národní
úložiště
šedé
literatury

Metodika pro tvorbu balíčků SIP se zaměřením na digitalizáty tištěných dokumentů

Cubr, Ladislav,; Ostráková, Natalie; Kočišová, Pavlína
2020

Dostupný z http://www.nusl.cz/ntk/nusl-432324

Dílo je chráněno podle autorského zákona č. 121/2000 Sb.

Licence Creative Commons Uveďte autora-Neužívejte dílo komerčně-Nezasahujte do díla 3.0 Česko

Tento dokument byl stažen z Národního úložiště šedé literatury (NUŠL).

Datum stažení: 26.05.2024

Další dokumenty můžete najít prostřednictvím vyhledávacího rozhraní nusl.cz .

Metodika pro tvorbu balíčků SIP se zaměřením na digitalizáty tištěných dokumentů

2019

Ladislav Cubr, Natalie Ostráková, Pavlína Kočišová

Metodika vznikla na základě institucionální podpory dlouhodobého koncepčního rozvoje výzkumné organizace poskytované Ministerstvem kultury.

Obsah

Oponenti

1) Mgr. Zdeněk Hruška, Moravská zemská knihovna, Odbor digitalizace

2) Ing. Martin Lhoták, Knihovna Akademie věd ČR, v. v. i.

I. TEORETICKÁ ČÁST

Úvod

Metodika pro vytváření balíčku SIP pro digitalizáty tištěných dokumentů (dále jen jako „Metodika pro balíčky SIP“) je dokument, který popisuje procedurální postup pro užití standardů NDK pro metadata, formáty a obrazová data užívaná při digitalizaci tištěných dokumentů v českých knihovnách. Standardy NDK jsou závazné pro Národní knihovnu ČR a Moravskou zemskou knihovnu pro digitalizaci v rámci projektu Vytvoření Národní digitální knihovny, a dále pro knihovny digitalizující na základě podpory získané z podprogramu VISK7, případně pro další organizace, které budou odevzdávat svoje digitalizáty do LTP úložiště NK ČR k dlouhodobému uchovávání. Metodika pro balíčky SIP obsahuje i některá obecnější doporučení, která lze vztahovat i na jiné typy dokumentů než digitalizáty tištěných dokumentů (monografií a periodik), primárně se však zaměřuje na tento typ dokumentu.

Podobná metodika v českém prostředí chybí. Existuje pouze Metodika pro vytváření bezpečnostních kopií archiválií v digitální podobě (Dvořák a kol., 2015), která byla vytvořena Národním archivem ČR. Tato metodika se však vztahuje na jiný typ dokumentu (digitalizáty archiválií) a pokrývá celý životní cyklus dokumentu. V oblasti digitalizace podrobněji popisuje některé technologické otázky digitalizace (výběr a kalibrace skeneru, barevná specifikace apod.), ale nikoliv procedurální postup v podobě, jako tato metodika.

V zahraniční existují metodiky pro vytváření digitalizátů zaměřené na otázku obrazových dat a částečně metadat. Za nejvýznamnější v tomto směru lze považovat směrnice „Technical Guidelines for Digital Cultural Content Creation Programmes“ z roku 2008 a směrnice americké iniciativy FADGI (Federal Agencies Digital Guidelines Initiative), které byly vytvořeny pro paměťové instituce v USA a které nesou název „Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Files“ (FADGI, 2010; FADGI, 2016). Směrnice vyšly v roce 2010, v roce 2016 pak byla vydána jejich revidovaná verze. Tyto poskytují zejména technologické specifikace z hlediska vlastního snímání a otázky kontroly kvality, nejsou však specifické pro užití konkrétního metadatového profilu a formátu. Tyto směrnice lze doporučit pro potřeby výběru snímacího zařízení, jeho kalibrace, volbu prostorového rozlišení, způsob zpracování obrazových dat a práci s barevným prostorem. Tyto směrnice jsou také využity v teoretické a praktické části předkládané metodiky. Metodika pro

balíčky SIP obsahuje specifická procedurální doporučení pro vytváření konkrétního typu digitálního dokumentu s užitím konkrétního metadatového a formátového profilu (které jsou popsané v standardech NDK) a konkrétních nástrojů v českém prostředí digitalizační projektů knihoven.

Mimo záběr metodiky je dlouhodobé uchovávání i zpřístupňování digitalizátů. Pro dlouhodobé uchovávání lze užít Metodiku logické ochrany digitálních dat, která vznikla v projektu NAKI ARCLib[1]. Tato metodika obsahuje v teoretické a praktické části pasáže, které lze do určité míry užít pro dlouhodobé uchovávání digitálních dokumentů v různých organizacích užívajících různá softwarová řešení. Její implementační část je pak zaměřena specificky na konkrétní softwarový systém, systém ARCLib.

Specifičnost pro určitý typ dokumentu, konkretizace metadatových, formátových a datových (obrazových, textových ve smyslu OCR) otázek je to, co činí Metodiku balíčky SIP novou v českém prostředí. Je určena všem organizacím, zejména knihovnám, které využívají standard NDK, ať již povinně, neboť pak odevzdávají své balíčky SIP do LTP úložiště NK ČR k dlouhodobému uchovávání, nebo dobrovolně, jako způsob standardizace produkce. V obou případech je účelem standardů NDK i Metodiky pro vytváření balíčků SIP účelem vytvářet data a metadat při digitalizaci v takové podobě, aby při vytváření balíčku AIP nebylo třeba užít formátovou normalizaci (čímž dochází k minimalizaci nákladů pro archiv) a aby byly během produkce dokumentů zachyceny všechny důležité informace (metadata) pro následné dlouhodobé uchovávání a dodrženy postupy, přičemž obojí by mělo zaručit zvýšení kvality vytvářených digitalizátů, jejich autenticity a předjímaných požadavků dlouhodobého uchovávání i zpřístupňování.

1 Obecná část

1.1 Model OAIS

Základní konceptuálním rámcem pro oblast dlouhodobého uchovávání digitálních dokumentů (digitální archivace) je model OAIS (Open Archival Information System) obsažený v normě ISO 14721 (česky ČSN ISO 14721). Tento model je již od počátku milénia (kdy byl již znám jako návrh normy) základním a obecně přijímaným referenčním rámcem pro řízení

[1] https://invenio.nusl.cz/record/371612/files/content.csg.pdf

životního cyklu digitálních dokumentů z hlediska digitální archivace. Model OAIS se primárně zabývá archivací a zpřístupněním digitálních dokumentů, ale do určité míry vymezuje i otázku produkce digitálních dokumentů. Činí tak v podobě konceptu dohody o dodávání dat (submission agreement), která je vyjádřením toho, v jaké podobě budou digitální objekty dodávány do archivu, a je výsledkem dohody mezi producentem (vkladatelem) a archivem. Tato podoba je v normě ČSN ISO 14721 specifikována jako tzv. vstupní informační balíček (submission information package, balíček SIP). V praxi může docházet k tomu, že archiv nemá možnost ovlivnit tuto podobu způsobem, který odpovídá požadavkům digitální archivace, například tehdy, pokud má legislativou stanovené povinnosti přijímat a archivovat digitální publikace od vydavatelů a současně platí, že a) legislativa vydavatelům neukládá povinnost dodávat digitální publikace do archivu ve formátech, které archiv preferuje (z důvodů jejich vhodnosti pro archivaci), b) archiv nemá možnost provádět formátovou konverzi publikací do požadované podoby (např. z důvodu neexistence adekvátních nástrojů pro takovou konverzi). Optimální situací je, pokud archiv může specifikovat podobu balíčku SIP tak, aby při převodu do balíčku AIP nebylo potřeba provádět formátovou normalizaci. To je možné v případě, kdy producent i archiv jsou jedna a tatáž organizace (a tudíž za cílové formáty digitalizace lze zvolit formáty vhodné pro archivaci), a kdy technologie pro danou oblast jsou dostupné (tj. existují vhodné formáty a adekvátní konverzní nástroje), nebo kdy je archiv taková organizace, která má pravomoc nebo autoritu (jako metodické centrum) vydávat pokyny pro standardizaci pro producenty v nějaké oblasti (např. knihovnictví).

1.1.1 Koncept archivu

Norma ISO 14721 pojednává o specifickém modelu archivu, který označuje jako „otevřený archivační informační systém“ (open archival information system), zkráceně archiv OAIS a definuje jej jako „archiv (archive), který se sestává z organizace tvořené lidmi a systémy, jež přijala odpovědnost uchovávat informace a zpřístupňovat je cílové komunitě (designated community), přičemž tento archiv může být součástí větší organizace.“ (ISO 14721:2012, s. 24)

Norma ISO 14721 vymezuje šest obecných závazných povinnosti, které musí daná organizace plnit, aby mohla provozovat archiv OAIS. Tyto povinnosti odlišují pojem „archiv OAIS“ od jiných užití pojmu „archiv“. Jde o následující povinnosti (ISO 14721:2012, s. 39) :

1. Vyjednávat s producenty informací a přijímat od nich příslušné informace.

2. Získávat možnost s poskytnutými informacemi dostatečně nakládat, aby bylo možné zajistit jejich dlouhodobé uchovávání.
3. Určit, ať již samostatně nebo ve spolupráci s dalšími stranami, které komunity by se měly stát cílovými komunitami, a tudíž by měly být schopny porozumět poskytovaným informacím. Tím je vymezena znalostní základna dané skupiny.
4. Zajistit, aby informace určené k uchovávání byly pro cílovou komunitu srozumitelné samy o sobě. Cílová komunita by měla být schopna informacím porozumět bez využití odborných zdrojů, například bez rady odborníků, kteří informace vytvořili.
5. Dodržovat zdokumentovaná pravidla a postupy, které zajistí, že informace budou chráněny před všemi možnými nepředvídatelnými událostmi (včetně zániku archivu), a zajistit, že informace nebudou nikdy smazány (s výjimkou případu, kdy jejich smazání bude součástí schválené strategie). Nemělo by docházet k žádnému jednorázovému mazání dat.
6. Zpřístupňovat uchovávané informace cílové komunitě a umožňovat šíření informací v podobě kopií původně dodaných datových objektů nebo v takové podobě, aby bylo možné zpětně dohledat (*as traceable to*), ke kterým původně dodaným datovým objektům se vztahují, a to společně s doklady o jejich autenticitě.

První povinnost zahrnuje zejména dohodu o dodávání dat. Druhá povinnost zahrnuje vyřešení otázky práv duševního vlastnictví (pokud archiv nemá právo dodaná data měnit, například je převést do jiného formátu, je jeho činnost de facto znemožněna). Třetí povinnost se týká cílové komunity z hlediska jejího vymezení. Čtvrtá povinnost se týká cílové komunity s ohledem na související koncepty nezávislé srozumitelnosti (tj. informace srozumitelné samy o sobě) a interpretačních informací. Pátá povinnost specificky pojednává o problematice dlouhodobého uchovávání navzdory změnám technologií (datových nosičů, formátů apod.). Šestá povinnost se týká zpřístupňování informací a specificky zohledňuje otázku autenticity.

Za účelem plnění těchto funkcí norma přináší tzv. funkční model. Funkční model OAIS vymezuje šest základních funkčních celků (*functional entities*) archivu OAIS. Funkční celek je konceptuální model, představující kategorii činností (funkcí a služeb), které musejí být archivem vykonávány, aby byl archivem OAIS. Výkon těchto činností pro digitální archivaci zajišťují počítačové technologie stejně jako lidští pracovníci (nejen jako obsluha, ale i jako tvůrci návrhu softwaru apod.). Těmito funkčními celky jsou: Příjem, Archivní úložiště, Správa dat, Administrace, Plánování uchovávání a Zpřístupnění (ISO 14721:2012, s. 45). Sedmým funkčním celkem jsou základní služby (*common services*), jimiž jsou „podpůrné služby

nezbytné pro provoz archivu OAIS“ (ISO 14721:2012, s. 20). Funkční model není pro tuto metodiku podstatný, vzhledem k tomu, že se zabývá doporučeními pro producenta. Důležitější je vymezení prostředí, ve kterém se archiv nachází.

1.1.2 Prostředí archivu OAIS

Norma ISO 14721 vymezuje okolí, ve kterém se archiv OAIS nachází a které určuje jeho vstupy a výstupy, zejména z hlediska informací.

Producent (*producer*) je „úloha vykonávaná osobami nebo klientskými systémy poskytujícími informace určené k uchovávání; může se jednat o další archivy OAIS nebo také o osoby či systémy v daném archivu OAIS“ (ISO 14721:2012, s. 25). Producent je podle normy ISO 14721 role spočívající v dodávání informací do archivu OAIS. V praxi masové digitalizace knihoven vykonává úlohu producenta, archivu i managementu často jedna a tatáž knihovna.

Koncový uživatel (*consumer*) je „úloha vykonávaná osobami nebo klientskými systémy, které využívají služeb archivu OAIS za účelem nalezení a vlastního zpřístupnění uchovávaných informací; tuto úlohu mohou vykonávat další archivy OAIS nebo též osoby nebo systémy z daného archivu OAIS.“ (ISO 14721:2012, s. 21).

Management (*management*) je „úloha vykonávaná těmi, kdo určují celková pravidla archivu OAIS jako součást širších pravidel, například v rámci větší organizace“ ((ISO 14721:2012, s. 24). Management může například ve vztahu k archivu schvalovat zřizovací listinu, určovat rozsah působnosti, být hlavním zdrojem financování nebo vyhodnocovat výkon.

Z definice normy ISO 14721 vyplývá, že hlavní činností producenta je dodávat informace do archivu. Tato norma však dále vymezuje také některé další dílčí činnosti, ze kterých vyplývá, že producent je často též samotným producentem informací (a samozřejmě sám výraz „producent“ tuto funkci jasně implikuje).

Archiv OAIS podle normy ISO 14721 uzavírá s vkladatelem dohodu o dodávání dat (*submission agreement*), což je „dohoda uzavřená mezi archivem OAIS a producentem, která stanovuje datový model a další potřebná nastavení pro relaci dodávání dat (*data submission session*); datový model určuje formát/obsah a logické konstrukty užívané producentem a způsob, jakým jsou reprezentovány na všech dodaných datových nosičích nebo při všech telekomunikačních spojeních“ (ISO 14721:2012, s. 26). Relace dodávání dat je „jednotlivá

dodávka datového nosiče nebo telekomunikační spojení, kterými jsou archivu OAIS poskytována data“ (ISO 14721:2012, s. 22).

Tato dohoda by podle normy ISO 14721 měla vždy v nějaké podobě existovat, ale nemusí jít vždy o formální podobu (smlouvu). Uváděným příkladem je webarchiv (jako typ archivu OAIS, který uchovává sklizený webový obsah), kde dohoda o dodávání dat nabývá podobu nastavení sklízecího robota (ISO 14721:2012, s. 36).

S koncovým uživatelem pak archiv OAIS uzavírá dohodu o objednávce (*order agreement*), což je „dohoda mezi archivem a koncovým uživatelem, v níž jsou stanoveny údaje o dodání, například typ datového nosiče a formát dat“ (ISO 14721:2012, s. 24). Tato dohoda opět nemusí být formální a ustanovení normy lze interpretovat tak, že dohodou o objednávce může být jednoduše to, že uživatel v digitální knihovně (jako součásti archivu OAIS) vyhledá požadovaný dokument.[2] Rozdíl mezi koncovým uživatelem a cílovou komunitou spočívá v tom, že koncový uživatel je jakýkoliv subjekt, který interaguje s archivem OAIS s cílem získání informací (tedy i softwarový systém). Člen cílové komunity je takový koncový uživatel, na základě jehož znalostní základny se udržují informace tak, aby byly srozumitelné samy o sobě.

1.1.3 Informační model OAIS

Z hlediska této metodiky je klíčový informační model OAIS, protože ten se vztahuje i na balíčky SIP.

Pojem „informace“ (*information*) definuje norma ISO 14721 jako „jakékoliv znalosti (*knowledge*), které mohou být předmětem výměny (*exchange*)“ a udává, že při výměně jsou informace „vždy vyjádřeny (tj. reprezentovány) určitým typem dat“ (ISO 14721:2012, s. 30). Data jsou definována jako „opakovaně interpretovatelná reprezentace informací ve formalizované podobě vhodné pro komunikaci, interpretaci nebo zpracování.“ (ISO 14721:2012, s. 21). Pojmy informace a data jsou kategorie. Jednotlivý objekt spadající do první kategorie nazývá norma ISO 14721 informační objekt (*information object*), objekt druhé kategorie datový objekt (*data object*).

Při výměně tedy příjemce získává informace vždy z dat, v tom smyslu, že převádí datový objekt na informační objekt. Aby se tento proces mohl uskutečnit, musí příjemce disponovat odpovídající znalostní základnou (*knowledge base*), což je „množina informací,

[2] Zejména viz ISO 14721:2012, s. 38

které si osvojila osoba nebo systém a která této osobě nebo tomuto systému umožňuje porozumět přijímaným informacím“ (ISO 14721:2012, s. 23).

Porozumění informací příjemcem je tedy chápáno jako převod datového objektu na informační objekt užitím znalostní základny příjemce. Pokud si například český čtenář zapůjčí knihu „Information Science in Theory and Practice“ Briana Vickeryho, pak pozorovatelné znaky (vytištěná slova) představují datový objekt. Aby příjemce knize rozuměl, tj. mohl převést tyto znaky na informace, musí umět anglicky. Pokud tomu tak je, pak to znamená, že znalost angličtiny tvoří součást jeho znalostní základny. Čtenář, který anglicky neumí, bude potřebovat získat dodatečné informace, aby knize rozuměl. Tento typ informací norma OAIS označuje za interpretační informace[3] (*representation information*) a definuje je jako „informace, které převádějí datový objekt do smysluplnějších významových celků (*the information that maps a data object into more meaningful concepts*)“ (ISO 14721:2012, s. 25). Tato definice působí poněkud krkolomně, nicméně jejím smyslem je poukázat na skutečnost, že srozumitelnost lze posuzovat ve stupních. Ve výše uvedeném příkladu i čtenář, který anglicky neumí, rozumí tomu, že text je v angličtině; pokud by nevěděl, že text je v angličtině, mohl by rozumět alespoň tomu, že text je v cizím jazyce; na nejnižším stupni (např. pokud je negramotný) pak může stále rozumět alespoň tomu, že datový objekt je kniha. Interpretační informace jsou tedy informace, které popisují formu reprezentace informací v datech (v našem příkladu je jí anglický jazyk). Tyto informace potřebuje získat ten, jehož znalostní základna informace o formě reprezentace neobsahuje. Čtenář, který neumí anglicky, si například může opatřit českou učebnici angličtiny a anglicko-český slovník, což jsou datové objekty, ze kterých již na základě své znalostní základny dokáže získat informace: učebnice a slovník tedy čtenáři poskytují interpretační informace, které potřebuje pro porozumění anglicky psané knize.

Datový objekt norma ISO 14721 definuje výčtem – datový objekt je „buď fyzický objekt, nebo digitální objekt“ (ISO 14721:2012, s. 21). Fyzický objekt (*physical object*) je „objekt (například měsíční hornina, biologický vzorek, mikroskopické sklíčko) s fyzicky pozorovatelnými vlastnostmi, které reprezentují informace, jež je pro účely uchovávání, šíření

[3] V této metodice je užíván pro representation information český překlad „interpretační informace“. Není to doslovný překlad, ten by byl „informace o reprezentaci“. Podle soudu autorů však lépe vystihuje skutečnost, že veškerá data musejí být předmětem interpretace.

a samostatného využívání vhodné patřičně zdokumentovat“ (ISO 14721:2012, s. 24). Digitální objekt (*digital object*) je „objekt složený z množiny bitových posloupností (*a set of bit sequences*)“ (ISO 14721:2012, s. 22). Informační objekt (*information object*) je pak „datový objekt se svými interpretačními informacemi“ (ISO 14721:2012, s. 23).

Informační model přináší typologii informačních objektů, které se vyskytují v průběhu životního cyklu informací. Z nich musíme vyčlenit jako klíčový informační objekt ten, který norma ISO 14721 nazývá informační obsah a který je podle normy hlavním předmětem dlouhodobého uchovávání v archivu OAIS; dále pak zmíněné interpretační informace. Kategorie informací, které slouží jako nezbytné dodatečné informace, jež musejí být uchovávány spolu s informačním obsahem, nazývá norma archivační informace. Ty obsahují pět podkategorií informací. Dalším důležitým prvkem informačního modelu je informační balíček jako logická schránka, která zpravidla obsahuje informační obsah a archivační informace.

1.1.3.1 Informační obsah a interpretační informace

Informační obsah (*content information*) je podle normy ISO 14721 „množina informací, která je původním předmětem uchovávání nebo která obsahuje část těchto informací či všechny tyto informace; informační obsah je informační objekt složený ze svého datového objektu s obsahem a svých interpretačních informací.“ (ISO 14721:2012, s. 21) Původní vydání normy uvádí jako příklad informačního obsahu tabulku s čísly, která reprezentuje teplotní údaje (a která je chápána jako teplotní údaje), s tím, že součástí informačního obsahu již není dokumentace, který by vysvětlovala historii a původ těchto teplotních údajů nebo to, jak se vztahují k jiným pozorováním.

Datový objekt s obsahem (dále zkráceně jako objekt CDO) tedy nese informační obsah, pokud je příjemcem spojen s odpovídajícími interpretačními informacemi (které mohou, ale nemusejí být součástí jeho znalostní základny). Je třeba upozornit, že digitálním objektem podle normy ISO 14721 není nutně soubor – může jim být, stejně jako jím může být například množina souborů. V případě objektu CDO může jít o jeden soubor ve formátu PDF, který reprezentuje digitální knihu, nebo stovku souborů ve formátu JPEG, které dohromady reprezentují sto stran knihy – záleží na konkrétní implementaci. Jednotlivý digitální objekt CDO v archivu je tedy taková podmnožina bitových posloupností z celkové množiny všech bitů uložených na datových nosičích archivu, které jsou potřebné k transformaci bitů do daného konkrétního informačního obsahu.

Interpretační informace člení norma ISO 14721 na tři druhy:

Strukturální interpretační informace (*structure information*) jsou takové „interpretační informace, které udávají, jak jsou další informace složeny; mohou například převádět bitové toky (*bit streams*) na základní typy dat, jako jsou znaky, čísla a pixely, a na seskupení těchto typů dat, jako znakové řetězce a pole“ (ISO 14721:2012, s. 26).

Sémantické interpretační informace (*semantic information*) jsou takové „interpretační informace, které podrobněji popisují význam nesený strukturálními interpretačními informacemi.“ (ISO 14721:2012, s. 26).

Třetí druh interpretačních informací zavedlo až druhé vydání normy – jde o ostatní interpretační informace (*other representation information*), které jsou definovány jako takové „interpretační informace, které nelze jednoduše zařadit mezi strukturální interpretační informace nebo sémantické interpretační informace; například k porozumění datovému objektu s obsahem mohou být potřeba software, algoritmy, šifrování, psané pokyny atd., přičemž všechny tyto informace budou odpovídat definici interpretačních informací, byť nebude zřejmé, zda se vztahují ke struktuře nebo významu; informace udávající vztah mezi strukturálními interpretačními informacemi a sémantickými interpretačními informacemi nebo popisující software potřebný pro zpracování databázového souboru lze také považovat za ostatní interpretační informace.“ (ISO 14721:2012, s. 24).

Výše byl uveden příklad interpretačních informací potřebných k převedení fyzického objektu (tištěné knihy) do informačního obsahu (tedy člověku srozumitelného intelektuálního obsahu). Lze říci, že učebnice angličtiny a slovník jsou datové objekty, které reprezentují sémantické interpretační informace. Znalostní základna čtenáře, který si je opatřil, přitom již předtím obsahovala strukturální interpretační informace, jež mu umožňovaly ve fyzickém objektu vidět lineární text (strukturu).

Nás zde zajímá především otázka interpretace digitálních objektů. Lze říci, že problematika interpretačních informací tvoří jádro digitální archivace. Digitální objekt musí být v daný okamžik vždy uložen na nějakém datovém nosiči, který je sice fyzickým objektem (pevný disk, BluRay Disc apod.), ale data na něm zapsaná člověk vnímat nedokáže, natož je převést do srozumitelné podoby. Z tohoto hlediska je nutné, aby interpretaci digitálního objektu zprostředkovávala počítačová technologie.

V praxi digitální archivace se klade největší důraz na strukturální interpretační informace, zejména na formát souboru. Formáty patrně nejviditelněji odrážejí problém zastarávání digitálních technologií.

Z hlediska své činnosti musí archiv OAIS shromáždit dostatečné interpretační informace k objektu CDO. Způsoby konkrétního řešení se mohou lišit podle posouzení situace archivem. V případě formátu to může znamenat získat formátovou specifikaci a uložit ji spolu s objektem CDO nebo jen zaznamenat technické informace o formátu, na základě kterých je možno tyto interpretační informace dohledat (tj. předpoklad, že jsou tyto informace běžně dostupné, a tedy samy o sobě srozumitelné pro cílovou komunitu). Tyto technické informace jsou v praxi zaznamenávány do datových objektů, které se označují jako technická metadata. Z informačního modelu OAIS je zřejmé, že i k těmto metadatům (tak jako k jakýmkoliv jiným typům dat), je nutné získat dostatečné interpretační informace. Tento proces podle normy ISO 14721 v digitálním světě znamená rekurzivitu – interpretační informace o technických metadatech budou zaznamenány v dalším digitálním objektu (např. v PDF popisujícím metadatovým standard). Tato rekurzivita podle normy končí tehdy, pokud jsou interpretační informace zaznamenány v podobě fyzického objektu (tedy např. vytištěný standard). V praxi to v současnosti není považováno za větší problém vzhledem k tomu, že metadata jsou zpravidla ukládána v textových formátech (zejména v XML), jejichž interpretace se nepovažuje za problematickou.

Norma ISO 14721 uvádí dva typy specializovaných softwarových nástrojů pro interpretační informace (tyto nástroje samy patří mezi ostatní interpretační informace). Software pro zobrazení interpretačních informací (*representation rendering software*)[4] je software, který umožňuje interpretační informace reprodukovat v podobě srozumitelné lidem. Příkladem je prohlížeč Adobe Acrobat, který dokáže zobrazit formátovou specifikaci uloženou v PDF. Zpřístupňovací software (*access software*) je software, který dokáže prezentovat samotný informační obsah nebo jeho část. Jinými slovy, software pro zobrazení interpretačních informací reprodukuje interpretační informace a slouží jako pomůcka pro uchovávání informačního obsahu, zatímco zpřístupňovací software reprodukuje informační obsah a slouží jako prostředek pro zpřístupňování informačního obsahu cílové komunitě.

Na tomto místě je třeba upozornit, že norma ve své obecnosti nepopisuje odlišnosti užití různých typů informačního obsahu. Archiv OAIS musí uchovávat objekt CDO spolu

[4] Anglický název je zavádějící, měl by být spíše „representation information rendering software).

se zvolenou množinou interpretačních informací (které bude uchovávat v podobě dalšího digitálního objektu, zpravidla nazývaného metadata). Některé typy informačního obsahu mohou být určeny k reprodukci lidským uživatelům, jako tomu bývá v případě dokumentů digitálního dědictví (např. zobrazení v digitální knihovně), zatímco jiné nemusí být vůbec určeny pro vnímání člověkem, ale pro zpracování jinými systémy (to je příklad rozsáhlých datových sad získaných z pozorování, které mohou být reprezentovány v podobě tabulek, obsahovat numerické údaje a sloužit jako zdroj pro automatizované analýzy).

Za jeden ze způsobů, jak řešit otázku interpretačních informací, bylo určeno vytvoření mezinárodních registrů interpretačních informací (zejména pro formáty). Otázce těchto registrů a bližšímu přiblížení problematiky formátů se věnuje Kapitola 3.

Závěrem tohoto shrnutí je potřeba zopakovat důraz na rozlišování interpretačních informací na jedné straně a metadat, které je reprezentují, na straně druhé. Metadata jsou datové objekty a potřebují rovněž své interpretační informace. Rekurzivitu v praxi ukončuje užití souborů v textových formátech typu XML, jejichž interpretace se z hlediska digitální archivace považuje (vzhledem k jednoduchosti souborového formátu) za relativně bezproblémovou.

1.1.3.2 Archivační informace

Informační model OAIS definuje archivační informace (*preservation description information*) jako „informace, které jsou nezbytné k adekvátnímu uchovávání jednotlivého informačního obsahu“ (ISO 16363:2012, s. 25). Jde o kategorii informací, která obsahuje pět podkategorií.

Identifikační informace (*reference information*) jsou „informace, která jsou využívány jako identifikátor informačního obsahu“ (ISO 14721:2012, s. 25). Tyto informace mohou také zahrnovat „identifikátory, které vnějším systémům umožňují jednoznačně odkazovat na konkrétní informační obsah“ a udávat a popisovat způsob jejich přidělování (ISO 14721:2012, s. 74). Uváděným příkladem z oblasti knihoven je perzistentní identifikátor (konkrétním příkladem pak ISBN) a bibliografický popis. Tyto informace tedy mohou zahrnovat i globální perzistentní identifikátory, na základě nichž mohou uživatelé vyhledávat konkrétní informační obsah.

Provenienční informace (*provenance information*) jsou „informace, které dokumentují historii informačního obsahu; tyto informace vypovídají o původu nebo zdroji informačního obsahu, o veškerých změnách, které mohly od doby jeho vzniku nastat, a o tom, kdo o něj od

doby jeho vzniku pečoval“ (ISO 14721:2012, s. 25). Jak dále norma uvádí, archiv nese odpovědnost za vytváření a uchovávání těchto informací až od okamžiku jejich příjmu do archivu; provenienční informace z dřívější doby by měl poskytnout vkladatel. Uváděnými příklady z oblasti knihoven jsou metadata o procesu uchovávání (ukazatele k předchozím verzím jednotky, historie změn); metadata o procesu digitalizace.

Informace o neporušenosti (*fixity information*) jsou „informace, které udávají, jak je zajištěno, aby objekt s informačním obsahem nebyl nezdokumentovaným způsobem změněn“ (ISO 14721:2012, s. 22). Tyto informace „fungují jako obal nebo ochranný štít, který chrání informační obsah“ (ISO 14721:2012, s. 34). Příkladem je digitální otisk.

Kontextuální informace (*context information*) jsou „informace, které dokládají vztah informačního obsahu k jeho okolí; patří mezi ně důvod vytvoření informačního obsahu a jeho vztah k dalším objektům s informačním obsahem.“ (ISO 14721:2012, s. 21).

Informace o přístupových právech (*access rights information*) jsou „informace, které udávají omezení týkající se přístupu k informačnímu obsahu, a to včetně právního rámce, licenčních podmínek a řízení přístupu“ (ISO 14721:2012, s. 19).

1.1.3.3 Informační balíček

Informační balíček je „logická schránka, která může obsahovat informační obsah a archivační informace; k tomuto informačnímu balíčku jsou připojeny informace o zabalení (*packaging information*), které vymezují a určují informační obsah, a informace o popisu balíčku (*package description*), které usnadňují vyhledání informačního obsahu“ (ISO 14721:2012, s. 23).

Norma dále odlišuje tři typy informačních balíčků:

- Archivní informační balíček (*archival information package*) je „informační balíček, který je složen z informačního obsahu a přidružených archivačních informací a je uchováván v archivu OAIS“ (ISO 14721:2012, s. 23). Dále bude nazýván jen jako balíček AIP.
- Vstupní informační balíček (*submission information package*) je „informační balíček, který dodává producent do archivu OAIS tak, aby mohl být využit při sestavení nebo aktualizaci jednoho nebo více AIP a/nebo přidružených popisných informací (*descriptive information*)“ (ISO 14721:2012, s. 26). Dále bude nazýván jako balíček SIP.

- Výstupní informační balíček (dissemination information package) je „informační balíček odvozený z jednoho nebo více balíčků AIP a zaslaný archivem OAIS koncovému uživateli jako odpověď na jeho požadavek vůči tomuto archivu (ISO 14721:2012, s. 22). Dále bude nazýván jako balíček DIP.

Klíčový informační balíček z hlediska archivu OAIS je balíček AIP. Ten musí obsahovat informační obsah a přidružené archivační informace. Definice nadřazené kategorie (tj. informačního balíčku) uvádí, že tyto dvě složky může obsahovat. To znamená, že při dodávání balíčků SIP do archivu nebo vydávání balíčků DIP koncovým uživatelům nemusí každý jednotlivý balíček SIP nebo DIP vždy obsahovat informační obsah a archivační informace. Do archivu mohou být například dodávány odděleny balíčky SIP s informačním obsahem a balíčky SIP s archivačními informacemi. Podobně mohou být archivem vydávány, v závislosti na požadavcích koncových uživatelů například, jen balíčky DIP obsahující informační obsah, ale již nikoliv archivační informace.

Informace o zabalení (*packaging information*) jsou „informace, které slouží k propojení a popisu součástí informačního balíčku“ (ISO 14721:2012, s. 24). Popis balíčku (package description) jsou „informace určené pomůckám pro zpřístupnění“ (ISO 14721:2012, s. 24). Těmito pomůckami norma míní „softwarový program nebo dokument, který koncovým uživatelům umožňuje najít, analyzovat, objednat nebo získat informace z archivu OAIS“ (ISO 14721:2012, s. 19). Popisné informace (*descriptive information)* jsou „množina informací, která je složena především z popisů balíčků a je poskytována správě dat za účelem podpory koncových uživatelů při objednávání a získávání informačních jednotek z archivu OAIS“ (ISO 14721:2012, s. 22).

Hlavním účelem informací o zabalení je vymezit, které části balíčku AIP jsou informační obsah a které archivační informace. To znamená popsat jednak to, které soubory v balíčku tvoří objekt CDO, jež reprezentuje informační obsah (např. pět souborů ve formátu JPEG obsahujících článek) a které technická metadata reprezentující interpretační informace (např. jeden soubor v XML popisující formát JPEG). Objekt CDO a technická metadata dohromady tvoří informační obsah. Dále je třeba popsat, jaké soubory jsou archivační metadata reprezentující archivační informace (např. druhý soubor v XML obsahující bibliografický popis). V popisu balíčků a v popisných informacích se pak podle normy obvykle opakují identifikační informace. Pokud budeme uvažovat digitální knihu v PDF, pak informační obsah může tvořit jeden soubor ve formátu PDF (objekt CDO) a jeden soubor ve formátu XML obsahující technická metadata (např. informace o verzi a typu, příkladem je PDF/A-2u);

archivační informace jeden soubor v XML s archivačními metadaty (blíže o metadatech bude pojednávat Kapitola 1.3), přičemž jejich součástí bude identifikátor ISBN (který bude současně obsažen i v samotném informačním obsahu, jelikož je vydavatelskou praxí uvádět jej přímo v knize) a tento identifikátor bude také obsažen v popisu balíčku a popisných informacích. Na základě tohoto identifikátoru pak čtenář jakožto koncový uživatel může knihu vyhledat v archivu.

Předmětem dlouhodobého uchovávání je informační obsah (např. konkrétní kniha v PDF) i přidružené archivační informace uložené v balíčku AIP spolu s informačním obsahem. Archiv vytváří balíček AIP z balíčku SIP (nebo z více balíčků SIP), který mu dodá vkladatel. Na základě dohody o dodávání dat oba subjekty (archiv a vkladatel) definují datový model, tedy podobu dodávaných informací, jehož součástí by měl být datový slovník (popis všech typů dat, což zahrnuje i metadata). Osvědčeným postupem v komunitě digitální archivace je dohoda o užití takových formátů objektu CDO, které podporují dlouhodobé uchovávání, a užití mezinárodně rozšířených metadatových schémat.

1.1.4 Přístupy k uchovávání

Norma ISO 14721 popisuje dvě hlavní kategorie opatření pro uchovávání informačního obsahu vykonávaná v archivu OAIS nad balíčkem AIP jako digitální migraci a emulaci. Emulaci se blíže věnovat nebudeme (a norma se jí také detailněji nezabývá).

Digitální migrace (*digital migration*) je podle normy „přesun (*transfer*) digitálních informací v rámci archivu OAIS se záměrem tyto informace uchovat.“ (ISO 14721:2012, s. 22). Podle normy tento přesun od jiných typů přesunů odlišují následující tři atributy (ISO 14721:2012, s. 22):

- je zaměřen na uchování celého informačního obsahu,
- nová podoba informací v archivu nahrazuje podobu předchozí a
- archiv OAIS řídí všechny stránky přesunu a nese za ně plnou odpovědnost.

Norma ISO 14721 rozlišuje čtyři typy digitálních migrací. Renovace (*refreshment*) a replikace (*replication*) jsou migrace, při kterých nedochází ke změně datových objektů (tedy bitových posloupností), balíčkovací migrace (*repackaging*) a transformace (*transformation*) jsou migrace, při nichž ke změně datových objektů dochází. Definice těchto typů jsou následující (ISO 14721:2012, s. 103-104):

Renovace: Digitální migrace, při níž je jedna instance datového nosiče, která obsahuje balíček AIP, více balíčků AIP nebo části balíčků AIP, nahrazena jinou instancí datového nosiče stejného typu, a to zkopírováním bitů na datový nosič využitý k umístění balíčků AIP a ke správě a přístupu k datovému nosiči. Díky tomu dokáže stávající mapovací infrastruktura archivního úložiště beze změny stále nalézat balíček AIP a přistupovat k němu.

Replikace: Digitální migrace, při níž nedochází k žádným změnám balíčkovacích informací, informačního obsahu ani archivačních informací. Bity nesoucí tyto informační objekty jsou při přesunu na novou instanci stejného nebo nového typu datového nosiče zachovány. Renovace je také replikací. Replikace však může vyžadovat změny mapovací infrastruktury archivního úložiště.

Balíčkovací migrace: Digitální migrace, při níž dochází k změně bitů balíčkovacích informací.

Transformace: Digitální migrace, při níž dochází k změnám bitů informačního obsahu nebo archivačních informací, přičemž je současně vyvinuta snaha uchovat informační obsah v úplnosti.

Renovaci a replikaci se nebudeme dále věnovat. V praxi se souhrnně označuje jako **bitová ochrana** a znamená zkopírování dat na jiný datový nosič, rozdíl spočívá v tom, že replikace může vyžadovat úpravu mapovací infrastruktury (tato změna postihuje hardware a software, nikoliv data). Norma ISO 14721 tento rozdíl nepříliš srozumitelně vysvětluje, ale příklad z běžné praxe pomůže lepšímu porozumění. Zkopírování dat z jednoho nosiče CD na jiný je příkladem renovace. Příkladem replikace je zkopírování dat z deseti nosičů CD na pevný disk počítače.

Žádná z těchto dvou typů migrací nepředstavuje větší intelektuální problém, riziko spočívá především v nedostatku financí a nedostatečné kontroly stavu datových nosičů. Archiv musí mít nastaveny mechanismy pro monitorování stavu datových nosičů a plány pro migraci v případě, že je zaznamenáno riziko jejich degradace nebo zastarávání (tj. konec jejich hardwarové podpory). Rizikem samozřejmě může být i sama volba datových nosičů, které jsou zcela nevhodné svou povahou (názorným příkladem z minulosti je disketa, která nikdy nebyla bezpečným nosičem) nebo náročností údržby (např. z hlediska nedostatku finančních prostředků).

Ani balíčkovací migrace nepředstavuje větší riziko. Jedná se o změnu informací o zabalení. Příkladem může být jiné uspořádání v rámci adresáře. Riziko, které přináší

balíčkovací migrace, spočívá spíše v neudržení odlišení informačního obsahu a archivačních informací.

Transformace je nejdůležitějším a nejsložitějším typem digitální migrace. Pouze transformace podle normy zakládá novou verzi balíčku AIP. To znamená, že verze balíčku AIP, která prošla renovací, replikací nebo balíčkovací migrací, zůstává nezměněna.

Původně uložený balíček AIP (tj. informační balíček vytvořený z balíčku SIP dodaného vkladatelem) má být považován za první verzi balíčku AIP a norma jej označuje jako originál, resp. původní balíček AIP (*original AIP*). Tento původní balíček „může být udržován pro ověření uchovávání informací“ (ISO 14721:2012, s. 105). Norma takovýto postup (tj. uchovávat první verzi balíčku AIP i tehdy, když je vytvořena novější verze) tedy nepředepisuje, pouze uvádí jako možnost.

Transformaci norma dále dělí na dva dílčí typy. Vratná transformace (*reversible transformation*) je taková transformace, kdy je možná bezeztrátová zpětná transformace. Uváděným příkladem vratné transformace je migrace textového souboru (obsahujícího písmena anglické abecedy) v kódování ASCII do kódování UNICODE UTF-16 (ISO 14721:2012, s. 106).

Nevratná transformace (*non-reversible transformation*) je taková „transformace, u které nemůže být zaručeno, že se jedná o vratný převod“ (ISO 14721:2012, s. 24). Uváděný příklad může být nesrozumitelný, proto uveďme jednoduchý příklad z praxe – formátovou konverzi z formátu TIFF (nekomprimovaná verze) do formátu JPEG (formát JPEG nepodporuje jinou než matematicky ztrátovou kompresi). Nevratná transformace je podle normy rizikem pro zachování autenticity. V případě, že jde o transformaci archivačních informací, je možné učinit zobecnění, že v tomto případě by nikdy nemělo jít o nevratnou transformaci.

Norma systematicky nepopisuje rozdíly mezi transformací, která mění informační obsah, a transformací, která mění archivační informace. Její koncepty však lze specifikovat následujícím způsobem. V případě transformace informačního obsahu mohou nastat dvě základní varianty. Zaprvé může být potřeba změnit pouze interpretační informace o formátu. Například může dojít k tomu, že archivem doporučovaný nástroj pro zobrazení souboru ve formátu PDF zastará a nebude adekvátně reprodukovat formát PDF. Pak bude muset archiv provést průzkum a vybrat jiný nástroj, a informace o tomto nově doporučovaném formátu budou zdokumentovány a uchovávány spolu s objektem CDO. Tato transformace tedy zahrnuje pouze interpretační informace. Za druhé může nastat situace, kdy dochází ke změně bitů objektu

CDO, typicky v případě formátové konverze. Pokud půjde o konverzi z formátu TIFF (nekomprimovaná verze) do formátu JP2 (matematicky bezeztrátová komprese), půjde o vratnou transformaci, nicméně nová reprezentace (objekt CDO ve formátu JP2) bude vyžadovat jiné interpretační informace, neboť se jedná o jiný formát.

Jinou situací je proces obohacování archivačních metadat. V případě, že se mění objekt CDO, pak se vždy také mění archivační informace v tom smyslu, že se doplní o záznamy těchto změn (zejména jde o provenienční informace). Proto bude v průběhu uchovávání objem archivačních informací narůstat. Rovněž může dojít k tomu, že se v průběhu uchovávání obohatí archivační informace o nové informace (např. další perzistentní identifikátor).

1.1.5 Specifická standardizace informačních balíčků

Pro specifické typy digitálních dokumentů existují specifické metadatové standardy a doporučení pro výběr vhodných formátů. Datový objekt s obsahem může být tvořen jedním nebo více soubory, které mohou být uloženy v jednom nebo více formátech (souborových formátech). Formát je typem strukturálních interpretačních informací.

1.2 Formát objektu CDO

Americká studie o formátech souborů uvádí: „Většina souborů – s výjimkou souborů, které jsou jednoduchými datovými toky – obsahuje dvě základní komponenty: strukturální prvky a datové prvky. Formát souboru reprezentuje jedinečné a specifické uspořádání těchto strukturálních a datových prvků.“ (Lawrence, 2000 s. 2).

Formát je jedním z typů interpretačních informací modelu OAIS, konkrétně informací o způsobu, jakým je potřeba datový objekt uložený v konkrétním formátu interpretovat (softwarovou aplikací), a v případě digitalizátů knih také o způsobu, jak datový objekt v daném formátu reprodukovat, tj. adekvátně zobrazit.

V kontextu digitální archivace by v ideálním případě: a) měl vždy být k dispozici dostatek volně dostupných (specializovaných) softwarových aplikací, které znalostí daného formátu disponují (tj. aplikace, které mají dostatečné strukturální interpretační informace) a dokáží s ním pracovat; b) informace o formátu by měly být obsaženy v dostatečně dobře popsané dokumentaci (formátové specifikaci), která by měla být dostupná; c) formát by neměl být zatížen patenty. Počítačová realita má však k tomuto ideálu daleko a právě formáty a aplikace pro práci s nimi jsou předmětem častých změn, a tedy rizikem pro dlouhodobé uchovávání. Jeden z odhadů průměrné délky zastarání formátu (od doby uvedení) na trh

se pohybuje v rozmezí osmi až dvaceti let (Nielsen, Thirifays, 2011). Zastarávání se projevuje dvěma způsoby. Jednak narůstajícím rizikem ztráty dostupnosti softwarových aplikací, které dokáží formáty adekvátně reprodukovat (tj. zobrazit, přehrát nebo jiným způsobem prezentovat smyslům lidského uživatele). Zadruhé pak ztrátou schopnosti nové generace softwarových aplikací formát zpracovávat (upravovat data v daném formátu, provádět konverzi do jiného formátu apod.). V řadě případů také není dostupná formátová specifikace nebo se k formátům váží licenční omezení, což rovněž představuje velká rizika pro uchovávání informací. Připomeňme si druhou povinnost archivu OAIS (získat možnost s informacemi dostatečně nakládat), kterou lze zřejmě aplikovat i na tento případ, neboť nad informacemi uloženými ve formátu, jehož specifikace je nedostupná (uzavřená) nebo zatížená patenty, z principu nelze získat plnou kontrolu.

Pokud nedojde k včasné formátové konverzi v době, kdy ještě existují vhodné nástroje pro tuto transformaci, může dojít k nevratné ztrátě informačního obsahu (objekt CDO může existovat jako uložený objekt, ale nebude z něj možné získat informační obsah). Archiv by mohl teoreticky vytvořit nový nástroj na základě uložené formátové specifikace zastaralého formátu, prakticky je však takovou alternativu obtížné ověřit.[5]

1.2.1 Roviny a aspekty užití formátu

Z hlediska popisu formátu musíme odlišovat několik rovin, přičemž v případě rastrových formátů jde zejména o tyto roviny:

- Rodina formátů
- Konkrétní formát
- Verze formátu
- Komprese
- Profil

Příkladem rodiny formátů je RAW. Existuje celá řada konkrétních formátů RAW této rodiny, které vytvořili výrobci zařízení. Například fotoaparáty Canon užívají formát Canon

[5] Vzhledem k tomu, že je obtížné předvídat vlastnosti budoucího technologického prostředí, které mohou vytvoření takového nástroje znemožňovat, nebo vzhledem k tomu, že již nebudou dostupné další interpretační informace, které formátová specifikace předpokládala, ale nezaznamenala, neboť v době vytvoření formátové specifikace byly tyto další interpretační informace běžně dostupné.

RAW. Verze je dána historicky a odlišné verze formátu mohou znamenat odlišné požadavky na zobrazení. Různé formáty nabízejí odlišné možnosti komprese (např. TIFF verze 6 může být v nekomprimované podobě, zatímco JP2 je vždy komprimovaný). Profil znamená specifické nastavení v rámci formátu při jeho vytváření (např. volba ztrátové nebo bezeztrátové komprese), u některých formátů nastavení profilu vyžaduje specializovanou znalostí (což je zejména případ formátu JP2).

Z hlediska užití musíme u rastrových formátů odlišovat nejméně tyto aspekty: a) archivační formát; b) prezentační formát; c) prezentační meziformát.

Archivační formát je takový, který je aktuálně vhodný z hlediska potřeb dlouhodobého uchovávání. Někdy je nazýván jako archivní obrazová matrice (*archival master*). Koncept obrazové matrice (master) byl do digitální archivace převzat z komerčního sektoru. Obrazová matrice jsou v obrazovém průmyslu obvykle komprimovaná data, která slouží jako zdroj pro vytváření obrazových dat v různé kvalitě, v různých formátech a pro různé účely a nabízí nejvyšší možnou kvalitu dané produkce (např. fotografa). Archivační formát je volen mj. právě s ohledem na to, aby měl tuto funkci obrazové matrice, přičemž však jsou na jeho výběr kladeny další omezující znaky.

Prezentační formát je takový, který je aktuálně vhodný pro zpřístupňování z hlediska potřeb cílové komunity, v kontextu současné praxe formou prezentace v digitální knihovně. Prezentační formát můžeme rozdělit na hlavní prezentační formát (formát představující nejvyšší možnou kvalitu) a doplňkové prezentační formáty (např. malé náhledy obrázků).

Prezentační meziformát je meziformát, ze kterého digitální knihovna generuje cílový prezentační formát. Nejčastějším případem současné praxe je formát JP2 ve ztrátové kompresi, ze kterého se v digitální knihovně generuje formát JPEG jako výsledný hlavní prezentační formát. Jednou z výhod JP2 jako prezentačního meziformátu je řešení problému, že formát JP2 není nativně podporován v běžných internetových prohlížečích, a rovněž relativní jednoduchost takovéto formátové konverze do JP2.

1.2.2 Výběr archivačního formátu

Široce známým doporučeným postupem pro řízení životního cyklu digitálních dokumentů je, že prvním a zásadním krokem je výběr vhodného archivačního formátu, tj. formátu, ve kterém budou dokumenty uchovávány v archivu (a v případě, kdy dokumenty vytváří i uchovává jedna a tatáž organizace, vytváření finálních dat přímo v tomto formátu).

Archivační formát je volen z hlediska jeho (aktuální) vhodnosti pro uložení v balíčku AIP v digitálním archivu. Cílem je uložit obsah v takovém formátu, o kterém se předpokládá, že jeho užití v současnosti a blízké budoucnosti nebude představovat větší riziko. Pokud vkladatel do archivu dodá data v jiném formátu než archivačním, je doporučeným postupem, aby archiv provedl normalizaci do archivačního formátu (Cubr, 2010, s. 83-86).

Volba archivačního formátu není jednoduchá záležitost, nicméně dnes již existuje řada doporučení od uznávaných organizací, kterými se lze řídit. Formátový registr Kongresové knihovny již dlouho patří mezi hlavní zdroje pro stanovení kritérií výběrů formátů (Library of congress, 2013). Z hlediska seznamu doporučených konkrétních formátů byl ještě před několika lety de facto jediným citovaným zdrojem v odborné literatuře dokument Floridského digitálního archivu (Florida Digital Archive). Tento dokument obsahoval výčet nejběžnějších formátů s hodnocením jejich spolehlivosti na třístupňové škále (vysoká, střední a nízká spolehlivost) pro potřeby uchovávání v tomto archivu.[6] Od té doby vzniklo více takových doporučení, která lze považovat za směrodatná, neboť je vydaly významné organizace.

Za jeden z nejvýznamnějších zdrojů pro výběr archivačního formátu lze v současnosti označit dokument Deklarace doporučených formátů (Recommended Formats Statement), který od roku 2014 vydává Kongresová knihovna (Library of Congress). Tento dokument obsahuje konkrétní seznam doporučených formátů (pro digitální i fyzické objekty), který je určen jak pro vnitřní potřeby Kongresové knihovny, tak i pro jiné archivy. Zahrnuje seznamy preferovaných a akceptovaných formátů. Dokument je každý rok aktualizován, poslední seznam byl vydán v září 2019 (9).

Dalším významným doporučením je studie americké iniciativy FADGI, která byla vydána v roce 2014 a zaměřuje se specificky na rastrové formáty (tedy formáty relevantní pro digitalizaci knih) (FADGI, 2014). Studie obsahuje zhodnocení nejběžnějších formátů (TIFF, JP2, PDF, PNG, JPEG) z hlediska digitální archivace a uvádí sadu podrobných srovnávacích kritérií seskupených do čtyř hlavních kategorií (udržitelnost, ekonomické faktory, požadavky na implementaci, nastavení a možnosti).

Kritéria, která se v různých doporučeních objevují nejčastěji, lze zobecnit do těchto kategorií: podpora, otevřenost a nezatíženost patenty.[7] Podporou formátu se zde rozumí míra

[6] Dokument nesl název „Recommended Data Formats for Preservation Purposes in the Florida Digital Archive“ a v současnosti již není dostupný (Cubr, 2010, s. 84).

[7] Srv. např. Cubr, 2010, s. 83; Fernie, 2008, s. 11-14.

jeho užívanosti v dané komunitě a dostupnost nástrojů pro vytváření, zpracování a reprodukci. Nízká úroveň podpory znamená, že formát zastarává nebo že se vůbec neujal. Otevřeností formátu se rozumí skutečnost, že je dostupná dokumentace formátu (tj. formátová specifikace). Směrnice MINERVA uvádějí, že užití otevřených formátů „napomůže interoperabilitě a zajistí, že zdroje lze opětovně využívat, vytvářet a upravovat celou řadou aplikací. Rovněž napomůže vyhnout se závislosti na konkrétním dodavateli.“ (Fernie, 2008, s. 32)

Za dostupnost se v praxi zpravidla považuje to, že formátová specifikace je buď volně dostupná online, nebo že je dostupná v nějaké normalizační organizaci, která ji udržuje. Například specifikace formátu TIFF je dostupná na webu jeho vlastníka (firmy Adobe) a je průmyslovým standardem. Specifikace formátu JP2 je mezinárodní normou, která je dostupná k zakoupení v normalizační organizaci ISO[8].

Nezatíženost patenty nutně neznamená, že formát není nikým vlastněn, pouze to, že výkon práv duševního vlastnictví není uplatňován.

Dalšími uváděnými kritérii pro výběr formátu jsou například: míra zpětné kompatibility; možnosti exportu do jiných formátů; míra nezávislosti na specifických hardwarových a softwarových platformách; rozumná rovnováha mezi nabídkou funkcí formátu na straně jedné a přiměřenou komplexitou na straně druhé (Cubr, 2010, s. 85).

Obecně lze říci, že archivační formát by měl splňovat mj. také funkce obrazové matrice. Mimo oblast digitální archivace je častou volbou (zejména profesionálních fotografů) uložení obrazové matrice ve formátu DNG. Takovéto užití je v digitální archivaci problematické jednak kvůli svázanosti formátu DNG s aplikacemi firmy Adobe (která je původcem tohoto formátu), jednak také k nemalé ceně těchto aplikací, která může být pro řadu paměťových organizací zásadní provozní překážkou. To ukazuje, jak již bylo uvedeno výše, že požadavky na archivační formát jdou nad rámec požadavků na obrazovou matrici.

1.2.3 Prezentační formáty

Volba prezentačního formátu závisí na požadavcích cílové komunity. Jelikož v současné praxi je hlavní formou zpřístupnění digitalizátů jejich zobrazení v digitální knihovně, tyto požadavky se řídí především podporou formátu v internetových prohlížečích (13

[8] Citace specifikace: ISO/IEC 15444-1:2004. Information technology - JPEG 2000 image coding system: core coding system. 2nd ed. Geneva: ISO, 2004.

str. 46). Směrnice MINERVA doporučují zpřístupňovat digitální kopie dokumentů v různých velikostech nebo formátech, aby zacílení na uživatele bylo co nejširší (Fernie, 2008, s. 73). Směrnice pro budování kvalitních digitálních sbírek (Framework of Guidance for Building Good Digital Collections) vydané americkou normalizační organizací NISO doporučují pro zpřístupňování digitalizovaných dokumentů užití formátů JPEG a PDF.

1.2.4 Formátové registry

V souvislosti s přijetím modelu OAIS se v komunitě paměťových institucí objevil navazující koncept globálního registru interpretačních informací. Idea takového registru spočívá v tom, že bude sloužit jako zásobník interpretačních informací, které jsou potřebné pro digitální objekty (zejména informací o formátech, souvisejících aplikacích a všech ostatních prvcích počítačového prostředí, jež podporuje adekvátní reprodukci digitálních objektů a jejich zpracování). Za tímto návrhem stála pragmatická úvaha, podle níž nejsou jednotlivé instituce schopny všechny potřebné interpretační informace shromažďovat vlastními silami.

Jako první vznikl registr PRONOM[9]. Byl založen a je již patnáct let provozován britskými Národními archivy (The National Archives). Ačkoliv PRONOM zdaleka nesplňuje své původní ambice (obsahuje poměrně rozsáhlou databázi formátů, ale jejich popis je většinou minimální), jde v současnosti o nejvýznamnější důvěryhodný projekt, který se alespoň snaží uskutečňovat původní vizi směrodatného globálního registru interpretačních informací. Později sice vznikly dva další registry, ale již zanikly. Prvním z nich byl GDFR, který skončil již před několika lety.[10] Registr UDFR provozovaný Kalifornskou digitální knihovnou (California Digital Library) měl spojit registry GDFR a PRONOM, ale byl ukončen v dubnu 2016 kvůli nedostatku financí.[11]

Klíčovou funkcí registru PRONOM je to, že (jako jediný registr vůbec) nabízí jednoznačný a jedinečný identifikátor formátu, přesněji řečeno, jak uvádí sám registr: „rozšiřitelné schéma pro poskytování perzistentních, jedinečných a jednoznačných identifikátorů pro jednotky interpretačních informací zaznamenané v registru PRONOM.“ (Brown, 2006, s. 4). Formát je tedy pouze jedním z typů interpretačních informací, o nichž registr vede údaje, nicméně nejrozšířenějším. Funkce identifikátoru PUID jsou dvě: propojení se záznamem jednotky interpretačních informací v registru PRONOM (tj. způsob identifikace

[9] http://www.nationalarchives.gov.uk/PRONOM/Default.aspx

[10] http://library.harvard.edu/preservation/digital-preservation_gdfr.html

[11] http://www.udfr.org/

záznamu, přičemž tento záznam by ideálně měl obsahovat co nepodrobnější informace o formátu nebo jiné jednotce interpretačních informací) a jedinečný perzistentní identifikátor, který odlišuje v maximální možné míře jeden formát od druhého (odlišuje se nejen typ formátu, ale často i verze[12]). Například PUID pro JPEG verze 1.00 je „fmt/42“, pro verzi 1.01 „fmt/43“ a pro verzi 1.02 „fmt/44“. Registr MIME,[13] který je nejužívanějším obecným registrem formátů (sloužícím i pro účely mimo kontext digitální archivace), odlišuje formáty jen na základě typu a názvu, například formát JPEG všech verzí má označení „image/jpeg“.

Za druhý významný zdroj informací o formátech lze považovat registr Kongresové knihovny[14]. Ten obsahuje nejen základní interpretační informace o formátech, ale také o rizicích s nimi spojených. (Library of congress, 2013)

Formátový registr PRONOM nabízí záznamy o formátech a dalších jednotkách interpretačních informací v podobě volně dostupných webových stránek. Současně poskytuje identifikační mechanismus, který obsahuje popis toho, kde se v daném formátu nachází tzv. „magické číslo“ (údaj o verzi formátu), a identifikátory PUID. „Magické číslo“ je interní mechanismus označení konkrétního formátu daný formátovou specifikací, zatímco identifikátor PUID je jedinečný externí identifikátor, který je přidělovaný registrem PRONOM. Registr PRONOM také nabízí svůj vlastní nástroj, DROID, který provádí formátovou identifikaci užitím uvedeného mechanismus na jednotlivé soubory.

1.3 Metadatové standardy

Norma ISO 14721 definuje metadata jako „data o jiných datech“ (ISO 14721:2012, s. 24). Metadata, tak jako datový objekt s obsahem, jsou předmětem interpretace softwarovými nástroji a dalšími prvky počítačového prostředí a v konečném bodě i lidskými uživateli. Pro normu jsou klíčové přidružené informace, které jsou v životním cyklu digitálního dokumentu zaznamenávány a udržovány spolu s datovým objektem s obsahem a které byly představeny v popisu normy výše. Současné metadatové standardy užívané pro specifikaci normy ISO 14721 v komunitě paměťových institucí jsou určeny pro záznam těchto specifických typů informací. Ačkoliv se pojem metadata může vztahovat na širokou škálu typů (např. i na

[12] Například u formátu Epub se jeho jednotlivé verze v registru PRONOM nerozlišují.

[13] http://www.iana.org/assignments/media-types/media-types.xhtml

[14] https://www.loc.gov/preservation/digital/formats/index.shtml

metadata uložená v obrazových datech nebo v databázi), standardy užívané v paměťových institucí jsou specifickým typem metadat, které se vyznačující následujícími charakteristikami:

- jsou vysoce strukturovaná;
- ukládající se do samostatných textových souborů ve formátu XML;
- uchovávají se spolu s datovým objektem s obsahem v informačním balíčku;
- jsou pečlivě zdokumentovaná mezinárodními standardy;
- užívá je většina organizací dané oblasti digitálního dědictví (např. knihovnictví).

Dodejme, že tento typ metadat má svůj původ v katalogizačních standardech, původně vytvářených pro popis tištěných knih a dalších typů fyzických dokumentů za účelem možnosti jejich strojového zpracování. Od té doby trvá linie vysoké standardizace a širokého osvojení ve většině organizací dané komunity.

Klíčovým prvkem těchto standardů je sada elementů (*element set*). Jak uvádí Zeng, metadatové standardy založené na sadě elementů obecně slouží k popisu zdrojů specifického typu nebo pro konkrétní účel, definují význam elementů a jejich vztahy a poskytují návod, jaké hodnoty a jakým způsobem by měly být připsány elementům při popisu konkrétního zdroje nebo účelu. Každý element je definován určitým počtem atributů (základním je název) a dalších nezbytných informací (definice, identifikátor[15] atd.). Z tohoto důvodu je klíčové, aby všechny sady elementů byly definovány v metadatovém formátu, který je konzistentní, srozumitelný a komunikovatelný mezi různými komunitami (Zeng, 2016, s. 38). Užití serializace digitálních metadat do formátu XML je typické pro paměťové instituce obecně (Zeng, 2016, s. 131).

Pro ilustraci aktuálně vnímané důležitosti úlohy metadat (výše uvedeného typu) v současné praxi poslouží následující citace, která pocházející přímo z jednoho z takových standardů: „Bez strukturálních metadat jsou obrázek stránky nebo textový soubor tvořící digitální dílo téměř k ničemu, a bez technických metadat zohledňujících digitalizační proces si badatelé nemohou být jisti, jak přesný odraz originálu digitální verze poskytuje. Pro účely vnitřní správy musí mít knihovna přístup k náležitým technickým metadatům, aby mohla pravidelně obnovovat a migrovat data, a tak zajistit zachování cenných zdrojů.“ (METS, 2016).

[15] Například PREMIS má u každého elementu uveden jedinečný identifikátor založený na hierarchické řazení, např. 1.2, 3.1.

1.3.1 Přehled metadatových standardů pro digitalizáty tištěných dokumentů

Již před érou digitálních metadat byla knihovnická komunita známá vysokou mírou jednotné standardizace na mezinárodní úrovni. Tak tomu je i v případě metadatových standardů pro digitální data. Knihovny pro správu digitalizátů se nejčastěji užívají tyto mezinárodní metadatové standardy: METS[16], PREMIS[17], MODS[18], MIX[19] a ALTO[20] .

Tyto standardy lze z hlediska rozsahu jejich možné aplikace rozdělit do dvou skupin: a) omezené užití (MIX; ALTO), b) univerzální užití (METS, PREMIS, MODS). První skupina je určena specificky pro digitalizáty knih (a některé další typy digitálních dokumentů), druhou skupinu lze aplikovat na většinu typů digitálních dokumentů. Všechny tyto standardy, s výjimkou PREMIS, jsou definovány specificky jako XML schémata. Standard PREMIS není svázán se schématem XML, ale lze jej jako XML vyjádřit (existuje oficiální XML schéma k tomuto standardu) a jako XML schéma se také užívá pro balíčky SIP a AIP.

Standard ALTO je primárně datový formát (zaznamenávající text získaný procesem OCR a jeho souřadnicové umístění vzhledem k obrazu), ale obsahuje některé metadatové prvky (např. informace o obrazovém zdroji pro OCR). Standard MODS je užíván pro zápis bibliografických informací, a je tedy určen k naplňování role identifikačních informací OAIS (a případně popisných informací OAIS). Zbylé tři standardy, které jsou nejdůležitější, popíšeme detailněji v následujících oddílech.

1.3.2 PREMIS

Standard PREMIS sám sebe označuje jako standard pro archivační metadata (*preservation metadata*), který „podporuje životaschopnost, reprodukovatelnost, srozumitelnost, autenticitu a identitu digitálních objektů v archivačním kontextu“ (PREMIS, 2015, s. 1). Slouží však nejen pro zápis archivačních informací, ale také interpretačních informací. Aktuální třetí verze PREMIS vyšla v roce 2015. Standard neobsahuje pouze sadu elementů, ale také vlastní komplexní datový model, terminologický slovník a podrobný text vysvětlující logiku a možnosti užití standardu v archivu. Ve své sebedefinici klade standard PREMIS také důraz na to, aby jeho elementy byly implementovatelné, což podle něj znamená,

[16] http://www.loc.gov/standards/mets

[17] http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf

[18] http://www.loc.gov/standards/mods/mods-overview.html

[19] http://www.loc.gov/standards/mix/

[20] https://www.loc.gov/standards/alto/description.html

že hodnoty většiny elementů musí být možné automatizovaně vyplňovat a zpracovávat archivem (Premis, 2015, s. 3). K takovémuto cíli ostatně směřují všechny zde uvedené metadatové standardy.[21]

Datový model v PREMIS definuje čtyři základní entity: objekt (*object*), činitel (*agent*), událost (*event*) a právní deklarace (*rights statement*). Objekt dále člení na čtyři úrovně: intelektuální entita (*intellectual entity*) je „jednotlivý intelektuální nebo umělecký výtvor (*creation*), který je považován za relevantní pro cílovou komunitu v kontextu digitální archivace"; reprezentace (*representation*) „množina souborů (včetně strukturálních metadat) potřebná pro úplnou reprodukci intelektuální entity"; soubor (*file*) „pojmenovaná a uspořádaná posloupnost bajtů, kterou dokáže rozeznat operační systém" a která je v určitém formátu; bitový tok (*bitestream*) „data v rámci jednoho souboru, která mají smysluplné společné vlastnosti pro archivační účely" (Premis, 2015, s. 8). Všechny úrovně (vyjma intelektuální entity) odpovídají pojmu „digitální objekt" v modelu OAIS, přičemž reprezentace v PREMIS odpovídá pojmu „objekt CDO". Intelektuální entita odpovídá informačnímu obsahu modelu OAIS s tím rozdílem, že v standardu PREMIS jde specificky o reprodukovaný informační obsah (tj. který může vnímat člověk).

Intelektuální entitu je možné podle modelu PREMIS také dále specifikovat podle úrovní abstrakce popsaných ve známém knihovnickém modelu FRBR. Model FRBR stanovuje tyto čtyři úrovně: (*work*), vyjádření (*expression*), manifestace (*manifestation*) a exemplář (*unit*)[22].

PREMIS obsahuje elementy, které odpovídají všem typům archivačních informací. Klíčové jsou zejména elementy pro zápis provenienčních informací. V tomto ohledu PREMIS vhodně předepisuje logiku metadatového zápisu: „metadata, soubory, bitové toky a reprezentace uchovávané v archivu se popisují jako statické množiny bitů. Není možné změnit soubor (nebo bitový tok nebo reprezentaci); lze pouze vytvořit nový soubor (nebo bitový tok nebo reprezentaci), který se vztahuje k zdrojovému objektu." (Premis, 2015, s. 22). Tento vztah mezi novým a předchozím objektem definuje jako vztah odvození (*derivation relationship*), u něhož musí být zaznamenán specifický typ události, odlišný od událostí, které nevytvářejí nový objekt. Standard odlišuje dva typy odvození ze zdrojového digitálního objektu do nového objektu: replikace (*replication*) a transformace (*transformation*) (Premis, 2015, s. 19). Replikace znamená vytvoření digitální kopie, která je bitově identická se zdrojovým digitálním

[21] Týká se to i metadat ve schématu MODS, které obvykle vznikají konverzí bibliografických záznamů ve formátu MARC.

[22] https://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf

objektem (Premis, 2015, s. 272), transformace má za výsledek vytvoření jednoho nebo více digitálních objektů, které nejsou bitově identické se zdrojovým objektem (Premis, 2015, s. 273).

Pro strukturální interpretační informace slouží sekce elementů popisujících formát (název formátu; verze formátu; název formátového registru; identifikátor záznamu formátu v tomto registru; role registru). Pro podrobnější popis interpretačních informací je ve standardu PREMIS vyčleněna možnost vnořit externí schéma.[23] Pro digitalizáty knih je za tímto účelem užíván standard MIX. PREMIS obsahuje i sekci signifikantních vlastností, která však není v praxi příliš užívána.

1.3.3 METS

Standard METS slouží primárně jako metadata zaznamenávající informace o zabalení modelu OAIS (tedy o zabalení balíčků SIP, AIP a DIP). Především umožňuje vnoření dalších metadatových schémat pro popis archivačních a interpretačních informací (a tím jejich identifikaci). Dále obsahuje sekci určenou pro zápis provenienčních informací (formou vnoření externího schématu) a zápis některých interpretačních informací (např. o chování objektu). V praxi je METS užíván zejména pro první funkci (záznam informací o zabalení) a také jako datový formát. Jeho sekce strukturální mapa se využívá pro záznam informací o všech obrazových souborech a jejich posloupnosti. Tyto informace tedy netvoří strukturální interpretační informace, ale vlastní datovou součást digitalizátu knihy, bez níž by objekt CDO byl neúplný.

METS je možno užít v kombinaci se standardem PREMIS, přičemž lze zvolit několik způsobů implementace. Americké směrnice NISO doporučují zaznamenat PREMIS do sekce METS pro zápis provenienčních informací (NISO, 2007, s.55).

1.3.4 MIX

Název standardu MIX je zkratkou pro „Metadata for Images in XML Standard“. Tento standard je XML schéma, které je založeno na americké normě ANSI/NISO Z39.87-2006. Podle vlastního popisu je účelem normy „standardizovaná sada metadatových elementů pro rastrová obrazová data“, přičemž tyto elementy „dokumentují digitální obrazová data vytvořená digitální fotografií nebo skenováním a též data, která byla pozměněna editováním nebo obrazovým převodem“ (ANSI/NISO, 2006, s. 1) Standard MIX obsahuje elementy této normy,

[23] V rámci elementu „objectCharacteristicsExtension“.

přidává několik dalších (např. rozděluje prostorové rozlišení do dvou elementů) a snižuje povinnost vyplnění elementů. Podle své vlastní definice MIX vznikl jako formát pro výměnu nebo uložení dat specifikovaných v uvedené normě NISO. Standard MIX se v praxi užívá pro záznam obrazových vlastností digitalizátů (tedy dalších typů interpretačních informací), a to jako externí schéma pro PREMIS.

Norma ANSI/NISO Z39.87-2006 uvádí, že není určena pro záznam provenience (ANSI/NISO, 2006, s. 1).. Kupodivu to není tak docela pravda vzhledem k tomu, že jedna její sekce elementů („Change History“) je určena pro záznam provenienčních informací z doby produkce (pro záznam generací dat vzniklých při vytváření finálních produkčních dat i užitých procesů), ale v praxi se za tímto účelem užívají spíše elementy standardu PREMIS, ačkoliv je PREMIS primárně určen pro záznamy operací v archivu, nikoliv pro digitalizaci.

1.3.5 Metadata v obrazových souborech

Speciální oblastí je zabudování metadat do obrazových souborů. Široce rozšířeným a obrazovým průmyslem podporovaným standardem pro zabudovaná metadata je EXIF[24]Dva hlavní snímkové formáty (TIFF a RAW) podporují záznam metadat ve formátu EXIF. Tato metadata do souborů zapíše snímací zařízení (ať již skener, nebo fotoaparát) automaticky. Obsahují velké množství elementů popisujících mj. snímací zařízení (včetně sériového čísla), způsob snímání nebo obrazové nastavení. Formát JP2 záznam EXIF metadat neumožňuje. Umožňuje však vnořit libovolná metadata v XML (do strukturálního prvku „XML Box“, jehož volba je součástí profilu tohoto formátu). V praxi se do něj zapisují například bibliografické údaje (Library of Congress, 2006).

2 Specifická část

2.1 Východiska standardů NDK

2.1.1 Digitalizační projekt

Digitalizací se zde rozumí specificky digitalizace tištěných dokumentů do podoby rastrových dat. Tento typ digitalizace směrnice FADGI definuje jako „konverze analogových barevných a jasových hodnot do nespojitých číselných hodnot. Číslo nebo množina čísel označuje barvu a

[24] http://www.cipa.jp/std/documents/e/DC-008-2012_E.pdf

jas každého pixelu v rastrovém obrázku.“ (FADGI, 2010, s. 44). Výstupem této digitalizace a předmětem následného uchovávání a zpřístupňování je digitalizát tištěného dokumentu, který je tvořen množinou souborů rastrových dat, jež reprezentují vizuální vlastnosti částí tištěné knihy (stránek, přebalu atd.) a jejich posloupnost, a který obvykle také zahrnuje textová data, jež jsou výstupem optického rozpoznávání znaků (OCR) v rastrových obrazech a která umožňují do určité míry pracovat s obrazy rovněž jako s textem (viz funkce plnotextového prohledávání digitalizátu).

Digitalizační projekt může mít různé cíle. Pokud je však cílem výsledné digitalizáty dlouhodobě uchovávat i zpřístupňovat čtenářům, je optimálním řešením takové, aby byla digitalizace nastavena z hlediska potřeb digitální archivace. Pro digitální archivaci je klíčové, aby zahájení digitalizačního projektu předcházela prvotní specifikace balíčku SIP a aby byly stanoveny vhodné transparentní postupy (včetně specifických nástrojů) pro vytváření dat a metadat, které budou určitou zárukou toho, že vytváření balíčku SIP proběhlo tak, aby nebyla narušena autenticita informačního obsahu v balíčku SIP.

Specifikaci balíčku SIP pro kontext této metodiky poskytují standardy NDK, určení vhodných postupů je pak vlastním cílem této metodiky. Digitalizační projekt by měl být sepsán po obeznámení se se standardy NDK a metodikou, a teprve na základě těchto požadavků by měla být navržena jeho realizace.

Digitalizační projekt, po stanovení podoby balíčku SIP a postupů, musí zahrnovat řadu obecných specifikací technické a organizační povahy. Mezi ně patří: výběr vhodného snímacího zařízení, vytvoření digitalizačního pracoviště (včetně stanovení rolí a jejich obsazení lidskými zdroji), výběr testování softwarových nástrojů pro vytvoření digitalizátů v požadované podobě (včetně metadat), volba / vývoj digitalizačního systému pro řízení celého průběhu digitalizace, testování zařízení a softwaru a kontrola kvality (včetně postupů řešení chyb). Tyto otázky jsou již dobře popsány v existující odborné literatuře. Jako základní zdroj pro tuto problematiku lze doporučit směrnice americké iniciativy FADGI (Federal Agencies Digital Guidelines Initiative), které byly vytvořeny pro paměťové instituce v USA a které nesou název „Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Files“. Směrnice vyšla v roce 2010[25], v roce 2016 pak byla vydána jejich revidovaná verze[26]. Jako

[25] http://www.digitizationguidelines.gov/guidelines/FADGI_Still_Image-Tech_Guidelines_2010-08-24.pdf

[26] http://www.digitizationguidelines.gov/guidelines/FADGI_Still_Image_Tech_Guidelines_2016.pdf

úvodní zdroj lze současně využít také Metodiku pro vytváření bezpečnostních kopií archiválií v digitální podobě vytvořenou Národním archivem (důležité jsou především kapitoly 4.1-4.6).

2.1.2 Formáty pro objekt CDO

Digitalizát tištěného dokumentu lze pojímat jako objekt CDO tvořený dvěma komponentami: obrazová komponenta (archivní kopie) a OCR komponenta. Třetí komponentou je strukturální komponenta (strukturální mapa v METS), která je však v praxi často označována za strukturální metadata, takže se v této metodice budeme držet toho zavedeného úzu.[27]

2.1.2.1 Archivační formáty pro obrazovou komponentu

Směrodatné zahraniční zdroje uvádějí formáty TIFF a JP2 jako dva hlavní archivační formáty pro obrazová (rastrová) data. Tyto formáty splňují v dostatečné míře uvedené tři základní podmínky kladené na archivační formát (podpora, otevřenost, nezatíženost patenty). Je nepochybně výhodou, že v oblasti rastrových formátů existuje možnost takové volby vzhledem k tomu, že existují jiné typy dat, pro které otevřené formáty zatím nejsou k dispozici nebo se neužívají.

Směrnice FADGI doporučují jako archivační formát TIFF, verze 6, profil bez komprese (FADGI, 2010, s. 79). Doporučení Floridského digitálního archivu označovalo za rastrové formáty s nejvyšší spolehlivostí nekomprimovaný TIFF, JP2 v bezeztrátové kompresi a PNG.[28] V Deklaraci doporučených formátů Kongresové knihovny pro 2019-2020 se pro digitální obrazová data (digitální fotografie a další typy digitálních obrazových dat) uvádějí jako preferované archivační formáty: 1. TIFF, 2. JP2, 3. PNG (Library of Congress, 2019). Směrnice pro budování kvalitních digitálních sbírek (Framework of Guidance for Building Good Digital Collections) vydané americkou normalizační organizací NISO doporučují jako archivační formáty nekomprimovaný TIFF a bezeztrátově komprimovaný JPEG 2000 (NISO Framework working group, 2007, s. 28). Podle rozsáhlého průzkumu provedeného v letech 2012-2013 považují severoamerické knihovny formát TIFF za nejspolehlivější archivační formát vůbec (bez ohledu na typ dokumentu) (Rimkus, 2014).

[27] Přesnější by bylo označení strukturální data, protože informace zaznamenané ve strukturální mapě METS popisují vlastní objekt CDO, nikoliv metadata.

[28] Citováno dle Cubr, 2010, s. 84.

V současně praxi je nejužívanějším archivačním formátem TIFF.[29] Nekomprimovaný TIFF jako archivační formát využívají pro své digitalizační projekty například Kongresová knihovna (FADGI, 2014, s. 5), pro některé projekty i Národní knihovna Švédska[30]. Do roku 2014 jej jako archivační formát využívala i Národní knihovna Francie, od roku 2014 je jejich archivačním formátem JPEG2000 (Duploy, 2017). Převaha formátu TIFF jako archivačního formátu digitalizačních projektů souvisí nejen s důvěrou, které se těšil a stále těší, ale také s jeho dlouhou historií (specifikace aktuální, tj. šesté verze byla vydána v roce 1992) a širokou podporou v technologickém prostředí od 90. let 20. století do současnosti.

Přibližně posledních deset let se začíná v novějších digitalizačních projektech (zejména evropské provenience) rozšiřovat užití formátu JP2 jakožto archivačního formátu (Van der Knijff, 2011; FADGI, 2014). Jistý vliv na to může mít i příznivá studie italských odborníků, která byla v roce 2008 vydána v periodiku D-Lib Magazine (Buonora, Liberati, 2008). Tato studie srovnávala tři rastrové formáty (TIFF, JP2 a JPEG) z hlediska jejich vhodnosti pro archivaci. Za nejvhodnější archivační formát označila právě JP2, mj. díky jeho nejlepší robustnosti (odolnosti vůči menšímu poškození bitů). Formát JP2, který byl původně v knihovnách užíván spíše jako formát pro zpřístupnění, se stal archivačním formátem například pro digitální obrazové fondy knihovny Wellcome Library (Buckley, 2009) nebo pro masovou digitalizaci norské národní knihovny (Brygfjeld, 2010). Prvně jmenovaná knihovna užívá profil formátu JP2 ve ztrátové kompresi, druhá v bezeztrátové. Užití ztrátové komprese není výjimečné. V praxi jsou voleny různé profily formátu JP2, zahrnující nejen volbu typu komprese, ale některé další parametry specifické pro tento formát (JPEG 2000 profiles, 2010). Vhodné nastavení těchto parametrů vyžaduje specialistu; z toho důvodu si některé knihovny nechaly vypracovat profily na zakázku u externích odborníků.[31] V posledních letech bylo uspořádáno několik seminářů věnujících se této problematice. K formátu JP2 se však objevují i kritické názory (Van der Knijff, 2011) a diskuze o tomto formátu jako archivačním formátu pokračuje (FADGI, 2014, s. 3). Obecně lze říci, že v USA převažuje volba formátu TIFF a v Evropě se rozšiřuje užití formátu JP2. Užití formátu PNG jako archivačního formátu je v současné praxi řídké, což může být vzhledem k jeho kvalitám překvapivé.

[29] Srv. např. FADGI, 2014, s. 3-4; Rimkus, 2014; Van der Knijff, 2011.

[30] Digitalizované noviny jsou archivovány ve formátu JPEG2000, ostatní dokumenty většinou ve formátu TIFF. (Neiss, 2017)

[31] Například profil formátu JP2 pro Wellcome Library vytvořil Robert Buckley, jeden z autorů specifikace formátu JP2 (viz. Buckley, 2009). Buckley vytvořil specifikaci formátového profilu JP2 i pro řadu dalších knihoven.

2.1.2.2 Prezentační formáty pro obrazovou komponentu

V praxi je nejužívanějším hlavním prezentačním formátem digitálních knihoven JPEG (FADGI, 2014). Tato volba je logická vzhledem k jeho vysoké podpoře v internetových prohlížečích nebo zobrazovacích aplikacích. Formát JPEG jako prezentační formát užívá například Gallica, jedna z největších digitálních knihoven světa provozovaná francouzskou národní knihovnou (Bruys et al., 2019). Formát JP2 není podporován běžnými internetovými prohlížeči, pro jeho užití jako prezentačního formátu je nutno nainstalovat plugin. Běžným způsobem zpřístupnění, který řeší tento problém z hlediska komfortu cílové komunity, je využití formátu JP2 jako prezentačního meziformátu, ze kterého digitální knihovna generuje formát JPEG, v němž jsou obrazová data prezentovaná čtenářům internetovým prohlížečem (Buckley, 2009, s. 11). Jako doplňkové vedlejší formáty se užívají například formát PDF (pro možnost stažení digitalizátu knihy čtenářem v podobě jednoho souboru) nebo formát GIF (pro náhledy obrázků).

2.1.2.3 Archivační formát pro OCR komponentu

Optické rozpoznávání znaků (OCR) je metoda získávání textu z obrazu. V současné praxi se užívá k tomu, aby se z rastrových dat vzniklých digitalizací vytěžil textový obsah. Výstup z OCR se ukládá v podobě strukturovaného textového formátu, který obsahuje informace o pozici (obrazem vyjádřených) konkrétních písmen (slov) v obrazovém souboru, z něhož byl vytvořen, aby bylo zajištěno namapování textu na obraz. Tímto formátem je v současné praxi převážně ALTO. Formát ALTO sám sebe popisuje jako „standardizovaný formát XML k ukládání informací o rozložení (layout) a obsahu“ (ALTO Principles, 2016). Formát ALTO XML je navržen jako externí schéma pro standard METS. Jde však především o datový formát (obsahuje vlastní text předlohy a jeho strukturu), částečně slouží i jako metadatový formát (např. popis informací o zdrojovém obrázku). V případě, že se vytvářejí archivační formát i prezentační varianty, je nutné, aby měly stejnou pixelovou velikost (počet pixelů na šířku a výšku obrázku, jinak text nebude správně namapován).

Míra efektivity OCR závisí na několika faktorech: stav předlohy, kvalita softwaru a jeho slovníků a komprese. Podle některých výzkumů[32] přinášejí ztrátově komprimovaná obrazová data mírně lepší výsledky procesu OCR v porovnání s bezeztrátovou kompresí (Chapman 2007,

[32] Národní archivy Velké Británie v roce 2017 naopak uvádějí že ztrátová komprese optické rozpoznávání textů ovlivňuje negativně. (The National Archives, 2017, s. 3)

s. 39; Buckley, 2006, s. 6). Záleží však také na tom, v jakém formátu je ztrátová komprese – různé nástroje pro OCR mohou s různými formáty pracovat odlišně (např. Tessaract nepodporoval formát JP2). V praxi platí, že pro novodobé fondy v dobrém stavu předloh (a pro běžné jazyky) je efektivita OCR velmi vysoká.

Prezentace digitalizátu knihy v digitální knihovně zahrnuje nejen zobrazení, ale i reprodukci těchto strukturovaných textových informací získaných z OCR, která umožňuje čtenářům plnotextové prohledávání obrazových dat. Přidáním této strukturované textové složky je digitalizát knihy obohacen o funkci, kterou jeho tištěná předloha nikdy neměla.

2.1.3 Metadatový aplikační profil

V digitalizační praxi si paměťové instituce pro konkrétní projekt vytvářejí tzv. metadatový aplikační profil. Koncept tohoto profilu je založen na ideji, že pro konkrétní kontext je nutno metadatové standardy lokalizovat a optimalizovat (Zeng, 2016, s. 54). Metadatový aplikační profil je soubor metadatových elementů, které jsou vybrány z jednoho nebo více mezinárodních metadatových standardů a jsou spojeny do jednoho sloučeného schématu, který je uzpůsoben na míru funkčním požadavkům konkrétního užití, zatímco je zachována interoperabilita s původními standardy (Duval, 2002). Součástí profilu může být i vypracování vlastních metadatových elementů, v praxi digitalizátů tištěných dokumentů to však není obvyklé. Na webové stránce Kongresové knihovny je možno nalézt ukázky metadatových profilů různých významných paměťových institucí světa (včetně NK ČR).[33] Metadatový profil se může stát národním standardem. Tak je tomu v případě standardů NDK pro metadata.

Metadatový aplikační profil pro digitalizaci slouží k tomu, aby byly zachyceny zejména provenienční informace o původu a historii změn digitalizátu v průběhu jeho vytváření, počínaje snímáním a konče finalizací balíčku SIP. Jde zejména o operace, které byly vykonány (události standardu PREMIS) a informace o všech generacích obrazových dat. Pokud nejsou tyto informace zachyceny v průběhu digitalizace, jejich pozdější zjišťování archivem může být obtížné nebo přímo nemožné. Rovněž je důležité, aby balíček SIP obsahoval kvalitní identifikační informace, v praxi jde především o bibliografická metadata a perzistentní identifikátory. Garantem kvality těchto metadat by měla být vždy digitalizující knihovna – předlohy jsou popsány v jejím katalogizačním systému. Tyto záznamy jsou předmětem konverze do formátu MODS.

[33] http://www.loc.gov/standards/mets/mets-registered-profiles.html

2.1.4 Specifikace obrazových dat

V digitalizační praxi je třeba u obrazových dat rozlišovat několik dalších aspektů / vlastností, které jsou nad rámec formátu: generace dat; strukturální model obrazové reprezentace; formátový profil; typy komprese; obecné obrazové vlastnosti; prezentační varianty obrazových dat.

2.1.4.1 Generace obrazových dat

Prvotní výstup snímání skenerem nebo fotoaparátem budeme nazývat původní snímky. Původní snímky jsou soubory v rastrových formátech uložené po snímání na datový nosič pracovní stanice digitalizační linky. Tyto původní snímky zpravidla procházejí dalším zpracováním do doby, než je vytvořen konečný obrazový výstup digitalizace (finální produkční data). Každá transformace, kterou prošla obrazová data od původních snímků po vytvoření finálních produkčních dat, znamená vytvoření nové generace obrazových dat. K vytváření dalších generací dochází také v pozdějších etapách (archivace a zpřístupňování), nicméně cíle vytváření těchto generací jsou odlišné. Cílem digitalizačních transformací je vytvořit konečný digitální produkt a teprve tento konečný produkt lze považovat za plnohodnotnou obrazovou složku digitalizátu knihy. Cílem archivace a zpřístupnění je zachovat tento digitalizační produkt v požadované kvalitě, v případě archivace z hlediska uchovávání informačního obsahu navzdory rizikům technologického zastarávání, v případě zpřístupnění vytvoření takové podoby digitalizátu dat, která je vhodná pro aktuální potřeby cílové komunity a která se může lišit od dat uložených v balíčku AIP. Původní snímky v současné praxi mohou být v odlišném formátu než finální produkční data. Generace se od sebe liší změnami, které lze zaznamenat na bitové úrovni, přičemž však může platit, že některé transformace je možno vykonat společně před tím, než bude uložena nová generace obrazových dat.

2.1.4.2 Strukturální model obrazové reprezentace

Modelem obrazové reprezentace se zde rozumí vztah mezi strukturou tištěné knihy (posloupností stran) a způsobem digitální reprezentace této předlohy v digitalizátu knihy. Převažující strukturální model současné praxe lze charakterizovat následovně:

- Základní předmět reformátování = jedna stránka knihy.
- Základní objekt uložení = jeden soubor reprezentující tuto jednu stránku.

Tento model znamená, že jeden soubor v rastrovém formátu reprezentuje jednu stránku knižního bloku, přičemž všechny rastrové soubory mají stejnou velikost. Pro reprezentaci jiných částí knih, než jsou stránky (např. přebal), nebo nestandardních částí (např. stránka s mapou, kterou lze rozložit, takže její velikost bude jiná než ostatních stran), se v praxi užívají různé postupy. Běžnou současnou praxí je snaha zachytit všechny části knihy, včetně vakátů a prázdných přídeští (jedinou výjimkou v tomto směru je zadní část přebalu, která bývá bílá a nepovažuje se za smysluplné ji digitalizovat).

2.1.4.3 Formátový profil

Formátovým profilem se rozumí nastavení v rámci konkrétního formátu. Mezi hlavní prvky profilu rastrových formátů patří komprese (tj. její nastavení v rámci možností formátu, přičemž ne všechny rastrové formáty umožňují všechny typy kompresí). Nejčastějším profilem pro formát TIFF je nekomprimovaná varianta. Formátový profil pro JP2 zahrnuje (nutnou) volbu mezi ztrátovou a bezeztrátovou kompresí a dále několik dalších specifických nastavení typických pro tento formát.

2.1.4.4 Typy komprese

Klíčovým aspektem rastrových obrazových dat je komprese. Pro potřeby této práce postačí následující klasifikace čtyř základních typů:

1. Nekomprimovaná varianta (tj. komprese není užita)
2. Matematicky bezeztrátová komprese
3. Vizuálně bezeztrátová komprese
4. Vizuálně ztrátová komprese

Nekomprimovaná varianta může být teoreticky nejlepší možnou volbou pro archivační formát. Studie iniciativy FADGI uvádí: „Nekomprimovaná datová struktura má jednu velkou výhodu: je relativně transparentní. Transparentnost souvisí s ukazatelem udržitelnosti: nemělo by být složité vytvořit nástroj, který dokáže přečíst informaci o obalu (wrapper) a rozbalit rastrová data tak, aby je bylo možno zobrazit“ (FADGI, 2014, s. 4). Jediným problémem s užitím nekomprimovaných dat je jejich velikost. V současné praxi je však tento problém někdy velkou překážkou vzhledem k omezenému rozpočtu na úložné kapacity.

Matematicky bezeztrátová komprese obrazových dat v principu odpovídá konceptu vratné transformace modelu OAIS. Kompresní algoritmus snižuje transparentnost formátu, ale současně umožňuje snížení požadavků na úložné kapacity.

Vizuálně bezeztrátovou kompresí se myslí taková matematicky ztrátová komprese, která na základě nějakého přijatého psychofyziologického modelu stanoví kompresní poměr, jehož výstupem (při zobrazení) má být informační obsah, který by lidský vnímatel neměl rozeznat od výstupu matematicky bezeztrátové komprese, nebo jsou viděné rozdíly nepodstatné (Buckley, 2009, s. 4).

Vizuálně ztrátová komprese je taková matematicky ztrátová komprese, která přináší vizuálně patrné změny obrazové kvality. Míra komprese v rámci tohoto typu se může dále výrazně lišit, přičemž nejvyšší možnou vizuálně ztrátovou kompresi lze pochopitelně užít pouze pro náhledy obrázku (tedy pro doplňkové prezentační formáty).

Ne všechny rastrové formáty umožňují všechny výše uvedené volby. Formát TIFF umožňuje všechny varianty: nekomprimovanou variantu (která je také nejčastěji užívaná pro TIFF jako archivační formát), matematicky bezeztrátovou kompresi (algoritmus LZW nebo ZIP) a ztrátovou kompresi (algoritmus JPEG). Formát JPEG nabízí pouze matematicky ztrátovou kompresi (typy 3-4).[34] Formát PNG nabízí pouze bezeztrátovou kompresi, podle studie FADGI s vynikajícími výsledky (FADGI, 2014, s. 3). Formát JP2 nabízí pouze komprimované varianty (typy 2-4). Koncept vizuálně bezeztrátové komprese (typ 3) je v současné praxi spojován právě s tímto formátem. Skupina výzkumníků z několika významných knihoven světa provedla rozsáhlý test vnímání (různě vysoké) ztrátové komprese formátu JP2 mezi čtenáři s cílem navrhnout vhodné profily formátu JP2 ve vizuálně bezeztrátové kompresi (Chapman, 2006). Některé knihovny pak začaly v praxi využívat vizuálně bezeztrátovou kompresi pro archivační formáty.[35] Pro matematicky bezeztrátovou kompresi se udává, že kompresní poměr je obvykle zhruba 2:1 (Buckley, 2008, s. 6).

2.1.4.5 Obecné obrazové vlastnosti

Směrnice FADGI určují čtyři základní obrazové vlastnosti rastrových dat prostorové rozlišení (*spatial resolution*), bitovou hloubku (*bit-depth*), barevný model (*color model*) a barevný profil (*color profile*).

[34] Srv. Buonora, 2008.

[35] Například Wellcome Library (Buckley, 2009)

Prostorové rozlišení určuje množství informací v rastrovém souboru z hlediska počtu pixelů (obrazových prvků) na jednotku měření, obvykle palec (odtud zkratka PPI),[36] tj. „stanovuje, jak blízko od sebe jsou jednotlivé pixely umístěny“; bitová hloubka „stanovuje maximální počet odstínů nebo barev v digitálním obrazovém souboru“ (FADGI, 2010, s. 4). Prostorové rozlišení a rozměry digitálního obrazu určují celkový počet pixelů v souboru; při určení požadované velikosti souboru je nutno zadat hodnotu prostorového rozlišení a rozměry (např. 300 ppi + 8x10 palců). Současným standardem pro rastrové obrazy ve stupních šedi i barevné podobě je užití bitové hloubky 8 bitů pro jeden pixel pro jeden kanál. Obrazy ve stupních šedi mají jeden kanál, barevné obrazy tři a více kanálů. Barevný model je způsob číselné specifikace barev s užitím tří nebo více kanálů. Například barevný model RGB obsahuje tři kanály o bitové hloubce 8 nebo 16 bitů.

Prostorové rozlišení tedy vymezuje, jak detailně může být převedena tištěná kniha (např. čitelnost písma), a bitová hloubka, jak věrně mohou být zachyceny její barvy nebo odstíny šedi (tedy barevná věrnost) ve výsledném souboru. Prostorové rozlišení, ani bitová hloubka logicky nemohou udávat ani zaručovat kvalitu uložených informací, pouze vymezují rozsah možné kvality digitalizačního převodu. Kvalita digitalizátu se odvíjí od míry detailnosti (např. velikost písma) a barevnosti předlohy, která je digitálně zachycována, a nastavených hodnot prostorového rozlišení a bitové hloubky. Obecně platí, že rozmezí prostorového rozlišení a bitové hloubky je na jedné straně ovlivněno minimální hranicí (tj. aby kniha ještě byla čitelná a její obrazové prvky do určité míry rozpoznatelné), na druhé straně maximální hranicí rozlišení, tj. takové, nad jejíž rámec snímání nemůže v principu přinést již žádný pozorovatelný rozdíl, a digitalizace ve vyšším rozlišení by byla neekonomická či jinak neúčelná. Existující doporučení se tedy mohou pohybovat pouze mezi těmito krajními případy. Doporučení současné praxe však také zohledňují skutečnost, že při masové digitalizaci je prakticky nemožné vytvářet tato nastavení pro každou knihu zvlášť. Z tohoto důvodu se vydávají plošná doporučení pro minimální rozlišení a bitovou hloubku pro různě definované kategorie knih z hlediska jejich předpokládané velikosti písma a barevnosti.

Pro běžné knihy novodobých fondů obsahující barevné prvky se obecně doporučuje barevný model RGB, bitová hloubka 24 bitů a minimální prostorové rozlišení 300-400 PPI (FADGI, 2016; The Association for library collections and technical services, 2013). Digitalizační projekt může vytvořit odlišné pracovní postupy například pro tištěné knihy

[36] Zkratka pro počet pixelů na palec (pixel per inch), někdy se užívá též počet pixelů na milimetr nebo centimetr.

obsahující barevné ilustrace nebo fotografie a tištěné knihy obsahující pouze text (a druhé snímat jen ve stupních šedi). Při masové digitalizace však může být z hlediska přípravy obtížné kontrolovat, zda kniha neobsahuje barevné prvky. Z tohoto důvodu se často vytváří jedno plošné nastavení obrazových vlastností pro všechny knihy.[37]

V případě prostorového rozlišení nemohou být doporučení minimálního rozlišení ničím jiným než predikcí očekávané nejmenší velikosti písma u určité skupiny předloh. V případě, že se ve skupině vyskytne anomálie, pak tato plošná doporučení přirozeně nemohou zaručit kvalitní výsledek snímání. Hodnota prostorového rozlišení se nastavuje pouze na skeneru, pro fotografování tento údaj přirozeně nemá smysl. Na výsledné rozlišení fotografie má vliv kvalita senzoru fotoaparátu a objektivu, vzdálenost knihy od objektivu a zaostření.

Zatímco nastavení výše uvedených vlastností (prostorové rozlišení, bitová hloubka a barevný model) v současné praxi nepředstavuje větší problém, barevný profil je složitější problematika. Barevný profil „určuje interpretaci číselných hodnot popisujících pixely v obrázku tím, že popisuje chování zařízení nebo rozsah barevného prostoru“ (FADGI, 2010, s. 45). Barevný prostor je „geometrická reprezentace barev v prostoru, který lze vizuálně vnímat nebo vytvářet užitím konkrétního barevného modelu“ (FADGI, 2017) Barevný prostor je například vyžadován aplikacemi pro zobrazení. Snímací zařízení zpravidla užívají vlastní barevný profil, který je závislý na konkrétním zařízení nebo výrobci. Tento technologicky závislý profil lze převést do ICC profilu (standardu pro univerzální barevnou specifikaci) a uložit do obrázku.

2.1.4.6 Prezentační varianty

Hlavním cílem současné praxe je vytvoření finálních produkčních dat v archivačním formátu. Obrazová data v archivačním formátu se následně v nezměněné podobě uchovávají v archivu do doby, než bude z důvodů zastarávání technologií nutno přistoupit k archivačním opatřením. Běžnou praxí je, že se v digitální knihovně čtenářům nezpřístupňují obrazová data v archivačním formátu, ale vytváří se jejich prezentační varianta v prezentačním formátu. V případě užití archivačního formátu TIFF bývají prezentační variantou zpravidla obrazová data ve formátu JPEG jako hlavním prezentačním formátu,[38] v případě archivačního formátu

[37] Příkladem je digitalizace projektu NDK, kde je plošně pro všechny digitalizované dokumenty nastaveno rozlišení 300 PPI, barevný model RGB, bitová hloubka 8 bitů.

[38] Srv. např. Smith, 2006, s. 10 a Vychodil, 2010, s. 64.

JP2 bývá vytvořen prezentační meziformát ve formátu JP2 (ve ztrátové kompresi), ze kterého systém digitální knihovny vytváří za chodu prezentační varianty ve formátu JPEG.

Důvody vytváření prezentační varianty pro zpřístupnění jsou různé a zpravidla jsou kombinací více faktorů. Zaprvé jsou dány historicky. Digitalizace se prováděla již v dobách, kdy bylo internetové připojení ještě pomalé nebo nákladné, a tudíž prezentační varianta tvořená obrazovými soubory menší velikosti (zejména ve ztrátové kompresi) byla uživatelsky vhodnou formou prezentace. Druhým důvodem je, že formát TIFF, hlavní volba digitalizačních projektů, nebyl a dosud není podporován internetovými prohlížeči. Třetím důvodem může být následování praxe obrazového průmyslu, kdy je obvyklým způsobem ukládat obrazové matrice v nejvyšší možné kvalitě a užívat je jako zdroj pro generování obrazových dat v různé kvalitě pro různé účely (mj. také prezentace na webu, např. v online periodicích). Odůvodněním také mohou být výzkumy stanovující psychofyziologický model, podle něhož je informační obsah reprodukovaný z obrazových dat ve ztrátové kompresi určité úrovně čtenářem vizuálně nerozeznatelný od informačního obsahu reprodukovaného z obrazových dat v bezeztrátové kompresi nebo nekomprimované podobě.[39] Z toho se vyvozuje, že zpřístupňování obrazových matric v archivním formátu je neúčelné. Nikoliv výjimečným případem současné praxe je, že se prezentační varianty vytvářejí již při produkci, jako součást finálních produkčních dat (a tedy nikoliv až v archivu), a to z důvodu jednoduššího zpracování.[40]

2.2 Standardy NDK a související předpisy

Národní knihovna ČR byla v letech 2009-2014 příjemcem projektu Vytvoření Národní digitální knihovny (projekt NDK), který byl financován z Integrovaného operačního programu EU programového období 2007-2014. NK ČR se v souvislosti s tímto projektem začala poprvé soustavněji věnovat digitální archivaci. Výstupy projektu byly tři: digitalizáty tištěných dokumentů, jejich archivace v archivu a zpřístupnění uživatelům. V rámci projektu byl vybudován archiv pod názvem LTP úložiště NK ČR. V listopadu 2011 byla vydána nová zřizovací listina NK ČR, v jejímž článku II.41 je již explicitně uveden závazek digitální archivace i odkaz na koncept důvěryhodného digitálního repozitáře. Realizátory projektu byly Národní knihovna ČR a Moravská zemská knihovna.

[39] Viz např. Chapman, 2007.

[40] Viz např. Standardy pro obrazová data na https://www.ndk.cz/standardy-digitalizace/standardy-pro-obrazova-data.

[41] „Formuluje strategie a postupy dlouhodobé ochrany elektronických dokumentů a provozuje důvěryhodné digitální úložiště.“

Standardy NDK byly vytvořeny v souvislosti s realizací projektu NDK. Byly určeny jako specifikace pro digitalizáty tištěných monografií a periodik vytvářených Národní knihovnou ČR a Moravskou zemskou knihovnou v rámci toho projektu i v období jeho udržitelnosti a budou zachovány i po skončení udržitelnosti. Účelem standardů NDK bylo vytvářet balíčky SIP tak, aby objekt CDO byl již v archivačním formátu a LTP úložiště NK ČR nemuselo provádět formátovou normalizaci. Metadatový profil byl navržen tak, aby umožnil zaznamenání všech důležitých informací, kterou popisují proces produkce digitalizátů.

V rámci NK ČR má odbornou a kurátorskou stránku digitální archivace na starosti specializovaný odbor (ODIF, Odbor digitálních fondů).[42] ODIF při přípravě projektu NDK zavedl nové standardy a postupy ve třech klíčových oblastech digitální archivace (metadata, trvalé identifikátory a datové formáty). Pro metadata vytvořil standard NDK (metadatový aplikační profil pro digitalizaci).[43] Profil je určen k zaznamenávání metadat v průběhu digitalizace tištěných dokumentů, která jsou důležitá z hlediska archivace. Pro trvalou identifikaci digitálních dokumentů odbor navrhl a vybudoval systém nazvaný ČIDLO, který je založen na standardu URN:NBN. Jeho součástí je i národní resolver.[44] Jako datový formát pro digitalizáty tištěných dokumentů byl vybrán relativně nový formát JP2 (JPEG 2000, Část 1 specifikace).

2.2.1 Standardy pro metadata

Standardy NDK, kterých se týká tato metodika, jsou historicky nejstarší z metadatových standardů NDK. Jsou rozděleny pro dvě skupiny dokumentů: monografie a periodika. Oficiální název je Definice metadatových formátů (DMF). Hlavním obsahem je metadatový profil, jehož základem je METS a PREMIS. Je třeba si uvědomit, že označení standard zde znamená, že jde o metadatový profil, který je českým národním standardem pro knihovny. Tento metadový profil je ale založen na užití mezinárodních metadatových standardů, které jsou mu nadřazené. Dále nejde o standard povinný pro všechny knihovny, povinnost je dána pouze pro ty, které mají data ukládat do LTP úložiště NK ČR.

[42] Na počátku nesl název Odbor digitální ochrany.

[43] http://www.ndk.cz/standardy-digitalizace/metadata

[44] https://resolver.nkp.cz/

Standard PREMIS není primárně určen pro záznam operací před odevzdání balíčku SIP do archivu. Lze jej však aplikovat, protože pro záznam událostí v produkci de facto jiný vhodný standard neexistuje (s jistými výjimkami v případě MIX).

DMF je metadatový profil založený na více souborech v METS XML v balíčku SIP (jeden pro bibliografická metadata, ostatní vytvářené zvlášť pro každý soubor obrazové matrice v archivním formátu) (Národní knihovna, 2016). V jiných aplikacích v zahraničí tomu bývá často jinak. Může být užit pouze jeden soubor METS XML v balíčku SIP.

Součástí standardu je i užití identifikátoru URN:NBN pro identifikaci digitálních dokumentů. Jeho přidělování zajišťuje systém ČIDLO, systém perzistentní identifikace digitálních dokumentů, který byl rovněž vytvořen v Odboru digitálních fondů NK ČR, a sice v letech 2011-2013 a který je dále vyvíjen. V současnosti systém ČIDLO přidělil digitalizovaným knihám a číslům periodik více než jeden milion identifikátorů URN:NBN, což z něj činí patrně největší identifikační systém dokumentů digitálního dědictví v ČR. Z analýzy evropských národních identifikačních systémů URN:NBN vyplývá, že český systém patří z hlediska pravidel a funkcí mezi nejkomplexnější (Cubr a kol., 2016).

V roce 2014 byla existující pravidla systému ČIDLO zapracována do Metodiky pro přidělování a správu životního cyklu unikátních perzistentních identifikátoru digitálních dokumentů podle standardu URN:NBN, která byla 5. června 2015 uznána Ministerstvem kultury ČR jako certifikovaná metodika a v roce 2018 aktualizována ve verzi 2.0[45]. Metodika pro přidělování a správu životního cyklu unikátních perzistentních identifikátoru digitálních dokumentů podle standardu URN:NBN, podobně jako předkládaná metodika pro vytváření balíčků SIP, představuje ucelený komplexní návod, jakým způsobem využívat systém ČIDLO.

2.2.2 Standardy pro formáty

Hlavním typem digitálních dat, se kterými NK ČR již dlouhodobě pracuje, jsou rastrová obrazová data (tvořená především digitalizací tištěných dokumentů). Jako archivační formát byl před projektem NDK užíván JPEG a jako prezentační formát DjVu. Při přípravě projektu bylo třeba vhodný formát znovu zvážit, s ohledem na výši investice, masový záběr plánované digitalizace a vývoj v oblasti digitální archivace. Již tehdy bylo také zřejmé, že formát DjVu jako prezentační formát (jako archivační nebyl zvažován nikdy) již zastaral.[46] Výsledkem

[45] https://www.ndk.cz/archivace/Certifik_metodika_urnnbn_2018.pdf

[46] Jedním z důvodů byla prohra souboje s jeho hlavním konkurentem, formátem PDF.

interní analýzy byl zvolen JP2 jako nový archivační i prezentační formát pro digitalizované (novodobé) tištěné dokumenty NK ČR. Jako hlavní výhody byly označeny otevřená dokumentace, neproprietárnost, kompresní možnosti a využitelnost pro archivaci i zpřístupnění. Úspěchem projektu NDK byla možnost využít matematicky bezeztrátovou kompresi pro archivační formát. Celá řada institucí, která v současnosti přechází nebo se chystá přecházet na JP2 jako archivační formát (konverzí z formátu TIFF), totiž volí matematicky ztrátovou kompresi, a to z důvodu významné úspory úložných kapacit. I když jde o vizuálně bezeztrátovou kompresi (tedy uživatel nic nepozná), ztrátově komprimovaný formát je vždy rizikem pro budoucí migrace. Pro prezentaci byla zvolena implementační varianta s image serverem. Profil JP2 pro zpřístupnění je prezentační meziformát pro vytváření dočasných obrázků ve formátu JPEG pro uživatele při jeho procházení digitální knihovnou. Tím byl vyřešen problém s nutností pluginu. Archivační formát je v JP2 s profilem v bezeztrátové kompresi, prezentační meziformát v JP2 s profilem ve ztrátové kompresi.

Pro OCR komponentu byl zvolen formát ALTO XML. V DMF se uvádí, že „OCR (ALTO XML) bude vznikat z uživatelské kopie - OCR je lepší ze souborů s kompresí (méně šumu)“ a „je nutné zachovat velikost obrazu uživatelských a archivních kopií stejnou (počet pixelů, rozlišení) tak, aby ALTO XML odpovídalo“. První doporučení však v praxi není využíváno.

2.2.3 Standardy pro obrazová data

Vedle předpisu konkrétního formátu a jeho profilu, obsahuje Standard NDK pro data vzniklá v podprogramu VISK 7 další doporučení pro tvorbu obrazových dat, kterými jsou :[47]

- ořez dokumentů cca 1 mm vně okraje dokumentu
- narovnání podle řádků textu
- veškeré úpravy obrazů je třeba provádět na archivních souborech
- uživatelský soubor se bude generovat až po všech úpravách
- uživatelská i archivní kopie musí mít stejný rozměr (v pixelech) a stejné rozlišení (v DPI).
- skenování v rozlišení minimálně 300 PPI
- barevná hloubka 24 bitů

[47] http://www.ndk.cz/standardy-digitalizace/standardy-pro-obrazova-data

- barevný model RGB

2.2.4 Podmínky VISK7 pro 2019

Podmínky pro VISK7 jsou vydávány každý rok. Na jejich vydávání se podílí Odbor digitálních fondů standardy NDK, které jsou pro VISK7 závazné, a dalšími doporučeními, týkajícími se zejména způsobu dodávání balíčků SIP.

V podmínkách VISK7 pro rok 2019 jsou specificky zdůrazněny následující podmínky, které souvisí s vyplňováním metadat podle standardů NDK. Žadatel musí pro digitalizované dokumenty získat tyto identifikátory:

- platná čísla České národní bibliografie
- identifikátor URN:NBN (přes systém ČIDLO)
- u periodických dokumentů číslo ISSN (pokud u starších periodických dokumentů dosud nebylo ISSN přiděleno, musí zajistit jeho přidělení u České národní agentury ISSN při Národní technické knihovně)
- sigla instituce (pokud dosud nebyla přidělena, musí o ni zažádat prostřednictvím Národní knihovny ČR).

V podmínkách pro VISK7 pro rok 2016 byla uvedena podmínka: „Konverze obrazových souborů pomocí OCR do textového formátu s úspěšností rozpoznávání min. 95%“ (Česko, 2015). V podmínkách pro VISK7 pro roky 2018, 2019 a 2020 tato podmínka již není.

II. IMPLEMENTAČNÍ ČÁST

Úvod

Uvedená doporučení jsou určena pro knihovny, případně jiné paměťové instituce, které vytvářejí balíčky SIP obsahující digitalizáty tištěných dokumentů, které jsou určeny k dlouhodobému uchovávání (digitální archivaci) v repozitáři. Doporučení se zaměřují především na obecnější procedurální postupy, konkrétnější postupy jsou uvedeny pouze v souvislosti s plněním technických metadat.

Metodika obsahuje pouze doporučení, která nejsou závazná, a snaží se navrhovat optimální postup, kterého nemusí být v současné praxi vždy možné dosáhnout, nicméně předpokladem je, že v budoucnosti to možné bude.

Metodika je určena jak pro knihovny, které digitalizují podle aktuálních standardů NDK pro digitalizáty periodik a monografií (tj. metadatových profilů vydávaných Národní knihovnou ČR), tak pro digitalizace podle jiných metadatových profilů, pokud zachovávají některé základní rysy standardů NDK.

Ve vztahu ke standardům NDK tato metodika navrhuje postupy, které jsou nad rámec toho, co je povinné v rámci standardů NDK, resp. specifikuje postupy, které standardy NDK nepopisují,

ale předpokládají. V případě rozporu této metodiky s podmínkami VISK 7 platí, že podmínky tohoto podprogramu jsou závazné pro knihovny, které z něj získávají dotaci na svou digitalizaci.

Terminologie

Archivační formát

formát digitalizátu, který je považován za aktuálně vhodný pro zajištění dlouhodobého uchovávání

Archivní kopie

konečný digitalizát v archivačním formátu

Datový validátor

nástroj pro validaci technických vlastností dat nad rámec validace formátu

Digitalizace

převod fyzické předlohy do digitální podoby

Digitalizační zařízení

fyzické zařízení zajišťující digitalizaci předlohy

Digitalizát

digitální dokument vzniklý digitalizací fyzické předlohy (tištěné, zvukové)

Digitální otisk

mechanismus pro kontrolu neporušenosti digitálních objektů, např. MD5, SHA-1

Formátová identifikace

jednoznačné určení formátu, optimálně prostřednictvím identifikátoru PUID

Formátový identifikační nástroj

nástroj pro formátovou identifikaci

Formátový validátor

nástroj pro validaci formátu, tj. pro kontrolu zda kontrolovaný formát soubor odpovídá specifikaci formátu

Charakterizace

proces zjištění informací (zejména technických informací, např. informace o formátu) o souboru, které jsou extrahovány přímo ze souboru

JP2

rastrový formát rodiny JPEG2000 podle Part I. specifikace JPEG2000

Kontrola neporušenosti

kontrola neporušenosti souboru užitím digitálního otisku

Metadatový extraktor

nástroj pro charakterizaci, jenž ze souboru získá především jeho technické vlastnosti

Metadatový validátor

nástroj pro validaci metadat

Prezentační formát

formát, ve kterém je digitalizát zpřístupňován uživatelům (např. v digitální knihovně)

Prvotní digitalizát

data, která jsou bezprostředním výstupem digitalizačního zařízení, tj. data, která digitalizační zařízení uloží do souboru nebo souborů na datový nosič po skončení procesu snímání

Předloha

fyzický objekt, ze kterého je digitalizačním zařízením vytvářen digitalizát

PUID

identifikátor registru PRONOM jednoznačně označující formát i jeho jednotlivé verze případně jiné parametry formátu (typ komprese apod.)

Původní snímek

prvotní digitalizát, který vznikl snímáním

Snímací zařízení

digitalizační zařízení pro digitalizaci tištěných předloh (skener, fotoaparát)

Snímání

digitalizace užitím snímacího zařízení

Snímkový formát

formát, ve kterém je uložen původní snímek

Snímek

digitalizát vzniklý snímáním

Standard NDK

standard pro digitální dokumenty vydaný Odborem digitálních fondů Národní knihovny ČR

Technická metadata

informace o vlastnostech souboru, jako je velikost, formát, komprese, zařízení, kterým byl vytvořen, obrazové a další specifické vlastnosti

Uživatelská kopie

konečný digitalizát v prezentačním formátu

Validace

automatická kontrola toho, zda jsou data nebo metadata vytvořena v souladu s deklarovanými specifikacemi

Validátor

nástroj pro validaci

3 Digitalizační projekt

3.1 Technické zajištění

3.1.1 Snímací zařízení

Mezi prvotní rozhodnutí při přípravě digitalizačního projektu patří určení toho, zda bude snímání provedeno formou skenování (skener), nebo fotografování (digitální fotoaparát).[48] Následně je potřeba věnovat dostatečné úsilí výběru adekvátního skeneru nebo fotoaparátu, který dokáže vytvářet digitalizáty v předepsané podobě. Optimálním variantou je vybraná snímací zařízení před jejich pořízením přímo otestovat.

Mezi základní požadavky na výběr skeneru patří dostatečné prostorové rozlišení skeneru a možnost ukládat původní snímky v nekomprimované podobě.

Jako základní zdroj pro otázky výběru vhodných snímacích zařízení lze doporučit směrnici FADGI (Federal Agencies Digital Guidelines Initiative) nazvanou „Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Files“ (aktuální verze byla vydána v roce 2016[49]), a českou „Metodiku pro vytváření bezpečnostních kopií archiválií v digitální podobě“[50] (zejména kapitoly 4.1-4.6).

3.1.2 Softwarové nástroje pro tvorbu digitalizátů

Pro vytváření digitalizátů tištěných dokumentů je potřeba využívat řadu vysoce specializovaných nástrojů, zejména pro zpracování obrazových dat, formátové konverze nebo optické rozpoznávání znaků. Specializované nástroje pro tvorbu digitalizátů je potřeba vybrat již ve fázi přípravy digitalizačního projektu a provést testování, zda nástroje skutečně dokáží vytvářet digitalizáty v předepsané podobě. Toto testování se může opakovat, dokud nebude nalezen vhodný nástroj. Vzhledem k tomu, že některé kvalitní nástroje jsou komerční, musí výběr zohledňovat nejen kvalitu výstupů softwarových nástrojů, ale také finanční náklady spojené s jejich pořízením a užitím. Finanční náklady spojené se softwarovými nástroji musejí být důkladně propočítány a zahrnuty do rozpočtu digitalizačního projektu. Je špatnou praxí neověřovat si náklady spojené se specializovaným softwarem, který je nezbytný pro vytváření

[48] Tato metodika se blíže zaměřuje na skenování.

[49] http://www.digitizationguidelines.gov/guidelines/digitize-technical.html

[50] http://cesarch.cz/wp-content/uploads/2015/06/metodika-pro-bezpecnostni-digitalizaci_v1.pdf

digitalizátů tištěných dokumentů, a teprve při zahájení digitalizace zjistit, že rozpočet projektu není dostatečný.

3.1.3 Validační nástroje

Pro kvalitní digitalizaci, která zohledňuje požadavky dlouhodobého uchovávání, jsou nezbytné nejen produkční softwarové nástroje, ale také specializované validační nástroje. Zejména důležitá je validace metadat, souborových formátů a balíčku SIP.

Ve fázi přípravy digitalizace je zapotřebí, aby zvolené validátory byly otestovány, zejména jejich zapojení do celkového digitalizačního systému. Je nutné, aby digitalizační systém dokázal integrovat tyto specializované nástroje. Pokud nedojde k testování validačních nástrojů, může se stát, že po zahájením digitalizace v daném projektu budou zjištěny dodatečné náklady na integraci validátorů, se kterými původní rozpočet projektu nepočítal.

3.1.4 Kontrola předloh

Tištěné předlohy, které mají být zdigitalizovány v daném projektu, by měly být důkladně zkontrolovány z hlediska kvality a úplnosti před zahájením digitalizace. V případě, že součástí projektu musí být seznam konkrétních předloh, které mají být zdigitalizovány (a tedy existuje závazek uvedené tituly zdigitalizovat), měly by být tyto předlohy zkontrolovány ještě před podáním projektového záměru. Pro digitalizaci by měl být vždy vybírán takový exemplář, který je úplný a nejzachovalejší (v případě digitalizace periodik by měl být kompletní minimálně ročník). Pokud knihovna nedisponuje žádným úplným exemplářem, případně jsou všechny její exempláře ve špatné kvalitě, měla by zvážit, zda je nutné, aby takovou předlohu digitalizovala – je zde vždy možnost přenechat digitalizaci daného titulu monografie nebo ročníku periodika jiné knihovně. Další variantou je zapůjčit úplný a kvalitní exemplář z jiné knihovny. Nedoporučuje se kompletovat jeden titul monografie nebo čísla periodika z více exemplářů.

3.2 Základní standardizační doporučení

3.2.1 Stanovení základní intelektuální entity a granularity

Základní intelektuální entitou se rozumí intelektuální entita, která je obsažena v úplnosti jak fyzicky (v podobě dat), tak z hlediska bibliografického popisu (v podobě odpovídajících metadat) v jednom balíčku SIP (např. číslo periodika). V balíčku SIP mohou být obsaženy i nižší intelektuální entity, které tvoří části základní intelektuální entity (na úrovni popisu i

fyzicky – např. články), a popsány vyšší intelektuální entity, jejichž částí je základní intelektuální entita (např. ročník nebo titul). Základní intelektuální entita by měla vždy tvořit smysluplný celek (např. číslo periodika, a nikoliv svazek několika čísel).

Toto doporučení se týká digitalizací, kdy si knihovna sama specifikuje podobu balíčku SIP. V případě standardů NDK je výše uvedené doporučení obsaženo jako předpis, který je nutno dodržovat.

3.2.2 Perzistentní identifikátory tištěné předlohy

Pro propojení tištěných předloh s digitalizáty je klíčové užití perzistentních identifikátorů tištěné předlohy na úrovni titulu. Odpovídající identifikátory tohoto typu jsou ISBN (International Standard Book Numbering), ISMN (International Standard Music Numbering), ISSN (International Standard Serials Numbering) a čČNB (číslo České národní bibliografie), který by měly být zaznamenány do metadat digitalizátu tištěné předlohy, pokud to pravidla těchto identifikačních systémů umožňují. S výjimkou posledního jde o identifikátory řízené mezinárodními organizacemi prostřednictvím národních agentur.

Identifikátor ISBN je přidělován českým knihám od roku 1989. Řídí jej Národní agentura ISBN v ČR při Národní knihovně ČR.[51]

Identifikátor ISMN je přidělován českým hudebninám od roku 1996.[52] Řídí jej Národní agentura ISMN v ČR při Národní knihovně ČR.[53]

Identifikátorem ISSN jsou označována periodika v ČR (ČSSR) od 70. let 20. století. Přiděluje jej České národní středisko ISSN při Národní technické knihovně. [54]

Ve všech případech je účast vydavatelů dobrovolná a neplatí, že veškerá česká produkce daných typů publikací od doby zavedení těchto identifikačních systémů do českého prostředí obsahuje některý z těchto identifikátorů. Identifikátory ISBN a ISMN jsou většinou uvedeny v samotném

[51] https://www.nkp.cz/sluzby/sluzby-pro/isbn-ismn-issn

[52] https://www.nkp.cz/soubory/ostatni/prirucka-ismn.pdf

[53] https://www.nkp.cz/sluzby/sluzby-pro/isbn-ismn-issn/oma#ISMN

[54] https://www.techlib.cz/cs/2844-ceske-narodni-stredisko-issn

dokumentu (knize, hudebnině) a nelze je přidělovat zpětně (tj. po vydání)[55]. Identifikátor ISSN lze přidělovat zpětně a o jeho přidělení může požádat nejen nakladatel periodika, ale i knihovna.

Identifikátor čČNB je užíván v ČR od roku 2010.[56] Jedná se o kód národní bibliografie, pro který je ve formátu MARC 21 vyhrazeno podle 015 a který některé jiné národní knihovny přidělují již desítky let záznamům svých národních bibliografií. Je přidělován popisné jednotce České národní bibliografie, nelze jej tedy přidělit jednotlivým číslům periodika, jednotlivým svazkům vícesvazkového díla bez významných názvů části a monografické řadě/edici, která zahrnuje díla s vlastními názvy.[57] Aby byl perzistentní, musejí být zachovány i neplatné identifikátory čČNB (zneplatněné např. po sloučení záznamu). Identifikátor čČNB by podle současných pravidel měly mít „veškeré publikované dokumenty vydané na území ČR od roku 1801 do současnosti“, na stránkách Souborného katalogu je dále uveden výčet dokumentů, které nemají nárok na přidělení, s tím, že v případě pochybností se tento identifikátor nepřidělí.[58]

Pokud tištěné periodikum nemá přidělený identifikátor ISSN, měla by knihovna požádat o jeho přidělení České národní středisko ISSN, a to s dostatečným předstihem před zahájením digitalizace. Skutečnost, zda dané periodikum má přidělen identifikátor ISSN, lze ověřit v národní databázi ISSN, kterou spravuje Národní technická knihovna.[59]

Pokud pravidla pro přidělování identifikátoru čČNB umožňují přidělení tohoto identifikátoru[60] a tištěná předloha na úrovni záznamu titulu jej zatím nemá, měla by knihovna požádat o jeho přidělení Souborný katalog ČR.[61] I zde je třeba, aby se tak stalo s dostatečným předstihem před zahájením digitalizace.

V případě digitalizací financovaných z podprogramu VISK 7 je přítomnost identifikátorů čČNB a ISSN povinná.

---

[55] Je-li dokument zaveden do národní bibliografie, pak mu jsou i zpětně přiděleny tyto identifikátory a jsou tedy součástí katalogizačních záznamů knihovny NK ČR.

[56] https://www.caslin.cz/caslin/spoluprace/sluzby-souborneho-katalogu-cr/cislo-cnb-v-sk-cr

[57] http://www.registrdigitalizace.cz/rdcz/info/data/ccnb

[58] https://www.caslin.cz/caslin/spoluprace/sluzby-souborneho-katalogu-cr/cislo-cnb-v-sk-cr

[59] http://aleph.ntkcz.cz/F/?func=find-b-0&local_base=stk02

[60] https://www.caslin.cz/caslin/spoluprace/sluzby-souborneho-katalogu-cr/cislo-cnb-v-sk-cr

[61]https://www.caslin.cz/caslin/spoluprace/sluzby-souborneho-katalogu-cr/jak-spravne-postupovat-nez-zacne-knihovna-digitalizovat-dokument

3.2.3 Perzistentní identifikátory digitalizátu

Jako hlavní perzistentní identifikátor digitalizátu tištěných dokumentů na úrovni intelektuální entity se doporučuje užívat identifikátor URN:NBN. Ten v českém prostředí zajišťuje služba ČIDLO (Český identifikační a lokalizační systém).[62]

Identifikátor URN:NBN lze přidělit v případě digitalizovaných periodik článku, číslu a ročníku (titulu digitalizovaného periodika URN:NBN přidělit nelze), v případě digitalizovaných monografií svazku monografie (jednodílové nebo vícesvazkové), příloze monografie a vnitřní části monografie (kapitole). Pro odlišnou úroveň granularity musí být přidělen jiný identifikátor URN:NBN.

Podle aktuálních standardů NDK pro digitalizáty tištěných dokumentů (verze 1.7.1 pro digitalizovaná periodika a 1.3.1 pro digitalizované monografie)[63] musí být identifikátor URN:NBN povinně přidělen základní intelektuální entitě (číslu periodika, resp. svazku monografie).

Podrobná pravidla pro přidělování identifikátorů URN:NBN i řízení celého jejich životního cyklu jsou obsažena v certifikované metodice popisující pravidla systému ČIDLO, která byla vydána v roce 2015 a aktualizována v roce 2018. Účastníkem (registrátorem) systému ČIDLO se může stát jakákoliv knihovna nebo jiná instituce, která má v systému ADR (Centrální adresář knihoven a informačních institucí v ČR)[64] Národní knihovny ČR přidělenu siglu. Podmínkou je dodržování pravidel systému. Registraci zařizuje kurátor systému ČIDLO (urnnbn@nkp.cz). V případě, že pro knihovnu provádí digitalizaci externí dodavatel, musí mít souhlas dané knihovny se zastupováním v systému ČIDLO.

Jednou z funkcí systému ČIDLO je i trvalé přesměrovávání na aktuální URL adresu digitalizátu tištěného dokumentu v digitální knihovně. Za tímto účelem musí knihovna poskytnout systému ČIDLO součinnost v procesu sklízení aktuálních adres URL přes protokol OAI-PMH (Open Archives Initiative Protocol for Metadata Harvesting).

[62] https://www.ndk.cz/archivace/resolver-urn-nbn-sluzba-cidlo

[63] https://www.ndk.cz/standardy-digitalizace/metadata

[64] https://aleph.nkp.cz/cze/adr

3.2.4 Projektová dokumentace

Knihovna by měla k digitalizačnímu projektu vytvořit podrobnou dokumentaci, kterou by optimálně měla zpřístupňovat spolu s digitalizáty tištěných dokumentů. Tato dokumentace by měla přinejmenším obsahovat informace o užitých metadatových standardech (např. konkrétních metadatových standardech NDK včetně jejich verze) a zvolených archivačních a prezentačních formátech a optimálně též informace o užitých validačních nástrojích.

4 Digitalizace

4.1 Příprava bibliografických záznamů

Před zahájením vlastního snímání předlohy musí existovat její bibliografický záznam v katalogu knihovny, která digitalizuje. Tento záznam by také měl být důkladně zkontrolován. V případě nekvalitního nebo neúplného záznamu by měla být provedena rekatalogizace. Kontrolu záznamů (resp. rekatalogizaci) předloh určených k digitalizaci by měl vždy provádět školený a zkušený katalogizátor. Z toho důvodu by digitalizační projekt měl počítat s vyčleněním odpovídající pracovní síly (katalogizátorem na alespoň částečný úvazek). Nedoporučuje se kontrola pracovníky, kteří nejsou katalogizátory. Kvalita katalogizačního záznamu má zásadní vliv na kvalitu digitálních bibliografických metadat uložených v balíčků SIP a následně metadat zobrazených v digitální knihovně čtenářům. Šetření prostředků na katalogizátory při digitalizaci může významným způsobem snižovat kvalitu výstupů digitalizačního projektu. Také platí, že pozdější opravy bibliografických metadat (tj. opravy v repozitáři a digitální knihovně) jsou mnohem komplikovanější a technicky, personálně i finančně nákladnější.

Bibliografický záznam by již měl obsahovat identifikátory ISBN, resp. ISMN, ISSN nebo čČNB, pokud lze některý z těchto identifikátorů na základě pravidel daných identifikačních systémů přidělit.

Pro převod bibliografických údajů do digitálních metadat (standardy MODS a DC) by měl být užit vlastní katalog knihovny, nikoliv jiné katalogy (souborný katalog, báze ČNB apod.), byť by tyto jiné katalogy obsahovaly agregované záznamy knihoven zapojených do digitalizace.

4.2 Snímání předloh

4.2.1 Věrnost reprodukce tištěné předloze

Digitalizát by měl být co nejvěrnější digitální reprodukcí tištěné předlohy. Pro účely specifikace věrnosti doporučujeme využít směrnici DLF (Digital Library Federation) nazvanou „Benchmark for Faithful Digital Reproductions of Monographs and Serials“[65] Směrnice uvádí, že cílem věrnosti je „přesně reprodukovat výchozí zdrojový dokument, s ohledem na jeho úplnost, vzhled původních stránek (včetně tonality a barvy) a správnou (tj. původní) posloupnost stránek.“ (The digital library federation benchmark working group, 2002, s. 2). Směrnice DLF také uvádí několik zásad, které musejí splňovat obrazová data reprezentující tištěnou předlohu. Tyto zásady doporučujeme využít v maximální možné míře, pokud směrodatné podmínky pro digitalizační projekt (např. pravidla podprogramu VISK7 nebo standardy NDK) neuvádějí jinak.

4.2.2 Základními parametry pro skenování

Jako minimální parametry pro skenování doporučujeme prostorové rozlišení nejméně 300 PPI[66], barevnou hloubku nejméně 24 bitů (tj. 3 x 8 bitů) a barevný model RGB. Tato doporučení jsou v souladu s aktuálním standardem NDK pro obrazová data.[67]

Pokud je to časově možné, doporučujeme provést zběžný průzkum míry detailnosti předloh určených pro digitalizační projekt a v případě potřeby rozlišení zvýšit paušálně na 400 PPI[68], případně i vyšší. Za tímto účelem lze užít tzv. Quality Index (QI) (viz doporučení AIIM TR26-1993 Resolution as it Relates to Photographic and Electronic Imaging, původně určené pro mikrofilmy). Tento výpočet bere v úvahu velikost písmen[69] předlohy a plánovanou kvalitu výsledného obrazu na stupnici od špatné kvality po kvalitu excelentní. Index udává kolik obrazových bodů (pixelů) je potřeba pro reprezentaci nejmenšího písmene ve zdrojovém textu. Barevné a šedé obrazy vyžadují nejméně 16 obrazových bodů (Quality Index =8) pro excelentní, detailní, zobrazení nejmenšího písmene zdrojového textu, bitonální obrazy potřebují 24 obrazových bodů. Pro písmeno o velikosti 1 mm je tak ideální snímací rozlišení 400 PPI,

[65] http://old.diglib.org/standards/bmarkfin.pdf

[66] Rozlišní 300 PPI postačuje pro text s velikostí písmen 1,4mm.

[67] https://www.ndk.cz/standardy-digitalizace/standardy-pro-obrazova-data

[68] Dle směrnice FADGI je rozlišení 400 PPI použitelné i pro vzácné knihy.

[69] V případě netextových materiálů (mapy, nákresy apod.) se výpočet odvíjí od šířky nejtenčí čáry, tahu.

písmeno pak bude reprezentováno 16ti obrazovými body u obrazů v plných barvách a v odstínech šedi.

Pokud takový průzkum není časově možný, doporučujeme řídit se obecnými doporučeními, které jsou obsaženy ve směrnici FADGI.[70] Tato směrnice pro různé kategorie dokumentů (rukopisy, další vzácné dokumenty, knihy, noviny, fotografie aj.) doporučuje optimální hodnotu rozlišení v závislosti na požadované výsledné kvalitě. Doporučujeme využít uváděné hodnoty rozlišení pro nejvyšší kategorii kvality („4 Star“).

4.2.3 Snímkový formát

Jako snímkový formát obrazových dat (formát původních snímků) by měl být vždy zvolen nekomprimovaný formát. V případě skenování by měl být užit formát TIFF, verze 6, bez komprese; v případě fotografování formát rodiny RAW (například fotoaparáty firmy Canon využívají formát Canon RAW). Tento požadavek je obzvláště důležitý vzhledem k tomu, že jako archivační formát (výstup digitalizace určený k dlouhodobému uchovávání) by měl vždy vybrán formát s bezeztrátovou kompresí. Užití ztrátové komprese pro snímkový formát znamená, že archivační formát bude pouze formálně bezeztrátový, ale reálně bude obsahovat ztrátová obrazová data.

4.2.4 Zabudování EXIF metadat do souborů

Pro digitalizaci je vhodné využívat EXIF metadata zabudovaná do obrazových dat. EXIF je široce rozšířeným a obrazovým průmyslem podporovaným standardem pro zabudovaná metadata pro rastrová data. Tato metadata jsou obsažena přímo v souboru a obsahují informace týkající se procesu snímání. EXIF není podporován ve všech obrazových formátech, ale jen pro JPEG, TIFF a RAW. Formáty TIFF a RAW jsou doporučenými formáty pro původní snímky. Je vhodné, aby tyto původní snímky tato metadata obsahovaly. Zpravidla je tato funkce již nastavena na snímacích zařízeních (skenery, fotoaparáty), případně je třeba zkontrolovat, zda funguje nebo tuto funkci na snímacím zařízení nastavit. Tyto údaje nejsou sice zachovány v doporučovaném formátu pro archivní kopie (JP2), který standard EXIF nepodporuje, ale jsou doporučeným zdrojem pro plnění technických metadat o formátu TIFF, resp. RAW.

[70] http://www.digitizationguidelines.gov/guidelines/digitize-technical.html

4.2.5 Barevný profil

Při digitalizaci by měl být vybrán adekvátní barevný profil. Na základě směrnice FADGI doporučujeme využít pro archivní kopie digitalizovaných novin barevný profil sRGB, pro ostatní tištěné dokumenty Adobe RGB 1998. Snímek musí vždy obsahovat informaci o barevném prostoru. Buď tuto informaci obsahuje snímek z procesu skenování nebo je v následných procesech nutné snímku barevný prostor přiřadit či původní snímek převést do vhodného barevného profilu. (FADGI, 2016, s. 61)

4.3 Zpracování dat

Data, které tvoří konečný digitalizát, lze rozdělit do tří skupin – obrazová komponenta (archivní a uživatelské kopie), OCR komponenta (textová data vzniklá optickým rozpoznáváním znaků, která souřadnicově sedí na obrazová data) a textová komponenta (prostý text pro vyhledávání).

4.3.1 Zpracování obrazové komponenty

Zpracování obrazové komponenty začíná zpracováním původních snímků a končí vytvořením archivních a uživatelských kopií. Každý krok zpracování, který končí uložením souboru, vytváří novou generaci obrazové komponenty.

Ořezy by se měly provádět na původním snímku ve formátu TIFF a měly by být provedeny cca 1 mm vně okraje stránky[71]. Vyrovnávání zešikmení by se mělo provádět rovněž na původním snímku, a to podle prostředního řádku textu stránky[72].

V případě fotografování doporučujeme provádět uvedené úpravy na původním snímku RAW v aplikaci výrobce fotoaparátu a po jejich skončení provést konverzi do formátu TIFF (verze 6, bez komprese).

Po skončení ořezů by měly být soubory ve formátu TIFF převedeny do archivních a uživatelských kopií, přičemž soubor archivní i uživatelské kopie by měl být vytvořen z téhož souboru TIFF předchozí generace. Soubor archivní i uživatelské kopie musí mít stejnou

[71] Ořezy vně uchovají hrany stránek, které mohou nést potenciálně zajímavé informace pro koncové uživatele, badatele. Je však zpravidla nutné takový ořez provádět manuálně nebo poloautomaticky. U novodobé literatury, kde se informace o hraně považuje za nevýznamovou je tedy akceptovatelný i ořez dovnitř. Takový ořez by pak měl být minimálně, tj. neořezávat víc než je nutné. (Vychodil, 2012)

[72] Případně dle převažující části textu, není-li stránka tištěna rovně, cílem je aby strana působila rovně.

pixelovou velikost i stejné rozlišení. Jedním z důvodů je namapování OCR komponenty pro současná i budoucí (vzniklá budoucí formátovou konverzí v repozitáři) obrazová data.

Jako formát archivních i uživatelských kopií doporučujeme využít formát JP2 (obrazový rastrový formát specifikovaný v první části standardu ISO/IEC 15444-1:2004) v souladu s aktuálními standardy NDK pro obrazová data (blíže viz další oddíly).

Všechny generace obrazové komponenty by měly být uchovány nejméně do doby úspěšného dodání balíčku SIP do repozitáře. Důvodem zachování je možnost případných oprav (chyb, které vznikly až ve zpracování), kontrola kvality (dohledatelnost postupu) a vytváření a validace technických metadat.

V řetězci datových transformací obrazové komponenty od původních snímků do archivních kopií nesmí být nikdy užita ztrátová komprese. Změny ve zpracování od původních snímků do archivních kopií budou nutně zahrnovat několik datových transformací (formátové konverze, ořezy apod.). Žádná z nich nesmí zahrnovat ztrátovou komprimaci obrazových dat.

4.3.2 Archivační formát (formát pro archivní kopie)

Archivační formát (formát archivních kopií) je v projektu Národní digitální knihovna bezeztrátový JP2. Následující text popisuje nastavení parametrů pro formátový profil JP2 pro archivní kopie.

4.3.2.1 Druh komprese a transformace

Druh komprese (Compression)	Bezeztrátová
Transformace (Transformation)	5-3 reversible filter

Obrazy budou komprimovány bezeztrátově, aplikací filtrů 5-3.[73]

4.3.2.2 Kompresní poměr

Výsledný kompresní poměr (Compression ratio)	Záleží dle charakteristik obrazu.

[73] Aplikace filtrů 9-7 na řádky a sloupce obrazu vede k výsledné ztrátové kompresi.

Kompresní poměr se u migrace do bezeztrátového JP2 explicitně neudává, nástroj je na základě vlastností obrazů spočítá sám.[74] Barevné obrazy mohou mít při bezeztrátové kompresi kompresní poměru okolo 1:2, strana s textem okolo 1:5 a prázdná, bílá strana i 1:400.

4.3.2.3 Dlaždice

Dlaždice (Tiling)	4096x4096

Vstupní obraz může být v začátku transformačního procesu zpracováván buď jako celek (tj. jeden obraz odpovídá jedné dlaždici) nebo může být rozdělen na dlaždice čtvercového tvaru. Každá dlaždice je pak zpracovávána separátně. Rozdělení obrazu na dlaždice urychluje proces komprese a dekomprese obrazu. Minimální povolená velikost jedné dlaždice je 128x128 obrazových bodů, tato velikost vede ke generaci velkého počtu dlaždic a naopak ke zpomalení procesů. Pro archivační obrazy není důležitá doba dekomprese obrazu a není tedy nutné obraz dělit na menší dlaždice, jako je tomu například u zpřístupňujících kopií JP2 v NDK[75]. Zároveň také dnes používané systémy velmi rychle dlaždice o velikosti 4096x4096 dekomprimují, není tedy nutné datový tok archivních kopií dále zesložiťovat rozdělením obrazu na menší dlaždice.

4.3.2.4 Průběh zobrazení

Průběh zobrazení (Progression order)	RPCL

Parametr průběh zobrazení (Progression order) udává, jak budou posílány pakety při přenosu dat a při jejich dekompresi. Standard definuje 5 způsobů zobrazení: LRCP, RLCP, RPCL, PCRL a CPRL, kde jednotlivá písmena odpovídají přenášeným datům v paketu (L =vrstvě kvality, R= rozlišení, C= barevná komponenta a P= pozici). V případě hodnoty zvolené pro projekt Národní digitální knihovny, tj. RPCL, dochází k seskupování paketů dle rozlišení, tj. nejdříve jsou shromážděna všechna data odpovídající první vrstvě rozlišení, která je tak

[74] U zpřístupňujících kopií ve formátu JP2 je kompresní poměr explicitně zadán, jedná se však o hodnotu aplikovanou na vstupní data nikoliv o výslednou hodnotu komprese obrazu, ta se může, a je to pro projekt NDK přijatelné, od předepsané aplikované hodnoty mírně lišit, vliv na konečný kompresní poměr mají totiž vlastnosti konkrétního obrazu. Aby výsledná hodnota kompresního poměru odpovídala předepsané hodnotě, musel by se pro každý jednotlivý obraz spočítat bitrate, což je proveditelné, nikoliv však pro projekt NDK nutné.

[75] Doba dekomprese dlaždic o velikosti 1024x1024 je rychlejší než u velikosti dlaždic 4096x4096.

zobrazen ve své maximální možné kvalitě, následně se nahrává další vrstva rozlišení a jí odpovídající barevnost, kvalita apod. V případě tohoto pořadí zobrazení uživatel vidí postupně se zvětšující obraz.

4.3.2.5 Dekompoziční úrovně

Počet dekompozičních úrovní (Decomposition level)	5 nebo 6

Počet dekompozičních úrovní se odvíjí od velikosti vstupního obrazu a od požadavků organizace na zpřístupnění (velikost náhledu, počet „obrazů“ mezi náhledem a plnou velikostí obrazu). Přítomnost několika dekompozičních úrovní má příznivý vliv na výsledný obraz a obecně se doporučuje do obrazu jich několik dát, standardem se zdá být minimální počet 5 dekompozičních vrstev.

4.3.2.6 Vrstvy kvality

Počet vrstev kvality (Quality layers)	1

Počet vrstev kvality zjednodušeně určuje, kolik obrazů s různou úrovní kvality (s různým kompresním poměrem) je možné z jednoho datového proudu extrahovat. Protože archivní kopie jp2 nejsou primárně určeny pro zpřístupnění je dostatečná jedna úroveň kvality.

4.3.2.7 Regiony

Velikost regionů (Precinct size)	256x256 pro první dvě dekompoziční úrovně, 128x128 pro nejnižší úrovně

Okrsky (příp. regiony, angl. precincts) sdružují bloky kódů souvisejících dat, jež se ukládají do jednotlivých balíčků a slouží k tomu, aby místně související data byla umístěna v jednom balíčku, přičemž jeden okrsek může být rozdělen do více paketů. Okrsky, stejně tak jako dlaždice, umožňují přístup k vybraným částem obrazu, tj. například načtení jednotlivých vybraných okrsků. Použitím okrsků lze tedy pro zpřístupnění rozdělit dlaždice na menší části.

4.3.2.8 Zájmové oblasti

Zájmové oblasti (Regions of Interests)	Ne

Zájmové oblasti jsou části obrazu, které jsou dekomprimovány prioritně vůči ostatním částem obrazu, jež jsou pro uživatele méně významné (např. pozadí). Při dekompresi se tedy tato část zobrazuje obvykle nejdříve a ve vyšší kvalitě. (Vrtělová, 2017) V projektu NDK se funkce zájmových oblastí nevyužívá.

4.3.2.9 Velikost bloků

Velikost bloků (Code block size)	64 x 64

Bloky kódu (codeblock) jsou čtvercového tvaru a jsou na sobě nezávisle kódovány do výsledného datového toku. Velikost bloků může nabývat hodnot od 4x4 pixelů do 1024x1024 pixelů. Nejobvyklejší velikost těchto bloků v paměťových institucích je 64x64 obrazových bodů.

4.3.2.10 Lokalizace dlaždice

Značka lokalizující dlaždice TLM (Tile Length Markers)	Ano

Značkovací segment TLM nese informaci o délce dlaždic, resp. částí dlaždic v celém toku dat jednoho souboru. Tato informace může následně posloužit k rychlejší lokalizaci a orientaci v proudu dat při použití více dlaždic.

4.3.2.11 Přemostění

Přemostění (Bypass)	Ano

Parametr přemostění, tj. BYPASS se týká procesu komprese a dekomprese, jedná se o režim, v jakém bude obraz zpracován. Hodnota BYPASS znamená, že kodér při kompresi

vynechá kompresi některých, méně významných dat, čímž se urychlí proces komprese i následné dekomprese (i o 20%)[76]. Výsledná komprese obrazu je pak o něco menší.

4.3.2.12 ICC profily

ICC profily (ICC Profile)	Ano

Obrazy by měly vždy obsahovat informaci o svém barevném profilu, což zajistí, že barvy budou v následných zobrazovacích aplikacích správně interpretovány. Formát JP2 (tj. Part 1 standardu JPEG2000) podporuje barevný prostor sRGB a vybrané ICC profily. Pro účely digitální archivace se kvůli co nevěrnějšímu zachování barev doporučují profily Adobe RGB a ProPhotoRGB, které mají větší rozsah barev (gamut).

4.3.2.13 Hlavička segmentu packetů

Značka začátku hlavičky segmentu paketů SOP (Start of Packet Header)	Ano (Cuse_sop=yes)
Značka konce hlavičky segmentu paketů EPH (End of Packet Header)	Ano (Cuse_eph=yes)

Značky SOP a EPH označují začátek a konec paketů tvořících tok jednoho souboru. Zvyšují odolnost souboru proti přenosovým chybám, přijímající systém (protokol) díky nim dokáže rozpoznat, že mu nějaká data chybí.

4.3.2.14 Vložená metadata

Vložená metadata (Embedded Metadata)	Ne

Do obrazů ve formátu JPEG2000 je možné vložit související metadata[77], například identifikátor, informace o souvisejících právech apod. Takové soubory pak mohou obsahovat dostatečné informace a nehrozí jejich ztráta v systému. Aktuálně se tato funkcionalita formátu

[76] Nástroj OpenJpeg má až do verze 2.2.0 tento režim rozbitý

[77] Nejedná se o metadata EXIF, formát JP2 nepodporuje EXIF standard.

v projektu NDK nevyužívá, obrazy jsou ale pojmenovány s využitím jednoznačného identifikátoru související intelektuální entity.

4.3.3 Prezentační formát (formát pro uživatelské kopie)

Pro zpřístupňování je doporučeno užít formát JP2 ve ztrátové kompresi, jež umožní rychlejší přenos a zobrazení dokumentu uživateli[78]. Tento JP2 je prezentačním meziformátem (zvaný někdy jako production master copy), tj. v prezentačním systému z něj aplikace (tzv. image server) generuje obrázky ve formátu JPEG[79].

PARAMETRY PRO FOPRMÁTOVÝ PROFIL PRO UŽIVATELSKÉ KOPIE	
Druh komprese (Compression)	Ztrátová
Transformace (Transformation)	9-7 reversible filter
Výsledný kompresní poměr (Compression ratio)	1:8 až 1:30
Dlaždice (Tiling)	1024x1024
Průběh zobrazení (Progression order)	RPCL
Počet dekompozičních úrovní (Decomposition level)	5 nebo 6
Počet vrstev kvality (Quality layers)	12 (logaritmicky)
Velikost regionů (Precinct size)	256x256 pro první dekompziční úrovně, 128x128 pro nejnižší dekompoziční úrovně
Zájmové oblasti (Regions of Interests)	Ne
Velikost bloků (Code block size)	64x64

[78] Tento kompresní poměr je ideální zvolit dle skenovaných dokumentů a ověřit experimentálně. V projektu NDK bylo zjištěno, že i při kompresním poměru 1:20 nese obraz významné informace. V některých případech je však vhodnější kompresní poměr nižší.

[79] Tímto odpadá na straně uživatelů nutnost obrazy stahovat nebo instalovat plugin pro zobrazení obrazů JP2. Prohlížeče nativně formát JPEG podporují.

Značka lokalizující dlaždice TLM (Tile Length Markers)	Ano („R“)
Přemostění (Bypass)	Ano
ICC profily (ICC Profile)	Ano
Značka začátku hlavičky segmentu paketů SOP (Start of Packet Header)	Volitelné
Značka lokace hlavičky segmentu paketů EPH (End of Packet Header)	Volitelné
Vložená metadata (Embedded Metadata)	Ne

4.3.4 Vytváření OCR komponenty

OCR komponenta (Textová data získaná nástrojem pro optické rozpoznávání znaků) musí být zapsána ve formátu ALTO XML.[80] OCR komponenta (výstup OCR) by měla být vytvářena až z archivních kopií ve formátu JP2. Důvodem je zachování mapování textu na pixelovou pozici v obraze. V metadatové části souboru s OCR by mělo být uvedeno, z jakého souboru byl daný soubor s OCR vytvořen (element „sourceImageInformation“).

4.4 Vytváření metadat

4.4.1 Převod bibliografických metadat

Bibliografická metadata (MODS, DC) by měla být automaticky přebírána z katalogizačního záznamu z katalogu knihovny, která je vlastníkem předlohy. Údaje na vyšší míře granularity, než kterou popisuje katalogizační záznam (např. číslo periodika), musejí být vytvářeny manuálně pracovníkem v digitalizaci. Při popisu periodik by měl mít pracovník k dispozici kompletní celý ročník fyzické předlohy.

Převod katalogizačního záznamu do formátu MODS zajišťuje série transformačních šablon. Tyto šablony využívají jazyk xsd a pomocí série příkazů a podmínek jsou schopny z jazyka katalogizačních záznamů MARC 21 vytvořit záznam v MODS.

[80] https://www.loc.gov/standards/alto/

Nejprve se katalogizační záznam převede z MARC 21 do MARCXML, kdy již dostane podobu xml jazyka, ovšem se zachovanou informací o polích a indikátorech, které slouží k další transformaci z MARCXML do MODS.

Dále by hlavní METS dokument měl obsahovat fyzickou a logickou mapu záznamu. Fyzická mapa slouží k identifikaci každé jednotlivé naskenované strany v souborech master copy, user copy, alto, ocr, txt a vedlejší METS soubor amd_sec. Logická mapa naproti tomu mapuje hierarchické pořadí jednotlivých popisovaných úrovní dokumentu pomocí zanořených a slouží jako podpora správného zobrazování digitalizátu v aplikacích pro čtení.

4.4.2 Získávání technických metadat

Technická metadata by měla být v maximální možné míře převzata z metadatových extraktorů, které tyto údaje získávají přímo ze souborů. Postup získávání metadat by tedy primárně neměl být založen na údajích, které se přednastaví do digitalizačního systému (např. název skeneru nebo skenovací aplikace pro jednu linku) a systém je pak jen automaticky přiděluje do metadat všech dokumentů dané linky. Pokud jsou tyto informace zaznamenávány do metadat z přednastavených hodnot uložených v digitalizačním systému, hrozí vždy lidská chyba (např. při výměně skeneru nebo aktualizaci skenovací aplikace se neprovede změna přednastavených hodnot). Dále platí, že pro získávání informací o formátu nelze užít prostý opis koncovky souboru.

V případě technických metadat, která se plní o původních snímcích (ve formátu TIFF nebo RAW) je vhodné využívat EXIF metadata, která jsou obsažena přímo v souboru. EXIF metadata jsou relevantní pouze pro popis formátu původního snímku (tedy TIFF nebo RAW) a některých dalších vlastností původních snímků, nikoliv pro další generace dat. Rovněž se doporučuje užít EXIF metadata pro informace o snímacím zařízení a způsobu snímání. Iniciativa FADGI vydala doporučení, jaké minimální elementy EXIF metadat ve formátu TIFF je vhodné zaznamenat do metadat (Embedded metadata working group-Smithsonian institution, 2010).

Pro převod ze schématu metadatového extraktoru (např. JHOVE) do technických metadat v MIX a PREMIS je potřeba předem ověřit možnosti namapování, a poté toto namapování nastavit v digitalizačním systému. Pro různé metadatové extraktory, včetně různých verzí téhož nástroje, se může namapování lišit. V některých případech nemusí být zcela jednoznačné.

Doporučené minimální metadatové extraktory jsou JHOVE a jpylyzer. Příloha této metodiky obsahuje doporučené namapování pro aktuální verze nástroje jpylyzer a JHOVE.

4.4.3 Nástroje pro formátovou identifikaci

Jako základní formátové identifikační nástroje musí být užity nástroje DROID[81] a JHOVE.

Nástroj DROID čerpá informace z formátového registru PRONOM. Tento registr obsahuje identifikátor PUID, který je klíčový pro jednoznačnou identifikaci formátu (Brown, 2006, s. 4). Název formátu není pro identifikaci formátu dostatečný. PUID je „rozšiřitelné schéma pro poskytování perzistentních, jedinečných a jednoznačných identifikátorů pro jednotky interpretačních informací zaznamenané v registru PRONOM.“ (Brown, 2006, s. 4). Formát je tedy pouze jedním z typů interpretačních informací, o nichž registr vede údaje, nicméně nejrozšířenějším. Funkce identifikátoru PUID jsou dvě: propojení se záznamem jednotky interpretačních informací v registru PRONOM (tj. způsob identifikace záznamu, přičemž tento záznam by ideálně měl obsahovat co nepodrobnější informace o formátu nebo jiné jednotce interpretačních informací) a jedinečný perzistentní identifikátor, který odlišuje v maximální možné míře jeden formát od druhého (odlišuje se nejen typ formátu, ale i verze). Například PUID pro JPEG verze 1.00 je „fmt/42“, pro verzi 1.01 „fmt/43“ a pro verzi 1.02 „fmt/44“. Registr MIME,[82] který je nejužívanějším obecným registrem formátů (sloužícím i pro účely mimo kontext digitální archivace), odlišuje formáty jen na základě typu a názvu, například formát JPEG všech verzí má označení „image/jpeg“.

Nástroj DROID využívá ke své činnosti metadatové soubory obsahující záznamy z registru PRONOM, které se nazývají „signature files“. Z důvodu průběžné aktualizace registru PRONOM se pro účely identifikace doporučuje použít vždy nejnovější verzi nástroje DROID a nejnovější verzi „signature files“.

Standard NDK doporučuje využít formátový identifikační nástroj, který pracuje s registrem PRONOM, a tedy dokáže souboru přidělit identifikátoru PUID. Tomu pak musí odpovídat záznam v metadatech PREMIS. Aktuální doporučovaný nástroj je uveden v aktuální příloze.

[81] Aktuálně je možné využít i nástroj FIDO (http://openpreservation.org/technology/products/fido/), jenž je též založen na registru PRONOM. Preference nástroje DROID v této metodice se odvíjí od jeho delší existence a faktu, že je vyvíjen stejnou institucí jako registr PRONOM. Nicméně i nástroj FIDO je vyvíjen důvěryhodnou institucí, jež se zabývá digitální archivací a spravuje i jiné nástroje pro digitální archivaci.

[82] http://www.iana.org/assignments/media-types/media-types.xhtml

Níže jsou uvedeny doporučené způsoby získávání a záznamu informací o formátové identifikaci.

Elementy PREMIS „formatName“ a „formatVersion“ by měly být vyplněny užitím nástroje JHOVE.

Elementy PREMIS v metadatovém kontejneru „formatRegistry“ by měly být vyplněny následujícím způsobem:

- element „formatRegistryName“ by měl vždy obsahovat hodnotu „PRONOM“;
- element „formatRegistryKey“ by měl obsahovat identifikátor PUID, který by měl být získán nástrojem DROID.

4.4.4 Propojování událostí s objektem a agentem

Pro propojování událostí s objektem je třeba důsledně v souladu se standardem PREMIS odlišovat vztah odvození od ostatních vztahů. Standard PREMIS předepisuje následující logiku metadatového zápisu pro změny objektu: „metadata, soubory, bitové toky a reprezentace …. se popisují jako statické množiny bitů. Není možné změnit soubor (nebo bitový tok nebo reprezentaci); lze pouze vytvořit nový soubor (nebo bitový tok nebo reprezentaci), který se vztahuje k zdrojovému objektu.“ (Premis, 2015, s. 22). Tento vztah mezi novým a předchozím objektem definuje jako vztah odvození (*derivation relationship*). Standard odlišuje dva typy odvození ze zdrojového digitálního objektu do nového objektu: replikace (*replication*) a transformace (*transformation*) (Premis, 2015, s. 19). Replikace znamená vytvoření digitální kopie, která je bitově identická se zdrojovým digitálním objektem (Premis, 2015, s. 272), transformace má za výsledek vytvoření jednoho nebo více digitálních objektů, které nejsou bitově identické se zdrojovým objektem (Premis, 2015, s. 273)

Událost, která je odvozením, musí být s dotčenými objekty (zdrojový a výsledný objekt) propojena na základě elementu „relatedEventIdentifierValue“.

Jiné události (tj. ty, které nejsou odvozením), musejí být propojeny na základě elementu „linkingEventIdentifierValue“.

Jako agenti musejí být vždy uvedeny všechny nezávisle existující nástroje, které jsou bezprostředním původcem události. V praxi bývá někdy při digitalizaci užit komplexní digitalizační systém, který řídí operace a integruje různé jiné nástroje pro dílčí operace. Pokud je pro událost užit nástroj, který existuje nezávisle na tomto systému, je dobrou praxí tento

nezávislý nástroj uvést jako samostatného agenta, tj. činitele události. Například nástroje pro formátovou identifikaci nebo validaci, které reálně vykonávají událost, digitalizační systém je pouze využívá.

4.4.5 Záznam událostí a nástrojů

Kongresová knihovna udržuje řízený slovník pro PREMIS, který obsahuje zejména typy událostí.[83] Doporučení níže uvádějí jak události povinné z hlediska standardů NDK[i], tak doporučené nad rámec standardů NDK. Objektem událostí, uvedených v tomto oddíle, je vždy soubor (úroveň „file“ v modelu PREMIS): k jednomu souboru se váže jedna nebo více událostí.

4.4.5.1 Snímání (skenování, fotografování)

Snímáním se rozumí vlastní proces skenování nebo fotografování, jehož bezprostředním výstupem jsou původní snímky. Tato událost se vztahuje k objektům, jimiž jsou soubory původních snímků (TIFF nebo RAW).

Podle standardů NDK se zapíše typ události (PREMIS: <eventType>) s hodnotou „capture“ a detail události (PREMIS: <eventDetail>) s hodnotou „digitization“; snímací zařízení (MIX: <captureDevice>) se v případě skeneru zapíše s hodnotou buď „reflection print scanner“ (nejčastěji používaný typ skeneru), nebo „transmission scanner“, v případě fotografování s hodnotou „digital still camera“. Propojení objektu s událostí se provede elementy metadatového kontejneru PREMIS „<relatedEventIdentification>“.

4.4.5.2 Formátová konverze z TIFF do JP2 archivní kopie

Tato událost se váže k archivním kopiím (tj. souborům ve formátu JP2 v bezeztrátové kompresi, které byly vytvořeny ze souborů předchozí generace ve formátu TIFF).

Podle standardů NDK se zapíše typ události (PREMIS: <eventType>) s hodnotou „migration“ a detail události (PREMIS: <eventDetail>) s hodnotou „MC_creation“; kodek, který byl užit k vytvoření souborů ve formátu JP2 (PREMIS: <creatingApplicationName>) se zapíše celým názvem (doporučené hodnoty pro nejčastěji užívané aplikace jsou „Kakadu“ a „OpenJPEG“) a zvlášť se zapíše i verze kodeku (PREMIS: <creatingApplicationVersion>). Údaje o kodeku se zapíší i do části PREMIS Agent; název agenta (PREMIS: <agentName>) se zapíše sloučením elementů PREMIS „<creatingApplicationName>“ a „<creatingApplicationVersion>“ (mezi

[83] http://id.loc.gov/vocabulary/preservation/eventType.html

nimi musí být mezera); typ agenta (PREMIS: <agentType>) se zapíše s hodnotou „software“; poznámka o agentovi (PREMIS: <agentNote>) by měla začínat hodnotou „command line: “ (za dvojtečkou je mezera) a následovat bude konkrétní příkazový řádek užitý v daném kodeku. Propojení objektu s událostí se provede elementy metadatového kontejneru PREMIS „<relatedEventIdentification>“.

4.4.5.3 Vytvoření ALTO XML z OCR

Tato událost se váže k souborům OCR komponenty, které vytvořil software pro optické rozpoznávání znaků z obrazových dat a které jsou ve formátu ALTO XML. Podle standardů NDK se zapíše typ události (PREMIS: <eventType>) s hodnotou „capture“ a detail události (PREMIS: <eventDetail>) s hodnotou „XML_creation“; software, který byl užit k vytvoření souborů OCR komponenty se zapíše celým názvem (ALTO: <processingSoftware>:<softwareName>), zapíše se i verze softwaru (ALTO: <processingSoftware>:<softwareVersion>).

Nad rámec standardů NDK doporučujeme zapsat název agenta (PREMIS: <agentName>) sloučením elementů ALTO „<processingSoftware>:<softwareName>“ a „<processingSoftware>:<softwareName>“ (mezi nimi musí být mezera); typ agenta (PREMIS: <agentType>) zapsat s hodnotou „software“. Propojení objektu s událostí se provede elementy metadatového kontejneru PREMIS „<relatedEventIdentification>“.

4.4.5.4 Formátová identifikace

Formátová identifikace je jeden z klíčových procesů digitální archivace. Doporučujeme proto nad rámec povinností stanovených standardy NDK zapsat do metadat událost formátové identifikace, a to jako opakovanou událost, při níž byly užity nejméně dva nástroje (JHOVE a DROID, viz 2.4.3). Tato událost se váže k souborům původních snímků (TIFF nebo RAW) a souborům archivních kopií (bezeztrátový JP2).

Jako typ události (PREMIS: <eventType>) se zapíše hodnota „format identification“. Pro užití nástroje JHOVE se zapíše název agenta (PREMIS: <agentName>) s hodnotou obsahující „JHOVE“ + číslo verze, například „JHOVE v1.20“. Pro užití nástroje DROID se zapíše název agenta (PREMIS: <agentName>) s hodnotou obsahující „DROID: version“ + číslo verze + verze souboru signature files, například „DROID: version: 6.4, Signature files: 1. Type:

Container Version: 20171130 File name: container-signature-20171130.xml 2. Type: Binary Version: 93 File name: DROID_SignatureFile_V93.xml“

Propojení objektu s událostí se provede elementem PREMIS „<linkingEventIdentifier>“.

4.4.5.5 Formátová validace

Formátová validace je další z klíčových procesů digitální archivace. Doporučujeme proto nad rámec povinností stanovených standardy NDK zapsat do metadat událost formátové validace. Tato událost se váže k souborům původních snímků (TIFF nebo RAW), kdy doporučujeme užít nástroj JHOVE jako formátový validátor, a souborům archivních kopií (bezeztrátový JP2), kdy doporučujeme užít nástroje JHOVE a jpylyzer jako formátové validátory (tato událost bude tedy opakovatelná).

Jako typ události (PREMIS: <eventType>) se zapíše hodnota „validation“ a jako detail události (PREMIS: <eventDetail>) hodnota „format validation“.

Pro užití nástroje JHOVE se zapíše název agenta (PREMIS: <agentName>) s hodnotou obsahující „JHOVE“ + číslo verze, například „JHOVE v1.20“. Pro užití nástroje jpylyzer se zapíše název agenta (PREMIS: <agentName>) s hodnotou obsahující „jpylyzer“ + číslo verze, například „jpylyzer v1.18.0“. Dále doporučujeme vyplnit element (PREMIS:<eventOutcomeInformation>) hodnotou obsahující textové výstupy validátorů JHOVE a jpylyzer popisující výsledek formátové validace. V případě aktuální verze nástroje JHOVE doporučujeme převzít hodnotu obsaženou v elementu schématu JhoveView „<status>“ (např. „Well-Formed and valid“). V případě aktuální verze nástroje jpylyzer, pokud jeho metadatové schéma obsahuje v elementu „<isValidJP2>“ hodnotu „True“, pak se zapíše hodnota „valid“, pokud jpylyzer vypíše hodnotu „False“ pak by se v metadatech měla objevit hodnota „not valid“. Propojení objektu s událostí se provede elementem PREMIS „<linkingEventIdentifier>“.

5 Kontrola kvality

Z hlediska kontroly kvality je pro oblast dlouhodobého uchovávání klíčová validace následujících čtyř oblastí:

- Validace metadat;

- Formátová validace;
- Datová validace;
- Validace balíčku SIP.

5.1 Digitální otisk

Podle standardů NDK je nutný digitální otisk (konkrétně MD5) pro soubory v konečném balíčku SIP. Digitální otisky pro jednotlivé soubory (vyjma souboru info.xml a souboru obsahujícího MD5) jsou uloženy v metadatech, pro celý balíček SIP (vyjma souboru info.xml a souboru obsahujícího MD5) pak rovněž v podobě samostatnému souboru MD5 v kořenovém adresáři balíčku SIP).

Postup, který musí být dodržen pro vytvoření digitálních otisků souborů balíčku SIP v souladu se standardy NDK, je následující. Musejí již existovat soubory v následujících podadresářích: adresář se soubory OCR komponenty v ALTO XML („ALTO“), adresář se soubory archivních kopií („masterCopy“), adresář se soubory uživatelských kopií („userCopy“) a adresář se soubory obsahujícími prostý text („TXT“). Digitální otisk k souborům v těchto podadresářích musí být zapsán do vedlejších metadatových souborů (podadresář „amdSec“). Hodnoty těchto digitálních otisků musejí být následně zapsány také do hlavního souboru v METS XML v kořenovém adresáři balíčků a k nim musejí být přidány digitální otisky vedlejších metadatových souborů. Teprve poté může být vytvořen soubor MD5 v kořenovém adresáři balíčku SIP (tento soubor obsahuje i digitální otisk hlavního souboru v METS XML). Nakonec musí být vytvořen soubor info.xml v kořenovém adresáři, který obsahuje výčet souborů včetně samostatného souboru MD5. Soubor MD5 v kořenovém adresáři tedy logicky nemůže obsahovat digitální otisk k samostatnému souboru MD5, ani k info.xml.

Balíčky SIP jsou však vytvářeny v několika fázích, ve kterých vzniká několik odlišných generací obrazových dat (minimálně tři základní generace původní snímky v TIFF / RAW, ořezané snímky v TIFF, archivní a uživatelské kopie v JP2). Tyto generace by měly být podle doporučení 2.3.1 této metodiky zachovány do doby úspěšného dodání balíčku SIP do repozitáře.

Pro bezproblémový proces vytváření souborů (všech typů) doporučujeme, aby bezprostředně po každém uložení nového souboru byl vytvořen digitální otisk (a to i k předchozím generacím souborů, které nejsou obsaženy v balíčku SIP). Tyto digitální otisky by měly být dočasně uchovávány v digitalizačním systému a při finalizaci balíčku SIP zapsány do metadat a

samostatného souboru MD5 podle postupu uvedeného v předchozím odstavci. Důvod, proč je důležité, aby byl digitální otisk vytvořen bezprostředně po vytvoření souboru, je ten, že později vytvořený digitální otisk již může být otiskem porušených souborů (v době mezi vytvořením souboru a finalizací balíčku SIP může z různých důvodů dojít k porušení souboru, například vadou datového nosiče dočasného digitalizačního úložiště), a tedy by mohl pozbýt své kontrolní funkce. Po vytvoření digitálního otisku by také měla být okamžitě spuštěna kontrola neporušenosti (jako kontrola toho, že samotný digitální otisk byl vytvořen správně). Dále je vhodné spustit kontrolu neporušenosti vždy před tím, než vznikne nová generace souboru (jako ověření toho, že zdrojový soubor není poškozen). To platí zejména pro formátovou migraci z TIFF do JP2.

Do elementu PREMIS <messageDigestOriginator> doporučujeme zapsat vždy software, který MD5 vytvořil.

5.2 Validace metadat

Validace metadat je kontrola souladu metadat s předepsanou podobou metadatového profilu v balíčku SIP. Metadatový profil je soubor metadatových elementů, které jsou vybrány z jednoho nebo více metadatových standardů a jsou spojeny do jednoho sloučeného schématu, který je uzpůsoben na míru funkčním požadavkům konkrétního užití, zatímco je zachována interoperabilita s původními standardy (Duval, 2002). Standardy NDK jsou metadatové profily, které jsou vytvořeny z mezinárodních metadatových standardů užívaných knihovnami i jinými institucemi pro potřeby digitální archivace (základem jsou mezinárodní standardy METS a PREMIS, k nimž se podle typu dokumentu / dat pojí mezinárodní standardy pro popisná a technická metadata). Podobně je tomu s metadatovými profily, které vydávají jiné národní knihovny i další instituce. Validace metadat v případě metadatového profilu zahrnuje jednak kontrolu užitím oficiálních validačních schémat mezinárodních standardů (zaznamenaných v XSD), jednak kontrolu podle profilu. Validaci metadat lze rozdělit na dva typy. Základní validace metadat je kontrola přítomnosti elementů, strukturálních vztahů mezi nimi a dodržení obecného omezení pro hodnoty elementů (např. pouze znakový řetězec). Rozšířená validace je specifičtější sémantická kontrola (např. konkrétní počet znaků určitého typu).

Pro účely validace metadat balíčků SIP vytvořených podle standardů NDK lze využít nástroj Komplexní validátor, který vyvíjí Národní knihovna ČR. Tento nástroj však nepokrývá všechny

historické verze standardů NDK a v případě vydání nových standardů NDK bude zřejmě vždy existovat určitá prodleva, než bude Komplexní validátor aktualizován o možnost validace podle nejnovějších standardů NDK. Z tohoto důvodu lze doporučit využívat jen takové verze standardů NDK, které dokáže validovat Komplexní validátor, a pokud to není možné, pak alespoň provádět validaci mezinárodních metadatových standardů, obsažených ve standardech NDK, dle jejich oficiálních validačních schémat (XSD). V případě vytváření balíčků SIP, které se neřídí standardy NDK, je doporučeno užívat vždy alespoň tuto validaci podle oficiálních validačních schémat mezinárodních standardů.

Rovněž by měla být provedena manuální validace metadat na úrovni popisu částí základní intelektuální entity (desky, přídeští, titulní strana apod.) a dodržení původní posloupnosti stran, optimálně s tištěnou předlohou v ruce.

5.3 Formátová validace

Formátová validace je kontrola, zda je souborový formát vytvořen dle požadavků daných jeho oficiální dokumentací, případně dalších požadavků (např. nastavení formátového profilu, tedy nastavení parametrů v rámci formátu).

Formátová validace by měla být provedena bezprostředně po vytvoření nové generace obrazových dat. Doporučené nástroje pro formátovou validaci jsou: pro formáty TIFF a JP2 nástroj JHOVE vždy nejnovější verze a pro formát JP2 dále ještě nástroj jpylyzer. Validace formátu TIFF (tedy souborů předchozích generací obrazových dat), které nebudou dlouhodobě uchovávány, je důležitá pro to, aby bylo ověřeno, že pro následnou formátovou migraci do formátu JP2 byly užity zdrojové soubory, které byly vytvořeny korektně (tj. ve validním formátu TIFF).

Pro kontrolu formátového profilu JP2 (lokální nastavení formátu v rámci možností oficiální specifikace, která umožňuje volbu různých parametrů) doporučujeme využívat nástroj Komplexní validátor. Případně je možné, aby instituce využila jiný nástroj, který dokáže využívat nástroj jpylyzer a výstupy srovnávat s oficiálním profilem JP2 doporučeným ve stadardech NDK pro obrazová data.

5.4 Datová (obrazová) validace

Datová validace je validace datových prvků souboru, která je nad rámec formátové validace. Datová validace, tak jak ji chápeme v této metodice, ověřuje, zda je soubor neporušený a zda

jej lze v odpovídající aplikaci otevřít. Tyto vlastnosti někdy formátové validátory neodhalí. Tuto validaci je možné provádět automatizovaně pomocí specializovaných nástrojů či manuálně (ověření otevřitelnosti). Pro obrazy ve formátu JP2 doporučujeme využít nástroj ImageMagick, případně Kakadu. Tyto nástroje pomáhají odhalit chyby v obrazovém datovém toku, které formátové validátory nemohou odhalit (např. otevřitelnost souboru). Optimálně pro datovou obrazovou validaci lze rovněž využít Komplexní validátor, který umožňuje oba uvedené nástroje zapojit do validačního procesu.

Rovněž by měla být provedena manuální kontrola kvality obrazů při dodatečném popisu (označení desek, přídeští apod.) digitalizačními pracovníky, případně již pracovníky skenování. Součástí této kontroly je vizuální inspekce obrazu (příp. poslech záznamu). Měly by být přehlédnuty náhledy všech obrazů, čímž se zkontroluje kompletnost a vizuální konzistence skenování. Z těchto obrazů by následně měl být vybrán vzorek a ty by měly projít důkladnější vizuální kontrolou[84].

Iniciativa FADGI doporučuje kontrolovat následující oblasti pro shodu s projektovou specifikací a pro detekci defektů:

-kontrola otevíratelnosti souboru

-kontrola vlastností souboru (komprese, barevný prostor, bitová hloubka) zda odpovídá zadání

-kontrola informací o barevném prostoru (zda jsou správné a kompletní)

-kontrola obrazu vůči analogové předloze (rozměry, rozlišení, orientace, dopad ořezů, kompletnost dokumentu)

-kontrola kvality obrazu (jas, kontrast, barevná věrnost, šum, artefakty, míra detailu apod.)

(FADGI, 2016, s. 87-89)

[84] Iniciativa FADGI doporučuje takto důkladněji zkontrolovat alespoň 10 obrazů nebo 10% z každé vzniklé dávky obrazů.

5.5 Validace balíčku SIP

Validace balíčku SIP je komplexní validace, která může zahrnovat všechny výše uvedené typy validací, a dále také kontrolu struktury balíčků (např. přítomnost předepsaných souborů a adresářů.

Základním prvkem validace je kontrola úplnosti a neporušenosti souborů před odesláním balíčku SIP do repozitáře. Tu lze jednoduše provést užitím souboru MD5 v balíčku SIP, který obsahuje digitální otisky i výčet všech souborů).

Důležité je však užít komplexní nástroj – pro standardy NDK jím může být Komplexní validátor, s omezením, které byly uvedeny výše (viz Metadatová validace). Optimálně by měl producent digitalizátů mít k dispozici také vlastní nástroj, který dokáže provést základní kontrolu struktury balíčku SIP.

Příloha – mapování výstupů metadatových extraktorů do metadat balíčků SIP

Mapování výstupů nástrojů

Doporučení pro mapování výstupů nástroje jpylyzer, JHOVE do metadat k souborům původních snímků (TIFF) a archivním kopiím v bezeztrátovém formátu JP2.

Nástroj jpylyzer

	Mapování
metadata MIX	
<fileSize>	fileInfo/fileSizeInBytes
<imageWidth>	properties/jp2HeaderBox/imageHeaderBox/width
<imageHeight>	properties/jp2HeaderBox/imageHeaderBox/height
<colorSpace>	properties/ jp2HeaderBox/colourSpecificationBox/enumCS (u properties/ jp2HeaderBox/colourSpecificationBox/meth =Enumerated) **nebo** properties/

	jp2HeaderBox/colourSpecificationBox/icc/colourSpace u properties/ jp2HeaderBox/colourSpecificationBox/meth =Restricted)
<iccProfileName>	properties/ jp2HeaderBox/colourSpecificationBox/description
<iccProfileVersion>	properties/ jp2HeaderBox/colourSpecificationBox/icc/profileVersion
<codec>	properties/contiguousCodestreamBox/com/comment
<codecVersion>	properties/contiguousCodestreamBox/com/comment
<codestreamProfile>	properties/contiguousCodestreamBox/siz/rsiz[85]
<tileWidth>	properties/ contiguousCodestreamBox/siz/xTsiz
<tileHeight>	properties/ contiguousCodestreamBox/siz/yTsiz
<qualityLayers>	properties/ contiguousCodestreamBox/cod/layers
<resolutionLevels>	properties/ contiguousCodestreamBox/cod/levels
<samplingFrequencyUnit>	properties/jp2HeaderBox/resolutionBox/ hodnota podle toho jaké rozlišení se plní, nejsnazší je nastavit in a brát rozlišení z vRescInPixelsPerInch a hRescInPixelsPerInch
<xSamplingFrequency>	Kontejnerový element, neobsahuje konkrétní hodnotu ale níže uvedené elementy, ty lze plnit z vícero elementů viz níže.
Varianta zápisu 1:	
<numerator>	properties/jp2HeaderBox/resolutionBox/captureResolutionBox/hResdInPixelsPerInch
<denominator>	Vždy 1
Varianta zápisu 2:	
<numerator>	properties/jp2HeaderBox/resolutionBox/captureResolutionBox/hRcN
<denominator	Hodnota elementu <denominator> se získá výpočtem mezi hodnotami elementů v kontejnerovém elementu

[85] Pokud hodnota zde odpovídá číslu 1 pak se jedná o Profile 0 a do metadat se vyplní P0, pokud rsiz obsahuje hodnotu 2, pak se jedná o profil 1 a do metadat se vyplní hodnota P1, pokud rsiz obsahuje hodnotu "ISO/IEC 15444-1" pak se nejedná o žádný profil a žádné omezení, jedná se o profil 2 a vyplní se hodnota P2.

	properties/jp2HeaderBox/resolutionBox/captureResolutionBox: hRcN / hRcD x 10 ^{hRcE} x 0,0254 (v palcích)
<ySamplingFrequency>	Kontejnerový element, neobsahuje konkrétní hodnotu ale níže uvedené elementy, ty lze plnit z vícero elementů viz níže.
Varianta zápisu 1:	
<numerator>	properties/jp2HeaderBox/resolutionBox/captureResolutionBox/vResdInPixelsPerInch
<denominator>	Vždy 1
Varianta zápisu 2:	
<numerator>	properties/jp2HeaderBox/resolutionBox/captureResolutionBox/vRcN
<denominator	Hodnota elementu <denominator> se získá výpočtem mezi hodnotami elementů v kontejnerovém elementu properties/jp2HeaderBox/resolutionBox/captureResolutionBox: vRcN / vRcD x 10 ^{vRcE} x 0,0254 (v palcích)
<bitsPerSampleValue>	properties/contiguousCodestreamBox/siz/ssizDepth
<bitsPerSampleUnit>	properties/ jp2HeaderBox/imageHeaderBox/bPCDepth
<samplesPerPixel>	properties/ jp2HeaderBox/imageHeaderBox/nC

Nástroj JHOVE

	Mapování
Metadata MIX	
<fileSize>	jhove/repInfo/size
<formatName>	jhove/repInfo/format
<formatVersion>	jhove/repInfo/version
<byteOrder>	mix/byteOrder
<compressionScheme>	properties/property/name=Transformation/value[86] pro formát TIFF: mix/BasicDigitalObjectInformation/compressionScheme
<imageWidth>	properties/property/name=XSize/value
<imageHeight>	properties/property/name=YSize/value
<colorSpace>	properties/property/property/name=EnumCS/value pro formát TIFF: mix/BasicImageInformation/BasicImageCharacteristics/PhotometricInterpretation/colorSpace
<tileWidth>	properties/property/name=Codestream/property/name=XTSize/value
<tileHeight>	properties/property/name=Codestream/property/name=YTSize/value
<qualityLayers>	properties/property/name=NumberOfLayers/value
<resolutionLevels>	properties/property/name=NumberDecompositionLevels/value
<scannerManufacturer>	mix/ImageCaptureMetadata/ScannerCapture/scannerManufacturer
<scannerModelName>	mix/ImageCaptureMetadata/ScannerCapture/ScannerModel/scannerModelName
<scannerModelSerialNo>	mix/ImageCaptureMetadata/ScannerCapture/ScannerModel/scannerModelName
<samplingFrequencyUnit>	property/name=NisoImageMetadata/mix/ImageAssessmentMetadata/SpatialMetrics/samplingFrequencyUnit
<scanningSoftwareName>	mix/ImageCaptureMetadata/ScannerCapture/ScanningSystemSoftware/scanningSoftwareName

[86] Value je číslo, kde 1=lossless, tj. 5-3 reversible a 0=lossy, tj. 9-7 irreversible.

<scanningSoftwareVersionNo>	mix/ImageCaptureMetadata/ScannerCapture/ScanningSystemSoftware/scanningSoftwareName číslo verze je součástí tohoto elementu
<orientation>	mix/ImageCaptureMetadata/orientation
<samplingFrequencyUnit>	mix/ImageAssessmentMetadata/SpatialMetrics/samplingFrequencyUnit
<xSamplingFrequency>	
<numerator>	Properties/property/name=VertResolution/property/name=Numerator/value **nebo** property/name=NisoImageMetadata/mix/ImageAssessmentMetadata/SpatialMetrics/xSamplingFrequency/numerator[87] Pro TIFF: mix/ImageAssessmentMetadata/SpatialMetrics/xSamplingFrequency/numerator
<denominator>	Properties/property/name=VertResolution/property/name=Denominator/value **nebo** property/name=NisoImageMetadata/mix/ImageAssessmentMetadata/SpatialMetrics/xSamplingFrequency/denominator Pro TIFF: mix/ImageAssessmentMetadata/SpatialMetrics/xSamplingFrequency/denominator
<ySamplingFrequency>	
<numerator>	Properties/property/name=HorizResolution/property/name=Numerator/value **nebo** property/name=NisoImageMetadata/mix/ImageAssessmentMetadata/SpatialMetrics/ySamplingFrequency/numerator Pro TIFF: mix/ImageAssessmentMetadata/SpatialMetrics/ySamplingFrequency/numerator

[87]Informace o rozlišení obrazu se může nacházet na dvou místech, dle toho, jaký kodek obraz konvertoval.

<denominator>	Properties/property/name=HorizResolution/property/name=Denominator/value **nebo** property/name=NisoImageMetadata/mix/ImageAssessmentMetadata/SpatialMetrics/ySamplingFrequency/denominator Pro TIFF: mix/ImageAssessmentMetadata/SpatialMetrics/ySamplingFrequency/denominator
<bitsPerSampleValue>	mix:ImageColorEncoding/mix:BitsPerSample/mix:bitsPerSampleValue
<bitsPerSampleUnit>	mix:ImageColorEncoding/mix:BitsPerSample/mix:bitsPerSampleUnit
<samplesPerPixel>	mix:ImageColorEncoding/mix:samplesPerPixel

Je možné využít i nástroje další. Například nástroj FITS agreguje několik nástrojů, mezi nimi i nástroje JHOVE a jpylyzer, stejně tak nástroj Kost-Val. Průběžně vznikají nové nástroje a inovují se existující. Výše uvedené nástroje JHOVE a jpylyzer jsou komunitou okolo digitální archivace hojně používané a průběžně aktualizované (zvl. JHOVE), považujeme je tedy za nástroje spolehlivé.

O Autorech

Hlavním autorem této metodiky je Ladislav Cubr za spolupráce s Natalií Ostrákovou a Pavlínou Kočišovou.

Citovaná literatura

ALTO Principles. *ALTO: Technical Metadata for Layout and Text Objects (Standards, Library of Congress)* [online]. Washington (DC): Library of Congress, June 8, 2016 [cit. 2017-03-13]. Dostupné z: https://www.loc.gov/standards/alto/description.html.

ANSI/NISO Z39.87-2006. Data Dictionary – Technical Metadata for Digital Still Images. Bethesda (MD): NISO Press, 2006, xiv, 107 s. ISBN 978-1-937522-37-7. ISSN 1041-5653.

BROWN, Adrian. *The PRONOM PUID Scheme: a scheme of persistent unique identifiers for representation information* [online]. London: National Archives, 27 July 2006, 9 s. [cit. 2017-03-21]. Digital Preservation Technical Paper, issue 2. Dostupné z: https://webarchive.nationalarchives.gov.uk/+/http://www.nationalarchives.gov.uk/aboutapps/pronom/pdf/pronom_unique_identifier_scheme.pdf

BRUYS, Alix, Bertrand CARON, Yannick GRANDCOLAS a Thomas LEDOUX. JPEG Got the Blues: Properly Rendering 32-bits JPEG. *Open Preservation Foundation* [online]. © The Open Preservation Foundation, 7th Nov 2019 [cit. 2019-11-08]. Dostupné z: https://openpreservation.org/blog/2019/11/07/jpeg-got-the-blues/

BRYGFJELD, Svein Arne. JP2K for preservation and access, experiences from the National Library of Norway. In: *JPEG 2000 for the Practitioner* [online]. Glasgow: Digital Preservation Coalition, 16 Nov 2010 [cit. 2017-03-15]. Dostupné z: http://www.dpconline.org/docman/miscellaneous/events/521-jp2knov2010brygfjeld/file.

BUCKLEY, Robert. *JPEG 2000 - a Practical Digital Preservation Standard?* [online]. Glasgow: Digital Preservation Coalition, February 2008, 21 s. [cit. 2017-03-31]. DPC Technology Watch Series Report, 08-01. Dostupné z: http://www.dpconline.org/docman/technology-watchreports/87-jpeg-2000-a-practical-digital-preservation-standard/file.

BUCKLEY, Robert. *JPEG 2000 as a Preservation and Access Format for the Wellcome Trust Digital Library* [online]. London: King's College London, Aug 2009, 17 s. [cit. 2017-03-15]. Dostupné z: http://wellcomelibrary.org/content/documents/22082/JPEG2000-preservationformat.pdf.

BUONORA, Paolo a Franco LIBERATI. A Format for Digital Preservation of Images. *D-Lib Magazine* [online]. 2008, **14**(7/8), - [cit. 2017-03-31]. DOI: 10.1045/july2008-buonora. Dostupné z: http://www.dlib.org/dlib/july08/buonora/07buonora.html.

CUBR, Ladislav. *Dlouhodobá ochrana digitálních dokumentů*. Praha: Národní knihovna ČR, 2010, 154 s. ISBN 978-80-7050-588-5.

CUBR, Ladislav, Iveta LODROVÁ, Martin ŘEHÁNEK a Zdeněk VAŠEK. Srovnání vybraných národních identifikačních systémů užívajících identifikátory URN:NBN. *ProInflow: časopis pro informační vědy* [online]. Brno: Masarykova univerzita, Filozofická fakulta, 2016, **8**(1) [cit. 2018-11-15]. DOI: https://doi.org/10.5817/ProIn2016-1-3. ISSN 1804–2406. Dostupné z: http://www.phil.muni.cz/journals/index.php/proinflow/article/view/2016-1-3

ČESKO, Ministerstvo kultury, Odbor umění, literatury a knihoven. Veřejné informační služby knihoven (VISK): podprogram č. 7: národní program ochrany a digitalizace dokumentů ohrožených degradací kyselého papíru - KRAMERIUS. In: *Veřejné informační služby knihoven* [online]. Praha: Národní knihovna ČR, 16. 9. 2015, 8 s. [cit. 2017-03-20]. Dostupné z: http://visk.nkp.cz/dokumenty/visk7/2016/VISK7-podm2016.doc

DUPLOY, Laurent. JPEG 2000 as a preservation format for digitization: lessons learned from a library. In: *Archiving2017: Final Program and Proceedings : May 15–18, 2017, Riga, Latvia.* Riga: Society for Imaging Science and Technology, 2017, s. 157–159. ISBN 978-0-89208-326-8.

DUVAL, Erik, Wayne HODGINS, Stuart SUTTON a Stuart L. WEIBEL. Metadata Principles and Practicalities. *D-Lib Magazine* [online]. 2002, **8**(4) [cit. 2017-03-13]. DOI: 10.1045/april2002-weibel. Dostupné z: http://www.dlib.org/dlib/april02/weibel/04weibel.html

DVOŘÁK, Tomáš, KOUCKÝ, Karel, ŠULC, Jaroslav a kol. *Metodika pro vytváření bezpečnostních kopií v digitální podobě* [online]. Verze 1.0. Praha: Národní archiv, Státní oblastní archiv v Praze, 2015 [cit. 2018-08-28]. Dostupné z: https://www.nacr.cz/wp-content/uploads/2019/05/metodika2015.pdf

EMBEDDED METADATA WORKING GROUP – SMITHSONIAN INSTITUTION. *Basic Guidelines for Minimal Descriptive Embedded Metadata in Digital Images. Federal agencies digital guidelines initiative* [online]. April 2010 [cit. 2018-11-15]. Dostupné z: http://www.digitizationguidelines.gov/guidelines/GuidelinesEmbeddedMetadata.pdf

FEDERAL AGENCIES DIGITIZATION INITIATIVE. *Color space - Glossary. Federal Agencies Digital Guidelines Initiative* [online]. Washington (DC):FADGI, 2017 [cit. 2017-03-16]. Dostupné z: http://www.digitizationguidelines.gov/term.php?term=colorspace

FEDERAL AGENCIES DIGITIZATION INITIATIVE, Still Image Working Group. *Raster Still Images for Digitization: A Comparison of File Formats* [online]. Part 3. Narrative and Summary Table. Washington (DC): FADGI, Revised, Aug 29, 2014, 9 s. [cit. 2017-03-15]. Dostupné z: http://www.digitizationguidelines.gov/guidelines/FADGI_RasterFormatCompare_p3_20140829_r.pdf

FEDERAL AGENCIES DIGITIZATION INITIATIVE**,** Still Image Working Group. *Technical Guidelines for Digitizing Cultural Heritage Materials: creation of raster image files.* [online] Washington (DC): FADGI, Aug 2010, 96 s. [Citace: 18. 08 2017]. Dostupné na WWW: http://www.digitizationguidelines.gov/guidelines/FADGI_Still_Image-Tech_Guidelines_2010-08-24.pdf

FEDERAL AGENCIES DIGITIZATION INITIATIVE**.** *Technical Guidelines for Digitizing Cultural Heritage Materials: creation of raster image files: September 2016*. [online] [Citace: 18. 08 2017]. Dostupné na WWW: http://www.digitizationguidelines.gov/guidelines/FADGI_Still_Image_Tech_Guidelines_2016.pdf

FERNIE, Kate, ed. *Technical Guidelines for Digital Cultural Content Creation Programmes*. Version 2.0. [Rome]: Minerva EC, 2008, 92 s. Dostupné také z: http://www.minervaeurope.org/publications/MINERVA%20TG%202.0.pdf.

CHAPMAN, Stephen, et al. 2007. *Page Image Compression for Mass Digitization* [online]. 2007 [cit. 2018-12-02]. Preprint textu příspěvku publikovaného ve sborníku Archiving 2007 (ISBN 978-0-89208-270-4). Dostupné také z: https://www.imaging.org/site/PDFS/Reporter/Articles/2007_22/Rep22_4_Arch2007_CHAPMAN.pdf

IFLA STUDY GROUP ON THE FUNCTIONAL REQUIREMENTS FOR BIBLIOGRAPHIC RECORDS. *Functional Requirements for Bibliographic Records: final report* [online]. Haag: IFLA, September 1997, as amended and corrected through Feb 2009, v, 137 s. [cit. 2017-03-27]. Dostupné z: https://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf.

ISO 14721:2012. Space data and information transfer systems - Open archival information system (OAIS) - Reference model. 2nd ed. Geneva: ISO, 2012, 126 s.

ISO 16363:2012. Space data and information transfer systems - Audit and certification of trustworthy digital repositories. Geneva: ISO, 2012, 70 s.

JPEG 2000 profiles – examples from a range of institutions. In*: JPEG 2000 for the Practitioner* [online]. Glasgow: Digital Preservation Coalition, 16 Nov 2010 [cit. 2017-03-15]. Dostupné z: http://www.dpconline.org/docman/miscellaneous/events/529-jp2knov2010parametercomparisonchart/file

LAWRENCE, Gregory W. et al. *Risk management of digital information: a file format investigation*. Washington (DC): Council on Library and Information Resources, c2000, vi, 75 s. ISBN 18-873-3478-5. Dostupné také z: https://www.clir.org/pubs/reports/pub93/pub93.pdf

LIBRARY OF CONGRESS. *Recommended Formats Statement 2019-2020* [online]. Washington (DC): The Library of Congress, [2019] cit. 2019-05-01]. Dostupné z: http://www.loc.gov/preservation/resources/rfs/RFS%202016-2017.pdf.

LIBRARY OF CONGRESS. *Sustainability of Digital Formats: Planning for Library of Congress Collections. Digital Preservation (Library of Congress)* [online]. Washington (DC): Library of Congress, Last updated 07/24/2013 [cit. 2017-03-16]. Dostupné z: http://www.digitalpreservation.gov/formats/index.shtml.

LIBRARY OF CONGRESS, Office of Strategic Initiatives. *JPEG 2000 Profile for the National Digital Newspaper Program* [online]. Washington (DC): The Library of Congress, April 27, 2006, 24 s. [cit. 2017-02-26]. Dostupné z: http://www.loc.gov/preservation/resources/rfs/RFS%202016-2017.pdf

METS: An Overview & Tutorial. Metadata Encoding and Transmission Standard (METS) [online]. Washington (DC): Library of Congress, February 9, 2016 [cit. 2017-03-13]. Dostupné z: http://www.loc.gov/standards/mets/METSOverview.v2.html.

NÁRODNÍ KNIHOVNA ČR. *Standardy pro metadata. Národní digitální knihovna* [online]. Praha: Národní knihovna ČR, 04.03. 2016 [cit. 2016-07-18]. Dostupné z: http://www.ndk.cz/standardy-digitalizace/metadata

NEISS, Bengt. *file format for still images* [elektronická pošta]. Message to: natalie.ostrakova@nkp.cz. 9. listopadu 2017 [cit. 2017-11-10]. Osobní komunikace.

NIELSEN, Anders Bo a Alex THIRIFAYS. Cost Aspects of Ingest and Normalization. In: *BORBINHA, José et al., ed. iPRES 2011: 8th International Conference on Preservation of Digital Objects* [online]. Singapore: National Library Board, 2011, s. 107-115 [cit. 2017-03-14]. ISBN 978-981-07-0441-4. Dostupné z: https://phaidra.univie.ac.at/detail_object/o:294293

NISO FRAMEWORK WORKING GROUP. *A framework of guidance for building good digital collections: a NISO recommended practice* [online]. 3rd ed. Baltimore (MD): National Information Standards Organization (NISO), 2007, iii, 95 s. [cit. 2017-03-15]. ISBN 978-1-880124-74-1. Dostupné z: http://www.niso.org/publications/rp/framework3.pdf

PREMIS EDITORIAL COMMITTEE. *PREMIS Data Dictionary for Preservation Metadata* [online]. Version 3.0. Washington (DC): Library of Congress, June 2015, rev. Nov 2015, viii, 273 s. [cit. 2016-10-14]. Dostupné z: http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf

RIMKUS, Kyle, Thomas PADILLA, Tracy POPP a Greer MARTIN. Digital Preservation File Format Policies of ARL Member Libraries: An Analysis. *D-Lib Magazine* [online]. 2014, **20**(3/4) [cit. 2017-03-13]. DOI: 10.1045/march2014-rimkus. Dostupné z: http://www.dlib.org/dlib/march14/rimkus/03rimkus.html

SMITH, Neil. Digitising Documents for Public Access. In: *MACDONALD, Lindsay, ed. Digital heritage: applying digital imaging to cultural heritage*. Amsterdam: Elsevier, 2006, s. 3-32. ISBN 0-75-066183-6.

THE ASSOCIATION FOR LIBRARY COLLECTIONS AND TECHNICAL SERVICES, Preservation & Reformatting Section. Minimum Digitization Capture Recommendations. In: *Association for Library Collections & Technical Services (ALCTS)* [online]. Chicago: ALA, June 2013 [cit. 2017-03-14]. Dostupné z: http://www.ala.org/alcts/resources/preserv/minimumdigitization-capture-recommendations

THE DIGITAL LIBRARY FEDERATION BENCHMARK WORKING GROUP. *Benchmark for Faithful Digital Reproductions of Monographs and Serials* [online]. Version 1. Washington (DC): Digital Library Federation, December 2002, 6 s. [cit. 2017-03-31]. Dostupné z: http://old.diglib.org/standards/bmarkfin.pdf.

THE NATIONAL ARCHIVES. *General hints and tips for digitisation for business use: September 2017* [online]. 2017 [cit. 2018-12-15]. Dostupné z: https://www.nationalarchives.gov.uk/documents/information-management/hints-tips-digitisation-for-business-use.pdf

VAN DER KNIJFF, Johan. JPEG 2000 for Long-term Preservation: JP2 as a Preservation Format.*D-Lib Magazine* [online]. 2011, **17**(5/6) [cit. 2017-03-13]. DOI: 10.1045/may2011-vanderknijff. Dostupné z: http://www.dlib.org/dlib/may11/vanderknijff/05vanderknijff.html

VRTĚLOVÁ, Lucie. *Analýza nastavení formátu JPEG 2000*. Brno, 2017. Bakalářská práce. Vysoké učení technické v Brně, Fakulta informačních technologií. Vedoucí práce Bařina David

VYCHODIL, Bedřich. *JPEG2000 - Aneb nemyslete si, že vás mine!.* Knihovna [online]. 2010, 21(2), s. 53-68. Dostupné z: http://oldknihovna.nkp.cz/knihovna102/10253.htm

ZENG, Marcia Lei a QIN, Jian. *Metadata*. 2nd edition. Chicago: Neal-Schuman, an imprint of the American Library Association, 2016. xxvii, 555 stran. ISBN 978-1-55570-965-5.

[i] Události v DMF

<eventType>	<eventDetail>	popis
capture	digitization	Vytvoření původního skenu
capture	XML_creation	Vytvoření souboru ALTO
capture	TXT_creation	Vytvoření prostého textu
migration	MC_creation	Vytvoření archivní kopie v JP2 z původního skenu
derivation[i]	UC_creation	Vytvoření uživatelské kopie
deletion	PS_deletion	Smazání původního skenu