

EKONOMICKÁ UNIVERZITA V BRATISLAVE
FAKULTA HOSPODÁRSKEJ INFORMATIKY

Evidenčné číslo: 103004/I/2018/3374129115

Štandard DDI v práci s dátami

Managing data by DDI standard

Diplomová práca

EKONOMICKÁ UNIVERZITA V BRATISLAVE
FAKULTA HOSPODÁRSKEJ INFORMATIKY

Štandard DDI v práci s dátami

Managing data by DDI standard

Diplomová práca

Študijný program: Informačný manažment

Študijný odbor: Kvantitatívne metódy v ekonómii

Školiace pracovisko: Katedra aplikovanej informatiky FHI

Vedúci záverečnej práce: doc. Dr. Ing. Miroslav Hudec

Čestné vyhlásenie

Čestne vyhlasujem, že záverečnú prácu som vypracoval samostatne a že som uviedol všetku použitú literatúru.

Dátum:

.....

(podpis študenta)

Pod'akovanie

Moje veľké pod'akovanie patrí vedúcemu diplomovej práce doc. Dr. Ing. Miroslavovi Hudecovi za pripomienky, odbornú pomoc, veľkú dávku trpezlivosti a strpenia, pochopenie a usmernenie pri vypracovaní záverečnej práce. Rovnako moje pod'akovanie patrí spoločnosti Colectica za udelenie evaluačnej licencie na softvér Colectica Designer.

ABSTRAKT

HORNÁČEK, Milan: *Štandard DDI v práci s dátami*. – Ekonomická univerzita v Bratislave. Fakulta hospodárskej informatiky; Katedra aplikovanej informatiky. – Vedúci záverečnej práce: doc. Dr. Ing. Miroslav Hudec. – Bratislava: FHI, 2018, 64s.

Cieľom práce je preskúmanie štandardu DDI v práci s dátami od zberu, cez spracovanie až po archiváciu a na ilustračných príkladoch ukázať jeho výhody a obmedzenia. Práca je rozdelená do 3 kapitol, obsahuje 17 obrázkov a 2 kódy. Prvá kapitola je teoretická časť práce, v ktorej sme definovali základné termíny a uviedli sme pojem DDI štandardu. Druhá kapitola popisuje cieľ práce, metodiku práce a metódy skúmania. V záverečnej kapitole sme popísali editory vhodné na prácu s DDI štandardom a na ilustračných príkladoch ukázali životný cyklus dát.

Kľúčové slová:

aliancia DDI, DDI, štandard DDI, XML, editačný softvér, Colectica Designer

ABSTRACT

HORNÁČEK, Milan: *Managing data by DDI standard*. – University of Economics in Bratislava. Faculty of Economic Informatics; Department of Applied Informatics. – Supervisor: doc. Dr. Ing. Miroslav Hudec. – Bratislava: FHI, 2018, 64p.

The aim of the thesis is to review the DDI standard in working with data from collection, through processing to archiving and using illustrative examples to show its benefits and restrictions. The thesis is divided into 3 chapters, contains 17 pictures and 2 codes. The first chapter is the theoretical part of the thesis, in which we defines the basic terms and we introduced the term DDI standard. The second chapter describes the aim of the work, the methodology of work and methods of investigation. In the final chapter we described editors suitable for working with DDI standard and using illustrative examples we showed the life cycle of data.

Key words:

DDI Alliance, DDI, DDI standard, XML, editing software, Colectica Designer

OBSAH

ÚVOD	9
1 Súčasný stav riešenej problematiky	10
1.1 Aliancia DDI.....	10
1.2 Súčasný pracovný produkt spoločnosti DDI Alliance	10
1.2.1 DDI-Codebook	11
1.2.2 DDI-Lifecycle.....	12
1.2.3 RDF slovníky	13
1.2.4 Kontrolované slovníky	14
1.2.5 DDI-Lifecycle MD (riadený modelom) ako budúcnosť DDI štandardu	14
1.3 Metadáta.....	17
1.4 XML.....	19
1.4.1 Definícia.....	19
1.4.2 Prvky a atribúty XML	20
1.5 Termíny a definície súvisiace s pochopením DDI.....	21
2 Cieľ práce, metodika práce a metódy skúmania	31
3 Výsledky práce a diskusia.....	32
3.1 Softvér používaný na prácu s XML	32
3.2 Codebook ako nevyhnutný dokument pre používateľa dát	37
3.3 Nástroje používané na prácu s DDI	38
3.4 Programové prostriedky spoločnosti Colectica	40
3.5 Práca s dátami pomocou štandardu DDI.....	44
3.5.1 Zber dát prostredníctvom programu Colectica Designer	46
3.5.2 Práca s dátami prostredníctvom programu Colectica Designer	55

3.5.3 Archivácia dát prostredníctvom programu Colectica Designer	61
Záver	62
Bibliografické odkazy	63

ÚVOD

V súčasnosti sa zbierajú, spracúvajú a archivujú veľké množstvá dát. Spôsobil to fenomén Big Data, ktorý je nazývaný aj informačnou lavínou. Denne sú generované obrovské objemy dát prostredníctvom internetu, smartfónov, kamerových systémov a zariadení Internetu vecí. Tieto zozbierané dáta môžu predstavovať pre ich vlastníkov obrovské možnosti na zisk. Manipulácia s dátami však nie je vždy efektívna a to spôsobuje veľké finančné straty.

Na trhu je aktuálne veľký počet štandardov na prácu s dátami, mnohé z nich fungujú na báze voľne dostupných štandardov. Medzi ne patrí aj štandard DDI (Data Documentation Initiative), ktorého popis činnosti sme si vybrali za predmet našej diplomovej práce, konkrétne jeho poslednú verziu 3.2. Štandard sa neustále vyvíja a v marci roku 2018 bola ohlásená verzia 4. Podporuje komplexný opis dát pre objavovanie a analýzu a poskytuje podporu pre efektívne zdieľanie dát. Tvorcovia vyvíjajú štandard pod heslom „vytvor jedenkrát, používaj často“ čo v preklade znamená, že metadáta vytvorené týmto štandardom sú vo formáte aplikovateľnom na viacero štúdií viacerými agentúrami.

V prvej kapitole našej práce sa venujeme súčasnému stavu riešenej problematiky. Opisujeme organizáciu aliancia DDI, ktorá stojí za vznikom štandardu DDI, súčasné pracovné produkty spoločnosti, definujeme pojem metadáta a značkovací jazyk XML, ktorý je podstatou štandardu DDI a v záverečnej podkapitole sa venujeme termínom nevyhnutným na pochopenie problematiky DDI štandardu.

Časti ciele práce, metodika práce a metódy skúmania sme spojili do jednej kapitoly. V tejto časti sme si stanovili ciele, ktoré sme chceli našou prácou dosiahnuť, metodiku práce a metódy skúmania, ktoré sme zvolili k dosiahnutiu našich cieľov.

V tretej kapitole sme zhrnuli výsledky našej práce. Časť výsledky práce sme spojili s časťou diskusia. V prvej časti sme otestovali a opísali softvér používaný na prácu s jazykom XML, nástroje na prácu s DDI a pomocou obdržanej evaluačnej licencie na softvér Colectica Designer sme opísali prácu so štandardom DDI od zberu, cez spracovanie až po archiváciu dát. V závere sme zhodnotili naplnenie stanovených cieľov.

1 Súčasný stav riešenej problematiky

V tejto kapitole sme uviedli definície základných termínov a oboznámili sme čitateľa s problematikou štandardu DDI. Popísali sme Alianciu DDI, ktorá stojí za vznikom tohto štandardu a stručne sme opísali všetky produkty tejto organizácie.

1.1 Aliancia DDI

Hlavným cieľom Aliancie DDI je vytvoriť medzinárodný štandard pre opis dát zo sociálnych, behaviorálnych a ekonomických vied. Od svojho založenia v roku 1995 sa vývoj pracovných produktov DDI Alliance pohyboval v súlade s vyvíjajúcimi sa potrebami užívateľských komúní, ktoré sa snažia zdokumentovať údaje a procesy okolo svojej zbierky.

Pracovné produkty sa tiež vyvíjali spoločne s vývojom v oblasti technológií a výpočtovej infraštruktúry, ktorá to podporuje. XML, v ktorom sú vyjadrené aktuálne pracovné produkty, sa sám zmenil a rozvinul. Normy, ku ktorým sú zosúladené pracovné produkty, ako napríklad Dublin Core, ISO / IEC 11179 a naposledy GSIM, sa taktiež vyvinuli a technológie, ktoré sú k dispozícii na realizáciu pracovných produktov, dozreli.

Takéto široké ciele si vyžadujú úroveň zložitosti, ale so základnou logikou toho, čo bolo vyrobené a ako produkty súvisia. Tento dokument objasní účel každého pracovného produktu a pokúsi sa vysvetliť, prečo sa niekedy prekrýva, ale zriedka sa zdvojuje.

Tento prehľad pomôže používateľom určiť, ktoré pracovné produkty majú implementovať, a vedieť, čo očakávať s ohľadom na budúci vývoj.¹

1.2 Súčasné pracovné produkty spoločnosti DDI Alliance

Existuje niekoľko pracovných produktov DDI Alliance, z ktorých každý existuje pre iný účel. Každý je diskutovaný samostatne. Treba poznamenať, že ako organizácia pre normalizáciu, DDI Alliance nepožaduje, aby používatelia realizovali určitý pracovný produkt alebo konkrétnu verziu pracovného produktu. Ponúka každý pracovný produkt na zamýšľaný účel spôsobom, ktorý je užitočný pre niektorých členov komunity užívateľov DDI. Tieto pracovné produkty sú jednoducho nástroje, ktoré sa majú použiť - nie je potrebná žiadna

¹ [cit. 2018-01-28]

hodnota pri inovácii na novšiu verziu akéhokoľvek pracovného produktu alebo špecifikácie DDI, pokiaľ to nie je prínosom pre implementátora (odlišného od softvéru, kde je často aktualizácia na najnovšiu verziu odporúčaná).

Z tohto dôvodu budú všetky existujúce pracovné produkty DDI zachované v priebehu času, aby podporili používateľov, ktorí implementovali daný pracovný produkt. Súčasné pracovné produkty zahŕňajú:

- **DDI-Codebook** - DDI-Codebook je štruktúra XML na popis kódových kníh (alebo dátových slovníkov) pre jednu štúdiu.
- **DDI-Lifecycle** - DDI-Lifecycle (tiež vyjadrený v XML) rozširuje pokrytie jednej štúdie pozdĺž životného cyklu dát a môže opísať niekoľko vlín zhromažďovania údajov a dokonca aj ad hoc zbierky súborov údajov zoskupených na účely porovnania. Je veľmi užitočný pri riešení zberu sériových údajov, čo sa často prejavuje pri tvorbe údajov v štatistických úradoch a dlhodobých výskumných projektoch.
- **Názvy RDF slovníkov** - Názvy RDF pre sémantický web v súčasnosti pokrývajú tri oblasti: jeden pre popis štatistických klasifikácií, jeden na opis dátových súborov na účely zisťovania na webe a jeden na popis štruktúr jednotlivých súborov údajov. Tieto slovníky sú založené na štruktúre DDI-Codebook a DDI-Lifecycle XML, ale nie sú totožné s ani jednou. Sú navrhnuté na použitie v kombinácii a ako také nie sú duplicitné.
- **Kontrolované slovníky** - DDI riadené slovníky sú odporúčané súbory pojmov a definícií. Používajú sa na popísanie rôznych typov činností alebo artefaktov, ktoré existujú v štruktúre DDI-kódexu, DDI-Lifecycle a DDI-RDF Discovery.

1.2.1 DDI-Codebook

Špecifikácia kódu DDI je pôvodným pracovným produktom Aliancie a prešiel niekoľkými zmenami verzie. Najnovšia verzia je verzia 2.5.1. Kódex DDI nebol vždy označovaný ako taký - pri vytváraní DDI-Lifecycle bola pridaná časť "codebook" titulu, ktorá rozlišuje medzi týmito dvoma. Kódex DDI je štruktúra XML na popis kódových kníh (alebo dátových slovníkov) pre jednu štúdiu (štúdia v DDI parlance je jediná vlna zberu údajov). Kódex

DDI je navrhnutý tak, aby sa používal po vykonaní: predpokladá sa, že súbor údajov existuje ako výsledok kolekcie a potom sa opíše pomocou štruktúry XML poskytnutej špecifikáciou.

Kódex DDI je silne používaný, pretože je štruktúrou za nástrojmi ako Nesstar, ktorý používajú archívy údajov v rámci CESSDA a kanadské akademické inštitúcie. Najväčšou jednotnou používateľskou základňou je pravdepodobne Medzinárodná sieť pre prieskum domácností (IHSN), ktorá poskytuje nástroje na dokumentovanie štúdií vypracovaných štatistickými agentúrami v rozvojovom svete, z ktorých niektoré sú založené na nástrojoch Nesstar.

Kód DDI je - a bude - udržiavaný v rámci obmedzeného rozsahu jeho špecifikovaného pokrytia. Akékoľvek chyby budú opravené v budúcich vydaniach a môžu sa vykonať drobné zmeny, ktoré umožnia zaradiť DDI-Codebook do iných pracovných produktov. Napríklad, keď bol prepustený DDI-Codebook 2.5, systém identifikácie bol aktualizovaný tak, aby umožnil zahrnutie identifikátorov DDI-Lifecycle tak, aby metadáta, ktoré sa prekrývali medzi týmito dvomi pracovnými produktmi, mohli byť vyjadrené pomocou špecifikácie kódov alebo životného cyklu. Navyše bol pridaný obsah na podporu dodatočného pokrytia popisných objektov GSIM, ktoré boli pridané aj do DDI-Lifecycle.

1.2.2 DDI-Lifecycle

DDI-Lifecycle je výsledkom náročnejšieho súboru požiadaviek vyplývajúcich z používania knižnice DDI. Je to znova komplexná štruktúra metadát vyjadrená pomocou XML, ale je určená na rôzne účely. Životný cyklus je schopný opísať nielen výsledky zberu údajov, ale opisovať metadáta v celom procese zhromažďovania údajov od počiatočnej koncepcie až po archiváciu výsledných údajov. Môže opísať niekoľko verzií údajov a metaúdajov, ktoré sa menia v priebehu životného cyklu údajov, a teda názov "DDI-Lifecycle".

DDI-Lifecycle môže opísať niekoľko vlín zhromažďovania údajov a dokonca ad hoc zbierky súborov údajov zoskupených na účely porovnania. Je veľmi užitočné pri riešení zberu sériových údajov, čo sa často prejavuje pri tvorbe údajov v štatistických úradoch a dlhodobých výskumných projektoch.

DDI-Lifecycle bol prvýkrát vydaný ako DDI verzia 3.0 a teraz je vo verzii 3.2. Na účely údržby (opravy chýb a zosúladenia) budú existovať ďalšie vydania DDI-Lifecycle a bude existovať jednorazová funkčná expanzia zahŕňajúca prácu, ktorá predchádza nadchádzajúcemu DDI 4, pokrývajúcemu metodológiu prieskumu. Toto bude zahrnuté do verzie DDI 3.3 beta (v súčasnosti sa vyvíja).

Treba poznamenať, že kvôli účelu DDI-Lifecycle (ktorý zahŕňa opis dát po jednotlivých štúdiách ako nevyhnutnú súčasť jeho širšieho rozsahu), všetky metaúdaje, ktoré môžu byť vyjadrené pomocou DDI-Codebooku, sa môžu tiež opísať pomocou DDI-Lifecycle. DDI-Lifecycle je - z dôvodu účelu, ktorý bol navrhnutý taktiež - opäť zložitejší ako DDI-Codebook a vyžaduje komplexnejšiu výpočtovú infraštruktúru.

DDI-Lifecycle môže byť tiež použitý na dokumentovanie špecifických súborov metaúdajov mimo popis jednej štúdie alebo súboru štúdií. Napríklad oblasti spoločne zdieľaných metaúdajov, ako sú koncepty, štatistická klasifikácia alebo geografické štruktúry, môžu byť popísané a odkazované na ľubovoľný počet štúdií. Činnosti spracovania možno opísať a používať na podporu metadátového prístupu k zhromažďovaniu, spracovaniu a uverejňovaniu údajov.

1.2.3 RDF slovníky

RDF je Resource Description Framework, súbor webových technológií na opis rôznych informácií spôsobom, ktorý je možné spustiť strojom. Ako taký poskytuje alternatívu (pre sémantický web a prepojené dáta) s tradičnými štruktúrami XML špecifikovanými DDI Alliance - DDI-Lifecycle a DDI-Codebook. Na používanie RDF sa pre konkrétnu doménu vyvíja "slovník" (alebo ontológia) s rôznymi typmi informačných objektov a ich vzťahmi a vlastnosťami. DDI Alliance publikovala tri takéto slovníky: jeden pre popis štatistických klasifikácií (XKOS), jeden na opis dátových súborov na účely zisťovania na webe (DDI-RDF Discovery) a jeden na opis štruktúr jednotlivých súborov údajov (PHDD).

Tieto slovníky sú založené na štruktúre DDI-Codebook a DDI-Lifecycle XML, ale nie sú totožné s ani jednou. Sú navrhnuté na použitie v kombinácii a ako také nie sú duplicitné. Používajú aj iné slovné zásoby RDF vyvinuté na iných miestach, kde existujú a sú vhodné - považuje sa to za najlepšiu prax v komunite RDF na podporu interoperability.

RDF je určená ako súprava štandardov na publikovanie vecí na webe. Jeho implementácia sa rozširuje mimo komunitu, ktorá používala špecifikácie DDI XML. Tradične sa DDI XML používa za zatvorenými dverami na podporu uchovávaní, výmeny a správy dát (hoci šírenie je určite jedným aspektom správy dát). Takže hoci systémy, ktoré sú dnes založené na DDI XML, môžu byť čiastočne implementované pomocou slovníkov a technológií RDF, špecifikácie RDF Alliance sú určené na poskytovanie možností na úrovni technologickej platformy a nie na duplikovanie špecifikácií XML.

V budúcnosti pre model DDI-Lifecycle budú existovať ekvivalentné reprezentácie RDF a XML pre všetky obsahy metadát DDI. Napriek tomu sa ako základ pre niektoré časti DDI-Lifecycle (MD) použijú špecifikácie RDF slovníkov a budú sa v priebehu času udržiavať v závislosti od potreby opravy chýb.

1.2.4 Kontrolované slovníky

Kontrolované slovníky produkované DDI Alliance sú množiny výrazov a definícií používaných na opis rôznych typov činností alebo artefaktov, ktoré existujú v štruktúre DDI-Codebook, DDI-Lifecycle a DDI-RDF Discovery. Môžu byť použité buď s špecifikáciou XML (alebo s akýmkoľvek systémom, ktorý potrebuje takéto terminológie). Sú vyvinuté a udržiavané nezávisle od akýchkoľvek iných pracovných produktov DDI a sú skôr podporným pracovným produktom, než tým, ktorý má byť použitý samostatne. Kontrolované slovníky sú vyjadrené v Genericcode (XML) a v SKOS (vo vývoji).

Kontrolované slovníky DDI sú odporúčané súbory pojmov a definícií - organizácie môžu slobodne používať modifikované alebo úplne iné slovníky na svojom mieste, ak to lepšie vyhovuje potrebám ich organizácií.

1.2.5 DDI-Lifecycle MD (riadený modelom) ako budúcnosť DDI štandardu

Rovnako ako skúsenosti s kódom DDI vedú k širšiemu súboru požiadaviek, ktoré DDI-Lifecycle splnil, implementácie slovníkov DDI-Lifecycle a DDI RDF poskytlí nové požiadavky, ktoré splní DDI-Lifecycle MD, teraz v priebehu rozvoja. DDI-Lifecycle MD sa bude veľmi líšiť od iných pracovných produktov DDI, pretože bude založený na informačnom

modeli obsahu metaúdajov. Tento informačný model bude dostupný pre implementáciu v štandardných slovných zásobách RDF a štandardných štruktúrach XML, ktoré budú ekvivalentné. Táto forma modelového štandardu je najlepšou praxou vo svete štandardizácie a niekoľko ďalších štandardov v iných oblastiach používa tento prístup už mnoho rokov.

Vzhľadom na to, že DDI Alliance predpokladá, že môže byť veľa rôznych implementačných technológií využívajúcich DDI ako základ, mať explicitný informačný model vyjadrený pomocou Unified Modeling Language (UML) je prospešné. Príkladom je štandardná výmena štatistických údajov a metaúdajov (SDMX) založená na modeloch a má štandardné implementácie v XML, JSON a EDIFACT (syntax plochých súborov). Navyše, tento model bol použitý ako základ pre slovníka RDF vydané W3C (doma samotných základňa RDF štandardov).

Presun DDI do modelovej orientácie je teda spôsob ochrany štandardov pred technologickými zmenami a spôsob, ako zaručiť zosúladenie medzi rôznymi implementačnými technológiami. Vývoj a riadenie samotného modelu sa realizovalo spôsobom, ktorý umožňuje programové generovanie odvodených formátov, ako sú reprezentácie špecifikácií v XML Schema a OWL / RDF, dokumentačné a programové knižnice. Toto predstavuje model DDI-Lifecycle.

Ako dôležitý posun v spôsobe, akým je produkt DDI vyvinutý, úsilie DDI-Lifecycle MD sa rozvinulo niekoľko rokov v sérii sprintov a online stretnutí. Ako bolo poznamenané, úsilie zahŕňa vývoj nielen informačného modelu, ale aj infraštruktúru na budovanie tohto modelu, transformáciu na skupinu reprezentácií (pôvodne schému XML a RDF / OWL) a súvisiacu dokumentáciu. Modul DDI-Lifecycle MD sa rozvinie v sérii "uvoľnení". Počiatočné verzie budú testy výrobnej infraštruktúry rovnako ako aj príležitosť pre komunity DDI vyjadriť sa k smeru informačného modelu. Okrem revízií, ktoré sú založené na spätnej väzbe zo strany komunity, bude každé z prvých verzií pridávať aj nový obsah.

DDI-Lifecycle MD bude ďalším typom pracovného produktu DDI: ako DDI-Codebook a DDI-Lifecycle, bude to štandardná štruktúra pre metadáta súvisiace so životným cyklom dát vo všetkých fázach a tak bude zahŕňať veľa z toho, čo je k dispozícii v štruktúrach XML dnes. Bude tiež obsahovať funkčnosť slovníkov DDI RDF, pretože sú založené na štruktúrach DDI-

Codebook a DDI-Lifecycle. Podobne ako DDI-Codebook a DDI-Lifecycle, bude kompatibilný s produktmi DDI Controlled Vocabulary.

Bude však existovať rozšírenie pokrytia z hľadiska životného cyklu údajov (v porovnaní s DDI-Lifecycle 3.*). Napríklad DDI-Lifecycle MD podporuje okrem zberov iných typov zhromažďovania údajov, ako sú zhromažďovanie zdravotných informácií (vrátane údajov z elektronického zdravotného záznamu), údajov o strojoch (protokoly) a iných foriem udalostí a administratívnych údajov. Podpora pre iné zdroje údajov si vyžiadala opätovné preskúmanie objektov spracovania údajov DDI a vývoj modelu slobodného procesu, ktorý nie je zahrnutý do zberu údajov z prieskumu. Tento model nám umožňuje predstaviť nové typy dátových cyklov vrátane tých, ktoré uľahčujú budovanie a údržbu registrov. Registre neberú údaje vo vlnách a kolách. Namiesto toho sa môžu podieľať na nových logických a fyzických produktoch, ktoré často podporujú NoSQL a iné typy veľkých dátových skladov, ktoré sú schopné rozdeliť a zhromažďovať údaje v širokom rozsahu formátov, keďže sa registre vyvíjajú v priebehu času.

Rozšírenie pokrytia neznamena rozširovanie nových študijných objektov. Namiesto toho modelový informačný model často uľahčuje opätovné použitie, špecializáciu a systematictější prístup k štúdiu vývoja a organizácie objektov.

Rovnako, zatiaľ čo to môže znieť ako šírenie rôznych pracovných produktov DDI, v skutočnosti to nie je: po vytvorení informačného modelu sa špecifikácie XML a RDF automaticky vygenerujú z modelu. Tak, hoci tam bude zvýšený počet pracovných produktov ponúkaných Aliancie užívateľským komunitám, budú v skutočnosti vyžadovať menej úsilia na rozvoj, a budú mať garantované vyrovnanie, pokiaľ ide o ich štruktúru metadát.

To podporuje používanie podmnožín celého modelu na rôzne účely. Tieto podmnožiny sa nazývajú funkčné zobrazenia. Sú realizované v súvisiacich balíkoch v modeli a v špecifikáciách XML a RDF. Tento prístup umožňuje používanie DDI pre rôzne publikum, skryje zložitosť celého modelu a podporuje interoperabilitu.

Po preskúmaní inej dokumentácie zistíte, že tieto funkčné zobrazenia sú zosúladené s konkrétnymi prípadmi použitia. Na záver je cieľom, že DDI-Lifecycle založený na modeloch môže používateľom prezentovať DDI, takže používatelia sa môžu učiť a vlastniť iba tie časti DDI, ktoré potrebujú. Samozrejme, z dôvodu ich prípadov použitia sa niektorí používatelia

nevyhnutne stanú odborníkmi. Predpokladá sa však, že väčšina používateľov sa bude musieť naučiť len určité náhľady.²

1.3 Metadáta

Metadáta sú údaje, ktoré popisujú iné údaje. Meta je predpona, ktorá vo väčšine použitia informačných technológií znamená "základnú definíciu alebo opis".

Metadáta sumarizujú základné informácie o údajoch, čo uľahčuje hľadanie a prácu s konkrétnymi údajmi. Napríklad *autor*, *dátum vytvorenia* a *dátumy upravené* a *veľkosť súboru* sú príklady veľmi základných metaúdajov dokumentov. Mať schopnosť filtrovať cez tieto metaúdaje uľahčuje niekomu nájsť konkrétny dokument.

Okrem súborov s dokumentmi sa metadáta používajú pre obrázky, videá, tabuľky a webové stránky. Použitie metadát na webových stránkach môže byť veľmi dôležité. Metadáta pre webové stránky obsahujú popisy obsahu stránky, ako aj kľúčové slová spojené s obsahom. Tieto sú zvyčajne vyjadrené vo forme metatagov. Metadáta obsahujúce popis a zhrnutie webovej stránky sa často zobrazujú vo výsledkoch vyhľadávania vyhľadávacími nástrojmi a ich presnosť a podrobnosti sú veľmi dôležité, pretože určujú, či sa používateľ rozhodne navštíviť stránku alebo nie. Metatagy často hodnotia vyhľadávacie nástroje, aby pomohli rozhodnúť o relevantnosti webovej stránky a boli použité ako kľúčový faktor pri určovaní pozície vo vyhľadávaní až do konca 90. rokov. Zvýšenie optimalizácie pre vyhľadávače (SEO) ku koncu 90. rokov viedlo k tomu, že mnohé webové stránky "kľúčové slovo vypchávajú" svojimi metadátami, aby oklamali vyhľadávače, takže ich webové stránky sa zdajú byť relevantnejšie než iné. Odvtedy vyhľadávacie stroje znížili svoju závislosť na metatagoch, aj keď sú stále zohľadnené pri indexovaní stránok. Mnohé vyhľadávacie nástroje sa tiež pokúšajú zastaviť schopnosť webových stránok zamedziť svoj systém tým, že pravidelne menia svoje kritériá pre hodnotenie, pričom spoločnosť Google je známa tým, že často mení svoje vysoko utajené poradie algoritmov.

Metadáta je možné vytvoriť manuálne alebo automatickým spracovaním informácií. Manuálna tvorba má tendenciu byť presnejšia a umožňuje používateľovi vkladať akékoľvek informácie, ktoré považuje za relevantné alebo potrebné na opis súboru. Automatické

² [cit. 2018-01-28]

vytvorenie metadát môže byť oveľa jednoduchšie, zvyčajne zobrazuje len informácie o veľkosti súboru, rozšírení súboru, kedy bol súbor vytvorený a kto súbor vytvoril.³

Existujú tri rôzne typy metadát: opisné metadáta, štrukturálne metadáta a administratívne metadáta.

- Popisné metaúdaje opisujú zdroj na účely, ako je objav a identifikácia. Môže obsahovať prvky ako titul, abstrakt, autor a kľúčové slová.
- Štrukturálne metadáta sú metadáta o kontajneroch s údajmi a naznačujú, akým spôsobom sú objekty zlúčené, napríklad spôsob, akým sú strany zoradené na vytvorenie kapitol. Popisuje typy, verzie, vzťahy a iné charakteristiky digitálnych materiálov.
- Administrátorské metadáta poskytujú informácie, ktoré pomôžu spravovať zdroj, napríklad kedy a ako boli vytvorené, typ súboru a iné technické informácie a kto má k nemu prístup.

Iné zdroje používajú iný prístup k typom metaúdajov v závislosti od rôznych širokých účelov, na ktoré sa používajú:

- Deskriptívne vlastnosti metadát zahŕňajú titul, predmet, žáner, autora a dátum vytvorenia.
- Medzi metadáta práv môžu patriť stav autorských práv, držiteľa práv alebo licenčné podmienky.
- Vlastnosti technických metadát zahŕňajú typy súborov, veľkosť, dátum a čas vytvorenia a typ kompresie. Technické metadáta sa často používajú na riadenie a interoperabilitu digitálnych objektov.
- Metadáta uchovávaní sa používajú v navigácii. Príklady vlastností metadát uchovávaní zahŕňajú umiestnenie položky v hierarchii alebo sekvencii.
- Označovanie jazykov zahŕňa metadáta použité pre navigáciu a interoperabilitu. Vlastnosti môžu obsahovať nadpis, názov, dátum, zoznam a odsek.⁴

³ [cit. 2018-01-28]

⁴ [cit. 2018-01-28]

1.4 XML

1.4.1 Definícia

Extensible Markup Language (XML) sa používa na popis dát. Štandard XML je flexibilný spôsob vytvárania informačných formátov a elektronického zdieľania štruktúrovaných údajov prostredníctvom verejného internetu, ako aj prostredníctvom podnikových sietí.

XML kód, formálne odporúčanie z World Wide Web Consortium (W3C), je podobný Hypertext Markup Language (HTML). Obe jazyky XML a HTML obsahujú značkovacie symboly na opis obsahu stránok alebo súborov. HTML kód popisuje obsah webových stránok (hlavne textových a grafických obrázkov) iba z hľadiska toho, ako sa majú zobrazovať a ako sa má s nimi nárábať.

Dáta XML sú známe ako autorepresívne alebo samoregulujúce, čo znamená, že štruktúra dát je vložená s údajmi, takže pri príchode dát nie je potrebné predbežne stavať štruktúru na ukladanie dát; je dynamicky rozpoznaná v XML. Formát XML môže používať ktorýkoľvek jednotlivец alebo skupina jednotlivcov alebo spoločností, ktoré chcú zdieľať informácie konzistentným spôsobom. XML je vlastne jednoduchšia a ľahko použiteľná podmnožina Standard Generalized Markup Language (SGML), ktorý je štandardom pre vytvorenie štruktúry dokumentu.

Základný stavebný blok XML dokumentu je prvok definovaný značkami. Prvok má začiatočnú a koncovú značku. Všetky prvky dokumentu XML sú obsiahnuté v najvzdialenejšom prvku, ktorý je známy ako koreňový prvok. XML môže tiež podporovať vnorené elementy alebo elementy v rámci prvkov. Táto schopnosť umožňuje XML podporovať hierarchické štruktúry. Názvy prvkov opisujú obsah prvku a štruktúra popisuje vzťah medzi prvkami.

Dokument XML sa považuje za "dobro vytvorený" (čiže schopný čítať a rozumieť analyzátorovi XML), ak je jeho formát v súlade so špecifikáciou XML, ak je správne označený a ak sú prvky správne vnorené. XML tiež podporuje schopnosť definovať atribúty prvkov a opisovať charakteristiky prvkov v začiatočnej značke prvku.

Aplikácie pre XML sú nekonečné. Napríklad tvorcovia počítačov by sa mohli dohodnúť na štandardnom alebo bežnom spôsobe, ako opísať informácie o počítačovom produkte

(rýchlosť procesora, veľkosť pamäte atď.) A potom opísať formát informácií o produkte pomocou kódu XML. Takýto štandardný spôsob opisu údajov by používateľovi umožnil posielat' inteligentného agenta (program) na webovú stránku každého výrobcu počítača, zhromažďovať údaje a potom robiť efektívne porovnanie.

Výhody XML sa niekedy objavili v revolučnom rozsahu krátko po jej zavedení. Avšak, ako koncept, nedosahoval revolučný rozsah. Rovnako sa nestal všeliekom. Nadmerná aplikácia XML v mnohých oblastiach technológie znižuje jeho skutočnú hodnotu a vedie k zbytočnému zmätku. Pravdepodobne najviac škodlivé je predvídateľné správanie mnohých dodávateľov, ktorí chcú prepracovať XML pomocou vlastného súboru vlastných rozšírení. Hoci niektorí chcú pridávať hodnotu do XML, iní sa iba snažia, aby používatelia priľnuli k ich produktom.

Sila XML spočíva v jeho jednoduchosti. Môže mať veľké množstvo informácií a konsolidovať ich do dokumentu XML - zmysluplných častí, ktoré poskytujú informácie a štruktúru.⁵

1.4.2 Prvky a atribúty XML

V XML obsahu je možné buď zadávať ako prvok text (kratšie alebo dlhšie bloky textu objavujúce sa medzi počiatočným tagom a konečným tagom), alebo ako hodnoty atribútov (názov alebo pojem definujúci stav alebo hodnotu atribútu). Atribúty môžu byť buď vyčíslené (vyžadujúce, aby sa hodnoty vybrali zo zoznamu poskytnutého v DTD - kontrolovanom slovníku) alebo nevyčíslené (kde sa hodnoty volia voľnejšie). Vzhľadom na schopnosť riadiť obsah atribútov je dobré používať atribútové hodnoty pre riadenie softvérových procesov a textové elementy na doručenie správ ľudským čitateľom. DDI je vo veľkej miere v súlade s týmto princípom. TimeMeth element, ktorý by mal obsahovať opis časovej metódy alebo časového rozmeru zhromažďovania údajov, môže slúžiť ako príklad. Textový element môže slúžiť na poskytnutie popisu metódy v ľudskom jazyku, zatiaľ čo metodový atribút môže obsahovať riadený slovník, ktorý je ľahko pochopiteľný softvérovým systémom.

Príklad toho, ako môže byť táto štruktúra použitá pri konkrétnom zvýraznení, je uvedená nižšie:

⁵ [cit. 2018-01-28]

`<method> < dataColl > < timeMeth method = 'panel'> Štúdia je panelová štúdia, pri ktorej sa 50% vzorky nahradí nasledujúcou množinou ..... atď'. </ timeMeth > </ dataColl > </ method>`

1.5 Termíny a definície súvisiace s pochopením DDI

API

API (alebo aplikačné programovacie rozhranie) je formát jazyka a správ používaný aplikačným programom na komunikáciu s operačným systémom alebo s iným riadiacim programom, ako je napríklad databázový systém (DBMS) alebo komunikačný protokol. Aplikácie API sa implementujú písaním volaní funkcií v programe, ktoré zabezpečujú prepojenie s požadovaným podprogramom na vykonanie.

Katalóg

Katalógy obsahujú množiny subjektov metadát, identifikátorov a popisov súvisiacich položiek zahrnutých v registri. Registre možno považovať za inteligentné katalógy so zdokonalenými funkciami, ktoré umožňujú klasifikáciu objektov.

Občianstvo

Byť prispievajúcim členom celého životného cyklu dát a uvedomujúc si, že jeden je súčasťou väčšieho vedeckého obrazu.

Čistenie

Metódy používané na "vyčistenie" zberu údajov, napr. Kontrola konzistencie, kontrola divokých kódov atď'. Atribút "agentúra" umožňuje špecifikáciu agentúry vykonávajúcej čistenie údajov.

Codebook

Dokument, ktorý poskytuje informácie o štruktúre, obsahu a usporiadaní dátového súboru.

Spoločenstvo

V týchto osvedčených postupoch sa pojem spoločenstvo (komunita) používa na identifikáciu akéhokoľvek zoskupenia osobných alebo organizačných subjektov na rôznych

úrovniah formálnej organizácie, ktoré uvažujú alebo vykonávajú implementáciu DDI. Príklady: národná štatistická služba, výrobca údajov, archív, konzorcium archívov údajov.

Komponent

Softvér so špecifickým účelom s dobre definovaným vstupom a dobre definovaným výstupom.

Ovládacie operácie

Metódy na uľahčenie kontroly údajov vykonávané primárnym vyšetrovateľom alebo archívom údajov. V tomto prvku DDI by mali byť špecifikované všetky špeciálne programy používané na takéto operácie. Atribút "agentúra" možno použiť na označenie agentúry, ktorá vykonala riadiacu operáciu.

Kontrolovaná slovná zásoba (CV)

V širšom zmysle sa CV môže pohybovať od krátkeho zoznamu jasne vymedzených, vzájomne sa vylučujúcich a vyčerpávajúcich termínov, ktoré sú jedinými možnosťami použitia v konkrétnom kontexte (napr. Populácia určitých prvkov alebo atribútov DDI) prostredníctvom klasifikácie na niečo také zložité ako tezaurus s tisíckami termínov a termínových vzťahov. CV bol tiež opísaný ako "Súbor predmetových výrazov a pravidiel ich použitia pri priradovaní termínov k materiálom na indexovanie a vyhľadávanie." V CV sa termín skladá z jednej alebo viacerých slov, ktoré slúžia na vyjadrenie konceptu (príklad: "strach", "ženy", "starostlivosť o dieťa"). Termíny sú vyberané z prirodzeného jazyka na zaradenie do kontrolovanej slovnej zásoby.

Prechod pre chodcov (crosswalk)

Štruktúrovaný model, ako jeden zoznam položiek mapuje do súvisiaceho zoznamu položiek. Mapy v DDI Comparison sú príkladom prechodu.

Úložisko údajov a znalostí

Miesto (môže byť virtuálne), kde sa nachádzajú produkty (metadáta, údaje a ďalšie súvisiace informácie) životného cyklu údajov.

Cyklus životnosti údajov

Fázy dátovej zložky výskumného procesu, od štúdia konceptualizácie až po analýzu a archiváciu údajov, ktoré sa vrátili k predchádzajúcim etapám. Tento proces bol často znázornený ako lineárny, ale schéma vložená do najlepšej praxe “Discovery and Dissemination” ponúka z pohľadu používateľa iný pohľad naň.

DDI

Pri používaní bez verzie v týchto osvedčených postupoch sa DDI odvoláva na najnovšiu špecifikáciu DDI, aktuálne verziu 3.0. Keď sa odkazuje na staršie verzie, číslo verzie bude explicitne špecifikované.

Aplikácia DDI

Softvérová aplikácia, ktorá číta a / alebo zapisuje DDI XML.

Inštancia DDI

Instance DDI je obálka najvyššej úrovne pre akýkoľvek dokument DDI. Môže obsahovať súbor prvkov najvyššej úrovne, ktoré vo všeobecnosti zodpovedajú modulárnemu rozdeleniu v rámci DDI. Každá inštancia DDI použije tento obal bez ohľadu na jeho obsah.

Profil DDI

Mechанизmus popisujúci vybranú podskupinu prvkov a atribútov organizácie.

Schéma DDI

Schémy sú udržiavateľné zoznamy prvkov metaúdajov, ktoré môžu byť zverejnené samostatne a opakovane použité viacerými štúdiami. Sú základom pre zdroje, ako sú banky otázok, koncepčné banky a premenlivé banky. Konštrukcia schém zohľadňuje ich potenciálne opätovné využitie inými.

Schéma DDI s "vzťahmi"

Jedna z podskupín systémov DDI, ktoré sa môžu vzťahovať na iné schémy DDI.

Schéma DDI s "Závislosťami"

Jedna z podskupín systémov DDI, ktoré vyžadujú existenciu iných systémov DDI.

Objav (Discovery)

Stratégie a procesy, ktoré koncový používateľ používa na vyhľadávanie a prístup k produktom (metadáta, údaje a ďalšie súvisiace informácie) životného cyklu údajov. Šírenie informácií: distribúcia údajov s cieľom prístupu koncového užívateľa k produktom (metadáta, údaje a ďalšie súvisiace informácie) životného cyklu údajov.

DNS

System DNS (Domain Name System) prekladá doménu a názvy hostiteľov na adresy IP. Prekladá názvy domén zmysluplné pre ľudí do číselných (binárnych) identifikátorov spojených so sieťovými zariadeniami na účely lokalizácie a adresovania týchto zariadení na celom svete. Často používaná analógia vysvetľujúca systém doménových mien je, že slúži ako "telefónny zoznam" pre internet tým, že prekladá hostiteľské počítačové hostiteľské počítače do adres IP. Napríklad www.example.com sa prekladá na 208.77.188.166.

DTD

Definícia typu dokumentu je jeden z niekoľkých jazykových schém SGML a XML a je tiež výraz používaný na opis dokumentu alebo jeho časti, ktorá je vytvorená v jazyku DTD. DTD je formálnym výrazom (v XML) štrukturálnych obmedzení pre triedu dokumentov XML. Konštrukty jazyka DTD sú deklarácie prvkov a atribútov.

EarlyBound

Špecifikácia referencie, ktorá požaduje presnú verziu.

Zatmenie (Eclipse)

Eclipse je pokračujúci projekt na podporu integrovaného vývojového prostredia s otvoreným zdrojovým kódom (IDE). Eclipse poskytuje rámec a základnú platformu (tzv. Platformu Eclipse), ktorá umožňuje organizácii vytvoriť integrované vývojové prostredie zo zásuvných komponentov softvéru poskytovaných členmi Eclipse.

Koncový používateľ

Osoba vykonávajúca prácu v životnom cykle údajov, pre ktorú sa vyžadujú metadáta DDI. Koncový používateľ pravdepodobne nebude ani vedieť o metadátach DDI v aplikácii, ktorú používa. Koncoví používatelia preklenujú životný cyklus údajov. Medzi príklady patria výskumné rady / orgány financovania, výskumní pracovníci, výrobcovia údajov, archivári, knihovníci, analytici údajov, správcovia registrov, výskumní analytici / autori.

Federované vyhľadávanie (federated search)

Federované vyhľadávanie je súčasné vyhľadávanie viacerých online databáz alebo webových zdrojov a je novinkou v automatizovaných, webových knižniciach a systémoch vyhľadávania informácií. Tiež sa často označuje ako portál alebo federatívny vyhľadávací nástroj.

Genericode

Genericode definuje štandardný formát pre definovanie číselníkov (tiež známy ako výpočtov alebo riadených slovníkov). Genericode si kladie za cieľ poskytnúť nasledujúce:

- Štandardný model a XML reprezentácie pre obsahu číselníku
- Štandardný model a reprezentácia XML dát spojených s položkami v zozname kódov
- Štandardný model a XML reprezentácie, ako nových číselníkov sú odvodené z existujúcich číselníkov

GNU LGPL

GNU Lesser General Public License (predtým GNU Lesser General Public License) je slobodný softvér, licencie vydanej Free Software Foundation.

Vláda (Governance)

Termín vládnutia sa tu používa na označenie procedúr spojených s postupmi rozhodovaní, riadenie a správu sád DDI metadát.

GUI

GUI je grafické (nie čisto textové) používateľské rozhranie do počítača.

HTTP

Skratka pre Hypertext Transfer Protocol, podkladový protocol, ktorý používa World Wide Web. HTTP definuje, ako sú správy formátované a prenášané, a aké akcie webové servery a prehliadače by mali prijať v reakcii na rôzne príkazy. Napríklad, keď zadáte adresu URL v prehliadači, to skutočne odošle príkaz HTTP k webovému server a nasmeruje ju načítať a odovzdať požadovanú webovú stránku.

HTTPS

HTTPS je skratka pre Hypertext Transfer Protocol cez protokol SSL (Secure Socket Layer). Jedná sa o protokol TCP / IP používaný webovými servermi pre prenos a zobrazenie webového obsahu bezpečne. Prenesené dáta sú šifrované, takže nemôžu byť prečítané každým, okrem príjemcu.

Identifikovateľný (v kontexte DDI)

"Identifiables" sú tie prvky DDI, ktoré nesú len základnú úroveň identifikácie: URN, číslo a názov.

Zahrnutie inline vs. referencie

Materiál je považovaný za zahrnutý do textu, ak je obsah explicitne zahrnutý. Začlenenie odkazom znamená, že materiál sa odkazuje na jeden dokument, ale je publikovaný inde.

Interné zverejnenie režimov DDI

To sa týka zverejňovania režimov DDI ako zdrojových balíčkov v rámci zadaného projektu, pracovnej skupiny alebo organizácie. Konkrétny projekt môže zahŕňať viac ako jednu organizáciu, napríklad projekt Eurobarometra.

Internacionalizácia

Internacionalizácia je proces plánovania a realizáciu produktov a služieb tak, aby mohli byť ľahko prispôsobené špecifickým miestnym jazykom a kultúram, procesom zvaným lokalizácia.

IP

Internet Protocol (IP) je číselné označenie (logická adresa), ktorá je priradená k zariadeniam, ktoré pôsobia v počítačovej sieti s využitím internetového protokolu pre komunikáciu medzi jednotlivými uzlami. Hoci IP adresy sú uložené ako binárne čísla, zvyčajne sú zobrazené v ľudske čitateľných záznamov, ako je 208.77.188.166 (pre IPv4).

Java

Java je programovací jazyk výslovne určený pre použitie v distribuovanom prostredí Internetu. Bol navrhnutý tak, aby mal "vzhľad a dojem" ako v jazyku C ++, ale je jednoduchší použiť ako C ++ a tvorí objektovo orientovaný programovací model.

Prenos vedomostí (knowledge transfer)

Akt zdieľania poznatkov získaných počas životného cyklu dát.

LateBound

Špecifikácia pre odkaz, ktorý vyžaduje najnovšiu verziu k dispozícii.

Logický záznam

Odkaz na záznam dát, ktorý je nezávislý na jeho fyzické umiestnenie. To môže byť fyzicky uložené v dvoch alebo viacerých miestach.

Udržiavacie agentúry

Tieto organizácie vlastnia objekty metadát ktoré udržiavajú, a len oni môžu vykonávať zmeny v týchto objektoch.

Sťahovanie (Migration)

Migrácia v kontexte DDI sa rozumie ako pohyby z DTD na XML schémy z hľadiska štruktúry dokumentu; a od DDI 2 do DDI 3, ako aj z DDI 3 späť do DDI 2, pokiaľ ide o prenášanie obsahu.

NCubes

Popisujú logickú štruktúru n-rozmerného pola, v ktorom každí koordinát pretína každ druhú dimenziu v jednom bode. Ncube bol navrhnutý pre použitie v zdrojovom kóde agregovaných údajov.

OAIS (Open archival information system)

Referenčný model vesmírnej komunity, ktorá upravuje všeobecné archívne činnosti a politiky. patrí sem:

- SIP: Submission information package
- AIP: Archival Information package
- DIP: dissemination information package

Zverejnené metadáta

Zverejnené metadáta sú považované za použiteľné mimo komunitu, ktorá vytvorila originálny dokument. Toto širšie publikum môže byť interné pre projekt alebo organizáciu,

alebo externé. Metadáta, ktoré je zverejnené, musia byť zabalené v inštancii DDI, s verziou, a sú k dispozícii pre opätovné využitie alebo odvolanie z vonka inštalácie. Obal ako inštalácia DDI nemusí nutne znamenať obal pre zverejnenie. Metadáta môžu byť balené pre dôvody iné ako publikovanie počas procesu jeho vnútorného vývoja. V týchto prípadoch sa nevyžaduje vytváranie verzií.

Register

Virtuálne, centralizované a štruktúrované databázy alebo portál, ktoré umožňujú vytvárať zoznamy, urobiť štruktúrované vyhľadávanie, a identifikovať a získať metadáta a prípadne informácie, ktoré sú rozmiestnené v sieti. Registre sú miesta, kde sú indexované viaceré typy zdrojov a sú viditeľné a dostupné na použitie v celej komunite; nezahŕňajú klastre serverov a nezávisia na spôsoboch získavania pre prístup k ich obsahu. Príklady DDI registrov by mohli byť otázka banky, pojem banky, spoločenské vedy katalógy prieskumu dáta a variabilný banky.

Balíček zdrojov (Resource package)

Balíček zdrojov je prostriedkom obalovania akejkoľvek sady DDI metadát pre odkazovanie ako súčasť študijnej jednotky alebo skupiny. Balíček zdrojov konštruuje materiály pre publikáciu, ktoré sú určené na ďalšie použitie viacerými štúdiami, projektami alebo skupinami užívateľov. Balíček zdrojov používa skupinový modul s alternatívnym prvkom najvyššej úrovne zvaným Balíček Zdrojov, ktorý sa používa na opis udržiavateľných modulov alebo schém, ktoré môžu byť používané viacerými jednotkami vykonávajúcimi štúdie mimo štruktúry skupiny.

Úloha (task)

Aktivita, ktorú osoba podstupuje na vytvorenie, editáciu alebo zobrazenie dokumentácie o dátach.

TTL (time to live)

Odbor v internetovom protokole (IP), ktorý určuje, koľko skokov môže paket cestovať pred likvidáciou alebo vrátením.

Unicode

Unicode je štandard výpočtového priemyslu umožňujúci počítačom konzistentne reprezentovať a manipulovať s textom vyjadreným vo väčšine svetových systémov písma.

Jedinečnosť

Globálne jedinečný identifikátor je špeciálny typ identifikátora použitého v aplikáciách, ktorý poskytujú referenčné číslo, ktoré je jedinečné v akomkoľvek kontexte.

URL

URL (Uniform Resource Locator, predtým Universal Resource Locator) je jedinečná adresa pre súbor, ktorý je prístupný na internete. Obyčajný spôsob, ako sa dostať na webové stránky, je zadať URL svojej domovskej stránky súboru do adresného riadku webového prehliadača. Avšak, každý súbor v rámci tohto webu možno tiež zadať pomocou adresy URL.

URN

URN (Uniform Resource Name) je internetový zdroj s názvom, ktorý, na rozdiel od adresy URL, má trvalý význam - to znamená, že majiteľ URN zdroja môže očakávať, že niekto (alebo program) bude vždy môcť nájsť tento zdroj.

Verzia Rationale

Jedná sa o voliteľný DDI prvok vo všetkých prvkoch s verzou a poskytuje miesto pre uvedenie dôvodu pre zmenu, napríklad oprava tlačovej chyby alebo opravu nepresného obsahu, ktoré môžu mať vplyv na analýzy uskutočnené pomocou staršieho obsahu.

Versionable (v kontexte DDI)

"Versionables" zahŕňajú podmnožiny DDI "identifikovateľných"(identifiable) prvkov. To sú prvky, pri ktorých si treba zaznamenávať zmeny v obsahu, a teda ďalšie atribúty súvisiace s verzou sú povolené.

Váženie (weighting)

Použitie postupov vzorkovania môže vyvolať potrebu použiť váhy na presnosť štatistických výsledkov. Ak bol vyvinutý váhový vzorec alebo koeficient, treba poskytnúť tento vzorec, definovať jeho prvky, a ukázať, ako je vzorec použitý na údaje.

XML editačný softvér, alebo XML "editor"

Aplikácie, ktoré uľahčujú vytváranie XML dokumentov poskytovaním pokynov ohľadom správneho použitia značiek založených na XML schéme, ktoré môžu byť vopred načítané do softvéru. XML editory tiež overujú dokumenty XML a pomáhajú pri výrobe platných dokumentov poukazovaním na existujúce chyby a zvyčajne naznačujú, ako by mohli byť chyby opravené. Príklady komerčných editorov XML sú XMLSpy, oXygen, XMetaL. Voľné editory sú tiež k dispozícii.

XPath syntaxe

Výraz XML Path Language (XPath) používa notáciu cesty, ako tie používané v URL, pre adresovanie častí XML dokumentu. Výraz sa vyhodnocuje, čím sa získa objekt z množiny uzlov, Boolean hodnoty, číslo, alebo typ reťazca. Napríklad výraz kniha / autor vráti sadu uzlov prvku <author> obsiahnuté v prvku <book>, pokiaľ sú tieto prvky sú deklarované v zdrojovom dokumente XML. Okrem toho výraz XPath môže mať predikáty (výrazy filtra) alebo volanie funkcie. Napríklad výraz kniha [@ type = "fiction"] sa vzťahuje k prvku <kniha>, ktorých atribút typ je nastavený na "fiction" ⁶

⁶ [cit. 2018-01-28]

2 Cieľ práce, metodika práce a metódy skúmania

V súčasnosti sa zberom, spracovaním a archiváciou dát zaoberá množstvo agentúr a inštitúcií. Veľakrát sa stretávame so situáciou kedy sú dáta zberané prostredníctvom dotazníkov vytvorených v textových editoroch alebo prostredníctvom online editorov. Tento spôsob umožňuje zozbieranie informácií, ktoré neobsahujú dostatočné doplnkové informácie – metadáta.

Cieľom práce je využitie štandardu DDI najvyššej verzie 3.2 XML, ktorý poskytuje riešenie na tento problém, k ukážke zberu, spracovania a archivácie dát na ilustračných príkladoch. Rovnako potrebné je ilustrovať možnosť použitia rozličných editorov či už na prácu s jazykom XML alebo priamo so štandardom DDI. Príklady použitia sme ukázali predložením konkrétnych XML kódov vygenerovaných aplikáciou a zároveň ukážkou vytvorenia daných položiek prostredníctvom grafických nadstavieb zvolenej aplikácie.

Kapitolu výsledkov práce sme rozdelili do viacerých častí. V prvej časti sme na základe inštalácie programov, preštudovania dokumentácie a zoznámenia sa s používateľským rozhraním, rovnako prieskumom rôznych diskusných stránok v sieti internet, opísali 5 najpoužívanějších editovacích programov pre prácu so značkovacím jazykom XML.

Ďalšiu časť sme venovali editačným programom, ktoré sú schopné pracovať s jazykom XML v štandarde DDI. V čase písania práce existovali 3 programy schopné tejto činnosti. Vzhľadom na nedostupnosť dvoch z nich, sme oslovili spoločnosť Colectica, ktorá nám na požiadanie poskytla evaluačnú verziu produktu Colectica Designer.

V poslednej časti sme tento poskytnutý softvér využili na ilustrovanie zberu dát prostredníctvom dotazníka, ktorý sme vytvorili v programe Colectica Designer. Postup vytvorenia dotazníka bol ukázaný prostredníctvom grafickej nadstavby softvéru a zároveň je na ukážku uvedený vygenerovaný XML kód. V časti spracovanie sme popísali možnosti importu dát, dátové typy, s ktorými sa dá v softvéri pracovať a možnosti vytvorenia vlastných položiek. V záverečnej podkapitole sme stručne zhrnuli možnosti archivácie pomocou nástrojov spoločnosti Colectica.

3 Výsledky práce a diskusia

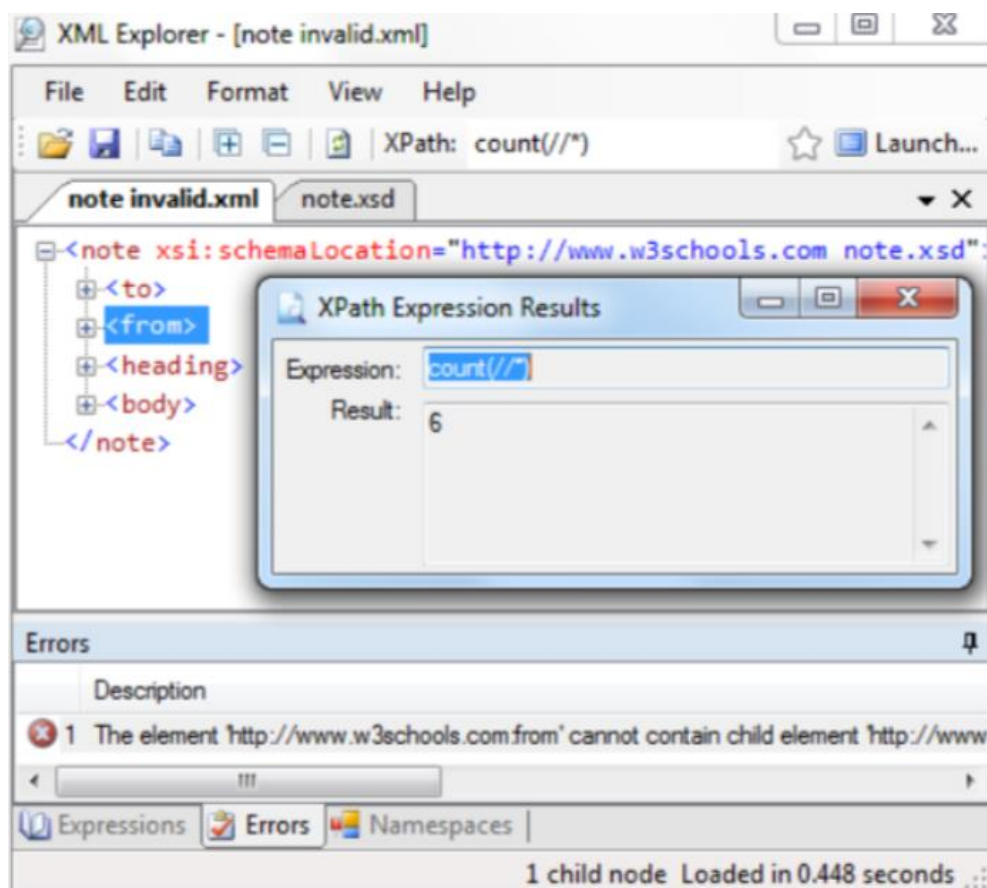
Časti výsledky práce a diskusia sme spojili do jedného celku. V tejto kapitole práce sme popísali možnosti práce so štandardom Data Documentation Initiative od zberu, cez spracovanie až po dokumentáciu, prípadne archiváciu dát. Operácie s dátami sme popísali s použitím softvérov schválených Alianciou Data Documentation Initiative pre prácu so spomínaným štandardom.

3.1 Softvér používaný na prácu s XML

V tejto podkapitole sme predstavili najpoužívanejší softvér na prácu s XML. Na trhu existuje viacero softvéru, ktorý je schopný čítať alebo editovať XML súbory, ide o softvér s platenou licenciou, ale takisto aj o voľne dostupné súbory. Ten softvér sa nazýva XML editačný softvér, alebo skráteno XML editor. Vzhľadom na to, že vytvorenie XML súboru nie je jednoduché pre bežného používateľa, editory sa stali nevyhnutnou súčasťou práce s týmto jazykom. Pre ilustráciu sme uviedli najpoužívanejšie a najlepšie hodnotené XML editory aktuálne dostupné na trhu:

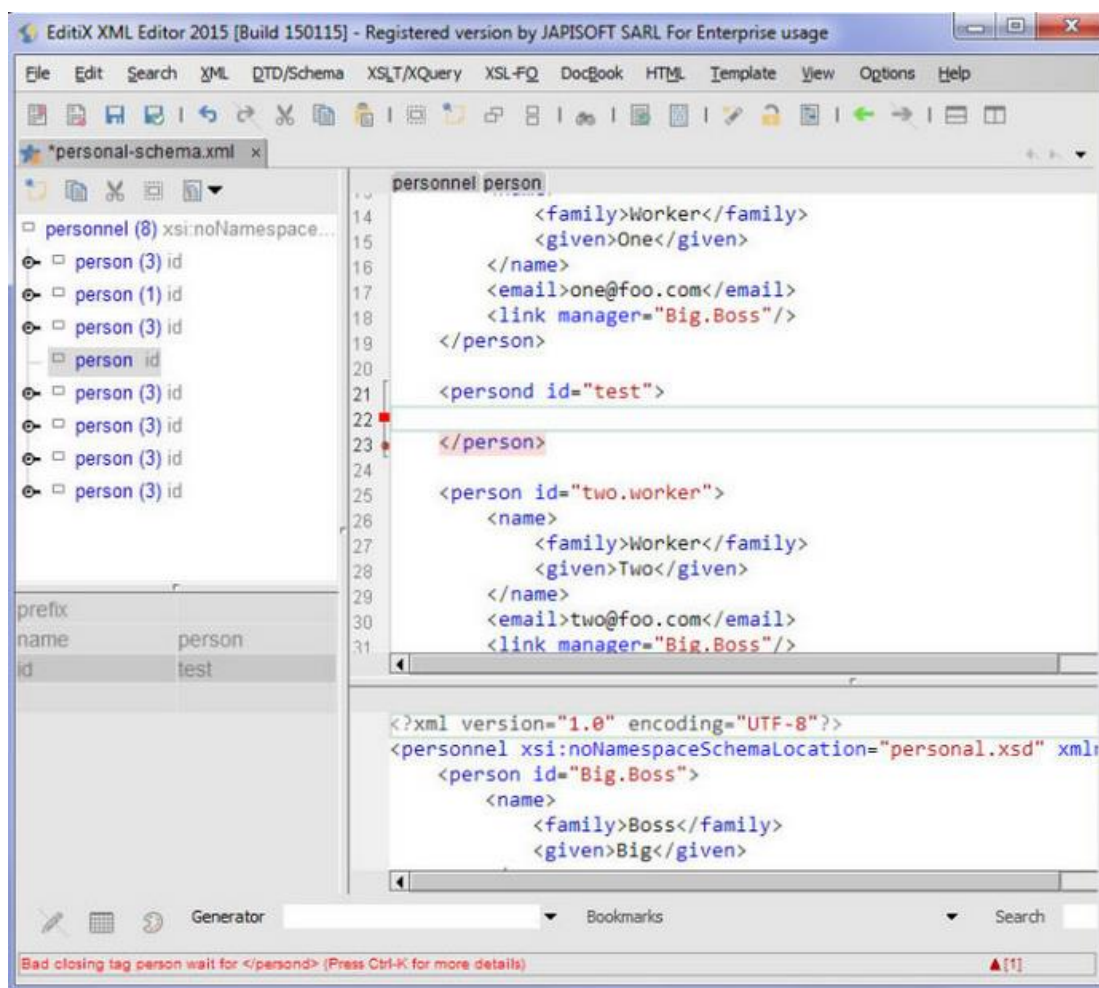
- XML Tree Editor - XML Tree Editor od spoločnosti SourceForge umožňuje základné operácie vrátane pridávania, úpravy či vymazávania textových uzlov spolu s ich atribútmi. Taktiež dokáže zobrazovať XML súbory vo forme stromu. Hlavným cieľom tohto softvéru je poskytnúť užívateľsky prívetivý nástroj na prácu s XML súbormi pre používateľov, ktorí nemusia mať nevyhnutne skúsenosti s XML. Medzi hlavné funkcie tohto softvéru patrí zabudovaný Free Pascal Lazarus, ktorý umožňuje ľahkú kompiláciu pre rôzne cieľové platformy a fakt, že XML Tree Editor podporuje jazykový preklad a tak jediný nástroj potrebný na vytvorenie nového prekladu je softvér samotný. Ďalšími dôležitými vlastnosťami sú prispôsobiteľnosť softvéru, vyhľadávanie, ktoré umožňuje vyhľadávať v textových hodnotách či skutočnosť, že textový uzol nie je oddelený od kontajnerovej značky, čo mu umožňuje obsahovať prakticky akúkoľvek hodnotu. Tento softvér predstavuje ideálny spôsob, ako bez problémov pracovať s XML súbormi bez plytvania časom.

- XML Explorer - XML Explorer je ľahký a rýchly nástroj, ktorý v súčasnej dobe umožňuje iba zobraziť XML súbory. Veľkou výhodou tohto softvéru je, že dokáže spracovať obrovské súbory - bol testovaný na súboroch s veľkosťou až 300 MB. Tento softvér umožňuje používateľom rýchle prezeranie a vyhľadávanie, kopírovanie formátovaných údajov XML, vyhodnotenie XPath výrazov a validáciu XSD schém. Medzi hlavné funkcie tohto produktu patrí validácia XML dokumentu pomocou XSD schémy špecifikovanej v dokumente, zobrazenie zoznamu chýb tejto validácie. Súčasťou XML Explorer je expresná knižnica, ktorá ukladá a spravuje bežne používané výrazy XPath. Softvér podporuje zobrazenie viacerých kariet dokumentov naraz a taktiež používa dokovateľné panely. Nevýhodou XML Explorer je momentálne neexistujúca podpora pre úpravu XML súborov.



OBRÁZOK 1: Ukážka používateľského rozhrania programu XML Explorer

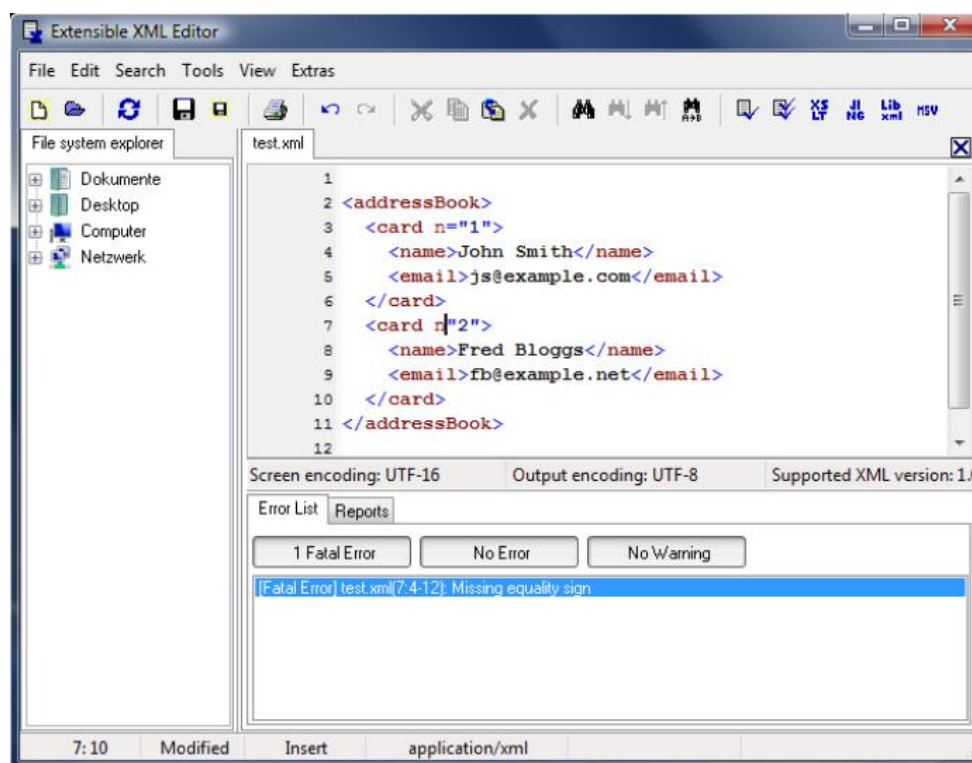
- XML Notepad 2007 - XML Notepad 2007 ponúka používateľom jednoduché a intuitívne rozhranie, ktoré umožňuje prehliadať a upravovať XML dokumenty. Tento nástroj vyvinula spoločnosť Microsoft. Medzi vybrané dôležité funkcie softvéru patrí synchronizácia jednotlivých druhov zobrazenia XML dokumentu tak, aby bolo možné rýchlo upraviť názvy a hodnoty uzlov. Nástroj XML Notepad 2007 využíva inkrementálne vyhľadávanie v textových i stromových zobrazeniach, čo znamená, že pri písaní hľadaného textu sa automaticky naviguje na zodpovedajúce uzly. Podporovanými funkciami sú vystrihnutie, kopírovanie, vkladanie a dokonca i „drag&drop“, čo umožňuje jednoduchšiu manipuláciu so stromovou štruktúrou. Funkcionalita „drag&drop“ funguje i pri prenášaní zo súborového systému i naprieč rôznymi inštanciami nástroja XML Notepad 2007. Pre jednotlivé kroky je neobmedzený počet krokov vrátiť. Zaujímavou funkcionalitou je možnosť nájdenia a nahradenia textu, ktorá podporuje regulárne výrazy a XPath výrazy. Dôležitou súčasťou je validácia okamžitá validácia XML schémy pri úprave dokumentu. XML Notepad 2007 obsahuje tiež HTML prehliadač, ktorý dokáže spracovať a zobrazíť XML dokumenty.
- Editix XML Editor- Tento nástroj je nielen XML editor, ale aj XSLT editor, ktorý je kompatibilný so systémom Windows, ale i s ďalšími operačnými systémami. Editix XML Editor je navrhnutý tak, aby pomohol autorom webových stránok a programátorom využívať najnovšie technológie XML a XML – podobné technológie, ako sú XSLT/FO, DocBook a XSD Schema. Editix XML Editor ponúka používateľom rozsiahlu škálu funkcií XML v prepracovanom IDE, ktoré je schopné navigovať používateľov s inteligentným pomocníkom, čo znamená, že aj začiatočníci môžu ľahko získať povedomie o tom, ako tento nástroj funguje. Softvér dokáže v reálnom čase vyhodnocovať XPath výrazy a detekovať syntaktické chyby. Nástroj má podporu pre rôzne šablóny a projektové riadenie. Editix XML Editor je určený najmä pre autorov webových stránok, tvorcov aplikácií a programátorov.



OBRÁZOK 2: Ukážka používateľského rozhrania programu EditiX XML Editor

- Essential XML Editor - Essential XML Editor je ľahký nástroj pre textové editovanie XML dokumentov. V tomto editore je k dispozícii dostatok kľúčových funkcií, ktoré používateľom poskytnú všetko, čo potrebujú. Medzi hlavné funkcie tohto nástroja patrí zabudovaný XML validátor spolu s DTD validátorom. Editor tiež obsahuje niektoré pokročilejšie funkcie, ako sú pluginy pre externé organizácie ako napr. W3C XML Schema validátor. Medzi ďalšie funkcie, ktoré tento softvér podporuje, patrí vrátenie / opätovné vykonanie kroku, vyhľadanie a nahradenie, skratky pre každý príkaz, vstavaný súborový prehliadač a ďalšie. K dispozícii sú tiež podrobné nastavenia stránok

a dialógové okno pre ukážku tlače. Program je dodávaný aj s externým hexadecimalným editorom, ktorý je možné prispôbiť podľa predvolieb užívateľa.



OBRÁZOK 3: Ukážka používateľského rozhrania programu Essential XML editor

Týchto päť programov je označovaných viacerými používateľmi za najlepšie editačné programy pre XML pre rok 2018. Ako môžeme vidieť na obrázkoch so zachytenými používateľskými rozhraniami, programy sú si veľmi podobné. Pravú časť rozhrania tvorí prehľadný strom riešenie kde sa na základe zložiek a podmnožín týchto zložiek dá veľmi efektívne postupovať pri selektovaní žiadaného projektu alebo súčasti projektu. Vrchná strana ponúka vždy určitú paletu tlačidiel pre konkrétne akcie a hlavnou časťou je samotná editačná časť, ktorá nám vo forme textu zobrazuje daný kód XML súboru. Ostatné moduly programov sa líšia; niektoré poskytujú možnosť kontroly chýb, vizuálne predstavujú vývoj verzii súboru s časovými značkami (história vývoja dokumentu) alebo jednoducho ponúkajú možnosť zobrazenia tzv. progresového panela, ktorý zaznamenáva všetky akcie, ktoré používateľ

vykonan. Výber toho správneho softvéru je založený na individuálnych preferenciách daného používateľa.

3.2 Codebook ako nevyhnutný dokument pre používateľa dát

Termín codebook nie je prekladaný do slovenčiny pretože je veľmi často používaný v práci s DDI štandardom a ďalšími podobnými štandardmi aj v slovenskej praxi a jeho preklad teda nemá veľký zmysel. Codebook predstavuje nevyhnutný dokument, ktorý informuje používateľa dát o štúdií, dátových súboroch, premenných, kategóriách atď., ktoré v konečnom dôsledku tvoria kompletný dátový set. Codebook môže obsahovať záznamovú vrstvu dátového setu, zoznamy mien premenných a značiek používaných na označovanie jednotlivých sekcií z dôsledku jednoduchšej orientácie dátovým setom, konceptuálne modely, prípady použitia, kódy pre chýbajúce hodnoty, poznámky, štatistiky frekvencie najčastejšie položených otázok a odpovedí k nim a takzvané reakcie „univerza“. Univerzum predstavuje cieľovú skupinu ľudí, na ktorých je štúdiá alebo teda sada štúdií zameraná.

Štandard DDI je veľmi náročný na vytvorenie bez pomoci softvéru, pretože už samotné XML a jeho schémy obsahujú mnoho predpisov, ktoré musia byť dodržané v vytvorení validného súboru XML. Štandard DDI množinu týchto predpisov ďalej rozširuje. To však ale neznamena, že jeho použitie prácu sťažuje. Naopak, po spracovaní vo vhodnom softvéri je tvorba štúdií, dotazníkov a ďalších projektov pre človeka ďaleko zrozumiteľnejšia. Práve vytvorenie codebooku pomáha dátovým setom byť omnoho zrozumiteľnejšie. A čo je možno ešte dôležitejšie, vzhľadom na to, že trend veľkých dát za posledné roky spôsobil obrovský niekoľkonásobný nárast veľkosti dát v sieti internet, vytvorenie codebooku umožňuje archiváciu a dokumentáciu dát pre zjednodušenie opätovného použitia v budúcnosti. Codebook dáva možnosť jednoduchých citácií a rovnako poskytuje inštrukcie pre používateľa o tom ako čítať, analyzovať, interpretovať a overovať dáta z dôvodu presnosti a pre možnosť replikácie.

Editačný softvér, spĺňajúci štandardy DDI dokáže načítavať dáta z viacerých zdrojov, či už ide o textový súbor alebo o balík dátových súborov, vždy musia byť známe informácie o štúdií prípadne sade štúdií, o premenných a dátových súboroch.

3.3 Nástroje používané na prácu s DDI

Funkcionalita bežných XML editačných programov neumožňuje zároveň aj editovanie súborov XML so štandardom DDI. V čase písania práce existovalo niekoľko programov, ktoré umožňovali určitý typ operácií so štandardom DDI. V tejto podkapitole sme popísali tie najrozšírenejšie, ktoré sú oficiálne schválené aj Alianciou Data Definition Initiative. Väčšina týchto programov je aj navzájom kompatibilná a dopĺňajú sa funkcionalitou.

Programy umožňujúce export a transfér dát vrámci tohto štandardu:

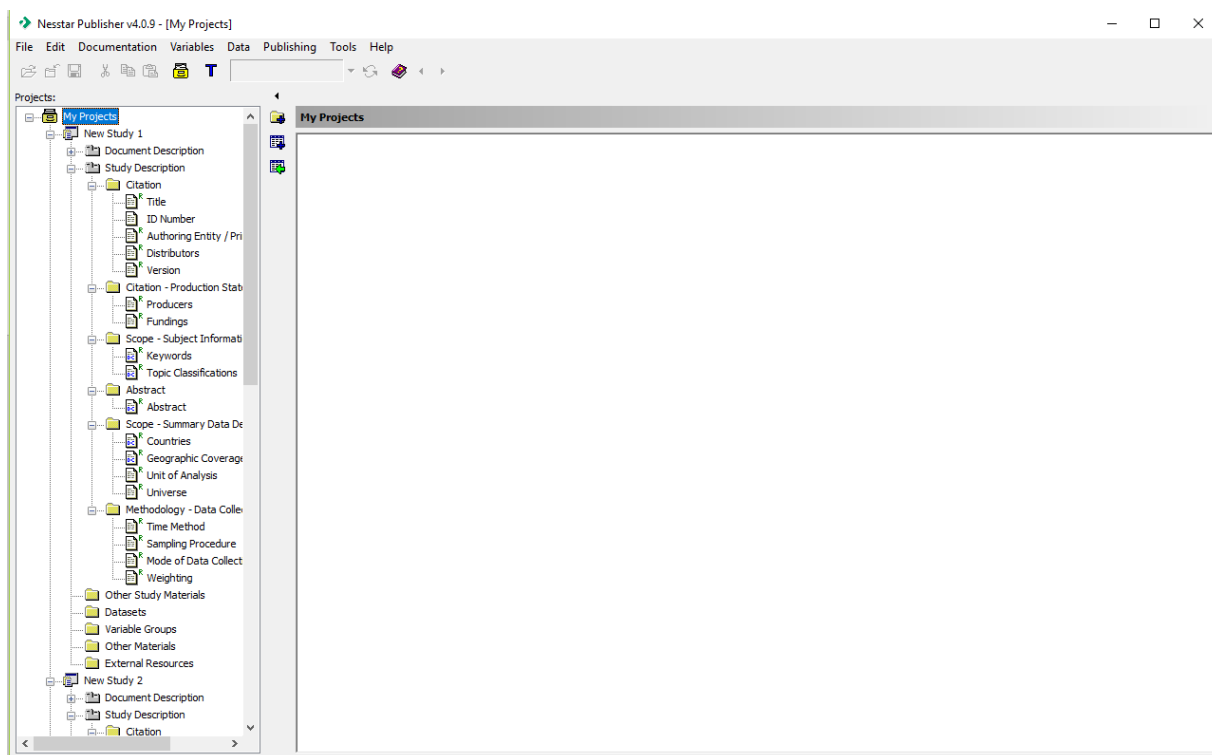
- StatTransfer – softvér je vyvinutý na zjednodušenie transféru štatistických dát medzi rozličnými programami, keďže dáta vytvorené jedným programom sú často potrebné v inom kontexte, či už ide o prezentáciu, ďalšiu analýzu, čistenie alebo korekciu. Z tohto dôvodu je pri transfére týchto dát potrebné aby boli premenné opísané značkami, menami, hodnotami a chýbajúcimi hodnotami. Bez použitia softvéru ako StatTransfer je manuálna príprava takéhoto procesu je veľmi časovo náročná. StatTransfer je schopný exportovať rozličného formáty do DDI-Lifecycle verzie 3.1 XML (najnovšia verzia 3.2 nie je podporovaná)
- SledgeHammer – softvér zameraný na dátovú konverziu a produkciu metadát. Je časťou balíka nástrojov nazývaného OpenDataForge (OpenDF). Medzi hlavné funkcie patrí čítanie a extrakcia metadát z rozličných dátových zdrojov a formátov. Je schopný zapisovať dáta v ASCII textovom formáte s rozličnými optimalizačnými možnosťami. Pre DDI štandard je zaujímavý najmä preto, že dokáže produkovať štandardizované metadáta zo zdrojových súborov vo všetkých aktuálnych verziách DDI Codebooku a DDI LifeCycle. V poslednom rade je možnosť vytvárania skriptov na načítanie ASCII dát do rôznych platforiem

Pre situácie kedy má používateľ existujúce DDI metadáta v XML formáte a je požiadaný aby tieto dáta prezentoval, boli vytvorené „štýlové listy“, ktoré sú schopné zobrazit' tieto metadáta v codebooku na čítanie vo formáte PDF alebo HTML. Spomenuli sme dva najčastejšie používané konvertory

- DDI-Lifecycle verzia 3.1 konverter používajúci XSLT štýlové listy – konvertovanie DDI 3.1 XML na MARC-XML, DataCite 2.2 XML, XHTML (DDA) alebo generický XHTML, DDI-Codebook XML a PDF
- Generický DDI-Codebook HTML konvertor – tento konvertor umožňuje konvertovať DDI metadáta vo formáte DDI-Codebook v akejkoľvek verzii na HTML súbor a teda umožňuje zobrazenie napr. v internetovom prehliadači (používaný najmä pri štúdiách kde sa autor odkazuje na internetový zdroj)
- XPDF – tento jednoduchý konvertor umožňuje konvertovať vytvorený codebook vo formáte PDF na ASCII text

Najdôležitejším softvérom na prácu s DDI štandardom sú pochopiteľne DDI editačné programy. V čase písania práce sme mali možnosť využiť tri existujúce editory:

- DdiEditor – tento editor dokáže produkovať DDI-Lifecycle XML súbory; v čase písania práce bola dokumentácia neprístupná a editor nebolo možné stiahnuť, nainštalovať a otestovať prácu s ním. Išlo o dočasný výpadok na strane spoločnosti vyvíjajúcej tento softvér DDA – Danish Data Archive.
- Nesstar Publisher – je nástroj na prácu s dátami a metadátami, prípravu a publikovanie dát z prieskumov a spracovanie ostatných materiálov pre Nesstar systém (Nesstar Server). Takisto môže byť použitý aj samostatne ako nástroj na spracovanie metadát konzistentných s DDI štandardom. Program je schopný tvoriť codebooky vo formáte PDF a DDI-Codebooky avšak iba vo verzii 1.2 XML. Na softvér Nesstar Publisher a Nesstar Server sa nám nepodarilo zohnať licenciю, preto sme mali možnosť zoznámiť sa iba s trial verziou programu, ktorá bola značne obmedzená. Free verzia nepodlieha podpore nesstar teamu preto ich nebolo možné ani kontaktovať.



OBRÁZOK 4: Používateľské rozhranie programu Nesstar Publisher

- Colectica – predstavuje súbor programových prostriedkov, ktoré dokážu produkovať PDF codebooky aj DDI-Lifecycle vo všetkých dostupných XML verziách. Z tohto dôvodu sme si vybrali tento softvér na ilustrovanie práce s DDI štandardom. Podrobnejšie sa softvéru venujeme v nasledujúcej podkapitole

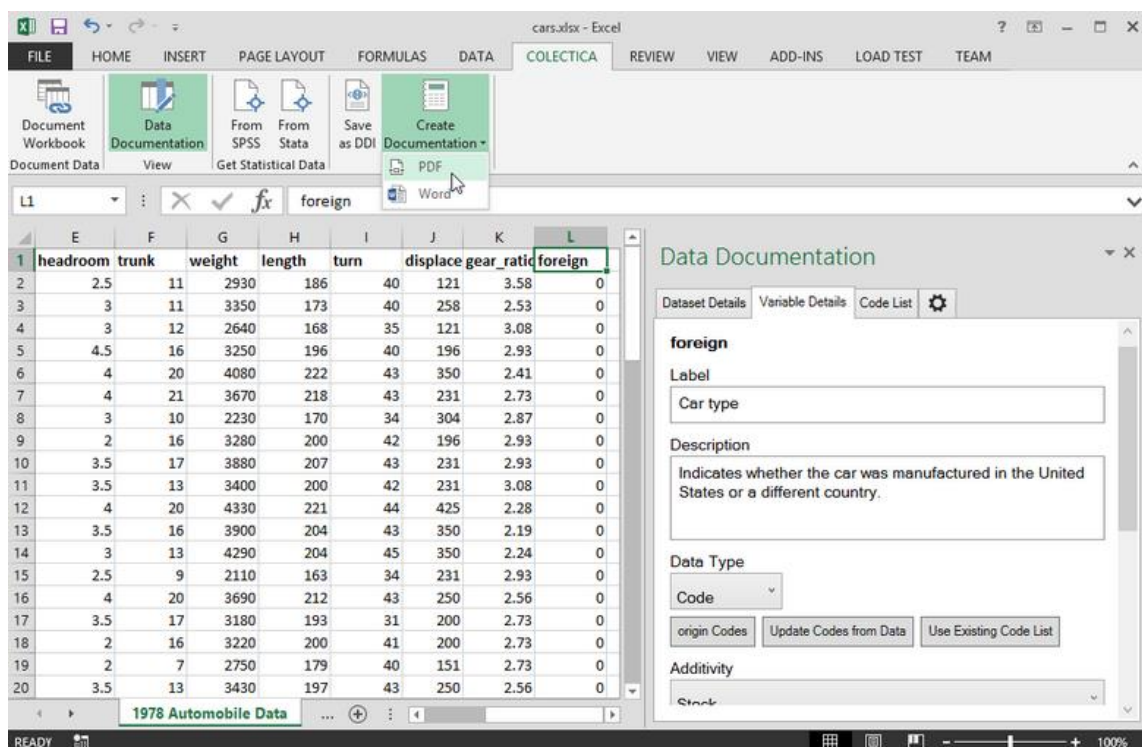
3.4 Programové prostriedky spoločnosti Colectica

Vzhľadom na skutočnosť, že na vytvorenie XML dokumentu v štandarde DDI je potrebné použitie softvéru (táto skutočnosť bola popísaná v predchádzajúcich podkapitolách) a softvér spoločnosti Nesstar (Publisher, Server) a DdiEditor boli nedostupné na prácu, či už z licenčných dôvodov alebo z dôvodov nedostupnosti serverov, zvolili sme si možnosť využitia programových prostriedkov spoločnosti Colectica. Spoločnosť sme oslovili so žiadosťou o pridelenie dočasnej licencie na niektorý zo softvérov, ktorý spoločnosť vyvinula. Našej požiadavke vyhovel a poskytli nám evaluačnú verziu softvéru Colectica Designer.

Spoločnosť vyvinula viacero produktov:

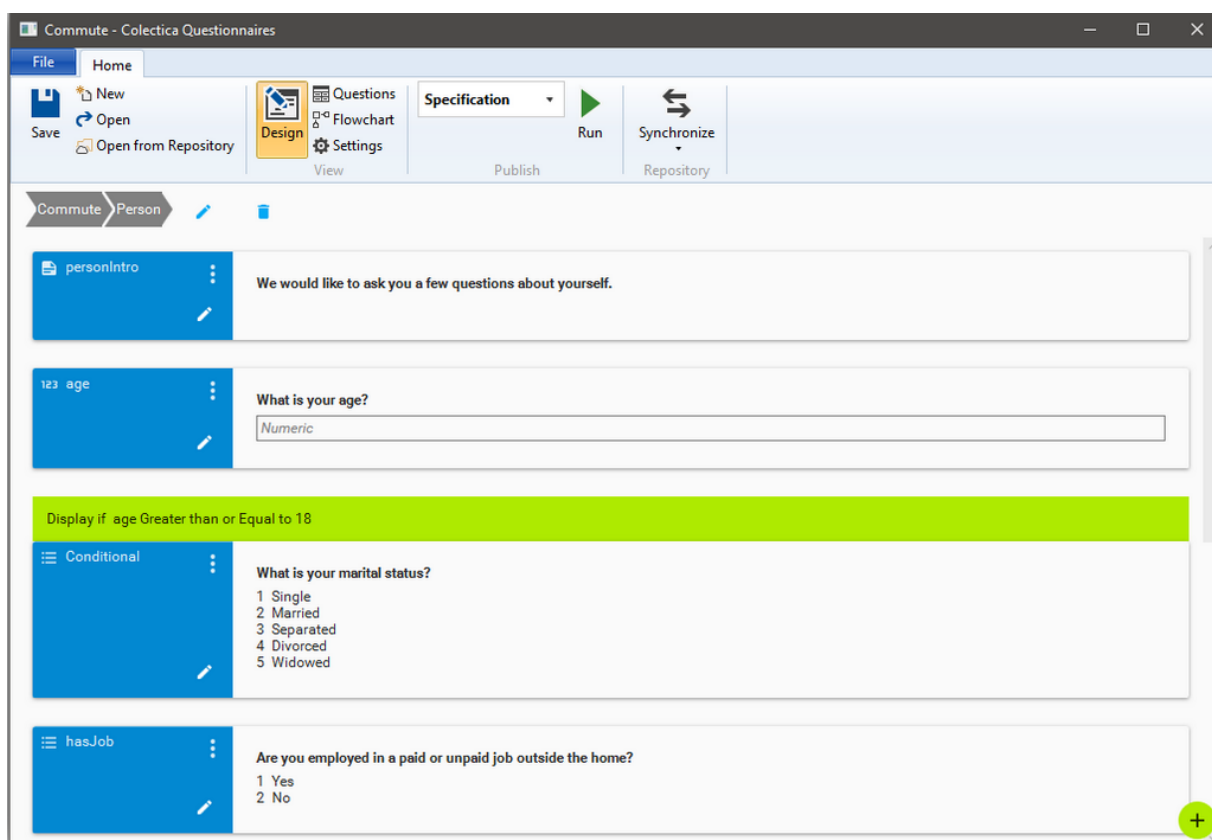
Desktopové aplikácie:

- Colectica Reader – predstavuje voľne dostupný softvér (je potrebná registrácia) na zobrazenie DDI metadát verzie 3. Slúži ako istý doplnkový softvér pre Colectica Designer a Colectica Questionnaires (a iné DDI aplikácie spoločnosti), ktorý umožňuje prezentáciu vytvorených XML dokumentov na zariadeniach, ktoré nespádajú pod licenčné dohody. Veľkou prednosťou softvéru aj napriek tomu, že nedokáže DDI metadáta vytvárať je schopnosť overiť správnosť zápisu týchto metadát podľa DDI 3 XML schémy
- Colectica for Excel – predstavuje add-in pre programový produkt spoločnosti Microsoft – Microsoft Excel. Umožňuje dokumentáciu dátových setov, premenných, publikovanie týchto dokumentáciu do rôznych štandardov, import SAS, Spss alebo Stata dokumentov do Excelu. Produkt je dostupný vo viacerých verziach, od štandardnej až po profesionálnu (štandardná verzia je voľne dostupná, ale jej funkcie sú značne limitované)



OBRÁZOK 5: Vizuálna ukážka add-inu Colectica for Excel priamo v Microsoft Excel

- Colectica Questionnaires – umožňuje tvoriť vizuálne reprezentácie prieskumov prostredníctvom dotazníkov (prieskumy tvoria hlavnú časť zberu štatistických dát pre tvorbu štúdií) bez potreby poznať DDI štandard alebo XML. Vytvorené otázky je možné zhlukovať do otázkových „baniek“, v ktorých je možné priradovať k otázkam dáta. Projekt je možné exportovať do DDI metadátového formátu a ďalej upravovať prostredníctvom Colectica Designera a iných aplikácií na to určených.



OBRÁZOK 6: Používateľské prostredie programu Colectica Questionnaires

- Colectica Designer – predstavuje hlavný produkt spoločnosti Colectica. Prostredníctvom programu je možné spravovať štatistické štúdie, vytvárať otázky do dotazníkov pre prieskumy, analyzovať všetky výsledky, generovať XML kód v DDI štandarde verzie 3, vytvárať dokumentáciu pre dátové sety, importovať a exportovať metadáta so štandardných formátov a automaticky

manažovať všetky verzie metadát pre zlepšenie procesu archivácie. Ilustračné príklady prípravy a spracovania metadát si uvedieme v ďalších podkapitolách

Serverové aplikácie:

- Colectica Repository – predstavuje jadro celej Colectica platformy. Priamo spolupracuje s klientskými aplikáciami (ako napr. Colectica Designer – v evaluačnej verzii, ktorú sme mali k dispozícii bola možnosť uloženia iba na lokálnych diskoch). Automaticky ukladá verzie projektov a jednotlivých metadách, s ktorými používateľ pracuje, poskytuje možnosť autentifikácie používateľa a tak zvyšuje bezpečnosť údajov
- Colectica Portal – predstavuje webovú aplikáciu založenú na Colectica Repository, ktorý umožňuje používateľom publikovať svoje štúdie a spolupracovať na základe autorizácie aj celým tímom ľudí
- Colectica Workflow Services – umožňuje používateľom podávať žiadosti o publikovanie ich súkromných DDI metadát na základe čoho administrátor môže presúvať informácie z jednej Repository do druhej
- Colectica RDF Services – je add-in dopĺňajúci zvyšné produkty umožňujúci volanie metadát DDI verzie 3 ako RDF

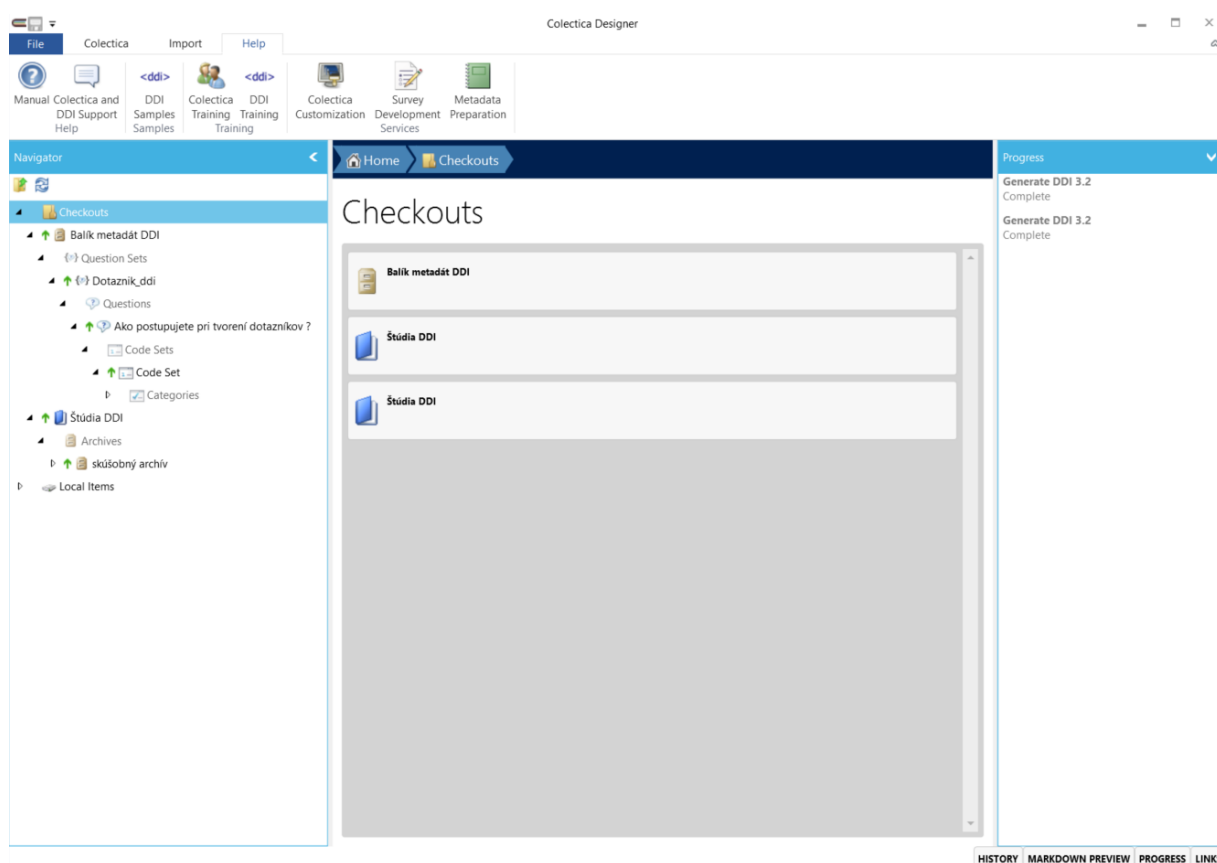
Nástroje pre developerov:

- Colectica Toolkit – je kolekcia nástrojov príkazového riadka navrhnutého tak, aby pomáhala organizácii vytvárať a spracovávať metadáta automatickým spôsobom; umožňuje konvertovanie DDI 3 súborov na DDI 2 a naopak, umožňuje vytvárať PDF dokumentácie a overovať správnosť metadát DDI 3
- Colectica SDK – predstavuje sadu .NET4 knižníc umožňujúcich overovať, čítať a zapisovať metadáta štandardu DDI 3 XML

Colectica umožňuje vytvorenie balíka pod názvom Colectica Fusion, ktorý umožňuje zákazníkov navoliť si komponenty, softvérové nadstavby a integrácie podľa svojej potreby a zároveň poskytuje možnosť konzultácie pri vytváraní metadátovej systémov architektúry.

3.5 Práca s dátami pomocou štandardu DDI

V tejto podkapitole sme na ilustračných príkladoch predstavili možnosti práce so štandardom DDI pomocou softvéru spoločnosti Colectica - Colectica Designer. K dispozícii sme mali evaluačnú verziu licencie pre tento softvér, ktorá mala obmedzené možnosti na prácu. Po obdržaní licencie sme sa zoznámili s programom. Celé používateľské prostredie je veľmi intuitívne. Spoločnosť Colectica nám poskytla dokumentáciu, ktorú sme si dôkladne preštudovali.

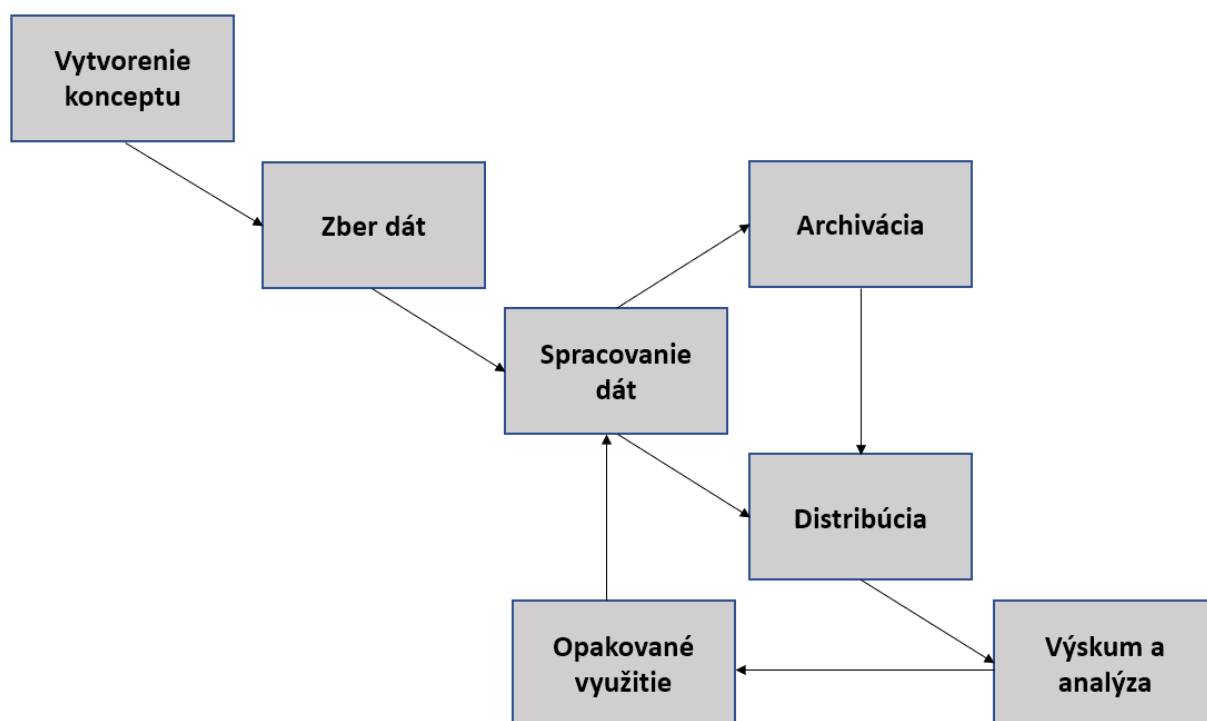


OBRÁZOK 7: Používateľské rozhranie programu Colectica Designer

Ako môžeme vidieť na obrázku vyššie, používateľské rozhranie programu Colectica Designer pripomína prostredie XML editorov, o ktorých sme písali v predchádzajúcich kapitolách. Vrchná časť ponúka tlačidlá na exekúciu konkrétnych akcií s programom alebo konkrétnym súborom. Ľavá strana ponúka prehľadný zoznam projektov a ich súčastí a umožňuje

jednoduchšiu navigáciu medzi konkrétnymi položkami. Stredná časť programu predstavuje samotný editor, do ktorého je možné zadávať textové informácie, priradovať dáta a metadáta do jednotlivých sekcií a spravovať jednotlivé projekty a ich súčasti. Práva časť umožňuje jednoduché prepínanie medzi zobrazovaním histórie (verzie dokumentov sú automaticky ukladané do Repository pri plnej verzii programu) a progresu v spracovaní dokumentov (každá akcia je zaznamenaná pre lepšiu orientáciu v práci).

Životný cyklus dát spracovaných štandardom DDI sa líši podľa použitia softvéru, spôsobu prístupu k spracovaniu dát, konceptu dát, ktorý si navrhne a ďalších faktorov. Nasledujúcim obrázkom sme sa pokúsili vystihnúť podstatu tohto procesu a jeho jednotlivých fáz.



OBRÁZOK 8: Životný cyklus dát spracovaných štandardom DDI

3.5.1 Zber dát prostredníctvom programu Colectica Designer

V tejto podkapitole si ukážeme jeden z možných spôsobov zberu dát prostredníctvom programu Colectica designer. Nový projekt je vždy vytváraný kliknutím na tlačidlo *Create New Checkout*. Hierarchiu dokumentov v projektoch môžeme sledovať v ľavom bočnom paneli programu v sekcii Checkouts. Plná verzia programu umožňuje vytvorenie projektu s pomenovaním, značkou a opisom. Do tejto položky je následne možné vkladať skupiny navzájom súvisiach položiek ako napríklad:

- Štúdie – dátovo centrované štúdia obsahujúce informácie o danej problematike
- Sady štúdií – skupiny štúdií, ktoré sú buď opakované alebo na seba tématicky nadväzujú
- Metadátové balíky – kolekcie znovupoužiteľných metadátových definícií, ktoré nie sú viazané iba na konkrétnu štúdiu ale ako položka sa dajú použiť vo viacerých štúdiách alebo publikovať ako samostatný XML dokument
- Inštrument – ide o samostatnú položku, ktorá môže obsahovať otázky dotazníka
- Dátový súbor – súbor obsahujúci informácie o premenných danej štúdie alebo sady štúdií

Podmnožinou týchto položiek sú položky obsahujúce konceptuálne a základné informácie:

- Koncept – opisuje základné myšlienky, ktoré štúdia preskúmava
- Konceptová sada – predstavuje súbor navzájom nadväzujúcich konceptov danej štúdie, pri ktorých sú automaticky ukladané ich verzie
- Univerzum – táto položka opisuje populáciu alebo objekty, na ktoré sa daná štúdia zameriava
- Sada Univerz – predstavuje súbor univerz, ktoré majú podobné vlastnosti
- Organizácia – predstavuje informácie o inštitúcii alebo spoločnosti vykonávajúcej danú štúdiu
- Sada organizácii – predstavuje súbor organizácii so spoločnými vlastnosťami

Poslednou podmnožinou základných položiek sú dátové položky a kolekcie

- Dátové kolekcie – informácie o procese zberu dát
- Otázka – patrí do súboru Inštrumentu, predstavuje otázku kladenú respondentovi
- Sada otázok – najčastejšia položka vytváraná v dotazníku, predstavuje súbor otázok, ktoré na seba tématicky nadväzujú
- Sekvencia – ide o znovupoužiteľnú sekvenciu otázok a kontrolovaných konštrukcií
- Premenná – znovupoužiteľná položka predstavujúca logický opis stĺpca dát
- Sada premenných – súbor podobných premenných
- Dátové vzťahy – položka opisujúca rozloženie dátového súboru
- Rozloženie záznamov – opisuje, ktoré premenné a v akom poradí vstupujú do dátového setu
- Fyzická štruktúra – opisuje formát dátového setu
- Dátový set – súbor obsahujúci dáta k spracovaniu programom

Evaluačná verzia programu Colectica Designer neumožňuje vytvorenie položky projekt, ale iba samostatných položiek, ktoré sme spomenuli v rozdelení vyššie. Táto skutočnosť však nespôsobuje ťažkosti pri snahe predstaviť fungovanie softvéru.

Každá nová hlavná vytvorená položka umožňuje pridávanie informácií pre lepšiu orientáciu. Vzhľadom na to, že štandard DDI pracuje s metadátami, vyplnenie týchto položiek je veľmi dôležité, pretože po publikácii štúdie poskytujú čitateľovi dodatočné informácie pre pochopenie problematiky.

- Názov a podnázov – kompletný a vysvetľujúci názov štúdie
- Alternatívny názov
- Autor – osoba alebo agentúra zodpovedná za tvorbu opisovanej položky
- Vydávateľ – osoba alebo agentúra zodpovedná za uverejnenie zdrojov v ich aktuálne dostupnej podobe
- Prispievateľ – osoba alebo agentúra, ktorá spolupracovala s autorom práce
- Dátum publikácie

- Identifikátor
- Licenčné práva autora

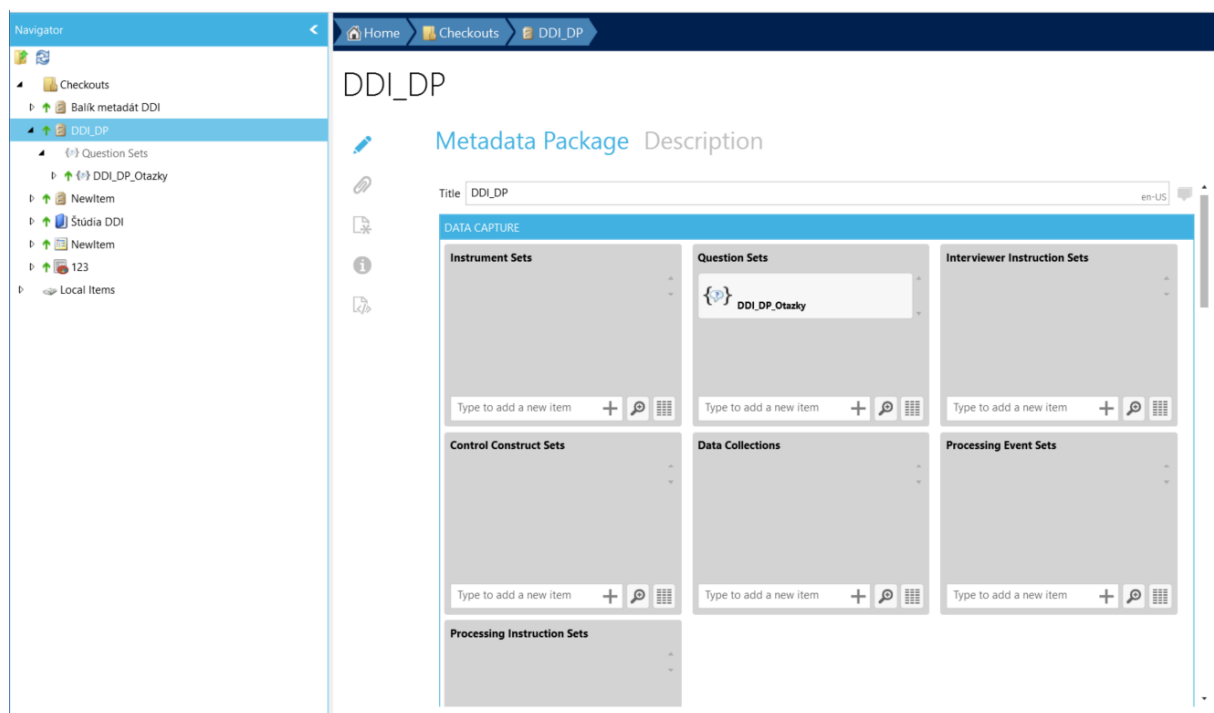
Jedným z najpoužívanejších spôsobov zberu dát je použitie dotazníkov. Samotný štandard DDI bol vyvinutý pre lepšiu manipuláciu s dátami pri prieskumoch, ktoré poväčšine bývajú vykonávané prostredníctvom už spomínaných dotazníkov. Vytvoríme si nový Checkout s typom položky Metadátový balík s pomenovaním „DDI_DP“. Názov agentúry ponecháme na „example.org“ čo predstavuje predvolený názov v prípade, že žiadnej agentúre nepodliehame. Táto akcia spôsobí vytvorenie novej položky v prehľadovom zozname Checkout, a zároveň nám otvorí editačné okno. V tomto kroku máme možnosť vybrať si jednu z podmnožín položiek. Pre našu prácu je zaujímavá sekcia „Zachytenie dát“. Medzi možnosťami tejto sekcie patria:

- Sady inštrumentov
- Sady otázok
- Sady inštrukcii anketára
- Sady kontrolovaných konštrukcii
- Dátové kolekcie
- Sady procesných udalostí
- Sady procesných inštrukcii

Medzi ďalšie sekcie položiek patria:

- Data
- Foundational
- Quality
- Packaging

Z ponúkaných možností sme si pridali novú Sadu otázok s pomenovaním „DDI_DP_Otazky“



OBRÁZOK 9: Vytvorenie sady otázok v programe Colectica Designer

Softvér Colectica zároveň generuje XML kód každej položky, ktorú vytvoríme. Pre ilustráciu uvádzame kód pre vytvorený Metadátový balík:

```
<Fragment xmlns:r="ddi:reusable:3_2" xmlns="ddi:instance:3_2">
  <ResourcePackage isUniversallyUnique="true" versionDate="2018-05-
01T16:03:02.7719203Z" xmlns="ddi:group:3_2">
    <r:URN>urn:ddi:example.org:9092f9ad-a01b-414b-8a8a-d955b85a6351:1 </r:URN>
    <r:Agency>example.org</r:Agency>
    <r:ID>9092f9ad-a01b-414b-8a8a-d955b85a6351</r:ID>
    <r:Version>1</r:Version>
    <r:Citation>
      <r:Title>
        <r:String xml:lang="en-US">DDI_DP</r:String>
      </r:Title>
    </r:Citation>
  </ResourcePackage>
</Fragment>
```

```

</r:Citation>
<r:QuestionSchemeReference>
  <r:Agency>example.org</r:Agency>
  <r:ID>7ba46fe3-e440-4a15-9eb7-d460136fbf24</r:ID>
  <r:Version>1</r:Version>
  <r:TypeOfObject>QuestionScheme</r:TypeOfObject>
</r:QuestionSchemeReference>
</ResourcePackage>
</Fragment>

```

KÓD 1: XML kód pre vytvorený metadátový balík „DDI_DP“

Následne máme možnosť pridávať:

- Samostatné otázky (Questions)
- Otázkové mriežky (Question Grids)
- Otázkové bloky (Question Blocks)
- Otázkové skupiny (Question Groups)

Zvolili sme si pridávania samostatných otázok. Zvolené otázky sú ilustratívneho charakteru. Zvolili sme si ich z dôvodu prezentácie možností odpovedí. Otázky, ktoré sme pridali sú nasledovné:

1. Aké je Vaše pohlavie?
2. Aký je Váš dátum narodenia?
3. Odkiaľ pochádzate?
4. V akom meste pracujete?
5. Na akej pozícii pracujete?
6. Aký je Váš priemerný mesačný príjem?
7. Akú značku notebooku by ste si vybrali z nasledovných alternatív ?
8. Preferujete matné displeje pred lesklými?
9. Koľko hodín denne strávite pri Vašom notebooku?
10. Aké je hlavné mesto Slovenskej republiky?

Každá z otázok obsahuje určité parametre. Názov otázky (značka pre lepšiu orientáciu), samotný text otázky a možnosti odpovedí (pozostávajúce z viacerých typov, ktoré je ďalej možné rozvíjať).

OBRÁZOK 10: Dialógové okno vytváranie novej otázky v programe Colectica Designer

Vygenerovaný XML kód vo svojom tele nesie názov otázky, značku pre lepšiu orientáciu v hierarchii projektu a taktiež aj kódovú skupinu pre pohlavie (M-Muž, Ž-Žena).

```
<Fragment xmlns:r="ddi:reusable:3_2" xmlns="ddi:instance:3_2">
  <QuestionItem isUniversallyUnique="true" versionDate="2018-05-
01T17:26:20.6923639Z" xmlns="ddi:datacollection:3_2">
    <r:URN>urn:ddi:example.org:167a7b53-7705-4fdf-a84f-115a4df75c92:1</r:URN>
    <r:Agency>example.org</r:Agency>
    <r:ID>167a7b53-7705-4fdf-a84f-115a4df75c92</r:ID>
    <r:Version>1</r:Version>
    <r:UserAttributePair>
      <r:AttributeKey>extension:QuestionInstruction</r:AttributeKey>
      <r:AttributeValue>{"en-US":"Respondent musí uviesť svoje pohlavie podľa príslušného
```

```

    kódu stanoveného v kóde odpovede"}</r:AttributeValue>
  </r:UserAttributePair>
  <QuestionItemName>
    <r:String xml:lang="en-US">Pohlavie</r:String>
  </QuestionItemName>
  <QuestionText audienceLanguage="en-US">
    <LiteralText>
      <Text>Aké je Vaše pohlavie?</Text>
    </LiteralText>
  </QuestionText>
  <CodeDomain blankIsMissingValue="false">
    <r:CodeListReference>
      <r:Agency>example.org</r:Agency>
      <r:ID>b744c262-b569-4a0d-ab9c-39f966017b51</r:ID>
      <r:Version>1</r:Version>
      <r:TypeOfObject>CodeList</r:TypeOfObject>
    </r:CodeListReference>
    <r:Label>
      <r:Content xml:lang="en-US">Pohlavie respondent</r:Content>
    </r:Label>
    <r:Description>
      <r:Content xml:lang="en-US">Zadajte Vaše pohlavie</r:Content>
    </r:Description>
    <r:ResponseCardinality minimumResponses="1" maximumResponses="1" />
  </CodeDomain>
  <r:ResponseCardinality minimumResponses="1" maximumResponses="1" />
</QuestionItem>
</Fragment>

```

KÓD 2: XML kód pre otázku „Aké je Vaše pohlavie?“ spolu s kódovanou odpoveďou

Zadefinovanie správneho formátovania odpovedí je jednou z najdôležitejších častí zberu dát prostredníctvom DDI štandardu. Pri nesprávnom type odpovedí nebude následne možné spracovať údaje do požadovaného formátu, resp. DDI validátor – jeden zo vstavaných nástrojov programu Colectica Designer nemusí overiť správnosť zápisu dát.

Ku každej otázke pristupujeme samostatne (pri väčšom počte významovo rovnakých otázok je možné zhľukovanie do skupín a priradovanie formátu odpovedí priamo skupine otázok). V koncepte je nutné uviesť si všetky možné druhy odpovedí, pretože nie každý respondent bude korektne dodržiavať pravidlá a predpisy dotazníka. Program musí následne vedieť identifikovať chybné zápisy. Prázdne odpovede môžu byť zadané ako chýbajúce hodnoty, v opačnom prípade bude dotazník vyhodnotený ako neplatný.

Odpoveď sme pridali kliknutím na tlačidlo pridania odpoved'ovej domény v dialógovom okne nami vytvorenej otázky. Je možné priradiť už predom vytvorený kód, ktorý spája kategórie s hodnotami. Pracovali sme s týmito typmi odpovedí:

- Textové domény – umožňujú nastavenie minimálneho a maximálneho počtu znakov odpovede tak aby odpoveď bola platná; nedodržanie tejto hranice vyhodnotí odpoveď ako neplatnú; väčšinou sa používa na špecifikáciu požiadaviek napr e-mailový účet

Ako príklad sme uviedli otázku „Na akej pozícii pracujete?“ – odpoveďou je textové pole s minimom znakov 10 a maximom 500

- Numerické domény – DDI štandard podporuje určité typy numerických hodnôt ako odpoved'ové domény: BigInteger (číslo s neobmedzenou veľkosťou), Integer (celé číslo s povolením pre negatívne hodnoty alebo s pravidlom len kladných čísel), Long (Integer s veľkosťou do 32 bitov), Short (Integer s veľkosťou do 16 bitov), Decimal (reálne číslo), Float (reálne číslo uložitelné vo vedeckých notáciách), Double (Float s veľkosťou do 32 bitov), Count (Ordinálny počet objektov v konečnej sade), Incremental (Hodnota ktorá je nekonečná, interval alebo pomer); opäť je možné stanoviť hranice minimálnej a maximálnej hodnoty, hodnoty je možné škálovať

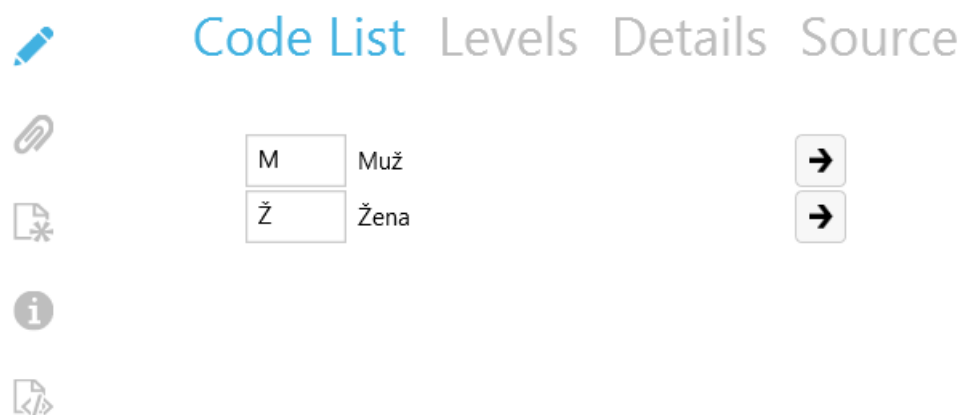
Ako príklad sme uviedli otázku „Aký je Váš priemerný mesačný príjem?“ – odpoveďou je Integer s povolením len kladných čísel

- Dátumové a časové domény – formát musí byť jasne stanovený v popise otázky pre respondenta aby nedošlo k neplatnému zápisu. Možnosti sú Date and Time, Date, Time, Year, Month, Day, Month and Day, Year and Month, Duration a Timespan

Ako príklad sme uviedli otázku „Aký je Váš dátum narodenia?“ – odpoveďou je Date

- Kódové domény – predstavujú dopredu vytvorené kódy, ktoré spájajú hodnotu s kategóriou

Ako príklad uvádzame otázku „Aké je Vaše pohlavie?“ – odpoveďou je M- Muž, Ž-Žena; 2 unikátne kódy kde M a Ž predstavuje hodnoty a Muž a Žena predstavuje kategórie k nim priradené



OBRÁZOK 11: Ukážka možného kódového listu pre použitie v otázkach

Na záver sme vygenerovali finálnu dokumentáciu ktorá slúži ako podklad pre realizátorov prieskumov. Ako príklad sme uviedli jednu z otázok, každý s parametrov otázky zadaných v XML kóde je pretransformovaný do vizuálnej reprezentácie otázky.

Pozícia

Question

Na akej pozícii pracujete?

Type

Text

Label

Pozícia

Description

odpoveďou je pracovná pozícia

Maximum Length

500

Minimum Length

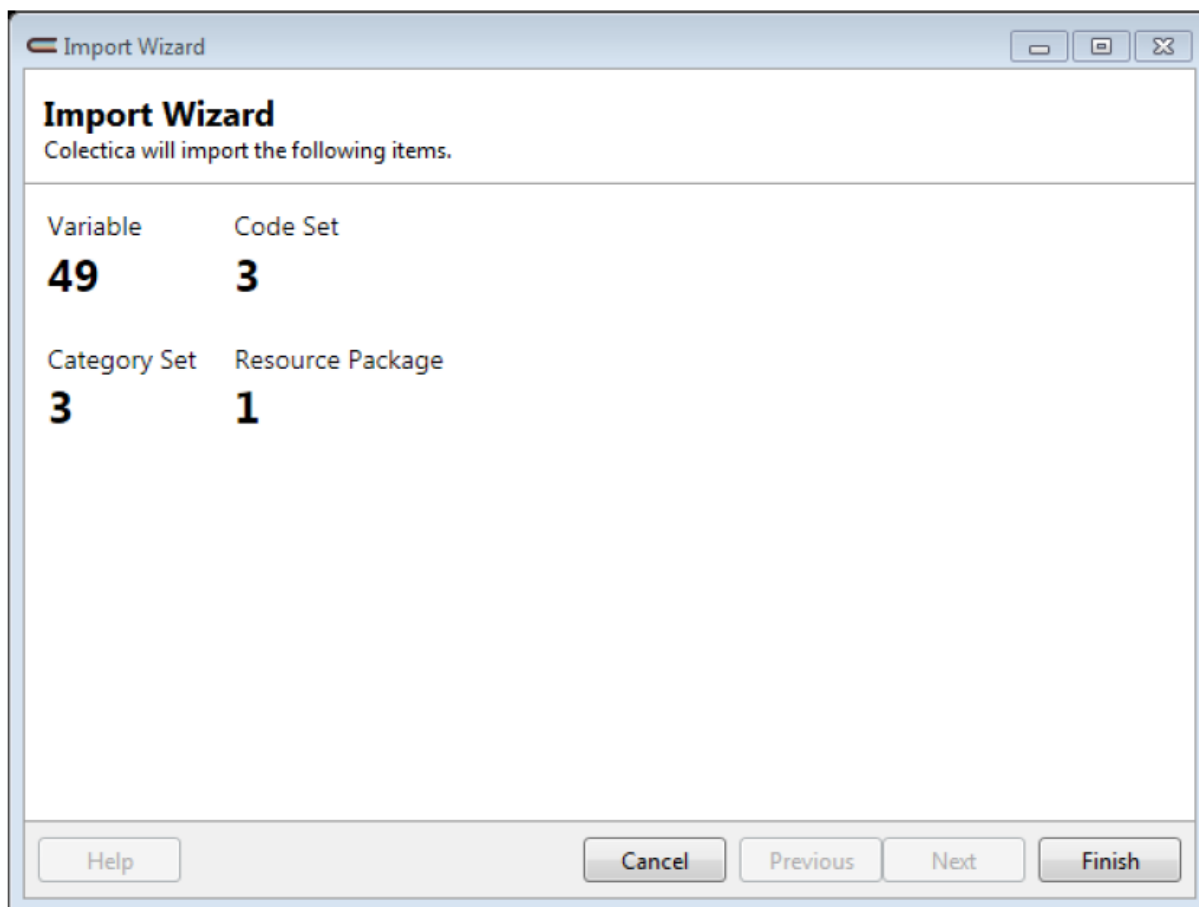
10

OBRÁZOK 12: Ukážka vygenerovanej dokumentácie ku otázke v softvéri Colectica Designer

3.5.2 Práca s dátami prostredníctvom programu Colectica Designer

V tejto podkapitole sme ukázali možnosti práce s dátami, ktoré zozbierame či už prostredníctvom nami vytvorených dotazníkov, alebo iným druhom výskumu. Program Colectica Designer dokáže importovať viacero formátov dát a následne s nimi pracovať. Základným pravidlom je, že dáta musia byť vo formáte DDI štandardu. Na overovanie tejto skutočnosti slúži DDI validátor, ktorý upozorní používateľa v prípade nevhodného formátu dát. V prípade, že formulár na zber dát bol vytvorený programom Colectica Designer a všetky odpovedové domény boli správne zadefinované, dátový balík je automaticky vyhodnotený ako platný a vhodný pre štandard DDI.

Colectica Designer ako jediný program z programových produktov spoločnosti Colectica má možnosť importovať dáta vo formáte DDI verzia 3. Pre ostatné produkty je používanie súborov vo formáte DDI verzie 3 natívnym formátom. Na import dát sa používa jednoduché dialógové okno, ktoré po výbere správneho súboru vypíše počty premenných, kódových sád, sád kategórií a zdrojových balíkov.

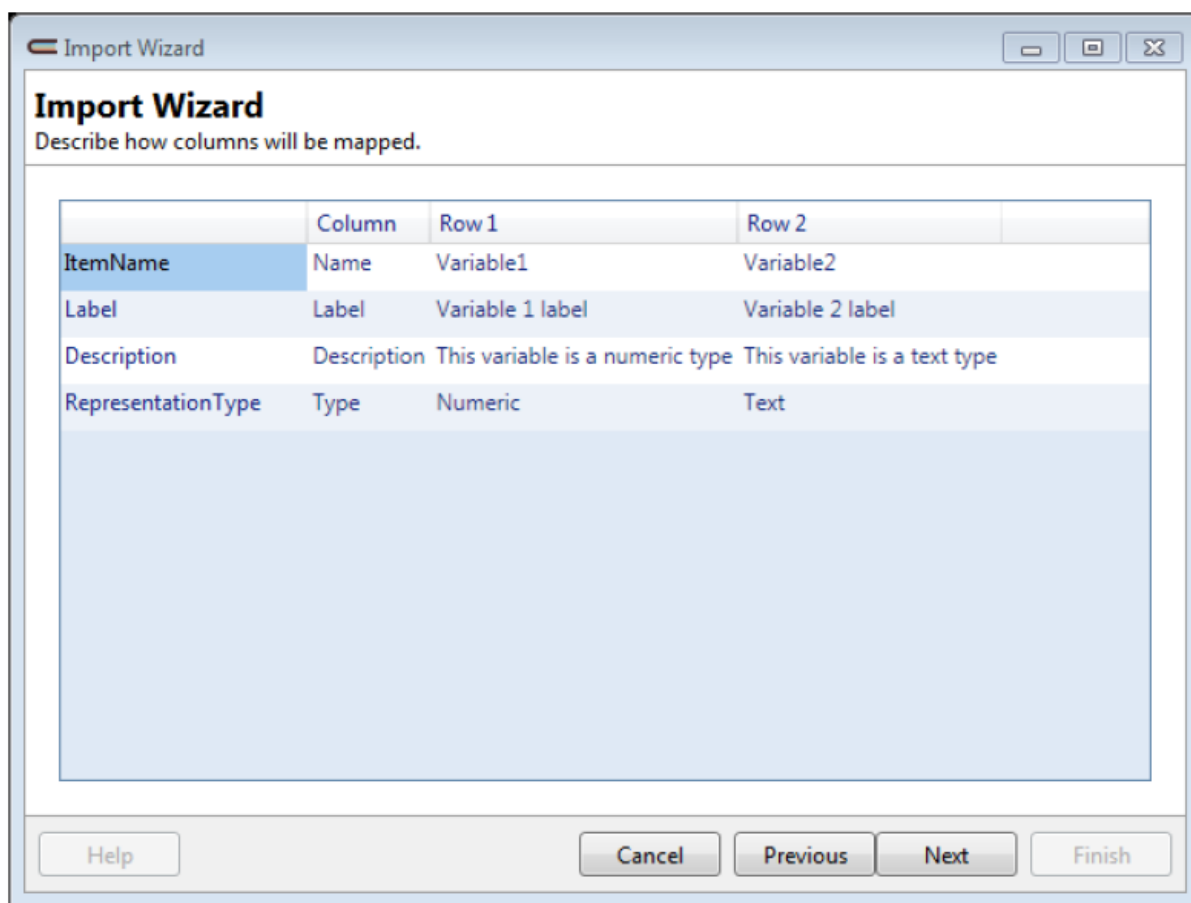


OBRÁZOK 13: Ukážka dialógového okna na import dát vo formáte DDI 3 XML

Ďalšie možnosti importu dát predstavujú:

- DDI 2 / Nesstar – Colectica Designer je schopný importovať dáta vo formáte DDI verzie 2 vytvorené softvérom Nesstar, ktorý sme spomenuli v predchádzajúcich kapitolách; import ostatných DDI 2 súborov je možný ale nie je priamo podporovaný ani testovaný. V tejto situácii je potrebné kontaktovať support spoločnosti Colectica, ktorý nám pomohli prekonvertovať náš zdrojový súbor do požadovaného formátu, ktorý vedel softvér prečítať. Je možné importovať viacero súborov naraz. Tieto súbory budú združené do spoločnej štúdie.
- Premenné – import premenných programom Colectica Designer je možný len v prípade že dané premenné sú v správnom formáte a v súboroch vytvorených

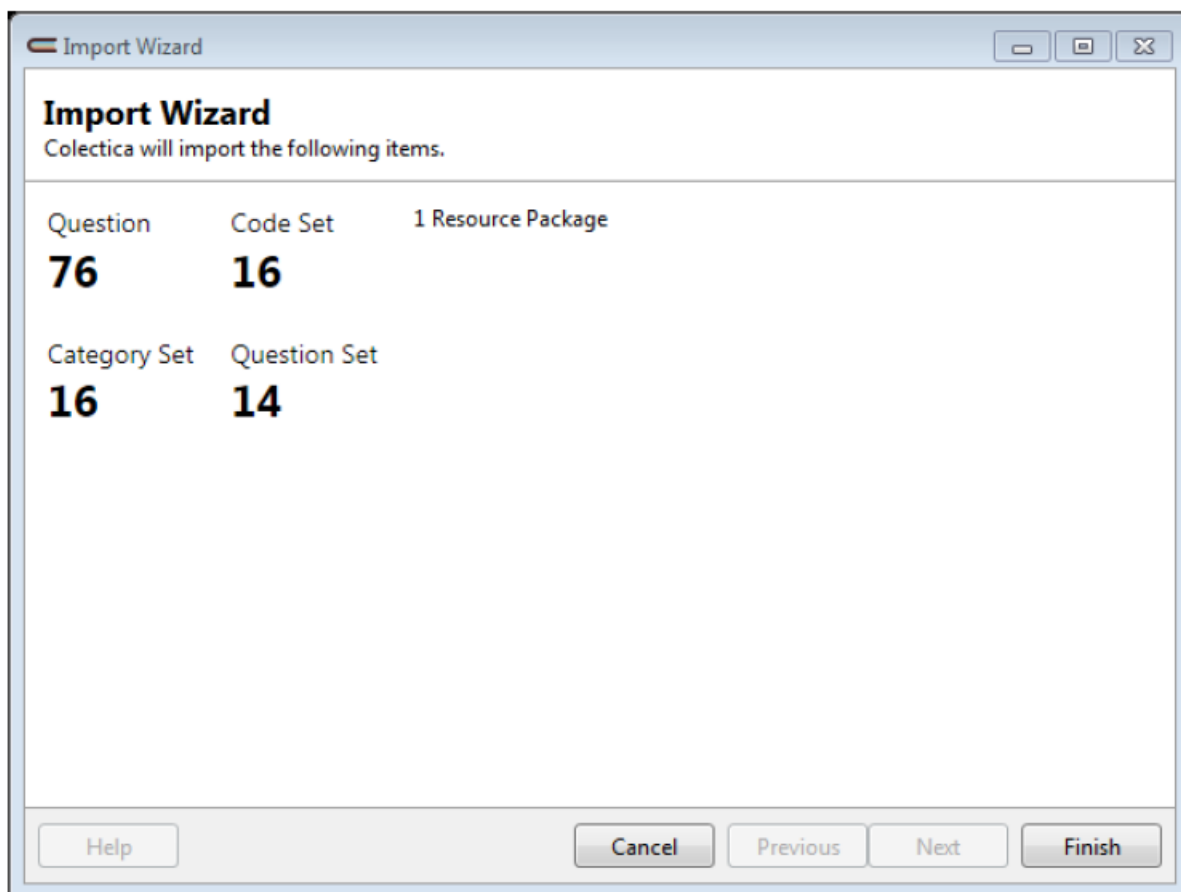
tabuľkovým editorom. Program pri zadaní súboru na import vypíše všetky možné stĺpce (premenné), ktoré našiel a indikuje metadátové polia do ktorých budú tieto premenné mapované. Tento spôsob nefunguje vždy dostatočne spoľahlivo, softvér nedokázal zachytiť každú z premenných, prípadne ich mapovanie nebolo presné. Z tohto dôvodu je lepšou voľbou vytvorenie nových premenných priamo v softvéri a ich priradenie do štúdie, v ktorej dané premenné potrebujeme. Na nasledujúcom obrázku je zobrazené dialógové okno pre import premenných ako ilustračný príklad:



OBRÁZOK 14: Import premenných do programu Colectica Designer

- Dotazníky vytvorené softvérom Blaise a CASES – každý vytvorený dotazník jedným z týchto dvoch softvérov priamo zameraných na tvorbu

prieskumových dotazníkov je možné importovať do programu Colectica Designer. Dotazníky sú importované ako položka Instrument. Po úspešnej validácii súboru s dotazníkom nám program v dialógovom okne vypísal počet otázok, kódových sád, sád kategórii a otázkových sád, ktoré je možné importovať.



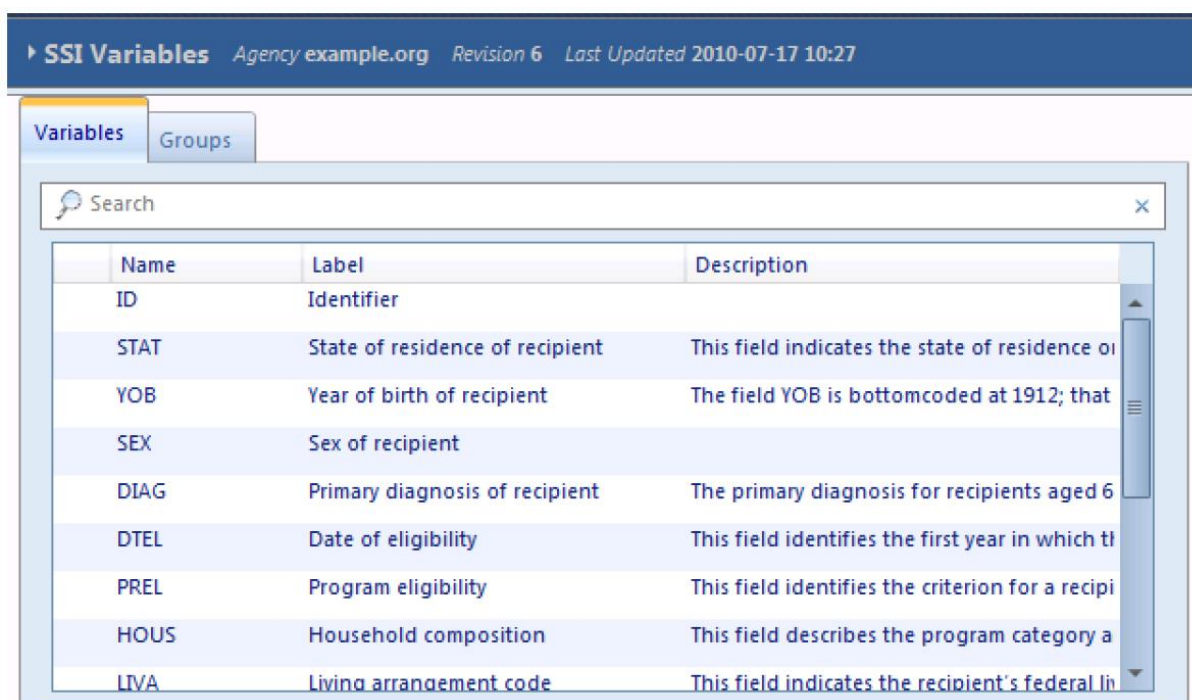
OBRÁZOK 15: Ukážka importu dotazníkov vytvorených softvérami Blaise a CASES do programu Colectica Designer

- SPSS súbory spoločnosti IBM, Stata súbory spoločnosti StataCorp – import týchto súborov prebieha podobne ako pri predchádzajúcich prípadoch

Dôležité je taktiež spomenúť, že pri každom druhu importu nám v záverečnom kroku ponúkne program Colectica Designer možnosť zmeny jazyka. Podľa predvolených nastavení detekuje jazyk automaticky.

V časti importu premenných sme spomenuli, že tento spôsob tvorby premenných nie je ideálny. Importované premenné sa nie vždy podarí načítať s kompletnými opismi a mapovaním. Je preto lepšie voliť možnosť vytvorenia vlastnej premennej. Evaluačná licencia programu Colectica Designer nám bohužiaľ neumožnila prácu s premennými v rozsahu ako by si záverečná práca vyžadovala.

Softvér umožňuje pridávať, mazať a editovať premenné v metadátovom balíku každého projektu. Každá z premenných je použiteľná v akejkolvek štúdii, nie je viazaná na jednu, pre ktorú bola vytvorená.



Name	Label	Description
ID	Identifier	
STAT	State of residence of recipient	This field indicates the state of residence of
YOB	Year of birth of recipient	The field YOB is bottomcoded at 1912; that
SEX	Sex of recipient	
DIAG	Primary diagnosis of recipient	The primary diagnosis for recipients aged 6
DTL	Date of eligibility	This field identifies the first year in which t
PREL	Program eligibility	This field identifies the criterion for a recipi
HOUS	Household composition	This field describes the program category a
LIVA	Living arrangement code	This field indicates the recipient's federal li

OBRÁZOK 16: Ukážka knižnice premenných v programe Colectica Designer

Každú premennú je možné otvoriť v dialógovom okne a editovať jej parametre. Medzi parametre patrí:

- Názov premennej
- Označenie pre lepšiu orientáciu v projekte
- Popis a definícia premennej
- Typ premennej – dátový typ (dátové typy boli bližšie rozpísané v časti odpoved'ových domén pri tvorbe dotazníku)
- Úloha premennej – opisuje funkciu premennej (identita, váha, čas, dátum)
- Analyzovaná jednotka – entita, na ktorú sa dáta viažu (inštitúcie, rodiny, skupiny, jednotlivci)
- Meracie jednotky, v ktorých je premenný zapísaná (míle, kilogramy)
- Jednotky pôvodu – informácie o poskytovateľovi informácie (respondent)
- Koncept – referencia na koncept viazaný na konkrétnu premennú
- Univerzum – referencia na definíciu univerza v metadátovom balíčku
- Zdrojové otázky – referencia na otázky použité na zber údajov
- Zdrojové premenné – referencia na premenné použité na zostrojenie danej premennej

V konečnom dôsledku po zozbieraní potrebných dát či už formou dotazníkov alebo importom predpripravených dát, Colectica Designer umožňuje vytváranie deskriptívnych štatistík a generovanie štatistických súborov. Táto možnosť v evaluačnej licencií programu nie je zahrnutá, ale po naštudovaní dokumentácie sme uviedli nasledovné štatistické jazyky, pre ktoré je možné vytvorenie súborov pozostávajúcich z definícií premenných a značiek:

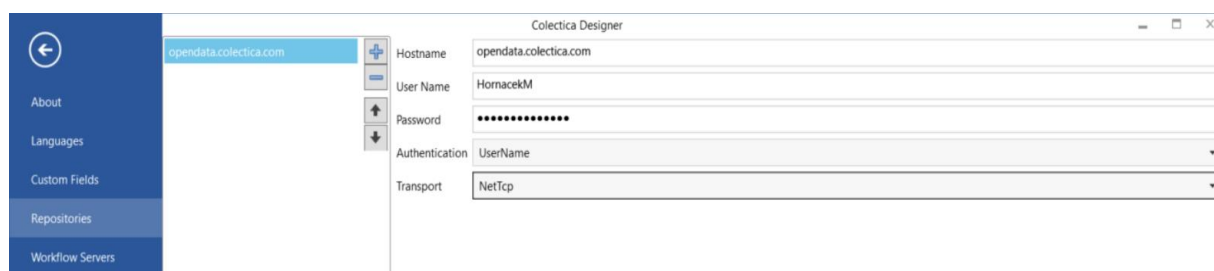
- R
- SAS
- SPSS
- Stata

3.5.3 Archivácia dát prostredníctvom programu Colectica Designer

V poslednej podkapitole sme stručne popísali možnosti archivácie a zálohovania dát prostredníctvom programu Colectica Designer. Ako sme už spomenuli v predchádzajúcich kapitolách, softvér Colectica Designer umožňuje export dát do rôznych formátov a teda aj jeho uloženie pre ďalšie použitie. Každý dokument alebo súbor vytvorený Colectica Designerom môže byť znova použitý ako zdrojový súbor pre ďalšie štúdie.

Najjednoduchší spôsob archivácie poskytovaný priamo spoločnosťou Colectica a plne kompatibilný s programom Colectica Designer je produkt Colectica Repository. Na najnižšom levelle dovoľuje program publikovať položky na lokálne disky. Tento spôsob zálohy ale nie je postačujúci pri dôležitých dokumentoch z dôvodu možnosti poškodenia pevných diskoch čo by viedlo ku strate dát. Colectica Repository ako priama súčasť programu Colectica Designer umožňuje uloženie dát na serveroch spoločnosti a teda záruku zachovania dát.

Colectica Repository je nutné v prvom kroku nakonfigurovať. Po zakúpení licencie je nutné kontaktovať prideleného admina, ktorý poskytne prihlasovacie meno a heslo a URL pridelenej Repository. Po vytvorení ľubovolnej položky v sekcii Checkouts je možné túto položku publikovať vybratím možnosti „Publish to Repository“. Položky, či už ide o dátové sety, metadátové balíky alebo kompletne štúdie, sú dostupné na vyžiadanie hneď po autentifikácii používateľa. Taktiež je možné priamo vytvoriť položku „Archív“ a pridať ju do metadátového balíčka štúdie. Táto položka je schopná uchovávať všetky informácie o položkách, na ktoré je odkázaná. Colectica Repository nie je k dispozícii v evaluačnej verzii softvéru, preto sme jej funkčnosť opísali bez vizuálnych ukážok.



OBRÁZOK 17: Ukážka konfigurácie Colectica Repository pre archiváciu dát

Záver

Výhodou štandardu DDI je jeho schopnosť vytvárať dátové štruktúry a metadáta, ktoré opisujú doplnujúce informácie k zozbieraným dátam a pomáhajú používateľom vytvárať kvalitnejšie štúdiá a hlavne umožňujú opakované použitie dát pri nadväzujúcich štúdiách.

Pri napĺňaní niektorých čiastkových cieľov sme narazili na problémy, kde programové prostriedky podliehali vysokým licenčným poplatkom alebo absolútne neposkytovali skúšobné verzie, prípadne boli tieto skúšobné verzie natoľko obmedzené, že demonštrácia funkčnosti nebola možná. Softvér spoločnosti Colectica – Colectica Designer – nám ale umožnil naplnenie hlavného cieľa našej práce, ktorým bol opis zberu, spracovania a archivácie dát prostredníctvom štandardu DDI. Evaluačná licencia pochopiteľne so sebou niesla určité obmedzenia, ale napriek tomu sme mali možnosť vytvorenia ilustračných príkladov. Program Colectica Designer a všeobecne štandard DDI je navrhnutý veľkou mierou hlavne na zber dát a ich archiváciu. Je kompatibilný s množstvom najpoužívanejších štatistických programov a formátov. Na spracovanie a výmenu dát je vhodnejší štandard SDMX.

V našej záverečnej práci sme mali jasne vytýčený cieľ a to oboznámiť čitateľa s možnosťou používania najpoužívanejších XML editačných programov a zároveň ukázať výhody štandardu DDI a programových prostriedkov určených na prácu so štandardom DDI. Následne ilustrovať použitie programu Colectica Designer na vytvorenie dotazníka na zber údajov, ukázať spracovanie a na záver archiváciu dát. Aj vzhľadom na obmedzenia softvérového vybavenia sme tento cieľ a čiastkové ciele naplnili a dodržali sme metodiku práce, ktorú sme si stanovili v druhej kapitole.

Bibliografické odkazy

1. RYSSEVIK, Jostein. *The Data Documentation Initiative (DDI) metadata Specification* [elektronický zdroj]. Nesstar Ltd, 12s. [cit. 2018-01-28]. Dostupné na: http://www.ddialliance.org/sites/default/files/ryssevik_0.pdf
2. DDI Alliance, kolektív [elektronický zdroj]. 2015, [cit. 2018-01-28]. Dostupné na: <https://www.ddialliance.org/about/work-products-of-ddi-alliance>
3. ROUSE, Margaret. *Metadata* [elektronický zdroj]. 2014, [cit. 2018-01-28]. Dostupné na: <https://whatis.techtarget.com/definition/metadata>
4. GREGORY, Arofan. - HEUS, Pascal. - RYSSEVIK, Jostein. *Metadata* [elektronický zdroj]. 2009, 24s. [cit. 2018-01-28]. Dostupné na: https://www.ratswd.de/download/workingpapers2009/57_09.pdf
5. ROUSE, Margaret. *XML (Extensible Markup Language)* [elektronický zdroj]. 2014, [cit. 2018-01-28]. Dostupné na: <https://searchmicroservices.techtarget.com/definition/XML-Extensible-Markup-Language>
6. DDI Alliance, kolektív [internetový zdroj]. 2018, [cit. 2018-01-28]. Dostupné na: <http://www.ddialliance.org/bp/definitions>
7. GREGORY, Arofan. – HEUS, Pascal. *DDI and SDMX: Complementary, Not Competing, Standards* [elektronický zdroj]. 2007, 17s. Dostupné na: http://odaf.org/papers/DDI_and_SDMX.pdf
8. DDI Aliance, kolektív. *Create a codebook* [elektronický zdroj]. 2018, Dostupné na: <http://www.ddialliance.org/training/getting-started-new-content/create-a-codebook>

Dokumentácie

9. Colectica, kolektív. *Dokumentácia* [elektronický zdroj]. 2017, Dostupné na: <https://docs.colectica.com/>

10. Nesstar, kolektív. *Nesstar Publisher v4 User Guide*, dokumentácia [elektronický zdroj]. 2011, Dostupné na:
http://www.nesstar.com/help/4.0/publisher/download_resources/Publisher_UserGuide_v4.0.pdf
11. Danish Data Archive, kolektív. Dokumentácia [elektronický zdroj]. 2011, Dostupné na:
<https://code.google.com/archive/p/ddieditor/wikis/Documentation.wiki>
12. Circle Systems, kolektív. *Stat/Transfer User manual version 14* [elektronický zdroj]. 2015, 205s. Dostupné na: <https://www.stattransfer.com/stman.pdf>
13. OpenDataForge, kolektív. *DataForge SledgeHammer Technical user guide* [elektronický zdroj]. 52s. Dostupné na:
<https://docs.google.com/document/d/16EMyzzw6bvwj0JsWITIfuqWjkY104P21opcgL3NYVZc/edit>
14. Editix XML Editor, dokumentácia [elektronický zdroj]. 2012, dostupné na:
<http://www.editix.com/doc/manual10/>
15. Essential XML Editor, dokumentácia [elektronický zdroj]. Dostupné na:
<http://www.philo.de/xmledit/>
16. XML Explorer, dokumentácia [elektronický zdroj]. Dostupné na:
<https://archive.codeplex.com/?p=xmlexplorer>
17. XML Tree Editor, dokumentácia [elektronický zdroj]. Dostupné na:
<https://sourceforge.net/p/xmltreeeditor/wiki/XMLTreeEditHelp/>