Název projektu:
Redesign Statistického informačního systému v návaznosti na zavádění eGovernmentu v ČR
Příjemce: Česká republika – Český statistický úřad
Registrační číslo projektu: CZ.1.06/1.1.00/07.06396

Příloha k zadávací dokumentaci veřejné zakázky
„Integrační nástroje, vstupní a výstupní subsystém“

Příloha č. 54

Specifikace hromadné aktualizace SMS-KLAS

Název souboru: RSIS_ZD001P54_KLAS_HROMAD_AKTUALIZACE.pdf
Počet stran přílohy (bez tohoto krycího listu): 8
Administrace přílohy: Ing. Ebbo Petrikovits

Verze ke zveřejnění

Hromadná aktualizace KLAS

(Specifikace řešení)
Vytvořeno: 29. 5. 2008
Poslední aktualizace: 29. 1. 2009

Anotace: *Materiál popisuje návrh řešení hromadné aktualizace obsahu číselníků a vazeb v KLAS externími datovými soubory. Materiál vychází ze zadání [1] a upřesnění na jednání týmu KLAS 7.5.2008.*

1. Požadavky

Požadavky hromadné aktualizace daty z externích datových souborů zahrnují:

- načtení dat z externího datového souboru;
- uživatelskou volbu mapování dat zdrojového souboru na strukturu zvoleného cílového číselníku;
- hromadnou aktualizaci položek číselníků obsahem souboru;
- hromadnou aktualizaci vazeb zvoleného číselníku obsahem souboru;
- data zpracovávat ve vstupním editačním rozhraní;
- na načítaná data uplatnit plné kontroly KLAS, jako by byly položky editovány ručně;

1.1 Požadovaný formát zdrojových dat

Požadováno je zpracování dat ve strukturované podobě ve formátu XML. Původní požadavek obsahoval i možnost načítání dat ve formátu xls (list aplikace Excel), který ale není dostatečně standardizován a popsán. Jeho zpracování je proto problematické a zadavatel od jeho zpracování upustil.

2. Návrh řešení

2.1 Základní charakteristika

Hromadná aktualizace číselníků KLAS je podle zadání určena k ovlivnění obsahu položek číselníků a jejich vazeb. Aplikace nebude schopná ovlivňovat obsah dat v katalogu KLAS. Všechny katalogové položky, které hromadná aktualizace bude využívat, musí proto být předem v katalogu KLAS v požadovaném stavu.

Řešení hromadné aktualizace obsahu číselníků a vazeb v KLAS je založeno na standardních principech hromadné (dávkové) aktualizace KLAS. Předpokládá proto založení nové kategorie aktualizačních úloh KLAS – Hromadná aktualizace z externího souboru. V této kategorii vzniknou postupně úlohy:

- zpracování dat ze souboru ve formátu univerzálního rozhraní KLAS;
- zpracování dat z pevně definovaných souborů;
- zpracování dat z obecného vstupního souboru;

V rámci takto založených úloh bude probíhat běžný mechanismus hromadné aktualizace, tedy:

- založení nového běhu s výběrem datového souboru ;
- dávkové zpracování datového souboru;
- uložení chybných vět do chybových tabulek;
- prohlížení a případně oprava vět v chybových tabulkách;

2.2 Struktura univerzálního rozhraní KLAS (UR_KLAS)

Struktura univerzálního rozhraní KLAS je řešena jako XML soubor s definovaným schématem. Struktura souboru umožňuje uložení všech dat potřebných k hromadné správě obsahu jednoho

číselníku v KLAS. V jednom souboru jsou položky pouze jednoho číselníku a pro každou položku číselníku je možné uložit data v souladu s datovým modelem KLAS. Ke každé položce tak v souboru lze uložit:

- **Kód položky**
  (znakový i numerický)
- **Název položky**
  (zkrácený i plný včetně historického vývoje a jazykových mutací);
- **Ostatní texty položky**
  (ostatní typy textů položky číselníku včetně historického vývoje a jazykových mutací)
- **Atributy položky**
  (znakové, numerické i textové včetně historického vývoje a jazykových mutací);
- **Texty atributů položky**
  (odlišené typem textu včetně historického vývoje a jazykových mutací);
- **Vazby položky**
  (položky vázaných číselníků včetně historického vývoje);

Struktura univerzálního rozhraní KLAS je popsána v popisu schématu XML v [2]. Graficky je zobrazena v dokumentu [4].

2.3 Zpracování dat ze souboru ve formátu univerzálního rozhraní KLAS

Formát UR_KLAS představuje jediný vstupní formát pro hromadnou aktualizaci KLAS. Další varianty aktualizace (viz. 2.4) jsou tvořeny jako nadstavba nad UR_KLAS a využívají datové transformace souboru do formátu UR_KLAS.

2.3.1 Postup zpracování

Struktura UR_KLAS umožňuje následující činnosti nad číselníky KLAS:

- **Naplnění obsahu číselníku:**
  založení položek zvoleného číselníku včetně všech textů, atributů a vazeb v souladu s definicí katalogové položky zvoleného číselníku;
- **Aktualizace obsahu číselníku:**
  historická editace všech textů položek číselníků, založení a editace atributů číselníků a jejich textů, založení a editace položek vazeb;
- **Ukončení/prodloužení platnosti položek číselníků:**
  ukončení platnosti položky a všech podřízených entit, prodloužení platnosti položky číselníku a všech podřízených entit, jejichž konec odpovídal původnímu konci platnosti položky;
- **Ukončení platnosti textů, atributů a vazeb:**
  ukončení platnosti textů (kromě názvu a zkráceného názvu položky), ukončení platnosti atributů a jejich podřízených entit, ukončení platnosti položek vazeb;
- **Zrušení (smazání) položek číselníků:**
  označení položky číselníku všech podřízených entit příznakem ZRUŠENO;
- **Zrušení (smazání) textů, atributů a vazeb:**
  označení položky textu, atributu nebo vazby příznakem ZRUŠENO;

Soubor ve formátu UR_KLAS je v prvním kroku načítán do vstupních databázových struktur KLAS. Současně s tímto nahráváním je validován proti definičnímu schématu UR_KLAS [2]. Pokud vstupní soubor tomuto schématu nevyhoví byť jen v jediné položce, je soubor odmítnut jako celek a označen jako nepoužitelný. Žádná data z tohoto souboru nejsou do databáze KLAS načtena, a to ani do tabulek vstupního rozhraní hromadné aktualizace. Takový běh není možné opakovat.

Všechny aktualizační věty uložené v UR_KLAS jsou při běhu úlohy hromadné aktualizace zpracovány vůči aktuálnímu obsahu editovaného číselníku.

Hromadná aktualizace vždy pouze doplňuje nové informace do číselníku KLAS. Všechny položky cílového číselníku, které se nevyskytují v datovém souboru, proto zůstávají hromadnou aktualizací cílového číselníku nedotčené.

Aplikace zpracovává obsah UR_KLAS tak, že pro všechny věty vstupního souboru zajistí promítnutí všech požadovaných změn do obsahu číselníku. V případě, že aktualizace některé položky neproběhne, resp. případná editace by způsobila kolizi s integritními pravidly číselníků KLAS, je tato věta vstupního souboru označena jako chybná a přenesena do chybové tabulky pro další zpracování uživatelem. V tom případě je cílová položky číselníku ponechána nedotčena.

2.3.2 Obsah souboru

Struktura UR_KLAS je vytvořena do značné míry variabilní, aby nebylo nutné vytvářet zbytečně složité soubory pro pouze dílčí aktualizace. Minimální přípustný obsah struktury je:

- **Naplnění obsahu číselníku:**
  pro celý soubor: Kód číselníku, Platí Od, Neplatí Po, Kód jazyka;
  pro každou položku: Znaková hodnota, Zkrácený název položky, Plný název položky
- **Aktualizace obsahu číselníku:**
  pro celý soubor: Kód číselníku, Platí Od, Neplatí Po, Kód jazyka;
  pro každou položku: Znaková hodnota položky, identifikace a hodnota editované entity (text, atribut, text atributu, vazba);
- **Ukončení platnosti položek číselníku**:
  pro celý soubor: Kód číselníku, Platí Od, Neplatí Po, Kód jazyka;
  pro každou položku: Znaková hodnota položky;
- **Zrušení položek číselníku**:
  pro celý soubor: Kód číselníku;
  pro každou položku: Znaková hodnota položky;
- **Zrušení položek vazby nebo atributu**:
  pro celý soubor: Kód číselníku;
  pro každou položku: Identifikace atributu, resp. identifikace vazby;

Informace uvedené pro celý soubor není nutné opakovat v jednotlivých položkách xml souboru. Pokud jsou uvedené u položek, jsou nadřazené hodnotě platné pro celý soubor.

2.3.3 Interpretace dat souboru

Data externího souboru jsou při aktualizaci KLAS interpretována s ohledem na obsah vstupních polí jednotlivých položek (platí, že hodnoty neuvedené v konkrétním prvku xml dokumentu jsou nahrazeny hodnotou z nadřízené úrovně, pokud je definována):

2.3.3.1 Rušení (mazání) záznamů

Rušení záznamů probíhá pouze pokud je nalezena položka identifikované entity. Pokud není nalezena, je věta vstupního souboru označena za chybnou a žádná aktualizace neprobíhá.

- **Identifikace položky bez uvedení platnosti**
  (Kód číselníku, Znaková hodnota položky):
  zrušení (smazání) obsahu položky a všech podřízených entit
- **Identifikace položky a atributu bez uvedení platnosti**
  (Kód číselníku, Znaková hodnota položky, Kód atributu):
  zrušení (smazání) výskytu atributu v položce a všech podřízených entit;
- **Identifikace položky a vazby bez uvedení platnosti**
  (Kód číselníku, Znaková hodnota položky, Kód vazby):
  zrušení (smazání) výskytu vazby v položce a všech podřízených entit;
- **Identifikace položky a typu textu bez uvedení platnosti**
  (Kód číselníku, Znaková hodnota položky, Typ textu):
  zrušení (smazání) výskytu typu textu v položce;

2.3.3.2 Změna platnosti (zkracování i prodlužování)

Změna platnosti probíhá pouze pokud je nalezena položka identifikované entity. Pokud není nalezena, je věta označena za chybnou a žádná aktualizace neprobíhá.

- **Identifikace položky s uvedením platnosti**
  (Kód číselníku, Znaková hodnota položky, Platí od, Neplatí Po):
  - úprava platnosti položky na požadovaný interval;
  - úprava platností všech podřízených entit, prodloužení pouze pokud jejich původní platnost byla shodná s platností položky;
- **Identifikace položky a atributu s uvedením platnosti**
  (Kód číselníku, Znaková hodnota položky, Kód atributu, Platí od, Neplatí Po):
  - úprava platnosti atributu na požadovaný interval
  - úprava platností všech podřízených entit, prodloužení pouze pokud jejich původní platnost byla shodná s platností atributu;
- **Identifikace položky a vazby s uvedením platnosti**
  (Kód číselníku, Znaková hodnota položky, Kód vazby, Platí od, Neplatí Po):
  - úprava platnosti vazby na požadovaný interval
  - úprava platností všech podřízených entit, prodloužení pouze pokud jejich původní platnost byla shodná s platností vazby;
- **Identifikace položky a typu textu s uvedením platnosti**
  (Kód číselníku, Znaková hodnota položky, Typ textu, Platí od, Neplatí Po):
  - úprava platnosti textu na požadovaný interval;

2.3.3.3 Aktualizace a vložení hodnoty

Aktualizace hodnoty znamená, že pokud byla nalezena položka identifikované entity, je hodnota ze vstupního souboru vložena do příslušné entity. Pokud v době platnosti nově vkládané hodnoty existuje původní hodnota, je platnost původní hodnoty upravena tak, aby novou hodnotu bylo možné vložit. Pokud mají původní i nová hodnota shodný interval platnosti, je původní hodnota napřed zrušena a následně je vložena hodnota nová. Pokud není položka identifikované entity nalezena, je hodnota ze souboru vložena jako nová.

2.4 Zpracování dat z pevně definovaných souborů

Existuje množina datových souborů (například z externích zdrojů), jejichž struktura je pevná a správci číselníků neovlivnitelná. Pokud je těmito soubory KLAS aktualizován pravidelně a navíc se jedná o rozsáhlejší data, bude pro každý takový soubor vytvořena speciální úloha hromadné aktualizace. Hromadná aktualizace KLAS počítá se zdrojovým souborem ve formátu xml případně dbf. Pokud požadovaný datový zdroj ve formátu xml nebo dbf není, je nutné ho napřed do formátu xml nebo dbf přeměnit, například pomocí nástrojů MS Office – Excel.

2.4.1 Hromadná aktualizace ze souboru ve formátu XML

Definice úlohy dovolí vybrat jako datový zdroj soubor ve formátu xml v předem známé struktuře, odlišné od UR_KLAS. Tento soubor zpravidla nebude mít definované schéma, proto ho nebude možné již na vstupu validovat na správnost struktury a obsahu. Po výběru datového souboru provede aplikace hromadné aktualizace KLAS konverzi zdrojového souboru do formátu UR_KLAS. K tomu využije technologie XML transformace (XSLT). Výsledný datový soubor ve formátu UR_KLAS vznikne automaticky na dočasném úložišti a uživateli nebude k dispozici. Výsledný soubor bude dále zpracován jako standardní datový vstup, viz. 2.3. V rámci tohoto standardního zpracování bude také validován proti schématu UR_KLAS. Pokud tedy původní zdrojový soubor obsahoval neočekávaná nebo jinak neplatná data, může se stát, že po transformaci do UR_KLAS nebude výsledný soubor validní pro načtení do vstupních struktur. V takovém případě bude celá aktualizace ukončena a soubor označen za nepoužitelný, včetně informace o důvodu odmítnutí.

2.4.2 Hromadná aktualizace ze souboru dBase (DBF)

Soubory formátu dBase (DBF) mají řadu významných omezení, jež brání jejich užití jako plnohodnotný zdroj pro hromadnou aktualizaci KLAS. Jedná se především o:

- omezení délky textového řetězce na 254 znaků;
- nemožnost vytváření strukturovaných záznamů;

Vzhledem k omezení na délku textu je možné soubory DBF použít především jako zdroj aktualizací směřujících ke:

- změně platnosti položek
  (jak číselníků, tak vazeb a atributů);
- rušení (mazání) položek
  (jak číselníků, tak vazeb a atributů);
- vkládání a editaci hodnot atributů a vazeb;

Omezení na „plochou“ strukturu DBF souboru vede k omezení rozsahu spravovaných položek (atributů a vazeb).

2.4.2.1 Zjednodušené univerzální rozhraní KLAS

Pro soubory DBF bude definováno zjednodušené rozhraní, omezené dále uvedenou pevnou strukturou DBF souboru.

Atribut	**Rozměr**	**Význam**
KodCis	Number (5,0)	Kód číselníku
CHodnota	Char(15)	Znaková hodnota položky
Hodnota	Number(15,0)	Numerická hodnota položky
Text	Char(254)	Název položky
ZkrText	Char(60)	Zkrácený název položky
PlatiOd	Date	Počátek platnosti (statistické)
AdmPlatiOd	Date	Počátek administrativní platnosti
NeplatiPo	Date	Konec statistické platnosti
AdmNeplati	Date	Konec administrativní platnosti
KodAtr	Number(5,0)	Kód atributu
CAtrib	Char(15)	Znaková hodnota atributu
NAtrib	Number(15,0)	Numerická hodnota atributu
TAtrib	Char(254)	Textová hodnota pojmenovaného textového atributu
Kodvaz	Number(5)	Kód vazby
VazCHod	Char(15)	Hodnota položky vázaného číselníku

Soubor obsahuje data pouze jedné jazykové mutace (její volba proběhne před spuštěním úlohy načtení souboru. Po načtení datového souboru budou položky transformovány do standardního rozhraní UR_KLAS, takže výsledná úloha hromadné aktualizace KLAS bude shodná se standardním zpracováním (viz 2.3).

Úloha zpracování této struktury bude do KLAS implementována přednostně, aby bylo možné co nejdříve zahájit hromadnou aktualizaci položek číselníků v KLAS.

V případě nepravidelné aktualizace KLAS daty souboru mimo množinu předdefinovaných formátů souborů, bude nutné provést transformaci takového souboru do struktury UR_KLAS jinými nástroji mimo aplikaci KLAS.

2.4.3 Ruční zadání parametrů běhu úlohy

Pro zpracování vstupního souboru bude možné zvolit následující společné atributy celé úlohy:

- Kód číselníku
- Administrativní platnost
- Statistická platnost
- Jazyková mutace

Tyto statické informace nemusí být obsaženy v datovém souboru a mohou být zadány jako společné informace pro celý soubor. V případě jejich zadání a současné existence hodnot v datovém souboru, má přednost hodnota nesená v datovém souboru.

2.5 Zpracování dat z obecného vstupního souboru

Zpracování dat z obecného vstupního souboru znamená identifikovat strukturu tohoto souboru a namapovat ji na strukturu UR_KLAS. Uživatel tedy vybere požadovaný datový soubor a aplikace mu zobrazí informace o datové struktuře souboru. K tomu lze využít pouze takové soubory, které mají svou strukturu čitelnou buď z hlavičkových údajů (DBF), nebo z popisného souboru (schéma XML) případně textové soubory s oddělovačem nebo pevnou délkou po zobrazení prvních několika řádek.

Zobrazenou informaci o struktuře vstupního souboru musí uživatel jednoznačně spojit s prvky UR_KLAS. Vstupní modul hromadné aktualizace KLAS podle tohoto spojení načte zvolený soubor do struktury UR_KLAS. Další zpracování již probíhá standardním postupem (viz 2.3).

Vzhledem k obecnosti UR_KLAS je nutné, aby v případě, kdy externí soubor bude typu XML, bylo propojení elementů souboru na prvky UR_KLAS dodržovalo hierarchii UR_KLAS i vstupního dokumentu.

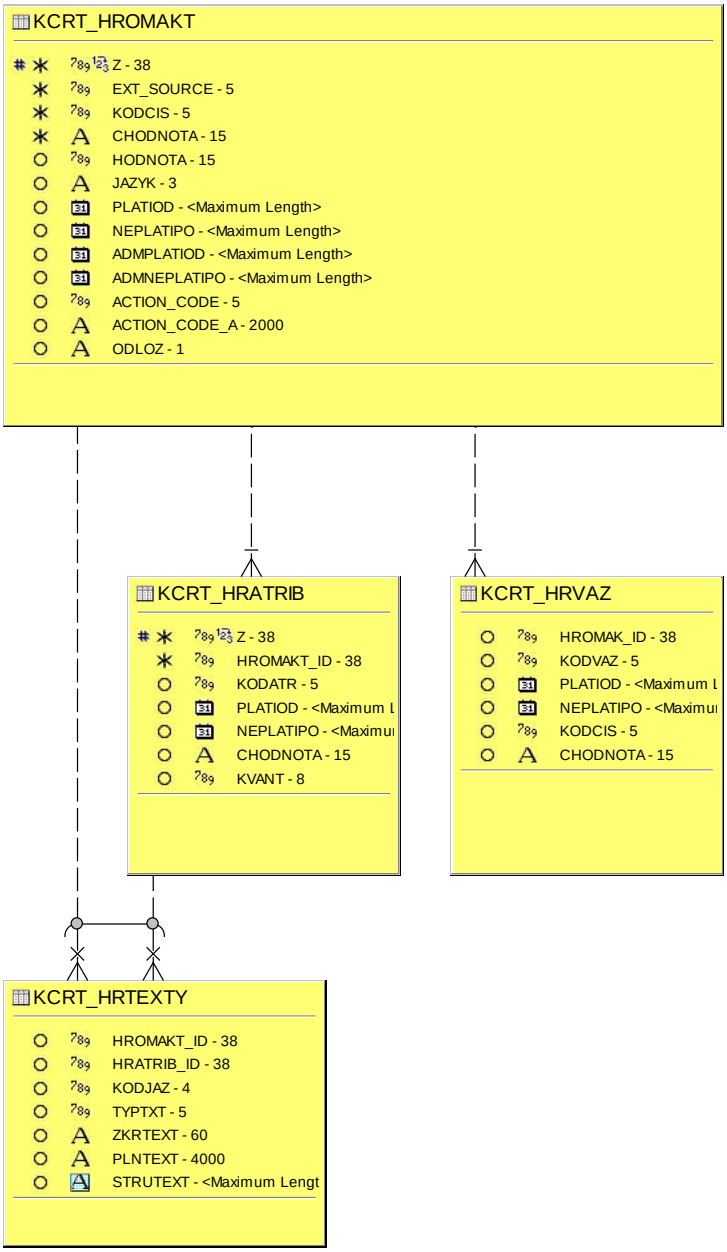
Pro soubory s plochou strukturou (DBF i CSV) budou možnosti propojení elementů předem omezeny tak, aby jedna řádka vstupního souboru nemohla generovat více záznamů UR_KLAS.

Pro zpracování obecného vstupního souboru bude možné zvolit společné atributy celé úlohy (viz 2.4.3).

3. Dopady do aplikace

3.1 Datový model

Datový model UR_KLAS je tvořen tabulkami dle následujícího schématu:



Detailní popis struktury tabulek je v [3].

3.2 Zpracování dat aplikací

Zpracování bude dostupné standardními prostředky dávkové aktualizace KLAS pomocí menu Aktualizace | Hromadná a následně volbou aktualizační úlohy.

3.2.1 Zpracování standardních XML souborů

XML soubory ve struktuře UR_KLAS nebo v pevných strukturách podporovaných aplikací (viz 2.4.1) bude možné zpracovávat on-line (z klientského PC) i autonomně, z dat uložených na databázovém

serveru (viz plnění z METIS). Soubory uvedených struktur budou načteny bez dalších doplňujících dialogů a postoupeny ke zpracování.

3.2.2 Zpracování dat souboru obecného formátu

Zpracování dat v obecném formátu bude možné pouze z klientského pracoviště, protože takové soubory vyžadují interakci uživatele, které při autonomním zpracování nelze poskytnout.

V rámci vytvoření nového běhu úlohy uživatel běžně vybere soubor. Zvolený soubor (jeho hlavička) bude načten aplikací a uživateli bude zobrazena informace o struktuře souboru. Následně bude uživatel nucen jednoznačně namapovat seznam přípustných prvků pro aktualizaci s jednotlivými prvky aktualizačního souboru. Prvky, které mohou být společné pro celý běh úlohy (viz 2.4.3) bude možné zadat ručně bez nutnosti propojení na prvky vstupního souboru. Bez tohoto propojení nemůže aktualizační funkce správně fungovat. Rozsah propojených prvků může ovlivnit dopady promítnutí zvoleného souboru do KLAS (viz 2.3.3).

3.2.3 Editace chybných dat a opakované běhy

Bez ohledu na strukturu souboru končí všechna data všech úloh hromadné aktualizace KLAS v jednotné struktuře rozhraňových, a tedy i chybových tabulek. Proto je možné vytvořit společný editor chybových tabulek úloh hromadné aktualizace.

Editace chybových tabulek umožňuje uživateli změnit hodnoty, které budou zpracovány úlohou tak, aby při opakovaném běhu nevznikala chyba a věta se do KLAS promítla.

4. Další využití formátu UR_KLAS

Formát UR_KLAS je možné využít nejen jako datový zdroj pro hromadnou aktualizaci KLAS, ale také jako standardizovaný exportní formát libovolného číselníku z KLAS. Tento formát by bylo možné zapojit do exportů z KLAS, do menu standardní exporty. Dále je tento exportní formát možné zapojit do internetové prezentace KLAS jako standardní formát pro ukládání číselníků.

Výhodou tohoto řešení je možnost exportovat data z KLAS ve stejném formátu, jaký umí zpracovat aplikace hromadné aktualizace KLAS. Pokud by exportní formát vyžadoval nějaké úpravy datového schématu, bude nutné tyto změny doplnit i do popisu schématu pro vstup do hromadné aktualizace KLAS.

5. Literatura

[1] KLAS-hromadná aktualizace-20080429.doc, ČSÚ, 28.4.2008

[2] KC_HromAkt.xsd, BIOS, 26.01.2009

[3] KLAS_HROMAKT_tabulky.pdf, BIOS, 29.1.2009

[4] KC_hromakt_xsd.pdf, BIOS, 29.01.2009

Vypracoval: Jan Olšanský