

MANUSCRIPTORIUM

základní informace o kompatibilitě

DRAFT v 01

Úpravy

Datum	Verze	Typ	Autor
2009-01-18		Revize staršího dokumentu	StP
2009-03-24		Úpravy	ŠČ, FS
2009-03-28	D v 01	Finalizace	StP, FS, SC, JK

Obsah

1	Úvod o dokumentu	3
1.1	Účel	4
2	Cíle zavedených konvencí.....	5
3	Pravidla tvorby názvu adresářů a souborů	6
3.1	Základní konvence	6
3.2	Komponenty používané pro tvorbu jmen	6
3.2.1	Signatura normalizovaně	6
3.2.2	kód majitele	6
3.2.3	CRC	6
3.3	Názvy média (volume).....	7
3.4	Názvy adresářů	7
3.5	Název textového (XML) souboru	Chyba! Záložka není definována.
3.6	Názvy obrazových souborů	8
4	Digitální dokumenty na médiích	9
4.1	Popis struktury a obsahu adresářů archivních médií.....	9
4.1.1	Adresář s metadaty, (archivní verze)	9
4.2	Popis struktury a obsahu adresářů uživatelských médií	10
4.2.1	Adresář s metadaty (uživatelská verze)	10
4.2.2	DTD podadresář obsahuje DTD soubory s definicemi typu dokumentů potřebných pro XML soubory Metadat.....	10
5	Doporučené zdroje a související literatura.....	11
	Zkratky.....	12

Předpokládaný čtenář

Tento dokument je určen pro základní orientaci ve tvorbě Komplexních digitálních dokumentů (KDD) pro všechny, kteří se zabývají tvorbou digitálních dokumentů určených pro dlouhodobé zachování a použitelnost.

Toto platí zejména pro dokumenty vznikající v přímé souvislosti s Manuscriptorium.

1 Úvod o dokumentu

Literární dokumenty od jejich vzniku doprovází otázky způsobu jejich využívání, zachovávání a ochrany. Tématem současnosti je mimo jiné digitalizace.

Digitalizace a zpřístupňování historických literárních dokumentů v digitální probíhá v České republice již od devadesátých let dvacátého století.

Záhy bylo rozpoznáno, že digitální dokumenty jsou jen přirozeným využitím nové informační prostředí, využitím jiného media,. Digitalizace je určitou novou podobou ‚reedice‘ dřívějších záznamů myšlenek a idejí zaznamenaných vždy způsoby odpovídajícími své době .Digitální prostředí však dovoluje nové způsoby nakládání s daty, jejich sdílení, nové pohledy na využitelnost a souvislosti.

Digitální prostředí je z historického hlediska na samém počátku vývoje, mnohé se zde rychle mění, mnohé se teprve objevuje. Mnohé ještě není a ještě dlouho nebude tak stabilizováno, ani obecně přijato, na rozdíl od knižního prostředí, které má za sebou historii mnoha tisíc let.

Digitální prostředí je realita, uprostřed které žijeme a na kterou reagujeme. Proto byla záhy rozpoznána nezbytnost ustanovení základních paradigmat, která jsou základem pro dlouhodobé zachování a využívání kulturního dědictví v době digitálních technologií. Cílem je dosáhnout následujícího:

- Převod existujících informací do digitální podoby,
- zachování existujících a tvorba nových digitálních dat,
- zajištění dlouhodobé morální životnosti digitálních dat,
- zajištění dostupnosti digitálních dat bez závislosti na zdrojích a nástrojích.

Pro Manuscriptorium byla zvolena koncepce, která minimalizuje závislost na lokálních odborných názorech a zájmech, na názorech technických a obchodních.

Zvolená koncepce vychází z těchto východisek:

- Informace vznikaly a budou vznikat v dobré víře a vůli svých tvůrců.
- Informace reflektují úroveň dobového poznání a možností.
- Informace se zachovávají do budoucnosti v otevřené a nezávislé formě.
- Informace je nutno zpřístupňovat.

Koncepce Manuscriptoria na těchto základech staví a zároveň respektuje platné právní normy a etické zvyklosti.

1.1 Účel

Účelem dokumentu je poskytnout zájemcům o účast v Manuscriptoriu základní vodítko jak se stát partnery, pokud jejichž data nevznikají v přímé spolupráci s NKČR. Pro tyto partnery je obvyklá nezávislá správa primárních dat, nezávislé zajištění jejich dlouhodobé archivace a zpřístupnění na internetu. Všechny tyto činnosti (i jednotlivě) mohou být řešeny individuálními smlouvami s NKČR s provozovatelem Manuscriptoria Aip Beroun s.r.o. nebo s dalšími stranami.

Předmětem tohoto dokumentu je poskytnout základní informace o konvencích tvorby názvů obrazových souborů a jejich organizaci do svazků. Tyto základní atributy jsou součástí ‚vazby‘ digitálního dokumentu a mají napomáhat reflexi originálního dokumentu v KDD (viz Komplexní digitální dokument [2]).

Pokud potenciální partner z jakýchkoli důvodů nemůže či nechce přijmout navržené konvence, může nabídnout svá data k jednání o krocích nutných k jejich přijetí v existující formě. Vyhovují-li vznikající nebo již existující data/metadata, lze je konvertovat, odkazovat či importovat do Manuscriptoria různými způsoby. Postup vedoucí ke kompatibilitě může být řešen individuálně samotným partnerem nebo může být předmětem individuálních smluv s provozovatelem Manuscriptoria. Této alternativně se věnuje kapitola 5.

Koncepce Manuscriptoria (MnS) umožňuje agregovat, zpracovávat a zpřístupňovat informace z heterogenních zdrojů při respektování různých lokálních konvencí a podmínek, přičemž umožňuje homogenizovat prostředí pro uživatele.

Návrhy změn těchto konvencí je možno zasílat na mailem na adresu (viz [5]). Periodicky se provádí analýza návrhů, jejich přínosů i souvisejících nákladů. Změny konvencí se vyhláší a implementují jednorázově na delší období.

Plná přímá kompatibilita je zjednodušením spolupráce s Manuscriptorium. V evropském měřítku je však běžná spíše obecná strukturální a obsahová příbuznost souvisejících informací.

V Manuscriptoriu jsou oddělovány tři oblasti:

1. Komplexní digitální dokumenty (KDD)
2. Zpřístupňování KDD
3. Správa KDD

KDD jsou vnímány jako samostatné objekty prezentující v digitálním prostředí reálné objekty. KDD má v sobě sdružovat informace o digitalizovaném objektu maximálně nezávisle na nástrojích použitých při jejich vzniku, uchování a zpřístupňování. Totéž platí o struktuře KDD.

Pro správnou interpretaci KDD jsou kromě dat a odborných metadat potřeba i metadata dokumentující vlastnosti těchto informací, včetně jejich historie a souvisejících technologických informací.

To je základem dobré morální životnosti a obecné použitelnosti KDD.

2 Cíle zavedených konvencí

Zvolené konvence dovolují dosáhnout u KDD dlouhodobé nezávislosti na měnícím se digitálním prostředí, na nástrojích pro jejich vytváření, správu a zpřístupnění.

Z odborného hlediska je pro popis dokumentů soustředěn do souboru XML, jehož struktura a vlastnosti jsou deklarovány v dokumentu [2] DTD TEI5.

Při tvorbě KDD je třeba zachovat mnoho dalších informací kromě samotných dat - obrazů, textů. Důležitá jsou metadata vztahující se k digitalizovanému objektu i metadata technického charakteru a mimo jiné i informace o technologických okolnostech digitalizace (metodika digitalizace, typ zařízení a jeho charakteristiky, kalibrace,...).

Detailní popis informačního obsahu KDD není předmětem tohoto dokumentu.

V MnS se připouští více KDD vztahujících se k jedinému souboru digitálních dat (nejčastěji obrazů).

Při pojmenovávání souborů je snahou zachovat maximum dostupných informací o originálu a jeho struktuře. Důležité je respektovat v odborné komunitě obvyklý způsob identifikace dokumentů a to až na úroveň jednotlivých stran. Tyto informace mají napomoci především pro specialisty co možná nejvěrněji reflektovat originální objekt a umožňovat obvyklou orientaci a odkazování v literatuře.

K tomu je nezbytné:

1. zavést a dodržovat jednoznačnou identifikaci každého jednotlivého obrazového souboru z hlediska příslušnosti ke konkrétnímu digitálnímu dokumentu a vytvořit tak základnu dovolující přistupovat individuálně k jednotlivým obrazům digitálního dokumentu.
Příkladem využití této zásady je tvorba virtuálních dokumentů.
2. Reflektovat fyzické vlastnosti skladby/členění originálního objektu
3. Zachovat nezprostředkovaně všechny informace o objektu digitalizace v nově vznikajícím digitálním dokumentu a to nezávisle na nástrojích použitých při digitalizaci, nezávisle na nástrojích použitých při správě a zpřístupnění.

3 Pravidla tvorby názvu adresářů a souborů.

Pravidla odrážejí velmi širokou škálu možností potenciálních partnerů. Autoři dokumentu a provozovatelé Manuscriptoria se snaží neomezovat možnosti vytváření a lokální správy významných informací nesouvisejících bezprostředně s předmětem digitalizace a snaží se nezasahovat do lokálních záležitostí partnerů.

Aby byly dokumenty převoditelné do Manuscriptoria, je třeba zajistit jejich technickou kompatibilitu a využitelnost.

Popis těchto pravidel je uveden v dokumentu Manuscriptorium Technical compatible [3] a vznikl s ohledem na předpokládaný vývoj a partnery.

3.1 Základní konvence

Názvy smí obsahovat pouze znaky obsažené v ISO646

- a) velká písmena bez diakritiky 'A'.. 'Z' (0x41..0x5A)
- b) číslice '0'.. '9' (0x31..0x39)
- c) podtržítko '_' (0x5F)

Důvodem je zajištění trvalé přenositelnosti dat mezi různými operačními systémy. Délka názvů je přijatelná pro OS Windows, Linux, MacOS i ISO 9660 .

3.2 Komponenty používané pro tvorbu jmen

Pro signaturu, pro kód místa uložení a pro CRC platí následující pravidla.

3.2.1 Signatura normalizovaně

V signatuře se odstraní diakritika, malá písmena se převedou na velká, číslice se ponechají a ostatní znaky se převedou na znaky '_' (podtržítko - 0x5F).

Poté se odstraní tyto znaky ze začátku a konce signatury.

Rovněž se odstraní jejich vícenásobné výskyty za sebou.

V tomto tvaru a délce je string použit spolu s kódem majitele k určení CRC.

Pokud je normalizovaná signatura delší než 15 znaků, znaky zprava se odstraní.

Je-li délka signatury kratší než 15 znaků, doplní se zprava znaky '_' .

3.2.2 Kód majitele

6 místný kód jednoznačný splňující výše uvedené podmínky na povolené znaky.

Každý použitý kód musí být zaregistrován v Seznamu majitelů, který je vytvářen centrálně ve spolupráci s NKČR podle dohodnutých pravidel. Je-li kód kratší než 6 znaků, doplní se zprava znaky '_' .

3.2.3 CRC

Vícemístná rezerva pro odstranění identifikačních duplicit. Tuto část může partner vynechat nebo ji využít dle vlastních potřeb. Při nabídce dat do Manuscriptoria je však třeba rozsah této oblasti specifikovat a také dlouhodobě dodržovat.

Správce Manuscriptoria si vyhrazuje právo tuto část informace nezachovávat a využívat ji dle své potřeby.

3.6 Názvy obrazových souborů

SSSSSSSSSSSSSSSSSSCCCCQTFFFFF.EEE (30 znaků):

- S Signatura (15 znaků)
- C CRC (4 znaky),
- Q typ kvality (1 znak), kde
 - 'N' Normal – RGB základní, ke čtení a prohlížení
 - 'P' Preview - náhled pro orientaci na stránce
 - 'G' Gallery – náhled pro orientaci v dokumentu
 - 'S' Special - speciální úpravy, např. optimalizovaný B/W
 - 'E' Excellent - Nejvyšší kvalita obrazu která je k dispozici
- T třída kvality ('O'-základ, '1'..-vyšší, 'X'- pro nejvyšší)
- F foliace či paginace zleva dorovnaná znaky '0' na 5 znaků, viz následující odstavec.
- E přípona souboru v návaznosti na formát obrazu

Jméno obrazu je vytvořeno dle výše uvedených pravidel, přičemž pět znaků identifikujících stránku (F) je generováno takto:

Část rukopisu/tisku CZ	EN	pro paginaci	pro foliaci
běžný list, strana	Folio, Page	...G00001P.JPG	...G00001R.JPG ...G00001V.JPG
Vložený list	Enclosed Sheet	...G0ES01P.JPG	...G0ES01R.JPG ...G0ES01V.JPG
Zpevňovací proužek	Reinforcing Strip	...G0RS01P.JPG	...G0RS01R.JPG ...G0RS01V.JPG
Hřbet	Spine	...G0000SP.JPG ...G0000HE.JPG ...G0000SE.JPG ...G0000BE.JPG ...G0000FC.JPG ...G0000FS.JPG ...G0000BC.JPG ...G0000BS.JPG	
Horní ořízka	Head Edge		
Boční ořízka	Side Edge		
Dolní ořízka	Bottom Edge		
Přední desky	Front Cover		
Přední přideštiny	Front end-sheet		
Zadní desky	Back Cover		
Zadní přideštiny	Back end-sheet		
Římské číslov. Přední	Front roman pagination / foliation	...G0F001P.JPG	...G0F001R.JPG ...G0F001V.JPG
Římské číslov. Zadní	Back roman pagination / foliation	...G0B001P.JPG	...G0B001R.JPG ...G0F001V.JPG

Příklad pro kvalitu G0 a formát JPEG:

Soubory uložené na disku CD

```

MS_E_1_____2Q7IG0000BC.JPG
MS_E_1_____2Q7IG0000BS.JPG
MS_E_1_____2Q7IG0ES01P.JPG
MS_E_1_____2Q7IG0ES02P.JPG
MS_E_1_____2Q7IG0ES03P.JPG
MS_E_1_____2Q7IG0ES04P.JPG
MS_E_1_____2Q7IG0000FC.JPG
MS_E_1_____2Q7IG0000FS.JPG
MS_E_1_____2Q7IG0000SP.JPG
MS_E_1_____2Q7IG00001P.JPG
MS_E_1_____2Q7IG00002P.JPG
  
```

4 Digitální dokumenty na mediích

Následující informace je určena především pro partnery, kteří chtějí a mohou využít možnosti archivovat svá data na pevných mediích v digitálním archivu CD-R v NKČR. CD-R disky jsou vytvářeny ve struktuře, která dovoluje do Manuscriptoria přímo importovat kompatibilní data a metadata.

Pro ostatní partnery je určena jako metodické vodítko.

Digitální dokumenty se vytvářejí ve dvou typech.

Archivní je určen pro archivaci v CD-R archivu

Uživatelský je určen především k přímému využití pro uživatele.

Archivovány jsou oba typy disků.

Rozlišení disků je promítnuto do Názvu média a do struktury těchto médií.

4.1 Popis struktury a obsahu adresářů archivních médií

V kořenovém adresáři media jsou umístěny tyto soubory:

MNSXDEF.XML identifikační soubor (formát XML), obsahuje údaje o jazycích popisu, identifikaci dokumentu a informace o kolekci médií ze kterých se digitální dokument fyzicky skládá (např. série CD-R)

Start.htm startovací soubor (formát HTML) pro prohlížení média v prohlížeči

4.1.1 Adresář s daty a metadaty, (archivní verze)

je adresář, který obsahuje popisné soubory týkající se digitalizovaného dokumentu.

Obsahuje data nejvyšší kvality a je hlavním primárním zdrojem dat a metadat.

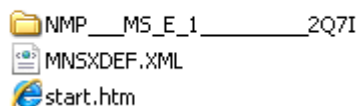
Struktura názvu se řídí dle pravidel pro [Název adresáře](#) s metadaty, viz výše..

Jeho obsahem jsou podadresáře s významem:

CZ, SK, ...	jazykové verze generovaných HTML souborů elektronické vazby
EN	anglická verze generovaných HTML souborů elektronické vazby
MISC	obsahuje soubory s popisem dokumentu a soubory s technickým popisem obrazů
	*.XML různé popisy dokumentu (Metadata), struktura názvů se řídí dle pravidel pro Název textového (XML) souboru
	DTD podadresář obsahuje DTD soubory s definicemi typu dokumentů pro XML soubory metadat
	.ICC profil ICC (= název vytvořeného profilu)
	gretag.jpg JPEG soubor s kalibrační tabulkou Gretag
	gretag.txt textový soubor s kalibrační tabulkou Gretag
	hexa.jpg JPEG soubor s kalibrační tabulkou Hexa
	hexa_tab.txt textový soubor s kalibrační tabulkou Hexa
G0	soubory potřebné pro zobrazování HTML stránek o dokumentu. Galerie náhledů, názvy obrázků se řídí dle pravidel pro Název obrazového souboru v minulé kapitole
EX	digitální obrazy nejvyššího rozlišení, tj. rozlišení, které bylo použito při digitalizaci, názvy obrázků se řídí dle pravidel pro Název obrazového souboru viz výše.

Příklad:

Soubory uložené na disku CD



4.2 Popis struktury a obsahu adresářů uživatelských medií

V kořenovém adresáři media jsou umístěny soubory:

Start.htm startovací soubor (formát HTML) pro prohlížení média v prohlížeči
MNSXDEF.XML identifikační soubor (formát XML), obsahuje údaje o jazycích
popisu, identifikaci dokumentu a informace o kolekci médií
ze kterých se dokument skládá

4.2.1 Adresář s daty a metadaty (uživatelská verze)

je adresář, který obsahuje popisné soubory týkající se digitalizovaného dokumentu. Obsahuje data uživatelské kvality, odvozená z nejvyšší kvality a určen pro přímé použití uživateli. Archivuje stejně jako primární zdroj.

Struktura názvu se řídí dle pravidel pro [Název adresáře](#), viz výše.

Jeho obsahem jsou podadresáře s významem: CZ, SK, ... jazykové verze generovaných HTML souborů elektronické vazby

EN anglická verze generovaných HTML souborů elektronické vazby
MISC obsahuje soubory s popisem dokumentu a soubory s technickým
popisem obrazů

*.XML různé popisy dokumentu (Metadata),
struktura názvů se řídí dle pravidel pro
[Název textového \(XML\) souboru](#) viz výše

DTD podadresář obsahuje DTD soubory s definicemi
typu dokumentů potřebných pro XML soubory Metadat.

G0 Galerie náhledů, soubory pro zobrazování HTML stránek o dokumentu.

Názvy obrázků se řídí dle pravidel pro [Názvy obrazových souborů](#) viz výše.

N3 soubory digitalizovaných stránek rukopisu v kvalitě definované pro uživatele.
Názvy obrázků se řídí dle pravidel pro [Názvy obrazových souborů](#) viz výše.

5 Minimální informační obsah dat a metadat

Pro partnery, kteří nemohou vlastními silami a prostředky dosáhnout kompatibility s Manuscriptorium je v následujícím naznačen minimální akceptovatelný rozsah dat a metadat. Využití této cesty předpokládá přímé jednání mezi provozovatelem MnS AiP Beroun s.r.o.s a zájemcem o partnerství. Popis dokumentu, jeho rozsah a kvalita bude konzultována s NKČR. Výsledkem spolupráce bude dohoda o vybudování individuálního konektoru.

Konektor lze vytvořit za předpokladu, že je zajištěna obsahová kompatibilita primárních textových metadat/dat partnera. Informace partnera musí být strukturované a musí v nich být k dispozici minimálně základní informace o popisovaných dokumentech.

Strukturovanými informacemi se rozumí dokumenty ve formátu XML, v databázových formátech či knihovnických formátech, přičemž lze stanovit algoritmus vedoucí k jejich sparaci do formy kompatibilní s Manuscriptorium.

Změna formy primárních dat je základem pro zahájení jednání o vytvoření nového konektoru pro partnera.

Minimální informace v metadatech o rukopisném dokumentu jsou tyto:

Instituce a místo uložení,

Signatura dokumentu,

název,

autor,

rok vzniku,

jazyk dokumentu,

iluminace,

obsah,

vazba,

materiál,

rozsah a rozměry,

pro staré tisky to jsou ještě tyto informace (impresum):

místo tisku,

jméno tiskaře,

datum tisku dokumentu

místo vydání,

jméno nakladatele,

datum vydání dokumentu

s ohledem na kompatibilitu s knihovnickými formáty pak:

země vydání

prvek věcného třídění – aspoň 1. výskyt Konspektu, předmětové heslo...

všeobecná poznámka.

Pokud jsou součástí dat obrazy, musí být v záznamu obsaženy i odkazy na ně. Kvalitativní hladiny obrazů mají korespondovat s odpovídajícími kvalitativními úrovněmi používanými v Manuscriptoriu. Podrobnější informace jsou uvedeny v dokumentu [Manuscriptorium Technical compatible](#) [3].

6 Doporučené zdroje a literatura

odkazy

- [1] Manuscriptorium v 2.0, Komplexní digitální dokument, říjen 2005
- [2] DTD TEI5
- [3] [Manuscriptorium Technical compatible, verze 1.2](#)
- [4] [Manuscriptorium Quality, Kvalita obrazových dat, verze 1.2](#)
- [5] info@manuscriptorium.cz

Zkratky

MnS - Manuscriptorium

KDD - komplexní digitální dokument

NKČR – Národní knihovna České republiky