Definice metadatových formátů pro digitalizaci periodik pro ANL

[materiál k debatě]

jméno	datum	verze dokumentu	provedené změny
Jan Hutař - NK	21.7.2011	draft 0.1	první znění

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 periodika nebo svazek monografie (viz dále).
- PSP balíček – producer submission package
    - balíček dat a metadat, který přichází od producenta dat (tedy např. z workflow digitalizace
    - PSP balíček bude obsahovat kompletní intelektuální entitu tj. číslo periodika nebo svazek monografie
    - z workflow digitalizace lze poslat více PSP balíčků v balíku např. [.zip], [.tar] apod.
    - 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

- SIP balíček – submission information package – je balíček dat a metadat v podobě, ve které je akceptovatelný pro LTP systém nebo pro aplikaci zpřístupnění.
- 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 UC i 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.)

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 jako jsou ročník, titul periodika, tj. je pro uživatele i pro systém možné spojit jednotlivá čísla do ročníků a titulů

4. Názvová konvence složek a souborů

pojmenování PSP balíčku

- každý PSP balíček přicházející z digitalizace by měl obsahovat pouze jedinou intelektuální entitu (číslo periodika a/nebo svazek monografie). Pak by název balíčku měl vycházet z identifikátoru této entity, např. URN:NBN, číslo čárového kódu použitého na fyzické jednotce apod.

pojmenování složek

- viz návrh struktur PSP balíčku (kap. 7)

pojmenování souborů

- názvy jakýchkoliv souborů náležejících k jedné základní entitě (svazek nebo číslo) musí být založeny na jednom typu identifikátoru
- pro číslo periodika, svazek monografie by takovým identifikátorem mohlo být 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 (podobný princip může být použit s využitím čárového kódu nebo jiného identifikátoru):

typ souboru	název souboru	vysvětlení
PSP balíček (číslo, svazek)	URN_123456	název celé složky PSP balíčku, u základních int. entit bude v názvu využito vždy URN:NBN
archivní kopie	MC_URN_123456_0013.jp2	archivní JPEG2000 stránky 13 čísla periodika/svazku monografie s URN:NBN 123456
uživatelská kopie	UC_URN_123456_0013.jp2	uživatelská kopie ve formátu JPEG2000 stránky 13 čísla periodika/svazku monografie s URN:NBN 123456
ALTO XML	ALTO_URN_123456_0013.xml	ALTO soubor náležející ke 13té stránce z čísla periodika/svazku monografie s URN:NBN 123456
OCR TXT	TXT_URN_123465_0013.txt	TXT soubor s OCR náležející ke 13té stránce z čísla periodika/svazku monografie s URN:NBN 123456
info.xml	INFO_URN_123456.xml	info xml k celému PSP balíčku čísla periodika/svazku monografie
MD5.xml	MD5_URN_123456.xml	xml s kontrolními součty k celému PSP balíčku čísla periodika/svazku monografie
Hlavni_METS.xml	METS_URN_123456.xml	hlavní METS záznam k celému číslu periodika/svazku monografie s URN:NBN 123456
AMD_METS.xml	AMD_METS_URN_123456_0013.xml	METS záznam s technickými metadaty pro stránku 13 z čísla periodika/svazku monografie s URN:NBN 123456

5. Transportní balík pro jeden nebo více PSP balíčků

Pokud bude jeden PSP balík obsahující 1 základní intelektuální entitu (číslo periodika a/nebo svazek monografie) přemísťován např. jako zip nebo tar, měl by název souboru zip/tar odpovídat názvu PSP balíčku (tedy z čČNB, ISSN nebo ISBN).

Výstupem workflow digitalizace ale může také být balík (např. zip nebo 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_123456.zip nebo CCNB_123456.zip 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_123456_YYYY kde YYYY může být pořadové číslo, datum, doba vzniku jednoho z více balíků obsahujících čísla určitého titulu s identifikátorem ISSN 123456).

Transportní balík by měl obsahovat následující části:

- balíčky PSP (svazků nebo čísel)
- kontrolní součty všech PSP balíčků
- seznam entit, které balík má obsahovat

Do úvahy mohou přijít balící metody jako BagIt[1)], tar, zip apod.

6. 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

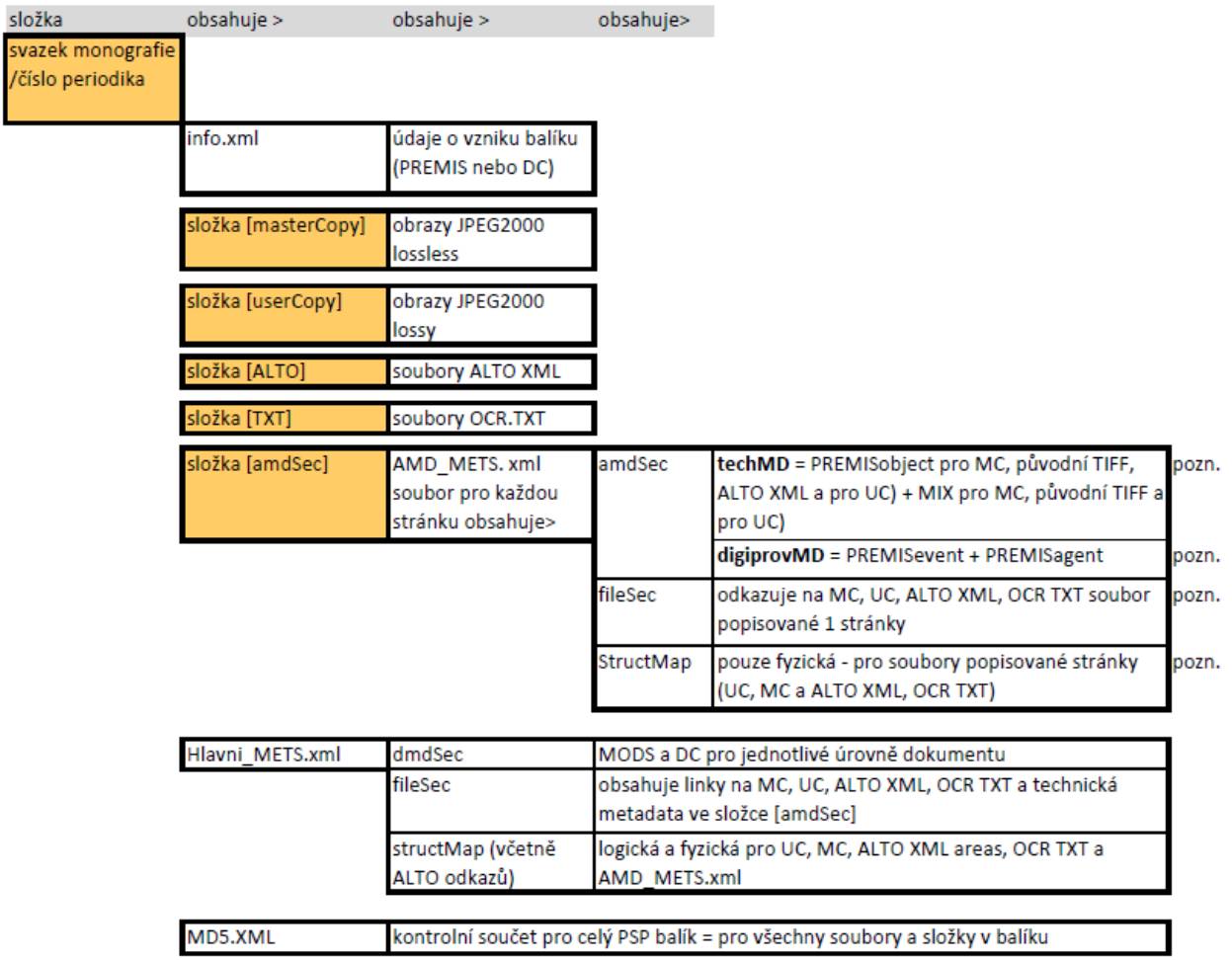
[1] https://confluence.ucop.edu/display/Curation/BagIt

Bylo by ideální, aby během digitalizace byl přidělován (generován přímo nebo vyžádán z aplikace Resolver URN:NBN) identifikátor URN:NBN:

- přidělován bude logickým úrovním (entitám)
    - u periodik tedy: číslo, případně ročník a titul a vnitřním částem
    - u monografií tedy: svazek, případně titul, vnitřní části

7. Struktura PSP balíčku

V kapitole je návrh struktury balení dat a metadat v jednom PSP balíčku na výstupu z workflow digitalizace. Možnost přenosu více PSP balíčku např. v zip souboru není tímto dotčena (více viz kap. 4).



složka	obsahuje >	obsahuje >	obsahuje>		
svazek monografie /číslo periodika					
	info.xml	údaje o vzniku balíku (PREMIS nebo DC)			
	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>	amdSec	**techMD** = PREMISobject pro MC, původní TIFF, ALTO XML a pro UC) + MIX pro MC, původní TIFF a pro UC)	pozn.
				digiprovMD = PREMISevent + PREMISagent	pozn.
			fileSec	odkazuje na MC, UC, ALTO XML, OCR TXT soubor popisované 1 stránky	pozn.
			StructMap	pouze fyzická - pro soubory popisované stránky (UC, MC a ALTO XML, OCR TXT)	pozn.
	Hlavni_METS.xml	dmdSec	MODS a DC pro jednotlivé úrovně dokumentu		
		fileSec	obsahuje linky na MC, UC, ALTO XML, OCR TXT a technická metadata ve složce [amdSec]		
		structMap (včetně ALTO odkazů)	logická a fyzická pro UC, MC, ALTO XML areas, OCR TXT a AMD_METS.xml		
	MD5.XML	kontrolní součet pro celý PSP balík = pro všechny soubory a složky v balíku			

Jedná se o variantu, kdy jeden hlavní metadatový záznam METS obsahuje metadata pro uživatelské i archivní kopie obrazových dat.

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 to 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.

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

složka [userCopy]
složka s uživatelskými kopiemi, pro každou naskenovanou stránku čísla periodika nebo monografie 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 nebo stránek svazku monografie.

složka [TXT]
obsahuje ke každé stránce 1 OCR soubor jako čistý text. Tj. tolik OCR.TXT souborů kolik je stránek čísla periodika nebo stránek svazku monografie.

složka [amdSec]
složka s technickými metadaty – **obsahuje pro každou naskenovanou stránku čísla časopisu nebo monografie 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
    - technických metadat (techMD), která ve formátu PREMISobject popisuje vlastnosti archivních kopií, uživatelských kopií, ALTO XML, původního TIFF souboru, ze kterého vznikly archivní kopie. Dále je přítomen záznam technických metadat v MIX formátu pro archivní a uživatelské kopie a pro původní TIFF.
    - metadat o provenienci digitálních objektů (digiProvMD) – v této části je využit formát PREMISevent a PREMISagent.
    - 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, uživatelská kopie a ALTO XML a OCR TXT. Jde o povinnou sekci METS záznamu.

- structMap – **pouze fyzická** strukturální mapa, povinná část METS záznamu. Bude ukazovat strukturu souborů k dané stránce, tj. opět archivní, uživatelské kopie i 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 nebo svazku monografie 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 záznam v Dublin Core.
- fileSec – hlavní část s linky na všechny digitální objekty (archivní, uživatelské kopie a ALTO XML a OCR TXT), které se váží k jednomu číslu periodika nebo monografii. Obsahuje také linky na administrativní metadata AMD_METS.xml do složky [amdSec].
- structMap – strukturální mapa pro celý dokument, tj. pro jedno číslo periodika nebo monografii. Obsahuje:
  - logickou část pro archivní, uživatelské kopie, ALTO XML, OCR TXT a AMD_METS.xml
  - fyzickou část pro archivní, uživatelské kopie, ALTO XML, OCR TXT a AMD_METS.xml
  - mapování na ALTO XML areas

soubor MD5.XML

poslední částí PSP balíčku je xml soubor s kontrolními součty pro celý balíček, pro každý objekt a pro každou složku. Z tohoto důvodu nejsou kontrolní součty součástí složek s objekty. Kontrolní součty jsou také samozřejmě v technických metadatech.

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

8.1 Bibliografická metadata

- použit bude 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 článek (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 článek 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>
- ad příloha (Supplement) - přílohou se rozumí volně vložená entita do jednotlivého čísla, např. mapa, obsah celého ročníku, CD/DVD 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 úrovní kde je to potřeba (vnitřní část, příloha apod.) se budou opakovat <dmdSec> části tolikrát, kolik je konkrétních částí
- 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**

8.1.1 Pole MODS a Dublin Core pro jednotlivé části periodika

Obsah pole „Popis“:
- vysvětlení a příklad
- doporučené plnění tam, kde je to možné uvést
- povinnost plnění dle NK ČR (slovní vyjádření: povinné, doporučené, nepovinné)

Pole MODS a DC pro titul periodika

Element MODS	**Atributy**	**Popis**	**Element Dublin Core**
<titleInfo>	ID	název titulu periodika povinné – použít názvové autority nebo katalogizační záznam --- ID musí vyjadřovat název úrovně, tj. např. „MODS_TITLE“	
<title>		názvová informace – název periodika povinné – převzít z katalogu	<dc.title> povinné
<subTitle>		podnázev periodika poviné pokud lze uvést	<dc.title> povinné pokud lze uvést
<partNumber>		číslo části, např. určité řady/edice (část 1, řada B), k použití u ročenek apod. doporučené	<dc:description> doporučené
<partName>		jméno edice nebo speciální ediční řady, např. Hygiena. k použití u ročenek a specializovaných periodik doporučené	<dc:description> doporučené
<name>	type	údaje o odpovědnosti za titul periodika povinné pokud lze uvést; nepočítá se s vyplněním u deníků, ale např. u ročenek, zvláštních vydání apod., které mají autora, editora --- type: použít jeden z typů - personal - corporate	

		- conference - family	
<namePart>	type	údaje o křestním jméně a příjmení apod. povinné kde lze uvést 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>	<dc:creator> povinné pokud lze uvést nutno do jednoho pole DC spojit jméno i příjmení
<role>		specifikace role osoby nebo organizace uvedené v elementu <name> povinné kde lze uvést	
<roleTerm>	type authority	popis role nutno použít kontrol. slovník např. z MARC21 povinné kde lze uvést --- 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"	
<typeOfResource>		popis charakteristiky typu nebo obsahu zdroje doporučené jedna z hodnot: - text - cartographic - notated music	<dc:type> doporučené

		- sound recording-musical - sound recording-nonmusical - sound recording - still image - moving image - three dimensional object - software, multimedia - mixed material pro periodika a monografie hodnota text; mělo by se vyčítat z MARC21 katalogizačního záznamu z pozice 06 návěští	
<genre>		bližší údaje o typu dokumentu povinné hodnota: title	<dc:type> povinné
<originInfo>		informace o původu předlohy povinné 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.	
<place>		údaje o místě spojeném s vydáním, výrobou nebo původem popisovaného dokumentu povinné pokud lze uvést	<dc:coverage> povinné pokud lze uvést
<placeTerm>	type	konkrétní určení místa, např. Praha povinné pokud lze uvést odpovídá hodnotě z katalogizačního záznamu, pole 260, podpole „a“ ---------- type – bude vždy text	<dc:coverage> povinné pokud lze uvést

<publisher>		jméno entity, která dokument vydala, vytiskla nebo jinak vyprodukovala povinné pokud lze uvést odpovídá poli 260 podpoli „b“ katalogizačního záznamu v MARC21	<dc:publisher> povinné pokud lze uvést
<dateIssued>		datum vydání předlohy, nutno zaznamenat v případě titulu roky v nichž časopis vycházel (např. 1900-1939) povinné odpovídá hodnotě z katalogizačního záznamu, pole 260, podpole „c“	<dc:date> povinné
<dateCreated>		datum vytvoření předlohy bude použito pouze při popisu tiskaře, viz poznámka u elementu <originInfo> odpovídá hodnotě z katalogizačního záznamu, pole 260, podpole „g“ doporučené	
<issuance>		údaje o vydávání povinné hodnota continuing odpovídá hodnotě uvedené návěští MARC21 na pozici 07	
<frequency>		údaje o pravidelnosti vydávání odpovídá údaji MARC21 v poli 310 nebo pozici 18 v poli 008 doporučené	
<language>		údaje o jazyce dokumentu povinné	
<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 povinné --- type: použít hodnotu code authority: použít hodnotu „iso639-2b“	<dc:language> povinné
<physicalDescription>		obsahuje údaje o fyzickém popisu zdroje/předlohy povinné	
<form>	authority	údaje o fyzické podobě dokumentu, např.	<dc:format>

		print, electronic apod. povinné- pro periodika hodnota print odpovídá hodnotám pozice 23 a 29 v poli 008 MARC21 --- authority: hodnota „marcform“	povinné
<extent>		údaje o rozsahu (stran, svazků nebo rozměrů); použití spíše u ročenek apod. doporučené pokud lze uvést odpovídá hodnotám v poli 300 podpolích „a“ a „c“ MARC21, pokud jsou vyplněna obě pole, bude se element <extent> opakovat	
<note>		poznámka o fyzickém stavu dokumentu; pro každou poznámku je nutno vytvořit nový <note> element doporučeno pokud lze vyplnit	
<abstract>		shrnutí obsahu dokumentu doporučené odpovídá poli 520 MARC21	
<note>		obecná poznámka k dokumentu doporučeno pokud lze vyplnit odpovídá poli 500 v MARC21	<dc:description> doporučené pokud lze vyplnit
<classification>	authority	klasifikační údaje věcného třídění podle Mezinárodního desetinného třídění povinné odpovídá poli 080 MARC21 --- authority: vyplnit hodnotu „udc“	<dc:subject> povinné
<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 doporučené pokud lze uvést Poznámka: element <relatedItem> může obsahovat jakýkoliv jiný element 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 povinné ---------- 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 - urn - pro URN:NBN - uuid - jiný interní identifikátor, hodnota atributu „local“, lze použít např. k vyjádření čárového kódu	<dc:identifier> povinné
<location>		údaje o uložení popisovaného dokumentu, např. signatura, místo uložení apod. povinné	
<physicalLocation>	authority	údaje o instituci, kde je fyzicky uložen popisovaný dokument, např. NK ČR povinné nutno použít kontrolovaný slovník – sigly knihovnen (ABA001 atd.) odpovídá poli 040 v MARC21 ---------- authority: hodnota „siglaADR“	<dc:source> povinné

<shelfLocator>		sigla nebo lokační údaje o dokumentu povinné	<dc:source> povinné
<part>	type	popis částí dokumentu, bude využit jen na popis ročníku (volume) periodika povinné --- type: hodnota bude vždy „volume“	
<detail>	type	upřesnění popisu části povinné --- type: hodnota bude vždy „volume“	
<number>		číslo části (ročníku) povinné pokud lze uvést	<dc:description> povinné pokud lze uvést; nutno doplnit slovo „volume number“, viz <dc:description>volume number: 25</dc:description>
<date>		datum vztahující se k části povinné v případě, že se ročník vycházel během více let (přelom roku), nutno uvést oba roky, např. 1920-1921	
<recordInfo>		údaje o metadatovém záznamu – jeho vzniku, změnách apod. povinné	
<recordContentSource>		kód nebo jméno instituce, která záznam vytvořila nebo změnila; nutno vytvořit kontrolovaný slovník doporučené	
<recordCreationDate>	encoding	datum prvního vytvoření záznamu, na úroveň minut povinné --- encoding: záznam bude podle normy ISO 8601, hodnota atributu tedy iso8601	
<recordChangeDate>	encoding	datum změny záznamu, na úroveň minut doporučené --- encoding: záznam bude podle normy ISO	

		8601, hodnota atributu tedy iso8601	
<recordOrigin>		údaje o vzniku záznamu doporučené hodnoty: machine generated nebo human prepared	

Pole MODS a DC pro číslo periodika

Element MODS	**Atributy**	**Popis**	**Element Dublin Core**
<titleInfo>	ID	název titulu periodika, kterého je číslo součástí povinné – použít názvové autority nebo katalogizační záznam --- ID musí vyjadřovat název úrovně, tj. např. „MODS_ISSUE“	
<title>		názvová informace – titul periodika povinné – převzít z katalogu	<dc:title> povinné
<subTitle>		podnázev periodika doporučené pokud lze uvést	<dc:title> povinné pokud lze vyplnit
<partNumber>		pořadové číslo vydání (čísla), např. 40; nebo u ročenek číslo určité řady/edice (část 1, řada B) povinné pokud lze vyplnit	<dc:description> povinné pokud lze vyplnit
<partName>		jméno edice nebo speciální ediční řady, např. Hygiena. k použití u ročenek a specializovaných periodik doporučené	<dc:description> doporučené
<name>	type	údaje o odpovědnosti za číslo periodika povinné pokud lze uvést; nepočítá se s vyplněním u deníků, ale např. u ročenek, zvláštních vydání apod. --- type: použít jeden z typů - personal - corporate - conference - family	

<namePart>	type	údaje o křestním jméně a příjmení apod. povinné kde lze uvést 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>	<dc:creator> povinné pokud lze uvést nutno do jednoho pole DC spojit jméno i příjmení
<role>		specifikace role osoby nebo organizace uvedené v elementu <name> povinné kde lze uvést	
<roleTerm>	type authority	popis role nutno použít kontrol. slovník např. z MARC21 povinné kde lze uvést --- 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 povinné hodnota: issue, supplement --- type: pro upřesnění typu čísla a jednotlivých vydání povinné hodnota může být:	<dc:type> povinné

		- 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)	
<originInfo>		informace o původu předlohy doporučené kde lze vyplnit (např. u ročenek, kde se vydavatel měnil) nepovinné pro deníky 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 povinné pokud lze uvést	<dc:coverage> povinné pokud lze uvést
<placeTerm>	type	konkrétní určení místa, např. Praha povinné pokud lze uvést odpovídá hodnotě z katalogizačního záznamu, pole 260, podpole „a“ ---------- type – bude vždy text	<dc:coverage> povinné pokud lze uvést
<publisher>		jméno entity, která dokument vydala, vytiskla nebo jinak vyprodukovala povinné pokud lze uvést odpovídá poli 260 podpoli „b“ katalogizačního záznamu v MARC21	<dc:publisher> povinné pokud lze uvést
<dateIssued>		datum vydání předlohy, v případě čísla	<dc:date>

		datum dne, kdy vyšlo; musí vyjádřit den, měsíc a rok povinné nutno zapsat v podobě DD.MM.RRRR	povinné
<dateCreated>		datum vytvoření předlohy bude použito pouze při popisu tiskaře, viz poznámka u elementu <originInfo> odpovídá hodnotě z katalogizačního záznamu, pole 260, podpole „g“ doporučené	
<language>		údaje o jazyce dokumentu povinné	
<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 povinné ---------- type: použít hodnotu code authority: použít hodnotu „iso639-2b“	<dc:language> povinné
<physicalDescription>		obsahuje údaje o fyzickém popisu zdroje/předlohy povinné	
<extent>		údaje o rozsahu (stran, svazků nebo rozměrů); použití spíše u ročenek apod. doporučené pokud lze uvést odpovídá hodnotám v poli 300 podpolích „a“ a „c“ MARC21, pokud jsou vyplněna obě pole, bude se element <extent> opakovat	
<note>		poznámka o fyzickém stavu dokumentu; pro každou poznámku je nutno vytvořit nový <note> element doporučeno pokud lze vyplnit	
<note>		obecná poznámka k dokumentu doporučeno pokud lze vyplnit odpovídá poli 500 v MARC21	
<identifier>	type	údaje o identifikátorech čísla, obsahuje unikátní identifikátory mezinárodní nebo lokální povinné	<dc:identifier> povinné

		---------- type: budou se povinně vyplňovat následující hodnoty, pokud existují: - doi - hdl - handle - isbn - převzít z katalogizačního záznam NK ČR (ročenky apod.) - urn - pro URN:NBN - uuid - jiný interní identifikátor, hodnota atributu „local“, lze použít např. k vyjádření čárového kódu	
<location>		údaje o uložení popisovaného dokumentu, např. signatura, místo uložení apod. doporučené - pro ročenky apod., kde se signatury jednotlivých čísel liší	
<physicalLocation>	authority	údaje o instituci, kde je fyzicky uložen popisovaný dokument, např. NK ČR povinné nutno použít kontrolovaný slovník – sigly knihovnen (ABA001 atd.) odpovídá poli 040 v MARC21 ---------- authority: hodnota „siglaADR“	<dc:source> doporučeno pokud lze vyplnit
<shelfLocator>		sigla nebo lokační údaje o dokumentu povinné pokud lze uvést	<dc:source> doporučeno pokud lze vyplnit

Pole MODS a DC pro číslo pro vnitřní část periodika (článek a ilustrace)

Element MODS	**Atributy**	**Popis**	**Dublin Core Element**
<titleInfo>	ID	názvová informace vnitřní části povinné ---------- ID musí vyjadřovat název úrovně, tj. např. „MODS_PICTURE“ pro obrázek v textu, „MODS_ARTICLE“ pro článek apod.	
<title>		vlastní název vnitřní části (článku) povinné	<dc:title> povinné
<subTitle>		podnázev vnitřní části (článku); za podnázev lze považovat i krátký text,	<dc:title> povinné

		který se před článkem objevuje tučným písmem (shrnutí obsahu článku) povinné pokud lze vyplnit	
<name>	type	údaje o odpovědnosti za vnitřní část povinné pokud lze uvést; --- type: použít jeden z typů - personal - corporate - conference - family	
<namePart>	type	údaje o křestním jméně a příjmení apod. povinné kde lze uvést 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>	<dc:creator> povinné pokud lze uvést nutno do jednoho pole DC spojit jméno i příjmení
<role>		specifikace role osoby nebo organizace uvedené v elementu <name> povinné kde lze uvést	
<roleTerm>	type authority	popis role nutno použít kontrol. slovník např. z MARC21 povinné kde lze uvést --- 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 povinné hodnota: article nebo picture ---------- type: doporučené hodnota „ArticleCategory" – možnost vyplnit bližší určení typu článku (možnost použít DTD periodika, Article Types) - news - advertisement - abstract - introduction - review - dedication - remark - bibliography - editorsNote - preface - aj. hodnota „PictureCategory" – možnost vyplnit další určení typu obrazu - table - illustration - chart - photograph - graphic - map - aj.	<dc:type> povinné
<language>		údaje o jazyce článku povinné	
<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 povinné ----------	<dc:language> povinné

		type: použít hodnotu code authority: použít hodnotu „iso639-2b“	
<physicalDescription>		obsahuje údaje o fyzickém popisu zdroje/předlohy povinné	
<form>	authority	údaje o fyzické podobě dokumentu, např. print, electronic apod. doporučené odpovídá hodnotám pozice 23 a 29 v poli 008 MARC21 --- authority: hodnota „marcform“	<dc:format> doporučené
<abstract>		shrnutí obsahu dokumentu doporučené	
<note>		obecná poznámka k dokumentu doporučeno pokud lze vyplnit odpovídá poli 500 v MARC21	<dc:description> doporučené
<classification>	authority	klasifikační údaje věcného třídění podle Mezinárodního desetinného třídění doporučené odpovídá poli 080 MARC21 --- authority: vyplnit hodnotu „udc“	<dc:subject> doporučené
<identifier>	type	údaje o identifikátorech, obsahuje unikátní identifikátory mezinárodní nebo lokální, které vnitřní část má – viz přehled typů atributů níže povinné --- type: budou se povinně vyplňovat následující hodnoty, pokud existují: - doi - hdl - handle - urn - pro URN:NBN - uuid - jiný interní identifikátor, hodnota atributu „local“, lze použít např. k vyjádření čárového kódu	<dc:identifier> povinné
<part>		popis částí dokumentu, bude využito na záznam rozsahu	

		povinné	
<extent>		upřesnění popisu části – rozsah na stránkách povinné	
<start>		první stránka, na které vnitřní část začíná povinné pokud lze uvést	<dc:coverage> povinné
<end>		poslední stránka, na které vnitřní část končí povinné pokud lze uvést	<dc:coverage> povinné

Pole MODS a DC pro přílohu

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

		- 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>	
<role>		specifikace role osoby nebo organizace uvedené v elementu <name> povinné kde lze uvést	
<roleTerm>	type authority	popis role nutno použít kontrol. slovník např. z MARC21 povinné kde lze uvést --- 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"	
<typeOfResource>		popis charakteristiky typu nebo obsahu přílohy doporučené 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 CD/DVD se SW	<dc:type> doporučené

		- mixed material	
<genre>		bližší údaje o typu dokumentu povinné hodnota: supplement	<dc:type> povinné
<originInfo>		informace o původu přílohy povinné - pokud lze vyplnit a *pokud se liší od údajů v popisu čísla (platí i pro jednotlivé sub-elementy)* 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 přílohy povinné pokud lze uvést	<dc:coverage> povinné pokud lze uvést
<placeTerm>	type	konkrétní určení místa, např. Praha povinné pokud lze uvést odpovídá hodnotě katalogizačního záznamu, pole 260, podpole „a“ ---------- type – bude vždy text	<dc:coverage> povinné pokud lze uvést
<publisher>		jméno entity, která přílohu vydala, vytiskla nebo jinak vyprodukovala povinné pokud lze uvést odpovídá poli 260 podpoli „b“ katalogizačního záznamu v MARC21	<dc:publisher> povinné pokud lze uvést
<dateIssued>		datum vydání přílohy, musí vyjádřit den, měsíc a rok povinné nutno zapsat v podobě DD.MM.RRRR možno použít hodnotu z katalogizačního záznamu, pole 260, podpole „c“	<dc:date> povinné
<dateCreated>		datum vytvoření přílohy bude použito pouze při popisu tiskaře, viz	

		poznámka u elementu <originInfo> nebo např. u popisu CD/DVD apod. doporučené odpovídá hodnotě z katalogizačního záznamu, pole 260, podpole „g“	
<frequency>		údaje o pravidelnosti vydávání doporučené pokud lze vyplnit odpovídá údaji MARC21 v poli 310 nebo pozici 18 v poli 008	
<language>		údaje o jazyce dokumentu povinné	
<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 povinné --- type: použít hodnotu code authority: použít hodnotu „iso639-2b“	<dc:language> povinné
<physicalDescription>		obsahuje údaje o fyzickém popisu zdroje/předlohy povinné	
<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“	<dc:format> povinné
<extent>		údaje o rozsahu (stran, svazků nebo rozměrů) doporučeno pokud lze uvést odpovídá hodnotám v poli 300 podpolích „a“ a „c“ MARC21, pokud jsou vyplněna obě pole, bude se element <extent> opakovat	
<note>		poznámka o fyzickém stavu dokumentu; pro každou poznámku je nutno vytvořit nový <note> element	

		doporučeno pokud lze vyplnit	
<abstract>		shrnutí obsahu dokumentu doporučené pokud lze vyplnit odpovídá poli 520 MARC21	
<note>		obecná poznámka k dokumentu doporučeno pokud lze vyplnit odpovídá poli 500 v MARC21	<dc:description> doporučené pokud lze vyplnit
<classification>	authority	klasifikační údaje věcného třídění podle Mezinárodního desetinného třídění povinné odpovídá poli 080 MARC21 --- authority: vyplnit hodnotu „udc“	<dc:subject> povinné
<identifier>	type	údaje o identifikátorech, obsahuje unikátní identifikátory mezinárodní nebo lokální, které příloha má – viz přehled typů atributů níže povinné pokud lze uvést --- 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 - urn - pro URN:NBN - uuid - jiný interní identifikátor, hodnota atributu „local“, lze použít např. k vyjádření čárového kódu	<dc:identifier> povinné

8.2 Technická a administrativní metadata

- 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 metadatech provedeny
- technická a administrativní metadata budou zabalena v části <amdSec> formátu METS ve vlastních formátech (MIX, PREMIS – části object; events; agent)
- **pro každý obrazový soubor v METS záznamu konkrétních metadat bude existovat vlastní <amdSec> část, která bude obsahovat metadata v <techMD> a <digiprovMD> podčástech.**
- všechny PREMIS záznamy budou obsaženy v tzv. vedlejším METS záznamu (AMD_METS.xml), který je určen pro administrativní a technická metadata (spolu s MIX záznamy).
    - 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.)**

8.2.1 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, 4) uživatelská kopie)
- 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>.
    - 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)
    - 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)
    - 0-1 element je nepovinný, neopakovatelný
    - 0-n element je nepovinný, opakovatelný
    - 1-n element je povinný a opakovatelný
    - element je povinný a neopakovatelný
- povinnost plnění dle NK ČR (slovní vyjádření: povinné, doporučené, nepovinné)

Obsah pole „Použití pro“
- použití jednotlivých elementů pro popis MC, UC, PS (původní sken), XML (ALTO)

Element	Popis	Použití pro
<objectIdentifier>	**identifikátor k jednoznačnému odlišení objektu v určitém kontextu; 1-n povinné**	**MC, XML, PS, UC**
<objectIdentifierType>	popis kontextu, ve kterém je identifikátor unikátní, např. NDK, ANL nebo název repozitáře; nutno použít kontrolovaný slovník; 1-1 povinné	MC, XML, PS, UC
<objectIdentifierValue>	vlastní hodnota identifikátoru, např. img0001-master, urn.nbn.cz-123465 apod.; 1-1 povinné	MC, XML, PS, UC
<objectCategory>	typ objektu, ke kterým se metadata (PREMIS object) vztahuje, např. file pro soubor, representation pro dig. reprezentaci, bitstream pro bitstream; 1-1 povinné	MC, XML, PS, UC
<preservationLevel>	**údaje o úrovni ochrany souboru, která se na něj vztahuje; některé soubory nejsou**	**MC, XML, PS, UC**

	tak důležité jako jiné, mají menší úroveň ochrany; **0-n** **povinné**	
<preservationLevelValue>	hodnota úrovně ochrany, která je pro soubor relevantní, pro původní sken PS hodnota deleted, pro MC a XML hodnota preservation, pro UC hodnota browsing; 1-1 povinné	MC, XML/TXT, PS, UC
<preservationLevelDateAssigned>	datum, kdy byla přiřazena hodnota úrovně ochrany, zápis v ISO 8601, na úroveň dne (DD-MM-RRRR) 0-1 doporučené	MC, XML/TXT, PS, UC
<objectCharacteristics>	**technické údaje o souboru** **1-n** **povinné**	**MC, XML/TXT, PS, UC**
<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 kodovani, podobně pak hodnota 2; 1-1 povinné	MC, XML/TXT, PS, UC
<fixity>	**údaje o kontrolním součtu** **0-n** **povinné**	**MC, XML/TXT, PS, UC**
<messageDigestAlgorithm>	použitý algoritmus kontrolního součtu, např. MD5 aj. 1-1 povinné	MC, XML/TXT, PS, UC
<messageDigest>	hodnota kontrolního součtu 1-1 povinné	MC, XML/TXT, PS, UC
<messageDigestOriginator>	agent (osoba, instituce, stroj, SW), který kontrolní součet vytvořil (např. JHOVE apod.) 0-1 povinné	MC, XML/TXT, PS, UC
<size>	údaje o velikosti souboru v bytech	MC,

	0-1 povinné	XML/TXT, PS, UC
<format>	**údaje o formátu souboru** **1-n** **povinné**	**MC,** **XML/TXT,** **PS, UC**
<formatDesignation>	**identifikace formátu souboru, výstup z JHOVE, PRONOM služeb apod.** **0-1** **povinné**	**MC,** **XML/TXT,** **PS, UC**
<formatName>	jméno formátu, např. image/tiff nebo Adobe PDF 1-1 povinné	MC, XML/TXT, PS, UC
<formatVersion>	verze formátu, např. 6.0 0-1 povinné	MC, XML/TXT, PS, UC
<formatRegistry>	**identifikace formátu – dodatečná informace o záznamu formátů v registrech formátů (např. PRONOM aj.)** **0-1** **povinné**	**MC,** **XML/TXT,** **PS, UC**
<formatRegistryName>	jméno použitého registru formátů, např. UDFR, PRONOM aj. 1-1 povinné	MC, XML/TXT, PS, UC
<formatRegistryKey>	unikátní identifikátor (označení) formátu v registru, např. fmt/155 z PRONOM 1-1 povinné	MC, XML/TXT, PS, UC
<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 a UC, 0-n povinné	MC, XML/TXT, PS, UC
<creatingApplicationName>	název aplikace, např. ImageGear, Kakadu apod.; 0-1 povinné	MC, XML/TXT, PS, UC
<creatingApplicationVersion>	verze aplikace, např. 15.03.000 0-1	MC, XML/TXT,

	povinné	PS, UC
<dateCreatedByApplication>	datum a čas vytvoření, např. 2008-11-10T12:37:46; musí být ve tvaru ISO 8601 (na úroveň vteřin); 0-1 povinné	MC, XML/TXT, PS, UC
<originalName>	původní jméno souboru , např. digibok_2007081301091_0011.jp2 0-1 povinné	MC, XML/TXT, PS, UC
<relationship>	vyjádření vztahu popisovaného souboru k jiným souborům a událostem (eventům) 0-n povinné	MC, XML/TXT, UC
<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 povinné	MC, XML/TXT; UC
<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 povinné	MC, XML/TXT; UC
<relatedObjectIdentification>	identifikace souvisejícího souboru 1-n povinné pro MC, XML/TXT, UC pro vyjádření vztahu k původnímu objektu (skenu), v případě vytváření UC z PS nebo MC, je nutno tento vztah také vyjádřit	MC, XML/TXT, UC
<relatedObjectIdentifierType>	specifikace kontextu, ve kterém je identifikátor souboru jedinečný, např. URN; temporary filepath; objectID	MC, XML/TXT, UC

	1-1 povinné	
<relatedObjectIdentifierValue>	vlastní řetězec identifikátoru, např. URN:NBN:cz-1301091_011#0001 nebo název souboru, cesta k souboru apod. 1-1 povinné	MC, XML/TXT, UC
<relatedEventIdentification>	identifikace s popisovaným souborem související události (eventu); seznam událostí viz PREMIS event 0-n povinné	MC, XML/TXT, UC
<relatedEventIdentifierType>	typ události, např. interní číslovací systém událostí jako no.nb.evt; NK repository event ID, UUID apod. 1-1 povinné	MC, XML/TXT, UC
<relatedEventIdentifierValue>	hodnota identifikátoru události, např. NK_EVT_005 nebo hodnota UUID aj. 1-1 povinné	MC, XML/TXT, UC
<relatedEventSequence>	pořadí události, např. 003; k určení pořadí lze určit datum události 0-1 doporučené	MC, XML/TXT, UC
<linkingEventIdentifier>	identifikátor události týkající původního skenu PS; typy událostí mohou být např. vytvoření, smazání 0-n povinné – pro PS nutný link na události vytvoření (digitalizace) a jeho vymazání	PS
<linkingEventIdentifierType>	typ identifikátoru události, např. UUID, NK_eventID, vlastní číslovací systém apod. 1-1 povinné	PS
<linkingEventIdentifierValue>	hodnota identifikátoru, např. event_01; img0001-master-event001 apod. 1-1 povinné	PS

8.2.2 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
    - digitalizace – vytvoření prvního skenu (např. do TIFF)
    - vytvoření ALTO XML
    - vygenerování MC
    - vygenerování UC
    - 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
- povinnost plnění dle NK ČR (slovní vyjádření: povinné, doporučené, nepovinné)

Pole záznamu PREMIS Event

Element	Popis
<eventIdentifier>	údaje o identifikátoru události v kontextu digitalizace nebo repozitáře 1-1 povinné
<eventIdentifierType>	typ identifikátoru, např. no.nb.evt; NK_eventID, UUID apod. 1-1

	povinné
<eventIdentifierValue>	hodnota identifikátoru, např. EVT_001; event_019 apod. 1-1 povinné
<eventType>	kategorizace události, nutno použít kontrolovaný slovník; typy událostí, které musí být zaznamenány: capture, migration, derivation, deletion 1-1 povinné
<eventDateTime>	datum a čas kdy byla událost provedena; nutno zapsat v ISO 8601 na úroveň vteřin 1-1 povinné
<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 povinné
<eventOutcomeInformation>	**informace o výsledku události** **0-n** **doporučené**
<eventOutcome>	kategorizace výsledku události, např. slovy jako successful nebo failure, možno použít kódy – nutno používat kontrolovaný slovník nebo seznam kódů 0-1 povinné
<linkingAgentIdentifier>	**identifikace jednoho nebo více agentů spojených s událostí** **0-n** **povinné**
<linkingAgentIdentifierType>	označení typu identifikátoru, např. NK_AgentID, UUID apod. 1-1 povinné
<linkingAgentIdentifierValue>	hodnota identifikátoru, např.

	agent_softwareName_5.2; agent_novakJ apod. 1-1 povinné
<linkingAgentRole>	role agenta ve vztahu k události, např. software; SW component; operator; nutno používat kontrolovaný slovník 0-n doporučené
<linkingObjectIdentifier>	**informace o objektu/souboru spojeného s událostí, link na něj** **0-n** **povinné**
<linkingObjectIdentifierType>	označení typu identifikátoru, např. PhysUnitID; URN, NK_OBJ, OBJ_001 apod.; hodnoty by se měly brát z kontrolovaného slovníku 1-1 povinné
<linkingObjectIdentifierValue>	hodnota identifikátoru, např. URN:NBN:cz-_0011#0001 aj. 1-1 povinné

8.2.3 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 adminstrátora nebo oprávněné osoby)**
    - **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>

    - 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)
    - 0-1 element je nepovinný, neopakovatelný
    - 0-n element je nepovinný, opakovatelný
    - 1-n element je povinný a opakovatelný
    - element je povinný a neopakovatelný
- povinnost plnění dle NK ČR (slovní vyjádření: povinné, doporučené, nepovinné)

Element	**Popis**
<agentIdentifier>	**popis identifikátoru, který jednoznačně označuje agenta v rámci jednoho kontextu (repozitář např.)** **1-n** **povinné**
<agentIdentifierType>	označení typu identifikátoru, např. NK_AgentID, UUID apod. 1-1 povinné
<agentIdentifierValue>	hodnota identifikátoru, např. agent_softwareName_5.2; agent_novakJ apod. 1-1 povinné
<agentName>	textové upřesnění agenta, např. přesný název SW, plné jméno osoby apod. - FixImage1.3; Jan Novák; CCS docWorks 6.2.1; 0-n doporučené
<agentType>	obecné označení agenta – pro osoby např. osoba, pro SW např. software apod. hodnoty: organization; person; software 0-1 povinné
<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 povinné pokud lze vyplnit

8.2.4 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) další MIX záznam pro uživatelskou kopii a 3) další MIX záznam pro původní soubor vzniklý prvotním skenováním (nejčastěji TIFF**) a to i přesto, že tento TIFF se v průběhu výroby maže a není archivován
- tyto tři 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 svazku monografie nebo 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)
    - 0-1 element je nepovinný, neopakovatelný
    - 0-n element je nepovinný, opakovatelný
- povinnost plnění dle NK ČR (slovní vyjádření: povinné, doporučené, nepovinné)

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

Element	Popis	Použití pro
<BasicDigitalObjectInformation>		
<ObjectIdentifier>	**údaje o identifikátoru obrazového dokumentu, který je formátem MIX popsán; 0-n doporučené**	MC, PS, UC
<objectIdentifierType>	např. jméno souboru, nebo jiný identifikátor; 0-1 povinné	MC, PS, UC
<objectIdentifierValue>	hodnota identifikátoru, např. 20110306_001.jp2 nebo urn:nbn:123456; 0-1 povinné	MC, PS, UC
<fileSize>	velikost souboru 0-1 doporučené	MC + PS
<FormatDesignation>	**údaje o formátu obrazového souboru 0-1 povinné**	**MC, PS, UC**
<formatName>	název formátu, např. lze využít MIME types[2] (Image/jp2 apod.) 0-1 povinné	MC, PS, UC
<formatVersion>	verze formátu, např. 1.0 0-1 povinné	MC, PS, UC
<byteOrder>	endianita, možnosti jsou little endian, middle (mix) endian a big endian 0-1 povinné	MC + PS
<Compression>	**údaje o kompresi obrazového souboru (pokud 0-n povinné**	**MC, PS, UC**
<compressionScheme>	informace o kompresním schématu, vyjádřeno číslem (např. 34712 je komprese JPEG2000) nebo slovy (např. JP2 Lossless) 0-1	MC, PS, UC

[2] http://www.iana.org/assignments/media-types/index.html

	povinné	
<BasicImageInformation>	**základní technické údaje o obrazovém dokumentu** **0-1** **povinný**	**MC, PS, UC**
<BasicImageCharacteristics>	**0-1**	**MC, PS, UC**
<imageWidth>	šířka obrazu v pixelech, např. 3987 0-1 povinné	MC, PS, UC
<imageHeight>	výška obrazu v pixelech, např. 2345 0-1 povinné	MC, PS, UC
<PhotometricInterpretation>	**photometrická interpretace** **0-1** **povinné**	**MC, PS, UC**
<colorSpace>	barevný prostor, např. RGB 0-1 povinné	MC, PS, UC
<ColorProfile>	**údaje o barevném profilu** **0-1** **povinné pro dokumenty, kde je nutno uchovat přesnou reprezentaci barvy původního dokumentu a používá se ICC profil)**	**MC + PS**
<IccProfile>	**ICC profil** **0-1** **povinné**	**MC + PS**
<iccProfileName>	jméno profilu, např. sRGB, Adobe RGB aj. 0-1 povinné	MC + PS
<iccProfileVersion>	verze profilu, např. sRGB IEC61966-2.1 0-1 povinné	MC + PS
<iccProfileURI>	odkaz na profil, např. www.profily.cz/sRGB_v4_ICC_pref.icc; 0-1 doporučené	MC + PS
<SpecialFormatCharacteristics>	**speciální technické údaje o obrazovém dokumentu, použití pro formát JPEG2000** **0-1** **povinný pro JPEG2000**	**MC, UC**

<JPEG2000>	**0-1** **povinné**	**MC, UC**
<CodecCompliance>	**údaje o kodeku** **0-1** **povinné**	**MC, UC**
<codec>	název kodeku, např. Kakadu, LuraWave aj. 0-1 povinné	MC, UC
<codecVersion>	verze kodeku, např. 3.1 0-1 povinné	MC, UC
< codestreamProfile >	popis codestream profilu JPEG2000, např. P0 a P1 (viz ISO/IEC 15444-4); 0-1 povinné	MC, UC
< complianceClass >	specifikace největší výšky, šířky a počtu komponentů, které dekodér dokáže dekódovat, lze použít hodnoty C0, C1 a C2; 0-1 povinné	MC, UC
<EncodingOptions >	**obsahuje informace o kodování JPEG2000** **0-1** **povinné**	MC, UC
<Tiles >	**popis pixelové velikosti dlaždic formátu JPEG2000** **0-1** **povinné**	MC, UC
< tileWidth>	šířka dlaždice, např. 128 0-1 povinné	MC, UC
< tileHeight>	výška dlaždice, např. 128 0-1 povinné	MC, UC
< qualityLayers>	číselná hodnota počtu vrstev, do kterých byl JPEG2000 rozdělen, např. 12 0-1 povinné	MC, UC
< resolutionLevels>	popis počtu nižších rozlišení, které lze z obrazu získat, např. 6 0-1 povinné	MC, UC

< ImageCaptureMetadata>	**popis procesu skenování, je důležité vyplnit, protože tyto údaje nelze zjistit z finálního master/archivního souboru** **0-1** **povinné**	**PS**
<SourceInformation>	**informace o předloze** **0-1** **doporučené**	**PS**
<sourceType>	Book, Newspaper aj.; nutno používat kontrolovaný slovník 0-1 povinné	PS
<SourceID>	**identifikátor předlohy** **0-n** **doporučené**	**PS**
<sourceIDType>	typ identifikátoru, např. čČNB, URN:NBN 0-1 povinné	PS
<sourceIDValue>	vlastní hodnota identifikátoru 0-1 povinné	PS
<GeneralCaptureInformation>	**základní údaje o skenování** **0-1** **povinné**	**PS**
<dateTimeCreated>	údaj o datu a čase skenování, např. 2009-01-03T08:25:28; zapsat v ISO 8601 na úroveň vteřin 0-1 povinné	PS
<imageProducer>	entita provádějící skenování, např. The National Library of the Czech Republic, osoba apod. 0-1 povinné	PS
<captureDevice>	typ skenovacího zařízení, např. reflection print scanner; doporučené využívání hodnot z kontrolovaného slovníku 0-1 povinné	PS
< ScannerCapture>	**údaje o skeneru** **0-1**	PS

	povinné	
<scannerManufacturer>	výrobce skeneru, např. 4DigitalBooks, Treventus, Zeutschel 0-1 povinné	PS
<ScannerModel>	**údaje o konkrétním typu skeneru** **0-1** **povinné**	PS
<scannerModelName>	jméno modelové řady skeneru, např. DL 0-1 povinné	PS
<scannerModelNumber>	číslo/označení modelu, např. 3000 0-1 povinné	PS
<scannerModelSerialNo>	výrobní číslo skeneru, např. E4R0003649 0-1 povinné	PS
<MaximumOpticalResolution>	**údaje o maximálním optickém rozlišení skeneru** **0-1** **povinné**	PS
< xOpticalResolution>	optické rozlišení na ose x, např. 300 0-1 povinné	PS
< yOpticalResolution>	optické rozlišení na ose y, např. 300 0-1 povinné	PS
< opticalResolutionUnit>	jednotka optického rozlišení, např. inch (in.) 0-1 povinné	PS
<scannerSensor>	popis typu snímacího senzoru skenovacího zařízení, např. matrix, linear, undefined aj. 0-1 povinné	PS
<ScanningSystemSoftware>	**údaje o softwaru skenovacího zařízení** **0-1** **povinné**	PS
<scanningSoftwareName>	název softwaru, např. Copinet 0-1 povinné	PS
<scanningSoftwareVersionNo>	číslo verze softwaru, např. 3.7	PS

	0-1 povinné	
<DigitalCameraCapture>	**údaje o snímacím zařízení (fotoaparát)** **0-1** **povinné, pokud je používán fotoaparát a není používán skener**	PS
<digitalCameraManufacturer>	výrobce fotoaparátu, např. Canon 0-1 povinné	PS
<DigitalCameraModel>	**popis modelu fotoaparátu** **0-1** **povinné**	PS
<digitalCameraModelName>	název modelové řady, např. EOS 0-1 povinné	PS
< digitalCameraModelNumber>	označení modelu fotoaparátu, např. 1000D 0-1 povinné	PS
< digitalCameraModelSerialNo>	výrobní číslo přístroje, např. E12345 0-1 povinné	PS
<camerarSensor>	typ senzoru fotoaparátu, např. matrix aj. 0-1 povinné	PS
<CameraCaptureSettings>	**údaje o nastavení fotoaparátu použitého ke snímání předloh** **0-1** **povinné**	PS
<ImageData>	**v rámci tohoto kontejnerového elementu budou použity následující sub-elementy:** **fNumber** **exposureTime** **isoSpeedRatings** **shutterSpeedValue** **apertureValue** **brightnessValue** **exposureBiasValue** **maxApertureValue** **subjectDistance** **meteringMode** **lightSource**	PS

	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 povinné	PS
<ImageAssessmentMetadata>	**informace o digitálním obrazu pro jeho hodnocení a využití z hlediska dlouhodobé ochrany apod.** **0-1** **povinné**	**MC, PS, UC**
<SpatialMetrics>	**rozměry obrázku, 2 rozměrná projekce objektů tak jak ji „vidí“ snímací zařízení** **0-1** **povinné**	**MC, PS, UC**
<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 doporučené	MC + PS
<samplingFrequencyUnit>	jednotka měření sampl. frekvence, např. hodnoty 1= žádná pevná jednotka ; 2= inch, 3=centimetr; 0-1 povinné	MC, PS, UC
<xSamplingFrequency>	**údaje o počtu pixelů na jednotku samplovací frekvence pro šířku obrázku** **0-1**	**MC, PS, UC**

	povinné, pokud hodnota samplingFrequencyUnit je 2 nebo 3	
<numerator>	čitatel, číselné vyjádření, např. 300 0-1 povinné	MC, PS, UC
<denominator>	jmenovatel, číselné vyjádření např. 1 0-1 povinné	MC, PS, UC
<ySamplingFrequency>	**údaje o počtu pixelů na jednotku samplovací frekvence pro výšku obrázku** **0-1** **povinné, pokud hodnota samplingFrequencyUnit je 2 nebo 3**	**MC, PS, UC**
<numerator>	čitatel, číselné vyjádření, např. 300 0-1 povinné	MC, PS, UC
<denominator>	jmenovatel, číselné vyjádření např. 1 0-1 povinné	MC, PS, UC
<ImageColorEncoding>	**doplňující údaje o barvě obrazu** **0-1** **povinné**	**MC, PS, UC**
<BitsPerSample>	**počet bitů na kanál** **0-1** **povinné**	**MC, PS, UC**
<bitsPerSampleValue>	hodnota počtu bitů, např. 8, 1, 4 nebo 8,8,8 apod. 0-n povinné	MC, PS, UC
<bitsPerSampleUnit>	specifikace jednotky, např. integer nebo floating point 0-1 doporučené	MC, PS, UC
<samplesPerPixel>	počet barevných komponentů na pixel, např. 1, 3, 4 0-1 povinné	MC, PS, UC
<TargetData>	**informace o kalibračních tabulkách** **0-1** **povinné pro obrazy, kde se dělá kontrola oproti kalibrační tabulce**	**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 povinné	MC
<targetID>	**údaje o původu kalibrační tabulky** **0-n** **povinné**	MC
<targetManufacturer>	výrobce/původce kalibrační tabulky, např. Eastman Kodak nebo NK ČR, oddělení kontroly kvality apod. 0-1 povinné	MC
<targetName>	název kalibrační tabulky, např. ColorChecker, MicrofilmScanTarget aj. 0-1 povinné	MC
<targetNo>	číslo nebo verze kalibrační tabulky 0-1 povinné	MC
<targetMedia>	údaj o tom, na jakém médiu je kalibrační tabulka, např. film, paper aj. 0-1 doporučené	MC
<externalTarget>	údaje o externí kalibrační tabulce; např. link na http://urn.fi/URN:NBN-fi-fd2009-target-00000001 nebo název a cesta ke konkrétnímu souboru 0-n povinné v případě, že byla použita externí kalibrační tabulka (targetType = 0)	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 doporučené	MC
<ChangeHistory>	**dokumentace procesů provedených na**	MC

	obrazovém souboru v jeho životním cyklu **0-1** **povinné**	
<ImageProcessing>	**údaje o zpracování obrazového souboru** **0-n** **povinné**	MC
<dateTimeProcessed>	2009-01-04T15:12:06; zapsat v ISO 8601 na úroveň vteřin 0-1 povinné	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 povinné	MC
<processingAgency>	The National Library of the Czech Republic 0-n doporučené	MC

8.3 METS

8.3.1 METS <fileSec>

file group

- pro obrazy i texty (ALTO XML) budou použity elementy <fileGrp>, 1 pro všechny obrazy, 1 pro všechny texty (ALTO XML) a 1 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

5. <fileGrp> pro obrazy uživatelských kopií, bude mít tyto atributy: ID="UC_IMGGRP" USE="Access"
    - každý soubor bude mít vlastní element <file> s následujícími atributy:
        - ID – identifikátor souboru jp2 uživatelské kopie 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 uživatelské kopie (xlink:href) a atribut LOCTYPE

8.3.2 Strukturální metadata a ALTO XML – METS <structMap>

- zaznamenávají hierarchické informace o dokumentu, včetně vazeb na fyzické soubory, ze kterých se skládají jednotlivé úrovně dokumentu
- 1 strukturální mapa popisuje 1 číslo periodika a musí popisovat strukturu až na úroveň všech článků čísla
- strukturální mapa METS včetně linků na ALTO XML bude v hlavním METS záznamu hlavni_METS.xml
- pro každou stránku seskupuje METS 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 „vrchní" , který obsahuje tyto atributy:
    - LABEL- může obsahovat titul periodika
    - TYPE – např. newspaper
    - ID – identifikátor div
    - DMDID – identifikátor části popisných metadat
- jednotlivé stránky jsou zanořeny do „vrchního" elementu jako další elementy
    - 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_Link031EEEA0)
        - ID – identifikátor div
        - ORDERLABEL – pořadové číslo stránky, jak je na ní vytištěno
        - ORDER – pořadí stránky v čísle periodika
    - pro soubory stránky vždy obsahují link <ftpr> na soubor obrazu archivní a uživatelské 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 obrazový soubor uživatelské kopie má v elementu <area> následující atributy: FILEID, který obsahuje ID souboru uživatelské 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 úroveň článků nebo např. ilustrací se popisuje pomocí do sebe zanořených elementů
- stránky sestávající pouze z jedné oblasti (area), jsou popsány jedním div elementech (TYPE="page")
- stránky obsahující více oblastí (area) jsou popsány jedním elementem, který má vnořené elementy pro každou oblast (area), která odpovídá např. článku, ilustraci. Atribut TYPE těchto sub-divů bude např. "ARTICLE" aj.
    - v tomto jsou dále zanořeny jednotlivé textové bloky (odstavce apod.)
    - u každého bloku je odkaz do ALTO XML souboru – 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>

Může to vypadat např. následovně…

```
<structMap TYPE="LOGICAL">
<div>  TYPE´"TITLE"
        <div> TYPE="ISSUE"
                <div> TYPE="TITLE SECTION"
                        <div> TYPE="HEADLINE"
                                <ftpr><area>
                        <div> TYPE="TEXTBLOCK"
                                <ftpr><area>
                <div> TYPE="CONTENT"
                        <div> TYPE="ARTICLE"
                                <div> TYPE="HEADING"
                                        <div> TYPE="TITLE"
                                                <ftpr><area>
                                <div> TYPE="BODY"
                                        <div> TYPE="BODY_CONTENT"
                                                <div> TYPE="PARAGRAPH"
                                                        <div> TYPE="TEXT"
                                                                <ftpr><area>
                                        <div> TYPE="ILLUSTRATION"
                                                <div> TYPE="IMAGE"
                                                        <ftpr><area>
```

kde jednotlivé části obsahují a popisují…

** **	**atributy**	**popis**
TITLE	LABEL TYPE ID DMDID	obsahuje údaje o titulu periodika povinné --- LABEL – název titulu periodika TYPE – hodnota TITLE ID – identifikátor DMDID – obsahuje identifikátor DMD popisné části MODS titulu
ISSUE	LABEL TYPE ID DMDID	obsahuje údaje o čísle periodika povinné --- LABEL – název titulu periodika TYPE- hodnota ISSUE ID – identifikátor DMDID – obsahuje identifikátor DMD popisné části MODS čísla
TITLE_SECTION	TYPE ID	obsahující údaje o titulní části čísla periodika; tj. o části, kde je titul, logo apod. doporučené --- TYPE – hodnota TITLE_SECTION ID – identifikátor elementu
HEADLINE	TYPE ID ORDER	s odkazy (pomocí ftpr/area) na přesná místa s textem nadpisu v ALTO XML souboru; povinné pokud se popisuje TITLE_SECTION --- TYPE – hodnota HEADLINE ID – identifikátor elementu ORDER – pořadí nadpisu
<ftpr> <area>	FILEID BEGIN BETYPE	FILEID – ID ALTO XML souboru BEGIN – ID textového bloku v ALTO XML souboru BETYPE – hodnota IDREF

TEXTBLOCK	TYPE ID ORDER	s odkazy (pomocí ftpr/area) na přesná místa s textem objevujícím se v nadpisové části první strany periodika v ALTO XML souboru; povinné pokud se popisuje TITLE_SECTION --- TYPE – hodnota TEXTBLOCK ID – identifikátor elementu ORDER – pořadí textového bloku
<ftpr> <area>	FILEID BEGIN BETYPE	FILEID – ID ALTO XML souboru BEGIN – ID textového bloku v ALTO XML souboru BETYPE – hodnota IDREF
CONTENT	TYPE ID	obsahující údaje textové části periodika – obsahuje články, ilustrace apod. povinné --- TYPE – hodnota CONTENT ID – identifikátor elementu
ARTICLE	LABEL TYPE ID DMDID	obsahující údaje o článku (popis nadpisu, těla článku apod.) povinné --- LABEL – název článku TYPE – hodnota ARTICLE ID – identifikátor elementu DMDID – identifikátor popisných metadat
TYPE="ARTICLE" tak může obsahovat další různých typů popisující různé části článku (heading, body apod.)		
HEADING	TYPE ID	obsahující popis nadpisu článku povinné pokud se vyskytuje --- TYPE – hodnota HEADING ID – identifikátor elementu
TITLE	TYPE	určující typ nadpisu

	ID	doporučené --- TYPE – hodnota TITLE ID – identifikátor elementu
<ftpr> <area>	FILEID BEGIN BETYPE	FILEID – ID ALTO XML souboru BEGIN – ID textového bloku v ALTO XML souboru BETYPE – hodnota IDREF
BODY	TYPE ID	popisující vlastní text článku, odstavce a ilustrace povinné --- TYPE – hodnota BODY ID – identifikátor elementu
TYPE="BODY" tak může obsahovat další různých typů popisující různé části článku (body_content, illustration apod.)		
BODY_CONTENT	TYPE ID	obsahující vnořený s odkazy na odstavce v ALTO XML povinné --- TYPE – hodnota BODY_CONTENT ID – identifikátor elementu
PARAGRAPH	TYPE ID ORDER	obsahující s typem text a odkazy na ALTO XML povinné --- TYPE – hodnota PARAGRAPH ID – identifikátor elementu ORDER – pořadí odstavce
TEXT	TYPE ID	určující typ odstavce povinné --- TYPE – hodnota TEXT ID – identifikátor elementu
<ftpr> <area>	FILEID BEGIN BETYPE	FILEID – ID ALTO XML souboru BEGIN – ID textového bloku v ALTO XML souboru BETYPE – hodnota IDREF
dalším typem v TYPE="BODY" může být např. illustration		
ILLUSTRATION	LABEL	obsahující údaje ilustraci a

	TYPE ID DMDID ORDER	linky do ALTO XML souboru povinné pokud se ilustrace vyskytuje --- LABEL – název ilustrace pokud je TYPE - ILLUSTRATION ID – identifikátor elementu DMDID – link na bibliogr. popis ORDER – pořadí ilustrace
IMAGE	TYPE ID	určující typ ilustrace, obsahuje ftpr/area odkazy na ALTO XML povinné --- TYPE – hodnota IMAGE ID – identifikátor elementu
<ftpr> <area>	FILEID BEGIN BETYPE	FILEID – ID ALTO XML souboru BEGIN – ID textového bloku v ALTO XML souboru BETYPE – hodnota IDREF

Jednotlivé elementy lze kombinovat a vytvářet nové struktury i nové typy. Tj. popisovaná struktura a typy jednotlivých elementů lze měnit , doplňovat.

8.4 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 úroveň stránky; OCR se nebude dělat úroveň článků
- 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
- ALTO XML se bude ukládat v LTP systému, ale nebude provázáno s archivní kopií – tj. není potřeba mít ve <structMap> záznamu METS, který půjde do LTP systému, odkazy na ALTO XML
- 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 jen pro hlavní text stránky, tj. nebudou popsány např. čísla stránek, běžící nadpisy ani jiné části vyskytující se na okrajích stránky mimo hlavní text (top, left, top a bottom margin)
    - elementy topMargin, leftMargin, rightMargin, bottomMargin tedy nebudou obsahovat elementy <TextBlock> ani jiné, jen atributy (viz níže)
- pokud je na konci věty dělící 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.)
    - např. ilustrace bude popsána v elementu /alto/Layout/Page/PrintSpace/ComposedBlock/GraphicalElement, kde ComposedBlock TYPE je Illustration
    - reklama s textem v rámečku bude popsána v elementu Layout/Page/PrintSpace/ComposedBlock/TextBlock, kde ComposedBlock TYPE je Advertisement
    - 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, vlastní plnění bude specifikováno pravidly a kontrolovanými slovníky

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
- povinnost plnění dle NK ČR (slovní vyjádření: povinné, doporučené, nepovinné)
- u elementů, které obsahují atributy, jsou atributy rozepsány pod čarou (vysvětlení, povinnost uvádění apod.)

Element	Atribut	Popis
<Description>		
<MeasurementUnit>		měřící jednotka pro souřadnice v ALTO XML; možné hodnoty – dpi, pixel, inch1200 a mm10); inch1200 = 1/1200 inche; doporučené plnění je „mm10“ nebo „pixel“; 0-1 povinné
<sourceImageInformation>		informace o obrazovém souboru, ze kterého vzniklo ALTO XML; 0-1 povinné
<fileName>		jméno obrazového souboru, ze kterého bylo ALTO XML vytvářeno; ideálně i s filesystem cestou jeho uložení; např. nlaImageSeq-33386-b.tif//produkce/OCR/digibok_XY/XY_011.tiff 0-1 povinné
<fileIdentifier>		jedinečný identifikátor obrazového souboru; 0-n doporučené
<OCRProcessing>	ID	popis procesu vzniku OCR; 0-n povinné ---------- ID OCR procesu, např. <OCRProcessing

		ID="OCRPROCES_1">; povinné
<preProcessingStep>		procesy před vznikem OCR 0-n povinné
<processingDateTime>		určení času procesu, který předcházel samotnému OCR; např. 2008-03-29T19:42:23 dle ISO 8601 na úroveň vteřin; 0-1 povinné
<processingAgency>		jméno nebo kód instituce, např. NK CZ; doporučujeme použít kontrolovaný slovník hodnot; 0-1 doporučené
<processingStepDescription>		popis procesu (např. zarovnání, ořez apod.); 0-n doporučené
<processingStepSettings>		např. CCS OCR Processing Filter 0-1 doporučené
<processingSoftware>		popis SW, který upravoval obrázek před vznikem OCR; 0-1 povinné
<softwareCreator>		výrobce softwaru - např. CCS Content Conversion Specialists GmbH, Germany; 0-1 povinné
<softwareName>		jméno softwaru - např. CCS docWORKS; 0-1 povinné
<softwareVersion>		verze SW, např. 6.2-1.16; 0-1 povinné
<ocrProcessingStep>		popis procesu vzniku OCR 1-1 – povinné pole povinné
<processingDateTime>		okamžik kdy bylo OCR vytvořeno; nutno zapsat v ISO 8601 na úroveň vteřin; 0-1 povinné
<processingAgency>		jméno nebo kód instituce, např. NK CZ

		doporučujeme použít kontrolovaný slovník hodnot; 0-1 povinné
<processingSoftware>		popis SW, který dělal vlastní OCR; 0-1 povinné
<softwareCreator>		výrobce softwaru - např. ABBYY, Russia ; 0-1 povinné
<softwareName>		jméno softwaru - např. FineReader; 0-1 povinné
<softwareVersion>		např. 8.0; 0-1 povinné
<Styles>		styly definují vlastnosti jednotlivých grafických prvků stránky. styl definovaný v elementu vrchní úrovně je použit jako výchozí pro podřízené elementy; 0-1 povinné
<TextStyle>	ID FONTSTYLE FONTFAMILY FONTSIZE	definuje font textu; 0-n povinné --- ID pro každý text style použitý v OCR souboru – povinné FONTSTYLE – např. bold, italics apod.; doporučujeme používat 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 povinné

		---------- ID pro každý odstavec + zarovnání; např. PAR_LEFT, PAR_RIGHT, PAR_CENTER, PAR_BLOCK; povinné ALIGN – zarovnání; povolené hodnoty: Left, Right, Center, Block aj.; povinné
<Layout>		layout - rozložení struktur (slov, odstavců apod.) na jedné stránce dokumentu; 1-1 povinný výskyt element není opakovací povinné
<Page>	ID ACCURACY POSITION QUALITY PHYSICAL_IMG_NR HEIGHT WIDTH PC	element popisující jednu stránku dokumentu; 1-n povinné ---------- ID – vygenerovaný identifikátor stránky, např. PAGE1, nebo P1 apod.; povinné ACCURACY – procentuální odhad přesnosti OCR (0-100); 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 povinné ---------- ID: unikátní ID pro element 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
<LeftMargin>	ID HPOS VPOS WIDTH HEIGHT	levý okraj – prostor mezi levým okrajem stránky a textem; 0-1 povinné ---------- 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
<RightMargin>	ID HPOS VPOS WIDTH HEIGHT	pravý okraj – prostor mezi pravým okrajem stránky a textem; 0-1 povinné ---------- 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
<BottomMargin>	ID HPOS VPOS WIDTH HEIGHT	pravý okraj – prostor mezi spodním okrajem stránky a textem; 0-1 povinné ---------- 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; povinné
<PrintSpace>	ID HPOS VPOS WIDTH HEIGHT	popis tvaru pokrývajícího textové pole stránky; 0-1 povinné ---------- 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
<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 > povinné ---------- 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 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>; doporučeno – v případě, že je tvar textového bloku nestandardní (víceúhelník)
<Polygon>	POINTS	popis (souřadnice) tvaru víceúhelníku; 0-1 povinné --- POINTS – vyjádření jednotlivých bodů víceúhelníku; povinné
<TextLine>	ID STYLEREFS HPOS VPOS WIDTH HEIGHT	popis jedné řádky textu v rámci textového bloku; 1-n povinný - 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 HPOS: horizontální pozice řádky; povinné VPOS: vertikální pozice řádky; povinné WIDTH – šířka řádky; povinné HEIGHT – výška řádky; povinné
<String>	ID	řetězec znaků – vlastní obsah OCR;

	CONTENT HEIGHT WIDTH HPOS VPOS CC WC V případě dělení slov také: SUBS_TYPE SUBS-CONTENT	znaky tvoří jednotlivá slova a více tagů <String> větu <TextLine>; 1-n v rámci <TextLine> povinné ---------- 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 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 *HypPart1* se vyskytuje při rozdělení slova u jeho první OCR části (u první části tagu <CONTENT> ve větě (stringu) první; *HypPart2* se vyskytuje u následujícího tagu <CONTENT> v následující větě (stringu), který obsahuje druhou část rozděleného slova/řetězce; *Abbreviation* – typ substituce používaný při rozepisování zkratek 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; nepovinné
<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 dělícího znaku; doporučené HPOS: horizontální pozice dělícího znaku; doporučené VPOS: vertikální pozice dělícího znaku; doporučené
<SP>	ID WIDTH HPOS	prázdný prostor mezi řádky; 0-n v rámci jednoho <TextLine>; vždy mezi řádky, tj. mezi tagy <String>;

	VPOS	povinné --- 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/TextBlock, 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 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>		tvar 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 tvar komponovaného bloku nestandardní (víceúhelník)
<Polygon>	POINTS	popis tvaru víceúhelníku; 0-1 povinné --- POINTS – vyjádření jednotlivých bodů víceúhelníku povinné
<TextBlock>	ID STYLEREFS HPOS VPOS WIDTH HEIGHT	v případě, že komponovaný blok (např. orámovaný tvar) obsahuje text; 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>; povinné 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/ComposedBlock označuje rozměry tvaru v rámci něhož je tabulka, ilustrace, reklama apod.; 0-1 - pro jeden výskyt /alto/Layout/Page/PrintSpace/ComposedBlock 0 nebo max. 1 výskyt <GraphicalElement>; povinné - 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