Manuscriptorium v. 1.0
Výběr a popis dokumentů

Verze 1.2

za Národní knihovnu ČR
PhDr. Zdeněk Uhlíř

23.2.2006

Obsah

O dokumentu

Účel

Tento dokument obsahuje základní pravidla pro výběr dokumentů k digitalizaci, vstupní informaci o standardech pro přípravu katalogových popisných záznamů a parametry dat a metadat pro zařazení evidenčních záznamů (metadat) a digitálních kopií dokumentů (dat) do systému Manuscriptorium.

Předpokládaný čtenář

Tento dokument je určen všem těm, kteří budou připravovat evidenční záznamy a digitální kopie dokumentů pro zařazení v systému Manuscriptorium.

Termíny a konvence

Termíny a konvence použité v tomto dokumentu jsou popsány a definovány v dokumentu [1], kap. 6.2. Kontext zpřístupnění dat a metadat je popsán v dokumentu [2].

Reference

V dokumentu se odkazujeme na následující literaturu:

[1] Manuscriptorium v.2.0 - analýza systému, prosinec 2004
[2] Manuscriptorium v. 2.0 – komplexní digitální dokument, říjen 2005
[3] Memoriae mundi series Bohemica, dostupné z URL: http://digit.nkp.cz
[4] Manuscript Access through Standards for Electronic Records (MASTER), dostupné z URL: http://xml.coverpages.org/master.html
[5] Reference Manual for the MASTER Document Type Definition, dostupné z URL: http://www.tei-c.org.uk/Master/Reference/oldindex.html
[6] M-Edit, dostupné z URL: http://www.memoria.cz/download/medit_cz.asp
[7] TorXmlValid, dostupné z URL: http://www.memoria.cz/site_cz/download.asp
[8] jEdit, dostupné z URL: http://www.jedit.org/
[9] Emacs, dostupné z URL: http://www.gnu.org/software/emacs/emacs.html
[10] NoteTabLight, dostupné z URL: http://www.webmasterfree.com/notetablight.html

Výběr dokumentů v Manuscriptoriu

Manuscriptorium kromě toho, že je otevřeným/souborným katalogem historických (knižních, resp. knihovních) fondů, tj. rukopisů a tištěných knih, příp. map do roku 1800 včetně, je také digitální knihovnou, tzn. zdrojem zpřístupňujícím úplné digitální kopie tohoto typu dokumentů. Možnosti výběru pro Manuscriptorium jsou tedy z obsahového hlediska neomezené. Praktickým omezením jsou však finanční možnosti poskytovatele dokumentů pro Manuscriptorium. Je důležité, aby výběr dokumentů k digitalizaci a ke zpřístupnění v Manuscriptoriu byl konzistentní, tzn. aby postihoval fondy poskytovatele z nějakého jasného hlediska. Skutečný výběr tak musí být založen na prioritách každého poskytovatele.

Jsou tři hlavní obecné možnosti, jak výběr provádět.

1) Dokumenty, které jsou pro fyzický fond poskytovatele charakteristické či typické, tzn. profilující složka jeho historického fondu, která bude digitalizována a posléze zpřístupňována pokud možno systematicky. V tomto případě je možno uvažovat také o připojení sekundárních dokumentů, tj. komentářů, shrnujících článků apod., třebaže to není podmínkou.
2) Dokumenty, které jsou pro fyzický fond poskytovatele považovány za nejvýznamnější z historického hlediska nebo nejkrásnější z výtvarného hlediska apod. V takovém případě je rovněž možno uvažovat o připojení vysvětlujících textů, třebaže to není podmínkou.
3) Dokumenty, o které je mezi uživateli největší zájem ať už z důvodu jejich reprodukce v tištěných materiálech, nebo z důvodu konkrétního badatelského zájmu (např. pořizování edic). V tomto případě je možno uvažovat i o propojení digitální obrazové kopie originálního dokumentu s jeho plným textem-edicí, třebaže to není podmínkou.

V případě dobrého finančního zázemí poskytovatele lze všechny tyto možnosti kombinovat. Naproti tomu v případě finančního zázemí skromnějšího je vhodné zvolit pouze jedinou z nich. V ideálním případě je preferována systematická digitalizace dokumentů pro fyzický fond poskytovatele typických či charakteristických (obecná možnost 1) s připojením dokumentů nejvýznamnějších (obecná možnost 2) a nejžádanějších (možnost 3), není to však podmínkou.

Popisy dokumentů

Popis dokumentů zásadně předchází vlastní digitalizaci. Popisy digitalizovaných dokumentů musí být dosti podrobné, protože jsou určeny nejen ke katalogizaci, ale také k propojení digitálních obrazů do formy elektronického dokumentu, jehož typickou podobou je virtuální kniha. Neosvědčil se však ani popis digitálních obrazů z důvodu absence mnoha informací, které jsou digitálním obrazem (kopií) nenahraditelné a které nese jen sám originální dokument, ani tradiční kodikologický způsob katalogizace, protože nejde primárně o náhradní slovní reprezentaci originálu, která má v tištěném prostředí nesporné oprávnění. Jestliže je v elektronickém prostředí dostupná reprezentace originálu v podobě obrazu a očekává se připojování dalších informací, pak se slovní prezentace (až interpretace) originálu v počátku stává postradatelnou, pokud není jejím účelem poskytnutí signifikantních informací vedoucích k nalezení dokumentu a informací v něm obsažených. Popis předcházející digitalizaci má výhodu mimo jiné i v tom, že při popisu dokumentu se současně provádí kontrola, zda je rukopis bez rizik způsobilý pro digitalizaci.

Princip pevně strukturovaného popisu DOBM (Digitization of Old Books, Manuscripts, and Other Documents) využívajícího SGML, který byl zaveden v počátcích digitalizace, má sice výhodu v

tom, že je jednoduchý, a usnadňuje tudíž rutinní až industriální práci. Nevýhodou však je to, že využívá pouze tvrdě strukturovaných dat na jedné nebo vůbec nestrukturovaných dat na druhé straně: bibliografické údaje a údaje o některých snadno typizovatelných vnějších znacích jsou ve formě tvrdě strukturovaných dat, zatímco ostatní údaje včetně údajů o intelektuálním obsahu originálního dokumentu jsou v podobě volného textu. V případě podrobného (a tedy rozsáhlého) popisu se tak klade překážka jak prosté orientaci v zobrazeném záznamu, tak sofistikovanějšímu vyhledávání. Ukázalo se, že pevná struktura záznamu je pouhým mezistupněm k takové formě záznamu, která bude využívat také dat semistrukturovaných.

Z toho důvodu byl princip pevně strukturovaného popisu v dalším vývoji opuštěn a byl nahrazen standardem MASTER (Manuscript Access through STandard for Electronic Records) vytvořeným v rámci evropského projektu, jehož řádným partnerem byla i Národní knihovna ČR.

Standard MASTER (nejprve na bázi SGML, posléze XML) umožňuje vytváření a využívání zejména semistrukturovaných dat, tzn. je přizpůsobivější jak variabilitě popisovaného materiálu, tak orientaci při zobrazení a postupům při vyhledávání. Je založen na strukturaci obsahových elementů do hloubky i na relativně volném využívání funkčních elementů, které se mohou vztahovat k různým horizontálním i vertikálním místům ve struktuře celého popisu. Pevná jsou pouze pravidla syntaxe. Ve standardu MASTER lze tudíž pořizovat jak zcela jednoduché, minimálně informačně nasycené záznamy, tak záznamy jdoucí do hloubky popisovaného originálního dokumentu. To znamená, že jeho praktické využití je velice široké a flexibilní, adaptovatelné pro různé účely i různou míru znalostí o materiálu, aniž je to na překážku využití v informačním systému.

MASTER

Zavedení popisů ve formě XML v rámci projektu MASTER **[4]** vedlo k prvotnímu popisu dokumentů ve volné a uživatelským potřebám přizpůsobenější struktuře využívající jen XML. Tyto prostředky jsou nyní volně dostupné a využitelné pro kohokoli, kdo se rozhodne poskytovat dokumenty pro Manuscriptorium. Důležité je, že původně podstatnou většinu popisů dokumentů zajišťoval dosti rozsáhlý kolektiv spolupracovníků, zatímco nyní významné množství dokumentů mohou popisovat jejich správci sami. K úspěšné spolupráci je nutné jen velmi krátké a jednoduché zaškolení, jaká je nutno dodržovat formální pravidla.

Záznam ve standardu MASTER je založen na základním členění na 1) identifikaci dokumentu, 2) záhlaví, 3) obsah, 4) fyzický popis, 5) historii dokumentu a 6) dodatečnou informaci. Toto členění je dále strukturováno, takže umožňuje vytvořit jak zcela elementární, tj. informačně minimálně uspokojivý záznam, tak záznam informačně bohatý a detailně strukturovaný. Jsou využívána tvrdě strukturovaná i semistrukturovaná data, takže lze nad tímto základem snadno vytvořit efektivní vyhledávání. Velkou výhodou standardu MASTER je to, že dovoluje integraci dat různých původců, a tedy být základem pro digitální knihovnu evropského rozsahu.

MASTER+

MASTER+ je extenzí standardu MASTER pro vytváření virtuálních knih, tzn. spojení popisného záznamu se sekvencí digitálních obrazů, které jsou kopiemi stránek originálního dokumentu. Pro vytvoření šablony dokumentu, k níž budou připojeny digitální obrazy, je nutno:

1) zjistit počet číslovaných listů/stránek a pokud se odlišuje od fyzického počtu listů/stránek (v případě, že některé z původně číslovaných listů/stránek chybí, nebo naopak v případě, že některé listy/stránky nejsou očíslovány), určit číslo posledního listu/poslední stránky;

2) zjistit mimořádně číslované listy/stránky (v případě, že některé listy/stránky nejsou číslovány, doplnit je v číselné řadě za poslední přecházející očíslovaný list/očíslovanou stránku s využitím speciálního označení);
3) zjistit chybějící listy.

Tuto šablonu lze vytvořit pomocí speciálního nástroje M-Edit, který vygeneruje příslušný počet stránkových úrovní s příslušným označením, na něž se navážou digitální obrazy. Pokročilejším nástrojem M-Tool lze vytvořit cesty od vygenerovaných stránkových úrovní přímo k digitálním obrazům. Tak lze odkazovat i na digitální obrazy, které nejsou fyzicky umístěny společně s příslušnými popisnými záznamy.

Standard MASTER+ za použití nástroje M-Tool umožňuje snadnou integraci zdrojů, a tedy vytvoření digitální knihovny v celoevropském rámci.

Praxe při psaní záznamů

Vytvářet jednoduché (minimálně informačně uspokojivé) popisné záznamy lze již po krátkém zaškolení. Vytvářet záznamy informačně bohatší již vyžaduje jistou praxi. Připravovat záznamy s vysokou informační hloubkou vyžaduje již pracovníka s delší zkušeností, který dobře ovládá význam jednotlivých elementů značkování i jejich strukturální syntax. Všechny tři základní stupně záznamů však jsou využitelné v jednom informačním systému. Všechny tři základní stupně záznamů jsou využitelné k vyhledávání, třebaže v rozdílné míře.

Velkou výhodou práce se standardy MASTER a MASTER+ je snadná integrace informačně heterogenních dat, pokud jsou formálně náležitě standardizována. Výhodou je i možnost pružného managementu při pořizování popisných záznamů, tzn. možnost snadno ovlivňovat produktivitu připravovatelů popisných dat podle aktuální potřeby.

XML editory

Popisné záznamy lze připravovat za použití různých nástrojů:

1) M-Edit **[6]** je editor umožňující vytvářet pouze minimálně informačně uspokojivé záznamy, přičemž není možno využít flexibility značkování. Ale zároveň dovoluje generovat základní šablonu dokumentu (stránkové úrovně). Vytvořené záznamy není nutno validovat.
2) NoteTabLight **[10]** umožňuje vytvářet složitější, tj. informačně hlubší záznamy, tzn. dovoluje využít flexibility značkování. Nevýhodou však je, že vytvořené záznamy je třeba validovat speciálním nástrojem TorXmlValid. Nevýhodou také je, že nedovoluje generovat základní šablonu dokumentu (stránkové úrovně).
3) Emacs **[9]** nebo jEdit **[8]** jsou sofistikované XML editory umožňující vytvářet nejsložitější popisné záznamy se zapojenou validací. Jsou to nástroje vyžadující profesionálně zkušeného pracovníka. Nevýhodou je, že nedovolují generovat základní šablonu dokumentu (stránkové úrovně).
4) M-Tool [http://www.memoria.cz/mtool/] je komplexní editor umožňující sice jen vytváření minimálně informačně uspokojivých záznamů, ale dovolující nejenom generovat základní šablonu dokumentu (stránkové úrovně), nýbrž také vytvářet cesty od vygenerovaných stránkových úrovní přímo k digitálním obrazům. Popisný záznam není nutno validovat, je implementována kontrola úplnosti dokumentu.

Závěr

K dispozici je tak dostatek nástrojů vhodných pro různě kvalifikované a profesionálně zdatné připravovatele dat. Je vypracována metodika (best practice) pro budování integrované digitální knihovny. Z technického obsahového hlediska je otevřena cesta k evropské digitální knihovně historických fondů.