

Fragmentace obsahu edic plných textů a možnosti využití crowdsourcingu při tvorbě a strukturaci obsahu plných textů v oblasti historických fondů

Úvod

O dokumentu

Tento dokument navazuje na dokument *Principy komunikace při poskytování plných textů dokumentů*

(http://www.manuscriptorium.com/sites/default/files/docs/VaV_2016/fragmented_fulltext_access_cs.pdf), který se zabývá problematikou tvorby plných textů. V něm jsou popsány problémy dokumentově orientovaného modelu a zároveň načrtnuto jejich řešení pomocí aplikace vztahově orientovaného modelu.

Pozornosti doporučujeme zejména kapitolu *Příklady situací, které je obtížné řešit v dokumentově orientovaném systému* a navazující *Výhody navrženého řešení*.

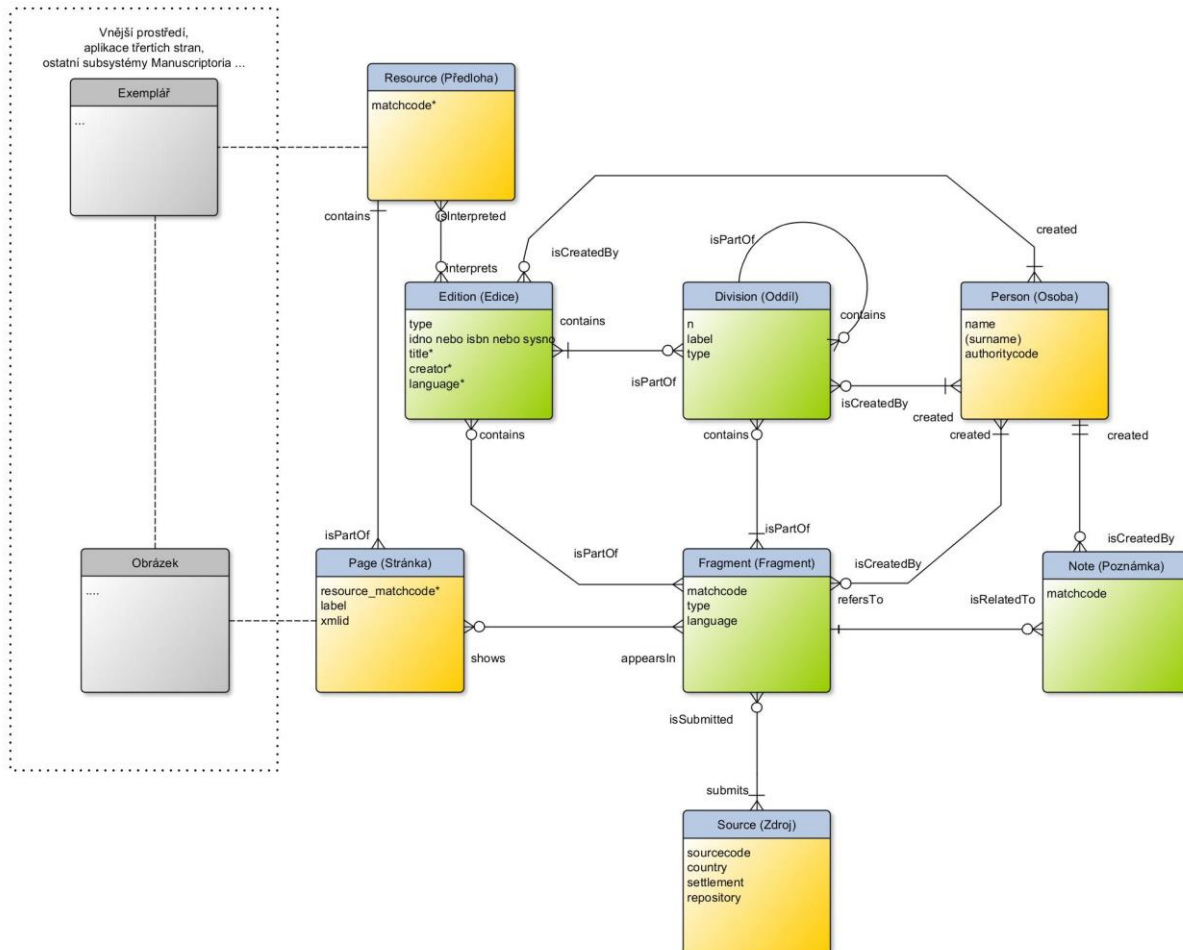
Tento článek popisuje způsob konkrétní implementace grafové (vztahové) databáze plných textů s využitím principů fragmentace obsahu edic plných textů a navazuje pojednáním o možnosti využití crowdsourcingu při tvorbě a strukturaci obsahu plných textů v oblasti historických fondů, kterou vyvíjený systém umožní.

Terminologie

V tomto dokumentu používáme nově termín *fragment* místo původního, poněkud zavádějícího termínu *odstavec*. *Fragmentem* textu pak může být odstavec, řádek textu nebo jiná smysluplná jednotka (závisí na povaze a struktuře textu, samotný typ fragmentu je indikován TEI P5 kódem, který je hodnotou příslušné vlastnosti fragmentu).

Databázový model

Naše implementace databáze plných textů pracuje se strukturou, kterou dokumentuje následující diagram.



Obr. 1 Model databáze plných textů v režimu poloprovozu

Edice

Klíčovým prvkem návrhu je entita *Edice*. Tato entita reprezentuje jednu konkrétní skutečnou textovou edici a k identifikaci edice slouží v našem návrhu trojkombinace informací Titul, Autor/Autoři, typ edice.¹

Skutečnou edici tvoří skupina *fragmentů*, které vytvořil stejný autor v rámci textu jednoho typu edice (překlad, transliterace či transkripce a podobně – výčet typů edic není konečný a volba typu edice je zcela v režii autora). Příslušnost k *edici* je tedy u *fragmentů* vkládaných do systému určena právě uvedením této trojkombinace informací.

¹ Systém samozrejme prideliuje vsem entitám systémové ID, to je ale dosud používané výhradne pro vnútorné potreby systému. Vrchná vrstva aplikácie sa systémovým ID nijak nepracuje.

Edice je oproti původním konvenčním řešením zcela oproštěna od vazby na existující fyzickou *Předlohu*. Vazbu na fyzickou *Předlohu* přes její *Stránky*² je možné do systému vkládat v libovolném čase. To mimo jiné znamená, že v systému mohou vznikat edice plných textů k fyzickým dokumentům, které dosud nebyly digitalizovány, zveřejněny, katalogizovány apod. Dokonce mohou vznikat edice, ke kterým fyzická předloha zcela chybí (neexistuje). Řešení je tedy velmi flexibilní a umožňuje koexistenci přepisů historických dokumentů s moderními odbornými plnými texty, což považujeme za výraznou výhodu v oblasti agregace informací (viz primární cíle projektu Manuscriptorium), ze které budou těžit všechny cílové skupiny uživatelů (všechny existující relevantní informace budou soustředěny bezešvě na jednom místě).

Předloha

Pokud je předloha známá, pak může být identifikována pomocí informací o uložení exempláře platného v čase *t*. Díky *Indexu digitalizovaných historických fondů s Digitálními konkordancemi*³ může být tento identifikátor fyzického exempláře považován za persistentní. Systém tento identifikátor dopočítává automaticky na základě dostupných informací o exempláři.

Předlohou však mohou teoreticky být i moderní dokumenty (například pro moderní odborné texty), u kterých z principu nemá smysl aplikovat exemplářově orientovaný způsob identifikace předlohy. Proto předloha může být identifikována více způsoby:

- kombinací informací pro výpočet identifikátoru fyzického exempláře,
- ISBN,
- jiným systémovým číslem.

Součástí identifikační informace je pak i uvedení typu identifikátoru.

Oddíl

Implementací Oddílů jsme umožnili strukturovat obsah edice do logických obsahových jednotek. Je pak na uvážení autora, zda a jak tuto entitu využije.

Fragment

Jak bylo popsáno v předešlých dokumentech, *Fragment* je nejmenší logickou jednotku intelektuálního obsahu textu.

Protože jsme nechtěli vnucovat autorům textů vlastní náhled na to, co je nejmenší intelektuální jednotkou obsahu textu, předpokládali jsme zavedení vazby *Fragment obsahuje nula nebo více Fragmentů*.

Tento přístup k modelování obsahu je již poměrně divoký a ze zkušeností s implementací aktuálního řešení nad texty více než 300 různých edic plných textů vychází náš návrh, který říká: autor rozhoduje, co je nejmenší logickou jednotkou a tedy Fragmentem. Je tedy na autorovi, zda Fragment reprezentuje odstavec, řádku, nebo v extrémním případě slovo nebo písmeno. Domníváme se, že autor zvolí vhodnou fragmentaci s ohledem na povahu plného textu a efektivitu vlastní práce.

² Je otázkou, zda v modelu umožnit vazbu mezi *Fragmentem* a *Předlohu* přímo, bez vazby vedoucí přes *Stránky*. To patrně ukážou zkušenosti z poloprovozu a detailní poznání praxe dalších původců/autorů plných textů.

³ Oba systémy jsou výstupem předchozích výzkumných aktivit (2016) a v době zveřejnění článku (listopad 2017) jsou provozovány v režimu poloprovozu. Začátkem roku 2018 budou převedeny do plného provozu.

Součástí vlastností *Fragmentu* je pak informace o typu fragmentu s konečný výčet typů nehodláme zavádět.

Tento flexibilní přístup k fragmentaci spolu s možností organizovat *Fragmenty* do *Oddílů* je podle našeho názoru vstřícný k autorům a zároveň činí obsah přehledným pro správce agregovaného obsahu a přehledným pro cílové koncové uživatele.

Poznámka

Implementovaný model umožňuje vkládat poznámky dvěma způsoby:

- 1) Jako autorkou poznámku, která je přímo součástí markupu TEI P5 v hodnotě vlastnosti *body* konkrétního *Fragmentu*, a tedy neexistuje v systému jako samostatná entita,
- 2) Jako samostatnou entitu *Poznámka*, jejímž autorem je libovolný uživatel (nejen autor původní *Edice/Fragmentu*), kterou autor poznámky vztáhne k libovolnému *Fragmentu*.

Zatímco první typ poznámek je garantovanou informací, která je přímo součástí původní textové edice, druhý typ poznámky umožňuje libovolnému autorovi komentovat libovolný text jiných autorů, aniž by bylo zasahováno do původního autorského obsahu.

Oba typy poznámek jsou vhodně prezentovány, aby byla povaha poznámky jasně rozlišena.

Předloha, Stránka, Osoba, Zdroj

Zatímco entity *Edice*, *Oddíl*, *Fragment* a *Poznámka* reprezentují v modelu intelektuální obsah, entity *Předloha*, *Stránka*, *Osoba* a *Zdroj* reprezentují reálné fyzické entity a slouží k propojení intelektuálního obsahu agregovaného v databázi se skutečným světem⁴.

Systém umožňuje v budoucnu doplňovat vztahy s dalšími typy entit dle potřeb a možností navazujících systémů (autoritní databáze, prosopografické databáze, digitální kartotéky, slovníky, geografické systémy apod.). Současná ověřovací implementace se přirozeně omezuje na propojení s fyzickými předlohami, systémy pro jejich prezentaci (digitální knihovna Manuscriptoria), autory a zdroji obsahu.

Praktická aplikace modelu

Vytvoření a naplnění databáze

Jako vhodný databázový stroj jsme zvolili NoSQL databázi neo4j (<https://neo4j.com/>) verze 3.3.0. kterou jsme zprovoznili v dedikovaném prostředí. Dotazovacím jazykem této databáze je Cypher Query Language (<https://neo4j.com/developer/cypher/?ref=product>).

Pro účely plnění databáze jsme definovali mapovací pravidla, která transformují obsah XML v TEI P5 do příkazů jazyka Cypher, jejichž prováděním v neo4j je instance databáze naplněna.

Pravidla jsou zapsána v jazyce XSL. Smyslem pravidel je analyzovat obsah TEI P5 XML, detektovat jednotlivé entity, vytvořit je, naplnit jejich vlastnosti hodnotami a vzájemně je propojit. Výstupem při provádění XSL transformace nad libovolnými vstupy ve formátu TEI P5 XSL je jeden aktualizací Cypher script. Mapovací pravidla lze adaptovat pro libovolné jiné formáty pro zachycení plných textů.

Vstupem pro plnění databáze může být jakýkoliv více či méně komplexní TEI P5 text v XML souboru. Při importu během zpracování každého vstupního souboru detekujeme a vytváříme (pokud ještě neexistují) jednotlivé entity v tomto pořadí:

- 1) detekujeme a vytváříme *Edici*,
- 2) detekujeme a vytváříme odpovědné *Osoby*, které spojujeme s *Edicemi*,
- 3) detekujeme *Předlohy* (fyzické exempláře) související s edicí a generujeme jednotlivé *Stránky*, pokud jsou tyto obsaženy v dané *Edici* (a rovnou provázeme *Stránky* s *Předlohou*),
- 4) generujeme strukturu *Oddílů*, které provázeme mezi sebou nebo s *Edicí*,
- 5) jednotlivé *Fragmenty* obsahu, které provázeme s *Oddíly* či přímo s *Edicí*.

Během hromadného importu nejsou vytvářeny *Poznámky*, protože případné poznámky jsou součástí poznámkového aparátu zapsaného jako markup jednotlivých Fragmentů (viz druhy poznámek).

Soubor s aktuálním mapováním XSL je k dispozici k nahlédnutí zde:

<http://www.manuscriptorium.com/cs/g2-fulltexty>.

Pro hromadné importy plných textů aktuálně platí:

- V jednom souboru by měla být uvedena právě jedna Edice.
- V jednom souboru může být vytvořena vazba na jednu nebi více Předloh nebo na žádnou předlohu.
- V souboru musí být uveden jazyk Edice.
- V souboru musí být uvedeno autorství Edice.
- V jednom dávkovém zpracování může být na vstupu více XML souborů tvořících jednu edici.
- Vstupní soubory musí být TEI P5 kompatibilní.

Uvedená pravidla jsou aplikována v rámci poloprovozu a nejde nutně o dogma. Pravidla lze adaptovat dle potřeby a na základě zkušeností získaných provozem (dle praxe poskytovatelů obsahu).

Hromadný import lze kdykoliv opakovat.

Uživatelské rozhraní

Úvodní stránka do uživatelského rozhraní je v době psaní článku dostupná na adrese

<http://www.manuscriptorium.com/cs/g2-fulltexty>

Vstupními body do aplikace jsou:

- Úvodní stránka: používá uživatel při přímém vstupu do aplikace
- Stránka čtení kopie na stránce: používá uživatel při vstupu z digitální knihovny Manuscriptoria (nebo jiné digitální knihovny)

Vyhledávání v edicích

Součástí aplikace je základní vyhledávání ve fragmentech textů jednotlivých edic.

Výsledkem vyhledávání je seznam edic. Kliknutím na edici v seznamu edic je tato otevřena.

Edice může být otevřena různými způsoby:

- 1) Na první stránce
- 2) Na první stránce s vyhledaným slovem

Při čtení edice lze v edici vyhledávat. Výsledkem vyhledání jsou pak fragmenty odpovídající hledanému výrazu.

Režimy čtení

V aplikaci jsou implementovány tyto režimy čtení:

- Čtení souvislého textu edice,
- Listování po stránkách s podporou paralelního čtení obraz-text
- Zobrazení galerie náhledu relevantních stránek digitální kopie předlohy.

Screenshot.

Způsob řešení propojení edic na digitální kopie v různých systémech

Pro podporu funkce paralelního čtení obraz-text je v současné době aplikace napojena přímo na API digitální knihovny Manuscriptoria⁵, odkud přebírá jednotlivé obrazy stránek, na které je obsah textů provázán.

Tato vazba na existující digitální kopie předlohy je řešena pomocí identifikátoru fyzického exempláře, který je dynamicky dopočítáván při vložení informace o existenci *Předlohy* do databáze plných textů. Výhodou takového řešení je, že nejsou vkládány pevné odkazy na existující digitální kopie. Edice tak může být kdykoliv provázána s *Předlohou* a digitální obrazy stran předlohy jsou dynamicky provázány s obsahem až v okamžiku zobrazení edice v aplikaci. Jinými slovy autor neřeší existenci digitální kopie v době tvorby textů, pouze prováže obsah s *Předlohou* a prezentační aplikace sama obrazy prováže ve stejném okamžiku, kdy je v Manuscriptoriu zveřejněna digitální kopie předlohy.

Díky takto řešené vazbě se lze v budoucnu na digitální kopii předlohy dotazovat přes API systému IDHF⁶, který vrátí jak informaci o existenci digitální kopie předlohy, tak i o příslušném API pro přístup k obrazům. V budoucnu tak bude možné implementovat API rozhraní dalších digitálních knihoven a bude možno snadno realizovat paralelní čtení obrazu předlohy a textu edice.

Virtuální edice

Aplikace umožňuje uživateli zobrazení tzv. virtuálních edic. Tuto funkci považujeme za jednu ze zásadních přidaných hodnot celého řešení.

Systém může dohledávat související edice a informuje uživatele o jejich existenci. Například se uživatel při čtení úryvku Kosmovy kroniky dozví, že existuje jiný úryvek téhož textu, nebo stejný úryvek v jiném jazyce (překlad) a může si zobrazit a zorganizovat fragmenty z obou edic v rámci vlastní edice.

V tuto chvíli je implementováno sledování vazeb mezi Edicemi, jejichž Fragmenty se vztahují ke stejným Předlohám.

⁵ Toto API bylo do systému Manuscriptoria nově doplněno právě pro potřeby paralelního čtení plných textů. Přebírá identifikátor fyzického exempláře a popis, číslo nebo folio stránky a vrací první odpovídající obraz. Toto řešení není bezchybné (pokud existuje v jednom dokumentu více stránek se stejným popisem, dojde k chybě), ale přináší velkou míru nezávislosti autora na výstupech digitalizace a je tak v souladu s koncepcí celého řešení.

⁶ Takto by v režimu plného provozu mělo být propojení realizováno v každém případě, protože garantuje správné propojení bez ohledu na to, v jakém čase je platné místo uložení fyzické kopie dokumentu (viz IDHF a Digitální konkordance).

Pokud uživateli připadá nabízená souvislost jako zajímavá, může v interface spojit jednotlivé edice nebo jejich části a zobrazit si edici zcela novou – virtuální. Může tak například paralelně číst obraz stránky, transliteraci textu stránky z edice A a moderní přepis textu stránky z edice B.

Sledování dalších vztahů může být dynamicky doplňováno a uvažujeme i o možnosti doplnění interface pro dotazování vazeb, který by byl dostupný přímo uživateli⁷. Předpokládáme, že potenciál funkce virtuálních edic bude umocněn později při implementování podpory crowdsourcingu (virtuální edice může být doplněna o řadu informací vytvořených crowdem – viz níže).

Možnosti využití crowdsourcingu

Již z výše uvedeného vyplývá, že celý model je extrémně dobře připraven na využití potenciálu crowdsourcingu.

V rámci crowdsourcingu předpokládáme tvoření obsahu následujících typů:

- uživatelské *Poznámky* (komentáře) k *Fragmentům* jednotlivých *Edic* (i cizích),
- doplňování nových vazeb:
 - *Fragmentů* na *Stránky Předloh*,
 - mezi *Edicemi* či *Fragmenty* (*souvisíS* apod.),
 - na externí zdroje (*odkazovánV*, *zobrazenV* apod. na autoritní zdroje, mapové portály, digitální knihovny a kartotéky atd.).
- Vytvoření nových *Fragmentů*, i celých nových *Edic*.

Pro realizaci samotné fáze tvoření obsahu uživateli stačí doplnit příslušný interface pro výrobu vazby a výrobu entity a interface pro správu vlastních entit a vazeb, které uživatel vytvořil. Návrh interface bude samostatným úkolem.

Jednotliví uživatelé budou oprávněni měnit výhradně vlastní obsah.

Obsah vytvořený uživateli pak bude prezentován jako související (a zřetelně odlišený) obsah při prohledávání a čtení edic a bude kdykoliv k dispozici ostatním uživatelům jako obsah pro obohacení virtuální edice.

Uživatelské poznámky

Poznámky budou existovat jako solitérní entity nebo jako entity propojené s ostatním obsahem – dle úvahy a potřeb uživatele.

Uživatelské edice

Vložením prvního fragmentu neexistující edice vznikne nová edice, jejímž autorem je uživatel. Uživatel v takovém případě bude systémem vyzván, aby doplnil vhodné rozlišující informace. Uživatelská edice je totiž určena stejnou kombinací informací, jako edice agregovaná z TEI P5 XML souborů při hromadném importu agregovaných zdrojů spolupracujících autorů.

Edice vznikající v režimu crowdsourcingu jsou vždy edicí specifického typu. Například překlad je vždy označen typem jako „překlad, crowdsourcing“⁸.

⁷ Takové funkce budou předmětem dalšího dílčího rozvoje.

⁸ Případně doplníme do modelu samostatnou vlastnost pro rozlišení původu obsahu.

Předpokládáme, že uživatelé uvítají doplnění funkce exportu edice do TEI (defragmentace).

Redakční rada a motivace uživatelů

Oprávnění tvořit obsah bude spojeno s konkrétní rolí a uživatel bude muset požádat člena Redakční rady o příslušné oprávnění (přiřazení role).

Vzhledem ke specifčnosti prostředí a vlastnostem komunity odborných uživatelů doporučujeme v rámci dalšího výzkumu a vývoje navrhnout a realizovat takové řešení, které využije vlastností modelu v plné míře tím, že umožní okamžité přidávání obsahu uživateli (bez kontroly redakční radou) s okamžitým zveřejněním takového obsahu do prezentační aplikace.

Redakční rada bude mít k dispozici interface ke kontrole vytvořeného obsahu a bude mít možnost zpětně mazat nevhodný obsah a blokovat uživatele, kteří jej vytvoří.

Samostatnou kapitolou je pak vytváření takového prostředí v komunitě uživatelů, aby motivace obsah tvořit byla dostatečně vysoká pro vznik relevantního korpusu plných textů. Toto téma přesahuje hranice tohoto článku a spadá do oblasti spolupráce např. oddělení Manuscriptorium a PR oddělení Národní knihovny České republiky, jako poskytovatele služeb systému Manuscriptorium.