Manuscriptorium v. 1.0

Manuscriptorium Technical compatible

Technická kompatibilita

Verze 1.2

za AiP Beroun s.r.o:

Ing Karel Kučera
Ing František Šibrava
Bc. Martin Majer

Obsah

O dokumentu

Účel

Tento dokument obsahuje základní pravidla 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.

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] MEdit, 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
[11] Manuscriptorium – Základy a kompatibilita: http://www.memoria.cz/docs/manuscriptorium_basics_and_compatibility_CZE.pdf
[12] Manuscriptorium – Výběr a popis dokumentů: http://www.memoria.cz/docs/manuscriptorium_document_description_CZE.pdf
[13] Manuscriptorium – Kvalita obrazu: http://www.memoria.cz/docs/manuscriptorium_image_quality_CZE.pdf

Úvod

Systém Manuscriptorium je koncipován jako digitální knihovna rukopisů, starých tisků a dalších vzácných dokumentů. Jako každá knihovna se skládá z katalogu, kterým je v tomto případě OKHF (Otevřený Katalog Historických Fondů) a vlastních digitálních dokumentů, které jsou uloženy v repository. OKHF soustřeďuje informace (metadata) o fyzických dokumentech (rukopisech atd.) ve formě tzv. evidenčních záznamů ve formátu XML. Repository obsahuje digitální kopie podmnožiny z těchto katalogizovaných dokumentů ve formě tzv. komplexních digitálních dokumentů (KDD). Principiálně systém Manuscriptorium centrálně soustřeďuje metadata v OKHF a zajišťuje přístup k digitalizovaným dokumentům na datových úložištích provozovatele i vzdálených úložištích dalších přispěvatelů.

Technická kompatibilita metadat

Aby jednotliví přispěvatelé mohli přispívat do systému Manuscriptorium jak evidenčními záznamy, tak s nimi souvisejícími digitálními daty, je nutno stanovit základní technické podmínky a parametry vstupních dat. Přispěvatel musí dosáhnout toho, že jím poskytovaná data i metadata jsou kompatibilní se systémem Manuscriptorium.

Pak provozovatel systému garantuje jejich import do Manuscriptoria.

Kompatibilita přispěvatelů se systémem Manuscirptorium musí být zajištěna ve dvou úrovních. Přispěvatelé musejí do Manuscriptoria přispívat jednak kompatibilními metadaty a v případě spolupráce se systémem v oblasti digitálních dat také formátovou kompatibilitou, dodržením jmenných konvencí pro názvy souborů apod.

Formátová kompatibilita metadat

Data pro evidenční záznamy musí být dodávána ve formátu XML a v kódování UNICODE UTF-8. Pro strukturu XML souboru je předepsán standard MASTER. Pro generování digitálních dokumentů umožňujících přístup k distribuovaným datům na uložištích přispěvatelů bude použit formát XML MASTER+, který bude popsán dále. Obrazová data musí být uložena ve formátech, které jsou přímo podporované internetovými prohlížeči (browsery). Jsou to formáty JPEG, GIF a PNG.

Obsahová kompatibilita

Aby bylo technicky možno do centrální DBEZ ukládat evidenční záznamy od jednotlivých přispěvatelů, musí tyto obsahovat informace alespoň na úrovni **povinných elementů DTD MASTER** [4] v elementu <msDescription>. To jsou elementy <settlement>, <repository> a <idNo>.

Protože toto jsou z praktického hlediska pro práci s katalogem informace nedostačující, byl ve spolupráci s Národní knihovnou České republiky stanoven tzv. minimální záznam. Ten udává *optimální minimum informací*, které by měl obsahovat platný evidenční záznam (pokud jsou ovšem tyto informace k dispozici).

Místo uložení Obsahuje město (nebo jinou sídelní jednotku), v níž je popisovaný dokument uložen, nikoli instituci, pro tento údaj slouží položka Majitel.

`msDescription/msIdentifier/settlement`

Knihovna Obsahuje instituci, v níž je popisovaný dokument uložen.

`msDescription/msIdentifier/repository`

Signatura Obsahuje signaturu, popř. jinou identifikaci dokumentu (např. v případě archivních dokumentů kromě vlastní signatury i jméno archivního fondu).

`msDescription/msIdentifier/idno`

Hlavní název Název dokumentu – v případě souboru více textů v jednom dokumentu je vhodné používat souborné označení, např. Textus varii, Právní sborník apod.

`msDescription/msHeading/title`

Autor Autor dokumentu nebo některých jeho částí. Autorem je míněn intelektuální původce textu, nikoli např. písař. Může obsahovat více jmen.

`msDescription/msHeading/author`

Rok vydání Udává dobu vzniku dokumentu. Může být zadáno jak přesným datem, tak libovolným časovým rozmezím.

`msDescription/msHeading/origDate`

Jazyk originálu Jazyk, jímž je dokument napsán. Lze zadat více jazyků.

`msDescription/msHeading/textLang`

Poznámka Libovolné další údaje, které popisovatel dokumentu uzná za vhodné uvést.

`msDescription/msHeading/note`

Obsah Umožňuje popsat obsah dokumentu krátkým, shrnujícím způsobem (např. Soubor právních textů týkajících se oblasti jihoněmeckého městského práva).

`msDescription/msContents/overview>/p`

Iluminace Umožňuje zadat libovolné informace o výzdobě rukopisu, tématicky sem náležejí např. i rytiny v tiscích. Je možné použít jak krátký shrnující popis, tak rozpis s uvedením jednotlivých folií rukopisu.

`msDescription/physDesc/decoration/decoNote/p`

Notace Položka pro údaje o notaci, pokud je v dokumentu obsažena.

`msDescription/physDesc/musicNotation/p`

Vazba Obsahuje popis vazby dokumentu.

`msDescription/physDesc/bindingDesc/binding/p`

Materiál Popisuje materiál dokumentu (obvykle papír, pergamen, kombinace obou).

`msDescription/physDesc/support/p`

Rozsah Zachycuje počet stran (folií) dokumentu včetně případných předních a zadních předsádek. Je vhodné uvádět i případné chyby ve foliaci (paginaci) – chybějící strany či vícenásobně číslovaná folia.

`msDescription/physDesc/extent`

Rozměry Umožňuje zadat rozměry jednotlivých listů (pokud jsou stejné nebo podobné) či libovolné rozmezí krajních hodnot.

`msDescription/physDesc/extent`

Literatura Umožňuje zapsat literaturu vztahující se k popisovanému dokumentu (edice, katalogy, monografie věnované konkrétním dílům, časopisecké články atd.)

`msDescription/additional/listBibl/bibl`

Rozšíření formátu pro popis starých tisků:

Místo tisku

```
<msDescription>
  <msHeading>
    <respStmt>
      <resp>printer</resp>
      <name type="place" role="printer">Místo tisku</ name>
    </respStmt>
  </msHeading>
</msDescription>
```

Jméno tiskaře

```
<msDescription>
  <msHeading>
    <respStmt>
      <resp>printer</resp>
      <name type="person" role="printer">Jméno tiskaře</ name>
    </respStmt>
  </msHeading>
</msDescription>
```

Místo vydání

```
<msDescription>
  <msHeading>
    <respStmt>
      <resp>publisher</resp>
      <name type="place" role="printer">Místo vydání</ name>
    </respStmt>
  </msHeading>
```

Jméno vydavatele

```
<msDescription>
  <msHeading>
    <respStmt>
      <resp>publisher</resp>
      <name type="person" role="printer">Jméno vydavatele</
name>
    </respStmt>
  </msHeading>
</msDescription>
```

Formální kompatibilita obrazových dat

Kompatibilitě obrazových dat je věnován samostatný dokument **[13]**.

Systém Manuscriptorium předpokládá jednu nebo více kvalitativních úrovní digitálních obrazů, ve kterých jsou po internetu zpřístupněny uživatelům. Jsou to obvykle:

NORMAL Nejvyšší kvalita, ve které se obrazy poskytují na internetu. (dig. obrazy z digitalizačního pracoviště NKP pro tuto kvalitativní úroveň mají rozlišení cca 220 dpi a jsou uloženy ve formátu JPEG). Tato obrazová kvalita je nezbytná pro použití v systému Manuscriptorium.

GALLERY Malé obrázky sloužící pro vytvoření galerie jednotlivých stránek digitalizovaného dokumentu. V Manuscriptoriu se používají obrázky ve formátu JPEG s výškou 100 pixelů (se zachováním poměru stran). Pokud tato obrazová kvalita chybí, v systému Manuscriptorium se nebude galerie vytvářet.

PREVIEW Náhledové obrázky pro navigaci ve velkém obrazu (kvalita NORMAL). V Manuscriptoriu se používají JPEG obrázky o šířce 200 pixelů (se zachováním poměru stran). Pokud tato obrazová kvalita chybí, jsou náhledové obrázky systémem vytvářeny automaticky.

Systém může poskytovat i obrazy v libovolných dalších obrazových kvalitách, pokud na ně budou uvedeny reference v popisných metadatech (MASTER+).

Základní konvence

Aby bylo možno efektivně generovat XML dokumenty pro popis digitální kopie, musí být dodržena základní pravidla pro názvy obrazových souborů a pravidla pro jejich umístění v adresářových strukturách.

Názvy adresářů a souborů smí obsahovat pouze znaky obsažené v normě ISO646

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

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

Jeden soubor s obrazem v dané kvalitě odpovídá jedné skenované jednotce originálního dokumentu, který je identifikován signaturou (např. pro rukopis je to typicky jedna strana). Všechny takovéto soubory jsou umístěny v jednom adresáři, který může mít podadresářovou strukturu. Pro přístup k digitálnímu dokumentu po internetu protokolem http odpovídá tomuto adresáři tzv. „bázová adresa digitálního dokumentu“ (URL).

Obrazové soubory odpovídající jedné skenované jednotce v různých kvalitativních úrovních musí být rozlišeny v názvu souboru nebo musí být umístěny v různých podadresářích.

Obrazové soubory jsou v názvu identifikovány foliací nebo paginací originální skenované jednotky (stránky dokumentu).

Název obrazového souboru

Pro název obrazového souboru v systému Manuscriptorium jsou přípustné dva formáty, s uvedením kvality obrazu nebo bez uvedení kvality. Pokud existuje více kvalitativních úrovní obrazů pro jednu skenovanoou jednotku, musí být v názvech souborů uvedena kvalita obrazu nebo musí být jednotlivé obrazové kvality uložený v různých podadresářích. Název obrazového souboru bez udání kvality má strukturu

```
ppppppppFFFFF.XXX
```

v názvu souboru s uvedením kvality jsou mezi prefix a foliaci/paginaci vloženy dva znaky určující kvalitativní úroveň

```
ppppQQFFFFF.XXX
```

kde

- **ppp…** je prefix jména souboru o délce 0..20 znaků, který by měl identifikovat originální dokument
- **QQ** udává obrazovou kvalitu s použitím dvou znaků. V systému Manuscriptorium jsou použity např. pro galerii G0, pro kvalitu „NORMAL“ N0 nebo N1, „PREVIEW“ má P0 atd.
- **FFFFF** (5 znaků) - foliace či paginace zleva dorovnaná znaky '0' na 5 znaků.
- **.XXX** je přípona názvu souboru (tečka + 3 znaky). Jsou povoleny přípony JPG, GIF a PNG.

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

Část rukopisu (tisku) česky	Část rukopisu (tisku) anglicky	Jméno souboru ze scanneru	FFFFF
Běžný list	Standard Sheet	0001r.JPG 0001v.JPG 00001.JPG	0001R foliace 0001V foliace 0001P paginace

Vložený list	Enclosed Sheet	ESnnn.JPG	ES001
Zpevňovací proužek	Reinforcing Strip	RSnnn.JPG	RS001
Hřbet	Spine	SP.JPG	000SP
Horní ořízka	Head Edge	HE.JPG	000HE
Boční ořízka	Side Edge	SE.JPG	000SE
Dolní ořízka	Bottom Edge	BE.JPG	000BE
Přední desky	Front Cover	FC.JPG	000FC
Přední přídeští	Front end-sheet	FS.JPG	000FS
Zadní desky	Back Cover	BC.JPG	000BC
Zadní přídeští	Back End sheet	BS.JPG	000BS
Římské číslov. přední	Front roman page	Frrrr.JPG	F001R foliace F001P paginace
Římské číslov. zadní	Back roman page	Brrrr.JPG	B001V foliace B001P paginace

Název obrazového souboru ve formátu JPEG v nejvyšší poskytované kvalitě (kterou bude přispěvatel označovat např. „EX“) pro stránku s foliací „1r“ rukopisu, který je identifikován např. prefixem „RUK1“ může vypadat takto:

```
RUK1_EX0001R.JPG
```

kde

- **RUK1_** je prefix, identifikující rukopis, ke kterému obraz patří
- **EX** označení obrazové kvality
- **0001R** identifikace stránky dokumentu pomocí foliace
- **.JPG** přípona obrazového souboru ve formátu JPG

Pokud v názvech obrazových souborů není uvedena informace o kvalitě obrazu, je třeba soubory s různými obrazovými kvalitami umístit do různých adresářů (název adresáře potom nese informaci o kvalitě obrazu v něm uložených souborů). Celá cesta vzhledem ke kořenovému adresáři pro jeden digitalizovaný fyzický dokument potom vypadá:

```
\QQ\ppppFFFFF.XXX
```

Název souboru včetně podadresáře pro kvalitu tento případ bude vypadat takto:

```
\EX\RUK1_0001R.JPG
```

Název souboru nyní neobsahuje informaci o obrazové kvalitě, ale soubor je přímo umístěn v adresáři EX, který obsahuje všechny obrazové soubory daného digitálního dokumentu pro tuto obrazovou kvalitu. Název je potom jako v předchozím případě složen z prefixu označujícího digitální dokument (RUK1_), identifikátoru digitalizované stránky (0001R) a přípony vyjadřující formát obrazu (.jpeg)

Obsahují-li názvy souborů informaci o obrazové kvalitě, potom mohou být soubory pro všechny obrazové kvality umístěny v jednom společném adresáři pro celý digitalizovaný dokument.

Přístup k obrazovým datům po internetu

Aby bylo možno sdílet systémem Manuscriptorium digitální data (obrazové soubory) od jednotlivých přispěvatelů po internetu z jejich webových serverů, musejí být dodrženy tyto konvence.

Přímý přístup

Data přispěvatelů musí být dostupná pomocí protokolu http zadáním jednoznačné adresy (URL), která může pro výše uvedený příklad vypadat např. takto:

```
http://www.memoria.cz/images/RUK1/EX/RUK1_0001R.JPG
```

Podobně, pokud jsou v názvech souborů uvedeny obrazové kvality, není nutné, obvzláště pro malé dokumenty, aby byly soubory v různých adresářích a adresa může vypadat:

```
http://www.memoria.cz/images/RUK1_EX0001R.JPG
```

Nepřímý přístup

Kromě takovéhoto přímého přístupu k souboru je umožněn i přístup k obrazům přes rozšíření webového serveru (cgi, asp, php apod.), jejichž odezvou bude předaný obrazový soubor. URL pak bude vypadat např. takto:

```
http://www.memoria.cz/system/img.cgi?DocId=RUK1_&IQ=EX&page=0001R
```

kde parametry skriptu **img.cgi** jsou:

- **DocId** identifikátor digitálního dokumentu
- **IQ** obrazová kvalita
- **Page** identifikátor stránky

Skript může mít podle potřeby další parametry, jako např. autentikační údaje v případě, že obrázky nejsou poskytovány volně apod.

Konektory

Protože praktické možnosti řešení ukládání dat jsou širší, než jsou naznačené možnosti přímého a zvláště pak nepřímého přístupu, lze dosáhnout kompatibility vybudováním individuálního konektoru zajišťujicího styk mezi Manuscriptoriem a úložištěm partnera. Tento konektor může být jak na straně Manuscriptoria, tak na straně poskatovatele.

Konektor na straně Manuscriptoria může být využíván více poskytovateli.

Využívání konektorů je vždy předmětem dvoustranných dohod mezi poskytovatelem dat a provozovatelem Manuscriptoria.

Správa datového úložiště

Správa každého datového úložiště pro uložení digitálních kopií nebo komplexních digitálních dokumentů je plně v režii provozovatele datového úložiště. Vlastník digitálních dat nese plnou zodpovědnost za správnou strukturu a obsah dat uložených na datovém úložišti. Správce datového úložiště je pracovník zodpovědný za umístění digitálních kopií dokumentu na datové úložiště a za jejich správnou strukturu a přístupnost.

Proces ověřování kompatibility

Aby bylo možno data a metadata přispěvatelů bezpečně začleňovat do systému Manuscriptorium během jeho provozu, je nezbytné zajistit jednotný proces ověřování kompatibility dodávaných dat a metadat před jejich začleněním do systému.

Proces ověřování kompatibility probíhá takto:

Přispěvatel předá provozovateli Manuscriptoria reprezentativní vzorek evidenčních záznamů a dat připravených ve smyslu tohoto dokumentu. Provozovatel provede formální a obsahovou kontrolu dodaných dat. Zjištěné nedostatky sdělí přispěvateli, který data upraví a předá znovu vzorek provozovateli. Tento proces může proběhnout vícekrát. Jeho výsledkem budou metadata přispěvatele importovatelná a akceptovatelná systémem Manuscriptorium.

Po dosažení tohoto stavu přispěvatel obdrží ceritifikát „Manuscriptorium Technical Compatible“ potvrzující technickou kompatibilitu dat a jejich využitelnost v Manuscriptoriu.
Tato kompatibilita je nutnou podmínkou pro přiznání nároku na udělení bezplatné licence na neomezené využívání celého obsahu Manuscriptoria.