# Informačný model pre zostavenie štruktúrovaných popisov štatistických údajov 1. časť

*Ján Švolík*
*Národná banka Slovenska*

*Cieľom tohto článku je naznačiť proces, ako efektívne popísať, čo znamenajú, resp. obsahujú štatistické údaje. Po stručnom úvode do problematiky metadát, t. j. informácií o údajoch, nasleduje ukážka tradičného spôsobu prezentácie údajov do tabuliek, kde sa údaje a metadáta navzájom prelínajú. S narastajúcim objemom údajov vznikla potreba ich efektívnejšej organizácie. Na tento účel sa vytvorilo viacero systémov, ako spájať údaje s ich popismi a ako ďalej s nimi pracovať. V ďalšej časti je jednoduchý pohľad na dva systémy na prenos údajov, GESMES/TS a XBRL. Tieto systémy vznikli z potreby efektívneho prenosu údajov medzi inštitúciami. Ich informačné modely (hlavne GESMES/TS) boli inšpiráciou pri tvorbe zásad informačného modelu NBS. Popis princípov tohto informačného modelu, načrtnutie procesu budovania dátového modelu NBS a zhodnotenie, čo a ako sa podarilo premietnuť do štatistického zberového portálu, bude predmetom pokračovania v nasledujúcom čísle.*

S rastúcim objemom rôznych štatistických údajov vzniká potreba efektívneho prístupu k nim. Samotný údaj bez popisu obsahu je bezcenný. Aby sa poznal význam jednotlivých údajov, pripájajú sa k nim informácie o ich obsahu, tzv. metadáta. Forma týchto metadátových informácií môže byť rôzna. Tradičnou formou je usporiadanie údajov do tabuliek, kde názvy tabuliek, záhlavia riadkov a stĺpcov tvoria prvotný metadátový pohľad, ktorý sa dopĺňa odkazmi na rôzne metodické materiály a poznámky pod tabuľkami. Pri práci s jednou alebo s malým počtom tabuliek je tento prístup postačujúci a aj pre používateľa údajov vyhovujúci. Komplikácie vznikajú pri distribúcii týchto tabuliek, kde zakaždým treba stanovovať detaily prenosu a výmenné formáty. Aj pri zavedení určitých konvencií v názvosloví a úložisk v prípade veľkého počtu dátových tabuliek (stovky, tisíce až milióny) vznikajú problémy vybrať a pospájať správne údaje z rôznych zdrojov. Ďalším problémom je vznik duplicít údajov, pretože rovnaké údaje majú často rôzne metadátové popisy. Dochádza aj k zámene údajov, lebo rôzne údaje mávajú rovnaké popisy. Pri tradičnom prístupe cez nezávislé tabuľky sa tieto javy ťažko identifikujú, a preto aj často dochádza k omylom.

Tento problém sa rieši vo svete rôznymi postupmi. Spoločným menovateľom týchto postupov je snaha o centralizáciu, zjednotenie a štrukturalizáciu metadátových popisov. Hľadajú sa univerzálne nástroje na prístup a výmenu údajov pomocou metadátového popisu.

- *Centralizácia* metadátového popisu spočíva v sústredení všetkých metadátových informácií na jednom mieste, aby ich používateľ nemusel hľadať v rôznych tabuľkách a iných textových zdrojoch.
- *Zjednotenie* – je proces, kde za zavádza jednotné pomenovanie pre tú istú podstatu s tým, že rôzne informácie nie je možné pomenovať rovnako.
- *Štrukturalizácia* – v tomto procese rôzne časti popisu, ktoré navzájom nesúvisia, sa snažíme navzájom oddeliť na samostatné zložky, aby sa k údajom dalo dostať s postupným výberom jednotlivých zložiek (štruktúrnych informácií) a aby sa dali opakovane použiť ako časť popisu pri rôznych údajoch.
- *Univerzálnosť* – možnosť vytvoriť nástroj, ktorým sa vieme jednoducho dostať ku všetkým údajom iba na základe ich metadátového popisu bez nutnosti použitia informácií o ich zdrojoch (odkiaľ, ako a v akej forme prišli).

## Ukážka určenia metadátových informácií z tabuľky

Na ilustráciu postupov pri určovaní metadátových informácií sme si zvolili údaje vybrané zo štatistického výkazu NBS z roku 2008. Zámerne sme vybrali staršie údaje, lebo umožňujú sústrediť sa viac na ilustráciu informačného modelu ako na analýzu ich hodnôt. Tab. 1 obsahuje iba samotné údaje, bez určenia ich reálneho obsahu. O údaji v označenej bunke vieme povedať iba toľko, že je to číslo s hodnotou 3 596 421. Čo konkrétne toto číslo znamená, je pre nás skryté.

Aby sme pochopili význam tohto čísla, potrebujeme získať informácie o jeho obsahu. V tab. 2 je pôvodná tab. 1 doplnená o záhlavia riadkov a stĺpcov spolu s ďalšími popisnými údajmi. V tejto tabuľke sú vyznačené miesta, kde sme identifikovali možné informácie o obsahu vybranej bunky.

Keďže nie je zrejmé, čo znamená kategória III, okrem uvedených popisov je potrebné použiť aj

*Tabuľka 1 Ukážka číselnej tabuľky bez popisov*

| | | | | |
|---|---|---|---|---|
| 722 534 150 | 703 162 308 | 19 371 842 | 22 787 490 | 100 322 597 |
| 158 357 897 | 154 698 352 | 3 659 545 | 1 389 640 | 25 861 387 |
| 11 065 759 | 10 912 879 | 152 880 | 157 837 | 2 918 355 |
| 1 115 601 801 | 1 096 103 541 | 19 498 260 | 22 827 127 | 100 322 604 |
| 160 524 235 | 156 864 690 | 3 659 545 | 1 389 640 | 25 861 387 |
| 12 446 218 | 12 293 338 | 152 880 | 157 837 | 2 918 944 |
| 309 207 379 | 293 775 895 | 15 431 484 | 9 685 527 | 74 758 882 |
| 129 063 285 | 125 466 864 | 3 596 421 | 1 351 420 | 22 917 363 |
| 9 409 076 | 9 256 254 | 152 822 | 154 713 | 1 931 591 |
| 23 458 198 | 22 385 376 | 1 072 822 | 92 156 | 3 696 567 |
| 10 814 150 | 10 238 507 | 575 643 | 4 737 | 820 981 |
| 447 951 | 447 949 | 2 | 6 | 35 757 |

externý zdroj Metodické vysvetlivky na vypracúvanie výkazu V(NBS) 33 – 12, kde zistíme, že kategóriou III sa rozumie majetok oceňovaný individuálne s identifikovaným znehodnotením. Popis obsahu danej bunky je potom nasledujúci: číslo 3 596 421 vyjadruje objem eurových úverov kategórie III (majetok oceňovaný individuálne s identifikovaným znehodnotením) v bankách za rezidentov zo sektora S.11 (nefinančné spoločnosti) ku koncu mesiaca 9/2008 prepočítaný na tisíce SKK. Následne sa vykoná očistenie týchto metadátových informácií: odstránenia sa nadbytočné informácie, spresní sa význam a typ ostatných informácií a komplexné informácie sa rozdelia na nezávisle zložky.

Z uvedeného príkladu vidieť, že metadátový popis našej bunky obsahuje informácie rôzneho charakteru, sú tu teda informácie rôzneho typu, ktoré medzi sebou viac alebo menej súvisia, a niektoré sú navzájom úplne nezávislé. Cieľom informačného modelu je navrhnúť nástroj, ktorý umožní jednotným spôsobom zapísať metadátové informácie rôzneho typu a rôznych vzájomných súvislostí.

## Popis údajov v systéme SDMX (časť GESMES/TS )

GESMES/TS (*GEneric Statistical MESsage for Time Series*) predstavuje informačný model, ktorý definuje:

- štruktúru a vzájomné súvislosti medzi údajmi – dátový model,
- formát súborov na distribúciu dátového modelu (metadátového popisu),
- formát súborov na výmenu samotných štatistických údajov.

*Tabuľka 2 Ukážka číselnej tabuľky s popismi*

| | | | | | | | V( NBS )33-12 |
|---|---|---|---|---|---|---|---|
| MESAČNÝ VÝKAZ O ÚPLNOM SEKTOROVOM ČLENENÍ ÚVEROV | | | | | | Stav ku dňu: | 30.9.2008 |
| MENOVITÁ HODNOTA | | | | | | Kód banky: | Spolu ( údaje v tis. Sk ) |
| EKONOMICKÉ SEKTORY | MENA | č. r. | ÚVERY SPOLU | | | | |
| | | | Spolu úvery stĺ. 5 až 13 | zo stĺpca 1: | | | Prečerpanie bežného účtu |
| | | | | kategória I a kategória II | kategória III | Zlyhané úvery | |
| a | b | c | 1 | 2 | 3 | 4 | 5 |
| REZIDENTI - SPOLU | | | | | | | |
| Spolu ( r.3 +7+11+15+19 +24 až 26) | SKK | 1s | 722 534 150 | 703 162 308 | 19 371 842 | 22 787 490 | 100 322 597 |
| | EUR | 1e | 158 357 897 | 154 698 352 | 3 659 545 | 1 389 640 | 25 861 387 |
| | OCM | 1x | 11 065 759 | 10 912 879 | 152 880 | 157 837 | 2 918 355 |
| SPOLU (všetky sektory) | SKK | 2s | 1 115 601 801 | 1 096 103 541 | 19 498 260 | 22 827 127 | 100 322 604 |
| | EUR | 2e | 160 524 235 | 156 864 690 | 3 659 545 | 1 389 640 | 25 861 387 |
| | OCM | 2x | 12 446 218 | 12 293 338 | 152 880 | 157 837 | 2 918 944 |
| S.11 Nefinančné spoločnosti | SKK | 3s | 309 207 379 | 293 775 895 | 15 431 484 | 9 685 527 | 74 758 882 |
| | EUR | 3e | 129 063 285 | 125 466 864 | 3 596 421 | 1 351 420 | 22 917 363 |
| | OCM | 3x | 9 409 076 | 9 256 254 | 152 822 | 154 713 | 1 931 591 |
| S.11001 verejné | SKK | 4s | 23 458 198 | 22 385 376 | 1 072 822 | 92 156 | 3 696 567 |
| | EUR | 4e | 10 814 150 | 10 238 507 | 575 643 | 4 737 | 820 981 |
| | OCM | 4x | 447 951 | 447 949 | 2 | 6 | 35 757 |

Gesmes správy sú založené na štandarde EDIFACT (medzinárodný EDI – *Electronic Data Interchange* – štandard vyvinutý OSN), ktorý bol navrhnutý na účely výmeny informácií medzi rôznymi inštitúciami. S nárastom popularity XML formátu na výmenu údajov súvisí SDMX iniciatíva (*Statistical Data and Metadata eXchange*), ktorá okrem iného definovala nový výmenný formát v XML. Takto vzniknutý štandard obsahuje časť pre výmenu údajov v XML formáte označovanú ako SDMX_ML a časť pre výmenu údajov v Gesmes/ts formáte označovanú ako SDMX-EDI.

Podrobnosťami týchto výmenných formátov sa nebudeme zaoberať, ale sústredíme sa na spôsob, ako sú v tomto štandarde popísané údaje. Informačný model je navrhnutý tak, aby umožňoval spracovať časové rady mnohodimenzionálnych číselných polí. Toto predstavuje tzv. hviezdicové usporiadanie údajov, kde v strede je tabuľka faktov obsahujúca samotnú číselnú hodnotu a sadu cudzích kľúčov ukazujúcich na tabuľky dimenzií. Takýto model oddeľuje údaje v tabuľke faktov, ktorá obsahuje merateľné kvantitatívne údaje (merania špecifických javov) od dimenzií, ktoré obsahujú popis (metadáta) týchto údajov.

V ekonomických meraniach má čas výnimočné postavenie. Z časového pohľadu sa v tabuľke faktov vyskytujú tri odlišné typy údajov, a to záznamy udalostí (napr. obchody), stavy (napr. hodnota v nejakom čase) a toky (akumulácia za určité obdobie).

V tomto systéme všetky metadáta (okrem vysvetliviek k meraniam), ktoré sa môžu použiť na popis údajov, sú vopred definované v dátovom modeli. Dátový model definuje oblasti, ktorých sa dané údaje týkajú a ktoré sú ďalej spresnené pomocou štruktúrnych definícií pre danú oblasť, t. j. použité dimenzie a atribúty pre dané údaje. Dimenzie môžu nadobúdať iba povolené hodnoty, ktoré sú definované v zozname kódov.

## Stručné zhrnutie informačného modelu SDMX

1. **Definícia pojmov (popisy).** Aby sme pochopili zmysel niektorých štatistických údajov, potrebujeme poznať pojmy, ktoré sú s nimi spojené. Napríklad samotné číslo 1,2953 je celkom bezvýznamné, ale ak vieme, že to je kurz amerického dolára voči euru z 23. novembra 2006, dostáva zmysel. Preto sú v systéme zavedené definície používaných pojmov.
2. **Zoskupovanie informácií o údajoch.** Pre redukciu objemu prenášaných informácií sa zaviedlo určité zoskupovanie údajov tak, aby sa spoločné informácie zasielali len raz. Toto zoskupovanie nemá na samotný dátový model vplyv, slúži len na optimalizáciu prenosu.
   Podľa toho, kde sú pripojené informácie o štatistických údajoch, rozoznávame tieto úrovne:
   - pozorovanie (platia pre jednotlivé merania nejakého javu),
   - časový rad (informácie spoločné pre všetky merania nejakého javu v pravidelných intervaloch, t. j. časového radu danej periodicity),
   - skupina/sibling (skupinu tvoria všetky časové rady bez ohľadu na periodicitu, s rôznymi frekvenciami),
   - dátová zostava – spoločné informácie sú súčasťou jej definície.
3. **Dimenzie a vlastnosti.** Existujú dva typy informácií: dimenzie, ktoré slúžia na identifikáciu a popis údajov, a vlastnosti, ktoré sú čisto popisné.
4. **Kľúče.** Dimenzie pre každú dátovú zostavu majú definované poradie nazývané rodina kľúčov. Z nich sa zostavujú kľúče umožňujúce identifikáciu údajov. Hodnoty kľúčov (zostave-

*Tabuľka 3 Ukážka definície jednej zostavy údajov: (Key family) ECB_AME1 / AME*

| n | Pripája sa | C/M | Format | Koncept – pojem | Code list – číselník |
|---|---|---|---|---|---|
| 1 | | | AN1 | Frequency | CL_FREQ |
| 2 | | | AN..4 | Ameco reference area | CL_AME_AREA_EE |
| 3 | | | AN1 | Ameco transformation | CL_AME_TRANSFORM |
| 4 | | | AN1 | Ameco aggregation method | CL_AME_AGG_METHOD |
| 5 | | | AN..3 | Ameco unit | CL_AME_UNIT |
| 6 | | | AN..3 | Ameco reference | CL_AME_REFERENCE |
| 7 | | | AN..10 | Ameco item | CL_AME_ITEM |
| 8 | | | AN..35 | Time period or range | |
| 9 | | | AN3 | Time format code | |
| 10 | | | AN..15 | Observation value | |
| 11 | Observation | M | AN1 | Observation status | CL_OBS_STATUS |
| | ... | | | ... | |
| 20 | Time series | C | AN..1050 | Source publication (Euro area only) | |

*Tabuľka 4 Ukážka zoznamu kódov (číselníka): Frequency code list (CL_FREQ)*

| Kód | Popis | Použitie | Vytvorený |
|---|---|---|---|
| A | Annual | ESCB | 9/24/1999 |
| B | Business | ESCB | 9/24/1999 |
| D | Daily | ESCB | 9/24/1999 |
| E | Event (not supported) | ESCB | 9/24/1999 |
| H | Half-yearly | ESCB | 9/24/1999 |
| M | Monthly | ESCB | 9/24/1999 |
| N | Minutely | ESCB | 4/6/2005 |
| Q | Quarterly | ESCB | 9/24/1999 |
| S | Half-yearly, semester (value H exists but change to S in 2009 ...) | ESCB | 10/2/2013 |
| W | Weekly | ESCB | 9/24/1999 |

né z hodnôt jednotlivých dimenzií) tvoria jedinečnú kombináciu pre každú sériu. Konvenčne, frekvencia sa uvádza ako prvá dimenzia v popise danej rodiny kľúčov. Čiastkové kľúče sa používajú pri práci so skupinami.

5. **Zoznamy kódov (číselníky).** Všetky možné hodnoty dimenzie sú definované v zozname kódov. Každá hodnota v tomto zozname má jedinečnú skratku (kód) a popis. Vlastnosti môžu byť niekedy zadávané cez kódy a niekedy ako voľný text. Pretože účelom vlastnosti je len bližšia charakterizácia údajov a nie ich identifikácia, nerobí to žiadny problém.
6. **Štruktúrna definícia.** Štruktúrna definícia (súbory kľúčov) špecifikuje súbor pojmov, ktoré popisujú a identifikujú zostavu údajov. Popisuje, ktoré pojmy sú dimenzie a ktoré sú vlastnosti, stanovuje, na ktorej úrovni sa ktoré pojmy pripájajú k údajom (skupina, séria a pozorovanie) a určuje povinnosť ich použitia. Tiež určuje, ktoré číselníky poskytujú povolené hodnoty pre dimenzie a vlastnosti (pokiaľ sú z číselníkov).

**Informačný model SDMX je vhodný na spracovanie:**
- číselných údajov,
- údajov s pevnou štruktúrou dátových zostáv,
- agregovaných údajov.

**Obmedzenia informačného modelu SDMX:**
- nie je vhodný na výmenu textových informácií (texty sa dajú vymieňať iba ako vlastnosti k nejakému číselnému údaju),
- nie je vhodný na výmenu individuálnych údajov,
- neumožňuje vymieňať kontrolné informácie,
- neobsahuje hierarchické informácie,
- nie je vhodný na statické údaje (ktoré sa nemenia v čase),

## Popis údajov pomocou XBRL

XBRL (*eXtensible Business Reporting Language*) je špecializovaná nadstavba nad XML, ktorá bola vyvinutá na účely výmeny údajov medzi podnikmi, aby bola zabezpečená obsahová čitateľnosť vymieňaných údajov. XBRL špecifikácia zavádza dva druhy dokumentov, a to taxonómiu a inštančné dokumenty. Taxonómiu tvorí štandardizovaná XML schéma, v ktorej sa definujú pojmy (*concepts*) a vzťahy medzi nimi. Taxonómia takto určuje výslednú štruktúru výkazov. Skutočne vykazované hodnoty, v XBRL nazývané fakty, sa nachádzajú v inštančných dokumentoch. Inštančný dokument je množina faktov, ktoré spĺňajú reštrikcie stanovené zavedenými pojmami danej taxonómie.

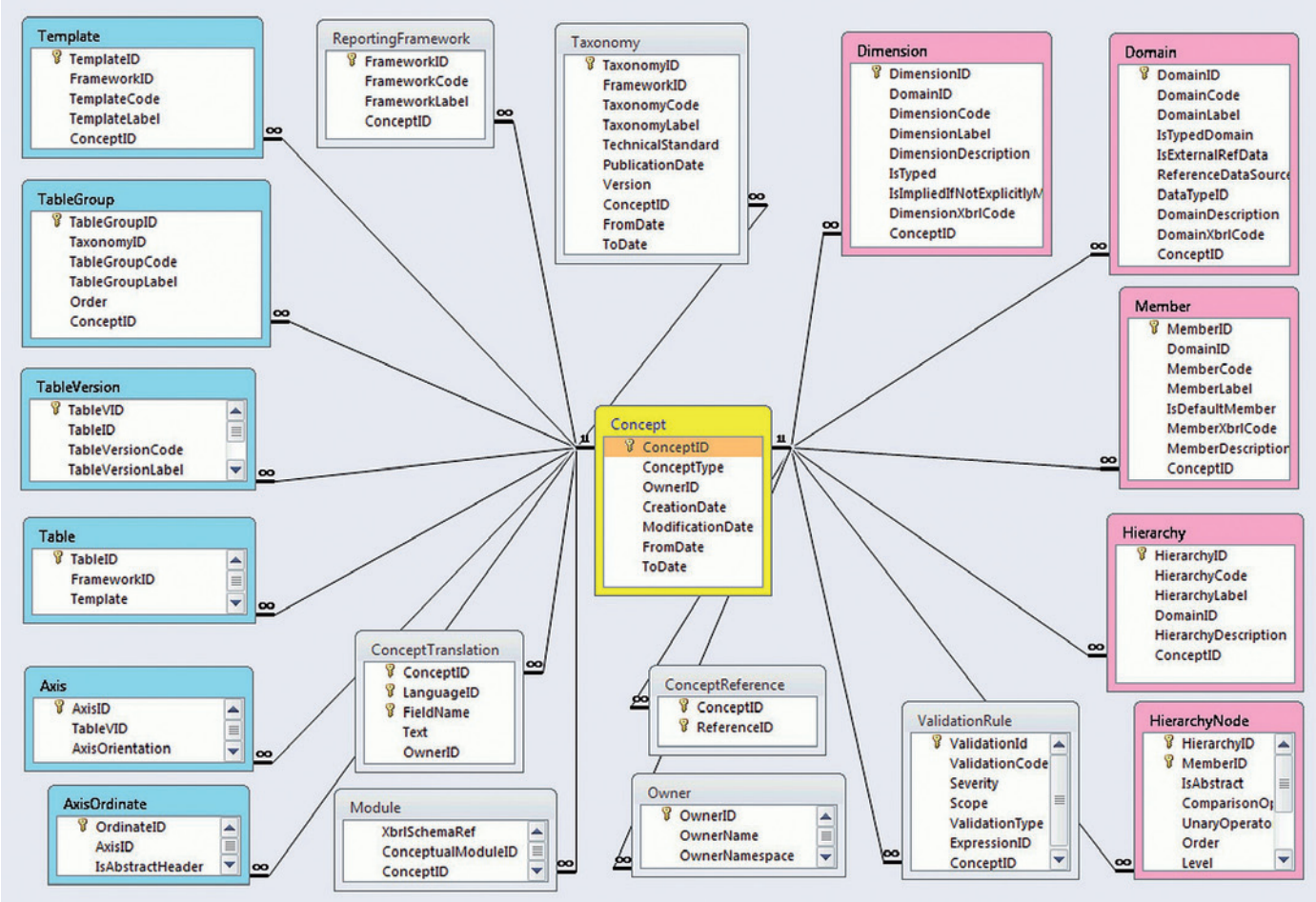
Taxonómiu môže tvoriť jeden alebo viacero dokumentov. Vždy sa začína definovaním pojmov, kde sa každému pojmu pridelí jeho názov, ID a typ. Následne sa definujú detaily týchto pojmov, vzťahy medzi nimi a väzby na vonkajšie zdroje. My sa budeme zaoberať taxonómiou EBA, ktorú tvoria stovky XML dokumentov. Z tejto taxonómie sa sústredíme iba na informačný model.

V systéme Gesmes sa zaviedli ekonomické pojmy, ktoré sa využívali iba na definovanie dimenzií a vlastností. V XBRL musí byť každý popisný prvok definovaný ako samostatný pojem. Teda nielen dimenzie a vlastnosti, ale napr. aj číselníky a ich jednotlivé prvky. V XBRL taxonómií nie je špeciálna kategória pre pojmy. Tie sú vyjadrené pomocou definícii tagov. Nasledujúci obrázok ilustruje na taxonómií EBA, kde všade sa definované pojmy využívajú. Modrou farbou je na ňom zvýraznená časť, ktorá sa týka grafickej prezentácie, a ružovou farbou časť, ktorá popisuje informačný model.

## Stručné zhrnutie informačného modelu XBRL

1. **Definícia pojmov (popisy).** Slovník pojmov v tomto modeli obsahuje primárne pojmy pre metriku, dimenzie, domény (číselníky) a prvky domén. Sekundárne pojmy sú pre rodiny a perspektívy. Každý pojem v slovníku je verejným prvkom v taxonómii. Pomocou týchto pojmov sa popisujú údaje (data point).
2. **Metrika (metric).** Časť povinná pre každý údaj, definuje povahu merania. Obsahuje význam meraných údajov, typ meraných údajov a časovú charakteristiku voči periodicite (či ide o okamžitú hodnotu alebo o intervalový údaj).

*Zobrazenie postavenia pojmov v EBA taxonómii*




3. **Dimenzie.** Dimenzie ďalej, resp. bližšie špecifikujú význam danej metriky. Hodnoty priradené jednotlivým dimenziám sú ich členovia. Členovia dimenzií sa zoskupujú v doménach. Každá dimenzia môže mať prvky iba z jednej domény. Na popis jednotlivých údajov môže byť použitý rôzny počet dimenzií. Na popis údaja možno použiť každú dimenziu iba raz. Dimenzie majú definované štandardné, resp. predvolené hodnoty, ktoré sa použijú, ak nie je dimenzia explicitne uvedená v popise údaja.
4. **Domény (číselníky).** Domény sú zoskupenia prvkov (členov), ktoré majú nejaké spoločné vlastnosti. Jedna a tá istá doména môže byť použitá vo viacerých dimenziách.
5. **Prvky domén.** Prvky môžu byť dvoch typov a to také, ktoré v doméne tvoria explicitný zoznam a také, ktoré sa zadávajú ako text vyhovujúci určitým podmienkam (napr. kód ISIN).
6. **Hierarchie.** Predstavujú zoskupenia prvkov explicitnej domény, ktoré majú hierarchické vlastnosti. Tieto hierarchie môžu tvoriť:
- podmnožiny členov domény (subdomény),
- schematické usporiadanie prvkov na prezentačné účely.
- základné aritmetické vzťahy medzi prvkami, či nadradený prvok hierarchie je (=, ≤ alebo ≥) vo vzťahu k podriadeným prvkom. Pri podriadených prvkoch určujeme, či budú pripočítane alebo odpočítané pri agregácii.
7. **Rodiny.** Rodiny sú skupiny dimenzií vytvárané na prezentačné účely.
8. **Perspektívy.** Perspektívy predstavujú iné kritériá pre zoskupovanie dimenzií. Tu sa zoskupujú podľa účelu, na ktorý sa budú používať (napr. na štatistické účely, na finančné účely a pod.).

### Informačný model XBRL je vhodný na spracovanie:

- číselných aj textových údajov,
- údajov s premenlivým počtom popisných dimenzií,
- individuálnych a agregovaných údajov.

### Obmedzenia informačného modelu XBRL:

- generuje pomerne rozsiahle súbory,
- automatizácia silne závisí od úrovne prípravy taxonómie,
- na spracovanie potrebuje nákladný procesor,
- definovaný je iba jednoduchý systém kontrol,
- nie je zaručená časová stálosť taxonómií.

*Pokračovanie v č. 9/2015.*