AiP Beroun s.r.o., http://www.aipberoun.cz, tel.: +420 311 611 237, fax: +420 311 611 238
Talichova 807, Beroun, 266 01; IČO: 25778943, DIČ: 25778943

Principy komunikace při poskytování plných textů dokumentů

verze 1.0.1

Tomáš Psohlavec & kol., AiP Beroun s.r.o.

Úvod

V současné době existují doporučitelné standardy, které lze dobře využívat pro zápis plných textů historických dokumentů. Vzniká tak množství digitálních plných textů, které mají formu transkripce či transliterace, případně autorské edice opatřené poznámkami apod. Jedná se o komplexní díla, pro jejichž tvorbu je vyžadována vysoká míra odborných znalostí. Specializovaní odborníci vytvářejí obsah v rámci různých projektů, v různých místech a různých časech.

Výsledky jejich práce jsou v průběhu času prezentovány, překládány, doplňovány, komentovány, upravovány – samotnými autory či jejich kolegy. S tím, jak roste množství souvisejících digitálních plných textů a jejich metadat, začínají se přirozeně objevovat problémy jak se správou, tak s prezentací takového obsahu.

V oblasti historických fondů může jeden plný text obsahovat fragmenty textů pocházející z různých exemplářů a zároveň i jeden fragment textu může být nalezen v různých variantách v různých exemplářích. Navíc je potřeba u vznikajících plných textů uvádět a udržovat vazbu na exempláře, ve kterých se texty původně vyskytovaly (včetně informace o tom, kde se v exempláři text nacházel).

Situace správce plných textů je pak poměrně složitá a konvenční přístup ke správě dokumentů na úrovni správy souborů s plným textem, přestává být v oblasti historických fondů udržitelné.

V současné době jsou totiž databáze plných textů typicky orientované dokumentově: hlavní logickou entitou, která je spravována a prezentována, je plný text uložený v jednom souboru – typicky jeden plný text je uložen v souboru např. jako XML ve formátu TEI P5.

Formát TEI je navržen jako standard pro reprezentaci textu v digitální podobě. Zatímco struktura formátu TEI perfektně ošetřuje možnost zachycení *jednoho* textu *v jednom* fyzickém exempláři *jedním* autorem (přepisovačem) *v jednom* XML souboru, v reálném prostředí fulltextů agregovaných z více zdrojů se běžně setkáváme se situacemi mnohem komplikovanějšími, které samotná existence standardu TEI nemůže dosti dobře řešit.

Příklady situací, které je obtížné řešit v dokumentově orientovaném systému

Zde uvádíme příklady problematických situací, které je obtížné řešit metodami a nástroji, které jsou dnes běžně dostupné a úspěšně používané v oblasti fondů moderních.

Tyto situace jsme identifikovali na základě reálných zkušeností čerpaných (nejen) během provozu systému Manuscriptorium, který se zaměřuje na agregaci existujících digitálních informací v oblasti historických knižních fondů.

P1.1: Jeden stejný text se vyskytuje ve více různých exemplářích

Formát TEI předepisuje způsob, jak tuto situaci ošetřit – viz následující příklad dle praxe Národní knihovny České republiky.

Autor plného textu nejprve uvádí, ve kterých skutečných pramenech se plný text vyskytuje a tyto označí:

```
<sourceDesc>
	<listBibl>
		<bibl>
			<idno>INC.85</idno>
			<note>Zdroj fulltextu. Uložení: <seg type="country">Espaňa</seg>
				<seg type="settlement">Madrid</seg>
				<seg type="repository">Biblioteca Nacional de Espaňa</seg>
				<seg type="ed">1</seg>
			</note>
		</bibl>
		<bibl>
			<idno>INC 1811</idno>
			<note>Zdroj fulltextu. Uložení: <seg type="country">Espaňa</seg>
				<seg type="settlement">Madrid</seg>
				<seg type="repository">Biblioteca Nacional de Espaňa</seg>
				<seg type="ed">2</seg>
			</note>
		</bibl>
		<bibl>
			<idno>INC 1101</idno>
			<note>Zdroj fulltextu. Uložení: <seg type="country">Espaňa</seg>
				<seg type="settlement">Madrid</seg>
				<seg type="repository">Biblioteca Nacional de Espaňa</seg>
				<seg type="ed">3</seg>
			</note>
		</bibl>
	</listBibl>
</sourceDesc>
```

Následně v plném textu používá odkazy na předěly stránek s označením, které v rámci souboru s TEI pro jednotlivé zdroje zavedl:

```
<body>
	<div>
		<pb ed="1" n="183"/>
		<pb ed="2" n="197"/>
		<pb ed="3" n="178"/>
		<head>Panegyricus dictus Olybrio et Probino consulibus</head>
		<p>Sol, qui flammigeris mundum conplexus habenis
volvis inexhausto redeuntia saecula motu,
sparge diem meliore coma, crinemque repexi
blandius elato surgant temone iugales
efflantes roseum frenis spumantibus ignis
iam nova germanis vestigia torqueat annus
```

```
consulibus, laetique petant exordia menses.
Scis genus Auchenium, nec te latuere potentes
Amniadae, nam saepe soles ductoribus illis
instaurare vias et cursibus addere nomen
his neque per dubium pendet Fortuna favorem
<pb ed="2" n="198"/>
nec novit mutare vices, sed fixus in omnes
cognatos procedit honos. Quemcumque require
hac de stirpe virum: certum est de consule nasci.
Per fasces numerantur avi semperque renata
nobilitate virent, et prolem fata secuntur
continuum simili servantia lege tenorem.
Nec quisquam procerum temptat, licet aere vetusto
floreat et claro cingatur Roma senatu,
se iactare parem, sed prima sede relicta
Aucheniis de iure licet certare secundo:
<pb ed="1" n="184"/>
<pb ed="3" n="179"/>
haud secus ac tacitam Luna regnante per aethram
sidereae cedunt acies, cum fratre recusso
aemulus … harenas.
Non, mihi centenis pateant si vocibus ora
<pb ed="2" n="199"/>
multifidusque ruat centum per pectora Phoebus,
acta Probi narrare queam, quot in ordine gentes
rexerit, ad summi quotiens fastigia iuris
venerit, Italiae late cum frena teneret
Illyricosque sinus et quos arat Africa campos.
Sed nati vicere patrem, solique merentur
<pb ed="1" n="185"/>
<pb ed="3" n="180"/>
victores audire Probi. Non contigit illi
talis honos, prima cum parte viresceret aevi,
nec … flos iuvenilis inumbret
oraque ridenti lanugine vestiat aetas.
Tu, precor, ignarum doceas, Parnasia, vatem,
quis deus ambobus tanti sit muneris auctor.
… lumine, quem tota variarat Mulciber arte:
hic patris Mavortis amor fetusque notantur
Romulei; pius amnis inest et belva nutrix;
electro Thybris, pueri formantur in auro;
fingunt aera lupam; Mavors adamante coruscat.
<pb ed="2" n="200"/>
Iam, simul emissis rapido velocior Euro
fertur equis, stridunt Zephyri cursuque rotarum
<pb ed="1" n="186"/>
…
<pb ed="3" n="185"/>
O bene signatum fraterno nomine tempus,
o consanguineis felix auctoribus anne,
incipe quadrifidum phoebi torquere laborem.
Prima tibi procedat hiemps non frigore torpens,
non … duces; te cuncta loquetur
tellus; te variis scribent in floribus Horae,
longaque perpetui ducent in saecula fasti.</p>
            </div>
</body>
```

Příklad 1: Panegyricus dictus Olybrio et Probino consulibus, elektronická edice, kódování XML: Matthew Gan

Co už TEI však přirozeně neřeší: jak ošetřit situaci, kdy plný text vytvoří Autor A přepisem z exempláře A, zatímco později Autor B nalezne stejný text nebo jeho část v exempláři B. Má tuto informaci přidat do stejného zdrojového souboru k obsahu A? Dostane se k takovému zdrojovému souboru? A pokud se dostane, smí vůbec Autor B doplňovat plný text do souboru autora A? Kdo o tom rozhoduje? Podobných otázek může vyvstávat celá řada.

Přitom podobné situace, kdy budou v různých fondech, v různých časech a různými badateli nalézány fyzické exempláře s příbuznými či stejnými texty, které bude žádoucí propojovat s plnými texty již

existujícími, budou zcela běžné (čím více plných textů vznikne a čím více exemplářů bude digitalizováno, tím častěji tato situace nastane).

P1.2: Různé texty se vyskytují v jednom nebo více fyzických exemplářích

Uvažovaná situace je variantou předchozího problému, jen je situace ještě složitější.

Například exempláře, které jsou konvoluty fragmentů mnoha historických dokumentů, mohou obsahovat různé texty, které už existují v digitální podobě, jako plné texty jiného nebo mnoha jiných exemplářů.

Pokud tedy budeme chtít, aby čtenář při prohlížení takového digitalizovaného konvolutu měl informaci i o již existujících fulltextech, jak toho dosáhneme? Budeme se snažit o doplnění dílčích metadat do původních XML TEI souborů? Pak budeme řešit podobné otázky, jako jsou nastíněny výše, ovšem řešení bude ještě obtížnější.

Narazíme navíc na další problémy. Například na to, že u dokumentů uvažovaného typu bude téměř vždy zcela jiné nejen *stránkování*, ale i samotné *řazení* fragmentů textů (jinak svázané konvoluty či seřazené fragmenty). A to je situace, kterou ani TEI formát nedokáže dost dobře ošetřit. Teoreticky bychom se mohli snažit rozlišit rozdílná pořadí tak, že budeme (poměrně neprakticky) vždy uvádět značky <pb/> i pro všechny předchozí a následující fragmenty textu v jednotlivých exemplářích (které však ještě v existujícím plném textu ani nemusí být zachyceny) a nechat rekonstrukci pořadí fragmentů v jednotlivých exemplářích na algoritmu stroje. Kromě toho, že jako autoři pak budeme doufat, že neuděláme chybu (kontrola při tvorbě by byla problematická), zcela jistě se dostáváme na limit i tak pokročilého aparátu, jako je TEI P5 – takové použití nebylo zamýšleno. Navíc návrh pravidel řazení podle <pb/> by sám o sobě byl výzvou (neexistuje univerzálně platný předpis a abecední či číselné řazení dle názvu stránky neodpovídá řazení v exempláři).

Vycházíme tedy z toho, že popsanou situaci nelze efektivně řešit editací jediného XML TEI P5 souboru. Pokud tedy toto není možné, musíme vytvářet nový (redundantní) XML TEI soubor, jehož texty budou sice jinak poskládané, ale jinak obsahově shodné s již existujícími?

Ani tento jeden způsob evidentně není doporučitelný.

P2.1: Pro část textu vytvořeného Autorem A chce Autor B doplnit vlastní poznámky

Předpokládejme, že autor A v čase T_A vytvoří přepis dokumentu, zatímco autor B jej v čase T_B chce opatřit poznámkami o různočtení, jak je vidět v následujícím příkladu z Manuscriptoria:

```
<sourceDesc>
        <msDesc xml:lang="cs">
                <msIdentifier>
                        <country key="xr">Česko</country>
                        <settlement>Praha</settlement>
                        <repository>Knihovna Národního muzea v Praze</repository>
                        <idno>I H 51</idno>
                </msIdentifier>
                <msContents>
                        <msItem>
                                <title>[Kuchařka]</title>
                        </msItem>
                </msContents>
                <history>
                        <origin>
                                <origDate>začátek 16. století</origDate>
                        </origin>
                </history>
        </msDesc>
</sourceDesc>
…
<div>
        <head>Kaše z zvěřiny takto se má dělati</head>
        <p>Vezmi jeleninu kýtní, vosole pec na rožníku, když se upeče, tluc v moždieři a rozpusť vlaským vínem anebo malvazím. Potom protiehni skrze hartuch. Vezmi cukru nebo medu, ať jest sladko, dajž tam vína řeckého, zpera čistě, a sádla čistého nalí, nalíž vajec, což by se dalo, že by husto bylo, a protiehni skrze hartuch, dajž kořenie všecka. Chceš li žlutú mieti, daj šafránu. Pakli chceš, nechajž, když bude vřieti, vlíž tam ta vajce a miechaj, ať se nesrazie. A když se koli zsadí, odstav do ohně a potom daj na mísu.</p>
</div>
<div>
        <head>Kaše fíková</head>
        <p>Přistav fíky u víně neb v malvazí, vařiž<note>
                        <choice>
                                <corr>vařiž</corr>
                                <sic>wawrziz</sic>
                        </choice>
                </note> je, ať vyvrú všecky. Potom voceď ztluc v moždieři, nasuše topének z bielého chleba, rozmočiž v tom, v čem si fíky vařil, vytáhniž to skrze hartuch, vondajž do čistého kotla, dajž do toho medu a<pb n="6r"/>nebo cukru. Pak dajž kořenie k tomu – pepř, zázvor, skořici, hřebíčky, květu málo, anézu tlučeného, chceš li omasť, daj na mísu a přisol, zdá liť se.</p>
</div>
<div>
        <head>Kaše z játr volových</head>
        <p>Vezmi játry a vař je v hovězí jíše, ztluc je v moždieři, naspi tam jalovce, což by se zdálo. A když to ztlučeš namiesto, nalíž vajec, coť se zdá, zhřeje víno <supplied>…</supplied>
                <note>pravděpodobně vynecháno slovo</note> nebo vlaské, rozpustíš a protiehni skrze hartuch, ať jest nehusto, ani velmi žídko, dajž pak kořenie – pepř, zázvor, skořice a květu nemnoho. Potom vezmi čistého chleba bielého, nastruž a prosej skrze čistý duršlák. Vezmi anézu a kmínu, ztluc to v moždieři, nasypiž do té kaše ten chléb, vařiž a miechaj, ať se nesrazí. Potom daj na mísu.</p>
</div>
…
```

Příklad 2: [Kuchařka], elektronická edice, editor: Černá, Alena M., kódování TEI: Lehečka, Boris; oddělení vývoje jazyka Ústavu pro jazyk český AV ČR, v. v. i.

Zatímco z příkladu je patrné, že TEI umožňuje připojovat k textu poznámky, opět však bude v praktickém světě obtížné připojovat poznámky z různých zdrojů (bylo by nutno řešit společnou editaci jednoho plného textu více autory).

P2.2: Existence variantních plných textů

V tomto případě se jedná o variantu problémů nastíněných v předchozích textech. Původní texty přepisované v průběhu věků a zachycené nyní v digitální podobě existují v mnoha variantách - různě datované přepisy, moderní edice, překlady – některé se liší zásadně, jiné se mohou lišit jen v drobnostech. Případně mohou existovat texty, ve kterých pasáže zcela shodné s existujícími texty střídají texty zásadně odlišné.

Těžko je odhadovat (natož závazně určit), kde leží hranice, za kterou už má smysl vytvářet plné texty jako samostatně stojící dokumenty a kdy jde o práci z větší části nadbytečnou s ohledem na korpusy již existujících textů?

P3: Nesnadná agregace dokumentů

I kdybychom dokázali výše uvedené problémy řešit v měřítku lokálních projektů, v heterogenním prostředí, kde digitální texty vznikají za různých podmínek, se velmi rychle dostaneme do problémů. Proč?

Ačkoliv formát TEI pro tvorbu plných textů doporučujeme, není pravděpodobné, že všechna odborná pracoviště budou pracovat s touto formou. I kdyby to tak bylo, již vůbec nelze zodpovědně očekávat, že všechna pracoviště převezmou naše doporučení, jak s tímto formátem nakládat při správě fragmentů textu a udržování vazeb na fyzické exempláře.

Naprosto zbytečná je pak úvaha nad tím, zda by se taková pracoviště řídila centralizovaným rozhodováním o tom, co je a co ještě není tvorba redundantního obsahu, ke kterým textům má které pracoviště oprávnění modifikace a podobně.

P4: Nemožná implementace principů crowdsourcingu

Vytváření plných textů je odborná a časově náročná činnost. Je tedy i poměrně nákladná. Proto může být vhodné zahrnout do tvorby obsahu principy crowdsourcingu: například studenti specializovaných vysokých škol jistě mohou vytvářet dostatečně kvalitní přepisy textů historických dokumentů.

Aktuálně ovšem toto není možné, právě proto, že masová a neorganizovaná tvorba obsahu povede dříve či později ke stejným problémům, jako popisujeme výše (spíše dříve - množství vznikajícího obsahu popsané problémy rychle umocní). I proto jsme dnes svědky toho, že v oblasti historických dokumentů se zavádění crowdsourcingu vyhýbáme.

Návrh řešení pro správu a zpřístupnění plných textů

Vztahově orientovaný model

Je zřejmé, že na globální úrovni neexistuje funkční aparát k udržení informací o vztazích mezi dílčím obsahem textů, původními fyzickými exempláři, informacemi o původu textů i exemplářů, o

autorech, právech atp. Takový aparát ani nemůže z principu vzniknout, pokud budou digitální texty dále organizovány jako jednotlivé samostatné XML soubory.

Přitom ale naším dlouhodobým cílem je umožnit centralizovaný přístup k veškerému digitálnímu obsahu oblasti historických fondů. Není v zájmu odborné ani laické obce uživatelů, abychom při plnění tohoto zámětu opomíjeli plné texty. Jak tedy našich cílů dosáhnout?

Domníváme se, že správným řešením je opustit *dokumentově orientovaný* model a pro potřeby agregace a správy agregovaného obsahu zavedeme model *vztahově orientovaný*. Navrhujeme tedy nový způsob fragmentace logického obsahu fulltextů do dílčích samostatných entit, které bude možno jednoznačně identifikovat, opatřit potřebnými metadaty a především organizovat do vztahových map (lze říci, že velmi specializovaných ontologií), které mohou, při vhodné implementaci do moderních informačních systémů, řešení dané problematiky usnadnit. V konečném důsledku tak umožníme dosažení vytčeného cíle – vznik prostředí pro kooperativní tvorbu a snadné užívání digitálních plných textů historických dokumentů.

Vhodná logická fragmentace dostupného obsahu

V našem návrhu tedy vycházíme z toho, že onou fundamentální entitou, kterou lze snadno pracovat, nemůže být jednotka na úrovni plného textu - je tedy potřeba pracovat s jemnějším rozlišením.

Je zřejmé, že ani entita typu stránka není tou správnou elementární jednotkou pro správu, protože stránkování nemá vazbu k logické struktuře obsahu a navíc i zde jasná příliš těsná souvislost se strukturou konkrétního fyzického exempláře (předlohy). Proto je pomyslným základním stavebním kamenem při správě plných textů výběr jiná, vhodnější logická jednotka, která nemá přímou vazbu na strukturu exemplářů. Domníváme se, že touto jednotkou jsou odstavce textu.

Přirozeně lze namítnout, že i odstavce textu mají vazbu na strukturu konkrétního exempláře a jsou mnohdy formátovacím prvkem. To nelze popřít, a proto uvádíme, že pod pojmem odstavec rozumíme nikoliv graficky odlišený blok textu, ale *nejmenší logickou jednotku intelektuálního obsahu textu*, která je v předloze samozřejmě obvykle graficky odlišena.

Abychom ošetřili situaci, kdy nedojde mezi autory digitálních plných textů ke shodě ohledně toho, co je vhodnou nejmenší logickou jednotkou intelektuálního obsahu a abychom předešli potřebě diktovat jednotlivým autorům, co je a co už není nejmenší logická jednotka intelektuálního obsahu, zavádíme v navrhovaném datovém modelu možnost, aby *odstavec* obsahoval jeden nebo více *odstavců*.

Kromě toho lze oprávněně předpokládat, že u většiny textů se setkáme s identickou strukturací na odstavce (transkripce, transliterace a překlad dokumentu budou stejně strukturované).

Entity a jejich atributy

Fulltext

Atributy:

- Název fulltextu
- Typ fulltextu (transliterace, transkripce, edice, překlad..)
- ID fulltextu

Vazby:

- obsahuje *Odstavec*
- jeObsahem *Dokumentu*
- jeSpolutvořen *Uživatelem (*odvozená vazba - pokud existuje vazba přes *Odstavec* nebo přes *Poznámku* k *Odstavci* daného *Fulltextu*)

Dokument

Atributy:

- ID dokumentu

Vazby:

- obsahuje *Fulltext*
- obsahuje *Obrázek*
- jePoskytnut *Zdrojem*

Odstavec

Atributy:

- ID odstavce
- Markup
  kompletní XML markup daného odstavce, část atributů je vnořena do TEI markupu:
  například, má-li odstavec odkazovat na konkrétní poznámku, pak je součástí markupu kotva dané poznámky s odkazem na ID poznámky
- Jazyk markupu
  Indikátor jazyka použitého v obsahu atributu Markup

Vazby:

- jeVytvořen *Uživatelem*

- jePoskytnutZe *Zdroje*
- jeSoučástí *Fulltextu*
- jeZachycenNa *Obrázku*
- odkazujeNa Poznámku
- jeSoučástí *Odstavce*

Poznámka

Atributy:

- ID poznámky
- Markup
  kompletní XML markup dané pozámky – například v TEI; pokud je poznámka provázána s konkrétním místem v odstavci, je kotva s ID poznámky součástí TEI odstavce
- Jazyk markupu
  Indikátor jazyka použitého v obsahu atributu Markup

Vazby:

- jeVytvořena *Uživatelem*
- seVztahujeK *Odstavci*

Uživatel

Atributy:

- ID uživatele
- (další běžné atributy)

Vazby:

- jeAutorem *Odstavce*
- jeAutorem *Poznámky*
- má *Roli*
- spolutvoří *Fulltext*
  odvozená vazba (pokud existuje vazba přes *Odstavec* nebo přes *Poznámku* k *Odstavci* daného *Fulltextu*)

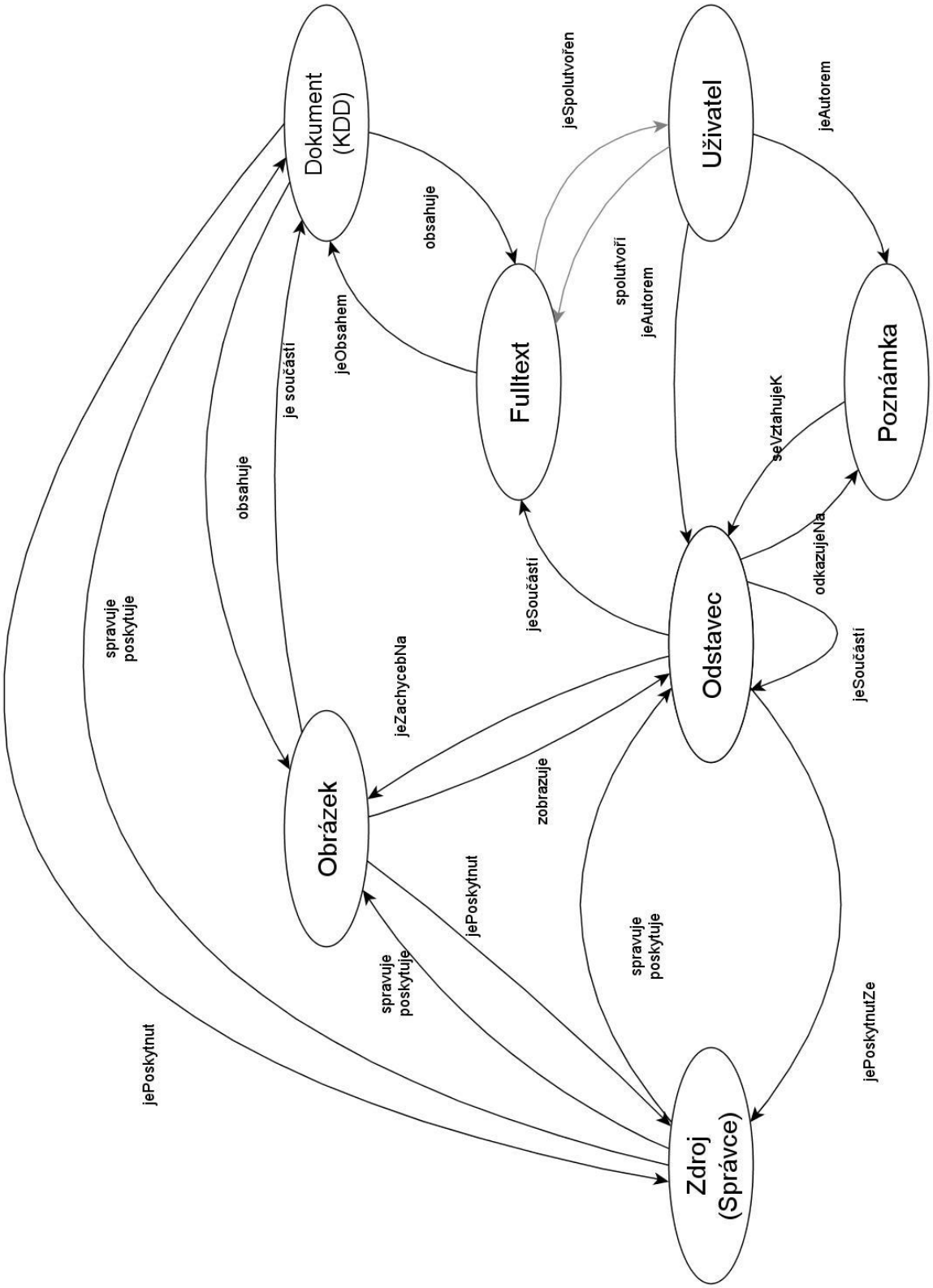
Obrázek

Atributy:

- ID obrázku
- Název předlohy
  typicky číslo stránky, ale model lze zobecnit na libovolné označení obecné předlohy zachycené v obrázku
- Typ předlohy

Vazby:

- jeSoučástí *Dokumentu*
- jePoskytnut *Zdrojem*
- zobrazuje *Odstavec*



Obrázek 1: schéma datového modelu

TEI markup jako hodnota vlastností klíčových typ entit

Jak je vidět z popisu jednotlivých typů entit, součástí atributů je i atribut s obsahem ve formě markupu ve zvoleném formátu. Předpokládáme využití TEI. To je výhoda návrhu popisovaného datového modelu, protože v navrhovaném uspořádání plně zachováváme výhody formátu TEI pro markup textů – rozsáhlé původní XML konkrétních entit Fulltextů pro naše potřeby pouze vhodně fragmentujeme.

Lze také předpokládat, že pokud budou v budoucnosti používány alternativní formáty pro zachycení plných textů historických dokumentů, budou obdobně logicky strukturovány (do *Odstavců*). Proto bude možné se změnou formátu fulltextů plně přejímat námi navrženou metodiku a datové struktury (a v budoucnu s minimálními korekcemi i implementované správní a prezentační nástroje).

Je evidentní, že v prostředí agregovaného obsahu mohou být pro tvorbu fulltextů současně použity v různých zdrojích různé formáty. Nic nebrání tomu, aby v systému, který bude pracovat s navrženým datovým modelem, koexistovaly obsahové entity, jejichž příslušné vlastnosti využívají pro zachycení atributů a vazeb rozdílné formáty. Je tedy zřejmé, že v navrženém řešení budou moci existovat fulltexty k exemplářům, kde jednotlivé fragmenty obsahu budou zachyceny v různých formátech, aniž bychom narazili na technické či organizační problémy.

Výhody navrženého řešení

Předpokládejme nyní existenci nástrojů, které umožní spravovat obsah plných textů a související informace ve formě vzájemně propojených entit. Pak je zřejmé, že v rámci takového systému lze nastíněné problémy řešit nepoměrně snadněji, než v systémech konvenčních dokumentově orientovaných.

P1.1: Jeden stejný text se vyskytuje ve více různých exemplářích
Problém je vyřešen tak, že entity odstavců jsou díky zachyceným vazbám snadno organizovatelné do jednoho fulltextu, přitom je ale snadno vysledovatelná příslušnost tohoto fulltextu k více exemplářům prostřednictvím vazeb na Obrázky, které jsou odvozeny právě z původních předloh a mají jednoznačné vazby na „své“ Dokumenty.

P1.1: Různé texty se vyskytují v jednom nebo více fyzických exemplářích
Problém je vyřešen tak, že entity odstavců jsou díky zachyceným vazbám snadno organizovatelné do různých souvislých fulltextů, přitom lze opět udržovat vazbu na strukturu původních dokumentů prostřednictvím vazeb na Obrázky, jak je popsáno v předchozím textu.

P2.1: Pro část textu vytvořeného Autorem A chce Autor B doplnit vlastní poznámky
Problém je vyřešen tak, že autor B vytváří poznámku k existujícímu obsahu zcela samostatně, aniž by jeho práce musela být koordinována s prací autora A a přitom nehrozí kolize. Poznámka autora B existuje samostatně, je zřejmé, že jde o poznámku jiného autora a s dílem autora A je „pouze“ provázána a vhodně prezentována. Je pak velmi snadné organizovat práci více uživatelů i řídit uživatelská oprávnění.

P2.2: Existence variantních plných textů
I v tomto případě je problém vyřešen – fragmenty (= Odstavce) variantních textů jsou přehledně organizovány, je zřejmá jejich příslušnost k fulltextům, dokumentům a fyzickým předlohám a díky atributům je zřejmý smysl existence konkrétních variant. Záleží pak pouze na autorovi obsahu a jeho vlastní tvůrčí/badatelské úvaze (v kombinaci s jeho uživatelskými oprávněními), jak bude nový obsah vytvářet a organizovat.

P3: Zjednodušení a automatizace agregace
Největší výhodou výše popsaného modelu je zásadní přínos pro řešení problémů při agregaci obsahu z různých zdrojů. Očekáváme, že libovolný plný text – ať vznikl za jakýchkoliv podmínek – je fragmentovatelný do samostatných odstavců. Při agregaci v reálném prostředí taková fragmentace může probíhat zcela automatizovaně – každý specifický zdroj fulltextů může být opatřen vlastním tzv. *fulltextovým konektorem*. V něm bude agregovaný fulltext rozdělen na odstavce, odstavce budou opatřeny identifikátory a na základě metadat agregovaných spolu s fulltexty budou naplněny jejich atributy a vytvořeny relevantní vazby. Tak vznikne samostatná vztahová mapa, která bude propojena s již existující mapou kompletního dostupného obsahu. Z existence takové vztahové mapy (velmi specializovaná ontologie) pak těží systémy pro prezentaci fulltextů.

P4: možné zavedení crowdsourcingu
Je zřejmé, že v uvedeném uspořádání můžeme s přiměřenou mírou rizik realizovat i projekty s prvky crowdsourcingu: i méně důvěryhodní uživatelé mohou vytvářet související obsah (přepisovat texty), aniž by ovlivnili obsah edic vyráběných odborníky apod. V rámci navrženého systému pak lze vystavět systém redakční rady, který tvorbu obsahu v crowdu umožní bezpečně řídit a kontrolovat.

Návrh technologií pro implementaci

Domníváme se, že pro implementaci podobného modelu lze využít grafové NoSql (not only SQL) databáze. Na základě výsledků předběžných testů se domníváme, že bude možné využít Neo4j databázi s dotazovacím jazykem Cypher.

Nad takovou databází by mělo být možné postavit dostatečně robustní a škálovatelný systém pro agregaci, správu a zpřístupnění fragmentů plných textů v Manuscriptoriu (nebo obecně v jakémkoliv projektu).

Principy komunikace

Komunikaci při poskytování plných textů vedenou mezi klientskou aplikací uživatele a repositářem plných textů, stejně jako komunikaci mezi spolupracujícími repositáři/systémy, navrhujeme realizovat po vzoru *International Image Interoperability Framework*, který navrhuje vlastnosti repositářů obrazových a předepisuje protokol pro poskytování jejich obsahu.

Obsahem fulltextového repositáře je síť entit pospojovaných vzájemnými vazbami a tedy poskytují přístup k mnohem komplikovanějšímu obsahu - přesto jsou principy komunikace podobné.

Navrhujeme tedy komunikaci přes REST rozhraní s tím, že protokol musí umožnit minimálně:

- dotázat se na vlastnosti repositáři jako celku
  (informuje o tom, co lze z repositáře získat, jakým způsobem, za jakých podmínek),
- požádat repositář o entity definovaných typů
  (poskytne výsledky vyhledání – například lze požádat o kompletní informaci o konkrétním Fulltextu, lze žádat o Odstavec a vysledovat, kterých dokumentů je součástí apod.),
- zpřesnit odezvu na dotaz omezujícím dotazem,
- dotázat se na vlastnosti (atributy i vazby) libovolné entity,
- získat hodnotu libovolného konkrétního specifikovaného atributu entity v kombinaci s jednoznačným identifikátorem dané entity.

Odpovědí pro dotaz na hodnotu vlastnosti je textová hodnota.

Odpovědí pro ostatní typy dotazů je vždy takzvaný manifest, který poskytuje kompletní dostupnou informaci o dané entitě. Tento manifest pak informuje klientskou aplikaci či partnerský systém o tom, co je v repositáři v dané souvislosti k dispozici a tazatel (jeho aplikace) má pak k dispozici všechny potřebné informace k tomu, aby s obsahem dále nakládal (například jej vhodně vizualizoval).

Závěr

Jak vyplývá z výše uvedeného textu, navrhovaný alternativní přístup k uspořádání plných textů by mohl řešit problémy nastíněné v úvodních kapitolách. V současnosti nám není znám žádný projekt, který by podobným způsobem obsah plných textů agregoval, spravoval či zpřístupňoval.

V budoucnu se tedy pokusíme uvést navržený model do praxe a ověřit jeho funkčnost v režimu poloprovozu s využitím dat připravovaných pro zpřístupnění v Manuscriptoriu, který také bude použit jako platforma pro implementaci souvisejících aplikací.