 1 (1): 7-18.

4 Viac o nástrojoch nájdete na http://tools.ddialliance.org/