

Definice metadatových formátů pro digitalizaci periodik

- vychází z obecné specifikace metadat v projektu NDK

Jan Hutař, NK ČR

VERZE 1.1 - 12.10.2011

jméno	datum	verze dokumentu	provedené změny
Jan Hutař - NK	20.9-12.10.2011	1.1	úpravy DC; povinností výskytu u některých elementů v MODS; forma zápisu data u vnitřních částí MODS; úpravy povinností plnění subelementů ALTO <preProcessingStep> <processingDateTime>; změna MD5 souborů, doplněna specifikace info.xml; nový sloupec pro povinnost elementů; <margin> v ALTO bude obsahovat text a textové bloky; upřesnění příloh; u všech tabulek přibyl sloupec povinnost; do logické mapy doplněn <div> pro přílohu; specifikace info.xml; + další drobnosti
Jan Hutař - NK	25.-7.9.2011	verze 1.0	kompletní úprava logické strukturální mapy; doplnění detailů o METS záznamu
Jan Hutař - NK	22-24.8.2011	draft 0.2	vymazání UC z celého dokumentu, úpravy textu, výměna obrázku; k vnitřní části přidán element recordInfo – nutno kvůli údajům o popisu článku; element <subject> přidán k popisu titulu, čísla, vnitřní části a přílohy; k číslu přidán abstrakt
Jan Hutař - NK	15.8.2011		vymazán element <name> na úrovni titulu; drobné úpravy
Jan Hutař - NK	21.7.2011	draft 0.1	první znění

1.	VÝCHODISKA	3
2.	VÝSTUPY DIGITALIZACE	3
3.	GRANULARITA METADATOVÉHO ZÁZNAMU	4
4.	IDENTIFIKÁTORY	4
5.	STRUKTURA PSP BALÍČKU	4
6.	NÁZVOVÁ KONVENCE SLOŽEK A SOUBORŮ	8
7.	TRANSPORTNÍ BALÍK PRO JEDEN NEBO VÍCE PSP BALÍČKŮ	10
8.	METADATA	10
8.1	<i>METS hlavička <metsHdr></i>	<i>11</i>
8.2	<i>METS část <dmdSec> - Bibliografická metadata – formát MODS a Dublin Core.....</i>	<i>11</i>
8.3	<i>METS část <amdSec> - Technická a administrativní metadata – formáty MIX a PREMIS</i>	<i>41</i>
8.4	<i>METS část <fileSec>.....</i>	<i>60</i>
8.5	<i>METS část <structMap> - Strukturální metadata a ALTO XML</i>	<i>61</i>
8.6	<i>OCR (ALTO XML a TXT OCR).....</i>	<i>69</i>

1. Východiska

- uživatelské kopie = UC
- archivní kopie = MC
- původní sken – PS – obrazový soubor vzniklý při digitalizaci, který se po zpracování (ořez, narovnání apod.), maže se a dále se neukládá
- u všech metadatových formátů budou použity verze aktuální v době implementace projektu NDK, nebo verze předchozí v případě, že nová verze je nová min. 3 měsíce
- základní intelektuální entita ve workflow digitalizace a následně i v LTP systému = číslo periodikaPSP balíček – producer submission package
 - o balíček dat a metadat, který přichází od producenta dat (tedy např. z workflow digitalizace
 - o PSP balíček bude obsahovat kompletní intelektuální entitu tj. číslo periodika
 - o z workflow digitalizace lze poslat více PSP balíčků v balíku např. [.tar] apod.
 - o pokud má dvousvazkové dílo v katalogu knihovny bibliografický záznam pro každý svazek, vznikne pro každý svazek PSP balíček a každý svazek bude brán jako jedna intelektuální entita; to samé platí i pro případ, že vícesvazkové dílo má pouze jeden záznam
- základní bibliografická metadata budou stahována přímo z knihovních katalogů do workflow digitalizace
- jako výchozí SW pro vytváření souborů JPEG2000 se bude používat Kakadu
- veškerá metadata musí pro zápis používat kodování UTF-8

2. Výstupy digitalizace

1. archivní kopie (1 MC pro každou stránku)
2. uživatelské kopie (1 UC pro každou vzniklou MC, tedy stránku)
3. OCR - ALTO XML soubor pro každou stránku
4. OCR TXT soubor - pro možnost stáhnout si jen text dokumentu (tam kde kvalita OCR je odpovídající), vyhledávání/indexace.
5. metadata pro MC
 - a. bibliografická metadata – MODS a DC
 - b. strukturální metadata – METS
 - c. technická metadata – MIX, PREMIS
 - d. administrativní metadata – PREMIS, METS
6. kontrolní metadatové soubory (s kontrolními součty a údaji o vzniku dat apod.)

Pozn.

Záznam METS nebude obsahovat žádná metadata pro uživatelské kopie. Tj. METS neobsahuje popisná, ani technická metadata pro UC. Obrazové soubory UC nejsou ani součástí strukturální mapy

<structMap> ani <fileSec>. Součástí PSP balíčku budou jen obrazy UC ve složce [userCopy]. Důvodem je to, že metadata pro UC budou vytvářena na vstupu do Krameria4 ve formátu FOXML (Fedora Object XML). Budou se vyrábět z METS záznamu pro MC, jehož specifikace je níže.

3. Granularita metadatového záznamu

Periodika

- základní intelektuální entitou periodik je 1 číslo
- každé číslo periodika má svůj vlastní metadatový záznam (=METS), který obsahuje údaje o nadřazených entitách čísla jako jsou ročník, titul periodika, tj. je pro uživatele i pro systém možné spojit jednotlivá čísla do ročníků a titulů

4. Identifikátory

Do workflow digitalizace budou přicházet bibliografická metadata, která již budou obsahovat následující identifikátory vrchních úrovní intelektuálních entit (úroveň titulu):

- ISBN – pouze pro titul monografie (jednosvazkové), nebo pro soubor monografií, které mají pouze jeden souborný záznam, ISBN není přiděleno vždy
- ISSN – pouze pro titul periodika, ISSN není přiděleno vždy (chybí např. u starých titulů z 19. století)
- ČČNB – identifikátor entity tak jak odpovídá katalogizačnímu záznamu, tj. každá entita se záznamem v katalogu NK/MZK má tento identifikátor

Nižší úrovně intelektuálních entit by měly mít také své identifikátory, ideálně URN:NBN, které bude přidělováno během digitalizace¹.

- přidělováno bude logickým úrovním (entitám)
 - o u periodik tedy: číslo, vnitřní část, příloha, případně ročník
- další možností je využití UUID

5. Struktura PSP balíčku

V kapitole je návrh struktury balení dat a metadat v jednom PSP balíčku na výstupu z workflow digitalizace.

¹ buď přímo v SW pro workflow digitalizace, nebo za pomoci aplikace Resolver URN:NBN

složka	obsahuje >	obsahuje >	obsahuje>								
svazek monografie /číslo periodika											
	info.xml	údaje o vzniku balíku									
složka [masterCopy]		obrazy JPEG2000 lossless									
složka [userCopy]		obrazy JPEG2000 lossy									
složka [ALTO]		soubory ALTO XML									
složka [TXT]		soubory OCR.TXT									
složka [amdSec]		AMD_METS.xml soubor pro každou stránku obsahuje>	<table><tr><td>amdSec</td><td>techMD = PREMISobject pro MC, původní TIFF, ALTO XML) + MIX pro MC, původní TIFF)</td></tr><tr><td></td><td>digiprovMD = PREMISevent + PREMISagent</td></tr><tr><td>fileSec</td><td>odkazuje na MC, ALTO XML, OCR TXT soubor popisované 1 stránky</td></tr><tr><td>StructMap</td><td>pouze fyzická - pro soubory popisované stránky (MC a ALTO XML, OCR TXT)</td></tr></table>	amdSec	techMD = PREMISobject pro MC, původní TIFF, ALTO XML) + MIX pro MC, původní TIFF)		digiprovMD = PREMISevent + PREMISagent	fileSec	odkazuje na MC, ALTO XML, OCR TXT soubor popisované 1 stránky	StructMap	pouze fyzická - pro soubory popisované stránky (MC a ALTO XML, OCR TXT)
amdSec	techMD = PREMISobject pro MC, původní TIFF, ALTO XML) + MIX pro MC, původní TIFF)										
	digiprovMD = PREMISevent + PREMISagent										
fileSec	odkazuje na MC, ALTO XML, OCR TXT soubor popisované 1 stránky										
StructMap	pouze fyzická - pro soubory popisované stránky (MC a ALTO XML, OCR TXT)										
Hlavní_METS.xml	dmdSec	MODS a DC pro jednotlivé úrovně dokumentu									
	fileSec	obsahuje linky na MC, ALTO XML, OCR TXT a technická metadata ve složce [amdSec]									
	structMap (včetně ALTO odkazů)	logická a fyzická pro MC, ALTO XML areas, OCR TXT a AMD_METS.xml									
MD5		kontrolní součty všech souborů v PSP balíku									

Jedná se o variantu, kdy technická a administrativní metadata nejsou obsažena v hlavním METS záznamu, ale pro každou stránku v jiném dalším METS záznamu (AMD_METS.xml). Důvodem je to, že pokud by bylo vše v hlavním METSu, byl by neúměrně dlouhý. Takto je z hlavního záznamu nalinkováno.

PSP balíček = 1 složka pro 1 číslo periodika.

Hlavní složka PSP balíčku obsahuje následující složky a soubory:

soubor info.xml

velmi krátce tu budou zaznamenány údaje o vzniku celého PSP balíčku – kdo, kdy ho vytvořil, jakou měl velikost, odkud kam byl nakopírován apod. Obsahovat by také měl informaci o stavu zpracování balíčku. Zaznamenány by také měly být údaje o obsahu PSP balíčku – počet a názvy souborů apod. Soubor info.xml by také mohl být vedle hlavního PSP balíčku. Údaje a struktura info.xml souboru:

1. vznik balíčku – datum dle ISO8601 na úroveň vteřin
2. ID balíčku – použit identifikátor čísla periodika (URN:NBN) – viz názvová konvence v kap. 6
3. ID titulu - ČČNB, ISBN nebo ISSN
4. údaje o větším celku, do kterého balíček patří - např. digitalizace pro ANL
5. název instituce, která je zadavatelem digitalizace
6. tvůrce balíčku – kód instituce (firmy), která balíček vytvořila
7. velikost balíčku – v kB
8. z jakého serveru bylo nahráno – URL
9. obsah balíčku
 - názvy souborů včetně directory path a koncovky (mime type)
 - počet souborů v balíčku celkem
10. stav zpracování – možné hodnoty
 - hotovo
 - opraveno
 - added OCR
 - added titles
 - added logical parts (issues, years)
 - updated xml (Mods, DC, identifikátory),
11. poznámka – např. o tom, že balíček neobsahuje OCR apod.

příklad balíčku, který obsahuje 2 soubory, jeden v rootu a druhý ve složce:

1. CREATED=2009-11-10T12:37:46
2. PACKAGEID=ANL_123456
3. TITLEID=ISSN1234-1236
4. COLLECTION=ANL
5. INSTITUTION=NKP
6. CREATOR=NazevFirmy
7. PACKAGESIZE=36000155kb
8. SOURCELOCATION= server123.firma.cz/baliky_hotovo/01/2011/12/000025456
9. ITEMLIST=scan01.jp2
 ITEMLIST=slozka/hotovo/27.9.2011/scan02.jp2
 ITEMTOTAL=2
10. STATUS=hotovo
11. NOTE=noOCR

složka [masterCopy]

složka s master kopiemi, obsahuje soubory JPEG2000 v neztrátové kompresi, 1 soubor = 1 stránka, tj. obsahuje všechny naskenované stránky čísla periodika

složka [userCopy]

složka s uživatelskými kopiemi, pro každou naskenovanou stránku čísla periodika obsahuje jeden JPEG2000 soubor se ztrátovou kompresí

složka [ALTO]

obsahuje ke každé stránce 1 ALTO XML soubor, tj. tolik ALTO XML souborů kolik je stránek čísla periodika.

složka [TXT]

obsahuje ke každé stránce 1 OCR soubor jako čistý text. Tj. tolik OCR.TXT souborů kolik je stránek čísla periodika.

složka [amdSec]

složka s technickými metadaty – **obsahuje pro každou naskenovanou stránku čísla časopisu 1 METS soubor (AMD_METS.xml)**. Záměrně nejsou tato metadata v hlavním METS záznamu (hlavni_METS.xml), protože ten by neúměrně narostl a bylo by obtížné s ním pracovat. Musí z něj být ovšem nalinkována (z části fileSec). Každý METS soubor AMD_METS.xml obsahuje následující části METS formátu:

- amdSec – administrativní metadata – obsahuje část
 - o technických metadat (techMD), která ve formátu PREMISobject popisuje vlastnosti archivních kopií, ALTO XML, původního TIFF souboru, ze kterého vznikly archivní kopie. Dále je přítomen záznam technických metadat v MIX formátu pro archivní kopie a pro původní TIFF.
 - o metadat o provenienci digitálních objektů (digiprovMD) – v této části je využit formát PREMISevent a PREMISagent.
 - o fileSec- sekce s odkazy na soubory – povinná část METS záznamu - v případě tohoto METS záznamu pro jednu stránku, který vzniká primárně k zachycení technických a administrativních metadat bude odkazovat na soubory, které jsou s tou konkrétní stránkou spojeny, tj. archivní kopie, ALTO XML a OCR TXT. Jde o povinnou sekci METS záznamu.
 - o structMap – **pouze fyzická** strukturální mapa, povinná část METS záznamu. Bude ukazovat strukturu souborů k dané stránce, tj. opět archivní kopie, ALTO XML a OCR TXT. Pro další mapování do LTP systému nebude potřeba.

soubor Hlavni_METS.xml

další částí PSP balíčku je hlavní METS dokument. Hlavní METS záznam tedy obsahuje:

- dmdSec – bibliografická metadata k číslu periodika včetně popisu nadřazených entit (např. ročník, titul) nebo naopak částí (např. kapitola). Základ bude z katalogu, případný další popis částí bude z digitalizace. Formátem hlavním bude MODS, nutná pro LTP je i přítomnost zkráceného záznamu v Dublin Core.
- fileSec – hlavní část s linky na všechny digitální objekty (archivní kopie, ALTO XML a OCR TXT), které se váží k jednomu číslu periodika. Obsahuje také linky na administrativní metadata AMD_METS.xml do složky [amdSec].
- structMap – strukturální mapa pro celý dokument, tj. pro jedno číslo periodika. Obsahuje:
 - o logickou část pro archivní, ALTO XML, OCR TXT a AMD_METS.xml
 - o fyzickou část pro archivní, ALTO XML, OCR TXT a AMD_METS.xml
 - o mapování na ALTO XML areas

soubor MD5

Poslední částí PSP balíčku je soubor s kontrolními součty pro celý balíček. Soubor .md5 je jeden na celou složku (jeden balíček s číslem periodika nebo svazkem monografie). Obsahuje kontrolní součet pro každý soubor obsažený v hlavní složce. Z tohoto důvodu nejsou kontrolní součty součástí podsložek balíčku. Kontrolní součty jsou také samozřejmě v technických metadatech.

6. Názvová konvence složek a souborů

pojmenování PSP balíčku

- každý PSP balíček přicházející z digitalizace by měl obsahovat pouze jedinou intelektuální entitu (číslo periodika). **Pak musí název balíčku vycházet z identifikátoru této entity, např. URN:NBN, číslo čárového kódu použitého na fyzické jednotce apod.**
- **každé číslo periodika musí mít svůj jednoznačný identifikátor, tím pádem pak každý PSP balíček a každý soubor v něm má vlastní jednoznačný identifikátor**

pojmenování složek

- viz návrh struktur PSP balíčku (kap. 5)

pojmenování souborů

- názvy jakýchkoliv souborů náležejících k jedné základní entitě (svazek nebo číslo) musí být založeny na jednom typu identifikátoru
- pro číslo periodika by takovým identifikátorem mohlo být URN:NBN, ČČNB, ISBN nebo ISSN titulu + další upřesnění (číslo výtisku apod.)
- podobně využitelným identifikátorem by mohlo být generované číslo UUID, které by se generovalo pro každý soubor. Tím by se ovšem ztratila vazba (i vizuální) na vrchní úroveň titulu i vazba na související soubory (stránka v jp2 a k ní náležející soubor ALTO XML apod.).

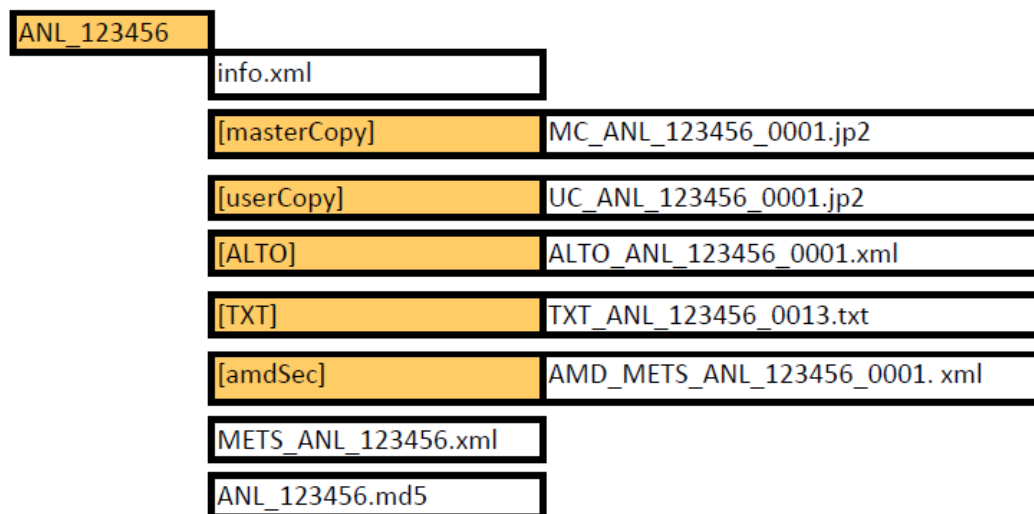
S využitím URN:NBN mohlo by to vypadat následovně (*POZOR: je použit příklad pojmenování pro projekt ANL+ - digitalizace periodik*):

typ souboru	název souboru	vysvětlení
PSP balíček (číslo, svazek)	ANL_123456	název celé složky PSP balíčku, u základních int. entit bude v názvu využito vždy URN:NBN
archivní kopie	MC_ANL_123456_0013.jp2	archivní JPEG2000 stránky 13 čísla periodika s urn:nbn:cz:anl-123456
uživatelská kopie	UC_ANL_123456_0013.jp2	uživatelská kopie ve formátu JPEG2000 stránky 13 čísla periodika s

		urn:nbn:cz:anl-123456
ALTO XML	ALTO_ANL_123456_0013.xml	ALTO soubor náležející ke 13té stránce z čísla periodika s urn:nbn:cz:anl-123456
OCR TXT	TXT_ANL_123456_0013.txt	TXT soubor s OCR náležející ke 13té stránce z čísla periodika s urn:nbn:cz:anl-123456
info.xml	INFO_ANL_123456.xml	info xml k celému PSP balíčku čísla periodika
MD5	ANL_123456.md5	soubor s kontrolními součty k celému PSP balíčku čísla periodika
Hlavni_METS.xml	METS_ANL_123456.xml	hlavní METS záznam k celému číslu periodika s urn:nbn:cz:anl-123456
AMD_METS.xml	AMD_METS_ANL_123456_0013.xml	METS záznam s technickými metadaty pro stránku 13 z čísla periodika s urn:nbn:cz:anl-123456

popis pojmenování složek viz struktura balíčku výše v kapitole 5

Složka jednoho balíčku PSP, který obsahuje jen jeden obrazový soubor k první stránce čísla periodika, pak může vypadat následovně (*příklad balíčku z digitalizace v projektu ANL+*):



7. Transportní balík pro jeden nebo více PSP balíčků

Pokud bude jeden PSP balík obsahující 1 základní intelektuální entitu (číslo periodika) přemísťován např. tar, měl by název souboru tar odpovídat názvu PSP balíčku (tedy vycházet z použitého identifikátoru pro číslo).

Výstupem workflow digitalizace ale může také být balík (např. tar), který obsahuje více PSP balíčků - toto sdružování bude omezeno jen kapacitou HW. Takovýto sdružený balík by měl být pojmenován na základě již užívaného identifikátoru.

- v případě, že balík obsahuje čísla titulu periodika, měl by název balíku vycházet z ČČNB nebo z ISSN
- v případě, že balík obsahuje svazky vícesvazkového díla, měl by název balíku vycházet z ČČNB nebo ISBN
- typ identifikátoru musí být vyjádřen v názvu souboru – např. ISSN_1234-5678.tar nebo CCNB_12345678910.tar apod.
- lze počítat s tím, že bude docházet k tomu, že sdružený balík nebude obsahovat např. všechny čísla určitého titulu periodika – tato skutečnost musí být patrná z názvu balíku (např. ISSN_1234-5678_YYYY kde YYYY může být pořadové číslo, datum, doba vzniku jednoho z více balíčků obsahujících čísla určitého titulu s identifikátorem ISSN 1234-5678).

Transportní balík by měl obsahovat následující části:

- balíčky PSP (svazků nebo čísel)
- informační soubor, který odpovídá specifikaci info.xml
- kontrolní součty všech PSP balíčků
- seznam balíčků v transportním balíku

Do úvahy mohou přijít balicí metody jako BagIt²⁾, tar apod.

8. Metadata

- veškerá metadata budou „zabalena“ pomocí kontejnerového formátu METS
- formát METS bude v aktuální verzi v době implementace nebo verzi předchozí (prosinec 2010 verze 1.9 - <http://www.loc.gov/standards/mets/mets-schemadocs.html>)
- veškerá metadata ve všech formátech musí být zapsána pomocí XML za použití kodování UTF-8
- **vložení metadatových formátů do kontejneru METS bude vždy formou <mdWrap>, tj. ne odkazováním z METS záznamu ven**

význam pole „povinnost“

- pole „povinnost“ uvádí, zda je plnění jednotlivých elementů povinné, doporučené nebo volitelné
- může nabývat následujících hodnot

² <https://confluence.ucop.edu/display/Curation/BagIt>

- M - mandatory (povinně plnit – element je součástí každého záznamu)
- MA - mandatory if available (povinně plnit pokud je to možné, pokud lze apod.)
- R - recommended (plnění hodnot elementu je doporučeno, není ovšem povinné)
- RA - recommended if available (doporučeno pokud lze plnit)
- O - optional (plnění hodnot elementu je zcela dle konkrétních potřeb)

8.1 METS hlavička <metsHdr>

- dokumentuje vznik a úpravy METS záznamu

element	atributy	popis	Povinnost
<metsHdr>	LASTMODDATE CREATEDATE	hlavička METS záznamu ----- LASTMODDATE – datum poslední úpravy záznamu, musí být ve tvaru ISO 8601 (na úroveň vteřin) CREATEDATE – datum vytvoření záznamu, musí být ve tvaru ISO 8601 (na úroveň vteřin)	M
<agent>	ROLE TYPE	údaje o tvůrci záznamu METS ----- ROLE – hodnota „CREATOR“ TYPE – hodnota „ORGANIZATION“	M
<name>		jméno jednotlivce nebo organizace; ----- tvůrce záznamu, buď dodavatel (firma XY) nebo v případě tvorby záznamu v knihovně bude využita sigla knihoven, tj. pro NK ČR hodnota „ABA001“	M

8.2 METS část <dmdSec> - Bibliografická metadata – formát MODS a Dublin Core

- každá část <dmdSec> musí mít ID a vnořený element <mdWrap> s atributy MDTYPE, MIMETYPE

element	atributy	popis	povinnost
<dmdSec>	ID	identifikátor <dmdSec> části METS záznamu ----- ID: pro <dmdSec> s popisem titulu periodika hodnota	M

		„MODSMD_TITLE“ pro záznam v MODS nebo „DCMD_TITLE“ pro záznam v Dublin Core pro <dmdSec> s popisem čísla periodika hodnota „MODSMD_ISSUE“ a „DCMD_ISSUE“ pro <dmdSec> s popisem vnitřní části periodika hodnota dle typů vnitřní části (článek, obraz) - hodnoty „MODSMD_ART“ a „DCMD_ART“ pro článek a hodnoty „MODSMD_PICT“ a „DCMD_PICT“ pro obraz pro <dmdSec> s popisem přílohy periodika hodnota „MODSMD_SUPPL“ a „DCMD_SUPPL“	
<mdWrap>	MDTYPE MIMETYPE	element obsahující vložené záznamy MODS ----- MDTYPE – hodnota „MODS“ pro záznamy v MODS, hodnota „DC“ pro záznam v Dublin Core MIMETYPE – hodnota „text/xml“	M

- na samotný bibliografický popis bude použit formát MODS, aktuální verze v době implementace, nebo verze předchozí (prosinec 2010 verze 3.4 viz <http://www.loc.gov/standards/mods/>) a formát Dublin Core (dále DC) kvalifikovaný (<http://dublincore.org/documents/dcmi-terms/>)
- DC je primárně určeno na poskytnutí dat přes OAI-PMH, bude odpovídat OAI XSD (viz http://www.openarchives.org/OAI/2.0/oai_dc.xsd) a bude se jednat o nekvalifikovaný Dublin Core
- DC bude použito, uloženo v METS apod. stejným způsobem jako formát MODS – viz struktura PSP balíčku výše
- pro vytvoření DC z MODS formátu může být použito oficiální mapování Kongresové knihovny – viz <http://www.loc.gov/standards/mods/mods-conversions.html>
- DC a MODS bude vložen v METS části dmdSec – viz struktura PSP balíčku v kap. 7
- základním zdrojem pro popisná metadata je katalog NK a MZK
- u digitalizovaných dokumentů je bibliografický popis vytvářen primárně z pohledu popisu fyzické předlohy, nejde o popis elektronického dokumentu

Periodika

- základní intelektuální entitou pro popis je číslo periodika, tj. v jednom METS záznamu, který bude obsahovat metadata a strukturu jednoho čísla periodika, budou MODS záznamy k tomuto číslu
- metadata budou popisovat následující entity:

- titul (Title)
 - číslo (Issue)
 - vnitřní část (InternalPart) – typy články (Article) a obraz (Picture)
 - příloha (Supplement)
- ad titul (Title) – MODS záznam bude obsahovat i číslo ročníku
 - ad číslo (Issue) – typy čísla jsou v elementu <genre> za použití atributu type
 - ad vnitřní část (InternalPart) - typy vnitřní části články a obraz by měly pokrýt veškerou variabilitu možností, které mohou texty a obrázky na tištěné stránce mít; bližší určení typů článku (novinky, zprávy, reklama apod.) a obrazu (fotografie, tabulka, ilustrace, graf apod.) bude možné vyjádřit pomocí atributů a výrazů kontrolovaného slovníku v elementu <genre>
 - u popisu vnitřní části je i element <recordInfo> - důvodem je to, že popis článků bude vznikat manuálně a je žádoucí o vzniku záznamu článku držet kontrolní údaje; element je volitelný
 - ad příloha (Supplement) - přílohou se rozumí volně vložená entita do jednotlivého čísla, např. mapa, obsah celého ročníku, CD/DVD apod.

Rozlišujeme 3 druhy příloh periodik:

1. příloha, která **se neskenuje**, ale chceme o ní vytvořit bibliografický záznam, dát najevo čtenáři, že existuje – to může být např. CD/DVD, pohlednice, plakát apod.
 - digitální podoba přílohy (pokud existuje) není součástí balíčku PSP čísla (Issue)
 - popis lze udělat v rámci popisu přílohy (Supplement) v MODS – viz specifikace níže
 - taková příloha není součástí logické strukturální mapy formátu METS
2. příloha podobného typu, tvaru a velikosti jako je popisované číslo periodika, která se spolu s číslem **skenuje**
 - digitální podoba přílohy je, spolu s číslem (Issue), součástí PSP balíčku čísla (Issue) a je součástí hlavního METS záznamu
 - popis lze udělat v rámci popisu přílohy (Supplement) v MODS – viz specifikace níže
 - taková příloha může mít vnitřní části (InternalPart) stejně jako číslo (Issue) a jejich text je součástí ALTO XML, které je společné pro číslo (Issue) i přílohu (Supplement)
 - **taková příloha je součástí logické strukturální mapy formátu METS**
 - **taková příloha je součástí fyzické strukturální mapy formátu METS (linky mezi jednotlivými soubory reprezentujícími stránky a popisnými metadaty)**
3. příloha odlišného typu, tvaru a velikosti než je popisované číslo periodika, která **se skenuje zvlášť** na číslo nezávisle
 - taková příloha se zpracovává z pohledu katalogizace jako „nezávislé“ periodikum, z pohledu digitalizace pak jako „nezávislý“ časopis
 - může se jednat o přílohy časopiseckého typu vycházející u různých deníků (Pátek u Lidových Novin, čtvrtletní příloha MF Dnes apod.)
 - k těmto přílohám vznikají metadata podobně jako pro jednotlivá čísla deníků nebo klasické časopisy, ovšem na původním čísle, ke kterému příloha patřila, nezávisle - tj. pro „původní“ číslo, u kterého byla příloha, vznikne 1 popis (PSP balíček s jedním

hlavním METS záznamem a ALTO XML souborem) a pro přílohu je vytvořen další 1 popis (a PSP balíček s METS záznamem), jako by šlo o běžný samostatný časopis

- příloha se pak popisuje jako číslo (Issue)
- jednotlivé MODS záznamy pro části (titul, číslo, vnitřní část a příloha) nejsou samopopisné, tj. neobsahují vždy údaje o vrchních entitách (článek neobsahuje informace o titulu apod.)
- pro každou entitu vznikne jeden MODS záznam s vlastním ID, které bude označovat i typ části (např. článek, ilustrace apod.) v případě opakování částí se bude opakovat odpovídající počet MODS záznamů
- každý MODS záznam bude uložen ve vlastní METS části <dmdSec> pomocí mdWrap
- u úrovně kde je to potřeba (vnitřní část, příloha apod.) se budou opakovat <dmdSec> části tolikrát, kolik je konkrétních částí
 - tj. v METS záznamu vznikne 1 část <dmdSec> pro bibliografický záznam titulu periodika, 1 <dmdSec> část pro bibliografický záznam čísla periodika, několik <dmdSec> částí pro vnitřní části (pro všechny články i obrázky) a odpovídající počet <dmdSec> částí pro přílohy, dle počtu příloh
 - bibliografický popis obrazů bude velmi minimalistický
- záznam periodika v katalogu – v katalozích NK a MZK existuje záznam pouze pro titul periodika, neexistují samostatné záznamy pro čísla, ročníky apod. – tj. vnitřní členění a popis musí vzniknout v digitalizaci, popis titulu periodika musí být stažen z katalogu do workflow digitalizace
- stránka se nebude popisovat, její logické i fyzické číslování i typ stránky je obsaženo ve struktuře METS dokumentu (část structMap)
 - typ stránky (Advertisement, Blank, Index aj.) budou odpovídat přesně seznamu typů z DTD periodika – viz <http://digit.nkp.cz/DigitizedPeriodicals/DTD/2.10/Periodical.xsd>
- **všechny top elementy MODS formátu jsou opakovatelné, kromě <recordInfo>**
- **všechny elementy Dublin Core jsou opakovatelné**

8.2.1 Pole MODS a Dublin Core pro jednotlivé části periodika

Obsah pole „Popis“:

- vysvětlení a příklad
- doporučené plnění tam, kde je to možné uvést
- povinnost platí pro elementy MODS tak pro elementy Dublin Core stejně
- pokud je rodičovský element např. doporučený, a dceřinný element povinný, znamená to že dceřinný element je povinný pouze tehdy pokud je použit element rodičovský

význam pole „povinnost“

- pole „povinnost“ uvádí, zda je plnění jednotlivých elementů povinné, doporučené nebo volitelné
- může nabývat následujících hodnot
 - M - mandatory (povinně plnit – element je součástí každého záznamu)

- MA - mandatory if available (povinně plnit pokud je to možné, pokud lze apod.)
- R - recommended (plnění hodnot elementu je doporučeno, není ovšem povinné)
- RA - recommended if available (doporučeno pokud lze plnit)
- O - optional (plnění hodnot elementu je zcela dle konkrétních potřeb)

Pole MODS a DC pro titul periodika

Element MODS	Atributy	Popis	povinnost	Element DC
<titleInfo>	ID	název titulu periodika pro plnění použít katalogizační záznam ----- ID musí vyjadřovat název úrovně, tj. např. „MODS_TITLE“	M	
<title>		názvová informace – název periodika hodnoty převzít z katalogu	M	<dc.title>
<subTitle>		podnázev periodika	MA	<dc.title>
<partNumber>		číslo části, např. určité řady/edice (část 1, řada B), k použití u ročenek apod.	R	<dc.description>
<partName>		jméno edice nebo speciální ediční řady, např. Hygiena. k použití u ročenek a specializovaných periodik	R	<dc.description>
<typeOfResource>		popis charakteristiky typu nebo obsahu zdroje jedna z hodnot: - text - cartographic - notated music - sound recording-musical - sound recording-nonmusical - sound recording - still image - moving image - three dimensional object - software, multimedia - mixed material	R	<dc.type>

		pro periodika a monografie hodnota text; mělo by se vyčítat z MARC21 katalogizačního záznamu z pozice 06 návěští		
<genre>		bližší údaje o typu dokumentu hodnota: title	M	<dc:type>
<originInfo>		informace o původu předlohy Poznámka: Jeden nebo více výskytů elementů se předpokládá pro vydavatele, další výskyt v případě nutnosti popsat tiskaře. Pokud je nutno vyjádřit tiskaře (pole 260 podpole „f“ a „e“ a „g“ v MARC21), je nutno element <originInfo> opakovat s atributem transliteration=“printer“ a elementy <place>, <publisher>, <dateCreated>, které budou obsahovat údaje o tiskaři. Pokud bylo za dobu vydávání více vydavatelů, nutno vzít z katalogizačního záznamu pole 260 indikátor 02 a údaje o vydavatelích opakovat.	M	
<place>		údaje o místě spojeném s vydáním, výrobou nebo původem popisovaného dokumentu	MA	<dc:coverage>
<placeTerm>	type	konkrétní určení místa, např. Praha odpovídá hodnotě z katalogizačního záznamu, pole 260, podpole „a“ ----- type – bude vždy text	MA	<dc:coverage>
<publisher>		jméno entity, která dokument vydala, vytiskla nebo jinak	MA	<dc:publisher>

		vyprodukovala odpovídá poli 260 podpoli „b“ katalogizačního záznamu v MARC21; v případě, že existovalo více vydavatelů, jsou uvedeni v poznámce v poli 500 a měli by se objevit v elementu top elementu <note>		
<dateIssued>		datum vydání předlohy, nutno zaznamenat v případě titulu roky v nichž časopis vycházel (např. 1900-1939), přebírat ve formě, jak je zapsáno v hodnotě pole v katalogu odpovídá hodnotě z katalogizačního záznamu, pole 260, podpole „c“	M	<dc:date>
<dateCreated>		datum vytvoření předlohy bude použito pouze při popisu tiskaře, viz poznámka u elementu <originInfo> odpovídá hodnotě z katalogizačního záznamu, pole 260, podpole „g“	R	
<issuance>		údaje o vydávání hodnota continuing odpovídá hodnotě uvedené návěští MARC21 na pozici 07	M	
<frequency>		údaje o pravidelnosti vydávání odpovídá údajům MARC21 v poli 310 nebo pozici 18 v poli 008	R	
<language>		údaje o jazyce dokumentu	M	
<languageTerm>	type authority	přesné určení jazyka – kódem nutno použít kontrolovaný slovník ISO 639-2, http://www.loc.gov/standards/iso639-2/php/code_list.php	M	<dc:language>

		----- type: použít hodnotu code authority: použít hodnotu „iso639-2b“		
<physicalDescription>		obsahuje údaje o fyzickém popisu zdroje/předlohy	M	
<form>	authority	údaje o fyzické podobě dokumentu, např. print, electronic apod. pro periodika hodnota print odpovídá hodnotám pozice 23 a 29 v poli 008 MARC21 ----- authority: hodnota „marcform“	M	<dc:format>
<extent>		údaje o rozsahu (stran, svazků nebo rozměrů); použití spíše u ročenek apod. odpovídá hodnotám v poli 300 podpolích „a“ a „c“ MARC21, pokud jsou vyplněna obě pole, bude se element <extent> opakovat	RA	<dc:format>
<note>		poznámka o fyzickém stavu dokumentu; pro každou poznámku je nutno vytvořit nový <note> element	RA	
<abstract>		shrnutí obsahu periodika jako celku odpovídá poli 520 MARC21	R	<dc:description>
<note>		obecná poznámka k periodiku jako celku odpovídá poli 500 v MARC21	RA	<dc:description>
<subject>	authority	údaje o věcném třídění předpokládá se přebírání z katalogizačního záznamu ----- authority: vyplnit hodnotu „czenas“	R	

<topic>		libovolný výraz specifikující nebo charakterizující obsah vnitřní části; použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (věcné téma) nebo obsah pole 650 záznamu MARC21	M	<dc:subject>
<geographic>		geografické věcné třídění použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (geografický termín) nebo obsah pole 651 záznamu MARC21	R	<dc:subject>
<temporal>		chronologické věcné třídění použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (chronologický údaj) nebo obsah pole 648 záznamu MARC21	R	<dc:subject>
<name>		jméno použité jako věcné záhlaví použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (jméno osobní) nebo obsah pole 600 záznamu MARC21	R	<dc:subject>
<classification>	authority	klasifikační údaje věcného třídění podle Mezinárodního desetinného třídění povinné odpovídá poli 080 MARC21 ----- authority: vyplnit hodnotu „udc“		<dc:subject>
<relatedItem>	type	informace o dalších dokumentech/částech/zdrojích, které jsou ve vztahu k popisovanému dokumentu; použití pro vyjádření edice, ve které je dokument vydán, údaj o edici musí obsahovat minimálně element <title> s jejím názvem Poznámka: element <relatedItem> může obsahovat jakýkoliv jiný element	RA	

		MODS – jejich použití se řídí pravidly popsanými pro tyto elementy; ----- type: hodnota „series“		
<identifier>	type	údaje o identifikátorech, obsahuje unikátní identifikátory mezinárodní nebo lokální, které titul periodika má – viz přehled typů atributů níže ----- type: budou se povinně vyplňovat následující hodnoty, pokud existují: - doi - hdl - handle - issn - převzít z katalogizačního záznam NK ČR - isbn - převzít z katalogizačního záznam NK ČR - ccnb – čČNB - převzít z katalogizačního záznam NK ČR - permalink záznamu z katalogu NK ČR, např. http://aleph.nkp.cz/F/?func=direct&doc_number=002186258&local_base=NKC - uuid - jiný interní identifikátor, hodnota atributu „local“, lze použít např. k vyjádření čárového kódu	M	<dc:identifier>
<location>		údaje o uložení popisovaného dokumentu, např. signatura, místo uložení apod. pozn. u dokumentů v digitální podobě není možné vyplnit	MA	
<physicalLocation>	authority	údaje o instituci, kde je fyzicky uložen popisovaný dokument, např. NK ČR	M	<dc:source>

		nutno použít kontrolovaný slovník – sigly knihoven (ABA001 atd.) odpovídá poli 040 v MARC21 ----- authority: hodnota „siglaADR“		
<shelfLocator>		sigla nebo lokační údaje o dokumentu	M	<dc:source>
<part>	type	popis částí dokumentu, bude využit jen na popis ročníku (volume) periodika ----- type: hodnota bude vždy „volume“	M	
<detail>	type	upřesnění popisu části ----- type: hodnota bude vždy „volume“	M	
<number>		číslo části (ročníku)	MA	<dc:description> povinné pokud lze uvést; nutno doplnit slovo „volume number“, viz <dc:description> volume number: 25 </dc:description>
<caption>		text před číslem ročníku, např. „ročník“, „roč.“, „volume“ apod.	O	
<date>		datum vztahující se k části v případě, že se ročník vycházel během více let (přelom roku), nutno uvést oba roky, např. 1920-1921	MA	
<recordInfo>		údaje o metadatovém záznamu – jeho vzniku, změnách apod.	M	
<recordContentSource>		kód nebo jméno instituce, která záznam vytvořila nebo změnila; nutno vytvořit kontrolovaný slovník	R	
<recordCreationDate>	encoding	datum prvního vytvoření záznamu,	M	

		na úroveň minut ----- encoding: záznam bude podle normy ISO 8601 na úroveň minut, hodnota atributu tedy iso8601		
<recordChangeDate>	encoding	datum změny záznamu ----- encoding: záznam bude podle normy ISO 8601 na úroveň minut, hodnota atributu tedy iso8601	R	
<recordOrigin>		údaje o vzniku záznamu hodnoty: machine generated nebo human prepared	R	

Pole MODS a DC pro číslo periodika

Element MODS	Atributy	Popis	povinnost	Element DC
<titleInfo>	ID	název titulu periodika, kterého je číslo součástí, převzít z katalogizačního záznamu titulu periodika použít názvové authority nebo katalogizační záznam ----- ID musí vyjadřovat název úrovně, tj. např. „MODS_ISSUE“	M	
<title>		názvová informace – titul periodika převzít z katalogu	M	<dc:title>
<subTitle>		podnázev periodika	RA	<dc:title>
<partNumber>		pořadové číslo vydání (čísla), např. 40; nebo u ročenek číslo určité řady/edice (část 1, řada B)	MA	<dc:description>
<partName>		jméno edice nebo speciální ediční řady, např. Hygiena; lze uvést i název tématického čísla nebo zvláštního vydání; k použití u ročenek a specializovaných periodik nebo u	R	<dc:description>

		tematických čísel nebo zvláštních vydání		
<name>	type	údaje o odpovědnosti za číslo periodika nepočítá se s vyplněním u deníků, ale např. u ročenek, zvláštních vydání čísel periodika apod., které mají vlastního autora/editora ----- type: použít jeden z typů - personal - corporate - conference - family	MA	
<namePart>	type	údaje o křestním jméně a příjmení apod. nutno vyjádřit pro křestní jméno i příjmení ----- type: použít jednu z hodnot: - date – doporučené pokud lze uvést - family – povinné pokud lze uvést - given – povinné pokud lze uvést - termsOfAddress – doporučené pokud lze uvést pokud nelze rozlišit křestní jméno a příjmení, nepoužije se type a jméno se zaznamená v podobě jaké je do jednoho elementu <namePart>	MA	<dc:creator> nutno do jednoho pole DC spojit jméno i příjmení
<role>		specifikace role osoby nebo organizace uvedené v elementu <name>	MA	
<roleTerm>	type	popis role	MA	

	authority	nutno použít kontrol. slovník např. z MARC21 ----- type: code – kód role z kontrolovaného slovníku rolí (http://www.loc.gov/marc/relators/relaterm.html) authority – údaje o kontrolovaném slovníku využitém k popisu role, k popisu výše uvedeného MARC seznamu nutno uvést authority="marcrelator"		
<genre>	type	bližší údaje o typu dokumentu hodnota: issue ----- type: pro upřesnění typu čísla a jednotlivých vydání povinné hodnota může být: - normal - běžné vydání - morning – ranní vydání - afternoon- odpolední vydání - evening – večerní vydání - sequence_X – pořadí vydání (sequence_1 = první vydání toho dne; sequence_2 = druhé vydání atd.) - corrected – opravené vydání - special – zvláštní vydání (např. k nějaké události) - supplement – v případě, že se příloha časopiseckého typu popisuje jako číslo	M	<dc:type>
<originInfo>		informace o původu předlohy doporučené kde lze vyplnit (např. u ročenek, kde se vydavatel měnil)	RA/O	

		nepovinné pro deníky a běžná čísla periodik Poznámka: Jeden nebo více výskytů elementů se předpokládá pro vydavatele, další výskyt v případě nutnosti popsat tiskaře. Pokud je nutno vyjádřit tiskaře (pole 260 podpole „f“ a „e“ a „g“ v MARC21), je nutno element <originInfo> opakovat s atributem transliteration="printer" a elementy <place>, <publisher>, <dateCreated>, které budou obsahovat údaje o tiskaři.		
<place>		údaje o místě spojeném s vydáním, výrobou nebo původem popisovaného dokumentu	MA	<dc:coverage>
<placeTerm>	type	konkrétní určení místa, např. Praha odpovídá hodnotě z katalogizačního záznamu, pole 260, podpole „a“ ----- type – bude vždy text	MA	<dc:coverage>
<publisher>		jméno entity, která dokument vydala, vytiskla nebo jinak vyprodukovala odpovídá poli 260 podpoli „b“ katalogizačního záznamu v MARC21	MA	<dc:publisher>
<dateIssued>	qualifier	datum vydání předlohy, v případě čísla datum dne, kdy vyšlo; musí vyjádřit den, měsíc a rok, dle toho jaké údaje jsou k dispozici nutno zapsat v následujících podobách: - DD.MM.RRRR – pokud	MA	<dc:date>

		víme den, měsíc i rok vydání - MM.RRRR – pokud víme jen měsíc a rok vydání - RRRR – pokud víme pouze rok - DD.-DD.MM.RRRR – vydání pro více dní - MM.-MM.RRRR – vydání pro více měsíců ----- qualifier – možnost dalšího upřesnění, hodnota „approximate“ pro data, kde nevíme přesný údaj		
<dateCreated>	qualifier	datum vytvoření předlohy bude použito pouze při popisu tiskaře, viz poznámka u elementu <originInfo> odpovídá hodnotě z katalogizačního záznamu, pole 260, podpole „g“ ----- qualifier – možnost dalšího upřesnění, hodnota „approximate“ pro data, kde nevíme přesný údaj	R	
<language>		údaje o jazyce dokumentu	M	
<languageTerm>	type authority	přesné určení jazyka – kódem nutno použít kontrolovaný slovník ISO 639-2, http://www.loc.gov/standards/iso639-2/php/code_list.php ----- type: použít hodnotu code authority: použít hodnotu „iso639-2b“	M	<dc:language>
<physicalDescription>		obsahuje údaje o fyzickém popisu zdroje/předlohy	M	
<extent>		údaje o rozsahu (stran, svazků nebo rozměrů); použití spíše u	RA	<dc:format>

		ročenek apod. odpovídá hodnotám v poli 300 podpolích „a“ a „c“ MARC21, pokud jsou vyplněna obě pole, bude se element <extent> opakovat; počet stránek bude vyjádřen ve fyzické strukturální mapě a bude tak vidět v aplikaci zpřístupnění i bez vyplnění tohoto pole		
<note>		poznámka o fyzickém stavu dokumentu; pro každou poznámku je nutno vytvořit nový <note> element	RA	
<abstract>		shrnutí obsahu dokumentu, zvláště pro ročenky, zvláštní vydání a tematická čísla plnit pouze v případech, že se liší od abstraktu na úrovni titulu odpovídá poli 520 MARC21	RA	<dc:description>
<note>		obecná poznámka k dokumentu odpovídá poli 500 v MARC21	RA	
<subject>	authority	údaje o věcném třídění plnit pouze pro tematická čísla, zvláštní vydání a ročenky – pouze pokud se liší od údajů v elementu <subject> na úrovni titulu ----- authority: vyplnit hodnotu „czenas“	RA	
<topic>		libovolný výraz specifikující nebo charakterizující obsah čísla; použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (věcné téma)	M	<dc:subject>
<geographic>		geografické věcné třídění použít kontrolovaný slovník - např. z báze autorit AUT NK ČR	R	<dc:subject>

		(geografický termín)		
<temporal>		chronologické věcné třídění použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (chronologický údaj)	R	<dc:subject>
<name>		jméno použité jako věcné záhlaví použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (jméno osobní)	R	<dc:subject>
<identifier>	type	údaje o identifikátorech čísla, obsahuje unikátní identifikátory mezinárodní nebo lokální ----- type: budou se povinně vyplňovat následující hodnoty, pokud existují: - doi - hdl - handle - isbn - převzít z katalogizačního záznam NK ČR (ročenky apod.) - urnnbn - pro URN:NBN, např. zápis ve tvaru urn:nbn:cz:anl-123456 pro projekt ANL+; pozor, musí odpovídat URN:NBN, podle kterého je pojmenovaný PSP balíček a jeho jednotlivé soubory - uuid - jiný interní identifikátor, hodnota atributu „local“, lze použít např. k vyjádření čárového kódu	M	<dc:identifier>
<location>		údaje o uložení popisovaného dokumentu, např. signatura, místo uložení apod. doporučené např. pro ročenky apod., kde se signatury jednotlivých čísel liší	R	
<physicalLocation>	authority	údaje o instituci, kde je fyzicky uložen popisovaný dokument,	MA	<dc:source>

		např. NK ČR nutno použít kontrolovaný slovník – sigly knihoven (ABA001 atd.) odpovídá poli 040 v MARC21 ----- authority: hodnota „siglaADR“		
<shelfLocator>		sigla nebo lokační údaje o dokumentu	MA	<dc:source>
<part>	type	popis částí dokumentu, bude využit jen na zaznamenání <caption> ----- type: hodnota bude vždy „issue“	O	
<caption>		text před označením čísla, např. „č.“, „číslo“, „No.“ apod.	RA	

Pole MODS a DC pro vnitřní část periodika (článek a obraz)

Element MODS	Atributy	Popis	Povinnost	Element DC
<titleInfo>	ID	názvová informace vnitřní části ----- ID musí vyjadřovat název úrovně, tj. např. „MODS_PICTURE“ pro obrázek v textu, „MODS_ARTICLE“ pro článek apod.	M	
<title>		vlastní název vnitřní části (článku, obrazu); u obrazu brát případně z popisku obrazu; pokud není titul, nutno vyplnit hodnotu „untitled“	M	<dc:title>
<subTitle>		podnázev vnitřní části (článku); za podnázev lze považovat i krátký text, který se před článkem objevuje tučným písmem (shrnutí obsahu článku)	MA	<dc:title>
<partNumber>		číslo vnitřní části např. článek na pokračování	RA	<dc:title>
<partName>		název pokračování vnitřní části (článku)	RA	<dc:title>
<name>	type	údaje o odpovědnosti za vnitřní	MA	

		část (článek i obraz) ----- type: použít jeden z typů - personal - corporate - conference - family		
<namePart>	type	údaje o křestním jméně a příjmení apod. nutno vyjádřit pro křestní jméno i příjmení ----- type: použít jednu z hodnot: - date – doporučené pokud lze uvést - family – povinné pokud lze uvést - given – povinné pokud lze uvést - termsOfAddress – doporučené pokud lze uvést pokud nelze rozlišit křestní jméno a příjmení, nepoužije se type a jméno se zaznamená v podobě jaké je do jednoho elementu <namePart>	MA	<dc:creator> nutno do jednoho pole DC spojit jméno i příjmení
<role>		specifikace role osoby nebo organizace uvedené v elementu <name>	RA	
<roleTerm>	type authority	popis role nutno použít kontrol. slovník např. z MARC21 ----- type: code – kód role z kontrolovaného slovníku rolí http://www.loc.gov/marc/relators/relaterm.html) authority – údaje o kontrolovaném	MA	

		slovníku využitém k popisu role, k popisu výše uvedeného MARC seznamu nutno uvést authority="marcrelator"		
<genre>	type	bližší údaje o typu vnitřní části povinné hodnota: article nebo picture ----- type: doporučené hodnota pro article – možnost vyplnit bližší určení typu článku (možnost použít DTD periodika, Article Types) - news - table of content - advertisement - abstract - introduction - review - dedication - bibliography - editorsNote - preface - main article - index (použije se pro všechny typy seznamů mimo hlavní obsah; např. seznam obrazů, tabulek apod.) - aj. hodnota pro picture – možnost vyplnit další určení typu obrazu - table - illustration - chart - photograph - graphic - map	M	<dc:type>

		- advertisement - cover - aj.		
<language>		údaje o jazyce vnitřní části nelze plnit u obrazu	MA	
<languageTerm>	type authority	přesné určení jazyka – kódem nutno použít kontrolovaný slovník ISO 639-2, http://www.loc.gov/standards/iso639-2/php/code_list.php nelze plnit u obrazu ----- type: použít hodnotu code authority: použít hodnotu „iso639-2b“	M	<dc:language>
<physicalDescription>		obsahuje údaje o fyzickém popisu zdroje/předlohy; určeno spíše pro články než pro obrazy	R	
<form>	authority	údaje o fyzické podobě vnitřní části, např. print, electronic apod. odpovídá hodnotám pozice 23 a 29 v poli 008 MARC21 ----- authority: hodnota „marcform“	R	<dc:format>
<abstract>		shrnutí obsahu vnitřní části	R	<dc:description> >
<note>		obecná poznámka k vnitřní části do poznámky by se měla dávat šifra autora vnitřní části, která se vyskytuje pod vnitřní částí odpovídá poli 500 v MARC21	RA	<dc:description> >
<subject>		údaje o věcném třídění	R	
<topic>	authority (volitelné)	libovolný výraz specifikující nebo charakterizující obsah vnitřní části; lze (není ovšem nutno) použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (věcné téma)	M	<dc:subject>

		----- při použití autoritních záznamů použít AUT NK ČR a atribut authority: vyplnit hodnotu „czenas“; při použití volných klíčových slov atribut authority nepoužívat		
<geographic>	authority	geografické věcné třídění použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (geografický termín) ----- authority: vyplnit hodnotu „czenas“	R	<dc:subject>
<temporal>	authority	chronologické věcné třídění použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (chronologický údaj) ----- authority: vyplnit hodnotu „czenas“	R	<dc:subject>
<name>	authority	jméno použité jako věcné záhlaví použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (jméno osobní) ----- authority: vyplnit hodnotu „czenas“	R	<dc:subject>
<classification>	authority	klasifikační údaje věcného třídění podle Mezinárodního desetinného třídění plnit pouze pro článek odpovídá poli 080 MARC21 ----- authority: vyplnit hodnotu „udc“	RA	<dc:subject>
<identifier>	type	údaje o identifikátorech, obsahuje unikátní identifikátory mezinárodní nebo lokální, které vnitřní část má – viz přehled typů atributů níže ----- type: budou se povinně vyplňovat	M	<dc:identifier> povinné

		následující hodnoty, pokud existují pro článek nebo obraz: - doi - hdl - handle - urnnbn - pro URN:NBN - uuid - jiný interní identifikátor, hodnota atributu „local“, lze použít např. k vyjádření čárového kódu		
<part>		popis částí vnitřní části, bude využito na záznam rozsahu nelze u obrazu	RA	
<extent>		upřesnění popisu části – rozsah na stránkách	MA	<dc:format>
<start>		první stránka, na které vnitřní část začíná	MA	<dc:coverage>
<end>		poslední stránka, na které vnitřní část končí	MA	<dc:coverage>
<recordInfo>		údaje o metadatovém záznamu vnitřní části – jeho vzniku, změnách apod.	M	
<recordContentSource>		kód nebo jméno instituce, která záznam vytvořila nebo změnila; nutno vytvořit kontrolovaný slovník	R	
<recordCreationDate>	encoding	datum prvního vytvoření záznamu vnitřní části ----- encoding: záznam bude podle normy ISO 8601 na úroveň minut, hodnota atributu tedy iso8601	M	
<recordChangeDate>	encoding	datum změny záznamu vnitřní části ----- encoding: záznam bude podle normy ISO 8601 na úroveň minut, hodnota atributu tedy iso8601	R	
<recordOrigin>		údaje o vzniku záznamu vnitřní části hodnoty: machine generated nebo human prepared	R	

Pole MODS a DC pro přílohu

- pozn. – přílohou se myslí příloha čísla periodika, která je jeho součástí a skenuje se s ním

Element MODS	Atributy	Popis	Povinnost	Element DC
<titleInfo>	ID	názvová informace přílohy použít názvové autority nebo katalogizační záznam ----- ID musí vyjadřovat název úrovně, tj. „MODS_SUPPLEMENT“	M	
<title>		názvová informace – název periodika, jehož součástí příloha je převzít z katalogu	M	<dc:title>
<partNumber>		číslo přílohy, pokud nějaké má doporučené pokud lze vyplnit	M	<dc:description>
<partName>		název přílohy	MA	<dc:title>
<name>	type	údaje o odpovědnosti za přílohu ----- type: použít jeden z typů - personal - corporate - conference - family	MA	
<namePart>	type	údaje o křestním jméně a příjmení apod. nutno vyjádřit pro křestní jméno i příjmení ----- type: použít jednu z hodnot: - date – doporučené pokud lze uvést - family – povinné pokud lze uvést - given – povinné pokud lze uvést - termsOfAddress – doporučené pokud lze uvést pokud nelze rozlišit křestní jméno a	MA	<dc:creator> nutno do jednoho pole DC spojit jméno i příjmení

		příjmení, nepoužije se type a jméno se zaznamená v podobě jaké je do jednoho elementu <namePart>		
<role>		specifikace role osoby nebo organizace uvedené v elementu <name>	MA	
<roleTerm>	type authority	popis role nutno použít kontrol. slovník např. z MARC21 ----- type: code – kód role z kontrolovaného slovníku rolí http://www.loc.gov/marc/relators/relaterm.html) authority – údaje o kontrolovaném slovníku využitém k popisu role, k popisu výše uvedeného MARC seznamu nutno uvést authority="marcrelator"	MA	
<typeOfResource>		popis charakteristiky typu nebo obsahu přílohy jedna z hodnot: - text – např. pro přílohu typu časopis, kniha, brožura apod. - cartographic – pro mapy - notated music - sound recording-musical - pro hudební CD/DVD - sound recording-nonmusical - sound recording - still image – fotografie, plakáty apod. - moving image – pro filmová DVD - three dimensional object - software, multimedia – pro	R	<dc:type>

		CD/DVD se SW - mixed material		
<genre>		blížejší údaje o typu dokumentu hodnota: - volume_supplement (příloha k ročníku, např. obsah celého ročníku) - issue_supplement (příloha k číslu)	M	<dc:type>
<originInfo>		informace o původu přílohy plnit pokud <i>se liší od údajů v popisu čísla (platí i pro jednotlivé sub- elementy)</i> Poznámka: Jeden nebo více výskytů elementů se předpokládá pro vydavatele, další výskyt v případě nutnosti popsat tiskaře. Pokud je nutno vyjádřit tiskaře (pole 260 podpole „f“ a „e“ a „g“ v MARC21), je nutno element <originInfo> opakovat s atributem transliteration=“printer“ a elementy <place>, <publisher>, <dateCreated>, které budou obsahovat údaje o tiskaři.	MA	
<place>		údaje o místě spojeném s vydáním, výrobou nebo původem přílohy	MA	<dc:coverage>
<placeTerm>	type	konkrétní určení místa, např. Praha odpovídá hodnotě katalogizačního záznamu, pole 260, podpole „a“ ----- type – bude vždy text	MA	<dc:coverage>
<publisher>		jméno entity, která přílohu vydala, vytiskla nebo jinak vyprodukovala odpovídá poli 260 podpoli „b“ katalogizačního záznamu v MARC21	MA	<dc:publisher>

<dateIssued>	qualifier	datum vydání přílohy, musí vyjádřit den, měsíc a rok, dle toho jaké údaje jsou k dispozici nutno zapsat v následujících podobách: - DD.MM.RRRR – pokud víme den, měsíc i rok vydání - MM.RRRR – pokud víme jen měsíc a rok vydání - RRRR – pokud víme pouze rok - DD.-DD.MM.RRRR – vydání pro více dní - MM.-MM.RRRR – vydání pro více měsíců možno použít hodnotu z katalogizačního záznamu, pole 260, podpole „c“ ----- qualifier – možnost dalšího upřesnění, hodnota „approximate“ pro data, kde nevíme přesný údaj	M	<dc:date>
<dateCreated>	qualifier	datum vytvoření přílohy bude použito pouze při popisu tiskaře, viz poznámka u elementu <originInfo> nebo např. u popisu CD/DVD apod. odpovídá hodnotě z katalogizačního záznamu, pole 260, podpole „g“ ----- qualifier – možnost dalšího upřesnění, hodnota „approximate“ pro data, kde nevíme přesný údaj	R	
<frequency>		údaje o pravidelnosti vydávání odpovídá údaji MARC21 v poli 310	RA	

		nebo pozici 18 v poli 008		
<language>		údaje o jazyce dokumentu	M	
<languageTerm>	type authority	přesné určení jazyka – kódem nutno použít kontrolovaný slovník ISO 639-2, http://www.loc.gov/standards/iso639-2/php/code_list.php ----- type: použít hodnotu code authority: použít hodnotu „iso639-2b“	M	<dc:language>
<physicalDescription>		obsahuje údaje o fyzickém popisu zdroje/předlohy	M	
<form>	authority	údaje o fyzické podobě dokumentu, např. print, electronic apod. povinné pro tištěné předlohy hodnota „print“, pro elektronické přílohy „electronic“ odpovídá hodnotám pozice 23 a 29 v poli 008 MARC21 ----- authority: hodnota „marcform“	M	<dc:format>
<extent>		údaje o rozsahu (stran, svazků nebo rozměrů) odpovídá hodnotám v poli 300 podpolích „a“ a „c“ MARC21, pokud jsou vyplněna obě pole, bude se element <extent> opakovat	RA	<dc:format>
<note>		poznámka o fyzickém stavu dokumentu; pro každou poznámku je nutno vytvořit nový <note> element	RA	
<abstract>		shrnutí obsahu dokumentu odpovídá poli 520 MARC21	RA	<dc:description> >
<note>		obecná poznámka k dokumentu odpovídá poli 500 v MARC21	RA	<dc:description> >

<subject>	authority	údaje o věcném třídění ----- authority: vyplnit hodnotu „czenas“	R	
<topic>		libovolný výraz specifikující nebo charakterizující obsah přílohy; použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (věcné téma)	M	<dc:subject>
<geographic>		geografické věcné třídění použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (geografický termín)	R	<dc:subject>
<temporal>		chronologické věcné třídění použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (chronologický údaj)	R	<dc:subject>
<name>		jméno použité jako věcné záhlaví použít kontrolovaný slovník - např. z báze autorit AUT NK ČR (jméno osobní)	R	<dc:subject>
<classification>	authority	klasifikační údaje věcného třídění podle Mezinárodního desetinného třídění odpovídá poli 080 MARC21 ----- authority: vyplnit hodnotu „udc“	M	<dc:subject>
<identifier>	type	údaje o identifikátorech, obsahuje unikátní identifikátory mezinárodní nebo lokální, které příloha má – viz přehled typů atributů níže ----- type: budou se povinně vyplňovat následující hodnoty, pokud existují: - doi - hdl - handle - issn - převzít z katalogizačního záznam NK ČR - isbn - převzít z katalogizačního záznam	MA	<dc:identifier>

		NK ČR - ccnb – čČNB - převzít z katalogizačního záznam NK ČR - permalink záznamu z katalogu NK ČR, např. http://aleph.nkp.cz/F/?func=direct&doc_number=002186258&local_base=NKC - urnnbn - pro URN:NBN - uuid - jiný interní identifikátor, hodnota atributu „local“, lze použít např. k vyjádření čárového kódu		
--	--	---	--	--

8.3 METS část <amdSec> - Technická a administrativní metadata – formáty MIX a PREMIS

- technická a administrativní metadata pro různé reprezentace jedné strany čísla periodika (původní TIFF, MC, ALTOXML a OCR.TXT) budou zabalena v části <amdSec> vedlejšího METS záznamu (AMD_METS.xml) ve vlastních formátech (MIX, PREMIS – části object; events; agent)
- **pro všechny reprezentace jedné strany čísla periodika bude v AMD_METS.xml existovat jedna část <amdSec>, která bude obsahovat metadata v <techMD> a <digiprovMD> podčástech pro jednotlivé soubory**
- část <amdSec> musí mít ID a vnořený element <techMD> nebo <digiprovMD>, oba s ID a vnořeným elementem <mdWrap> s atributem MDTYPE

element	atributy	popis	Povinnost
<amdSec>	ID	element obsahující technická metadata ve formátu PREMIS nebo MIX ----- ID – identifikátor konkrétní části <amdSec>, např. pro stránku 1 by hodnota mohla být „PAGE0001“	M
<techMD> nebo <digiprovMD>	ID	element rozlišující typy jednotlivých administrativních metadat ----- ID pro část <techMD>: - pro části obsahující PREMIS-object hodnota	M

		„OBJ_001“ – objekt 1 (PREMIS object pro smazaný TIFF, OBJ_002 by bylo pro MC, OBJ_003 pro ALTO XML - pro části obsahující MIX hodnota „MIX_001“ = MIX metadata pro původní TIFF, „MIX_002“ pro MC pro část <digiprovMD>: - pro části obsahující PREMIS-event hodnota „EVT_001“ apod. - pro části obsahující PREMIS-agent hodnota „AGENT_001“ apod.	
<mdWrap>	MDTYPE	element obsahující vložené záznamy PREMIS, MIX ----- MDTYPE - pro záznamy PREMIS object, event i agent vždy hodnota „PREMIS“ - pro záznamy MIX hodnota „NISOIMG“	M

- pro všechna digitalizovaná data se bude využívat formát PREMIS (jeho části object, event a agent), pro obrazová data dále i formát MIX
- technická a administrativní metadata budou vznikat i pro prvotní sken (většinou TIFF), který se po nutných úpravách maže a dále neuchovává – viz specifikace
- technická metadata jsou určena primárně pro zachycení technických informací o formátech souborů, o výsledcích validací a kontrol
- administrativní metadata zachycují veškeré změny, procesy apod., které byly na datech i metadatach provedeny
- všechny PREMIS a MIX záznamy budou obsaženy v tzv. vedlejším METS záznamu (AMD_METS.xml), který je určen pro administrativní a technická metadata
 - o celý METS záznam (AMD_METS.xml) a je linkován z hlavního METS záznamu dokumentu
- **plnění technických metadat se předpokládá z výstupů vzniklých využitím služeb třetích stran jako jsou JHOVE2, PRONOM aj.)**

8.3.1 PREMIS Objects

- bude odpovídat poslední aktuální verzi v době implementace (leden 2011 - PREMIS data dictionary v. 2.1), nebo verzi předchozí
- popisovat se pomocí PREMIS object budou soubory, tj. dle specifikace PREMIS vždy úroveň tzv. File (ne reprezentace ani bitstream)
- záznam v PREMIS object se bude vytvářet pro každý soubor 1) vzniklý v procesu digitalizace (původní sken, který se dále maže; 2) archivní obrazové kopie, 3) ALTO XML

- PREMIS object se nebude vytvářet pro OCR.TXT soubory
- pro každý záznam PREMIS object bude existovat vlastní podčást <techMD>
- záznam PREMIS Object pro jeden soubor bude obsahovat linky na eventy, které jsou popsány v PREMIS Events ve stejném METS metadatovém záznamu konkrétního dokumentu (číslo, svazek) v části <digiprovMD>; přes <premis:relatedEventIdentification>, to samé platí pro objekty, které budou nalinkovány v případě vztahu (např. UC vznikla z MC) s popisovaným objektem přes <premis:relatedObjectIdentification>.
 - o tj. např. PREMIS object popisující archivní soubor JPEG2000 je tímto způsobem nalinkován na původní sken ve formátu TIFF (resp. na jeho PREMIS object záznam) – pomocí tagu <relatedObjectIdentification>, který obsahuje ID původního objektu (např. TIFF)
 - o zároveň pomocí tagu <relatedEventIdentification> je záznam PREMIS object archivního souboru JPEG2000 nalinkován na událost, během které vznikl
- **POZOR – Premis Object bude vznikat a uchovávat se i pro neexistující data (původní a posléze smazaný TIFF)**

Pole záznamu PREMIS Object

Obsah pole „Popis“:

- vysvětlení a příklad
- doporučené plnění tam, kde je to možné
- výskyt elementu (jak je definováno formátem PREMIS – dle XSD)
 - o 0-1 element je nepovinný, neopakovatelný
 - o 0-n element je nepovinný, opakovatelný
 - o 1-n element je povinný a opakovatelný
 - o element je povinný a neopakovatelný

význam pole „povinnost“

- pole „povinnost“ uvádí, zda je plnění jednotlivých elementů povinné, doporučené nebo volitelné
- může nabývat následujících hodnot
 - o M - mandatory (povinně plnit – element je součástí každého záznamu)
 - o MA - mandatory if available (povinně plnit pokud je to možné, pokud lze apod.)
 - o R - recommended (plnění hodnot elementu je doporučeno, není ovšem povinné)
 - o RA - recommended if available (doporučeno pokud lze plnit)
 - o O - optional (plnění hodnot elementu je zcela dle konkrétních potřeb)

Obsah pole „Použití pro“

- použití jednotlivých elementů pro popis MC, PS (původní sken), XML (ALTO)

Element	Popis	Povinnost	Použití pro
<objectIdentifier>	identifikátor k jednoznačnému odlišení objektu	M	MC,

	v určitém kontextu; 1-n		XML, PS
<objectIdentifierType>	popis kontextu, ve kterém je identifikátor unikátní, např. NDK, ANL nebo název repozitáře; nutno použít kontrolovaný slovník; 1-1	M	MC, XML, PS
<objectIdentifierValue>	vlastní hodnota identifikátoru, např. img0001-master, urn.nbn.cz-123465 apod.; 1-1	M	MC, XML, PS
<objectCategory>	typ objektu, ke kterým se metadata (PREMIS object) vztahuje, např. file pro soubor, representation pro dig. reprezentaci, bitstream pro bitstream; 1-1	M	MC, XML, PS
<preservationLevel>	údaje o úrovni ochrany souboru, která se na něj vztahuje; některé soubory nejsou tak důležité jako jiné, mají menší úroveň ochrany; 0-n	M	MC, XML, PS
<preservationLevelValue>	hodnota úrovně ochrany, která je pro soubor relevantní, pro původní sken PS hodnota deleted, pro MC a XML hodnota preservation; 1-1	M	MC, XML, PS
<preservationLevelDateAssigned>	datum, kdy byla přiřazena hodnota úrovně ochrany, zápis v ISO 8601, na úroveň dne (DD-MM-RRRR) 0-1	R	MC, XML, PS
<objectCharacteristics>	technické údaje o souboru 1-n	M	MC, XML, PS
<compositionLevel>	údaj o tom, zda je nutné digitální objekt rozbalit nebo dekodovat; např. 0 (defaultně pro žádné zabalení nebo kodování); 1 pro jedno zabalení a kodování, podobně pak hodnota 2; 1-1	M	MC, XML, PS
<fixity>	údaje o kontrolním součtu 0-n	M	MC, XML, PS
<messageDigestAlgorithm>	použitý algoritmus kontrolního součtu, např. MD5 aj. 1-1	M	MC, XML, PS
<messageDigest>	hodnota kontrolního součtu	M	MC,

	1-1		XML, PS
<messageDigestOriginator>	agent (osoba, instituce, stroj, SW), který kontrolní součet vytvořil (např. JHOVE apod.) 0-1	M	MC, XML, PS
<size>	údaje o velikosti souboru v bytech 0-1	M	MC, XML, PS
<format>	údaje o formátu souboru 1-n	M	MC, XML, PS
<formatDesignation>	identifikace formátu souboru, výstup z JHOVE, PRONOM služeb apod. 0-1	M	MC, XML, PS
<formatName>	jméno formátu, např. image/tiff nebo Adobe PDF 1-1	M	MC, XML, PS
<formatVersion>	verze formátu, např. 6.0 0-1	M	MC, XML, PS
<formatRegistry>	identifikace formátu – dodatečná informace o záznamu formátů v registrech formátů (např. PRONOM aj.) 0-1	M	MC, XML, PS
<formatRegistryName>	jméno použitého registru formátů, např. UDFR, PRONOM aj. 1-1	M	MC, XML, PS
<formatRegistryKey>	unikátní identifikátor (označení) formátu v registru, např. fmt/155 z PRONOM 1-1	M	MC, XML, PS
<creatingApplication>	údaje o aplikaci, ve které byl popisovaný soubor vytvořen; nutno popsat skener, SW kde vzniklo ALTO XML/TXT, SW/kodek pro vytvoření JPEG2000 MC 0-n	M	MC, XML, PS
<creatingApplicationName>	název aplikace, např. ImageGear, Kakadu apod.; 0-1	M	MC, XML, PS
<creatingApplicationVersion>	verze aplikace, např. 15.03.000 0-1	M	MC, XML, PS

<dateCreatedByApplication>	datum a čas vytvoření, např. 2008-11-10T12:37:46; musí být ve tvaru ISO 8601 (na úrovni vteřin); 0-1	M	MC, XML, PS
<originalName>	původní jméno souboru, např. digibok_2007081301091_0011.jp2 0-1	M	MC, XML, PS
<relationship>	vyjádření vztahu popisovaného souboru k jiným souborům a událostem (eventům) 0-n	M	MC, XML
<relationshipType>	typ vztahu, doporučené hodnoty: derivation= vztah kde objekt je výsledkem změny jiného objektu; structural= vztah mezi částmi objektu; tj. např. ALTO vytvořené z TIFFU bude mít vztah derivation, podobně jako JPEG2000 z TIFFu vytvořený; 1-1	M	MC, XML;
<relationshipSubType>	upřesnění vztahu, doporučené hodnoty: created from; has source; is source of; has sibling; has part; is part of; has root; includes; is included in; apod.; tj. např. ALTO nebo JPEG2000 vytvořený z původního TIFFu budou mít vztah „created from“ 1-1	M	MC, XML;
<relatedObjectIdentification>	identifikace souvisejícího souboru 1-n pro MC, XML pro vyjádření vztahu k původnímu objektu (skenu)	M	MC, XML
<relatedObjectIdentifierType>	specifikace kontextu, ve kterém je identifikátor souboru jedinečný, např. URN; temporary filepath; objectID 1-1	M	MC, XML
<relatedObjectIdentifierValue>	vlastní řetězec identifikátoru, např. URN:NBN:cz-1301091_011#0001 nebo název souboru, cesta k souboru apod. 1-1	M	MC, XML
<relatedEventIdentification>	identifikace s popisovaným souborem související události (eventu); seznam událostí viz PREMIS event 0-n	M	MC, XML

<relatedEventIdentifierType>	typ události, např. interní číslovací systém událostí jako no.nb.evt; NK repository event ID, UUID apod. 1-1	M	MC, XML
<relatedEventIdentifierValue>	hodnota identifikátoru události, např. NK_EVT_005 nebo hodnota UUID aj. 1-1	M	MC, XML
<relatedEventSequence>	pořadí události, např. 003; k určení pořadí lze určit datum události 0-1	R	MC, XML
<linkingEventIdentifier>	identifikátor události týkající původního skenu PS; typy událostí mohou být např. vytvoření, smazání 0-n pro PS nutný link na události vytvoření (digitalizace) a jeho vymazání	M	PS
<linkingEventIdentifierType>	typ identifikátoru události, např. UUID, NK_eventID, vlastní číslovací systém apod. 1-1	M	PS
<linkingEventIdentifierValue>	hodnota identifikátoru, např. event_01; img0001-master-event001 apod. 1-1	M	PS

8.3.2 PREMIS Event

- bude odpovídat poslední aktuální verzi v době implementace (leden 2011 - PREMIS data dictionary v. 2.1), nebo verzi předchozí
- PREMIS event záznamy shromažďují informace o procesech a událostech, které se týkají jednoho nebo více objektů, v našem případě souborů. Primární použití je k zaznamenání událostí, které popisovaný soubor mění nebo upravují.
- bude vznikat pro události, které se dělaly na obrazových datech
 - o digitalizace – vytvoření prvního skenu (např. do TIFF)
 - o vytvoření ALTO XML
 - o vygenerování MC
 - o vygenerování UC
 - o vymazání PS
- popis událostí bude zachycovat informace o jejich výsledku/výstupu
- záznamy PREMIS event budou uloženy v METS záznamu určeném pro administrativní a technická metadata (AMD_METS.xml) v jeho části <amdSec>, podčást <digiprovMD>

- AMD_METS.xml je linkován z hlavního METS záznamu dokumentu
- pro každou událost bude vytvořena jedna <digiprovMD> část
- každý záznam PREMIS event je linkován na původce aktivity – tj. na PREMIS agent záznam

Obsah pole „Popis“:

- vysvětlení a příklad
- doporučené plnění tam, kde je to možné
- výskyt elementu (jak je definováno formátem PREMIS – dle XSD)
 - 0-1 element je nepovinný, neopakovatelný
 - 0-n element je nepovinný, opakovatelný
 - 1-n element je povinný a opakovatelný
 - element je povinný a neopakovatelný

význam pole „povinnost“

- pole „povinnost“ uvádí, zda je plnění jednotlivých elementů povinné, doporučené nebo volitelné
- může nabývat následujících hodnot
 - M - mandatory (povinně plnit – element je součástí každého záznamu)
 - MA - mandatory if available (povinně plnit pokud je to možné, pokud lze apod.)
 - R - recommended (plnění hodnot elementu je doporučeno, není ovšem povinné)
 - RA - recommended if available (doporučeno pokud lze plnit)
 - O - optional (plnění hodnot elementu je zcela dle konkrétních potřeb)

Pole záznamu PREMIS Event

Element	Popis	Povinnost
<eventIdentifier>	údaje o identifikátoru události v kontextu digitalizace nebo repozitáře 1-1	M
<eventIdentifierType>	typ identifikátoru, např. no.nb.evt; NK_eventID, UUID apod. 1-1	M
<eventIdentifierValue>	hodnota identifikátoru, např. EVT_001; event_019 apod. 1-1	M
<eventType>	kategorizace události, nutno použít kontrolovaný slovník; typy událostí, které musí být zaznamenány: capture, migration, derivation, deletion 1-1	M
<eventDateTime>	datum a čas kdy byla událost provedena; nutno zapsat v ISO 8601 na úroveň vteřin	M

	1-1	
<eventDetail>	další údaje o události, doporučené hodnoty pro výše uvedené <eventType> následují za /: - capture/digitization – vznik prvního skenu - capture/XML_creation - capture/TXT_creation - migration/MC_creation - derivation/UC_creation - deletion/PS_deletion 0-1	M
<eventOutcomeInformation>	informace o výsledku události 0-n	R
<eventOutcome>	kategorizace výsledku události, např. slovy jako successful nebo failure, možno použít kódy – nutno používat kontrolovaný slovník nebo seznam kódů 0-1	M
<linkingAgentIdentifier>	identifikace jednoho nebo více agentů spojených s událostí 0-n	M
<linkingAgentIdentifierType>	označení typu identifikátoru, např. NK_AgentID, UUID apod. 1-1	M
<linkingAgentIdentifierValue>	hodnota identifikátoru, např. agent_softwareName_5.2; agent_novakJ apod. 1-1	M
<linkingAgentRole>	role agenta ve vztahu k události, např. software; SW component; operator; nutno používat kontrolovaný slovník 0-n	R
<linkingObjectIdentifier>	informace o objektu/souboru spojeného s událostí, link na něj 0-n	M
<linkingObjectIdentifierType>	označení typu identifikátoru, např. PhysUnitID; URN, NK_OBJ, OBJ_001 apod.; hodnoty by se měly brát z kontrolovaného slovníku 1-1	M
<linkingObjectIdentifierValue>	hodnota identifikátoru, např. URN:NBN:cz-_0011#0001 aj. 1-1	M

8.3.3 PREMIS Agent

- bude odpovídat poslední aktuální verzi v době implementace (leden 2011 - PREMIS data dictionary v. 2.1), nebo verzi předchozí
- **využití PREMIS agent je spíše myšleno pro tzv. ochranné aktivity, které probíhají na archivních datech (AIP balíček) a je nutné pro každou událost na těchto datech mít přesnější informace o tom, kdo ji provedl (osoba administrátora nebo oprávněné osoby)**
 - o informace v PREMIS event a PREMIS object přicházející z procesu digitalizace v PSP balíčku jsou dostačující a dají nám dostatečné informace o události, kdy byla provedena, na jakém SW byla provedena (PREMIS object „creatingApplication“ + PREMIS event „eventDetail“ – tj. další upřesnění v PREMIS agent není nutné
- záznam PREMIS agent obsahuje charakteristiku tzv. agenta, který je spojen s provedenou a zaznamenanou událostí (PREMIS event)
- agent může být osoba, organizace nebo software
- z PREMIS Event je linkováno na agenta, který určitou akci provedl, typ ID agenta a jeho hodnota jsou uvedené v Premis Events (<premis:linkingAgentIdentifier>), plný popis agenta je pak v PREMIS Agent
- záznamy PREMIS agent budou uloženy v METS záznamu určeném pro administrativní a technická metadata (AMD_METS.xml) v jeho části <amdSec>, podčást <digiprovMD>
 - o AMD_METS.xml je linkován z hlavního METS záznamu dokumentu
- pro každého agenta, tj. jeden PREMIS agent záznam, bude vytvořena jedna <digiprovMD> část

Navrhovaná pole záznamu PREMIS Agent

Obsah pole „Popis“:

- vysvětlení a příklad
- doporučené plnění tam, kde je to možné
- výskyt elementu (jak je definováno formátem PREMIS – dle XSD)
 - o 0-1 element je nepovinný, neopakovatelný
 - o 0-n element je nepovinný, opakovatelný
 - o 1-n element je povinný a opakovatelný
 - o element je povinný a neopakovatelný

význam pole „povinnost“

- pole „povinnost“ uvádí, zda je plnění jednotlivých elementů povinné, doporučené nebo volitelné
- může nabývat následujících hodnot
 - o M - mandatory (povinně plnit – element je součástí každého záznamu)
 - o MA - mandatory if available (povinně plnit pokud je to možné, pokud lze apod.)
 - o R - recommended (plnění hodnot elementu je doporučeno, není ovšem povinné)
 - o RA - recommended if available (doporučeno pokud lze plnit)
 - o O - optional (plnění hodnot elementu je zcela dle konkrétních potřeb)

Element	Popis	Povinnost
<agentIdentifier>	popis identifikátoru, který jednoznačně označuje agenta v rámci jednoho kontextu (repozitář např.) 1-n	M
<agentIdentifierType>	označení typu identifikátoru, např. NK_AgentID, UUID apod. 1-1	M
<agentIdentifierValue>	hodnota identifikátoru, např. agent_softwareName_5.2; agent_novakJ apod. 1-1	M
<agentName>	textové upřesnění agenta, např. přesný název SW, plné jméno osoby apod. - FixImage1.3; Jan Novák; CCS docWorks 6.2.1; 0-n	R
<agentType>	obecné označení agenta – pro osoby např. osoba, pro SW např. software apod. hodnoty: organization; person; software 0-1	M
<agentNote>	použití pouze pokud je <agentType> Software a půjde o agenta souvisejícího s migrací TIFF na JPEG2000 (creation/migration Event); bude obsahovat příkaz k výrobě JPEG2000 souboru v aplikaci Kakadu 0-n	MA

8.3.4 Technická metadata MIX

- Bude využit formát MIX, verze aktuální v době implementace projektu, nebo verze předchozí (prosinec 2010 verze 2 – viz <http://www.loc.gov/standards/mix/>)
- **MIX záznam bude vznikat 1) pro archivní kopii, 2) pro původní soubor vzniklý prvotním skenováním (nejčastěji TIFF)** a to i přesto, že tento TIFF se v průběhu výroby maže a není archivován
- tyto dva MIX záznamy budou součástí jednoho METS záznamu AMD_METS.xml (v části <amdSec>, podčást <techMD>) pro administrativní a technická metadata, který vznikne ke každému obrazovému souboru a který je linkován z hlavního METS záznamu čísla periodika
- **MIX záznamy jednotlivých obrazových souborů se budou lišit – MIX záznam původního skenu nebude obsahovat např. element ImageProcessing, MIX záznam archivního souboru MC nebude naproti tomu obsahovat informace o procesu skenování, které se váží k původnímu skenu a budou v elementu ImageCaptureMetadata apod. – podrobnosti viz tabulka níže, sloupec „užití pro MC a PS“**

- pro každý záznam MIX bude vytvořena vlastní část <techMD>
- externí služby, jako např. JHOVE a PRONOM, budou využívány k plnění polí formátu MIX
- ve formátu MIX nebude uvedena informace o kontrolních součtech (fixity), která je obsažena v PREMIS object a není nutno ji opakovat (viz MIX profily Nizozemí, Finska a Norska)
- <fileSize> je pouze doporučené, údaj o velikosti souboru je součástí popisu PREMIS object

Navrhovaná pole formátu MIX pro popis archivní kopie a původního skenu

Obsah pole „Popis“:

- vysvětlení a příklad
- doporučené plnění tam, kde je to možné
- výskyt elementu (jak je definováno formátem MIX – dle XSD)
 - o 0-1 element je nepovinný, neopakovatelný
 - o 0-n element je nepovinný, opakovatelný

význam pole „povinnost“

- pole „povinnost“ uvádí, zda je plnění jednotlivých elementů povinné, doporučené nebo volitelné
- může nabývat následujících hodnot
 - o M - mandatory (povinně plnit – element je součástí každého záznamu)
 - o MA - mandatory if available (povinně plnit pokud je to možné, pokud lze apod.)
 - o R - recommended (plnění hodnot elementu je doporučeno, není ovšem povinné)
 - o RA - recommended if available (doporučeno pokud lze plnit)
 - o O - optional (plnění hodnot elementu je zcela dle konkrétních potřeb)

Obsah pole „Použití pro“

- použití jednotlivých elementů pro MC, PS (původní sken) – určuje, který element je a který není součástí MIX záznamu MC nebo MIX záznamu popisujícího původní obrazový dokument ze skeneru

Element	Popis	Povinnost	Použití pro
<BasicDigitalObjectInformation>			
<ObjectIdentifier>	údaje o identifikátoru obrazového dokumentu, který je formátem MIX popsán; 0-n	R	MC, PS
<objectIdentifierType>	např. jméno souboru, nebo jiný identifikátor; 0-1	M	MC, PS
<objectIdentifierValue>	hodnota identifikátoru, např. 20110306_001.jp2 nebo urn:nbn:123456; 0-1	M	MC, PS
<fileSize>	velikost souboru	R	MC + PS

	0-1		
<FormatDesignation>	údaje o formátu obrazového souboru 0-1	M	MC, PS
<formatName>	název formátu, např. lze využít MIME types ³ (Image/jp2 apod.) 0-1	M	MC, PS
<formatVersion>	verze formátu, např. 1.0 0-1	M	MC, PS
<byteOrder>	endianita, možnosti jsou little endian, middle (mix) endian a big endian 0-1	M	MC + PS
<Compression>	údaje o kompresi obrazového souboru (pokud 0-n	M	MC, PS
<compressionScheme>	informace o kompresním schématu, vyjádřeno číslem (např. 34712 je komprese JPEG2000) nebo slovy (např. JP2 Lossless) 0-1	M	MC, PS
<BasicImageInformation>	základní technické údaje o obrazovém dokumentu 0-1	M	MC, PS
<BasicImageCharacteristics>	0-1	M	MC, PS
<imageWidth>	šířka obrazu v pixelech, např. 3987 0-1	M	MC, PS
<imageHeight>	výška obrazu v pixelech, např. 2345 0-1	M	MC, PS
<PhotometricInterpretation>	photometrická interpretace 0-1	M	MC, PS
<colorSpace>	barevný prostor, např. RGB 0-1	M	MC, PS
<ColorProfile>	údaje o barevném profilu 0-1 povinné pro dokumenty, kde je nutno uchovat přesnou reprezentaci barvy původního dokumentu a používá se ICC profil)	MA	MC + PS
<IccProfile>	ICC profil 0-1	M	MC + PS
<iccProfileName>	jméno profilu, např. sRGB, Adobe RGB aj.	M	MC + PS

³ <http://www.iana.org/assignments/media-types/index.html>

	0-1		
<iccProfileVersion>	verze profilu, např. sRGB IEC61966-2.1 0-1	M	MC + PS
<iccProfileURI>	odkaz na profil, např. www.profil.cz/sRGB_v4_ICC_pref.icc ; 0-1	R	MC + PS
<SpecialFormatCharacteristics>	speciální technické údaje o obrazovém dokumentu, použití pro formát JPEG2000 0-1 povinný pro JPEG2000	MA	MC
<JPEG2000>	0-1	M	MC
<CodecCompliance>	údaje o kodeku 0-1	M	MC
<codec>	název kodeku, např. Kakadu, LuraWave aj. 0-1	M	MC
<codecVersion>	verze kodeku, např. 3.1 0-1	M	MC
< codestreamProfile >	popis codestream profilu JPEG2000, např. P0 a P1 (viz ISO/IEC 15444-4); 0-1	M	MC
< complianceClass >	specifikace největší výšky, šířky a počtu komponentů, které dekodér dokáže dekódovat, lze použít hodnoty C0, C1 a C2; 0-1	M	MC
<EncodingOptions >	obsahuje informace o kodování JPEG2000 0-1	M	MC
<Tiles >	popis pixelové velikosti dlaždic formátu JPEG2000 0-1	M	MC
< tileWidth>	šířka dlaždice, např. 128 0-1	M	MC
< tileHeight>	výška dlaždice, např. 128 0-1	M	MC
< qualityLayers>	číselná hodnota počtu vrstev, do kterých byl JPEG2000 rozdělen, např. 12 0-1	M	MC
< resolutionLevels>	popis počtu nižších rozlišení, které lze z obrazu získat, např. 6 0-1	M	MC
< ImageCaptureMetadata>	popis procesu skenování, je důležité vyplnit, protože tyto údaje nelze zjistit z finálního	M	PS

	master/archivního souboru 0-1		
<SourceInformation>	informace o předloze 0-1	R	PS
<sourceType>	Book, Newspaper aj.; nutno používat kontrolovaný slovník 0-1	M	PS
<SourceID>	identifikátor předlohy 0-n	R	PS
<sourceIDType>	typ identifikátoru, např. ČČNB, URN:NBN 0-1	M	PS
<sourceIDValue>	vlastní hodnota identifikátoru 0-1 povinné	M	PS
<GeneralCaptureInformation>	základní údaje o skenování 0-1	M	PS
<dateTimeCreated>	údaj o datu a čase skenování, např. 2009-01- 03T08:25:28; zapsat v ISO 8601 na úroveň vteřin 0-1	M	PS
<imageProducer>	entita provádějící skenování, např. The National Library of the Czech Republic, osoba apod. 0-1	M	PS
<captureDevice>	typ skenovacího zařízení, např. reflection print scanner; doporučené využívání hodnot z kontrolovaného slovníku 0-1	M	PS
< ScannerCapture>	údaje o skeneru 0-1	M	PS
<scannerManufacturer>	výrobce skeneru, např. 4DigitalBooks, Treventus, Zeutschel 0-1	M	PS
<ScannerModel>	údaje o konkrétním typu skeneru 0-1	M	PS
<scannerModelName>	jméno modelové řady skeneru, např. DL 0-1	M	PS
<scannerModelNumber>	číslo/označení modelu, např. 3000 0-1	M	PS
<scannerModelSerialNo>	výrobní číslo skeneru, např. E4R0003649 0-1	M	PS

<MaximumOpticalResolution>	údaje o maximálním optickém rozlišení skeneru 0-1	M	PS
<xOpticalResolution>	optické rozlišení na ose x, např. 300 0-1	M	PS
<yOpticalResolution>	optické rozlišení na ose y, např. 300 0-1	M	PS
<opticalResolutionUnit>	jednotka optického rozlišení, např. inch (in.) 0-1	M	PS
<scannerSensor>	popis typu snímacího senzoru skenovacího zařízení, např. matrix, linear, undefined aj. 0-1	M	PS
<ScanningSystemSoftware>	údaje o softwaru skenovacího zařízení 0-1	M	PS
<scanningSoftwareName>	název softwaru, např. Copinet 0-1	M	PS
<scanningSoftwareVersionNo>	číslo verze softwaru, např. 3.7 0-1	M	PS
<DigitalCameraCapture>	údaje o snímacím zařízení (fotoaparát) 0-1 povinné, pokud je používán fotoaparát a není používán skener	MA	PS
<digitalCameraManufacturer>	výrobce fotoaparátu, např. Canon 0-1	M	PS
<DigitalCameraModel>	popis modelu fotoaparátu 0-1	M	PS
<digitalCameraModelName>	název modelové řady, např. EOS 0-1	M	PS
<digitalCameraModelNumber>	označení modelu fotoaparátu, např. 1000D 0-1	M	PS
<digitalCameraModelSerialNo>	výrobní číslo přístroje, např. E12345 0-1	M	PS
<camerarSensor>	typ senzoru fotoaparátu, např. matrix aj. 0-1	M	PS
<CameraCaptureSettings>	údaje o nastavení fotoaparátu použitého ke snímání předloh 0-1	M	PS
<ImageData>	v rámci tohoto kontejnerového elementu budou použity následující sub-elementy: fNumber exposureTime	M	PS

	isoSpeedRatings shutterSpeedValue apertureValue brightnessValue exposureBiasValue maxApertureValue subjectDistance meteringMode lightSource flash focalLength backLight exposureIndex sensingMethod cfaPattern autoFocus PrintAspectRatio všechny hodnoty budou přebrány v případě použití fotoaparátu z údajů Exif		
<orientation>	popis orientace obrazu tak, jak je uložen vzhledem k jeho řádkům a sloupcům, např. normal*; normal, image flipper; normal, rotated 180°; unknown apod. 0-1	M	PS
<ImageAssessmentMetadata>	informace o digitálním obrazu pro jeho hodnocení a využití z hlediska dlouhodobé ochrany apod. 0-1	M	MC, PS
<SpatialMetrics>	rozměry obrázku, 2 rozměrná projekce objektů tak jak ji „vidí“ snímací zařízení 0-1	M	MC, PS
<samplingFrequencyPlane>	popis základní roviny, např. object plane (pro přímo ze předlohy digitalizované dokumenty), source object plane (pro digitalizaci mikrofilmů), camera/scanner focal plane (indikace sampl. frekvence fyzického senzoru); 0-1	R	MC + PS
<samplingFrequencyUnit>	jednotka měření sampl. frekvence, např. hodnoty 1= žádná pevná jednotka	M	MC, PS

	; 2= inch, 3=centimetr; 0-1		
<xSamplingFrequency>	údaje o počtu pixelů na jednotku samplerovací frekvence pro šířku obrázku 0-1 povinné, pokud hodnota samplingFrequencyUnit je 2 nebo 3	MA	MC, PS
<numerator>	čitatel, číselné vyjádření, např. 300 0-1	M	MC, PS
<denominator>	jmenovatel, číselné vyjádření např. 1 0-1	M	MC, PS
<ySamplingFrequency>	údaje o počtu pixelů na jednotku samplerovací frekvence pro výšku obrázku 0-1 povinné, pokud hodnota samplingFrequencyUnit je 2 nebo 3	MA	MC, PS
<numerator>	čitatel, číselné vyjádření, např. 300 0-1	M	MC, PS
<denominator>	jmenovatel, číselné vyjádření např. 1 0-1	M	MC, PS
<ImageColorEncoding>	doplňující údaje o barvě obrazu 0-1	M	MC, PS
<BitsPerSample>	počet bitů na kanál 0-1	M	MC, PS
<bitsPerSampleValue>	hodnota počtu bitů, např. 8, 1, 4 nebo 8,8,8 apod. 0-n POZOR – pro každou hodnotu je nutno element opakovat, tj. např. 3x element <bitsPerSampleValue> s hodnotou 8 <mix:BitsPerSample> <mix:bitsPerSampleValue>8</mix:bitsPerSampleValue> <mix:bitsPerSampleValue>8</mix:bitsPerSampleValue> <mix:bitsPerSampleValue>8</mix:bitsPerSampleValue> </mix:BitsPerSample>	M	MC, PS
<bitsPerSampleUnit>	specifikace jednotky, např. integer nebo floating point 0-1	R	MC, PS
<samplesPerPixel>	počet barevných komponentů na pixel, např. 1, 3, 4 0-1	M	MC, PS

<TargetData>	informace o kalibračních tabulkách 0-1 povinné pro obrazy, kde se dělá kontrola oproti kalibrační tabulce	MA	MC
<targetType>	typ kalibrační tabulky; 0= external (kalibrační tabulka se neobjeví na dig. obraze, je to oddělený dig. soubor); 1= internal (tabulka je naskenována spolu s přelohou a objeví se na dig. obraze); 0-n	M	MC
<targetID>	údaje o původu kalibrační tabulky 0-n	M	MC
<targetManufacturer>	výrobce/původce kalibrační tabulky, např. Eastman Kodak nebo NK ČR, oddělení kontroly kvality apod. 0-1	M	MC
<targetName>	název kalibrační tabulky, např. ColorChecker, MicrofilmScanTarget aj. 0-1	M	MC
<targetNo>	číslo nebo verze kalibrační tabulky 0-1	M	MC
<targetMedia>	údaj o tom, na jakém médiu je kalibrační tabulka, např. film, paper aj. 0-1	R	MC
<externalTarget>	údaje o externí kalibrační tabulce; např. link na http://skenservis.cz/target-00000001 nebo název a cesta ke konkrétnímu souboru 0-n povinné v případě, že byla použita externí kalibrační tabulka (targetType = 0)	MA	MC
<performaceData>	odkaz na soubor obsahující charakteristiku výkonu systému vzhledem k nastaveným hodnotám rozlišení atd.; možné hodnoty plnění – link URN nebo URL, nebo název souboru 0-n	R	MC
<ChangeHistory>	dokumentace procesů provedených na obrazovém souboru v jeho životním cyklu 0-1	M	MC
<ImageProcessing>	údaje o zpracování obrazového souboru 0-n	M	MC

<dateTimeProcessed>	2009-01-04T15:12:06; zapsat v ISO 8601 na úroveň vteřin 0-1	M	MC
<sourceData>	odkaz na původní zdrojová data, ze kterých byl vytvořen finální obrazový soubor; může to být např. URL nebo cesta do složky s původním skenem včetně názvu souboru; 0-1	M	MC
<processingAgency>	The National Library of the Czech Republic 0-n	R	MC

8.4 METS část <fileSec>

8.4.1 <fileSec> hlavního záznamu METS

file group

- pro obrazy i texty (ALTO XML/OCR.TXT) budou v hlavním METS záznamu použity elementy <fileGrp>, jeden element <fileGrp> bude existovat pro obrazy archivních kopií, další pro ALTO XML, další pro OCR.TXT soubory a další pro METS záznamy s technickými metadaty (AMD_METS.xml)
1. <fileGrp> pro obrazy archivních kopií, bude mít tyto atributy: ID="MC_IMGGRP" USE="Images"
 - každý soubor bude mít vlastní element <file> s následujícími atributy:
 - ID – identifikátor souboru jp2 jak je používán v METS záznamu
 - MIMETYPE – hodnota image/jp2
 - SIZE – velikost souboru jp2
 - CHECKSUMTYPE – hodnota MD5
 - CHECKSUM – hodnota kontrolního součtu
 - SEQ – pořadí souboru
 - CREATED – datum vytvoření, ISO8601 na úroveň vteřiny
 - subelementem pod <file> je element <Flocat>, který obsahuje link na obrazový soubor (xlink:href) a atribut LOCTYPE
 2. <fileGrp> pro ALTO XML bude mít následující atributy: ID="ALTOGRP" USE="Text"
 - každý ALTO XML soubor bude mít vlastní element <file> s následujícími atributy:
 - ID – identifikátor souboru ALTO XML jak je používán v METS záznamu
 - MIMETYPE – text/xml
 - SIZE – velikost souboru xml
 - CHECKSUMTYPE – hodnota MD5
 - CHECKSUM - hodnota kontrolního součtu

- CREATED - datum vytvoření, ISO8601 na úroveň vteřiny
- subelementem pod <file> je element <Flocat>, který obsahuje link na xml soubor obsahující ALTO (xlink:href) a atribut LOCTYPE
- 3. <fileGrp> pro soubory METS s technickými metadaty AMD_METS.xml bude mít následující atributy: ID="TECHMDGRP" USE="Technical Metadata"
 - každý METS xml soubor bude mít vlastní element <file> s následujícími atributy:
 - ID - identifikátor souboru AMD_METS.xml jak je používán v METS záznamu
 - MIMETYPE – text/xml
 - SIZE – velikost souboru xml
 - CHECKSUMTYPE – hodnota MD5
 - CHECKSUM - hodnota kontrolního součtu
 - SEQ – pořadí souboru
 - CREATED - datum vytvoření, ISO8601 na úroveň vteřiny
 - subelementem pod <file> je element <Flocat>, který obsahuje link na xml soubor AMD_METS.xml (xlink:href) a atribut LOCTYPE
- 4. <fileGrp> pro soubory OCR.TXT bude mít následující atributy: ID="TXTGRP" USE="Text"
 - každý OCR.TXT soubor bude mít vlastní element <file> s následujícími atributy:
 - ID - identifikátor souboru OCR.TXT jak je používán v METS záznamu
 - MIMETYPE – text/plain
 - SIZE - velikost souboru
 - CHECKSUMTYPE – hodnota MD5
 - CHECKSUM - hodnota kontrolního součtu
 - CREATED - datum vytvoření, ISO8601 na úroveň vteřiny
 - subelementem pod <file> je element <Flocat>, který obsahuje link na txt soubor (xlink:href) a atribut LOCTYPE

8.4.2 <fileSec> vedlejšího METS záznam AMD_METS.xml

- <fileSec> ve vedlejším METS záznamu AMD_METS.xml bude obsahovat jeden element <fileGrp> s vnořenými elementy <file> pro každou reprezentaci stránky, tj. MC, ALTO XML a OCR.TXT
- atributy jednotlivých <file> elementů odpovídají atributům pro jednotlivé typy dokumentů uvedených výše pro <fileSec> hlavního METS záznamu

8.5 METS část <structMap> - Strukturální metadata a ALTO XML

8.5.1 <structMap> hlavního záznamu METS

- zaznamenávají hierarchické informace o dokumentu, včetně vazeb na fyzické soubory, ze kterých se skládají jednotlivé úrovně dokumentu

- 1 strukturální mapa v hlavním METS záznamu popisuje 1 číslo periodika a musí popisovat strukturu až na úroveň všech článků čísla
 - součástí čísla mohou být přílohy – pokud se skenují spolu s číslem, popisuje strukturální mapa METS záznamu číslo včetně přílohy (bere se jako jedno číslo)
- strukturální mapa logická i fyzická včetně linků na ALTO XML bude v hlavním METS záznamu hlavní_METS.xml
- pro každou stránku seskupuje METS logická strukturální mapa odkazy na textové bloky (nebo ilustrace), které jsou součástí té stránky. Informace o blocích textu nebo ilustracích na stránce jsou uloženy v 1 ALTO XML souboru, který stránce odpovídá. Každý blok a každá ilustrace má unikátní identifikátor, který je použit jako odkaz v METS strukturální mapě.

Vyjádření fyzické strukturální mapy

- bude mít následující atributy <structMap LABEL="Physical_Structure" TYPE="PHYSICAL">
- fyzická strukturální mapa obsahuje rodičovský <div>, který obsahuje tyto atributy:
 - o LABEL- může obsahovat titul periodika
 - o TYPE – např. newspaper
 - o ID – identifikátor div
 - o DMDID – identifikátor části popisných metadat
- jednotlivé stránky jsou zanořeny do rodičovského elementu <div> jako dceřiné <div> elementy
 - <div> pro soubory stránky bude mít tyto atributy:
 - TYPE – bude se plnit typem stránky (viz typy stránek v DTD periodika http://digit.nkp.cz/DigitizedPeriodicals/DTD/2.10/DocumentationPeriodical/Periodical.html#element_PeriodicalPage_Link031EEEE0)
 - ID – identifikátor div
 - ORDERLABEL – pořadové číslo stránky, jak je na ní vytištěno
 - ORDER – pořadí stránky v čísle periodika
 - <div> pro soubory stránky vždy obsahují link <fptr> na soubor obrazu archivní kopie, na ALTO XML, na OCR.TXT a na AMD_METS.xml pomocí elementu <par>
 - link na obrazový soubor archivní kopie má v elementu <area> následující atributy: FILEID, který obsahuje ID souboru archivní kopie
 - link na ALTO XML má v elementu <area> následující atributy: FILEID, který obsahuje ID ALTO XML souboru, dále BEGIN="P1" kde P1 je ID elementu <page> z ALTO XML souboru; a atribut BETYPE="IDREF"
 - link na OCR.TXT soubor má v elementu <area> následující atributy: FILEID, který obsahuje ID souboru OCR.TXT
 - link na AMD_METS.xml soubor má v elementu <area> následující atributy: FILEID, který obsahuje ID souboru AMD_METS.xml

Vyjádření logické strukturální mapy

- bude mít následující atributy <structMap LABEL="Logical_Structure" TYPE="LOGICAL">
- logická struktura na úrovni článků nebo např. ilustrací se popisuje pomocí do sebe zanořených elementů <div>

- pokud stránka obsahuje jen obraz a žádný text, pak je popsána jedním elementem <div> a link do souboru ALTO XML vede přímo na element <ComposedBlock>
- stránky obsahující více logických oblastí jsou popsány jedním <div> elementem, který má vnořené <div> elementy pro každou logickou oblast, která odpovídá např. článku, ilustraci.
 - a. pokud se jedná o jednoduchý, celistvý článek na jedné straně, tak je popsán jen jedním <div> elementem s atributem TYPE="article"
 - o v tomto <div> jsou dále jako další <div> elementy zanořeny jednotlivé textové bloky (odstavce, nadpisy, obrazy apod.)
 - u každého bloku je odkaz do ALTO XML souboru na příslušný textový blok <TextBlock> – pomocí tohoto odkazu se v ALTO XML souboru nalezne jak text, tak i informace o jeho umístění na stránce (souřadnice), toto je realizováno pomocí struktury <area> v elementu <fptr>
 - u bloku tvořeného obrazem je odkaz do ALTO XML na příslušný komponovaný blok <ComposedBlock>; je realizováno pomocí struktury <area> v elementu <fptr>
 - v případě použití atributu ORDER umožňuje tento princip u článků vyjádřit i tzv. pořadí čtení jeho částí, jako jsou např. nadpis, autor, obrázek apod.
 - b. pokud článek není celistvý a je rozdělen na více částí, které se vyskytují na jedné nebo více stránkách, je nutné určit pořadí čtení těchto částí, opět pomocí atributu ORDER
 - o pro každou část článku existuje vlastní <div> element, podřízený hlavnímu <div> elementu článku
 - o element <div> každé části má atribut TYPE hodnotu „article-part“ a atribut ID musí vyjadřovat o jakou z částí se jedná, tj. např. ID="article5-1" odpovídá první části článku číslo pět
- do logické struktury PSP balíčku může být v případě její existence zakomponována i příloha (Supplement), která má vlastní <div> element s atributem TYPE="SUPPLEMENT"
 - o vnořené <div> elementy pro obraz a články i jejich použití je shodné se způsobem popisu logické struktury u elementu <div> s atributem TYPE="ISSUE"

Příklad

Logická mapa obsahující číslo periodika se 2 články a 1 přílohou. První článek je jen na straně jedna a má 3 součásti (titul a odstavec s normálním textem) a obrázek s popiskem i uvedeným autorem. Druhý článek začíná na straně první a jeho další odstavec je na straně druhé. První část děleného článku má titulní část a běžný text a druhá část na straně druhé obsahuje jen odstavec s běžným textem.

```

<structMap LABEL="Logical Structure" TYPE="LOGICAL">
  <div LABEL="Mladá fronta no. 5 29.06.1979" TYPE="PERIODICAL_TITLE" ID="TITLE_1" DMDID="XY">
    <div LABEL="Mladá fronta no.5 29.06.1979" TYPE="ISSUE" ID="ISSUE_1" DMDID="XY">
      <div LABEL="Boj o zrno" TYPE="ARTICLE" ID="ARTICLE_1" DMDID="XY" ORDER="0">
        <div TYPE="TITLE" ID="ARTICLE_PART_1" ORDER="1">
          <fptr>
            <area FILEID="ALTO_PAGE_1" BETYPE="IDREF" BEGIN="BLOCK1"/>
          </fptr>
        </div>
        <div TYPE="NORMAL_TEXT" ID="ARTICLE_PART_2" ORDER="2">
          <fptr>
            <area FILEID="ALTO_PAGE_1" BETYPE="IDREF" BEGIN="BLOCK2"/>
          </fptr>
        </div>
        <div LABEL="Obilí na poli" TYPE="PICTURE" ID="ARTICLE_PART_3" DMDID="XY" ORDER="3">
          <div TYPE="CAPTION" ID="ARTICLE_PART_4">
            <fptr>
              <area FILEID="ALTO_PAGE_1" BETYPE="IDREF" BEGIN="BLOCK3"/>
            </fptr>
          </div>
          <div TYPE="AUTHOR" ID="ARTICLE_PART_5">
            <fptr>
              <area FILEID="ALTO_PAGE_1" BETYPE="IDREF" BEGIN="BLOCK4"/>
            </fptr>
          </div>
          <div TYPE="IMAGE" ID="ARTICLE_PART_6">
            <fptr>
              <area FILEID="ALTO_PAGE_1" BETYPE="IDREF" BEGIN="COMPOSED_BLOCK1"/>
            </fptr>
          </div>
        </div>
      </div>
    </div>
    <div LABEL="XVI. sjezd strany" TYPE="ARTICLE" ID="ARTICLE_2" DMDID="XY" ORDER="1">
      <div TYPE="ARTICLE_PART" ID="ARTICLE_2-1" ORDER="1">
        <div TYPE="TITLE" ID="ARTICLE_PART_1" ORDER="1">
          <fptr>
            <area FILEID="ALTO_PAGE_1" BETYPE="IDREF" BEGIN="BLOCK5"/>
          </fptr>
        </div>
        <div TYPE="NORMAL_TEXT" ID="ARTICLE_PART_2" ORDER="2">
          <fptr>
            <area FILEID="ALTO_PAGE_1" BETYPE="IDREF" BEGIN="BLOCK6"/>
          </fptr>
        </div>
      </div>
      <div TYPE="ARTICLE_PART" ID="ARTICLE_2-2" ORDER="2">
        <div TYPE="NORMAL_TEXT" ID="ARTICLE_PART_1" ORDER="1">
          <fptr>
            <area FILEID="ALTO_PAGE_2" BETYPE="IDREF" BEGIN="BLOCK1"/>
          </fptr>
        </div>
      </div>
    </div>
  </div>
  <div LABEL="Mladá fronta no.5 29.06.1979" TYPE="SUPPLEMENT" ID="SUPPL_1" DMDID="XY">
    ... popis článků a obrazů stejně jako u TYPE="ISSUE"
  </div>
</div>
</structMap>

```


kde jednotlivé části obsahují a popisují...

význam pole „povinnost“

- pole „povinnost“ uvádí, zda je plnění jednotlivých elementů povinné, doporučené nebo volitelné
- může nabývat následujících hodnot
 - M - mandatory (povinně plnit – element je součástí každého záznamu)
 - MA - mandatory if available (povinně plnit pokud je to možné, pokud lze apod.)
 - R - recommended (plnění hodnot elementu je doporučeno, není ovšem povinné)
 - RA - recommended if available (doporučeno pokud lze plnit)
 - O - optional (plnění hodnot elementu je zcela dle konkrétních potřeb)

<div> type	Atributy	Popis	Povinnost
TITLE	LABEL TYPE ID DMDID	<div> obsahuje údaje o titulu periodika ----- LABEL – název titulu periodika, včetně čísla a data vydání čísla, např. Mladá fronta no. 5 29.06.1979 TYPE – hodnota „PERIODICAL_TITLE“ ID – identifikátor <div>, např. hodnota „TITLE_1“ DMDID – obsahuje identifikátor DMD popisné části MODS titulu	M
ISSUE nebo SUPPLEMENT	LABEL TYPE ID DMDID	<div> obsahuje údaje o čísle/příloze čísla periodika ----- LABEL – název titulu periodika, ve stejné podobě jako u titulu, tedy např. „Mladá fronta no. 5 29.06.1979“ TYPE- hodnota ISSUE nebo SUPPLEMENT ID – identifikátor <div>, např. hodnota „ISSUE_1“ nebo „SUPPL_1“ DMDID – obsahuje identifikátor DMD popisné části MODS čísla/přílohy	M
ARTICLE	LABEL TYPE ID DMDID ORDER	<div> obsahující údaje o jednom článku a jeho částech ----- LABEL – název článku TYPE – hodnota ARTICLE s pořadovým číslem, např. ARTICLE_1 ID – identifikátor <div> elementu DMDID – identifikátor popisných metadat ORDER – pořadí článku	M
<div> TYPE=“ARTICLE“ může obsahovat další vnořený <div> různých typů popisující různé části článku, rozlišujeme tyto části (typy):			

- TITLE - SUBTITLE - AUTHOR - TRANSLATOR - NORMAL_TEXT – běžný text bez dalšího upřesnění - PICTURE - ARTICLE_PART - u článků, které jsou rozděleny na více míst na jedné stránce nebo více stránkách - o tento <div> pro jednu součást rozděleného článku pak může obsahovat stejné části jako <div> pro článek, tj. (TITLE, SUBTITLE, AUTHOR, TRANSLATOR, NORMAL_TEXT, PICTURE)			
TITLE	TYPE ID ORDER	<div> obsahující link na textový blok s nadpisem ----- TYPE – hodnota „TITLE“ ID – identifikátor <div> elementu, který popisuje jednu část článku (nadpis), např. hodnota „ARTICLE_PART_1“ ORDER – pořadí části článku	MA
<fptr> <area>	FILEID BEGIN BETYPE	FILEID – ID ALTO XML souboru, např. „ALTO_PAGE_1“ BEGIN – ID textového bloku v ALTO XML souboru BETYPE – hodnota IDREF	
SUBTITLE	TYPE ID ORDER	<div> obsahující link na textový blok s podnadpisem ----- TYPE – hodnota „SUBTITLE“ ID – identifikátor <div> elementu, který popisuje jednu část článku (podnadpis), např. hodnota „ARTICLE_PART_2“ ORDER – pořadí části článku	MA
<fptr> <area>	FILEID BEGIN BETYPE	FILEID – ID ALTO XML souboru, např. „ALTO_PAGE_1“ BEGIN – ID textového bloku v ALTO XML souboru BETYPE – hodnota IDREF	
AUTHOR	TYPE ID ORDER	<div> obsahující link na textový blok se jménem autora ----- TYPE – hodnota „AUTHOR“ ID – identifikátor <div> elementu, který popisuje jednu část článku (autor), např. hodnota „ARTICLE_PART_3“ ORDER – pořadí části článku	MA
<fptr> <area>	FILEID BEGIN BETYPE	FILEID – ID ALTO XML souboru, např. „ALTO_PAGE_1“ BEGIN – ID textového bloku v ALTO XML souboru BETYPE – hodnota IDREF	
TRANSLATOR	TYPE ID	<div> obsahující link na textový blok se jménem překladatele	MA

	ORDER	----- TYPE – hodnota „TRANSLATOR“ ID – identifikátor <div> elementu, který popisuje jednu část článku (překladatel), např. hodnota „ARTICLE_PART_3“ ORDER – pořadí části článku	
<fptr> <area>	FILEID BEGIN BETYPE	FILEID – ID ALTO XML souboru, např. „ALTO_PAGE_1“ BEGIN – ID textového bloku v ALTO XML souboru BETYPE – hodnota IDREF	
NORMAL_TEXT	TYPE ID ORDER	<div> obsahující link na textový blok s běžným textem ----- TYPE – hodnota „NORMAL_TEXT“ ID – identifikátor <div> elementu, který popisuje jednu část článku (běžný text), např. hodnota „ARTICLE_PART_4“ ORDER – pořadí části článku	M
<fptr> <area>	FILEID BEGIN BETYPE	FILEID – ID ALTO XML souboru, např. „ALTO_PAGE_1“ BEGIN – ID textového bloku v ALTO XML souboru BETYPE – hodnota IDREF	
PICTURE	LABEL TYPE ID DMDID ORDER	<div> pro obraz náležející k článku plní se pokud se obraz vyskytuje ----- LABEL – název obrazu pokud existuje TYPE - PICTURE ID – identifikátor <div> elementu, který popisuje jednu část článku (běžný text), např. hodnota „ARTICLE_PART_5“ DMDID – link na bibliogr. popis obrazu ORDER – pořadí obrazu	MA
<div> element s typem PICTURE může obsahovat další <div> elementy s typy CAPTION, PICT_AUTHOR, PICT_TITLE a IMAGE; - CAPTION obsahuje text případného popisku k obrazu - AUTHOR obsahuje text se jménem případného autora obrazu - TITLE obsahuje text názvu obrazu, pokud nějaký název existuje - IMAGE – obsahuje link do souboru ALTO XML na blok popisující vlastní obraz			
CAPTION	TYPE ID	<div> obsahující link na textový blok s popisem obrazu ----- TYPE – hodnota CAPTION ID – identifikátor <div> elementu, např. „ARTICLE_PART_6“	MA
<fptr>	FILEID	FILEID – ID ALTO XML souboru	

<area>	BEGIN BETYPE	BEGIN – ID textového bloku v ALTO XML souboru BETYPE – hodnota IDREF	
PICT_AUTHOR	TYPE ID	<div> obsahující link na textový blok s autorem obrazu ----- TYPE – hodnota PICT_AUTHOR ID – identifikátor <div> elementu, např. „ARTICLE_PART_7“	MA
<fptr> <area>	FILEID BEGIN BETYPE	FILEID – ID ALTO XML souboru BEGIN – ID textového bloku v ALTO XML souboru BETYPE – hodnota IDREF	
PICT_TITLE	TYPE ID	<div> obsahující link na textový blok s názvem obrazu ----- TYPE – hodnota PICT_TITLE ID – identifikátor <div> elementu, např. „ARTICLE_PART_7“	MA
<fptr> <area>	FILEID BEGIN BETYPE	FILEID – ID ALTO XML souboru BEGIN – ID textového bloku v ALTO XML souboru BETYPE – hodnota IDREF	
IMAGE	TYPE ID	<div> obsahující link na komponovaný blok ALTO XML obsahující souřadnice vlastního obrazu ----- TYPE – hodnota IMAGE ID – identifikátor <div> elementu, např. „ARTICLE_PART_8“	MA
<fptr> <area>	FILEID BEGIN BETYPE	FILEID – ID ALTO XML souboru BEGIN – ID komponovaného bloku v ALTO XML souboru BETYPE – hodnota IDREF	
ARTICLE_PART	TYPE ID ORDER	<div> obsahující další vnořené <div> odkazující na jednotlivé části konkrétní části rozděleného článku; povinné pro dělený článek Pozn: pod <div> TYPE=“ARTICLE_PART“ lze vnořit všechny typy <div> jako pod <div> TYPE=“ARTICLE“ ----- TYPE – hodnota „ARTICLE_PART“ ID – identifikátor <div> konkrétní části, pro první část děleného článku např. „ARTICLE_2-1“, tj. první část článku 2 ORDER – pořadí konkrétní části děleného článku	MA

Jednotlivé <div> elementy lze kombinovat a vytvářet nové struktury.

8.5.2 <fileSec> vedlejšího záznamu METS (AMD_METS.xml)

- bude obsahovat pouze fyzickou strukturální mapu (TYPE="PHYSICAL")
- ta bude obsahovat pouze jeden <div> element s atributem TYPE="PERIODICAL_PAGE"
- do <div> budou vnořeny odkazy na jednotlivé reprezentace stránky periodika (MC, ALTO XML a OCR.TXT) pomocí elementu <fptr> s atributem FILEID

```
<structMap TYPE="PHYSICAL">
  <div TYPE="PERIODICAL_PAGE">
    <fptr FILEID="JP2_0001"/>
    <fptr FILEID="ALTOXML_0001"/>
    <fptr FILEID="OCRTXT_0001"/>
  </div>
</structMap>
```

8.6 OCR (ALTO XML a TXT OCR)

- bude použita poslední verze formátu ALTO XML aktuální v době implementace, nebo verze předchozí (prosinec 2010 verze 2 – viz <http://www.loc.gov/standards/alto/>)
- níže uvedená specifikace **neobsahuje všechny elementy a atributy formátu ALTO XML, obsahuje pouze ty, které jsou pro tuto konkrétní specifikaci relevantní – každý uvedený element má vyjádřenou míru relevance výrazy: povinné, doporučené a nepovinné**
- elementy a atributy, které v této specifikaci nejsou uvedeny, nepovažujeme pro účely specifikace za důležité
- ALTO XML i OCR TXT vzniknou pro všechny obrazové soubory náležející k jedné intelektuální entitě (svazku nebo číslu periodika) včetně prázdných stran, fotografií hřbetu, předsádky apod.
- ALTO XML i OCR TXT budou vznikat na úrovni stránky
- ALTO XML soubor pro zcela prázdné stránky bude obsahovat element /alto/Layout/Page/PrintSpace, ten ovšem podelementy /alto/Layout/Page/PrintSpace/TextBlock; /alto/Layout/Page/PrintSpace/TextBlock/Illustration; /alto/Layout/Page/PrintSpace/TextBlock/GraphicalElement ani /alto/Layout/Page/PrintSpace/TextBlock/ComposedBlock
- struktura ALTO XML bude generovaná na úrovni rozpoznání slova generovaná OCR
- kvalita rozpoznání znaků bude akceptována do určité hranice, výstupy nebudou ručně opravovány
- struktura ALTO umožní vyhledávání textu a jeho zvýraznění na úrovni slova, pokud bude použit odpovídající prohlížeč
- obrazy reprezentující stránku, které budou použity jako UC, musí odpovídat rozměry, orientací a natočením obrazu, který byl použit pro vytvoření OCR
- OCR TXT bude vznikat z hotových ALTO XML během procesu digitalizace
- ALTO XML se bude vytvářet pouze pro novodobé dokumenty, nebo dokumenty s určitou hranicí kvality OCR

- jméno OCR souboru musí odpovídat jménu obrazového souboru, ke kterému náleží; např. pr_0007.jp2 a al_0007.xml nebo např. 123456_006_alto.xml a 123456_006_archiv.jp2
- kódování ALTO XML i TXT OCR musí být v UTF-8
- souřadnice pozic (HPOS, VPOS, WIDTH, HEIGHT) musí být vyjádřeny v pixelech
- v této specifikaci ALTO XML se počítá s OCR i pro text mimo tzv. textové „zrcadlo“, tj. mimo hlavní text, jako jsou např. čísla stránek, běžící nadpisy ani jiné části vyskytující se na okrajích stránky (top, left, top a bottom margin)
 - o elementy topMargin, leftMargin, rightMargin, bottomMargin budou obsahovat elementy <TextBlock>, pro který platí stejná pravidla, jako pro element <textBlock> pro hlavní text stránky
 - o pozor: údaje z OCR mimo hlavní text stránky by neměly být vyhledávatelné v aplikaci zpřístupnění, docházelo by ke zmatení uživatele a výsledků (např. při hledání titulu kapitoly by byly zobrazeny výsledky pro každou stránku, která obsahuje běžící nadpis apod.)
- pokud je na konci věty dělicí znaménko, ALTO XML i OCR TXT musí obsahovat oba fragmenty slova s dělítkem a současně také kompletní slovo – je vysvětleno dále v tabulce
- ilustrace, reklamy a jiné grafické části stránky nebudou vyjádřeny v tazích /alto/Layout/Page/PrintSpace/Illustration ani Layout/Page/PrintSpace/GraphicalElement, tyto nejsou v popisu/tabulce níže vůbec uvedeny
- ilustrace, reklamy a jiné grafické části stránky budou vyjádřeny v tagu /alto/Layout/Page/PrintSpace/ComposedBlock/ s vyjádřením atributu TYPE, který bude označovat typ bloku (illustration, advertisement aj.)
 - o např. ilustrace bude popsána v elementu /alto/Layout/Page/PrintSpace/ComposedBlock/GraphicalElement, kde ComposedBlock TYPE je Illustration
 - o reklama s textem v rámečku bude popsána v elementu Layout/Page/PrintSpace/ComposedBlock/TextBlock, kde ComposedBlock TYPE je Advertisement
 - o tabulky, grafy obdobně
- elementy /alto/Layout/Page/PrintSpace/ComposedBlock/Illustration a Layout/Page/PrintSpace/ComposedBlock/ComposedBlock také nebudou využity
- /alto/Layout/Page/PrintSpace/ComposedBlock/TextBlock a /alto/Layout/Page/PrintSpace/ComposedBlock/GraphicalElement nebudou obsahovat elementy <Shape>; tvar těchto bloků je vyjádřen v elementu <Shape> samotného elementu <ComposedBlock>; logicky pak souřadnice tvaru <TextBlock> nebo <GraphicalElement> obsaženého v /alto/Layout/Page/PrintSpace/ComposedBlock jsou většinou shodné, pokud není tvarů nebo bloků v rámci /alto/Layout/Page/PrintSpace/ComposedBlock více
- všechny vyplněné hodnoty jsou příklady plnění, plnění v konkrétní instituci je nutno specifikovat vlastními pravidly a kontrolovanými slovníky
- ALTO XML bude využíváno pro tzv. pořadí čtení, tj. článek vyskytující se na více stránkách nebo na více různých místech jedné stránky bude možné zobrazit celý a ve správném pořadí. K tomu je nutno znát jeho strukturu. Struktura bude vyjádřena v korespondujícím METS záznamu v logické

strukturální mapě. Ta bude obsahovat odkazy na jednotlivé textové bloky článku, pomocí ID textových bloků použitých v ALTO XML.

Obsah pole popis:

- vysvětlení a příklad
- doporučené plnění tam, kde je to možné
- výskyt elementu (jak je definováno formátem ALTO XML – dle XSD)
 - 0-1 element je nepovinný, neopakovatelný
 - 0-n element je nepovinný, opakovatelný

význam pole „povinnost“

- pole „povinnost“ uvádí, zda je plnění jednotlivých elementů povinné, doporučené nebo volitelné
- může nabývat následujících hodnot
 - M - mandatory (povinně plnit – element je součástí každého záznamu)
 - MA - mandatory if available (povinně plnit pokud je to možné, pokud lze apod.)
 - R - recommended (plnění hodnot elementu je doporučeno, není ovšem povinné)
 - RA - recommended if available (doporučeno pokud lze plnit)
 - O - optional (plnění hodnot elementu je zcela dle konkrétních potřeb)
- POZOR - u elementů, které obsahují atributy, jsou atributy rozepsány pod čarou (vysvětlení, povinnost uvádění apod.)

Element	Atribut	Popis	Povinnost
<Description>			
<MeasurementUnit>		měřicí jednotka pro souřadnice v ALTO XML; možné hodnoty – dpi, pixel, inch1200 a mm10); inch1200 = 1/1200 inch; doporučené plnění je „mm10“ nebo „pixel“; 0-1	M
<sourceImageInformation>		informace o obrazovém souboru, ze kterého vzniklo ALTO XML; 0-1	M
<fileName>		jméno obrazového souboru, ze kterého bylo ALTO XML vytvářeno; ideálně i s filesystem cestou jeho uložení; např. nlImageSeq-33386-b.tif//produkce/OCR/digibok_XY/	M

		XY_011.tiff 0-1	
<fileIdentifier>		jedinečný identifikátor obrazového souboru; 0-n	R
<OCRProcessing>	ID	popis procesu vzniku OCR; 0-n ----- ID OCR procesu, např. <OCRProcessing ID="OCRPROCES_1">; povinné	M
<preProcessingStep>		procesy před vznikem OCR, které provádí SW pro OCR (např. natočení obrazu) 0-n	M
<processingDateTime>		určení času procesu, který předcházel samotnému OCR; např. 2008-03-29T19:42:23 dle ISO 8601 na úroveň vteřin; 0-1	O
<processingAgency>		jméno nebo kód instituce, např. NK CZ, název externí firmy apod.; doporučujeme použít kontrolovaný slovník hodnot; 0-1	R
<processingStepDescription>		popis procesu (např. zarovnání, ořez apod.); 0-n	O
<processingStepSettings>		nastavení kroku popsaného v <processingStepDescription>, např. CCS OCR Processing Filter 0-1	O
<processingSoftware>		popis SW, který upravoval obrázek před vznikem OCR; 0-1	M
<softwareCreator>		výrobce softwaru - např. CCS Content Conversion Specialists GmbH, Germany; 0-1	M

<softwareName>		jméno softwaru - např. CCS docWORKS; 0-1	M
<softwareVersion>		verze SW, např. 6.2-1.16; 0-1	M
<ocrProcessingStep>		popis procesu vzniku OCR 1-1 – povinné pole	M
<processingDateTime>		okamžik kdy bylo OCR vytvořeno; nutno zapsat v ISO 8601 na úroveň vteřin; 0-1	M
<processingAgency>		jméno nebo kód instituce, např. NK CZ doporučujeme použít kontrolovaný slovník hodnot; 0-1	M
<processingSoftware>		popis SW, který dělal vlastní OCR; 0-1	M
<softwareCreator>		výrobce softwaru - např. ABBYY, Russia; 0-1	M
<softwareName>		jméno softwaru - např. FineReader; 0-1	M
<softwareVersion>		např. 8.0; 0-1	M
<Styles>		styly definují vlastnosti jednotlivých grafických prvků stránky. styl definovaný v elementu vrchní úrovně je použit jako výchozí pro podřízené elementy; 0-1	M
<TextStyle>	ID FONTSTYLE FONTFAMILY FONTSIZE	definuje font textu; 0-n ----- ID pro každý text style použitý v OCR souboru – povinné FONTSTYLE – např. bold, italics apod.; doporučujeme používat	M

		kontrolovaný slovník; doporučené FONTFAMILY – např. arial, calibri apod.; doporučujeme používat kontrolovaný slovník; povinné FONTSIZE – velikost fontu, např. 10, 12 apod.; povinné	
<ParagraphStyle>	ID ALIGN	definuje formátování textových bloků; 0-n ----- ID pro každý odstavec + zarovnání; např. PAR_01, PAR_02 apod. povinné ALIGN – zarovnání; povolené hodnoty: Left, Right, Center, Block aj.; povinné	M
<Layout>		layout - rozložení struktur (slov, odstavců apod.) na jedné stránce dokumentu; 1-1 povinný výskyt element není opakovací	M
<Page>	ID ACCURACY POSITION QUALITY PHYSICAL_IMG_NR HEIGHT WIDTH PC	element popisující jednu stránku dokumentu; 1-n ----- ID – vygenerovaný identifikátor stránky, např. PAGE1, nebo P1 apod.; povinné ACCURACY – procentuální odhad přesnosti OCR (0-100);	M

		doporučené POSITION – pozice stránky; hodnoty k plnění: Left, Right, Foldout, Single, Cover; nepovinné QUALITY – krátký údaj o kvalitě předlohy stránky; hodnoty k plnění: OK, Missing, Missing in original, Damaged, Retained, Target, As in original; nepovinné PHYSICAL_IMG_NR - fyzické (pořadové) číslo stránky v dokumentu; vyjádřeno číslem, např. 1,2,3 apod.; povinné WIDTH – šířka stránky vyjádřená v pixelech; povinné HEIGHT – výška stránky vyjádřená v pixelech; povinné PC = Confidence level OCR souboru – hodnota mezi 0 (nejistá kvalita) a 1 (dobrá kvalita); nepovinné; pokud nevyplníte ACCURACY – tak je vyplnění doporučené	
<TopMargin>	ID HPOS VPOS WIDTH HEIGHT	horní okraj – prostor mezi vrchní hranou listu a vrchní linkou textu; 0-1 ----- ID: unikátní ID pro element	M

		TopMargin, např. P1_TM0001 (page 1, topMargin0001); povinné HPOS: horizontální pozice; povinné VPOS: vertikální pozice; povinné WIDTH – šířka vrchního okraje; povinné HEIGHT – výška vrchního okraje; povinné	
<TextBlock>	stejně plnění a pravidla jako pro element <TextBlock> vnořený do elementu <PrintSpace>		MA
<LeftMargin>	ID HPOS VPOS WIDTH HEIGHT	levý okraj – prostor mezi levým okrajem stránky a textem; 0-1 ----- ID: unikátní ID pro element LeftMargin, např. P1_LM0001 (page 1, leftMargin0001); povinné HPOS: horizontální pozice; povinné VPOS: vertikální pozice; povinné WIDTH – šířka levého okraje; povinné HEIGHT – výška levého okraje; povinné	M
<TextBlock>	stejně plnění a pravidla jako pro element <TextBlock> vnořený do elementu <PrintSpace>		MA
<RightMargin>	ID HPOS	pravý okraj – prostor mezi pravým okrajem stránky a	M

	VPOS WIDTH HEIGHT	textem; 0-1 ----- ID: unikátní ID pro element RightMargin, např. P1_RM0001 (page 1, rightMargin0001); povinné HPOS: horizontální pozice; povinné VPOS: vertikální pozice; povinné WIDTH – šířka pravého okraje; povinné HEIGHT – výška pravého okraje; povinné	
<TextBlock>	stejně plnění a pravidla jako pro element <TextBlock> vnořený do elementu <PrintSpace>		MA
<BottomMargin>	ID HPOS VPOS WIDTH HEIGHT	pravý okraj – prostor mezi spodním okrajem stránky a textem; 0-1 ----- ID: unikátní ID pro element BottomMargin, např. P1_BM0001 (page 1, bottomMargin0001); povinné HPOS: horizontální pozice; povinné VPOS: vertikální pozice; povinné WIDTH – šířka spodního okraje; povinné HEIGHT – výška spodního okraje;	M

		povinné	
<TextBlock>	stejné plnění a pravidla jako pro element <TextBlock> vnořený do elementu <PrintSpace>		MA
<PrintSpace>	ID HPOS VPOS WIDTH HEIGHT	popis tvaru pokrývajícího textové pole stránky; 0-1 ----- ID: unikátní ID pro element <printSpace>, např. P1_PS0001 (page 1, printSpace0001); - povinné HPOS: horizontální pozice; povinné VPOS: vertikální pozice; povinné WIDTH – šířka textového pole; povinné HEIGHT – výška textového pole; povinné	M
<TextBlock>	ID STYLEREFS HPOS VPOS WIDTH HEIGHT	popisy textových bloků na konkrétní stránce; 0-n pokud je stránka prázdná, TextBlock není potřeba uvádět; pokud je na stránce text tak ano ----- ID obsahuje identifikátor textového bloku na stránce, např. "BLOCK1" nebo P1_TB0002 (stránka 1, textový blok 2); povinné STYLEREFS: reference na ID definice formátování textových bloků <ParagraphStyle>; povinné	MA

		HPOS: horizontální pozice bloku; povinné VPOS: vertikální pozice bloku; povinné WIDTH – šířka textového bloku; povinné HEIGHT – výška textového bloku; povinné	
<Shape>		tvar textového bloku; 0-1 – pro jeden výskyt <TextBlock> jeden nebo žádný výskyt <Shape>; plnit v případě, že je tvar textového bloku nestandardní (víceúhelník)	RA
<Polygon>	POINTS	popis (souřadnice) tvaru víceúhelníku; 0-1 ----- POINTS – vyjádření jednotlivých bodů víceúhelníku; povinné	M
<TextLine>	ID STYLEREFS HPOS VPOS WIDTH HEIGHT	popis jedné řádky textu v rámci textového bloku; 1-n nutný alespoň jeden výskyt v rámci textového bloku ----- ID obsahuje identifikátor řádky textu v textovém bloku, např. "P1_TL0002 (stránka 1, řádka 2); povinné STYLEREFS: reference na ID definice formátování textových bloků <ParagraphStyle>; nepovinné	M

		HPOS: horizontální pozice řádky; povinné VPOS: vertikální pozice řádky; povinné WIDTH – šířka řádky; povinné HEIGHT – výška řádky; povinné	
<String>	ID CONTENT HEIGHT WIDTH HPOS VPOS CC WC V případě dělení slov také: SUBS_TYPE SUBS-CONTENT	řetězec znaků – vlastní obsah OCR; znaky tvoří jednotlivá slova a více tagů <String> větu <TextLine>; 1-n v rámci <TextLine> ----- ID obsahuje unikátní sekvenční číslo řetězce na stránce, např. "P3_ST0001" (strana 3, řetězec 1); povinné CONTENT – ukládá vlastní řetězec znaků (slovo); povinné HPOS: horizontální pozice řetězce; povinné VPOS: vertikální pozice řetězce; povinné WIDTH – šířka řetězce; povinné HEIGHT – výška řetězce; povinné CC – úroveň důvěry v přesnost	M

		OCR rozpoznání každého znaku v řetězci; jde o seznam čísel, každé z nich mezi hodnotami 0 (jistá) a 9 (nejistá) pro každý znak; např. CC="0001" pro CONTENT="TEXT"; povinné WC – úroveň důvěry v přesnost OCR výstupu celého řetězce - slova (word confidence); hodnota mezi 0 (nejistá) a 1 (jistá); např. WC="0,99"; povinné SUBS_CONTENT – obsah chybějící části řetězce v případě, že je slovo na konci řádku rozdělené i do druhého řádku; obsahuje celý řetězec - aby byl vyhledatelný i v případě, že slovo se na stránce vyskytuje, ale je rozděleno; povinné SUBS_TYPE – označení typu substituce; možné hodnoty: HypPart1; HypPart2; Abbreviation; povinné - při výskytu SUBS_CONTENT <i>HypPart1</i> se vyskytuje při rozdělení slova u jeho první OCR části (u první části tagu <CONTENT> ve větě (stringu) první; <i>HypPart2</i> se vyskytuje u následujícího tagu <CONTENT>	
--	--	---	--

		v následující větě (stringu), který obsahuje druhou část rozděleného slova/řetězce; <i>Abbreviation</i> – typ substituce používaný při rozepisování zkratk v textu na jejich plný text; při dělení slov v textu HypPart1 a HypPart2 povinné, abbreviation nepovinné	
<ALTERNATIVE>		alternativní hodnota OCR řetězce pro jednotlivá slova; 0-n lze použít v případě nejistoty rozpoznání řetězce;	O
<HYP>	CONTENT WIDTH HPOS VPOS	zápis znaku rozdělovníku slov 0-1 pro jeden výskyt <TextLine>; vždy pro poslední <String>; může se vyskytnout pouze na konci řádku (1x) ----- CONTENT – obsahuje řetězec znaků, které jsou v textu použity na rozdělení slova, nejčastěji „~“; povinné WIDTH – šířka dělicího znaku; doporučené HPOS: horizontální pozice dělicího znaku; doporučené VPOS: vertikální pozice dělicího znaku; doporučené	MA
<SP>	ID WIDTH HPOS VPOS	prázdný prostor mezi řádky; 0-n v rámci jednoho <TextLine>; vždy mezi řádky, tj. mezi tagy <String>;	M

		----- ID: unikátní ID pro prázdný prostor mezi řádky, např. P1_SP0001 (stránka 1, prázdný prostor 0001); povinné HPOS: horizontální pozice; povinné VPOS: vertikální pozice; povinné WIDTH – šířka prázdného prostoru; povinné	
<ComposedBlock>	ID TYPE HPOS VPOS WIDTH HEIGHT STYLEREFS	blok sestávající z jiných bloků; může obsahovat PrintSpace/ComposedBlock/Text Block, PrintSpace/ComposedBlock/Illustration, PrintSpace/ComposedBlock/GraphicalElement, /PrintSpace/ComposedBlock/ComposedBlock, tj. stejné elementy (bloky), které obsahuje samotný element /alto/Layout/Page/PrintSpace; 0-n povinné pro vyjádření bloků textu (např. orámovaný text, reklamy), pro vyjádření ilustrací, tabulek a grafik ----- ID: unikátní ID komponovaný blok, např. P6_CB0001 (stránka 6, komponovaný blok 0001); povinné	MA

		TYPE – označení typu komponovaného bloku; nutné používat kontrolovaný slovník (illustration, Advertisement, apod.); povinné HPOS: horizontální pozice bloku; povinné VPOS: vertikální pozice bloku; povinné WIDTH – šířka komponovaného bloku; povinné HEIGHT – výška komponovaného bloku; povinné	
<Shape>		tvár komponovaného bloku; 0-1 – pro jeden výskyt /alto/Layout/Page/PrintSpace/ComposedBlock jeden nebo žádný výskyt /alto/Layout/Page/PrintSpace/ComposedBlock/Shape; doporučeno – v případě, že je tvár komponovaného bloku nestandardní (víceúhelník)	RA
<Polygon>	POINTS	popis tvaru víceúhelníku; 0-1 ----- POINTS – vyjádření jednotlivých bodů víceúhelníku povinné	M
<TextBlock>	ID STYLEREFS HPOS	v případě, že komponovaný blok (např. orámovaný tvar) obsahuje text;	MA

	VPOS WIDTH HEIGHT	platí stejná pravidla jako pro normální element /alto/Layout/Page/PrintSpace/TextBlock; 0-n (pro jeden výskyt <ComposedBlock> 0 nebo více elementů /alto/Layout/Page/PrintSpace/ComposedBlock/TextBlock>; plnit pokud je v komponovaném bloku text ----- ID obsahuje identifikátor textového bloku v komponovaném bloku, např. P1_CB0002_SUB (stránka 1, textový blok 2, SUB značí komponovaný blok); povinné STYLEREFS: reference na ID definice formátování textových bloků /alto/Styles/ParagraphStyle; povinné HPOS: horizontální pozice bloku; povinné VPOS: vertikální pozice bloku; povinné WIDTH – šířka textového bloku; povinné HEIGHT – výška textového bloku; povinné	
<TextLine>	/alto/Layout/Page/PrintSpace/ComposedBlock/TextBlock/TextLine a ostatní elementy v rámci /alto/Layout/Page/PrintSpace/ComposedBlock/TextBlock mají stejná pravidla a výskyty jako jako ve vrchním		

	elementu /alto/Layout/Page/PrintSpace/TextBlock		
<GraphicalElement>	ID HPOS VPOS WIDTH HEIGHT	popis grafického tvaru; v případě využití v rámci /alto/Layout/Page/PrintSpace/Co mposedBlock označuje rozměry tvaru v rámci něhož je tabulka, ilustrace, reklama apod.; 0-1 - pro jeden výskyt /alto/Layout/Page/PrintSpace/Co mposedBlock 0 nebo max. 1 výskyt <GraphicalElement>; plní se pokud je na stránce a tedy v komponovaném bloku ilustrace, tabulka apod.; ----- ID – identifikátor grafického tvaru; povinné HEIGHT – výška grafického tvaru; povinné WIDTH – šířka grafického tvaru; povinné HPOS – horizontální pozice grafického tvaru; povinné VPOS – vertikální pozice grafického tvaru; povinné	MA