Zachováno navěky? Teorie a praxe dlouhodobého uchování digitálních dokumentů

Pavlína Kočišová — Zdeněk Vašek — Václav Jiroušek — Vojtěch Kopský — Jan Bilwachs — Filip Pavčík — Petr Cajthaml

Poděkování
Publikace vznikla v rámci institucionálního výzkumu Národní knihovny České republiky, financovaného z podpory Ministerstva kultury ČR na dlouhodobý koncepční rozvoj výzkumné organizace, konkrétně z projektu DKRVO 0137.

Národní knihovna České republiky

Zachováno navěky? Teorie a praxe dlouhodobého uchování digitálních dokumentů

Pavlína Kočišová — Zdeněk Vašek — Václav Jiroušek — Vojtěch Kopský — Jan Bilwachs — Filip Pavčík — Petr Cajthaml

Praha 2023

KATALOGIZACE V KNIZE – NÁRODNÍ KNIHOVNA ČR

Kočišová, Pavlína
Zachováno navěky? : teorie a praxe dlouhodobého uchování digitálních dokumentů / autorský tým: Pavlína Kočišová, Zdeněk Vašek, Václav Jiroušek, Vojtěch Kopský, Jan Bilwachs, Filip Pavčík, Petr Cajthaml.
-- 1. vydání. -- Praha : Národní knihovna České republiky, 2023. -- 1 online zdroj
České a anglické resumé
Obsahuje bibliografii a bibliografické odkazy

ISBN 978-80-7050-791-9 (online ; pdf)

* 004.6.056.5-022.314 * 930.25:[002.1:004.738.5] * 004.08:930.25 * 002.1:004.087 * 026/027+069+930.25 * 005.591 * (437.3) * (048.8:082)
– dlouhodobá ochrana digitálních dat
– archivace elektronických zdrojů
– digitální repozitáře
– elektronické zdroje
– paměťové instituce
– programy a projekty -- Česko -- 20.-21. století
– kolektivní monografie

004 - Počítačová věda. Výpočetní technika. Informační technologie [23]

Řešitel projektu:
Bc. Václav Jiroušek

Autorský tým:
Mgr. Pavlína Kočišová; PhDr. Zdeněk Vašek, Ph.D.;
Bc. Václav Jiroušek; Vojtěch Kopský, Ph.D.; Mgr. Jan Bilwachs;
Mgr. Filip Pavčík, PhD.; Mgr. Petr Cajthaml

Recenzoval:
MgA. Michal Indrák, Ph.D.

1. vydání

ISBN 978-80-7050-791-9 (online; pdf)

Obsah

Ediční poznámka

Byť se tato publikace v teoretické části zaměřuje na obecnou problematiku dlouhodobé digitální archivace, v konečném důsledku na ni nahlíží v kontextu zkušeností z velkých paměťových institucí v České republice. Z tohoto důvodu se bude v textu často hovořit o **Národní knihovně**, a pokud není uvedeno jinak, vždy se myslí Národní knihovna České republiky.

V částech, které budou popisovat českou aplikaci zejména mezinárodních metadatových a souborových formátů, se bude hovořit o **Standardu NDK**, čímž se myslí soubor metodik *Definice metadatových formátů* pro různé typy dokumentů (textové, zvukové, born digital) a *Pravidel pro popis dokumentů* pro tyto dokumenty, vydaných Národní knihovnou České republiky.

Úvod

V běžném životě jsme stále více obklopeni digitálním světem; dalo by se téměř říci, že pohybovat se v digitálním prostředí se pro nás v posledním desetiletí stalo stejně přirozené, jako pohybovat se ve světě fyzickém. Je to zajímavý fenomén, který se nenápadně rozvíjí od devadesátých letech minulého století s tím, jak se v našich kancelářích a domácnostech začaly masově objevovat počítače a připojení k internetu, a který s přelomem tisíciletí a příchodem webu 2.0 nabral nevídanou rychlost, až úplně ovládl prostor každodennosti. V digitálním prostoru komunikujeme, tvoříme nové informace, z digitálního prostředí nové informace přijímáme a zároveň digitální prostor používáme pro jejich šíření.

Úkolem paměťových institucí dneška se přirozeně stala ochrana digitálního kulturního dědictví a digitálních dokumentů, ať už těch, které v elektronickém prostředí vznikly, nebo těch, které do něj byly převedeny. Zhruba od druhé poloviny devadesátých let 20. století se mezinárodní komunita paměťových institucí zabývá vývojem standardů a postupů, které mají za úkol zajistit dlouhodobou ochranu digitálních dokumentů v důvěryhodných archivech. Výsledkem je vznik dlouhodobých úložišť a digitálních knihoven, které mají za cíl digitální dědictví uchovat a zpřístupnit pro příští generace.

V současné době se nacházíme nedlouho po celosvětové pandemii viru COVID-19, která nejen české veřejnosti ukázala nepopiratelnou roli paměťových institucí ve fungování společnosti a podpořila důležitost úsilí o budování jejich digitálních verzí. Lze očekávat, že i na základě této zkušenosti a zároveň vytrvalé digitalizace státní správy bude trend budování digitálních repozitářů všeho druhu v následujících letech nadále růst, a ukládání v dlouhodobých úložištích se stane běžnou součástí cyklu nakládání s dokumenty.

V zahraniční odborné literatuře na téma dlouhodobé archivace a digitalizace vyšlo v posledních patnácti letech mnoho výborných publikací. V českém prostředí bylo v odborných periodicích publikováno množství menších případových studií a článků od různých autorů, komplexních publikací ovšem vzniklo nemnoho. Zásadními

díly jsou monografie Dlouhodobá ochrana digitálních dokumentů Ladislava Cubra (2010a) a dále dvojice disertačních prací Digitalizace, popis pomocí metadat a jejich formáty Jana Hutaře (2012) a Autenticita a digitální informace (2017), kterou taktéž napsal Ladislav Cubr.

Co je obsahem této publikace:
Tato kniha by měla přinést ucelený pohled na danou problematiku z perspektivy aktuálních poznatků v oboru a zároveň představit praxi z knihoven a archivů na území České republiky. V následujících kapitolách se bude věnovat základnímu rámci doporučení k budování archivů, definici digitálních objektů a jejich identifikaci, teorii dlouhodobé archivace a bitové ochrany, budování důvěryhodných úložišť pomocí nástrojů LTP a formátům k dlouhodobému uchovávání jak v oblasti souborových formátů, tak metadat.

Teoretické kapitoly jsme se snažili doplnit příklady z konkrétních aplikací v projektu Národní digitální knihovny, kterou společně od roku 2010 tvoří Národní knihovna a Moravská zemská knihovna v Brně, a pohledem z prostředí českého archivnictví, které se spolu s knihovnami v současné době nejvíce podílí na masivní digitalizaci tuzemského kulturního dědictví.

Komu může být užitečná:
Ambicí této knihy je doplnit řadu jiných, prakticky orientovaných dokumentů, které v Národní knihovně vychází jako metodický materiál pro potřeby praktické digitalizace a dlouhodobé ochrany dokumentů v projektu Vytvoření Národní digitální knihovny a dalších paměťových institucí, které svoji digitalizaci financují z národních grantových programů.

Díky své orientaci na teorii a praxi budování důvěryhodných dlouhodobých úložišť je tato kniha zejména určena odborné komunitě z paměťových institucí, pro kterou může plnit funkci referenčního bodu pro lepší porozumění metodickým materiálům. Pro kolegy z institucí, které se teprve chystají zapojit do některého digitalizačního programu, může plnit funkci vstupního materiálu k oboru dlouhodobé archivace digitalizovaných dokumentů.

Doufáme, že se tato kniha zároveň stane publikací, kterou budou pro základní orientaci v problematice používat také studenti humanitních oborů, budoucí knihovníci, archiváři, informační specialisté, ale také informatici a další, kteří se během svého studia mohou setkat s tématem dlouhodobé archivace digitálních objektů.

V neposlední řadě má tato publikace ambici přiblížit obor dlouhodobé archivace v co nejúplnějším náhledu každému, kdo se s problematikou nakládání s digitálními dokumenty chce seznámit anebo potřebuje získat základní přehled z profesních důvodů.

1 Historie péče o digitální dědictví. Úvod do problematiky

Současně s rozvojem elektronických výpočetních strojů se začaly objevovat záznamy v digitální podobě. Nejdříve se jednalo o různé datasety, statické přehledy a podobně, ale s rozvojem techniky začaly vznikat regulérní dokumenty, které by v analogové podobě skončily nakonec v archivech.

Už v desetiletích, následujících po roce 1945, si byli zodpovědní pracovníci vědomi potřeby ochrany elektronických dat. Jejich aktivity byly primárně zaměřeny na ochranu nosičů digitálních informací. Od čtyřicátých do konce osmdesátých let bylo rozšíření výpočetní techniky omezené, byť postupně narůstalo. Většinou však měla omezenou funkčnost i pokud jde o vytváření obsahu, který by byl člověku srozumitelný. Postupem času byla (často náhodnými zjištěními a objevy) definována různá rizika a hrozby, ohrožující trvanlivost a kvalitu dat. Zároveň se ale na základě praxe vytvářela doporučení, jak těmto problémům čelit.

Od osmdesátých let 20. století do začátku 21. století byly postupně vytvořeny zásady pro dlouhodobé uchovávání digitálních dat a zrodil se i nový specializovaný obor na pomezí praxe a teorie, pro který se vžilo označení *Digital Preservation*. Pojem *Digital Preservation* v současné době slouží jako zastřešující množina, do které spadají všechny aktivity, jež jsou s oborem spojeny. V rámci snah o dlouhodobé uchovávání digitálních informací se do současnosti vyčlenilo několik dalších specializovaných podoborů, z nichž nejznámějším je zřejmě digitální kurátorství *(Digital Curation)*.

Počátek aktivit spojených s dlouhodobých uchováváním není přímo asociován s jednou událostí. (BAUCOM, 2019) Dokonce lze říci, že se již stačily kolem počátků tohoto oboru lidského vědění vytvořit mýty.

Jako iniciační moment pro vznik oboru *Digital Preservation* se často uvádí problémy spojené s přečtením dat ze sčítání lidu v USA

v roce 1960. V polovině osmdesátých let 20. století již tato data totiž byla takřka nečitelná kvůli nedostupnosti hardware, který by dokázal záznamová média přečíst. Ačkoliv se tento příběh jeví jako věrohodný,[1] není pravdivý (ADAMS a BROWN, 2000). Ve skutečnosti byla celá událost přibarvena a informace zkomoleny. Je ovšem pravda, že ke ztrátám informací dostupných pouze v elektronické podobě tímto způsobem docházelo. Např. i v České republice bylo zjištěno, že některé vládní materiály z osmdesátých let 20. století byly již v devadesátých letech nečitelné kvůli nedostupnosti specifického hardware. Právě stabilita software, ale naopak variabilita hardware před koncem osmdesátých let byla pro dlouhodobé uchovávání digitálních dat typická.

Jedním ze zlomů, který vedl k intenzivnímu promýšlení strategií trvalé dostupnosti informačního obsahu, byl případ dat ze sond Viking, které NASA vyslala v sedmdesátých letech na Mars. Na počátku devadesátých let agentura NASA zjistila, že na magnetických páskách leží mimo jiné několik tisíc fotografií, které obě sondy zaslaly, ale chybí jak hardware pro jejich zobrazení, tak zdrojový kód kódování těchto dat. (DELJANIN, 2011) Zobrazovací hardware se podařilo postavit, ale náročnější bylo pomocí reverzního inženýrství rekonstruovat kód, ve kterém byly snímky uloženy. Práce trvaly více než rok, ale přinesly úspěch (BLAKESLEE, 1990). Právě toto byl jeden z momentů, kdy si ve větším rozsahu odpovědní lidé uvědomili potřebu nejen data ukládat, ale zajistit i jejich čitelnost. Podobnou zkušenost jako NASA také udělala Laboratoř tryskového pohonu (JPL), která řídila zpracování dat z výzkumu Měsíce.

Obavy o zachování informací se neozývaly pouze z technologických center spojených s masivním využíváním digitálních technologií, ale i z dalších profesí, které byly zvyklé pracovat s dokumenty. Obavy z množství informací existujících pouze v digitální podobě projevovali také historikové, archiváři a další vědečtí výzkumníci (většina podnětů žádala tisknout dané dokumenty). Koncepční podnět nakonec vzešel z knihovnického světa. V roce 1994 se Commission

1 V zásadě odpovídá reálným zkušenostem se ztrátou různých dat v elektronické podobě, kdy nejčastější příčinou nebyla softwarová nečitelnost dat, ale právě nedostupnost hardwaru.

on Preservation and Access (CPA) a Research Libraries Group (RLG) dohodly na ustanovení pracovní skupiny, která se začala věnovat otázkám zpřístupnění a uchování digitálních dokumentů v knihovnách.

Pracovní skupina vypracovala závěry, v nichž doporučila zásady, které tvoří základy dnešního uchovávání digitálních dokumentů – tedy kopírování na nová média a zároveň sledování vývoj software tak, aby byl obsah z médií stále čitelný. Roku 1996 byla vydána závěrečná zpráva tohoto úkolu, nazvaná *Preserving Digital Information* (WATERS a GARRETT, 1996), která určila dvě další doporučení pro uchovávání digitálních informací: nutnost zapojit do ochrany digitálních informací samotné tvůrce obsahu (tak, aby data vytvářeli ve standardizovaných formátech a během životního cyklu dokumentu sledovali jeho srozumitelnost) a aby vznikly digitální repozitáře, archivy, jejichž posláním bude právě uchování digitálních dat. Tyto archivy by se měly řídit jednotnou metodikou a měly by získat certifikaci. Myšlenka spolupráce paměťových institucí s tvůrci dokumentů ještě v době před vznikem samotných dat do jisté míry překračovala obvyklé paradigma, na druhou stranu v archivnictví i knihovnictví měla určité paralely i v klasickém analogovém světě.

Ke vzniku standardu pro důvěryhodné digitální archivy předpokládané ve zprávě však nakonec vedla jiná cesta, byť se s myšlenkami knihovnické skupiny zcela protnula. Pod dojmem problémů s uchováním dat z kosmického výzkumu vyzvala v roce 1990 Mezinárodní organizace pro standardizaci (ISO) k vytvoření normy pro uchovávání digitálních dat. Výzvu přijal Consultative Committee for Space Data Systems (CCSDS) a začal pracovat na odpovídající normě. Šlo samozřejmě o nezmapovanou oblast poznání a koncept řešení se hledal poměrně dlouho. Výsledkem byl Open Archival Information System (OAIS) Reference Model.

První draft tohoto funkčního modelu důvěryhodného digitálního archivu byl zveřejněn v roce 1997, model byl podroben dalšímu kolu diskusí a v roce 2002 oficiálně přijat jako ISO norma. V roce 2012 pak vyšla druhá verze tohoto textu. Tím byl položen základ pro vytváření LTP systémů (repozitářů) a zásad práce s daty v těchto systémech. Nejen, že byly definovány principy a doporučení, ale uživatelé napříč obory i jednotlivými státy získali společnou platformu, v jejímž rámci

mohli rozvíjet nový obor. Na zprávu *Preserving Digital Information* navázala v roce 2002 zpráva *Trusted Digital Repositories: Attributes and Responsibilities*, kterou iniciovala opět RLG a která doplnila OAIS. RLG navázala spolupráci s archivním světem, konkrétně National Archives and Records Administration (NARA) a z jejich společné iniciativy vzešel v roce 2007 dokument *Trustworthy Repositories Audit and Certification* (TRAC), který byl později vydán jako ISO 16363 a dovršil původní myšlenku digitálních repozitářů, které budou moci být nezávisle auditovány a certifikovány. V průběhu praxe se však ukázalo, že nároky normy se rozchází s běžnou praxí a certifikaci podle ní až dodnes získalo jen minimum institucí.

Potřeba péče o digitální dokumenty byla od poloviny devadesátých let 20. století stále zřejmější, ve většině společností došlo ke ztrátám dat nebo jejich ohrožení. Začaly se vytvářet úžeji i šířeji zakotvené odborné skupiny, jejichž cílem bylo dále rozvíjet základy dlouhodobého uchovávání digitálních dat. V roce 2006 začal projekt *Preservation and Long-Term Access Through Networked Services Project* (PLANETS), na jehož základě se utvořilo zájmové uskupení Open Preservation Foundation, ve kterém jsou zapojeni zejména evropští odborníci na digitální archivaci z knihoven, archivů, univerzit i komerčních společností. V současnosti jde asi o nejdůležitější odborné uskupení, které provádí průzkumy, pořádá vzdělávací akce a přináší informace o doporučených postupech. V USA vzniklo konsorcium National Digital Stewardship Alliance, které je více formalizované a navázané na oficiální instituce. Poslední velkou organizací, kterou je třeba vést v patrnosti, je Digital Preservation Coalition, která se uskupila v roce 2002 ve Velké Británii. Tato organizace mj. stojí za vyhlášením *World Digital Preservation Day*, který připadá na první listopadový čtvrtek.

V roce 1997 začal vycházet první odborně zaměřený časopis RLG DigiNews (vycházel do roku 2007). Od roku 2006 jej doplňuje International Journal of Digital Curation, který nyní patří k hlavním platformám v oboru. Roku 2004 proběhl první ročník konference iPres, která se od té doby stala největším fórem pro setkávání pracovníků v oblasti Digital Preservation (dalšími akcemi s velkým dopadem jsou Preservation and Archiving Special Interest Group [PASIG] a Archiving).

Na počátku 21. století také začaly vznikat metadatové standardy, které tvoří základ pro současnou strukturu záznamů o uchovávaných datech (např. standard PREMIS byl vypracován v letech 2003–2005). Zároveň byly vytvořeny některé nástroje a znalostní báze pro dlouhodobé uchovávání, které jsou podrobněji popsány v samostatné kapitole níže (PRONOM, DROID etc.).

Pokud jde o české prostředí, lze stručně shrnout, že cílené aktivity směřující k dlouhodobému uchování digitálních dokumentů začaly být vyvíjeny přibližně od poloviny prvního decennia 21. století a vedly k vzniku projektů Národní digitální knihovny a Národního digitálního archivu. O obou projektech je pojednáno níže v samostatných kapitolách. Zároveň také byly položeny základy českého odborného názvosloví, zejména v souvislosti s překlady ISO norem 14721 a 16363. Z odborné literatury, která v českém prostředí vznikla, je třeba zmínit již v úvodu citované monografie Dlouhodobá ochrana digitálních dokumentů Ladislava Cubra (2010a) a disertační práce Digitalizace, popis pomocí metadat a jejich formáty Jana Hutaře (2012) a Autenticita a digitální informace (2017), opět od Ladislava Cubra.

2 Referenční rámec OAIS

Jak bylo v předchozí kapitole naznačeno, základním rámcem pro oblast dlouhodobého uchovávání digitálních dokumentů je model OAIS (Open Archival Information System), který vznikl na výzvu Mezinárodní organizace pro standardizaci mezi lety 1997–2002. Cílem původní myšlenky Consultative Committee for Space Data Systems (CCSDS) bylo vytvořit standard, který konzistentně popíše problematiku dlouhodobého uchovávání i procesy a koncepty, které k tomuto cíli povedou. Původní referenční model byl publikován roku 1999 pro veřejné připomínkování a poté fungoval jako interní dokument CCSDS (LAVOIE, 2000).

Model se však začal šířit a postupně jej adaptovaly i jiné obory, které potřebovaly standardizovat své úsilí o vytváření digitálních repozitářů. I z toho důvodu byl model OAIS již roku 2002 uznán jako mezinárodní norma ISO 14721 (poslední aktualizace proběhla v roce 2012) a v současnosti tvoří etalon pro celý obor ochrany digitálních dat. V českém prostředí byl referenční model OAIS adaptován velmi brzy v projektu Vytvoření Národní digitální knihovny. V srpnu 2014 byl vydán jeho český překlad Úřadem pro technickou normalizaci, metrologii a státní zkušebnictví jako česká technická norma ČSN ISO 14721.

Zásadní vlastností Open Archival Information System je skutečnost, že označuje *typ archivu* a zároveň je *konceptem archivace digitálních objektů* jako takových.

2.1 OAIS jako archiv

V případě OAIS jakožto druhu archivu rozumíme pod touto zkratkou digitální archiv sestávající z organizace lidí a systémů, kteří přijali odpovědnost za ochranu, uchovávání a zpřístupňování informací pro určitou cílovou uživatelskou komunitu. OAIS je v tomto případě množinou mnoha různých podtypů archivů fungujících v různých režimech. Může se jednat o archivy v různě restriktivním režimu od tzv. „open access“ přes archivy pro omezenou komunitu uživatelů

až po archivy nepřístupné, „dark archive“ nebo proprietární (CUBR et al., 2023).

Archivní repozitář podle OAIS má dvě primární funkce: zaprvé ochranu informací a zajištění její dlouhodobé ochrany a zadruhé poskytovat k archivované informaci přístup podle potřeb určeného okruhu uživatelů a za předem určených pravidel.

Takovýto repozitář má několik hlavních úkolů:

- vyjednat a získat informace od jejich producentů;
- zajistit takovou kvalitu informací, aby bylo možné je dlouhodobě archivovat;
- vymezit a dodržovat strategii a procesy v archivu tak, aby byla informace chráněna proti potenciálním hrozbám, a to i proti těm, které jsou nepředpokládané;
- zaručit, aby uchovávané informace byly lidskému uživateli srozumitelné bez použití druhotného výkladu nebo jiné pomoci;
- vymezit hlediska využití uživatelskou komunitou;
- zpřístupnit archivované informace uživatelům.

Při budování OAIS archivu je nutné si uvědomit, že ne všechny archivy mají stejné cíle a zaměření. Při budování archivu by měla být jasná východiska, podle kterých se pak celý projekt bude utvářet (REESE a BANERJEE, 2008, s. 8). Tato východiska lze obecně shrnout do čtyř množin: účel a funkce archivu, velikost archivu, provoz a technická řešení a možnosti implementace.

A) Účel a funkce repozitáře

Předtím než vůbec začne výstavba jakéhokoliv digitálního archivu, mělo by být jasné, jak bude ukotven. Jaký bude účel (mandát) zřízení archivu, kdo archiv pověřuje sběrem dat a jaký bude jeho status, tj. zda bude digitální archiv samostatnou institucí, nebo vznikne v rámci již existující organizace (například knihovny, archivu, muzea, univerzity atd.). Kdo je garantem a původcem archivu? Bude archiv budován jako příspěvková organizace, nebo jako komerční subjekt?

Na tyto otázky pak navazují témata spojená s právní situací státu, ve kterém bude archiv zřízen, respektive, právní prostředí zasahuje

již do vzniku samotného archivu a zároveň může výrazně ovlivňovat zejména samotný chod archivu. Nejviditelnějším příkladem je získávání obsahu a akvizice dat. Před zřízením archivu musí být jasné, jaké jsou legální možnosti a omezení při akvizici a jakým způsobem bude nutné ošetřovat zisk dat od producentů. Je třeba zjistit, zda budou data do archivu ukládat dobrovolně nebo zda je k tomu zavazují právní předpisy (a v případě kterých typů dokumentů) a jaké budou smluvní podmínky pro obě strany procesu. V neposlední řadě potom platné právní předpisy mohou definovat i to, kdo smí (a za jakých podmínek) data v archivu ukládat.

B) Velikost repozitáře

Velikost repozitáře zahrnuje mnoho dílčích problémů, které ve výsledku zásadně ovlivňují velikost a celkovou nákladnost projektu. V plánovací fázi procesu vzniku archivu je důležité promyslet, jaké jsou možnosti projektu a jeho potenciál, ale zároveň se držet realistického scénáře a soustředit se na to, co je možné skutečně vytvořit (REESE a BANERJEE, 2008, s. 9).

Digitální archiv se soustřeďuje kolem dat. V prvé řadě je nutné stanovit si **množství dat**. Je nutné kvantifikovat jak celkové množství dat, které by měl archiv pojmout, tak kolik dat bude přibývat ročně (nebo v jiném časovém úseku, např. grantovém období). V jakém řádu se budou počty uložených souborů pohybovat, v tisících, v milionech? V jakém řádu datového objemu je maximální hranice, v terabytech nebo petabytech? Jaká je škálovatelnost úložiště (a fyzického prostoru pro depozitář)? Další otázky ohledně archivovaných dat potom souvisí i s provozem archivu, kterému je věnován samostatný oddíl níže.

Digitální archiv je ale kromě dat tvořen hlavně **lidskými zdroji**, zaměstnanci, kteří se podílí na jeho utváření, správě, údržbě a provozu. V projektu digitálního archivu by neměla chybět úvaha o počtu zaměstnanců a jejich specializaci. Je nutné stanovit, zda archiv budou provozovat pouze oni, či bude jeho provoz zajišťovat i někdo zvenčí (např. externí firma). V této části plánování archivu je nutné vzít v úvahu také počet uživatelů, kteří budou instituci využívat pravidelně a jejich potřeby.

C) Provoz archivu

S ohledem na provoz archivu je třeba zaměřit se zejména na komplexnost dat a jejich specializaci.

Komplexnost dat – v tomto ohledu je nutné si stanovit, jaká data bude archiv přijímat (a jaká mu budou původci poskytovat). Mohou to být buď jednoduchá (textové formáty, obrázky, video), středně komplexní (složené dokumenty s množstvím vazeb mezi jednotlivými částmi) nebo velmi komplexní data (např. software, texty s vloženými tabulkami, celé internetové weby, databáze aj.). Od komplexnosti dat se bude odvíjet mimo jiné údržba fondu a požadavky na bitovou a logickou ochranu.

Specializace dat – ta určuje, do jaké míry je k použití a interpretaci dat v archivu nutné mít expertní znalosti. Jak to bude ovlivňovat jejich dlouhodobou ochranu? Zároveň s sebou specializace dat může nést otázku **oprávnění přístupu**. Archiv si musí na začátku určit, pro koho je vlastně určen a kterým uživatelům a za jakých podmínek dá přístup ke svým fondům. Oprávnění lze škálovat podle skupiny uživatelů.

D) Technická řešení a možnosti implementace

Také technická řešení provozu archivu je nutno vzít v potaz, zejména v následujících bodech:

Zdroj metadat – metadata mohou pocházet z několika zdrojů, od producentů či automatickou extrakcí pomocí konkrétních nástrojů. Získání metadat se také může lišit podle jejich druhu (popisná, administrativní, ochranná a jiná metadata)

Interoperabilita – při návrhu archivu je potřeba počítat s interoperabilitou dat, tj., jak bude digitální archiv komunikovat s ostatními systémy a jak bude interoperabilita podporována z pohledu použitých formátů dat a metadat.

Strategie ukládání – archiv může nebo nemusí mít mít vlastní datové úložiště; teoreticky mohou být data ukládána formou nákupu služby od externího dodavatele. V tom případě je potřeba vyřešit, zda je údržba úložiště povinností provozovatele archivu, nebo dodavatele služby. Zejména v případě vlastního úložiště je nutné počítat s tím, že se hardware musí průběžně obměňovat za nové.

Správa softwaru – kromě všeho ostatního je nutné si ujasnit, jak budou pro archiv získávány, udržovány a provozovány nástroje nezbytné pro jeho chod i správu a ochranu v něm uložených dat. Jiné mohou být podmínky v případě vlastního archivu, v případě úložného řešení pomocí nákupu služby. V obou případech může toto řešení ležet jak na provozovateli, tak na externím dodavateli.

2.2 OAIS jako referenční rámec digitální archivace

Pod pojmem OAIS zároveň rozumíme referenční rámec pro archivaci a zpřístupnění digitálních dokumentů (tzv. *high-level metadata framework for digital preservation*), který do určité míry vymezuje i otázku produkce digitálních dokumentů. Dále OAIS obsahuje terminologický slovník pro oblast digitální ochrany, archivy a repozitáře, který je v současné době základní referenční pomůckou pro odbornou veřejnost v oboru digitalizace. Informační model popisuje metadata, potřebná ve všech komponentách archivního systému a na všech stupních procesu archivace pro dlouhodobé uložení digitálního objektu. Referenční rámec OAIS je nezávislý na typu digitálního objektu, na technologii dlouhodobé ochrany i na podtypu samotného archivu. Model je univerzální pro všechny organizace, které nakládají s informacemi určenými pro dlouhodobou ochranu. Zároveň ale referenční rámec OAIS neřeší vlastní procesy tvorby metadat, tvorby dat jako takových a nezabývá se ani procesy mimo vlastní archiv – k tomu slouží například DCC Curation Lifecycle model, který je zaměřen na životní cyklus digitálního objektu.

Referenční rámec OAIS je v odborné komunitě využíván různými způsoby. Kromě obecně vzdělávacích účelů v odboru digitální ochrany pro odbornou veřejnost a paměťové instituce je využíván zejména jako základní plán architektury procesů při budování archivačních systémů, jelikož popisuje obecné principy a procesy jejich fungování. Umožňuje porovnání různých strategií a technik dlouhodobé ochrany. Model OAIS také funguje pro porovnávání různých datových modelů digitálních informací a v neposlední řadě je důležitým

základem pro vývoj metadatových standardů a metodik pro dlouhodobou ochranu digitálních dat (CUBR et al., 2023).

2.2.1 Prostředí OAIS archivu

Archiv OAIS je tvořen čtyřmi navzájem provázanými entitami a jejich vztahy. První entitou jsou producenti informací *(producers)*, druhou uživatelé *(consumers)*, třetí správci *(managers)* a čtvrtou je archiv samotný.

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."

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."

Správce *(manager)* je „úloha vykonávaná těmi, kdo určují celková pravidla archivu OAIS jako součást širších pravidel, například v rámci větší organizace" (CUBR et al., 2023).

Vztahy archivu OAIS s producentem nebo koncovým uživatelem vymezují vnější interakce archivu OAIS, vztahy mezi archivem OAIS a managementem jeho vnitřní interakce. V praxi masové digitalizace pak může úlohy producenta, správce a archivu zastávat jedna instituce, v českém prostředí typicky knihovna.

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". Relace dodávání dat je „jednotlivá dodávka datového nosiče nebo telekomunikační spojení, kterými jsou archivu OAIS poskytována data" (CUBR et al., 2023).

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 podoby nastavení sklízecího robota.

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. 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. Rozdíl mezi koncovým uživatelem a cílovou komunitou spočívá v tom, že koncový uživatel je jakýkoliv subjekt, který interaguje s archivem OAIS s cílem získat informace (může tedy jít i o softwarový systém). Člen cílové komunity je takový koncový uživatel, na základě jehož znalostní základny se udržují informace tak, aby byly srozumitelné samy o sobě (CUBR et al., 2023).

2.2.2 Moduly archivu OAIS

Referenční rámec OAIS pro zajištění úkolů archivace, ochrany a zpřístupnění informací navrhuje architekturu repozitáře, která se skládá z šesti navzájem provázaných funkčních komponent.

1) **Modul Příjem** *(Ingest)* zajišťuje procesy spojené s příjmem informačního balíčku *(Submission information package)* od producenta. V tomto modulu jsou data připravována po přijetí k uložení a další správě. Dochází k vytváření potřebných metadat a následnému vytvoření archivního balíčku *(Archival information package)*.
2) **Modul Archivní sklad** *(Archival Storage)* zajišťuje bezpečné uložení archivních balíčků a umožňuje jejich správu a vyhledávání.
3) **Modul Správa dat** *(Data Management)* zajišťuje přístup k datům a jejich úpravám, obecně umožňuje správu dat, popisných informací a systémových informací, které jsou používány na podporu archivních funkcí.
4) **Modul Administrace** *(Administration)* zajišťuje správu procesů, funkcí a nastavení repozitáře a uživatelských práv.
5) **Modul Zpřístupnění** *(Access)* zajišťuje rozhraní mezi archivem a uživatelem (což může být jak administrátor archivu, tak koncový uživatel). Tento modul je zodpovědný za funkce vyhledávání,

získání a zobrazení požadovaného obsahu pro koncového uživatele repozitáře. Samotné zobrazení dat pak už probíhá mimo archiv OAIS, zpravidla přes samostatnou zobrazovací aplikaci.

6) **Modul Plánování ochrany** *(Preservation Planning)* je modulem, ve kterém je možné plánovat a testovat metody dlouhodobé ochrany dat a testování příslušných nástrojů. Obecně slouží k provádění ochranných akcí v rámci archivu (HUTAŘ, 2012, s. 69).

3 Digitální objekt a jeho podoby

Referenční rámec OAIS ve své metodice stanovil základní informační model, podle kterého dnes fungují veškeré paměťové instituce v oblasti digital preservation. Základem tohoto modelu je *informace*, jež tvoří základní stavební jednotku informačních balíčků, kterým se budeme věnovat dále v rámci této kapitoly.

3.1 Informace

Informaci lze definovat několika způsoby. Norma ISO 14721 pojem informace vymezuje jako „jakékoliv znalosti *(knowledge)*, které mohou být předmětem výměny *(exchange)*" a udává, že při výměně jsou informace „vždy vyjádřeny (tj. reprezentovány) určitým typem dat". Data jsou pak definována jako „opakovaně interpretovatelné reprezentace informací ve formalizované podobě vhodné pro komunikaci, interpretaci nebo zpracování."

Pokud bychom chtěli možná poněkud abstraktnější vyjádření, lze říci, že *informace* je interpretací skutečnosti, kterou lze přenášet v čase a prostoru, přičemž *data* zde slouží jako jednotlivé atomy, ze kterých se informace skládá. Informace může být vyjádřena jako zpráva (komunikace) nebo jako výsledek pozorování (interpretace). Pro oblast digitální ochrany je pak důležitá ta část, která mluví o *přenášení v čase a prostoru* – což v prostředí nejen paměťových institucí plní digitální archiv.

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)*.

3.2 Digitální objekt

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, které si osvojila osoba nebo systém a která této osobě nebo tomuto systému umožňuje porozumět přijímaným informacím".

Norma OAIS nabízí definici i pro digitální objekt, který nazývá „množinou bitových posloupností" (CONSULTATIVE COMMITTEE, 2002 cit. podle CUBR, 2010a). Další pohled na digitální objekt poskytuje definice organizace UNESCO, která digitální objekt ukotvuje čistě v technickém smyslu jako "data uložená ve formě počítačových souborů, pro jejichž prohlížení, poslouchání (a jiné formy percepce) je nutný počítačový software.

3.3 Intelektuální entita

Intelektuální entitu lze definovat jako jednotlivý intelektuální nebo umělecký výtvor *(creation)*, který je považován za relevantní pro cílovou komunitu v kontextu digitální archivace (CUBR, 2017, s. 70).[23] Metadatový standard PREMIS zároveň intelektuální entitu vnímá jako reprodukovaný informační obsah, tedy ne originální informační obsah na fyzickém nosiči, ale jeho digitalizovanou reprezentaci (CUBR, 2017, s. 128). PREMIS pojímá intelektuální entitu jako smysluplnou část intelektuálního obsahu (dokumentu), což se shoduje s vnímáním intelektuálního obsahu v rámci OAIS (CUBR, 2017, s. 128).

2 Tato definice vychází ze standardu PREMIS a jeho pojetí rozdělení datového modelu na čtyři základní složky: intelektuální entitu, reprezentaci, soubor a bitový tok (CUBR, 2017, s. 70).

3 Koncept intelektuální entity má mnoho různých pojetí. V této publikaci se zaměřujeme na pojetí, které intelektuální entitu ukotvuje jako část digitálního prostředí, jsou ale i definice, které ji popisují čistě v prostředí fyzickém, jako fyzický nosič informace či zdrojový dokument, ze které vzniká digitální reprezentace.

Jiným způsobem lze intelektuální entitu vyjádřit například pomocí modelu FRBR, který ji stanovuje pomocí čtyř abstraktních úrovní: díla, vyjádření, manifestace a exempláře.

Pokud bychom tato obecná vyjádření chtěli konkretizovat, můžeme si pomoci příkladem z praxe Národní digitální knihovny – jako intelektuální entitu pro jednotlivé typy dokumentů stanovuje takovou část celku, která i samostatně stojící dává uživateli smysl. Například v případě klasické monografie je tedy intelektuální entitou celý svazek, v případě periodik jednotlivé číslo.

3.4 Granularita

Granularitou se myslí relativní velikost, rozsah, úroveň detailu či hloubka průniku, která charakterizuje nějaký objekt nebo činnost, v případě digitálního prostředí konkrétně intelektuální entity. Granularita určuje, co je základní entita informačních toků a procesů, co lze jak strukturalizovat a seskupovat, jak dělit vyšší celky na nižší a v jaké podrobnosti (CUBR, 2010a).

Na rozdíl od granularity ve fyzickém prostředí, kde lze pracovat s celky jako je například svazek, trilogie, edice nebo sbírka, nabízí granularita v digitálním prostředí mnohem větší hloubku. Za samostatný dokument lze vydělit samostatnou stranu, pokud to popis intelektuální entity (zpravidla celého svazku) vyžaduje. Neexistuje žádné univerzálně platné pravidlo, každá digitální knihovna může granularitu jednotlivých intelektuálních entit stanovit tak, jak vyhovuje jejím potřebám. Zpravidla je určující pouze omezení počtu souborů v informačním balíčku.

Pokud bychom navázali na předchozí příklad z praxe Národní digitální knihovny, intelektuální entitu v případě již zmíněné monografie můžeme rozdělit (granulovat) na nižší celky následovně: intelektuální entitou je svazek, následuje množina jeho vnitřních částí (kapitol či jednotlivých obrazů), příloh a nakonec samotných jednotlivých stran. Pokud je svazek součástí vícesvazkové publikace (jako například Dějiny zemí Koruny české), potom do granularity nad vrstvu svazku nadřadíme ještě titul, který zastřešuje jednotlivé svazky.

3.5 Typy digitálních dokumentů

V kontextu digitální ochrany obecně rozeznáváme tři digitálních dokumentů: digitalizáty, originálně digitální dokumenty a elektronické archiválie. Všechny tyto tři skupiny jsou velmi širokými množinami dokumentů, které v sobě zahrnují množství podtypů (BEAGRIE, 2008 cit. podle CUBR, 2010a).

3.5.1 Digitalizáty

Digitalizáty *(digital copies, digital surrogates)* označují dokumenty, které byly do digitální podoby převedeny z fyzických nosičů (analogových dokumentů). Mnohdy se jedná o dokumenty, které byly digitalizovány z důvodu ochrany své analogové verze před přílišným fyzickým namáháním či z důvodu degradace původního média. Může se jednat o tištěné knihy, periodika, mapy, rukopisy, zvuková média jako gramofonové desky, optické disky a mnoho dalších.

3.5.2 Originální digitální dokumenty

Výhradně digitální dokumenty nebo e-born dokumenty *(born digital documents;* také často označovány jako *eborn dokumenty* bez pomlčky*)* označují skupinu dokumentů, při jejichž vzniku se nepočítalo s tím, že budou mít svůj analogový ekvivalent (CUBR, 2010a). Vzhledem k této skutečnosti se jedná o velmi zranitelnou část digitálního dědictví, která byla jedním z původních impulzů pro vznik celého oboru dlouhodobé digitální ochrany.

Mezi e-born dokumenty patří také digitální publikace, určené ke všeobecnému šíření ať už zdarma, či za poplatek. Digitální publikace mohou mít online i offline podobu (šířenou na nosičích), mohou být statické (tj. v uzavřené podobě) nebo dynamické, pokud jejich obsah podléhá aktualizacím.

3.5.3 Elektronické archiválie

Elektronické archiválie *(electronic records)* jsou dokumenty vzniklé z činnosti nějaké organizace. Formálně jsou chráněny legislativou. Může se jednat o různé typy dokumentů od emailové korespondence přes databáze, webové stránky až po dokumenty zachycující výsledky činnosti organizace (CUBR, 2010a, s. 46).

3.6 Informační balíčky

Model OAIS pracuje s konceptem informačních balíčků. Tyto balíčky jsou množinami, ve kterých se nachází data určená k archivaci, metadata, která tato data popisují, a dále metadata, která popisují informační balíček a informace o jeho zabalení. Tato metadata tvoří dohromady informace, které jsou nutné k ochraně obsažených dat a digitálních objektů v balíčku. V rámci archivu OAIS rozeznáváme tři druhy datových balíčků, které de facto popisují stádia, kterými informace v datovém archivu procházejí.

1) Prvním druhem balíčku je **balíček SIP** *(Submission information package)*, který označuje datový balík, jenž vstupuje do archivačního prostředí OAIS po odevzdání producentem do příjmu repozitáře *(Ingest)*. Tento balík obsahuje data určená k archivaci a k nim náležící zejména popisná a technická metadata.
2) Druhým typem je **balíček AIP** *(Archival information package)*. Balíček AIP je po přijetí dat do archivu vytvořen z balíčku SIP dodáním dalších potřebných metadat či změnou struktury jeho obsahu. Takto vytvořený balíček je již vhodný k archivnímu uložení, protože je obohacen o tzv. popisné ochranné informace *(Preservation description information)* o obsahových informacích v balíčku.
3) Třetím typem je **balíček DIP** *(Dissemination information package)*, který může být vytvořen z částí jednoho nebo částí více archivních balíčků v momentu, kdy uživatel repozitáře (kterým může být jak osoba, tak aplikace) zadá požadavek na vyvolání dat z archivu. Tento balíček je pak exportován ven z archivu OAIS pro potřebu rozšíření informace (CUBR et al., 2023).
4) Mimo výše zmíněné tři typy balíčku, které vychází přímo z modelu OAIS se v digitalizační praxi používá ještě koncept **produkčního PSP balíčku** *(Producer submission package)*, který byl vytvořen v mezinárodní pracovní skupině LTP systémů, iniciované národními knihovnami Nizozemí a Německa (HUTAŘ, 2012, s. 71–72). Tato (dočasná) mezinárodní skupina, složená z členů odborné digitalizační komunity, vznikla za účelem obecné specifikace funkcionalit LTP systémů, které by mohly být využity pro potře-

by výběrových řízení pro tento typ systému. Výsledkem jednání bylo mimo jiné rozšíření obecného rámce OAIS v oblasti Ingest, který v případě referenčního rámce neřeší samotný vznik SIP balíku; předpokládá, že při příjmu je již balík hotový a vhodný k okamžitému použití. V realitě LTP systémů a jejich různých požadavků na formáty a strukturu toto ovšem není vždy reálné. Před samotným přijetím dat do archivu v rámci modulu Ingest se tedy data *předzpracují* v modulu či aplikaci mimo archiv OAIS a z produkčního PSP balíčku se stane balíček SIP. Modul, který tato pracovní skupina k modelu OAIS přidala, se nazývá *Pre-Ingest*.

4 Souborové formáty pro digitální archivaci a zpřístupnění

Pro praxi digitálního archivu je nutno vymezit, v jaké podobě bude data v SIP balících přijímat a jak budou tato data popsána. Tuto podobu definují souborové a metadatové formáty. V této a následující kapitole představíme kritéria výběru souborových i metadatových formátů, přístupy ke standardizaci a rizika, která plynou jak z přísně restriktivního přístupu, tak z liberálního. Pro souborové formáty kapitola stručně představí možnosti v oblasti formátové politiky a výběr formátu s přihlédnutím k udržitelnosti formátu.

Při dlouhodobé ochraně digitálního obsahu je třeba zajistit nejen neporušenost dat (bitovou ochranu), ale také to, že těmto datům i v budoucnu porozumíme a dokážeme jejich obsah zobrazit (logická ochrana). Pro tento účel je mimo jiné nutné rozumět souborovým formátům, v nichž je obsah uložen, a dokázat s nimi pracovat.

Paměťové instituce řeší v souvislosti s formáty dvě zásadní otázky. První z nich je volba archivačního master formátu, tj. formátu, do něhož budou ukládat výstupy svých vlastních digitalizačních projektů. Pro tento účel by měly zvolit formát, který je široce používán napříč paměťovými institucemi, je pro něj dostupný software nezbytný pro dlouhodobou ochranu a který má dobré předpoklady pro to, aby byl užíván dlouhodobě. Druhou otázkou je pak rozhodnutí, které formáty bude instituce pro dlouhodobé uložení přijímat od externích producentů. I v tomto případě může jít o produkty digitalizace, nebo se může jednat o objekty, které v digitální podobě již vznikly, např. e-knihy nebo digitální video. Na tento typ souborů pak někdy bývají kladeny nižší nároky než na archivační master formát, do kterého ostatně mohou být po přijetí transformovány.

Formát je specifické uspořádání datových a strukturních elementů souboru (LAWRENCE et al., 2000). U konkrétního souboru ovšem můžeme jeho strukturní i datové části definovat ještě podrobněji

a rozlišujeme tak dále verze formátu a formátové profily. Například u rastrového obrázku ve formátu TIFF bude v případě archivního souboru velmi pravděpodobně verzí tzv. revize 6,[4] obvykle označovaná jako TIFF_6, a formátovým profilem buď typ užité komprese, nebo nekomprimovaný obsah. U formátu JPEG 2000 je formátový profil detailnější a zahrnuje více než desítku parametrů, které lze pro kompresi nastavit. Paměťové instituce, které JPEG 2000 používají, mají pro archivační master formát předepsaný profil sestávající z konkrétních hodnot pro tyto parametry. Také Národní knihovna si v rámci Standardu NDK specifikovala konkrétní parametry, které má formát JPEG 2000 splňovat jak pro archivní, tak pro uživatelské kopie.[5]

Formáty, které obsahují různé typy datových složek, případně i několik datových složek najednou, označujeme jako kontejnerové (někdy se používá také termín *wrapper*[6]). Pokud je datová složka komprimovaná (tj. má zmenšený objem dat), je kompresní algoritmus uveden pod označením kodek, což je program, který slouží pro kódování i dekódování dat. V jednom druhu kontejneru mohou být různé kodeky (resp. data příslušným kodekem komprimovaná) a zároveň určitý kodek může být uložen v různých kontejnerech. Existují ovšem ustálené kombinace kodek-kontejner, které ve výskytu převažují.

Kontejnerovým formátem pro různá obrazová data je i formát TIFF, ale nejznámější jsou samozřejmě multimediální, typicky audiovizuální formáty, kde je obsah závislý na dvou (nebo i více) různých složkách (obrazové a zvukové), a tedy i na dvou kodecích, pokud ovšem není obsah uložen v nekomprimované podobě. Formátovým profilem v takovém případě bude kombinace kontejneru i všech použitých kodeků včetně jejich parametrů.

4 Označení této verze jako „revize“ souvisí s tím, že formát TIFF má u všech „verzí“ z tradičních důvodů číslo 42, takže je nelze podle něj rozlišit jako verze.

5 https://standardy.ndk.cz/standardy-digitalizace/standardy-pro-obrazova-data

6 Termín *wrapper* však kromě kontejnerového formátu může označovat i programy schopné integrovat jiné programy – viz dále v této kapitole.

4.1 Práce s formátem

Pro práci s formátem je vždy třeba mít vhodný software. Pokud usilujeme o dlouhodobou ochranu, pak musíme mít software, který nejenže dokáže vytvořit příslušné soubory a zobrazit jejich obsah, ale také ověří, zda jsou tyto soubory v pořádku a jestli jejich struktura odpovídá předepsanému profilu. Tento proces se nazývá validace a předchází mu určení, o jaký formát se jedná (identifikace). Oba tyto úkony mohou být součástí rozsáhlejších kontrolních procesů, při kterých dochází rovněž ke kontrole metadat příslušných k archivním souborům. Metadata jsou umístěna v samostatných souborech a připojena k archivním datům v tzv. informačním balíčku (CUBR et al., 2020).

Identifikace je jednoznačné určení formátu a jeho verze (HUTAŘ a MELICHAR, 2015a). Přípona souboru nerozliší mezi verzemi formátu a v některých případech nemusí ani formátu souboru odpovídat, lze ji ostatně i ručně přepsat. Proto je v procesech dlouhodobé ochrany nutné formát identifikovat pomocí specializovaného nástroje, který formát identifikuje na základě specifických bitových sekvencí uvnitř souboru, tzv. *signatures*, které jsou pro daný formát typické. Některé z těchto sekvencí, například tzv. *magic numbers*, jsou přímo určeny k identifikaci formátu.

Základním nástrojem pro identifikaci formátů je program DROID, který je propojený s formátovým registrem PRONOM a přiřazuje souborům perzistentní, jedinečný a jednoznačný identifikátor PUID (PRONOM Persistent Unique ID). U tohoto programu však někdy dochází k chybám, a proto např. metodika pro tvorbu SIP balíčku (CUBR et al., 2020), která je určena pro české knihovny, doporučuje identifikovat formát nejméně dvěma nástroji a kromě DROID použít pro identifikaci ještě nástroj JHOVE.

Validace je kontrola, zda soubor splňuje strukturní požadavky předepsané pro daný formát. Soubor je správně strukturovaný *(well-formed)*, pokud splňuje syntaktické požadavky pro daný formát. Například u formátu TIFF tedy platí, že začíná osmibajtovou hlavičkou,[7] po které následuje posloupnost *Image File Directories* (obrazo-

7 https://jhove.sourceforge.net/

vých složek – TIFF může obsahovat více obrazů v jednom souboru), z nichž každá obsahuje dvoubajtové pořadové číslo a za ním řadu osmibajtových značených položek.

Soubor je validní *(valid)*, pokud je *well-formed* a navíc dodržuje sémantické požadavky pro validitu daného formátu, např. že pro soubor v barevném prostoru RGB musí být každý pixel popsán třemi hodnotami.

Validační programy[8] jsou buď formátově specifické jako například Jpylyzer, který je určen pro validaci formátu JPEG 2000, nebo mohou pracovat s více formáty, jako JHOVE. Mohou pracovat samostatně nebo být součástí nástrojů *(*obecně označovaných jako *wrapper)*, které integrují několik validačních programů, jako například nástroj FITS (The File Information Tool Set).

Validační nástroje jsou rovněž využívány pro extrakci technických metadat, tzv. **charakterizaci**. Získaná metadata jsou jako XML soubory uložena spolu s dalšími metadaty, OCR soubory a archivními soubory do tzv. informačního balíčku pro pozdější kontrolu. V Národní knihovně je tato kontrola, označovaná jako balíčková validace, prováděna pomocí nástroje Komplexní validátor[9]. Ten navíc využívá knihovny Kakadu a ImageMagick k validaci obrazových dat a dále podle vlastních šablon, vytvořených na základě oficiálních validačních schémat, validuje metadata. Výstup nástroje Jpylyzer porovnává s předpisem pro formátový profil archivních souborů JPEG 2000.

Formátová migrace, označovaná též jako transformace, je přepisem dat z jednoho formátu do jiného. Používá se buď časně, v případě, kdy např. výstup z digitalizace nebo obsah dodaný externími producenty je v jiném než archivačním formátu, a nebo v pozdější fázi archivace, pokud archivačnímu formátu hrozí zastarání a je potřeba ho nahradit novým. Pro transformaci např. z formátu TIFF do JPEG 2000 existuje několik nástrojů: komerční Kakadu a open source OpenJPEG a Jasper. Větší rozsah formátů nabízí knihovna ImageMagick.

8 https://coptr.digipres.org/index.php/Validation

9 Komplexnímu validátoru se podrobněji věnuje kapitola o LTP nástrojích v českém prostředí, částečně potom také kapitola o digitální archivaci v českých knihovnách.

Archivní kopie určené pro dlouhodobé uložení jsou často nekomprimované nebo bezztrátově [10] komprimované soubory a jako takové jsou velké. Vzhledem k tomu, že posláním digitalizačních projektů není pouze ochrana kulturního dědictví, ale také jeho zpřístupnění uživatelům, je nutné vytvořit též uživatelské kopie, které používají ztrátovou kompresi a jsou proto menší a rychleji se stahují přes internet.

Rastrové obrazové formáty užívané v dlouhodobém uložení dat

TIFF *(Tag Image File Format) byl vyvinut, aby sjednotil různé proprietární formáty využívané pro skenování dokumentů, a původně nesl pouze bitonální informaci (British Library Digital Preservation Team, 2015). Vyvinula jej firma Aldus, kterou následně koupila firma Adobe. Postupně byl rozšiřován a dnes podporuje i barevnou hloubku 16bitů na kanál. Formát byl navržen jako rozšiřitelný, byla mu přidávána podpora pro různé typy obsahu a v současnosti je to v podstatě kontejner schopný nést obrazová data buď v nekomprimované podobě nebo v některé z řady různých kompresí, a to jak ztrátových (JPEG) nebo bezztrátových (PackBits, LZW). Navíc je schopen nést více obrazů v jednom souboru. Pro účely dlouhodobé ochrany bývá jeho profil omezen na jeden obraz a nekomprimovaný či bezztrátově komprimovaný obsah. Pro přísnější omezení v zájmu dlouhodobé archivace byl vyvíjen formát TI-A, který se ale narozdíl od analogicky motivovaného PDF/A (viz níže) zatím neprosadil. Specifikace formátu TIFF je dostupná na stránkách jeho vlastníka, firmy Adobe.*[11] *Patentová práva na kontejner TIFF nejsou uplatňována. V minulosti bylo třeba brát v úvahu patenty pro kompresi LZW, ale jejich platnost vypršela v roce 2004. Pro vytváření souborů ve formátu TIFF existuje řada nástrojů, protože se jedná o univerzálně rozšířený formát. V paměťových institucích je často používaná open source knihovna ImageMagick.*

Validátor: JHOVE, DPF manager

10 Kdekoli se v tomto textu mluví o bezztrátové kompresi bez dalších přívlastků, je myšlena matematicky bezztrátová komprese. U vizuálně bezztrátové a dalších, které nejsou přísně matematicky bezztrátové je vždy explicitně uvedeno, o jaký druh se jedná.

11 https://developer.adobe.com/content/dam/udp/en/open/standards/tiff/TIFF6.pdf

JPEG 2000 *je rastrový formát využívající pro kompresi vlnkovou transformaci (OSTRÁKOVÁ, 2018). Byl vyvíjen konsorciem JPEG (Joint Photographic Experts Group) jako nástupce formátu JPEG využívajícího kosinovou transformaci. Umožňuje uložení v bezztrátové nebo ztrátové kompresi, nekomprimovaný obsah nenabízí. Podporuje též tzv. progresivní zobrazování, tedy postupné stahování náhledu v nižším rozlišení a následné zobrazení v plném rozlišení po stažení zbytku dat. Jeho adopce stagnovala kvůli výpočetní náročnosti, dokud vývoj hardwaru nedosáhl na potřebný výkon. Současně s tím ale rostla i velikost paměťových médií, a tak nakonec fotografové kompresi nepotřebují a JPEG 2000 zaujal pozici tzv. „niche formátu", tedy formátu, který není univerzálně rozšířen, ale obsadil si svou „niku", tj. své vymezené pole působnosti. V případě JPEG 2000 to jsou medicínské zobrazovací aplikace, geologické a geografické aplikace a využití v paměťových institucích. JPEG 2000 je chráněn patenty, ale jejich držitelé je uvolnili pro otevřené použití. Aktuální specifikace je ISO/IEC 15444-1:2019 (ISO, 2019). Pro formát JPEG 2000 existují open source nástroje OpenJPEG a Jasper a komerční nástroj Kakadu.*
Validátor: Jpylyzer, JHOVE.

PNG *je univerzálně rozšířený rastrový formát,*[12] *který původně vznikl jako náhrada GIF. Využívá kompresní algoritmus DEFLATE, který umožňuje bezztrátovou kompresi. Stejně jako TIFF a JPEG 2000 podporuje bitovou hloubku až 16bitů na kanál (tj. 3 nebo 4x tolik na pixel, 24 nebo 32 v závislosti na tom, zda je využit alfakanál). Podporuje progresivní zobrazování. Má vlastní dokumentaci, ISO a W3C, která je zdarma,*[13] *používá ho Australský národní archiv a je docela rozšířený mezi menšími institucemi v USA. Mimo paměťové instituce je plošně rozšířený už proto, že je podporován většinou internetových prohlížečů. Je implementován ve většině grafických programů. Není chráněn patenty.*
Validátor: JHOVE, BadPeggy

12 http://www.libpng.org/pub/png/
13 https://www.w3.org/TR/png/

4.2 Udržitelnost formátu

Již v době analogových médií ohrožoval dlouhodobou ochranu digitálního dědictví fenomén zvaný *playback drift*, označující nástup nových médií a ústup těch starých. V digitální éře se tento problém znásobil, protože se již netýká pouze fyzických médií (byť těch samozřejmě také – i paměťová média zastarávají), ale i digitálních objektů, včetně formátů. K tomu, aby bylo možné přečíst daný formát i v budoucnosti, bude potřeba nejen funkční nástroj pro toto čtení, ale i operační systém, v němž bude nástroj fungovat, a hardware, na kterém systém poběží (SHIRKY, 2005).

Formáty zastarávají pomalu a nikdy se nedá naprosto spolehlivě říci, že daný formát nepůjde otevřít, protože někde pravděpodobně bude archivován program, který to dokáže (Knijff, 2014). Takový starý program ale poběží velmi pomalu, a i kdyby se ho podařilo spustit na současném počítači, zdaleka nebude schopen využít jeho plnou kapacitu, přinejmenším například proto, že nebude mít podporu užití vícejádrových procesorů *(multithreading)*. Takový postup je tedy spíše strategií pro záchranu dat než pro archivní praxi. Jako argument pro to, že žádný formát není přísně vzato nepoužitelný, to ale samozřejmě platí. Proto se také v anglosaské/sekundární literatuře v souvislosti s formáty používá spíše pojem *obsolescent*, tedy zastarávající, místo *obsolete*, zastaralý.

Navíc je pro archivní práci s formátem třeba nejen program, který dokáže formát otevřít, ale také všechny další programy nutné pro archivní práci. V delším časovém horizontu to může znamenat problém: pro nové operační systémy už tyto programy nikdo nevytvoří, což se nemusí projevit okamžitě po přechodu na jinou generaci, ale později ano. Již to, že je formát „zastarávající", je tedy dobrým důvodem vyvarovat se jeho užití.

4.3 DjVu – příklad obsolescentního formátu

Příkladem formátu, od jehož používání paměťové instituce upustily z důvodů zastarávání, je formát DjVu. Přinejmenším ve své době to byl velmi pokročilý formát určený pro ukládání a sdílení naskenovaných dokumentů, který byl v roce 2004 hodnocen dokonce lépe než PDF zejména proto, že dosahoval lepších výsledků v kompresi (DjVu, 2023). Standardizace v roce 2008 nicméně umožnila širší implementaci PDF, což spolu s vyšší uživatelskou vstřícností přispělo k jeho větší oblibě. Formát PDF se tak stal nejužívanějším formátem pro sdílení dokumentů a formát DjVu vytlačil do ústraní.

Zatímco přístup k formátu PDF byl postupně přímo implementován do internetových prohlížečů, pro formát DjVu bylo nutné instalovat plug-in. Po nástupu nového prohlížeče Edge v roce 2015 již plug-in pro tento prohlížeč ani nebyl vytvořen. V roce 2002 byl formát DjVu používán pro Million Book Project organizací Internet Archive, která od něj v roce 2016 ustoupila. Od roku 2011 od aktivního užití formátu DjVu upouští také Národní knihovna, která svá dříve vytvořená data postupně převádí do aktuálně využívaného JPEG 2000.[14]

4.4 Faktory udržitelnosti

Již před více než dvaceti lety identifikovali výzkumníci faktory, které udržitelnost formátu ovlivňují (BROWN, 2008). Následující výčet vychází z přehledu Kongresové knihovny (LOC, 2017). Mnohé faktory, neboli vlastnosti formátů, jsou přínosné nejen pro udržitelnost formátů, ale i pro bezprostřední práci s nimi.

Otevřenost *(Disclosure)* udává, do jaké míry je pro daný formát dostupná dokumentace a validační nástroje. Dokumentace (označovaná též jako specifikace) je nezbytná pro tvorbu softwaru pro práci s formátem, případně pro jeho implementaci do větších programů,

14 S formátem DjVu se lze stále setkat ve starší verzi digitální knihovny Kramerius 3 (http://kramerius.nkp.cz/kramerius/Welcome.do), většina jejího obsahu je však již souběžně zpřístupňována v modernějším rozhraní Krameria 5 (https://ndk.cz/).

které pracují s více formáty. Ideálně by měla být natolik podrobná, aby čistě na jejím základě bylo možné napsat program, který bude schopen s formátem pracovat a přinejmenším zobrazit obsah. V případě formátu JPEG 2000 byla kvalita specifikace prakticky ověřena tak, že bylo napsáním programu pověřeno několik programátorů, kterým se skutečně podařilo na základě ISO specifikace příslušný software vytvořit (LOC, 2023a). U videoformátu FFV1 provedl tento experiment spontánně Derek Buitenhuis (2019). Dostupnost specifikace samozřejmě usnadňuje i tvorbu programů pro archivní práci s formátem. Za přínosné je považováno, když je specifikace standardizovaná, tzn. prošla revizí u důvěryhodné standardizační instituce. Jako nejlepší forma dokumentace bývá některými institucemi hodnocena specifikace ISO (International Organization for Standardization), která má ale výraznou nevýhodu v tom, že není dostupná zdarma. Jinou variantou je např. využití specifikace IETF (Internet Engineering Task Force), která je dostupná bezplatně a využívá ji např. již zmíněný formát FFV1.

Rozšíření *(Adoption)* formátu mezi uživateli je patrně nejlepší indikátor šance na dlouhověkost. Je-li formát užívaný, mají výrobci softwaru zájem na jeho implementaci do svých programů. Jde vlastně o systém s pozitivní zpětnou vazbou. Některé instituce hodnotí zvlášť rozšíření formátu v paměťových institucích.

Patentové otázky *(Impact of patents, Legal issues).* Patentová ochrana souborových formátů může být na překážku vývoji otevřeného softwaru. Paměťové instituce obecně dávají přednost otevřeným formátům a otevřenému softwaru, protože vznikají v komunitě, která je s paměťovými institucemi v kontaktu, případně je rovnou jejich součástí, a je tu velmi silná zpětná vazba. Neopomenutelnou výhodou otevřených formatů a softwaru je samozřejmě i to, že jsou dostupné zdarma.

Patentová situace bývá často poměrně nepřehledná. Programy, především ty pro tvorbu komprimovaných souborů (kodeky), využívají řadu dílčích procesů, které jsou chráněny různými patenty. Patenty vztahující se k jednomu kodeku tedy může vlastnit i více držitelů. Z toho důvodu vznikají sdružení těchto držitelů označovaná jako patentové pooly. Poměrně intenzivní patentové spory probíhaly

u formátu MPEG, kde vzniklo patentových poolů hned několik. Jejich zvyšující se požadavky na licenční poplatky nakonec vedly k tomu, že se odtrhla skupina významných uživatelů formátu MPEG a vytvořila si vlastní formát AV1.

Ochrana autorských práv (*Digital Rights Management,* DRM*),* tedy mechanismy ochrany autorských práv, je samozřejmě na překážku dlouhodobé ochraně. Pokud jsou data zakódována, představuje to větší riziko, že i malé poškození dat znehodnotí větší část obsahu, a případné poškození nebo ztráta klíče by znamenaly ztrátu celého souboru. To neznamená, že formáty podporující ochranu práv není možné v dlouhodobém uložení použít, ale tato ochrana nesmí být aktivována/využita.

Autodokumentace *(Self-documentation)* představuje schopnost formátu nést metadata popisující jeho vlastní strukturu. To usnadňuje sledování stálosti souborů uložených v daném formátu.

Transparenost/komplexita. Pod těmito dvěma protichůdnými pojmy je pojednávána přístupnost formátu k analýze pomocí základních nástrojů, které nejsou založeny na znalosti formátu. V ideálním případě by měl být *human readable*, tedy přístupný analýze pouze prostřednictvím textového editoru. To samozřejmě není možné u kompresních formátů, které mají vysokou komplexitu, a tedy nízkou transparentnost, již ovšem může vyvažovat kvalitní dokumentace. Někteří autoři v rámci komplexity pojednávají i DRM mechanismy.

Závislost na vnějších zdrojích *(External dependencies).* Některé formáty umožňují používat pro zobrazení obsahu vnější zdroje informací, například fonty, které jsou umístěné na internetu. To samozřejmě není dobrá záruka pro zobrazení obsahu ve vzdálené budoucnosti, protože tyto zdroje se mohou ztratit úplně, anebo jen někdo přeorganizuje jejich uložení a efekt je stejný: odkaz nefunguje.

Textové formáty užívané v dlouhodobém uložení dat

PDF/A *Formát PDF je tzv. stránkově orientovaný formát, tedy formát, u kterého je jednoznačně definované rozložení textu i grafiky, což je v angličtině označováno pojmem „fixed layout" (fixní sazba).*[15] *PDF byl*

15 https://wiki.dpconline.org/images/f/ff/PDF_Assessment_v1.5.pdf

ovšem postupně rozšiřován o další možnosti obsahu včetně audia, videa, spustitelného kódu a externího obsahu. V důsledku toho se možnosti formátu staly nepřehlednými a došlo k tzv. formátové proliferaci, což je situace, kdy existuje mnoho variant souborů identifikovatelných jako určitý formát – v tomto případě PDF. Ta přináší riziko, že se v úložišti vyskytne soubor, který v budoucnu nebude možné otevřít.

Dvě asociace zabývající se standardizací tiskových předloh a zpracování obrazu proto ve spolupráci s firmou Adobe stanovily profil PDF pro archivaci, označovaný jako PDF-A (LOC, 2020). V něm jsou zakázána rozšíření, která by mohla ohrožovat přístup k obsahu. V současnosti existuje několik variant PDF/A, označovaných jako „flavours", které se liší přísností těchto omezení. Pro všechny platí zákaz videa, audia, JavaScriptu a spustitelných souborů a též zákaz využívání externího obsahu, na který lze pouze odkazovat, a dále povinnost vložit do souboru (embedovat) všechny použité fonty spojená s povinností použít pouze fonty, které lze embedovat legálně, tj. nejsou chráněny licencí, která by to zakazovala. Varianta PDF/A-1 je standardizována jako ISO 19005-1:2005 (ISO, 2005).
Validátor: VeraPDF, JHOVE

***EPUB** je formát určený pro e-knihy a jeho výchozí podoba je text bez fixní sazby, u kterého si čtenář může zvolit velikost písma a text se podle toho automaticky zalomí, což je v angličtině označováno jako „reflowable".[16] Od verze 3 má ovšem ePub i možnost fixní sazby, a navíc zde přibyla i další rozšíření, z nichž mnohá (například externí obsah) je třeba v archivní specifikaci zakázat. Verze 3.0.1 je standardizována jako ISO/IEC 23736:2020 (ISO, 2020).*
Validátor: EPUBCheck

4.5 Faktory kvality a funkcionality

Kromě faktorů, které ovlivňují udržitelnost formátu, hodnotí paměťové instituce i technické vlastnosti, které ovlivňují schopnost formátu

16 https://wiki.dpconline.org/images/7/73/EPUB_Assessment_v1.4a.pdf

zachovat signifikantní vlastnosti originálu (LOC, 2023a). Tyto faktory se liší podle typu obsahu, tedy zda se jedná o text, obrazová data, audio, video a jiné. V případě obrazových dat je například sledována podpora vysokého rozlišení, barevná hloubka, správa barev a další funkce.

Některé instituce posuzují formáty nejprve s ohledem na jejich technickou způsobilost nést zamýšlený obsah a až poté je pouze u vybraných formátů provedeno hodnocení po stránce udržitelnosti. Jiné instituce hodnotí oba druhy faktorů současně.

4.6 Zdroje informací o formátech

Dostupnost informací o formátech je nezbytná pro práci s nimi a instituce zabývající se dlouhodobým uložením dat proto začaly vyvíjet tzv. formátové registry, v nichž měly být tyto informace, případně odkazy na ně, uloženy. Z několika takových projektů přetrval pouze registr PRONOM, vytvořený a provozovaný britskými Národními archivy. Tento registr však obsahuje pouze minimální popisy formátů a je tedy vhodný především pro identifikaci formátů v součinnosti s programem DROID.

Skutečně obsáhlé formátové popisy tak obsahuje až databáze formátových popisů (LOC, 2022b) v sekci *Sustainability of digital formats* na stránkách **Kongresové knihovny** (Library of Congress). Popisy začínají obecnou identifikací a popisem, které jsou následovány poznámkou o lokálním užití, pokud LOC daný formát používá. Další část obsahuje informace o vlastnostech formátu vztahujících se k jednotlivým faktorům udržitelnosti, kvality a funkcionality. Následuje podrobná tabulka signifikátorů a identifikátorů, závěrečné poznámky a odkazy. Pokud existují, jsou nejprve uvedeny odkazy na specifikace formátu a poté na další užitečné zdroje.

Přestože databáze obsahuje velké množství formátů, nenajdeme zde formáty, které jsou pro dlouhodobé uložení nevhodné, už pro to, že u nich řada informací důležitých pro posouzení vhodnosti úplně chybí. To je například případ proprietárních RAW formátů výrobců

fotoaparátů, u kterých není dostupná dokumentace, což u proprietárních formátů nezřídka bývá záměr ze strany jejich původců.

Britská knihovna (British Library, BL) vypracovala hodnocení vybraných formátů, která jsou v současnosti umístěna na stránkách OPF (Open Preservation Foundation).[17] Tato hodnocení, označovaná jako *format assessment*, jsou nejdetailnější z dostupných a jsou strukturována podle vlastností důležitých pro udržitelnost. Hodnocení vznikla jen pro omezený počet formátů, které BL považuje za vhodné jako archivační master formáty, například z rastrových obrazů se jedná pouze o TIFF a JPEG 2000. U JPEG 2000 do hloubky rozebírá rizika s tímto formátem spojená a pro jeho uživatele připojuje doporučení každoročně monitorovat případné změny statusu tohoto formátu v otázkách souvisejících s jeho udržitelností.

KOST-CECO (Koordinační centrum pro dlouhodobou archivaci elektronických dokumentů Švýcarského federálního archivu) má na svých stránkách rovněž hodnocení vybraných formátů.[18] Hodnocení jsou strukturovaná podle faktorů, jsou poměrně stručná a doplněná bodováním, které je přeneseno do přiložené matice, kde jsou formáty navzájem porovnávány.

Knihovna Harvardovy univerzity (Harvard University Library)[19] a **Národní úřad pro archivaci a dokumentaci** (National Archives and Records Administration, NARA)[20] vytvořily pro hodnocení formátů matice, ve kterých mají jednotlivé faktory rozdělené do indikátorů, dílčích otázek týkajících se toho, zda formát splňuje podmínky pro užití v dlouhodobém uložení. Například faktor *otevřenost formátu* je u NARA rozdělen do pěti otázek: zda je formát proprietární, má publikovanou otevřenou specifikaci, jsou pro něj dostupné validační nástroje, zda je specifikace schválena a publikována mezinárodně uznávanou organizací pro standardy a jestli je specifikace kompletní a přesná. Tyto otázky jsou u NARA formulovány tak, aby na ně bylo

17 https://wiki.dpconline.org/index.php?title=File_Formats_Assessments

18 https://kost-ceco.ch/cms/de.html

19 https://docs.google.com/spreadsheets/d/1rR7HNoQswcOrl66yeRRI2qMGDKzYQxitrOmD7nfVFGQ/edit#gid=0

20 https://github.com/usnationalarchives/digital-preservation/tree/master/Digital_Preservation_Risk_Matrix

možno odpovědět ano nebo ne, a jsou hodnoceny pouze bodově od dvou do minus dvou bodů. V matici Knihovny Harvardovy univerzity jsou u jednotlivých položek odpovědi slovní a jejich význam pro hodnocení je zdůrazněn podbarvením buněk zeleně, oranžově a červeně, přičemž na vedlejším listu (jedná se o tabulku Excel) je uveden rozlišovací klíč pro určení významu podbarvení pro každou jednotlivou otázku.

Digitalizační iniciativa federálních úřadů (The Federal Agencies Digitization Initiative, FADGI) ve spolupráci s LOC vypracovala jednak srovnání pěti vybraných obrazových formátů a dále i studii pěti videoformátů, ve které jsou zvlášť hodnoceny kontejnery *(*pod označením *wrappers)* a kodeky *(encodings)*.[21] Formáty jsou hodnoceny na základě udržitelnosti, nákladů na správu, náročnosti na implementaci a technických vlastnosti.

4.7 Formátová doporučení

Kongresová knihovna vydává každoročně *Deklaraci doporučených formátů* (LOC, 2023d) pro všechny typy obsahu jak v digitální, tak ve fyzické formě. Neřeší se v ní ovšem pouze formáty, ale také doporučení ohledně technických charakteristik, metadat a dalších parametrů, byť právě v otázce technických parametrů zdaleka ne tak detailně jako níže uvedená metodika FADGI. Vyhláška řadí jak formáty, tak další parametry do dvou kategorií: doporučené *(Preferred)* a přijatelné *(Acceptable)*. V první kategorii pak například pro rastrové obrazy doporučuje formáty TIFF, JPEG 2000, PNG a JPEG.

Hodnocení formátů organizace KOST-CECO obsahují v závěrečném odstavci *(Results)* též doporučení ohledně možného užití formátu v dlouhodobém uložení. Ta jsou u některých formátů doplněná odstavcem *Preservation planning*, v němž navrhuje pro daný formát specifické parametry užití, které se ovšem někdy liší od obecně přijímaných zásad, kupříkladu pro formát TIFF připouští i použití souborů s více obrazy.

21 https://www.digitizationguidelines.gov/guidelines/File_format_compare.html

Metodika *FADGI–Technical Guidelines for Digitizing Cultural Heritage Materials* (Technické pokyny pro digitalizaci kulturního dědictví)[22] je zaměřena pouze na obrazová data a doporučuje formáty TIFF s nekomprimovanými daty nebo bezztrátovou kompresí ZIP (DEFLATE), JPEG 2000 v bezztrátové kompresi nebo vizuálně bezztrátové kompresi, PNG v kompresi ZIP a PDF/A buď s kompresí ZIP nebo JPEG 2000. Metodika obsahuje i detailní popis technických parametrů pro archivní soubory, které jsou odstupňované pro čtyři různé stupně kvality (označené jednou až čtyřmi hvězdičkami, přičemž čtyři označují nejvyšší kvalitu). V případě rastrových obrázků je to samozřejmě rozlišení, ale také například bitová hloubka nebo barevné prostory vhodné pro danou kvalitu.

4.8 Formátové politiky

Zatímco předchozí doporučení autoritních institucí jsou určena ostatním institucím jako vodítko pro jejich práci, formátové politiky lokálních institucí vyjadřují, jaké formáty jsou přijatelné pro jejich úložiště. Kyle Rimkus s kolegy provedl srovnání (RIMKUS et al. 2014) formátových politik členů Association of Research Libraries (ALR).

V Národní knihovně byl průzkum těchto formátových politik doplněný o další, které bylo možné dohledat na internetu. Přehledy těchto průzkumů byly zveřejněny jako doplňující informace v několika publikacích, týkajících se různých typů formátů. V článku o digitalizaci fonoválečků (BEŇAČKOVÁ et. al., 2020a) byl prezentován přehled přijetí zvukových formátů. V této práci je i přehled odkazů na sledované instituce. V další publikaci (KOPSKÝ, 2022) byl zveřejněn přehled videoformátů.

Přehled formátových politik u rastrových obrazových formátů je uveden v tabulce č. 1. Za pozornost stojí skutečnost, že kategorie přijetí *(categories of acceptance)* nejsou u všech institucí stejné, což indikuje, že i jejich formátové politiky se mohou lišit v důsledku rozdílných požadavků, které jsou na ně kladeny.

22 https://www.digitizationguidelines.gov/guidelines/FADGI%20Technical%20Guidelines%20for%20Digitizing%20Cultural%20Heritage%20Materials_3rd%20Edition_05092023.pdf

Tabulka č. 1 Srovnání formátových politik paměťových institucí pro rastrové obrazové formáty.
Význam zkratek: unc – nekomprimovaný, cmp – komprimovaný, l.less – bezztrátový, lossy – ztrátový

RASTER FORMATS CONSIDERED FOR ARCHIVAL USE →		TIFF (unc.)		TIFF (cmp.)		JP2 (l.less)		JP2 (lossy)		GIF		PNG		JPG		BMP		Categories of acceptance	
INSTITUTION	RELEASED	HI	ME	HI	ME	HI	ME	HI	ME	HI	ME	HI	ME	HI	ME	HI	ME	HIGH	MEDIUM
Alabama uni. lib.	2016	Sup					Kno		Kno	Sup		Sup		Sup				Supported	Known
Arthur Lakes Library	2018	Full					Lim		Lim		Lim		Lim	Full			Lim	Full support	Limited supp
Boston Uni Libraries	2011	For					Bit			For			Bit	For			Bit	Format	Bit-level
Canada library and archives	2015	Pre				Pre					Acc	Pre			Acc			Preferred	Accepted
Connecticut uni. lib.	2018	Sup								Sup		Sup		Sup			Kno	Supported	Known
Cornell University Li	2019	Hi			Me	Hi			Me		Me	Hi			Me		Me	High	Medium
Deep Blue (Michigan uni)	2011	L1					L2						L2	L1				Level 1	Level 2
Florida uni. libraries	2012	Hi			Me	Hi			Me		Me	Hi			me		Me	High	Medium
Hawai'i uni.	2019	Sup								Sup		Sup		Sup			Kno	Supported	Known
Houston uni, TX	2018	Hi					Me				Me		Me		ME			High	Medium
LOC	2016	Pre				Pre		Pre		Pre		Pre		Pre		Pre		Preferred	Acceptable
Minnesota University	2014	Full					Lim				Lim		Lim	Full			Lim	Full support	Limited supp
National Archives (USA)	2019	Pre				Pre					Acc	Pre			Acc			Preferred	Acceptable
North Carolina State Archives	2012	Rec				Rec					Acc		Acc		Acc			Recommended	Acceptable
North Carolina State University Libraries	2018	Sup								Sup		Sup		Sup			Par	Supported	Partially supported
Northwestern University	–	Hi.r.			Mo.r.	Hi.r.			Mo.r.				Mo.r.		Mo			Highly recom.	Moderately recom.
Purdue University Libraries	2012	Sus				Sus					Sup		Sup		Sup			Sustainable	Supported
McMaster, Ontario, CAN	–	Full											Bit		Bit			Full	Bit level preserv.
Simon Fraser University Archives	2017	Pre				Pre				Pre		Pre		Pre			Acc	Preferred	Acceptable
Smithsonian archives	–	Pre					Acc						Acc		Acc			Preferred	Acceptable
Southern Illinois uni.	2008	Sup								Sup		Sup		Sup			Kno	Supported	Known
Tasmanian Archives	2015	Rec				Rec					Acc	Rec			Acc			Recommended	Acceptable
Texas A&M uni., TX	2014	Pre					Acc								Acc			Preferable	Acceptable
UK Data service	2014	Rec			Acc		Acc				Acc		Acc		Acc		Acc	Recommended	Acceptable
W – University of Washington Libraries	2014	HI			ME	HI			ME		ME	HI			ME		ME	Highest	Medium
		24			5	11	9	1	6	8	12	13	10	11	13	1	12		

V loňském roce pak OPF (Open Preservation Foundation) zveřejnila přehled formátových politik (OPF, 2022),[23] ve kterém převažují evropské archivy, čímž se liší od našeho přehledu, v němž převažují severoamerické instituce, a to především univerzitní knihovny. Výstupem je součet bodů, které OPF přidělila formátům na základě hodnocení každé z 28 sledovaných institucí. Body byly přidělovány v rozsahu od –1 do +2 a maximálně bylo možné přidělit 56 bodů.[24] V našem vzorku srovnáváme formáty podle toho, kolik institucí je zařadilo do nejvyšší kategorie přijetí.

U rastrových formátů je ve vzorku OPF nejlépe hodnocen formát TIFF, který získal 50 bodů z 56 možných, následují JPEG 2000 (35) a PNG (22). V našem vzorku je rovněž na prvním místě TIFF (24 výskytů v nejvyšší kategorii přijetí ze 24 možných), ale JPEG 2000 a PNG se umístily v opačném pořadí (PNG 13 a JPEG 2000 11výskytů), přičemž JPEG 2000 se navíc o třetí pozici dělí s původním JPEG.

Ve skupině zvukových formátů získal ve studii OPF nejlépe hodnocený formát WAV 43 bodů, následují BWF (27), FLAC (30) a MP3 (19). Zvukové formáty jsou ve studii OPF v tabulce pro rok 2023 rozděleny na kontejnery a kodeky, nicméně z bodování je zjevné, že většina institucí kodeky nehodnotí (a pravděpodobně za vyjádření přijetí daného kodeku považují přítomnost odpovídajícího kontejneru ve své formátové politice). Výše uvedené hodnocení odpovídá kontejnerům, kde WAV má nejvyšší počet bodů, zatímco kódování LPCM, které je s ním nejčastěji spojeno, má nula bodů, zřejmě v důsledku toho, že se sledované instituce tímto rozlišením narozdíl od OPF nezabývaly. V našem vzorku má formát WAV 15 výskytů z 24 možných, BWF 9, FLAC 4 a MP3 pouze jeden výskyt, a to u organizace Archivematica, která v nejvyšší kategorii přijetí upozorňuje, že sem řadí i preferované zpřístupňující formáty.

23 https://docs.google.com/spreadsheets/d/1XjEjFBCGF3N1spNZc1y0DG8_Uyw18uG2j8V2bsQdYjk/edit#gid=893099148

24 Platí pro tabulku z roku 2022.

Zvukové formáty užívané při dlouhodobém uložení dat

WAVE *(též WAV) je kontejner pro zvuková data odvozený od obecného multimediálního kontejneru RIFF.*[25] *Nejčastěji je využíván v kombinaci s nekomprimovanými daty v kódování LPCM. Rozšířením tohoto formátu je BWF, který navíc obsahuje sektor (chunk) po uložení metadat využívaných v rozhlasovém vysílání (bext chunk). Jako specifikace formátu WAV je odkazována specifikace RIFF*[26] *vytvořená jako společný projekt IBM a Microsoftu, naopak formát BWF má vlastní specifikaci u European Broadcast Union (EBU).*[27] *WAV je univerzálně rozšířený formát a je implementován do většiny programů pro práci se zvukem.*
Validator: JHOVE, Shntool

FLAC *je otevřený formát pro uložení zvuku v bezztrátové kompresi vyvinutý nadací Xiph, která rovněž poskytuje jeho dokumentaci*[28] *a referenční implementaci libFLAC. Druhdy znám toliko jako předmět vášnivých debat v kroužcích audiofilů, dnes je FLAC univerzálně rozšířený formát s nativní podporou v operačních systémech včetně Windows 10 či Android 3.1. Má rovněž hardwarové implementace v audio a multimediální elektronice, byť, obzvláště v minulosti, ne tak rozšířené jako MP3, a je jedním z formátů komerční distribuce hudby. Může být též vložen do videokontejnerů jako zvuková složka filmu místo obvyklého nekomprimovaného LPCM. Samostatně se vyskytuje v nativním kontejneru s příponou .flac anebo kontejneru Ogg (.ogg). Formát FLAC má interní kontrolní součet.*[29] *Je implementován do řady programů, jak pro práci se zvukem, tak s videem.*
Validator: Lossless Audio Checker[30]

U videoformátů jsme rozlišení na kodeky a kontejnery použili i v našem přehledu, nicméně jak u nás, tak u OPF je zjevné, že informace získané z formátových politik jsou nekompletní, některé instituce

25 https://wiki.dpconline.org/images/4/46/WAV_Assessment_v1.0.pdf
26 http://www.tactilemedia.com/info/MCI_Control_Info.html
27 https://web.archive.org/web/20091229093941/http://tech.ebu.ch/docs/tech/tech3285.pdf
28 https://xiph.org/flac/format.html
29 https://wiki.dpconline.org/images/f/fe/FLAC_Assessment_v1.0.pdf
30 https://losslessaudiochecker.com/

uvádějí pouze kodek nebo pouze kontejner. V přehledu OPF se nejlépe umístil kodek FFV1 s 20 body (ne všechny sledované instituce se ovšem videoformáty zabývají), následovaný nekomprimovaným videem (13 bodů) a kodekem H.264 (MPEG2 s 12 body). V našem přehledu je nejužívanější videokodek JPEG 2000 u devíti institucí, nekomprimované video (6) a kodek FFV1 (3). U kontejnerů v OPF získal nejvíce bodů MPEG-4 (39), dále MKV (20) a MPEG-2 (18). V našem přehledu je nejužívanější kontejner AVI (9) a po něm MOV (7) a MKV (4).

Video formáty užívané v dlouhodobém uložení dat

FFV1 *je kodek pro bezztrátovou kompresi videa. Vyvinul ho Michael Niedermeyer v rámci projektu FFMPEG. FFV1 je velmi rychlý a z hlediska dlouhodobého uložení má velkou výhodu v možnosti vygenerování kontrolních součtů pro jednotlivé segmenty, do kterých jsou rozdělena obrazová políčka filmu. Pro účely dlouhodobé ochrany se používá v kombinaci s kontejnerem MKV (Matroska). Kodek byl standardizován u organizace IETF (Internet Engineering Task Force)*[31] *s významnou spoluprací a především zpětnou vazbou od uživatelské komunity. Je velmi populární v paměťových institucích, ale mimo ně je jeho rozšíření zanedbatelné. Pro tvorbu souborů ve formátu FFV1 slouží knihovna FFMPEG*
Validátor: MediaConch

JPEG 2000 *se používá nejen pro statická obrazová data, ale i pro video. Nejčastěji bývá kombinován s kontejnerem MXF, byť má i svůj vlastní kontejner MJPEG 2000, jehož název bývá někdy používán pro označení kodeku (LOC, 2021). Pro dlouhodobé uložení bývá užíván v bezztrátové kompresi, ve ztrátové kompresi byl velmi populární pro streaming. Standardizace je ISO/IEC 15444-3:2007*[32]*.*
Validátor: Jpylyzer

31 https://datatracker.ietf.org/doc/draft-ietf-cellar-ffv1-v4/
32 https://www.iso.org/standard/41570.html

4.9 Ostatní typy formátů

Kromě již zmíněných typů formátů: rastrových, textových, audio a video, ukládají některé paměťové instituce ještě další typy formátů, které jsou uvedené v jejich formátových politikách. Autoritní instituce ve svých doporučeních tyto typy formátů rovněž zahrnují. Jedná se o datasety, geografická data, designová a 3D data, software a videohry, prezentace, e-maily a webarchivy.

4.10 Restriktivní formátové politiky

Na zveřejnění výše popsané studie OPF reagoval polemický článek, který celý koncept formátových politik kritizoval jako příliš omezující (WHEATLEY, 2022). Jeho autor argumentuje tím, že pokud má úložiště zájem o data, která má jejich producent ve formátu, který úložiště nepřijímá, může se stát, že je producent vůbec neposkytne, protože nebude ochoten nebo ani schopen je sám převést do přijatelného formátu. Migrace by podle autora měla být vždy povinností úložiště. Autor ovšem neřeší situaci, kdy producent takové soubory teprve vytváří. Producent může například chtít obsah publikovat v PDF, ale vytváří jej ve Wordu nebo nějakém specializovaném nástroji. Pak je export tohoto obsahu daleko schůdnější pro něj než pro úložiště, které pak musí poměrně složitě konvertovat (HRZINOVÁ a JIROUŠEK, 2022). Proto například některé fakulty po svých studentech požadují odevzdávat diplomové práce v PDF/A.

Na druhou stranu však někdy úložišti nezbývá nic jiného, než přijmout obsah ve formátu, který pro dlouhodobé uložení není příliš vhodný. Jde zejména o případy, kdy obsah v takové podobě rovnou vznikl. Jedním takovým příkladem mohou být studentské projekty natáčené na amatérské videokamery, které zaznamenávají rovnou do ztrátově komprimovaných formátů (nejčastěji MPEG). V každém případě by se měla úložiště, která mají za úkol přijímat obsah od externích producentů, snažit s těmito producenty o přijatelném formátu komunikovat.

4.11 Formátové politiky NDK

Jako příklad formátových politik v českém prostředí může sloužit Standard NDK, specifikující požadavky Národní knihovny pro příjem souborových formátů v oblasti rastrových obrazů, zvukových nahrávek a textových elektronických publikací. Pro rastrové obrazy požaduje Národní knihovna u archivačních kopií využití formátu JPEG 2000 odpovídajícímu vlastnímu předepsanému formátovému profilu s bezztrátovou kompresí, pro uživatelské kopie pak týž formát s vizuálně bezztrátovou kompresí[33]. Jako archivační formáty pro zvuková data byly zvoleny formáty WAVE a BWF obsahující zvuk kódovaný bezztrátovou nekomprimovanou metodou LPCM, pro uživatelské kopie formát MP3[34]. Pro textové elektronické publikace je v době přípravy této publikace požadováno dodání formátu PDF/A-1, PDF/A-2, případně EPUB ve verzi 2.0.1.[35]

4.12 Ztrátová vs. bezztrátová komprese

Digitální ochrana kulturního dědictví, zvláště pokud jde o masovou digitalizaci tištěné produkce a audiovizuálních předloh, představuje ohromné množství dat, jejichž uložení vyžaduje značné finanční prostředky. Vhodnost použití komprese pro dlouhodobou ochranu proto odborníci řešili již v prvních dvou dekádách tohoto století, kdy byla finanční náročnost uložení dat podstatně palčivějším problémem než dnes. V teoretické rovině ji analyzovala Judith Rogová z nizozemské národní knihovny v roce 2007 (ROG, 2007).

Robert Buckley (2013) publikoval zprávu o užití ztrátové komprese JPEG 2000 u tří významných paměťových institucí. V úvodu zmiňuje, že zpočátku byl nejrozšířenějším archivačním formátem nekomprimovaný TIFF, což byla v té době vzhledem k jeho velikosti překážka pro masovou digitalizaci. Pro použití komprese pak uvažuje takovou

33 https://standardy.ndk.cz/ndk/standardy-digitalizace/standardy-pro-obrazova-data
34 https://standardy.ndk.cz/ndk/standardy-digitalizace/standardy-pro-zvukova-data
35 https://standardy.ndk.cz/ndk/standardy-digitalizace/standardy-pro-elektronicke-dokumenty

její míru, při které je daný soubor stále schopen plnit svou funkci, tedy kompresi označovanou obecně jako *funkčně bezztrátová*, v případě obrazových formátů nazývanou *vizuálně bezztrátová*. Buckley popisuje, jak jedna ze zmiňovaných institucí, Wellcome Digital Library, určovala hodnotu takové komprese tak, že výzkumníci postupně zvyšovali kompresní poměr a ve chvíli, kdy začali pozorovat kompresní artefakty (zkreslení obrazu, způsobená kompresí), vrátili se o krok zpět na předchozí kompresní poměr. Obrazy pozorovali ve 100% rozlišení na kalibrovaných monitorech. Jako vizuálně bezztrátový určili kompresní poměr 1:10 pro knihy a 1:8 pro ostatní materiály, což je podstatně větší úspora dat než matematicky bezztrátová komprese, u které je kompresní poměr 1:2 až 1:3. Buckley ovšem upozorňuje, že kvalita není při určitém kompresním poměru konstantní, ale závisí na charakteru předlohy, a představuje metody jak určit pro jednotlivé předlohy míru komprese potřebnou pro dosažení potřebné kvality.

Rogová ve výše citované práci (2007) rozebírá hlavní výhrady k použití komprese v digitální ochraně a zmiňuje relevantní protiargumenty. Již dříve jsme zmiňovali, že komprese vnáší do archivních souborů vyšší komplexitu, která ale může být vyvážena dostupností kvalitní dokumentace. Lze si také všimnout, že u formátů využívajících kompresi bývá zpravidla dokumentace kvalitnější a standardizovaná. U problému, kdy u komprimovaných souborů může i malé poškození dat mít za důsledek znehodnocení celého souboru, uvádí teoretickou úvahu, že protože jsou tyto soubory menší, je u nich menší pravděpodobnost takového poškození, a rovněž upozorňuje, že formát JPEG 2000 se osvědčil jako poměrně robustní. Poslední bod je jediný, který se týká pouze ztrátové komprese; všechny předchozí platí i pro matematicky bezztrátovou. Tento bod, obecně nazvaný *Komprese brání archivačním aktivitám*, se především zabývá problémem tzv. *generační ztráty*, který komplikuje jednu z hlavních archivačních aktivit – formátovou migraci.

Generační ztráta, tedy jev, kdy při pořizování kopií dat uložených ve ztrátově komprimované podobě dochází k dalším ztrátám informace (PALMER et al., 2013), je jedním z hlavních důvodů, proč mnozí odborníci dávají přednost bezztrátové kompresi. Pokud má instituce obsah uložený ve ztrátové kompresi, neměla by jej tedy migrovat do

dalšího ztrátového formátu. Místo toho se nabízí možnost migrovat obsah do bezztrátového formátu, ale to s sebou přináší podstatné zvětšení objemu dat, aniž by došlo ke zlepšení kvality. Další možností je tzv. rewrapping, kdy je datová složka souboru ponechána beze změny a mění se pouze kontejner *(wrapper)*, což se používá u některých videoformátů. V poslední době se potom objevila, byť zatím jen u jednoho formátu, možnost neztrátové rekomprese (ve smyslu, že nedochází k dalším ztrátám nad ty, ke kterým již došlo při vzniku původního souboru). Zatím se však vztahuje pouze na jeden formát: pro formát JPEG ji nabízí nový formát JPEG XL. Výsledný JPEG XL může být až o dvacet procent menší, a přesto nese všechnu informaci jako původní JPEG, což je potvrzeno tím, že tento proces je reverzibilní a ze vzniklého JPEG XL lze zpětně vytvořit soubor JPEG identický s tím výchozím (SNEYERS, 2022). Budoucnost ukáže, zda bude tento postup použitelný pro formátové migrace v rámci dlouhodobé ochrany kulturního dědictví a zda se podobné řešení podaří nalézt i pro další ztrátově komprimované formáty. Další možností, jak se generační ztráty vyvarovat, je migrace neprovádět a místo toho jít cestou emulace, tedy zakonzervování nástrojů pro zobrazení souboru. A samozřejmě nejspolehlivějším způsobem, jak se jí vyhnout, je nepoužívat ztrátovou kompresi. Další z uvažovaných strategií ve vztahu k použití komprese bylo používat nekomprimované nebo bezztrátově komprimované uložení pro obsah s vysokou důležitostí a pro méně důležitý obsah použít ztrátovou kompresi (ARMS a FLEISCHHAUER, 2005).

Autoritní instituce jako FADGI či KOST-CECO obecně podporují použití nekomprimovaných nebo bezztrátově komprimovaných souborů, ale v některých případech připouštějí i použití ztrátové komprese. Recommended format statement Kongresové knihovny má v kategorii „doporučené" *(preferred)* pro digitální fotografii doporučen nekomprimovaný obsah, pro textové soubory ve formátu PDF doporučuje bezztrátovou kompresi. V doporučení pro video se vyskytují i ztrátově komprimované formáty, ale vlastní digitalizační projekt LOC staví na bezztrátově komprimovaném JPEG 2000 a jeho autor George Blood uvádí generační ztrátu (pod synonymem *cumulative loss*) jako hlavní důvod této volby (BLOOD, 2011). U evropských

filmových archivů včetně českého Národního filmového archivu je oblíbený bezztrátový formát FFV1 (SVATOŠ, 2022). Organizace zabývající se digitalizací filmových materiálů obecně doporučují vyhýbat se ztrátové kompresi digitalizátů analogových předloh a Mezinárodní asociace zvukových a audiovizuálních archivů IASA dokonce varuje i před použitím bezztrátové komprese z důvodu vysoké komplexity a rizika nedostupnosti nástrojů pro dekódování takových souborů v budoucnu (PRENTICE a GAUSTAD, 2017).

4.13 Shrnutí

Při volbě archivačního master formátu je vhodné se nejprve seznámit s doporučeními autoritních institucí a z nabízených formátů vybrat takový, který nejlépe odpovídá potřebám dané instituce. Poté je třeba ověřit, zda jsou pro něj dostupné všechny nástroje potřebné pro archivní práci a tyto nástroje otestovat. I pokud již instituce s formátem pracuje, je stále třeba průběžně monitorovat, jestli se nezměnil jeho status z hlediska udržitelnosti, především zda je stále používán i dalšími paměťovými institucemi. Postup používaný při posuzování souborových formátů Národní knihovnou byl popsán v publikaci Ostrákové a Kopského (2020).

Při rozhodování o tom, které formáty přijímat od externích dodavatelů, se lze inspirovat formátovými politikami ostatních institucí, ale finální rozhodnutí mezi více či méně restriktivní formátovou politikou bude nakonec určeno především kvalifikací a odhodláním odborníků dané instituce, požadavky jejího zřizovatele a dohodou s externími dodavateli obsahu.

5 Metadatové formáty pro dlouhodobou archivaci a zpřístupnění

Druhým typem formátů, užívaných k popisu digitálních objektů, jsou metadata. Metadata nejsou součástí digitálního objektu, spíše jsou doprovodnými informacemi objektu, který popisují a uvádějí do kontextu.

Metadata jsou prostředkem pro třídění objektů podle různých charakteristik a jejich vyhledávání. V případě paměťových institucí se jedná o dodatečné, strukturalizované informace, které se pojí jak k digitálním, tak k analogovým dokumentům.

Metadata bývají obecně definována jako *„data o jiných datech“* (ISO 14721). Výstižnější je ovšem definice, která metadata popisuje jako *„informace, které vytváříme, ukládáme a sdílíme za účelem popisovat věci tak, abychom s nimi mohli interagovat a získávat z nich vědomosti, které potřebujeme“* nebo ve specifickém případě digitálních knihoven *„strukturované informace, které popisují, vysvětlují, lokalizují nebo jinak usnadňují vyhledávání, používání nebo správu informačního zdroje“* a zároveň *„data, spojovaná s informačním systémem nebo informačním objektem za účelem popisu, administrace, správy právních požadavků a technických funkcionalit, jejich použití a využití“* (ZENG a QIN, 2016).

V kontextu rámce OAIS metadata napomáhají k porozumění datům, osvětlují jejich původ, kontext a vymezení a v neposlední řadě data, opatřená metadaty, tvoří komplexní informace.

Metadata, která slouží k interpretaci, pochopení a užití digitální informace, OAIS definuje jako **reprezentativní informace**, což jsou informace, které mapují datové objekty do srozumitelných konceptů. Datový objekt potom spolu s reprezentativní informací tvoří informační objekt. Specificky v prostředí digitální ochrany se pak vytváří také **obsahová informace**, což je sada informací nebo jejich částí, původně určená k digitální ochraně, obsahující jak informační objekt, který je složený z obsahových dat, tak z reprezentativní informace.

5.1 Druhy metadat

Obecně se v digitalizaci knihovních fondů používají metadatové formáty několika typů: popisné, strukturální, technické, administrativní a právní, přičemž každý z těchto typů popisuje některý aspekt uchovávaného digitálního objektu.

Norma OAIS metadata také dělí do několika obecných množin. První množina se nazývá **referenční informace** *(Reference information)*. Tyto informace se používají pro identifikaci obsahových informací. Obsahují mimo jiné identifikátory, které umožňují na informaci odkazovat i mimo prostředí archivu. Nejznámějším příkladem referenční informace je ISBN, jednoznačný a jedinečný trvalý identifikátor, který odkazuje vždy na jednu konkrétní knihu.

Další množinou jsou **informace o původu** *(Provenance information)*, které slouží k dokumentaci historie nakládání s obsahovou informací a veškerých změn od jejího vzniku. Tyto informace také zaznamenávají údaje o kurátorech. Digitální archiv by měl vést veškeré údaje o nakládání s obsahovými informacemi od okamžiku, kdy přijdou do modulu Ingest.

Kontextuální informace *(Context information)* pak mapují vztah mezi obsahovou informací a jejím prostředím, včetně důvodu jejího vzniku, a toho, jak souvisí s jinými obsahovými informacemi.

Informace o celistvosti *(Fixity information)* dokumentují, že je technicky vzato obsahová informace kompletní, její obsah nebyl změněn ani manipulován jinak, než bylo zdokumentováno. Informace o celistvosti jsou obvykle vytvářeny algoritmy, které vygenerují řetězec znaků (součet), do kterých jsou tyto informace vloženy. Každá změna obsahové informace pak vyžaduje vygenerování nového řetězce. Celistvost obsahové informace je považována za základní signifikantní vlastnost, která určuje autenticitu informace.

Informace o přístupových právech *(Access Rights Information)* identifikují omezení, vázaná na přístup k obsahové informaci a související právní rámec, licence, zpřístupňovací politiku a další informace, které souvisí s rozšiřováním obsahové informace k uživatelům archivu. Informace o přístupových právech bývají podrobně deklarovány v *Submission Agreement* mezi producentem a archivem.

Poslední množinou metadat, která se ale na rozdíl od těch vyjmenovaných výše zabývá informačním balíčkem jako celkem, a ne pouze informacemi, které jsou v něm obsažené, jsou **Informace o datovém balíčku** *(Packaging information)*. Informace o datovém balíčku jsou potřebné k rozklíčování jeho obsahu, jelikož jenom zřídkakdy se balíček sestává pouze z jednoho digitálního objektu. Zabalení jednotlivých typů datových objektů a informací do jednoho balíčku je také způsobem, jak vytvořit vztahy mezi jeho různými součástmi. Informace o datovém balíčku pak kromě praktických informací o obsahu slouží také k jeho přesnějšímu vyhledávání (CUBR et al., 2023).

Všechny typy metadat ve výše zmíněném výčtu slouží k přesnějšímu popisu archivního informačního balíčku, respektive digitálních objektů, které obsahuje. obsahové informace, spolu s podrobným popisem z jednotlivých množin ochranných popisných informací (PDI), charakterizují balíček AIP, který je zároveň odvozen z informací o balíčku, jenž jej současně identifikují, popisují a rozlišují od ostatních balíčků. Všechna tato metadata potom slouží k potvrzení autenticity a kompletnosti obsahu.

5.2 Přehled různých formátů

V této podkapitole představíme nejobvyklejší metadatové formáty, používané v oblasti digital preservation se zvláštním přihlédnutím k těm, které jsou aplikovány v projektu Národní digitální knihovny v České republice.

5.2.1 PREMIS

Standard PREMIS[36] 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, ale také interpretačních informací. Aktuální třetí verze PREMIS vyšla v roce 2015. Standard PREMIS obsahuje

36 https://www.loc.gov/standards/premis/

vlastní komplexní data model, terminologický slovník a podrobný text vysvětlující logiku a možnosti užití standardu v archivu. Standard obsahuje také vlastní sadu elementů pro popis vztahů entitami, které definuje – agentem, objektem a událostí (a právní deklarací), přičemž nejčastěji se v zápisech metadat setkáváme s prvními třemi. Standard klade důraz na to, aby jeho elementy byly implementovatelné, což znamená, že hodnoty většiny elementů musí být možné automatizovaně vyplňovat a zpracovávat archivem (PREMIS, 2015, s. 3). Tento cíl je pro většinu metadatových standardů obecný.[37]

Datový model v PREMIS definuje **čtyři základní entity**: objekt *(object)*, činitel *(agent)*, událost *(event)* a právní deklarace *(rights statement)*. **Objekt** se dále člení na čtyři úrovně:

a) intelektuální entita *(intellectual entity)* je „jednotlivý intelektuální nebo umělecký výtvor *(creation)*, který je považován za relevantní pro cílovou komunitu v kontextu digitální archivace“;
b) reprezentace *(representation)* neboli „množina souborů (včetně strukturálních metadat) potřebná pro úplnou reprodukci intelektuální entity“;
c) soubor *(file)* je „pojmenovaná a uspořádaná posloupnost bajtů, kterou dokáže rozeznat operační systém“ a která je uložena v určitém souborovém formátu;
d) bitový tok *(bitstream)* představuje „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 ve standardu PREMIS jde specificky o reprodukovaný informační obsah (tj. obsah, který může vnímat člověk).

Intelektuální entitu je možné podle modelu PREMIS také dále specifikovat podle úrovní abstrakce popsaných ve známém knihov-

37 Týká se to i metadat ve např. schématu MODS nebo DC, které obvykle vznikají konverzí bibliografických záznamů ve formátu MARC21.

nickém modelu FRBR. Model FRBR stanovuje tyto čtyři úrovně: dílo *(work)*, vyjádření *(expression)*, manifestace *(manifestation)* a exemplář *(unit)*.[38]

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 jsou popsány 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 podle CUBR et al., 2023). 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).

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 v rámci elementu <objectCharacteristicsExtension>. Pro digitalizáty knih je ale za tímto účelem obecně užíván spíše standard MIX. PREMIS dále obsahuje sekci signifikantních vlastností, která však není v praxi zatím příliš užívána, mimo jiné z toho důvodu, že pro jejich použití je potřeba přípravná analýza každého digitálního objektu či kategorie zdrojového dokumentu.

38 https://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf

5.2.2 METS

Standard METS[39] *(Metadata Encoding & Transmission Standard)* slouží primárně pro zaznamenání informací o zabalení modelu OAIS (tedy o zabalení balíčků SIP, AIP a DIP). Především umožňuje začlenění dalších metadatových schémat pro popis archivačních a interpretačních informací (a tím jejich identifikaci). Obecně lze říci, že formát METS je kosterním formátem pro tvorbu metadatových záznamů. Kromě interpretačních informací obsahuje sekci určenou pro zápis provenienčních informací (opět pomocí externího schématu) a zápis některých interpretačních informací (např. o chování objektu). V praxi se METS užívá zejména pro první funkci (záznam informací o zabalení) a také jako datový formát. Jeho sekce strukturálních map se využívá pro záznam informací o všech obrazových souborech (fyzická mapa) a jejich hierarchické posloupnosti (logická mapa). Tyto informace jsou klíčové ke správnému zobrazení dokumentu v uživatelské aplikaci digitální knihovny.

Dále obsahuje část <fileSec>, která umožňuje výčet všech souborů v SIP balíku, z nichž se výsledný objekt CDO skládá a sekci <structLink>, která odkazuje na všechny objekty, které jsou v balíčku uloženy.

Tyto informace tedy v konečném důsledku netvoří strukturální interpretační informace, ale vlastní datovou součást digitalizátu knihy, bez níž by objekt CDO nebyl úplný. V praxi je METS 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). V české praxi se formát METS kromě PREMIS doplňuje také o níže popsané formáty MODS a DC (případně AES57 pro zvukové dokumenty) a RightsMD.

5.2.3 MODS

MODS *(Metadata Object Description Schema)*[40] je metadatový standard široce užívaný pro zápis deskriptivních metadat. Jde o výsledek pro-

39 https://www.loc.gov/standards/mets/

40 https://www.loc.gov/standards/mods/

jektu zaměřeného na vývoj standardu pro popis jakéhokoliv typu dokumentu a správu digitálních objektů v jazyce XML, který vedlo oddělení Network Development and MARC Standards Office, jež je součástí Kongresové knihovny. MODS vychází z katalogizačního standardu MARC 21, je ale jednodušší a snadno čitelný pro lidského uživatele. První verze MODS 1.2 byla zveřejněna v lednu 2002, aktuální verze 3.8 vyšla v září 2022. Rozvoj standardu stále řídí Network Development and MARC Standards Office ve spolupráci s mezinárodní komunitou vývojářů, přičemž návrhy na vylepšení a nové části standardu probíhají i pomocí celosvětové emailové konference (CUBR et al., 2023).

Vzhledem k historii vzniku umožňuje MODS téměř plnou konverzi záznamů v MARC 21 do metadatového zápisu v MODS pomocí převodní šablony MARCXML s minimální ztrátou informací. Formáty se však nepřevádí v poměru 1:1, jelikož slovník MODS je vyjádřen slovními značkami (elementy), na rozdíl od číselných značek MARC. Sada elementů MODS *(MODS Element Set)* ovšem umožňuje i vytváření kompletních originálních záznamů, nikoliv pouze konverze z existujících katalogizačních záznamů. Provázanost s MARC 21 je pak i v rozdílném vytváření záznamů a užitých hodnot v návaznosti na katalogizační pravidla, ve kterých byla zpracována předloha (AACR2 nebo RDA).

Elementová sada MODS obsahuje dvacet kontejnerových elementů *(top elements)*, které jsou zpravidla dále každý doplněny sadou vlastních podřízených elementů *(subelements)*. Každý element může být specifikován atributy, které konkretizují typ vyplněné hodnoty. Pokud se MODS použije v kombinaci s METS, je možné zápis rozčlenit do hierarchických úrovní, odpovídajících např. vnitřnímu členění dokumentu. Schéma MODS také počítá s doplněním elementů ze sad jiných standardů pomocí kontejnerového top elementu <mods:extension>, díky čemuž je možné v rámci jednoho zápisu popsat i specifické dokumenty nebo specifické informace (např. technického rázu), pro které MODS nemá ve vlastním element setu vhodné vyjádření.

5.2.4 Dublin Core

Dublin Core Metadata element set (také uváděný pod zkratkou DC)[41] vznikl původně jako soupis patnácti klíčových vlastností, kterými lze popsat libovolný digitální objekt včetně webových stránek. Těmito klíčovými vlastnostmi byly přispěvatel *(contributor)*, pokrytí/rozsah *(coverage)*, tvůrce *(creator)*, datum *(date)*, popis *(description)*, formát *(format)*, identifikátor *(identifier)*, jazyk *(language)*, vztah *(relation)*, vydavatel *(publisher)*, práva *(rights)*, zdroj *(source)*, předmět *(subject)*, název *(title)* a typ *(type)*.

Název standardu je odvozen od města Dublin ve státě Ohio, ve kterém se v roce 1995 konal OCLC/NCSA Metadata Workshop, na kterém bylo schéma vytvořeno.

Dublin Core (DC) byl formálně standardizován normami ISO 15836, ANSI/NISO Z39.85, a IETF RFC 5013. Jako metadatový standard je uznáván od roku 2002, kdy vznikla formální dokumentace DCMI Metadata Terms.[42] Standard v současné době spravuje iniciativa Dublin Core Metadata Initiative, která funguje na principu placeného členství.

Standard DC bylo původně možné rozdělit do dvou verzí. Základní sadu Jednoduchého DC *(Simple DC)* o patnácti elementech doplňuje rozšíření, tzv. Kvalifikované DC *(Qualified DC)*. Kvalifikovaná verze obsahuje navíc tři další elementy – *audience*, *provenance* a *rightsHolder*. Od roku 2012 byly tyto dvě verze sjednoceny do jednotného slovníku DCMI Metadata Terms.

Pomocí elementů DC lze univerzálně popsat široké množství digitálních objektů od textových přes obrazové, zvukové, audiovizuální až po webové stránky. Kromě popisu lze formát využít také jako klíč k propojení jiných metadatových standardů, respektive jejich elementových sad. Na rozdíl od jiných metadatových standardů nemá předepsaný syntax ani hierarchii mezi elementy. Díky své jednoduchosti slouží jako nástroj interoperability mezi standardy např. v oblasti linked data nebo v rámci sémantického webu. Kromě toho byl Dublin Core v minulosti inspirací pro vznik jiných popisných

41 https://www.dublincore.org/

42 https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

formátů, například níže zmíněného VRA CORE. Některé metadatové standardy (konkrétně MODS) mají pro konverzi z jednoho formátu do druhého vlastní mapování. I v současné době je přes svoji jednoduchost a nekomplikovanost (nebo právě díky ní) stále velmi široce využíván jako základní formát (zejména) bibliografického popisu v metadatech.

5.2.5 MIX

Standard MIX *(Metadata for Images in XML Standard)*[43] 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, 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. V současnosti se používá ve verzi 2.0 z roku 2008.

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 elementy jedné její sekce *(Change History)* jsou určeny pro záznam provenienčních informací z doby produkce (pro záznam generací dat vzniklých při vytváření finálních produkčních dat i užitých procesů). V praxi se však za tímto účelem užívají spíše elementy standardu PREMIS, ačkoliv ten je primárně určen pro záznamy operací v archivu, nikoliv pro digitalizaci (CUBR et al., 2023).

43 https://www.loc.gov/standards/mix/

5.2.6 ALTO

Standard ALTO *(Analyzed Layout and Text Object XML Schema)*[44] je primárně datový formát zaznamenávající text získaný procesem OCR *(Optical Character Recognition)* a jeho souřadnicové umístění vzhledem k obrazu, ale také on obsahuje některé metadatové prvky (např. informace o obrazovém zdroji pro OCR).

ALTO byl původně vyvinut pracovní skupinou METAe (The Metadata Engine Project) jakožto doplněk k formátu METS. Zatímco účelem METS je strukturování popisu dokumentu a jeho hierarchie (a z toho vyplývajících vztahů mezi jeho částmi), úkolem ALTO je vytvořit pomyslnou mapu vztahů a hierarchie pro každou jednotlivou stranu digitalizátu. Užití formátu je podmíněno využíváním technologie OCR, která dokáže pomocí neuronových sítí identifikovat polohu jednotlivých objektů a jejich řazení v naskenovaném obrazovém souboru a následně je převést do strojově čitelné podoby (CUBR et al., 2023).

Metadatová část standardu ALTO se dělí do tří sekcí (ALTO Principles, 2016): <Description> pro popis samotného souboru ALTO a jeho vzniku, <Styles> pro styly textu a jejich definici a nakonec <Layout>, který obsahuje informace o obsahu, rozděleném podle jednotlivých stran (<Pages>).

5.2.7 VRA CORE

VRA Core *(Visual Resources Association)*[45] je metadatový standard, užívaný v oblasti digitalizace umění a muzejních sbírek. Na rozdíl od Dublin Core a MODS není navržen jako obecné metadatové schéma, ale je designovaný na míru potřebám popisu obrazových děl a jejich muzejních předmětů a architektury v obecném smyslu, včetně jejich reprezentací (MILLER, 2022, s. 269). V naší práci jej krátce popíšeme jako jediný metadatový formát vhodný pro digitalizaci trojrozměrných předmětů v muzeích.

Popis muzejních sbírek má z velké části jiné potřeby, než jaké má popis sbírek v jiných paměťových institucích. Až do nedávné doby

44 https://www.loc.gov/alto/

45 https://www.loc.gov/standards/vracore/

muzea běžně nemusela řešit standardizaci popisu předmětů, jelikož s jinými institucemi nepotřebovala sdílet ani tento popis, natož pak metadata (tak jako například knihovny sdílí své katalogizační záznamy a v současné době i metadatový popis digitalizátů). Muzea také na rozdíl od jiných paměťových institucí typu knihovna či archiv mají ve svých sbírkách mnohem variabilnější škálu předmětů (nebo fyzických nosičů), od obrazů přes zoologické preparáty až po textilie, předměty z různých druhů materiálů a botanický materiál. Z tohoto důvodu mají muzea širší požadavky na popis svých objektů, jelikož kromě standardních informací potřebují zaznamenat například materiál předmětu, techniku vyhotovení, časové zařazení a historický umělecký sloh, historii vlastnictví a další údaje.

Jednotlivé verze formátu VRA CORE se od sebe v mnohém zásadně liší, proto je ve stručnosti popíšeme odděleně. **VRA CORE verze 3.0** je stavěn pouze na míru vizuálním objektům. Stejně jako Dublin Core je postaven na úrovni kategorií (elementů) homogenně, elementy nemají hierarchickou strukturu. S Dublin Core sdílí tento formát i další vlastnosti: poměrně stručnou elementovou sadu, opakovatelnost všech elementů, kvalifikátory a také inherentní mapování do Dublin Core. Klasickou strukturu formátu XML adaptoval tento standard až v nejnovější verzi **VRA CORE 4.0** (MILLER, 2022, s. 274). V této verzi se VRA podobá více MODS, elementy mají hierarchickou strukturu, subelementy a atributy, zapracoval též kontrolované slovníky (např. od Gettyho institutu), jejichž hodnoty jsou vyjádřeny číslem refid. Současnou verzi tvoří devatenáct top elementů, z nichž standard vyžaduje minimální záznam o třech z nich: <work> (dílo), <image> (vizuální reprezentace) a <collection> (sbírka).

5.2.8 AES57

Standard AES57 *(Audio Engineering Society)* poskytuje metadatový popis pro digitální a analogové formáty zvukových nahrávek pro účely digitální ochrany.[46] Umožňuje popsat nejen digitální zvukový soubor, ale i analogový nosič, neposkytuje však prostor pro zápis informací o okolnostech vzniku digitálního souboru, tak jak to známe ze sché-

46 https://www.aes.org/publications/standards/search.cfm?docID=84

matu MIX (BEŇAČKOVÁ et al., 2020a, s. 55). První snahy o standardizaci zápisu informací o zvukových nahrávkách začaly již na konci sedmdesátých let v pracovní skupině AES Standards Committee ve Spojených státech amerických. Standard AES57 byl potom oficiálně vydán v roce 2011 jako náhrada za starší AudioMD (OSTRÁKOVÁ a ŠÍR, 2017, s. 14). Standard je de facto kontrolovaným slovníkem pro popis zejména technických metadat. Umožňuje zápis základních informací o dané nahrávce, jako např. velikost, formát, délka, bitová frekvence (data o souboru) nebo typ komprese, kodek, počet kanálů, bitová hloubka nebo vzorkovací frekvence.

5.2.9 PBCore

Toto metadatové schéma bylo vytvořeno speciálně pro popis zvukových a audiovizuálních dokumentů[47]. Vzniklo specificky pro zajištění potřeb pro popis fondů rozhlasových a televizních stanic. Dnes schéma PBCore spravuje organizace American Archive of Public Broadcasting, která jej od roku 2013 používá pro popis svých dokumentů. Schéma PBCore vychází ze schématu Dublin Core a je vyjádřené v jazyce XML (OSTRÁKOVÁ a ŠÍR, 2017, s. 13). Stejně jako v případě AES57 a AudioMD v něm také lze zachytit technické údaje o nahrávce, ale ne v tak širokém rozsahu, jako u AES57. Jelikož zápis vychází z Dublin Core, není metadatový záznam v tomto schématu strukturovaný.

5.2.10 Právní metadata

Právní metadata tvoří celou skupinu metadatových formátů, které slouží k zápisu informací, souvisejících s licenčními podmínkami, vymezujícími vlastnictví, užívání a šíření dokumentů v digitální knihovně.

Nejspíš prvním metadatovým schématem, které obsahovalo elementy k zachycení informací ohledně právního statutu digitalizátu, byl jednoduchý **Dublin Core** (DC). Ten obsahoval element <rights>, jehož obsahem mělo být vyjádření ohledně různých vlastnických práv spojených se zdrojem (POMERANTZ, 2015, s. 107). Po doplnění DC o kvalifikátory bylo možné tento popis doplnit o *licence* (tj. právní

47 https://pbcore.org/

dokument), *rightsHolder* (jedinec nebo organizace) a *accessRights* (tj. jaká práva má *rightsHolder* k přístupu ke zdrojovému dokumentu, což se zakládá na předpokládané licenci). Vzhledem k tomu, jak obecné ale elementy pro DC jsou, je zřejmé, že pouze takový popis nemůže být dostatečný.

Mnohem rozšířenějším schématem je **Creative Commons Rights Expression Language** (CC REL). Creative Commons je projekt, který umožňuje sdílení a šíření uměleckých děl pomocí vlastních standardizovaných licencí, které jejich tvůrcům umožňují selektivně vymezit různá práva vedle samotného označení copyrightu[48]. Specifikace pro CC REL identifikuje ve svých licencích entity a vztahy mezi nimi, které jsou v copyrightu zahrnuty; obecně je lze rozdělit do skupin *vlastnosti díla* a *vlastnosti licence k dílu*. Dílo zahrnuje informace o názvu, zdroji a typu díla. Vlastnosti licence potom informují o povoleních, zákazech, požadavcích, legálním rámci, pod který licence spadá, a <legalCode> (což je samotný text licence). Na rozdíl od většiny právních metadat jsou Creative Commons v současné době velmi rozšířené napříč digitálními zdroji na internetu a široce přijímané komunitou tvůrců digitálních děl. Jedna z největších kulturních institucí na světě, která využívá CC, je v současné době digitální knihovna Europeana.

Standard **CopyrightMD** [49] je výsledkem pracovní skupiny při California Digital Library. Vznikl přímo na míru tehdejším potřebám pro dlouhodobé uchovávání dokumentů v dané instituci. V době svého vzniku (na začátku nultých let 21. století) byl mezi prvními normami, které identifikovaly klíčové informace k zachycení v metadatech, spojené s řádným užíváním *(fair use)* licencovaného obsahu. V současné době je již poněkud zastaralý (poslední aktualizace proběhla ve verzi 0.91 roku 2009), ale jako výhodu lze označit plnou kompatibilitu s METS, kam může být zápis právních metadat CopyrightMD vložen formou rozšíření.

V současné době je nejkomplexnějším formátem pro záznam právních metadat zřejmě **PREMIS:RIGHTS**, který vychází stejně jako

48 https://creativecommons.org/about/

49 https://cdlib.org/groups/rights-management-group-copyrightmd/

jeho základní verze z triády objekt–událost–činitel, a je tak schopen umožnit záznam informací o popisovaném dokumentu z několika hledisek a nadto reflektovat a zaznamenávat změny v licenci v čase.

5.3 Důležitost standardizace

V paměťových institucích, které přijímají do svých digitálních archivů zdigitalizované dokumenty a jiné digitální objekty, existuje různá míra požadavků na podobu a náležitosti metadatového popisu digitalizátu. Paměťové instituce bývají různě striktní také v případě přijímaných souborových formátů.

5.3.1 Důvody standardizace

Důvodů pro standardizaci je celá řada. Standardizace přispívá k určité homogenizaci přijímaných dat, což zjednodušuje jejich správu a ochranu. Standardizací kvality přijímaných dat zajišťuje digitální archiv úroveň věrnosti digitalizátu vůči originálu (přičemž ideální kvalita digitalizátu v maximální možné míře zachovává signifikantní vlastnosti předlohy). Kvalitní digitalizát je zároveň důležitým předpokladem pro archivy, které své sbírky zpřístupňují uživatelům a chtějí jim poskytnout co nejlepší výsledek k prezentaci. Standardizace (nebo omezení) přijímaných typů souborových formátů potenciálně zmenšuje rizika pro bitovou ochranu souborů.

5.3.2 Výhody a nevýhody standardizace

Liberální přístup umožňuje přijímat širší množinu zdigitalizovaných dokumentů v mnoha a mnoha typech souborových formátů a s různě obsáhlým metadatovým popisem ať už v technických nebo popisných metadatech. Nízký práh požadavků je vstřícnější vůči producentům SIP balíčků, jelikož jim klade menší překážky k zarchivování balíčku ve vybrané paměťové instituci, umožňuje jejich masivnější výrobu, představuje menší finanční a časovou zátěž, případně menší nároky na hardwarové a softwarové vybavení digitalizačního studia díky liberálnější politice ohledně technické kvality digitalizátu. Producent teoreticky nemusí provádět formátovou konverzi do souborových

formátů předepsaných archivem, balíčky mohou obsahovat pouze omezený nebo základní metadatový popis. V případě popisných metadat, která vznikají konverzí z katalogizačních lístků, není potřeba dělat jejich rekatalogizaci kvůli nedostatečnému zápisu. Cenou za tento liberální přístup jsou ale mnohdy velmi omezené možnosti bitové a zejména logické ochrany dokumentů a obtížnější správa svěřených dat. Pokud archiv neomezuje producenty ve velmi úzkém profilu přijímaných souborových dat, zvyšuje riziko, že např. z technických příčin nebude možné v budoucnosti obsah digitalizátů číst.

Striktní přístup pak producentům SIP balíčků předepisuje souborové formáty, rozsah metadatového popisu pro technická a/nebo bibliografická metadata. Příručky pro producenty SIP balíčků pak mnohdy obsahují i předepsané hodnoty pro jednotlivá pole, které musí splňovat výběr z kontrolovaných slovníků. Ty si digitální archiv buď spravuje sám, nebo je přejímá od jiných (paměťových) institucí. V závislosti na striktnosti předpisů ukládající paměťová instituce vytváří různě vysoký práh pro přijímání dat od producentů. Čím jsou požadavky pro příjem SIP balíčků vyšší, tím náročnější je jejich produkce na čas, finanční prostředky i technologické vybavení producenta. Vysoké požadavky na přijímaná data mohou negativně ovlivňovat množství archivovaných dokumentů, omezovat tvorbu digitalizátů a odrazovat producenty od k ukládání dat v archivu, který vyžaduje příliš striktní podobu obsahu SIP balíčku. Výběr souborových a metadatových formátů musí být mnohem pečlivější vzhledem k vysokému riziku, že data nebude možné při ztrátě technické podpory v budoucnosti dále číst. Pozitivem striktního přístupu a standardizované (homogenní) povahy ukládaných dat je jejich snadnější správa, bitová a logická ochrana.

5.4 Metadatový profil NDK a Standard NDK

Národní knihovna pro potřeby projektu Národní digitální knihovny vytvořila vlastní aplikační profil, založený na mezinárodních metadatových standardech, spravovaných americkou Kongresovou knihovnou.

Národní knihovna byla spolu s Moravskou zemskou knihovnou v Brně v letech 2010–2014 řešitelem projektu Vytvoření Národní digitální knihovny (projekt NDK), který byl financován z Integrovaného operačního programu EU programového období 2007–2014.[50] Národní knihovna 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. Prvním záměrem pro vytvoření Standardu NDK bylo formulovat specifikace pro digitalizáty tištěných monografií a periodik, vytvářené v rámci projektu NDK v Národní knihovně a Moravské zemské knihovně v Brně. Specifikace měly zajistit výrobu SIP balíčků tak, aby byl objekt CDO již v archivačním formátu a LTP uložiště NK ČR nemuselo provádět další formátovou normalizaci. Metadatový profil byl navržen tak, aby umožnil zaznamenání všech důležitých informací, které popisují proces produkce digitalizátů.

V současné době má vývoj Standardu NDK na starost Odbor novodobých digitálních sbírek (dříve Odbor digitálních fondů), konkrétně Oddělení standardů digitálních sbírek (dříve Oddělení pro standardy). Na vývoji se spolupodílí odborná komunita (která je zároveň uživatelem Standardu NDK), zastoupená v několika pracovních skupinách, podléhajících Formátovému výboru, který působí jakožto poradní orgán Národní knihovny. Návrhy na změny a další rozvoj Standardu jsou iniciovány jak vnitřními potřebami Národní knihovny, tak zvenčí ze strany odborné komunity.

Byť je Standard NDK primárně určen pro digitalizační linku NDK, je v praxi využíván také pro LTP systém ArcLib a pro tvorbu produkčních dat v systému ProArc. Kromě toho slouží také v digitální knihovně Kramerius, která funguje jako uživatelské prostředí pro zpřístupňování digitalizátů a na jejímž vývoji se Národní knihovna také podílí.[51]

50 Podrobněji se vzniku digitálního repozitáře pro Národní knihovnu a projektu NDK věnuje kapitola Dlouhodobé uchovávání digitalizátů v českých knihovnách.

51 Národnímu aplikačnímu profilu se dále věnuje část kapitoly Dlouhodobé uchovávání digitalizátů v českých knihovnách.

V praxi je metadatový aplikační profil souborem vybraných metadatových formátů, ze kterých si instituce extrahuje relevantní části elementové sady a následně je spojí do vlastního funkčního schématu, který je uzpůsoben potřebám a možnostem konkrétní aplikace. Podmínkou je, že aplikační profil musí fungovat v rámci pravidel, nastavených v původních (většinou mezinárodních) metadatových standardech, musí být navzájem interoperabilní (a zapsatelný v rámci formátu XML).

Metadatový aplikační profil Národní knihovny zahrnuje takové formáty, které umožňují zápis informací ze všech základních okruhů metadatového popisu. Konkrétní skladba element setu byla podřízena české katalogizační praxi a jejím standardům (v případě bibliografických metadat) a potřebě zaznamenat veškeré klíčové informace o procesu vzniku digitalizátů, jejich úprav a procesů souvisejících s tvorbou SIP balíčků. Zároveň jsou veškeré Definice metadatových formátů (jak se oficiálně nazývají normy, ze kterých Standard NDK sestává) uzpůsobeny na míru specifikům konkrétního typu dokumentu, ze kterých se digitalizát vytváří, a specifikům českého knihovního prostředí. V současné době (polovina roku 2023) Národní knihovna spravuje Definice metadatových formátů pro digitalizáty monografií, periodik, zvukových dokumentů (gramofonových desek a fonoválečků) a e-born dokumenty (e-monografie, e-periodika, skládaná periodika).

5.4.1 Složení SIP balíčku

Standard NDK předepisuje přesnou podobu a skladbu SIP balíčku.

Typický SIP balíček pro textový digitalizát obsahuje:

- **soubor info**, který popisuje jednotlivé komponenty SIP balíčku a základní informace o jeho vzniku;
- **soubor hlavní mets**, který ve formátu METS, MODS a DC obsahuje veškerá popisná a strukturální metadata, a slouží jakožto klíčový dokument k rozpoznání a řazení digitálních objektů v balíčku;
- **složky s master kopiemi** digitálních objektů (v archivní kvalitě) a **uživatelskými kopiemi** (v komprimačním formátu) digitálních objektů pro účely zobrazování v digitální knihovně;

- **složku amdsec** pro tzv. vedlejší mets soubory pro každý digitální objekt, které obsahují technická a administrativní metadata ve formátech PREMIS a MIX;
- **složku txt** pro přepis textů do jednoduchých textových souborů pro každou digitalizovanou stranu;
- **složku alto** pro OCR soubory pro každou stranu;
- **soubor s kontrolním součtem** ve formátu MD5 pro veškerý obsah SIP balíčku mimo souboru info.

V případě zvukových digitalizátů je v SIP balíčku obsažen navíc **soubor catalog_entry** pro původní katalogizační lístek v xml formátu a složky pro master kopii audio souborů a složka pro uživatelskou kopii audio souborů. V administrativních vedlejších souborech mets se kromě výše uvedených užívá i formát AES 57 pro zápis vlastností původního zvukového souboru.

Po obsahové stránce je nejméně rozsáhlý SIP balíček pro e-born dokumenty, který je od předchozích zároveň velmi odlišný. Neobsahuje totiž složky pro master a uživatelskou kopii digitálních objektů, ale pouze jednu **složku nazvanou original**. V souboru hlavní mets jsou obsažena jak popisná metadata ve formátech METS, MODS a DC, tak administrativní, technická a právní metadata ve formátech PREMIS, copyrightMD a ndkTech. Soubory info a MD5 zůstávají stejné jako u předchozích typů dokumentů, chybí složky pro OCR a txt soubory (CUBR et al., 2023).

6 Dlouhodobá archivace dat a bitová ochrana

Jedním z primárních úkolů paměťové instituce v oblasti digitalizace je zachování jejího digitálního obsahu po nejdelší možné období. Aby se předešlo potenciální ztrátě uložených dat, je nutné se zabývat jejich bitovou ochranou. Bitová ochrana je dlouhodobým procesem IT-infrastruktury a představuje komplexní problematiku hardwarových i softwarových řešení podílejících se na zachování datové integrity a autenticity dat. Lze ji považovat za jeden z elementárních předpokladů pro úspěšnost dlouhodobého uchovávání, neboť zajišťuje neporušitelnost a trvalý přístup k uloženým datům (LOC, 2022a).[52]

Z pohledu definice datového objektu (THIBODEAU, 2002) řeší bitová ochrana fyzickou rovinu uchovávané informace, a sice uchování datového toku bez ohledu na jeho obsah a význam. Tatáž informace může tudíž být uložena na libovolných paměťových médiích různého formátování prostřednictvím rozdílných technologií, aniž by změnila svůj význam. Smysluplnost a interpretace datového toku – například formátové strategie, kódování či archivační metadata – spadají naopak již pod problematiku logické ochrany.

6.1 Rizika pro dlouhodobou archivaci dat

Žádné úložné médium neposkytuje archivovaným datům ochranu po neomezenou dobu. Životnost archivovaných dat se odvíjí od mnoha faktorů. Data mohou být ohrožena lidským zásahem, například kybernetickým útokem, sabotáží či neúmyslnou chybou (například smazáním, znehodnocením dat) ze strany zaměstnance dané instituce. Kromě vyjmenovaných incidentů způsobených lidským faktorem

52 Viz definice bitové ochrany na stránkách LoC: https://www.loc.gov/programs/digital-collections-management/digital-formats/bit-level-preservation-and-long-term-usability/

jsou data uložená v paměťových institucích ohrožena také činiteli technického charakteru. Spolehlivost IT komponent je vyjádřena statistickou veličinou MTBF *(Mean Time Between Failures)*, která určuje průměrnou dobu mezi jednotlivými chybovými stavy. Veličina MTBF bývá zpravidla uváděna výrobcem dotyčného zařízení a předpokládá jeho užívání za normálních podmínek (SPEAKS, 2005, s. 2). V praxi bitové ochrany se jedná o nevyhnutelnou degradaci úložných médií a příslušných komponent. Kromě toho je nutné také brát v potaz zastarávání používaných technologií úložných médií, čtecích zařízení a používaného softwaru. V neposlední řadě je nutné zmínit také hrozby širšího kontextu, například války, politická rozhodnutí a živelné katastrofy (ROSENTHAL et al., 2005).

Altman a Landau (2020, s. 3–4) klasifikují hrozby pro dokumenty uložené v digitálním archivu do čtyř úrovní:

- **Selhání diskového sektoru** *(Disk Sector Failures)*: chyby malého rozsahu nacházející se na disku mohou zapříčinit částečnou nebo úplnou ztrátu kopie dokumentu
- **Nežádoucí vlivy okolního prostředí** *(Environmental Glitches)*: množství chyb vyskytujících se na úložném médiu je ovlivňováno vnějšími faktory, například selháním chlazení nebo elektrickými výboji
- **Selhání serverů** *(Server Failures)*: omezená životnost serverů a rizika institucionální úrovně, například vypovězení smlouvy poskytovatelem serverových služeb či ransomwarový útok
- **Dopady rozsáhlejšího charakteru** *(Major Shocks)*: vnější vlivy ekonomického, přírodního a politického charakteru – povodně, války, státní převraty, zemětřesení aj.

6.2 Metody bitové ochrany

Na základě dosavadních zkušeností světových paměťových institucí vznikly v uplynulých dvou dekádách dokumenty, které vymezují rámec nutných předpokladů pro důvěryhodný provoz digitálního archivu a některé i podrobněji stanovují metody pro naplnění požadavků na bitovou ochranu a řešení výše zmíněných rizik. Patrně nejvýznam-

nějším mezinárodním dokumentem je standard Open Archival Information System (OAIS), jehož druhá verze z roku 2012 formuluje šest základních funkcí (příjem dat, správu úložné hierarchie, nahrazení datových nosičů, kontrolu chyb, obnovu po havárii a poskytnutí dat) archivního úložiště *(Archival Storage)* v rámci funkčního modelu (OAIS, 2012, s. 4–8, 4–9). Model OAIS si klade za cíl vymezit základní rámec kritérií, které by digitální archiv měl splňovat. Nepopisuje však, jakým způsobem mají být zmiňované požadavky naplněny, neboť v každé instituci je volba konkrétních postupů a technologií ovlivněna mnoha faktory, především bezpečnostního, znalostního, ekonomického a právního charakteru (OAIS, 2012, s. 1–2).[53]

Konkrétněji se v souvislosti s realizací bitové ochrany vyjadřují novější dokumenty definující požadavky na digitální archiv. Jedním z nich je průběžně aktualizovaná sada postupů *Levels of Digital Preservation* sestavená sdružením National Digital Stewardship Alliance (NDSA) pod záštitou Kongresové knihovny (NDSA, 2019). Dokument *Levels of Digital Preservation* poskytuje přehled používaných postupů jednak pro jednotlivé okruhy problémů správy digitálního archivu a jednak pro požadovaný stupeň ochrany. Kromě *Levels of Digital Preservation* jsou průběžně doplňovány požadavky také v dokumentu *Preservation Storage Criteria* (PSC), jenž byl prvně prezentován na konferenci iPRES 2016 (SCHAEFER et al., 2018).

Realizace klíčových metod pro bitovou ochranu je motivována nutností řízení bezpečnosti digitálního archivu. Za východisko v řízení bezpečnosti lze považovat široce rozšířenou sadu doporučení osvědčených v IT infrastruktuře, nazývaných Information Technology Infrastructure Library (ITIL). Zaměření ITILu je velmi široké a lze jej aplikovat na jakoukoli instituci pracující s IT technologiemi. Z hlavních přínosných bodů ITILu pro řízení bezpečnosti paměťové instituce lze zmínit *procesní řízení*, v němž každý proces musí být náležitě popsán, a zároveň každému procesu musí být přiřazen jeho správce či pojmenovaná role, jíž je správce součástí; dále *řízení rizik*, v rámci něhož musí být každé riziko identifikováno, vyhodnoceno

53 Rezignace na popis realizace požadavků je v OAIS (2012, s. 1–2) výslovně zdůrazněna: „This reference model does not specify a design or an implementation. Actual implementations may group or break out functionality differently."

a ošetřeno (ITIL Foundation, 2019, s. 133) (viz popsaná rizika výše), a v neposlední řadě *řízení incidentů* a *disaster-recovery plan* (ITIL Foundation, 2019, s. 163–164). Vedle obecně pojatého modelu ITIL lze vycházet rovněž z dokumentu *Průvodce pro vytvoření institucionální politiky digitálního dlouhodobého úložiště,* pocházejícího z projektu *NESTOR,* zaměřeného konkrétněji na bezpečnost digitálního archivu – tudíž i na bitovou ochranu jeho obsahu (NESTOR, 2014). Kromě hlavních cílů, jimiž jsou zachování datové integrity, autenticity, úplnosti, čitelnosti, dohledatelnosti dat v archivu a důvěryhodnosti (NESTOR, 2014, s. 10), jsou zmíněny i klíčové problémy technické infrastruktury paměťové instituce s důrazem na zajištění stabilních, transparentních (tj. dostatečně popsaných) a bezpečnostně vyhovujících politik (NESTOR, 2014, s. 12–13).

Druhým společným aspektem při výběru konkrétních řešení je ekonomická udržitelnost digitálního archivu. Provoz infrastruktury si klade různé finanční nároky v závislosti na datovém objemu digitálního archivu, výpočetní náročnosti a užitých technologiích, vázaných licencemi. Z těchto důvodů je nutné brát na zřetel modularitu digitálního archivu (tj. jeho snadnou přizpůsobitelnost aktuálně zpracovávaným a uchovávaným datovým objemům) a nezávislost na třetích stranách při jeho provozu a správě.

Níže jsou pojednány základní a nejčastěji užívané metody pro docílení bitové ochrany a splnění požadavků v pojednaných dokumentech.

6.2.1 Replikace

Vlastní-li paměťová instituce pouze jednu kopii dat, existuje v případě poškození jediné kopie významné riziko jejich nenávratné ztráty. Data je z toho důvodu nutné *replikovat,* a sice ukládat ve více kopiích na různá úložná média nacházející se v geograficky oddělených lokalitách, aby v případě selhání či ohrožení jednoho úložiště bylo možné data zrekonstruovat. Replikace zahrnuje proces redundance (opakování objektu, jeho zkopírování), zároveň však zaručuje, že replika (zkopírovaný objekt) bude se svými dalšími kopiemi stále identická (JOSHI, 2019). K definici a náplni replikace se vyjadřují výše zmiňované dokumenty s požadavky na digitální archivaci. Model

OAIS například považuje replikaci za jeden z typů migrace, při němž nedochází ke změně bitů, avšak může vyžadovat změny v mapování infrastruktury informačního archivu (OAIS, 2012, s. 5–4, s. 5–5). Dokument *Levels of Digital Preservation* podrobněji stanovuje požadavky na replikaci dat ve čtyřech stupních ochrany; první stupeň *(Know your content)* předpokládá minimálně dvě kompletní kopie v oddělených lokalitách, dokumentaci úložných médií a stabilní úložiště; druhý stupeň *(Protect your content)* požaduje alespoň tři kopie, z nichž jedna kopie se nachází v geograficky oddělené lokalitě; v rámci třetího stupně *(Monitor your content)* je vyžadována nejméně jedna kopie, která se od ostatních nachází v geograficky oddělené lokalitě, a jedna kopie na jiném typu média; nejvyšší čtvrtý stupeň ochrany *(Sustain your content)* vyžaduje nejméně tři kopie v geograficky oddělených lokalitách s rozdílnými hrozbami, diverzifikaci úložišť a plán pro předcházení zastarávání hardwaru, softwaru a úložných médií (NDSA, 2019).

Větší množství replikací sice poskytuje vyšší úroveň bezpečnosti, avšak vyžaduje vyšší nároky na správu, rozpočet a představuje i významnější energetickou stopu (KINNAMAN a MUNSHOWER, 2022). Z toho důvodu vznikají konsorciální řešení požadavku na replikaci dat. Patrně nejvýznamnějším projektem je LOCKSS (Lots Of Copies Keep Stuff Safe) založený na principu mezi institucionálních sítí s cílem vytvořit decentralizovanou a distribuovanou síť, v níž každý její účastník (paměťová instituce) vlastní jednu kopii objektu. Jelikož LOCKSS je sítí decentralizovanou, každý z účastníků komunikuje s každým *(peer-to-peer)* a nikdo z účastněných nemá kontrolu nad všemi obsaženými kopiemi. Data jsou v síti LOCKSS pravidelně automaticky čtena, validována a opravována (LOCKSS, 2018b). Provoz sítě a procesů odpovědných za vzájemnou komunikaci je zajištěn open source softwarem – LOCKSS daemonem – který je spuštěn v každé zapojené instituci. Síť LOCKSS lze provozovat jak privátně, tak v rámci komunitních projektů, z nichž nejvýznamnějšími jsou Global LOCKSS a CLOCKSS (LOCKSS, 2018a).

6.2.1.1 Zrcadlení

Jednou z možností, jak realizovat replikaci na úrovni bitů, je zrcadlení bitového toku, jež umožňují disková pole RAID (viz podkapitola Decentralizované systémy úložišť 6.3.3). Výhodou zrcadlení jsou nižší nároky na výpočetní výkon, zejména je-li realizováno hardwarovými řadiči. Jeho nevýhodou je obsazení většího úložného prostoru.

6.2.1.2 Erasure Coding

Erasure coding je dopřednou opravou chyb (*Forward Error Correction*, FEC) a spočívá v rozdělení dat na menší buňky o počtu *k*, z nichž se pomocí matematické funkce vypočtou paritní buňky o počtu *m*. Fragmenty dat a paritní informace jsou následně uloženy na různých discích. V případě potřeby lze narušenou integritu dat opravit na základě paritní informace (SHENOY, 2020). Technologie erasure coding je používána ve filesystému Hadoop (Apache Hadoop, 2023b) i Ceph (Ceph Documentation, 2016b) a výpočty paritních informací implementuje například rovněž RAID 4 – 6. Zajištění datové integrity prostřednictvím erasure code lze ve srovnání se zrcadlením docílit s menším obsazením místa na discích, ovšem s nutností kódovat a dekódovat paritní bity narůstají nároky na výpočetní výkon.

6.2.2 Obnovení

Migrace je jednou ze strategií, která se v paměťových institucích odehrává jak na úrovni logické, tak i na úrovni bitové ochrany dat. Dělení migrace na logickou a bitovou rovinu je patrné již z referenčního modelu OAIS, jenž rozlišuje operace, při nichž dochází ke změně bitových posloupností (transformace, přebalení), a operace zachovávající bitovou posloupnost. V případě bitové ochrany se vedle výše zmíněné replikace jedná o *obnovení*, které nahrazuje původní úložné médium věrnou kopií, s níž úložiště pracuje nadále jako s původním médiem. Určitý požadavek na obnovení deklarují rovněž *Levels of Digital Preservation* v rámci čtvrté úrovně ochrany úložiště, když vyžadují „plán a prováděcí akce pro určení zastaralosti hardwaru, softwaru a úložných médií úložiště“ (NDSA, 2019). Obnovení se provádí zejména kvůli omezené životnosti úložných médií a každá paměťová instituce by měla pravidelnost jeho realizace stanovit v rámci řízení rizik na základě užívaných technologií.

6.2.3 Zálohování

Proces zálohování se od replikací liší tím, že kopírování neprobíhá v reálném čase, nýbrž v cyklicky stanovených termínech, mezi nimiž nejsou média, na kterých jsou zálohy uloženy, propojena se zbytkem infrastruktury. Vytváří tím možnost vrátit se ke stavu před nechtěným zápisem (například před kybernetickým útokem). Data, která jsou uložena na nepřepisovatelných médiích v dostatečném počtu replik, u nichž je pravidelně prováděna kontrola integrity uložených dat, zálohování nutně nevyžadují. Pro bitovou ochranu digitálního archivu je však nezbytné důsledně zálohovat jeho kritické součásti podléhající neustálým změnám souvisejícím se zpracováváním (importem a exportem) a ukládáním dat. Příkladem jsou katalogy distribuovaných úložišť (katalogové uzly systému iRODS či NameNody v úložišti Hadoop), jejichž ztráta by mohla zapříčinit, že uložená data již nepůjde dohledat.

6.2.4 Kontroly integrity

Integrita neboli celistvost (v anglosaské literatuře též *fixity*) dat znamená, že data nebyla mezi dvěma časovými body změněna (NOVAK, 2006). Kontrola integrity je jedním z klíčových požadavků mezinárodních doporučení a standardů týkajících se bezpečnosti a uchovávání dat. Z toho důvodu tvoří nedílnou součást auditních procesů a představuje stěžejní metodu k docílení bitové ochrany archivovaných dat a důvěryhodnosti digitálního archivu.

6.2.4.1 Požadavky na kontrolu integrity

Referenční model OAIS definuje v rámci funkčního modelu obecný požadavek na kontrolu chyb *(Error Checking)* v archivním úložišti, která má zajistit, že žádná součást AIP balíčku nacházející se na úložišti ani během vnitřních přenosů na úložišti nebyla porušena. Funkce kontroly chyb má dle OAIS zahrnovat notifikace všech hardwarových a softwarových incidentů a jejich zápis do logů. Kontrola se dle OAIS realizuje na základě informace o integritě *(Fixity Information)* AIP balíčku. Samotná informace o integritě má rovněž podléhat kontrole integrity (OAIS, 2012, s. 4–9).

Dokument *Levels of Digital Preservation* sestavený sdružením NDSA specifikuje požadavky na kontrolu integrity podrobněji ve čtyřech

úrovních. V prvním stupni kontroly se uvádí požadavek na antivirovou kontrolu, kontrolu informace o integritě, byla-li poskytnuta, a pokud nikoli, vytvoření takové informace. V druhé úrovni by měla být integrita kontrolována při každé manipulaci s digitálním obsahem, informace o ní musí být uložena separátně od obsahu a médium s originálními daty by mělo být opatřeno ochranou proti nechtěnému zápisu. Třetí úroveň požaduje kontrolu informace o integritě v pevně stanovených intervalech, dokumentaci provedených kontrol a jejich výsledků a na vyžádání schopnost provedení auditu informací o integritě obsahu. V nejvyšším stupni ochrany je vyžadována kontrola integrity při specifických událostech a v případě nutnosti oprava či nahrazení poškozeného obsahu (NDSA, 2019).

Dokument *Preservation Storage Criteria* stanovuje nejen periodické kontroly integrity každé kopie, nýbrž také kontroly napříč jednotlivými kopiemi (SCHAEFER et al., 2018).

6.2.4.2 Informace o integritě (Fixity Information), její získání a použití

Výše zmíněné požadavky se odvolávají na informaci o integritě, která je k poskytnutí výsledku kontroly klíčová, nedefinují však již konkrétní formát této informace. Informace o integritě jsou metadata přidružená k datovému obsahu, s nímž se manipuluje a jenž je uložen v digitálním archivu. Při kontrole jsou informace nezávisle přepočítány (opětovně vygenerovány) a porovnány s uloženými metadaty (NOVAK, 2006, s. 1).

Základní informací o integritě může být například velikost a počet souborů. Velikost souborů ani jejich počet ovšem – jsou-li tyto údaje uvedeny samostatně – nelze označit za zcela spolehlivé informace o integritě, neboť jsou velmi náchylné ke kolizi; velikost nezohledňuje změnu v bitovém toku při zachování jeho délky a počet souborů neuvažuje v potaz změnu obsahu. Společně se zmíněnými údaji se proto používají také kontrolní součty a kryptografické hašovací funkce poskytující vyšší úroveň zabezpečení. V následujícím přehledu jsou uvedeny nejčastěji užívané typy algoritmů:

- **CRC32** *(Cyclical Redundancy Check)* je kód s nízkou výpočetní zátěží, díky níž je používán pro detekci chyb během síťových přeno-

sů (De STEFANO et al., 2014, s. 5). Jeho výstupem je hash o délce 32 bitů. Jelikož se nejedná o kryptografickou funkci, lze jej do jisté míry možné použít „obousměrně“ pro opravu poškozených dat (STIGGE et al., 2006).

- **MD5** *(Message Digest Algorithm)* je paměťovými institucemi nejčastěji užívanou kryptografickou hašovací funkcí pro kontrolu integrity dat (BARNESS et al., 2018, s. 23). Její nároky na výpočetní výkon jsou nízké, liší se délkou vstupního bitového toku. Nehledě na délku vstupních bitů je výstupem funkce MD5 vždy hodnota o jednotné délce 128 bitů (De STEFANO et al., 2014, s. 5). Ačkoli kolize (tj. docílení stejného hashe pro odlišné posloupnosti bitů na vstupu) je méně pravděpodobná než v případě CRC32 (ARAVINDAN, 2014), v minulosti byl tento algoritmus prolomen a nelze jej používat pro bezpečnostní účely (KUZNETSOV, 2014).
- **SHA-1** *(secure hash algorithm 1)* je kryptografickou hašovací funkcí, jejímž výstupem je hash o velikosti 160 bitů. Z toho důvodu je výpočetně náročnější než funkce MD5. Jelikož byl tento algoritmus v minulosti prolomen, je nutné jej dnes považovat za zastaralý a nevhodný k praktickému využití (ROSENTHAL, 2017).
- **SHA-2** *(secure hash algorithm 2)* je kryptografickou hašovací funkcí, jejímž výstupem je hash o velikosti 256 nebo 512 bitů. Ačkoli tento algoritmus vyžaduje vysoké nároky na výpočetní výkon, poskytuje vysoce detailní digitální otisk. Zároveň jej lze označit za bezpečný, neboť doposud nebyl prolomen běžnými výpočetními prostředky (De STEFANO et al., 2014, s. 5).
- **SHA-3** *(secure hash algorithm 3)* je kryptografickou hašovací funkcí, která se od SHA-1 a SHA-2 liší použitím algoritmu KECCAK. SHA-3 lze obdobně jako jeho předchůdce SHA-2 použít pro bezpečnostní účely, protože doposud nebyl prolomen.
- Při volbě vhodného algoritmu pro vytvoření informace o integritě je nutné uvážit jeho nároky na výpočetní výkon na jedné straně a bezpečnost a náchylnost ke kolizím na straně druhé. Na ochranu vůči chybám technického (nezáměrného) charakteru postačuje MD5 funkce, která se z hlediska dlouhodobé praxe jeví jako vhodný kompromis mezi výkonem a odolností vůči kolizím. Jak bylo naznačeno výše, MD5 ani CRC32 neposkytují ochranu vůči záměrným

manipulacím s daty, kterou zaručuje SHA-2 a SHA-3 (ADDIS, 2020, s. 2–3). Tallmann (2021, s. 2–3) rozděluje v tomto ohledu kontrolu integrity do třech typů:

- **Transakční kontrola integrity** se provádí během manipulace s daty (při ingestu nebo replikaci) pro odhalení případných poškození. Pokud se jedná o manipulaci v rámci důvěryhodného systému, postačují k těmto účelům algoritmy CRC32 nebo MD5.
- **Autentizační kontrola integrity** prověřuje, zda nebyl soubor během delšího časového úseku záměrně či nezáměrně změněn. Vzhledem k bezpečnostním účelům kontroly a méně častým realizacím se doporučuje užívat komplexnějších neprolomených algoritmů SHA-2 nebo SHA-3.
- **Klidová kontrola integrity** *(Fixity-At-Rest)* sleduje stav uložených dat na discích. Jejím účelem je nacházet chyby způsobené degradací médií, lidským či softwarovým faktorem. Tento typ kontroly by měl být prováděn v pravidelných intervalech. Užití konkrétních kryptografických funkcí závisí na jejich dostupnosti v digitálním archivu.

Vytvořené kontrolní součty a kryptografické hashe jsou krátké textové řetězce, jež je možné bez větších problémů implementovat do databází, logů nebo metadat archivovaných AIP balíčků. Obdobně jako samotná data je doporučeno replikovat na oddělená média i hashe a kontrolní součty (ADDIS, 2020, s. 5).

Implementace informace o integritě v metadatech slouží nejen pro kontrolu integrity, nýbrž také pro kontrolu autenticity. V některých případech je změna v datech očekávaná a žádoucí (například v případě komprese či formátové migrace). Metadata musejí takovou změnu popsat a zároveň doložit autenticitu neboli potvrdit, že se stále jedná o tentýž dokument. Manipulaci s daty lze popsat standardem PREMIS, jenž zároveň elementem <premis:fixity> podporuje zápis informace o integritě příslušného souboru.

6.2.4.3 Kontrola integrity jako součást úložných technologií

Kontrola integrity se realizuje jak na hardwarové, tak na softwarové úrovni. Na hardwarové úrovni probíhá kontrola bitového toku například při přenosu přes Fibre Channel (MEGGYESI, 1994) nebo iSCSI (SHEINWALD, 2002). Kontroly integrity softwarové úrovně jsou

prováděny jednak v rámci konkrétního řešení dané paměťové instituce, představují však také klíčovou funkci softwarově definovaných úložišť (např. Ceph, ZFS a Hadoop).

6.3 Úložné technologie bitové ochrany

V této podkapitole jsou představeny používané úložné technologie pro bitovou ochranu dat v paměťových institucích. Vzhledem ke složitosti a rozsahu problematiky následující text rezignuje na kompletní přehled technologií a ve stručnosti jsou pojednána pouze témata, která nezbytně souvisejí s výše vyjmenovanými metodami (především pak s metodou replikace). Výklad je strukturován od nejmenších článků úložné infrastruktury – fyzických úložišť – přes disková pole tvořící z diskových jednotek skupiny, až po decentralizované systémy schopné distribuovat datový tok do geograficky vzdálených lokalit pomocí softwarově definovaných úložišť (*Software Defined Storage,* SDS).

6.3.1 Fyzické komponenty poskytující úložnou kapacitu

6.3.1.1 Magnetické pásky

Magnetické pásky jsou médiem, na nichž se data ukládají magnetizací tenké umělohmotné pásky potažené magnetickou vrstvou. Zápis na pásku je možné provádět v lineární nebo v lineárně „klikaté“ stopě (NEUROTH et al., 2010). Dnešní technologie LTO-9 *(Linear Tape Open)* založená na otevřeném standardu poskytuje maximální úložnou kapacitu jedné magnetické pásky o velikosti až 18 TB (LTO, 2023). Kromě LTO lze zmínit také proprietární řešení od firmy IBM, jejíž pásky TS1170 poskytují až 50 TB úložné kapacity, kterou je možné pomocí komprese ztrojnásobit (MELLOR, 2023). Páskové technologie zahrnují proces datové komprese, jenž šetří kapacitu na médiu a zefektivňuje přenos dat, a zároveň jsou opatřeny kontrolními postupy pro zajištění integrity dat (NEUROTH et al., 2010; ARSLAN et al., 2022). Jednotlivé magnetické pásky se zapojují do páskových knihoven, které disponují sofistikovanými zapisovacími a čtecími robotickými zařízeními schopnými vybírat požadovanou pásku pro čtení/zápis.

Při dodržení šetrné manipulace (tj. uchovávání ve vhodných klimatických podmínkách bez přítomnosti prachu a teplotních výkyvů, stejně jako použití šetrného čtecího/psacího zařízení) je doba životnosti magnetických pásek ve srovnání s jinými úložnými médii velmi vysoká; některé odhady udávají až 30 let (NEUROTH et al., 2010, s. 10–25). Tato doba se může lišit v závislosti na použitých materiálech a technologiích. Počínaje třetí generací zahrnuje technologie LTO schopnost WORM *(Write Once Read Many)*. Tato forma ochrany vůči zápisu poskytuje ochranu vůči nechtěným (chyby lidského faktoru) či záměrným (ransomware útok) zásahům do uložených dat.

Díky zmíněným vlastnostem jsou magnetické pásky pro paměťové instituce vhodným řešením pro uchování takových dat, která vyžadují dlouhodobou ochranu a méně častý přístup s jednorázovým zápisem. Obvyklým případem je zahrnutí páskové knihovny do infrastruktury prostřednictvím hierarchického úložného systému (viz níže podkapitola 6.3.3.4 níže o iRODS), v němž se pomocí stanovených politik data distribuují na adekvátní typ úložného média dle četnosti přístupových požadavků.

6.3.1.2 Pevné disky

Pevné disky jsou magnetická úložná média, která ukládají data na točící se plotny pomocí zapisovacích a čtecích hlav (NEUROTH et al., 2010, s. 10–28). Rychlost přístupu k datům je vyšší než u magnetických pásek, závisí však na technických parametrech pevného disku, především na rychlosti otáček ploten, rychlosti hlav a použitém rozhraní pro přenos dat (NEUROTH et al., 2010, s. 10–28). V době vzniku tohoto textu dosahuje kapacita magnetických disků přibližně 20 TB na jednotku (KLEIN, 2022). Pevné disky jsou velmi náchylné na magnetická pole a elektrostatické výboje a v porovnání s magnetickými páskami jsou významným rizikem mechanické nárazy, které v momentě spuštěného disku, kdy se nacházejí čtecí hlavy nad plotnami, mohou způsobit fatální poškození. Doba životnosti se v případě pevných disků pohybuje mezi 3 a 10 lety a bývá udávána veličinou MTBF (NEUROTH et al., 2010, s. 10–30) (viz také výše, podkapitola 4.1 o rizicích pro dlouhodobou archivaci dat).

V infrastruktuře paměťových institucí nacházejí pevné disky využití jako online nebo near-line [54] úložiště pro uchovávání informací vyžadujících častější přístup. Ačkoli dříve byly pevné disky označovány za nevhodné pro dlouhodobou archivaci jednak kvůli krátké životnosti a jednak z důvodu vysokých cenových nákladů (NEUROTH et al., 2010, s. 10–31), dnes se díky pokročilejším technologiím, nižším cenovým nákladům a s použitím metod redundance používají i pro uchovávání záloh v dlouhodobějším časovém horizontu.

6.3.1.3 SSD (Solid State Drive)

Základní paměťový článek o velikosti jednoho bitu v SSD tvoří hradlo MOSFET. Hradla se nacházejí ve vzájemně propojených paměťových čipech a jsou po jednom (*single-level cells,* SLC), dvou (*multi-level cells,* MLC) nebo po třech (*triple-level cells,* TLC) seskupovány do buněk. Modely s více bity v jedné buňce jsou cenově dostupnější a poskytují vyšší kapacitu, ovšem trpí delší dobou odezvy a větší náchylností k bit rot (GILLIS a KRANZ, 2021). Charakteristickou vlastností SSD disků je velmi krátká čtecí vybavovací doba a řádově delší doba zápisu, jež se při opakovaných zápisech prodlužuje v důsledku opotřebení buněk. Životnost média je v případě SSD vyjádřena veličinou TBW *(Tera Bytes Written)* udávající množství zapsaných dat. Rovnoměrnost opotřebení buněk je zajištěna pomocnými obvody. MOSFET hradla jsou schopna udržet informaci i po odpojení od napájení, avšak po delší době (v řádu nižších jednotek let) bez přívodu elektřiny může dojít ke ztrátě dat. Tato doba se zkracuje se zvyšující se teplotou při skladování (VÄTTÖ, 2015). Jelikož SSD neobsahují žádné pohyblivé komponenty, nejsou ve srovnání s pevnými disky náchylné k mechanickému poškození.

Vhodným využitím SSD v infrastruktuře paměťových institucí jsou online úložiště s vysokými nároky na výkon, například databázové systémy, katalogy a systémy zajišťující zpřístupnění. Důležitou roli

54 Pojmem „online“ zde není myšleno připojení k internetové síti, nýbrž zapojení média do (interní) sítě za účelem okamžitého přístupu k uloženým informacím. Pojem „near-line storage“ znamená takové úložiště, které sice neposkytuje přístup k informacím okamžitě, nicméně přístupu lze dosáhnout bez lidského zásahu. Příkladem mohou být pevné disky, jež se automaticky přepínají do režimu nečinnosti a vypínají otáčení ploten.

pak sehrávají v tierovaných (hierarchicky spravovaných) úložištích, v nichž se do SSD ukládají nejčastěji dotazované informace. Naopak kvůli vysokým pořizovacím nákladům, krátkodobé schopnosti uchovat data bez napájení a velmi komplikované obnově dat při selhání čipu se SSD nedoporučují k dlouhodobé archivaci.

6.3.2 Disková pole

Jednotlivá výše popsaná úložná média lze sdružovat do diskových polí. Disková pole jsou pak specializovaná infrastrukturní zařízení propojující lokálně i vzdáleně připojené disky, případně kapacity jiných diskových polí. Důležitou vlastností diskových polí je vyrovnávací cache s velmi rychlým zápisem a čtením, která je schopna zaručit rychlou odezvu i za přítomnosti „pomalejších" disků, na něž se data zapíšou později. Vyrovnávací cache bývá chráněna vůči smazání datového obsahu, k němuž může dojít například při nechtěném vypnutí zařízení. Diskové pole zároveň disponuje prostředky pro sledování stavu disků a v případě potřeby i pro opravu méně závažných chyb. V případě závažné chyby je schopno postiženou jednotku vyřadit, začlenit do pole nový disk ze seznamu náhradních jednotek (tzv. *hotspare* disků) a přepsat na něj obsah vyřazeného disku. Opravné procesy jsou umožněny na základě RAID geometrií vytvářejících redundanci mezi připojenými disky.[55]

6.3.3 Decentralizované systémy úložišť

Zatímco disková pole včetně RAID technologií operují na krátké vzdálenosti, decentralizované, softwarově definované úložné systémy vytvářejí nad servery s úložnou kapacitou další, abstraktní softwarovou vrstvu a mezi těmito jednotlivými uzly (servery) jsou schopny úložnou kapacitu efektivně distribuovat. Výhodou open source softwarově definovaných úložišť je absence problémů s proprietárním uzamčením a hardwarovou kompatibilitou jednotlivých komponent, která zároveň umožňuje naplnit jeden z požadavků bitové ochrany na hardwarovou diverzitu úložiště (TALLMAN a WANG, 2022, s. 342).

55 Redundanci vytvářejí geometrie RAID 1 až RAID 6. Naopak RAID 0 je pouhým spojením disků bez redundance.

6.3.3.1 CEPH

Ceph je otevřený, škálovatelný, softwarově definovaný systém, který poskytuje tři základní typy úložišť: blokové úložiště pomocí RADOS-Block-Device, souborové úložiště prostřednictvím CephFS a objektové úložiště přes RADOS-Gateway. Jeho klíčovou vrstvou je *Reliable Autonomous Distributed Object Store* (RADOS) – objektové úložiště, které je možno škálovat na mnoho úložných zařízení. Jeho pracovní náplní je distribuce dat a pracovní zátěže napříč dynamickým a heterogenním clustrem úložišť, zatímco navenek poskytuje aplikacím zdání jednoho logického objektového úložiště (WEIL et al., 2007, s. 1). Vrstva RADOS se skládá z množství *Object Storage Daemonů* (OSD) mapovaných na jednotlivé disky, s nimiž komunikují prostřednictvím backendu nazývaného *BlueStore*. OSD kontrolují svůj vlastní stav a stav ostatních OSD a vrací tyto informace Ceph-Monitorům, které udržují mapu clusteru a jeho stav (Ceph Documentation, 2016a). Kromě OSD a Ceph-Monitorů se cluster skládá ještě z Ceph-Manageru, jenž společně s Ceph-Monitorem řeší kritické stavy v clusteru. Replikace se provádí mezi jednotlivými uzly (servery) složenými z OSD daemonů, Ceph-Monitorů, Ceph-Managerů a Ceph-Metadata-Serverů, jejichž množina tvoří dohromady jeden cluster. Objekty spravované OSD daemony jsou společně s mapami clusteru replikovány mezi navzájem nezávislými uzly pomocí algoritmu CRUSH (WEIL et al., 2007, s. 1).

Obdobně jako jiné Software Defined Storage systémy, Ceph je hardwarově agnostický a nepředstavuje tudíž riziko vzájemných hardwarových nekompatibilit. Ceph je díky dalším svým výhodám, především transparentnosti, škálovatelnosti, ekonomické nenáročnosti a odolnosti vůči chybám, pro dlouhodobé úložiště paměťových institucí často voleným řešením.

6.3.3.2 Gluster

Gluster je otevřeným, distribuovaným, škálovatelným, softwarově definovaným systémem vyvíjeným společností Red Hat, Inc., poskytujícím souborové úložiště GlusterFS. Základní jednotkou GlusterFS jsou „cihly“ *(bricks)* poskytující úložnou kapacitu, jimiž jsou sdílené adresáře na serverech nebo diskových polích. Všechny servery nacházející se v GlusterFS musejí mít nainstalovaný a spuštěný řídící

daemon *(glusterd)*, který spravuje sdílené „cihly“ a jejich členství v úložišti. GlusterFS umožňuje několik konfigurací úložiště (GLUSTER DOCS, 2023):

- Distribuovaný GlusterFS svazek: výchozí nastavení, při němž jsou soubory libovolně distribuovány mezi zahrnutými „cihlami“. Toto nastavení slouží k zvětšení úložné kapacity a neaplikuje žádný typ redundance
- Replikovaný GlusterFS svazek: nastavení, které zajišťuje replikaci mezi jednotlivými „cihlami“
- Distribuovaný a replikovaný GlusterFS svazek: kombinace dvou předchozích nastavení, v němž jsou soubory distribuovány napříč replikovanými sadami „cihel“
- Rozptýlený GlusterFS svazek: data jsou rozptýlena po více „cihlách“ a opatřena redundancí v podobě erasure code pro ušetření úložné kapacity (viz výše, podkapitola Erasure Coding 6.2.1.2)
- Distribuovaný a rozptýlený GlusterFS svazek: nastavení, v němž jsou soubory distribuovány do více svazků, a zároveň jsou opatřeny redundancí v podobě erasure code

Klient přistupuje do filesystému Glusteru pomocí nativního klienta nebo protokolů NFS a CIFS. GlusterFS je klientem viděno jako jeden filesystém. Umístění souborů na GlusterFS není řízeno žádným centrálním uzlem, nýbrž je založeno na principu distribuované hašovací tabulky. Gluster podporuje rovněž funkčnost geo-replikace umožňující asynchronně replikovat data do geograficky vzdálených lokalit.

Gluster byl společně s Cephem v oblasti dlouhodobého uchovávání dat často zmiňovaným a doporučovaným řešením (TOMÁŠEK, 2018, s. 18–32; SCOTT, 2019). V současné době je ovšem nutné zohlednit blížící se konec vývoje Glusteru, který má skončit 31. prosince 2024 (Red Hat Customer Portal, 2023).

6.3.3.3 Hadoop

Hadoop je vyvinut společností Apache pro účely ukládání velkých objemů dat ve velkých objektech *(Big Data)*, které je nutno zároveň analyzovat a aktivně využívat. V porovnání s Cephem má tudíž Hadoop odlišné ambice a poskytuje pouze distribuované souborové

úložiště (*Hadoop Distributed File System,* HDFS). Výhodou HDFS je jeho navržení na práci s velkými soubory (v řádu TB) a vysoká datová průchodnost pro přístup aplikací. V souvislosti s tím Hadoop vychází z premisy, že výpočetní výkon pro práci s velkými daty je nejvýhodnější na místě jejich uložení. HDFS tudíž disponuje rozhraními, která umožňují aplikacím se souborovým úložištěm efektivně komunikovat. Jelikož se jedná o souborové úložiště, uživatel příslušné aplikace pomocí daného rozhraní pracuje s adresářovou strukturou (Apache Hadoop, 2023a).

Hlavními stavebními prvky HDFS clustru jsou uzly NameNode a DataNode. V uzlech DataNode jsou zapsána data v podobě bloků o konfigurovatelné velikosti, do nichž jsou soubory rozděleny. Hlavní uzel NameNode obsahuje informaci o rozmístění souborů v DataNodech, zároveň také disponuje informacemi o stavu DataNodů a o nutnosti replikace obsažených bloků. Předpokládá se, že jednou uložený velký soubor bude mnohokrát čten a nebude měněn/mazán. Proto je HDFS koncipován pouze pro jednorázový zápis souborů, který může být navíc prováděn jen jedním zapisovatelem v témže čase.

Možnost manipulace se soubory velkých objemů a vysoká průchodnost úložiště Hadoopu nabízí paměťovým institucím konkrétní případy využití. V kontextu dlouhodobé ochrany dat se typicky jedná o archivaci webu a její zpřístupnění (KVASNICA, 2015). Zranitelným článkem Hadoopu je uzel NameNode, s jehož ztrátou jsou ztracena i metadata s přístupy k uloženým datům.

6.3.3.4 iRODS

iRODS je uživatelskými politikami řízené úložiště. V porovnání s výše pojednanými systémy úložišť se jedná o middleware, který nevytváří žádnou úložnou kapacitu, nýbrž vytváří „zastřešující" vrstvu mezi kapacitou poskytnutou souborovými systémy a doménově-specifickými aplikacemi na straně druhé. Jedná se o distribuovaný systém, sestávající z katalogového uzlu a alespoň jednoho datového uzlu. K datovým uzlům jsou připojena vzdálená úložiště (disková pole, páskové knihovny) a lokální kapacity. Katalogový uzel vede databázi uložení objektů a soubor pravidel, za jakých podmínek a kam se data budou

kopírovat. Systém spoléhá na nativní (systémové) způsoby ukládání souborů. Obdobně jako v případě Hadoopu, kritickou součástí iRODSu je katalogový uzel obsahující metadata o pozicích uložených dat, v důsledku jehož ztráty nebudou data dohledatelná.

6.3.4 Úložiště od poskytovatelů cloudových služeb

Cloudová úložiště jsou optimálním řešením pro paměťové instituce, které nedisponují dostatečnými materiálními a personálními prostředky pro nákup a údržbu vlastní infrastruktury schopné poskytovat datům bitovou ochranu. Princip cloudového úložiště spočívá v pronájmu úložného prostoru a výpočetní kapacity, přičemž obojí může být škálováno na míru požadavkům dané instituce. Nevýhodou je skutečnost, že uploadem dat na cloudové úložiště externí společnosti ztrácí dotyčná paměťová instituce plnou kontrolu nad svými daty. Kromě toho poskytovatel služeb zpravidla účtuje za uložení dat periodické poplatky. Využití úložišť cloudových služeb tudíž představuje specifická rizika (omezení finančních prostředků, ukončení smlouvy ze strany poskytovatele), která by měla být zohledněna v exit-plánu dotyčné paměťové instituce. Z množství poskytovatelů cloudových služeb lze uvést především:

6.3.4.1 Amazon S3

Amazon S3 *(Simple Storage Service)* je jakožto součást *Amazon Web Services* (AWS) online-webovou službou, která poskytuje přes webové rozhraní objektově orientované cloudové úložiště. Jednotlivé objekty skládající se z jakéhokoliv typu dat o velikosti do 5 TB, metadat a globálního identifikátoru jsou seskupovány do konfigurovatelných jednotek *(bucketů)*. Uživatelskými nástroji pro práci s objekty a buckety je webová konzole S3, příkazový řádek, sada vývojářského softwaru (SDK) a metoda REST API (WITTIG a WITTIG, 2019, s. 237). Zatímco produkt S3 umožňuje okamžitý přístup k datům, AWS nabízí souběžně levnější službu *Amazon Glacier* pro dlouhodobé uchovávání velkých souborů s umožněním přístupnosti od jedné minuty do 12 hodin (WITTIG a WITTIG, 2019, s. 241).

6.3.5 Distribuce kapacity

6.3.5.1 Blokový přístup

Níže popsané protokoly podporující blokový přístup poskytují aplikačnímu serveru diskovou kapacitu vytvořenou na diskovém poli a poskytovanou jako souvislý počet bloků pojmenovaný LUN *(Logical Unit Number)*. Server může s LUN pracovat jako s fyzickým diskem, tj. dělit jej na oddíly, formátovat, zahrnout do virtualizace disku typu LVM *(Logical Volume Management)* a hospodařit s ním na úrovni operačního systému (např. monitoring zaplnění objemové kapacity a počtu souborů…). LUN má v porovnání se souborovým nebo objektovým přístupem nižší latenci. Blokovou kapacitu lze za provozu adaptovat měnícím se požadavkům, například zmenšovat či zvětšovat, migrovat mezi tiery atd. (Oracle, 2023).

Bloková kapacita může být distribuována prostřednictvím protokolu **FC** *(Fibre Channel)*, určeného pro bezeztrátové přenášení velkého objemu dat na delší vzdálenost po optických vláknech. Fibre channel je spolehlivou, výkonnou, avšak ekonomicky náročnou technologií propojující disková pole a servery. Levnější řešení poskytuje protokol **iSCSI** *(internet Small Computer System Interface)*, který umožňuje odesílat SCSI příkazy přes TCP/IP protokol a v porovnání s FC nevyžaduje k propojení diskových polí a serverů specifické switche a adaptéry (RAFFO, 2019). Další alternativou je technologie **FCoE** *(Fibre Channel over Ethernet)*, která umožňuje sloučit síť úložišť *(Storage Area Network, SAN)* využívající FC protokol s ethernetovou sítí LAN do jednoho konvergovaného adaptéru *(Converged Network Adapter,* CNA*)*. Použití technologie FCoE poskytuje vyšší přenosovou rychlost než iSCSI a nevyžaduje provoz specializovaných hardwarových komponent pro Fibre Channel síť (JASTRAB, 2008).

6.3.5.2 Souborový přístup

S nárůstem propustnosti ethernetových sítí roste obliba *Network Attached Storage* (NAS). Jedná se o kapacitu vytvořenou na serveru nebo diskovém poli a zpřístupněnou přes LAN současně několika aplikačním serverům. Uživatel připojuje jednotku již jako funkční filesystém. Proces poskytující přístup k souborovému systému musí

zohlednit řadu problémů, např. přístup k souborům, jejich zamykání, autorizaci a autentizaci uživatelů na NAS službě a mnohé další.

Účastníky procesu sdílení adresářů jsou server a několik klientů. Server definuje přístupová síťová oprávnění, klient připojuje filesystém definovaný na serveru ke svému prázdnému lokálnímu adresáři *(mount point)*. Poté, co je spojení klientem navázáno, klient má na svém lokálním adresáři sdílenou adresářovou strukturu, jejíž obsah může s příslušnými oprávněními modifikovat. Běžnými protokoly pro sdílení adresářů po síti jsou **NFS** *(Network File System)* a **SMB/CIFS** *(Server Message Block)* (HANNAN, 2016).

6.3.5.3 Objektový přístup

Pro úplnost výkladu je vhodné zmínit rovněž objektový přístup, jenž se týká výše jmenovaných objektově-orientovaných úložišť. Za rozhraní umožňující objektový přístup lze považovat například **Ceph Object Gateway**, které je prostřednictvím REST API schopno komunikovat s objektovými úložišti Amazon S3 a OpenStack Swift. Ke komunikaci je používán *Object Gateway Daemon*, který na jednu stranu interaguje s clustrem Cephu a na druhou stranu je rozhraní kompatibilní s API cloudu Amazon S3 a systému OpenStack Swift (Ceph Documentation, 2016c).

7 Nástroje LTP se zaměřením na praxi v České republice

Dlouhodobé uchování digitálních dokumentů představuje rozsáhlý komplex aktivit, jejichž hlavním cílem je uchovat čitelnost uložených dat (ve smyslu zobrazení běžně dostupnými softwarovými nástroji), zachovat jejich věrnost, důvěryhodnost a srozumitelnost nesené informace. K tomuto účelu jsou zřizovány a provozovány dlouhodobé repozitáře dle konceptu OAIS (viz výše), pro které se vžil název **LTP systémy**. Ty využívají (integrují) množství dílčích hardwarových komponent a softwarových nástrojů, aby mohly naplnit požadované cíle. Jádrem každého z těchto systémů je samozřejmě pokročilé úložiště a vhodné nástroje pro jeho správu. Základními úkoly správců dat je ochránit bitstream, uložená data správně identifikovat a popsat. Ochrana bitstreamu je realizována pomocí popsaného úložiště (RŮŽIČKA et al., 2019). Tato činnost je dobře zmapovaná a lze se řídit doporučenými postupy. V souladu s konceptem OAIS je třeba dlouhodobé uchovávání (logickou ochranu) chápat jako sled aktivit, které s pomocí k tomu určených nástrojů a definovaných politik zajišťují výše popsané cíle.

Z hlediska logické ochrany je třeba se zaměřit na dva nejdůležitější aspekty. Prvním z nich je uchování informačního obsahu a jeho kontextu (content data object dle OAIS). V tomto ohledu musí správci repozitářů především průběžně sledovat určenou komunitu a vyhodnocovat její znalosti a následně srozumitelnost uchovávaných digitálních objektů. Pro uchování informačního obsahu je pak nutné monitorovat zastarávání formátů a podle něj nastavovat ochranná opatření. Aktuálně jsou jimi nejčastěji formátové migrace, případně emulace softwaru. Aby bylo možné tato opatření správně nastavit, je nezbytné detailně znát obsah repozitáře a způsob jeho zpřístupnění. Nejefektivnějším způsobem (KEJSER et al. 2011) jak docílit dlouhodobého uchování je vytvářet digitální dokumenty podle dokumen-

tovaných metadatových a datových formátů popsaných veřejnými standardy (Národní archiv, 2022a). Taková produkce zpravidla zaručí, že data přijatá do systému pro dlouhodobé uchování, nebudou v krátkodobém časovém horizontu ohrožena, a tudíž u nich nebude nutné provádět ochranné operace v podobě formátových migrací, případně pro ně vytvářet emulační prostředí. Většinou se též doporučuje zvolit standard nezatížený patenty, s dostupnou dokumentací a s dostupnými nástroji (LOC, 2017).

Lze zmínit, že také v rámci České republiky již máme zkušenosti s nenávratnou ztrátou digitálních dat v důsledku zastaralosti formátu. Příkladem mohou být digitální data produkovaná v osmdesátých letech 20. století, která sice byla uchována na neporušených nosičích, ale v důsledku neznalosti principů LTP nebyla prováděna ochranná opatření, která vedle k zastarání hardware. Následně již nebylo možné data z nosičů získat kvůli absenci hardwaru i kompatibilního software (VOJÁČEK a KUNT, 2019).

Jen díky spolehlivým a podrobným informacím o obsahu OAIS archivu lze přistoupit k průběžnému hodnocení čitelnosti dat a rozhodnout o potřebě ochranných opatření, ať už jsou jimi formátové migrace nebo emulace. LTP systém je tak třeba chápat jako prolnutí sofistikovaných nástrojů a znalostí jejich uživatelů. Cyklus opatření, která mají zajistit dostupnost a čitelnost digitálních dokumentů, začíná již v okamžiku jejich vzniku, kdy by se měly v ideálním případě tvořit v doporučených formátech a jejich verzích s popisem odpovídajícím zvoleným standardům. Ochranná opatření tak začínají hned při příjmu dokumentů do OAIS archivu, protože pro jejich správné nastavení potřebují správci přesně identifikovat formáty uložených dat. Na základě této identifikace lze dále volit nástroje pro realizaci ochranných opatření. Při vstupu do archivu je tak nutné všechna data správně identifikovat (tedy rozpoznat jejich souborový formát), na základě identifikace provést validaci (ověřit, zda daný soubor skutečně odpovídá všem pravidlům formátu, ke kterému se hlásí, že digitální objekt je validní reprezentací daného typu formátu) a následně se pokusit soubor tzv. charakterizovat, tedy zjistit jeho signifikantní technické vlastnosti. Díky tomuto postupu lze uložená data spravovat a na základě znalostí specifikací jednotlivých formátů a jejich variant nastavit postupy pro další zpracování.

Všechny tři operace lze samozřejmě spouštět i nad již uloženými daty a ověřovat předchozí výsledky. Opakované spouštění je dokonce žádoucí, protože dochází jak k vývoji aplikací pro tyto účely, tak k rozvoji znalostí o formátech. Opakováním procesu identifikace lze po určitém časovém odstupu dojít k rozdílným výsledkům, které mají samozřejmě dopad na správu repozitáře. Výstupy z popsaných procesů by měly být zapisovány jak do databází pro správu repozitáře, tak do metadat archivních balíčků. Při volbě nástrojů je třeba přizpůsobit výběr vhodných nástrojů obsahu repozitáře. Vzhledem k velikosti a množství digitálních souborů, které vstupují do digitálních repozitářů, je nutné získávat potřebné informace co nejvíce automatizovaně. Pro velké instalace navíc platí, že pro některé činnosti, jako je např. identifikace, je vhodné zvolit více nástrojů, což na jednu stranu vede k přesnějšímu výsledku, na druhé to znamená další nároky na vyhodnocování výsledků a rozhodování o jejich relevanci.

Cílem této kapitoly je představit jednotlivé nástroje (a jejich skupiny), které lze využít pro doporučované procesy v průběhu správy uložených dat, a to v kontextu českých paměťových institucí. Vzhledem k počtu jednotlivých nástrojů bude představen jen reprezentativní výběr. Jak bylo řečeno výše, často jsou všechny nástroje integrovány do LTP systémů. Většina z nich však funguje i samostatně, lze je využívat nejen v repozitářích, ale již při produkci digitálních dokumentů anebo při samostatných analytických úkonech. Nejdříve budou představeny nástroje (tedy ty, které existují samostatně a obvykle se specializují na jednu konkrétní činnost) tak, jak odpovídají životnímu cyklu digitálního dokumentu. Následně budou popsány některé LTP systémy, které se v ČR užívají nebo jsou diskutovány. Popsané nástroje odpovídají aktuálnímu stavu v roce 2023. Výčet užívaných nástrojů samozřejmě nemusí být kompletní, budou zmíněny jen ty nástroje, které mají větší počet instalací, jejich nasazení je zdokumentováno, anebo mají specifické vlastnosti. Zároveň budou nastíněny pouze nástroje, které jsou speciálně určené pro operace za účelem dosažení dlouhodobého uchovávání digitálních objektů. Kromě dále uvedených specifických nástrojů spoléhají správci digitálního obsahu na běžné systémové nástroje

unixových operačních systémů nebo další běžné nástroje pro správu digitálních dat (jako jsou např. nástroje na kontroly kontrolních součtů, různé manažery etc.).

7.1 Formátové registry a znalostní báze

Trvalé uchovávání digitálních dat do značné míry závisí i na globální spolupráci a informační infrastruktuře. Ani velké repozitáře nedisponují takovým personálním zázemím, aby byly schopny obsáhnout a dostatečně zajistit veškeré oblasti. Aby kurátoři mohli vykonávat své úkoly, potřebují mít k dispozici znalostní báze informací o digitálních formátech a rizicích s nimi spojenými a databáze obsahující jedinečné technické znaky (charakteristiky) konkrétních formátů. Tyto databáze nemají vysloveně povahu nástroje, tak jak je jazykově chápán, ale bez jejich užití nelze dlouhodobé uchovávání realizovat, protože z nich čerpají další nástroje, které je využívají k vlastním úkonům. Typickým příkladem těchto informačních bází jsou formátové knihovny/formátové registry. Jedná se o rozsáhlé databáze obsahující podrobné technické specifikace jednotlivých formátů a jejich verzí digitálních dokumentů (souborů). Nástroje pro identifikaci formátů přebírají z těchto registrů informace o specifikaci a vlastnostech jednotlivých formátů a s jejich pomocí je určují. Správa těchto registrů je poměrně nákladná, jelikož vyžaduje velké množství analytické práce pro udržování a prohlubování jednotlivých záznamů digitálních souborů. Jen v nejrozsáhlejším registru PRONOM má záznam více než 2300 formátů, a to zdaleka nepokrývá plnou šíři existujících formátů. Zatímco některé další nástroje lze nahradit specifickými aplikacemi v režii jednotlivých repozitářů, formátové knihovny jsou natolik náročné na správu, že je třeba využívat externí zdroje a sdružovat síly při jejich údržbě. Význam existujících formátových registrů je tak pro plnění LTP funkcí zcela klíčový a repozitáře jsou na jejich informacích závislé.

V jednotlivých LTP systémech následně vznikají lokální formátové knihovny, které podle svého zaměření přebírají část nebo celý

formátový registr, případně kombinují více registrů a vytvářejí si tak svůj vlastní. Ten je průběžně aktualizován podle vývoje svých zdrojových systémů. Často jsou pak lokální registry doplňovány o vlastní definice formátů, které v globálních registrech záznam nemají. Tato situace nastává zejména tehdy, když existují specifické souborové formáty pouze s národním nebo dokonce oborovým uplatněním. Jejich význam pro globální registry je nižší, a proto musí jejich záznam spravovat příslušný repozitář a snažit se o jejich prosazení do globálního registru. Případně jde o situace, kdy se s daným formátem v lokálním prostředí pracuje odlišným způsobem a podařilo se díky tomu definovat nové vlastnosti (Národní archiv, 2022b). Příkladem může být dlouhodobá snaha Národního archivu ČR prosadit do registru PRONOM záznamy souborových formátů typických pro datové schránky v ČR (isdoc, isdocx,[56] zfo) a formáty z rodiny Software 602, které byly hojně využívány v devadesátých letech 20. století.[57]

Nejdůležitějším formátovým registrem je aktuálně registr PRONOM. Je základním zdrojem informací pro většinu LTP systémů a představuje tak zcela klíčovou infrastrukturu pro zajišťování dlouhodobého uchovávání digitálních dat. Registr PRONOM provozuje britský The National Archives, přičemž spolupracuje s Digital Preservation Coalition. Vývoj registru začal v roce 2002 a od roku 2004 disponuje webovou podobou. Obsahuje nejen technické informace o jednotlivých formátech a jejich verzích, ale také podrobný soupis dokumentace, informace o užití a doporučených nástrojích (samozřejmě za předpokladu, že tyto informace jsou pro daný formát k dispozici). PRONOM na počátku vznikl jako znalostní báze spojená s nástrojem DROID *(Digital Record Object Identification)*, zveřejněn byl až později. Dalším nástrojům, které jsou využívány pro identifikaci formátů, nyní registr poskytuje tzv. signature files, které slouží jako základ pro identifikaci. Každý záznam je identifikován pomocí unikátního kódu PUID *(PRONOM Persistent Unique Identifier)*, který

56 ISDOCX i ISDOC byly zařazeny v prosinci 2021 – https://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=detailReport&id=2396 ahttps://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=detailReport&id=2395.

57 https://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=detailReport&id=2063

je identifikátorem souborového formátu a u některých formátů rozlišuje i jejich jednotlivé verze. Identifikátor PUID je namapován na konkrétní bitové sekvence (signatures, magic numbers), které se nacházejí v těle souboru na místě typickém pro daný formát[58] (HUTAŘ a MELICHAR, 2015b).

Vzhledem k tomu, že PRONOM původně vznikl jako interní databáze, je pro ostatní instituce složité přispívat k jeho obsahu, což je poměrně nekomfortní za situace, kdy je na něm jejich fungování závislé. Neustále proto vznikají pokusy o vytvoření dalšího alternativního registru souborových formátů. Z různých důvodů se však tyto registry neprosadily, respektive se dlouhodobě neudržely. Společným jmenovatelem byla přílišná náročnost udržování registrů aktuálních a technicky dostupných (McKINNEY et al., 2014). Z historických projektů lze zmínit např. projekty GDFR (Global Digital Format Registry) a pozdější UDFR (Unified Digital Format Registry). Velké ambice měl projekt DPTR (Digital Preservation Technical Registry) realizovaný pod záštitou NSLA (National and State Libraries Australasia), který měl mít kromě informací o formátech i údaje o softwaru, hardwaru apod. Ani tento projekt neuspěl, ale některé z nápadů byly využity pro zlepšení registru PRONOM. Další realizovaný projekt měl název Preserv2 a byl pokusem o vylepšenou podobu PRONOM. Aktuálně nejpoužívanější alternativou k registru PRONOM je registr FDD, který provozuje Kongresová knihovna. Také pro ni je registr formátů především doplňkem k dalším aktivitám.

Jak již bylo uvedeno, kromě technických informací o formátech je zajištění dlouhodobé čitelnosti digitálních objektů řízeno především činností kurátorů, kteří posuzují rizika, vhodnost formátů a jejich perspektivy. Kurátoři proto potřebují mít informace pro rozhodování, kterým digitálním formátům mohou z dlouhodobého hlediska důvěřovat a proč. Vzhledem k různorodosti formátů je vhodné, aby se řídili doporučeními odborné komunity a využívali její znalostní báze. Právě takovou znalostní bázi poskytuje Kongresová knihovna (LOC, 2017), která rozlišuje sedm faktorů udržitelnosti digitálního

58 Signature je sekvence bytů nebo seznam více sekvencí, které mohou být v souboru určitého formátu obsaženy. Pro některé formáty může jít pouze o sled bytů na počátku souboru, pro jiné formáty hraje roli i poloha v rámci celého kódu.

formátu (typ formátu – např. proprietární; rozsah používání formátu; transparentnost, možnost vložit metadata přímo do souboru, vnější závislosti, vliv patentů, možnost technické ochrany). Pro činnost báze je využíván registr FDD. Na svém webu pak Kongresová knihovna udržuje rozsáhlou znalostní bázi s informacemi o digitálních formátech s odkazy na jejich dokumentaci a s hodnocením jejich vhodnosti pro dlouhodobé ukládání.

Kromě globálních a otevřených formátových knihoven existuje několik dalších komunit sdružených kolem konkrétních řešení, které společně budují vlastní formátové knihovny. Při tom využívají informace z PRONOM pro vybudování vlastního registru, následně tyto informace obohacují v souvislosti s praxí vlastních repozitářů, mezi členy své komunity pak sdílejí další informace o rizicích nebo pravidlech spojených se zpracováním formátů. Příklady takových komunit jsou uživatelské skupiny systémů jako jsou Archivematica, Ex Libris Rosetta nebo SDB/Preservica.

7.2 Nástroje využívané při ingestu do repozitáře

Počet činností, které je nutné realizovat v rámci dlouhodobého uchování je velmi značný. Pro některé lze využít značné množství obecných i specializovaných nástrojů, další činnosti jsou tak specifické (např. konkrétní činnost nebo dokonce jen aplikace pro specifický formát), že pro ně musí být vytvořena zcela specifická aplikace. Velká část těchto nástrojů je dostupná v režimu open source. Při budování LTP systému musí jeho tvůrci zvážit přínos nebo případné obtíže při využití jednotlivých nástrojů. Kombinace několika „drobných" specifických nástrojů může být ideální z hlediska flexibility a možností přizpůsobení požadavkům organizace, komplexní řešení však zřejmě bude méně náročné z hlediska provozu a údržby. Obdobné pravidlo je třeba uplatňovat i vzhledem k tvůrci daného nástroje – program vyvíjený větší uživatelskou komunitou nebo komerční firmou nebude zřejmě zcela přesně odpovídat všem požadavkům instituce, je u něj však vyšší pravděpodobnost stabilní podpory a dalšího rozvoje. Po-

zornost je třeba věnovat i pozici nástroje v rámci workflow zpracování dat – nástroj by se neměl stát „úzkým hrdlem“, tedy bodem, ve kterém je celé workflow výrazně zpomaleno například kvůli nárokům na výkon serveru.

Z pohledu dlouhodobého repozitáře začíná životní cyklus okamžikem vstupu do LTP systému. V této části systémů pro dlouhodobé uchovávání digitálních dat se uplatňují obecné softwarové nástroje, které se využívají i pro jiné účely mimo tuto oblast. Typicky jde o antivirové programy, nástroje pro kontrolu fixity (respektive nástroje, které jsou schopny vytvářet i kontrolovat kontrolní součty) nebo souborové manažery pro kontrolu úplnosti přenesených dat. Zejména u nástrojů pro kontrolu fixity má smysl zvážit, jaké typy kontrolních součtů se v rámci LTP systému uplatní. I přes obecnou kompromitaci algoritmu md5 je pro kontrolu datové integrity v rámci úložiště plně vyhovující a výpočetně méně náročný, i když standardem bývá již některý z pokročilejších algoritmů, obvykle SHA-256.

7.3 Význam identifikace v rámci dlouhodobého uchování

Pro realizaci dalších kroků vedoucích k zajištění čitelnosti uchovávaných digitálních objektů je klíčová jejich identifikace, respektive identifikace jejich souborových formátů. Ta se obvykle vykonává na vstupu do repozitáře, ale měla by se následně provádět opakovaně, na což některé strategie pro dlouhodobé uchovávání zapomínají. Také definice formátů prochází postupným vývojem, zdokonalují se nástroje pro jejich identifikaci a jsou doplňovány databáze s informacemi. Při správě dat je tudíž chybou, pokud nejsou identifikační nástroje spouštěny i nad uloženým obsahem. Formátem datového souboru se rozumí standard, na základě kterého je informační obsah uložen do elektronického souboru, který je zapsán pomocí binárního kódu. Při zpracování digitálních souborů na vstupu do repozitáře je uplatňována série kroků, během které se nejprve rozpozná, o jaký digitální formát se jedná, a dále se pomocí dalších nástrojů ověří, že je digitální objekt validní reprezentací daného typu formátu. Tento

krok je zcela zásadní, neboť umožňuje správně identifikovat, jaké soubory do archivu vstupují. Na základě identifikace lze patřičně nastavit strategii dalšího uchování, vyhodnotit riziko formátu a určit nástroje, které jsou pro práci s ním vhodné. Identifikace je také nezbytnou součástí kontroly úspěšnosti migrací. Jen u dat ve validních formátech lze spolehlivě užívat nástroje pro jejich zobrazení anebo následné úpravy.

V případě vytvoření souboru dle standardu určitého formátu může odchylka od normy v budoucnu vést k tomu, že kvůli ní selže hromadná operace, jako je např. formátová migrace, spuštěná nad dávkou dat s domněle stejnými vlastnostmi. Vedle toho přirozeně existuje nebezpečí, že se někdo omylem či záměrně pokusí repozitáři podvrhnout dokument, který se hlásí k jinému formátu, než ve kterém byl ve skutečnosti vytvořen. To s sebou nese provozní i bezpečnostní rizika. Pokud se pokusíte otevřít soubor nástrojem, který k tomu není určen, obvykle se to nepodaří nebo není možné srozumitelně zobrazit obsah. V případě, že je chyba odhalena až při detekci potřeby ochranných opatření, mohou kurátoři narazit na to, že již nejsou dostupné potřebné nástroje. Je tedy zřejmé, že úspěšná identifikace dat na vstupu je nezbytným a klíčovým krokem dlouhodobého uchování. Bez ní důvěryhodný digitální archiv provozovat nelze (validaci na rozdíl od identifikace lze teoreticky vynechat, byť se tím provozovatel vystavuje riziku). Výsledky identifikace se obvykle ukládají do AIP a používají se i pro volbu vhodného validačního nástroje, který potvrdí, že formát odpovídá své normě. Zatímco nástroje pro identifikaci mají spíše obecný charakter, validátory jsou specializované. Vyplývá to z charakteru jejich činností.

Z konceptu OAIS vyplývá, že je třeba uchovávaná data udržovat aktuální. Totéž platí i pro výsledky identifikace formátů. Za optimální se považuje období pěti let (HUTAŘ, 2016), po kterých by se měly identifikace opakovat. Pokud má archiv trvale uchovat nějaký digitální obsah, měl by také sledovat změny nástrojů pro LTP, validátorů, extraktorů a především technologií používaných k identifikaci formátu souborů. Tyto procesy jsou samozřejmě výpočetně a personálně náročné a výrazně zatěžují instituce provozující digitální archivy, přičemž nesmí omezit jejich další provoz. Jako kompromisní

řešení se tak nabízí výběr optimalizačních scénářů, podle kterých jsou vybrána jen některá data, u kterých je větší pravděpodobnost změny. Jednotlivé LTP systémy jsou pro tento účel odlišně disponovány, některé (např. Rosetta) s ním počítají, u jiných ho lze realizovat jen reingestem dat (Archivematica). Příkladem, který dokládá smysl reidentifikace, může být formát TIFF. Dnes je identifikován pomocí PUID fmt/353, ale do roku 2011 ho bylo možné popsat pomocí fmt/7 (TIFF v3), fmt/8 (TIFF v4), fmt/9 (TIFF v5) a fmt/10 (TIFF v6).

Tímto příkladem jsme se dostali k metodám identifikace souborových formátů. Při identifikaci rozeznáváme dvě hlavní metody: dle přípony *(extension)* a dle obsahu *(signatures)*. K základní identifikaci formátu slouží přípona, která bývá odvozena od jeho názvu. Při určování *signatures* se využívá tzv. magické číslo *(magic number)*, což je hodnota typická pro daný souborový formát. Jsou to v podstatě konkrétní bitové sekvence, které se musí nacházet v těle souboru na místech typických pro daný formát. Registr formátů uchovává signature pro daný formát a ta je pak využívána pro srovnání s identifikovaným souborem. Při identifikaci samozřejmě dochází k mnoha konfliktním situacím, se kterými se musí nástroje vypořádat:

- není dostupná přípona souboru;
- přípona souboru je v rozporu s výsledkem dle signature;
- není k dispozici signature, kterou by souborový formát uměl identifikovat;
- je identifikováno více verzí formátů, mezi nimiž není nástroj schopen dále rozlišit
- a další.

Různé výsledky identifikace mohou být také následkem funkčních vlastností nástrojů. Vývojáři musí vždy volit mezi rychlostí a spolehlivostí. Rychlejší nástroje obvykle používají jednodušší metody identifikace (hledají sekvenci bitů v kratší části souboru nebo kontrolují jen hlavičky apod.). Větší spolehlivost naopak často znamená vyšší nároky na čas a výpočetní výkon.

Jednotlivé nástroje využívají informace z globálních registrů obsahujících jedinečné technické znaky konkrétních formátů – dle konceptu OAIS jde o síť registrů interpretačních informací *(repre-*

sentation network). Tou nejvíce užívanou, jak bylo uvedeno výše, je databáze PRONOM. Klíčovým identifikátorem v registru PRONOM je Unique ID, tzv. PUID, což je nejčastěji užívaný přesný identifikátor formátu. V internetovém prostředí se také často používá jako identifikátor typu formátu tzv. MIMEtype. Ten dokáží zjistit též běžné nástroje operačních systémů/file systémů, ovšem pro dlouhodobou archivaci je tato informace nedostatečná. Obecně platí, že úspěšnost identifikace datových formátů je vyšší u dat produkovaných co nejblíže době identifikace. Čím jsou data starší, tím více naráží kurátoři dlouhodobých repozitářů na zastarávající formáty, které nejsou dostatečně popsané a jejich identifikace je obtížná. Nicméně v činnosti paměťových institucí se setkáváme i s těmito daty a někdy je jako výsledek validace třeba akceptovat zjištění, že jde o neznámý formát. Také z tohoto důvodu je vhodné provádět opakované identifikace, protože prohlubování interpretačních informací může vést k úspěšnější identifikaci.

7.4 Dostupné nástroje pro identifikaci datových formátů

Jak bylo řečeno, otázka identifikace formátů je pro dlouhodobé uchování digitálních dokumentů klíčová. Nástrojů pro její realizaci existuje omezené množství. Jejich programování a udržování v aktuálním stavu je poměrně náročné a obvykle bývají poskytovány uživatelům bezplatně. V ideálním případě je třeba, aby na tvorbě, a především průběžně údržbě těchto nástrojů pracovaly větší týmy. Realitou však je, že za většinou nejčastěji užívaných nástrojů je jen malý tým. Celý systém identifikace je třeba rozdělit do dvou úrovní. Tou první jsou zmíněné databáze interpretačních informací, tedy registry souborových formátů. Přes několik pokusů vybudovat větší počet registrů je dlouhodobě udržovaný pouze jeden, a to registr PRONOM. Stále se však objevují snahy o vytvoření veřejné alternativy (a také existují uzavřená řešení v rámci jednotlivých uživatelských komunit, jak bylo zmíněno výše). Přestože je registr PRONOM největší, neobsahuje zdaleka všechny souborové formáty. Pro vy-

tvoření záznamu je třeba, aby byl formát dobře zdokumentovaný, což vždy není možné. Zdaleka ne všechny záznamy jsou kompletní, signatures nejsou např. dostupné pro necelou polovinu záznamů. Z databáze PRONOM čerpají informace všechny užívané nástroje pro identifikaci souborových formátů. I u těchto nástrojů platí, že ne všechny vzniklé se udrží, některé zaniknou, respektive přestanou být inovovány. Příkladem může být kdysi nadějný nástroj pro identifikaci ffident.[59]

Jde většinou o volně dostupné nástroje, které fungují jako samostatné stojící aplikace ovládané přes příkazovou řádku, výjimečně mají nástroje uživatelské rozhraní. Velmi často jsou integrované v komplexních systémech pro dlouhodobou ochranu, tedy v rámci LTP systému. Cílem těchto nástrojů je formátu přiřadit jediný PUID z registru PRONOM, tj. co nejpřesnější a jednoznačnou identifikace formátu, na rozdíl od běžně dostupných nástrojů pro identifikaci formátu (např. příkaz file v linuxovém prostředí) (SPENCER, 2022).

Vedle identifikátorů PUID se v internetovém prostředí pracuje ještě s identifikátory MIMEtype, jež spravuje organizace IANA. Identifikátory MIMEtype[60] jsou pro dlouhodobou archivaci považovány za nedostatečné (HUTAŘ a MELICHAR, 2015b). Důvodů je hned několik: nikdy nerozlišují verze souborových formátů, jednomu formátu může odpovídat více MIME typů, neexistuje žádný centrální registr s úplným výčtem formátů a nejsou svázány s jednoznačným způsobem identifikace (tj. není určeno, jak se formát takto identifikuje, neexistují žádné signatures, záleží na konkrétní implementaci). Vedle identifikátorů PUID lze použít i identifikátory FDD (File Format Descriptions) přidělované Kongresovou knihovnou formátům v jejím registru souborových formátů (LOC, 2023b). Identifikátor FDD má prefix fdd a 6 čísel (např. fdd000001 pro formát WAVE), novému formátu je vždy přidělen první volný identifikátor v číselné řadě. Počínaje rokem 2018 začali zaměstnanci Kongresové knihovny přidávat do popisů formátů i další dostupné identifikátory

59 http://web.archive.org/web/20061106114156/http://schmidt.devlib.org/ffident/index.html

60 Akronym pro Multipurpose Internet Mail Extension.

formátů (např. PUID, identifikaci z Wikidat), aby došlo k propojení vícero registrů a informací, jež obsahují (LOC, 2023c).[61]

Z existujícího množství nástrojů pro identifikaci souborových formátů si v této kapitole představíme ty nejčastěji používané a v následující provedeme jejich srovnání. Open Preservation Foundation provedla průzkum ohledně přístupu k digitální archivaci v paměťových institucích. Průzkum probíhal od listopadu 2019 do února 2020 a odpovědi byly získány od 98 institucí (knihovny, archivy, muzea, vládní oddělení, výzkumná jednotka větší instituce, mediální instituce, galerie datová centra a další) z 31 zemí světa. Mj. výzkum zjišťoval, zda instituce používají dále zmíněné identifikační nástroje, či zda je zkoumají a testují. Z dostupných dat jsme zjistili, že nejčastěji je využíván nástroj DROID (47 institucí), nástroj Siegfried se objevuje v užívání 15 institucí a FIDO jedenáctkrát. Dvanáct z těchto sledovaných institucí používá v produkčním prostředí dva zmíněné nástroje, tři instituce mají v produkčním prostředí nasazené všechny tři nástroje. Kromě samostatně použitelných aplikací, které budou dále popsány, existují snahy o sdružení více nástrojů do jednoho celku (tzv. *wrapperu*). V jednom rozhraní je možné spouštět více nástrojů s různými funkcemi. Příkladem takového nástroje je aplikace FITS, která v sobě integruje nástroje pro identifikaci, validaci i extrakci metadat.[62] Jednotné rozhraní přináší jisté usnadnění, v jádru jde však jen o nadstavbu nad existujícími aplikacemi. Aplikace normalizuje a upravuje výstupy dílčích nástrojů a hlásí jejich chyby. Vytvořena byla na univerzitě v Harvardu pro tamní LTP systém.

7.4.1 DROID[63]

Nástroj DROID *(Digital Record Object Identification)* vytvořila společnost Tessella Scientific Software Solutions ve spolupráci s The National Archives. Další správu a rozvoj nástroj zajišťuje The National

61 Ne vždy lze však dosáhnout bezešvého propojení všech registrů, jelikož se liší granularitou popisu souborových formátů. Příkladem může být formát TIFF, který má v registru Kongresové knihovny záznamů více, jelikož Kongresová knihovna rozlišuje i jednotlivé revize (verze) formátu TIFF narozdíl např od PRONOM-u.

62 https://coptr.digipres.org/index.php/FITS_(File_Information_Tool_Set)

63 https://digital-preservation.github.io/droid/

Archives vlastními silami. DROID je v současnosti nejrozšířenějším nástrojem pro identifikaci formátů v paměťových institucích. Byl vytvořen v jazyce Java a je použitelný na operačních systémech Linux, Windows a MacOs. Je dostupný jako CLI i jako GUI. Výstupy nástroje jsou ve formátu CSV. Provádí automatickou identifikaci formátů jak u jednotlivých objektů, tak hromadně. Výstupem je informace o konkrétní verzi formátu digitálního objektu. DROID používá dvě definice signature files – jednu pro binární soubory a druhou pro kontejnery. Tyto definice jsou získávány z PRONOM.[64] Jak vyplynulo z výše odkázaného průzkumu, DROID zůstává trvale nejrozšířenějším nástrojem pro identifikaci souborových formátů. V rámci ČR ho jako primární zdroj identifikace používá většina paměťových institucí věnujících se dlouhodobému uchovávání digitální dat. Je využíván v Národním digitálním archivu (NDA), který provozuje Národní archiv ČR a to jak ve Validátoru SIP, tak v ingestovém workflow NDA. DROID je také nasazen v Národní digitální knihovně nebo v rámci digitálního archivu, který buduje Archiv hlavního města Prahy (AHMP).

7.4.2 FIDO[65]

Nástroj FIDO vytvořil Adam Fraquhar z The British Library v roce 2010. FIDO je vytvořen v jazyce Python a je také použitelný na operačních systémech Windows, Linux a MacOs. Výstupy nástroje jsou ve formátu CSV. Byl navržen tak, aby mohl být snadno integrován do automatizovaných workflow a vracel výsledky identifikace velmi rychle. Jeho rychlost je ve srovnání s DROID skutečně značně vyšší, cenou za tuto rychlost však je ne úplně spolehlivá identifikace. Nástroj je však průběžně vyvíjen, jeho spolehlivost se průběžně zlepšuje. Rozvoj nástroje nyní obstarává organizace Open Preservation Foundation ve spolupráci s dalšími programátory.

7.4.3 Siegfried[66]

Nástroj Siegried byl vytvořen v roce 2014 digitálním archivářem Robertem Lehanem. Je vytvořen v jazyce Golang a je použitelný na

64 https://www.nationalarchives.gov.uk/aboutapps/pronom/droid-signature-files.htm
65 https://openpreservation.org/tools/fido/; https://github.com/openpreserve/fido
66 https://www.itforarchivists.com/siegfried

operačních systémech Windows, Linux a MacOs. Výstup je primárně vytvořen ve formátu YAML, ale je možné vytvořit i CSV nebo DROID CSV (CSV formátované dle CSV vydávané nástrojem DROID). Siegfried je z dominantních nástrojů nejmladší, byl vytvořen s vizí odstranění nedostatků jiných nástrojů. Identifikace formátu zde probíhá přiřazovacím algoritmem odlišným od předcházejících nástrojů. Siegfried formát identifikuje na základě PRONOM a přiřadí mu PUID a MIMEtype, snaží se přiřadit FDD (beta verze) a identifikátor z Wikidat.[67] Nástroj Siegfried přebírá definici signatures file z DROID (resp. z PRONOM). Nástroj je poměrně rychlý a vykazuje vysokou úspěšnost. Jeho většímu rozšíření brání zejména stav podpory, jelikož nástroj vznikl a je udržován víceméně jako soukromý projekt, byť spolupracuje s některými institucemi. Nástroj byl volitelně implementován do LTP systému RODA a je využíván v její uživatelské komunitě. Toto zázemí dodává nástroji stabilitu, byť jeho vývoj spočívá jen na omezené skupině osob v čele s tvůrcem. V českém prostředí využívá Siegfried jako primární nástroj pro identifikaci formátů vznikající digitální archiv Univerzity Karlovy. Také je zapojen do nástroje Bitcurator, který slouží pro forenzní analýzu obsahu celého serveru nebo file systemu a je také jedním z nástrojů, které lze využít v rámci aktivit na vstupu do repozitáře.[68]

7.4.4 Apache Tika

Za zmínku dále stojí nástroj Apache Tika, který je vytvořen v jazyce Java a vedle identifikace provádí u některých formátů i charakterizaci a extrakci textového obsahu. Dokáže identifikovat více než 1400 souborových formátů z taxonomie MIME typů. Užívá mechanismus identifikace, který je postaven na obdobě signatures (Apache Tika, 2022a), tj. v souboru hledá určitou sekvenci, pokud ji najde, přiřadí MIMEtype, pokud ji nenajde, pokusí se soubor identifikovat dle koncovky souboru nebo filename patterns (Oracle, 2010).

67 Web/databáze/agregátor dat sbírající informace o souborových formátech z různých zdrojů, používá jednoznačný identifikátor s prefixem Q následovaném číslem – např. Q217570 pro WAVE (LOC, 2023c).

68 https://bitcurator.github.io/

7.4.5 Identifikace souborového formátu

V předchozích odstavcích jsme se věnovali důležitosti identifikace souborových formátů a nástrojům, které ji realizují. Ve stručnosti nyní shrňme, jak identifikace technicky probíhá. V ideálním případě probíhá identifikace souborového formátu na základě signatures. DROID má defaultně nastavenou velikost souboru, kterou má prohledávat (65 536 bytů), v tomto úseku se magic numbers obvykle vyskytují. Není tomu tak však vždy a může dojít ke špatné identifikaci, je-li prohledávaná oblast omezená jako v případě defaultního nastavení DROID. Proto je možné nastavit buffer size, tj. jak velkou část souboru DROID prohledá (hodnota –1 prohledá celý soubor) s tím, že čím větší část souboru prohledává, tím je běh nástroje pomalejší. DROID občas využívá pouze rozlišení dle přípony, pokud není schopen identifikovat formát pomocí signatures. Nástroj FIDO umožňuje také nastavení velikosti prohledávané oblasti. Siegfried naproti tomu nemá (ale může se mu navolit) nastavený limit prohledávané oblasti a pracuje s tak velkou částí souboru, jak je třeba, aby dosáhl jednoznačné spolehlivé identifikace. Není-li formát identifikován pomocí signatures, může být určen na základě přípony souboru (např. neexistuje-li pro formát signature) a má-li koncovku .txt nebo by výsledek předchozích metod byl UNKNOWN (neznámý typ souboru), aplikuje Siegfried metodu „Text“, aby alespoň určil, zda se jedná o soubor s *a generic plain text* (a přiřadil puid x-fmt/111). Toto řešení Siegfriedu je chápáno jako nouzové, dává kurátorovi alespoň základní informaci, že identifikovaný soubor je textového charakteru se známým kódováním. Přestože jsou všechny tři hlavní identifikační nástroje – DROID, FIDO a Siegfried – založené na registru PRONOM, tj. přiřazují PUID na základě nalezené bitové sekvence v souboru, případně dle koncovky, výsledky identifikace se mezi těmito nástroji liší.

7.5 DROID a Siegfried – srovnání a možnosti využití

Protože je nástroj pro identifikaci souborových formátů jednou z klíčových součástí LTP systému, představíme v této část use case výbě-

ru nástroje na příkladu Archivu Univerzity Karlovy. Archiv stál před rozhodnutím, který nástroj pro identifikaci používat. Archiv využíval, obdobně jako Národní archiv ČR, ve svém informačním systému nástroj DROID. Původně se zdál jako optimální řešení a Archiv UK v jeho využívání sledoval ověřenou praxi dalších paměťových institucí nejen v ČR. Nástroj byl zakomponován do ingestového workflow, jeho výsledky se zapisovaly do AIP. Byl také zapojen do samostatného nástroje na validaci SIP balíčků ze spisových služeb. Během testování větších dávek vstupujících dokumentů byly zjištěny vážné výkonové problémy. Důvodem byla velká režie na spuštění procesu identifikace (natažení JRE, závislostí apod.), kdy se nástroj DROID nad každým vstupujícím SIP spouští znova, protože neumí pracovat v režimu trvalého spuštění – je to přirozený důsledek toho, že je napsán v jazyce Java.

Archiv UK přijímá poměrně značné množství digitálních archiválií (v blízkém horizontu lze očekávat přejímky desítek tisíc SIP v jedné dávce), které však mají malý datový objem. Režie na opakované spouštění DROID by představovala až 80 % času zpracování balíčku. Jak bylo řečeno výše, DROID je konfigurovatelný a lze určit, jak velkou část souboru má zpracovávat. Na příkladu souboru ve formátu PDF/A, který je pro Archiv UK (a další archivy v ČR) stěžejní, se ukázalo, že defaultní nastavení není dostatečné. Pro přesnější identifikaci variant PDF/A je třeba implicitní hodnotu (parametr maxBytesToScan) vhodným způsobem navýšit, přičemž je třeba vzít do úvahy poměr přesnost vs. rychlost. Zpracování jedné archiválie s datovým souborem o velikosti stovky kb trvalo deset sekund. Taková časová náročnost byla nepřijatelná (při předpokládané přejímce s 20 tis. SIP by jen identifikace trvala minimálně 55 hodin).

Tato zjištění nejsou samozřejmě nová ani překvapivá, podobnou zkušenost učinila celá řada institucí. Některé se ji rozhodly ignorovat, protože obvykle zpracovávají datově objemnější soubory, u kterých režie DROID hraje jen zanedbatelnou roli, případně v jejich workflow příliš nezáleží na rychlosti zpracování. V jiných institucích, jako je např. Archives of New Zeeland, šli cestou optimalizace komunikace s DROID, respektive vytvořili samostatný plugin, který umožňuje

obejít omezení.[69] Při rozhodování, zda DROID i s jeho režií ponechat či nikoli, byla zásadní i informace o jeho spolehlivosti.

Nejdříve byl proveden srovnávací nejčastějších identifikačních nástrojů (DROID, FIDO, Siegfried) nad ZIP-kontejnerem (compression level 9, tedy ten nejnáročnější). Ten obsahoval: 2× PDF 1.7 o velikostech 227 Kb a 873 Kb; 1× PDF 1.3 o velikosti 878 Kb; 1× PDF/A-1a o velikosti 8.4 MB. Test probíhal pomocí přímého spuštění na koncové stanici, která měla dostatečný výkon a nástroje běžely podle svého optimálního nastavení. V případě DROID a FIDO byl měřen jak čas při defaultním nastavení, tak při podrobném testu nad celým souborem. Siegfried nativně prohledával celý soubor. Výsledek ukázal, že DROID potřeboval k realizaci identifikace v obou režimech takřka 10 s, zatímco FIDO i Siegfried vykázaly výsledky pod 1 s, druhý jmenovaný dokonce méně než 0,3 s. I opakovaný test nad větším množstvím dat vykázal obdobné parametry. Konkrétně byl využit testovací korpus, který byl shromážděn komunitou Digital Corpora z USA (https://digitalcorpora.org).[70] Obsahuje 981 souborů zabalených po jednom do ZIP archivu. Zatímco DROID zpracovával výsledek více než jednu a půl hodiny, Sidegfried vrátil odpověď za minutu a půl. Při testu kvality výsledku, který se zaměřil již jen na DROID a Siegfried, byl použit vzorek obsahující 1800 druhů souborových formátů (soubor obsahoval mnoho exotických a zastaralých datových formátů). Srovnání obou nástrojů ukázalo, že se nástroje liší v 30 % výsledků (u 546 souborů). Z tohoto počtu souborů Siegfried u 406 dospěl k rozhodnutí, že soubor je textového charakteru a identifikoval ho jako Plain Text File (x-fmt/111), zatímco DROID vykázal formát jako UNKNOWN. Naopak Siegfried označil jako neznámé 93 souborů, které DROID chybně vyhodnotil jako Standard Data Format (fmt/1555) – přičemž tak určil i některé soubor s příponou txt. Siegfried určil několik formátů jako Plain text, i když je bylo možné určit přesněji a konkrétně, což se DROIDu povedlo.

Výsledky neumožňují jednoznačný závěr. Siegfried má větší snahu určit soubor za všech okolností, i když riskuje chybu, DROID v přípa-

69 https://github.com/rosetta-format-library/Rosetta.Droid-FormatIdentificationPlugin/releases

70 http://soton.corpora.openplanetsfoundation.org/govdocs_selected.tar.gz

dě pochybností na identifikaci rezignuje. Oba nástroje mají určitou chybovost, především pokud jde o raritně se objevující formáty (jde např. o přípony: accdb, ai, bf, bf12, com, dll, drv, eps, esl, fil, fsl, kic, m4a, mda, mdb, mdt, nef, odg, ovl, partial, sam, svg, tdb, vxd, xlam, xps). Dominantní formáty určují správně oba. V případě testu na korpusu formátů z Digital Corpora došlo k rozdílné identifikaci u 57 souborů z 981. Rozdíly byly obdobného charakteru jako u testu ve spolupráci s AHMP – pouze u osmi souborů Siegfried konstatoval, že jde o neznámý formát, zbytek se pokusil určit, zatímco DROID vykázal 35 neznámých formátů (vesměs se jednalo o soubory, které neměly záznam v registru a bylo tak těžké rozhodnout, zda byla identifikace správná). Pokud jde o výše zmíněné soubory zpracované pomocí softwaru T602, v testovacím souboru byly použity tři soubory s příponou 602 v odlišných verzích. DROID dva soubory vyhodnotil jako neznámé, jeden jako fmt/1555. Siegfried naopak dva soubory hodnotil jako Plain text a jeden jako neznámý. Závěrem lze konstatovat, že spolehlivost identifikace obou nástrojů je možné chápat jako obdobnou. Archiv UK kromě výkonových výsledků zohlednil a jako výhodnější vyhodnotil skutečnost, že je provozně výhodnější pracovat s informací o možnosti falešné identifikace textových souborů, jak ji provádí Siegfried, než s neznámou identifikací pomocí DROID. Představené výsledky nelze interpretovat tak, že by Siegfried poskytoval kvalitativně lepší výstupy než DROID, ani obráceně. Oba nástroje pracují odlišnými metodami, které je třeba při vyhodnocování výstupů zohlednit a neřídit se jimi bez uvážení.

7.6 Prostředky validace

V návaznosti na výsledky identifikace souborových formátů je dalším doporučeným krokem provedení validaci těchto výsledků. Potřeba validace se však netýká jen formátů dat, ale i správnosti a úplnosti metadatového popisu. Až po ověření všech informací se lze spolehnout na následující úkony prováděné již zcela v režii nástrojů LTP systému. Všechny úkony týkající se dlouhodobé správy digitálních dokumentů jsou založeny na důvěře v informace, které mají kurátoři

v systémech dlouhodobého uchování k dispozici. Validací se v prostředí digitální archivace obecně zjišťuje, nakolik je daný digitální objekt v souladu s konkrétním předpisem (specifikací) nebo standardem pro daný typ formátu nebo předpisu metadat. Zjišťuje se, zda jsou naplněny syntaktické a sémantické požadavky. Na rozdíl od identifikace je základní vlastností nástrojů pro validaci formátů jejich specifičnost. Protože se v tomto případě jedná už o podrobnější zpracování formátů, je evidentní, že jeden nástroj nebude umět validovat větší množství souborových formátů. Nástroje se proto specializují podle toho, s jakými formáty umí pracovat. To se v širším kontextu týká jak souborových formátů, tak metadatových standardů. Validace znamená ověření toho, že celé tělo souboru má strukturu a obsah, které jsou stanovené nějakým předpisem. V praxi to znamená, že k ověření validity musí nástroje analyzovat třeba i celý obsah souborů. Také v případě validačních nástrojů je třeba mít takové nástroje, které jsou udržované, podporované a dále rozvíjené. Příkladem nástroje, který mohl hrát užitečnou roli v oblasti dlouhodobého uchovávání digitálních dat je aplikace DPF Manager, která byla vyvinuta v rámci projektu Preforma. Primárním účelem této aplikace byla validace souborů z rodiny formátu *.tiff.[71] Vzhledem k rozšíření tohoto formátu jako archivního formátu pro digitalizaci, přinášela aplikace žádané funkce. Bohužel s koncem financování projektu byl ukončen i vývoj aplikace, jejíž poslední release byl vydán v září 2017 (Easy Innova, 2017). Aplikace je stále funkční a dostupná, ale její podpora je takřka nulová a z hlediska spolehlivosti pro LTP systémy je tedy nutné ji hodnotit jako nedoporučenou.

7.6.1 Validace metadat

V prostředí paměťových institucí v České republice jsou vytvářena metadata zejména podle dvou předpisů pro SIPy, jehož naplnění je třeba ověřovat. V případě knihoven jsou to Definice metadatových formátů (Standard NDK), u archivů jsou SIPy ze spisových služeb tvořeny podle Národního standardu pro elektronické systémy spisové

71 K validaci souborů ve formátu *.tiff je tak nyní nejčastěji doporučovaný nástroj JHOVE, který má více funkcí, ale validace Tiffů k nim také patří.

služby – NSESSS (Ministerstvo vnitra, 2023). Tento standard předepisuje strukturu metadat SIP a obsah jednotlivých elementů. Struktura obou typů SIP balíčků je komplikovaná a vyžaduje speciální validační nástroj, který ověří správné vytvoření. V případě Standardu NDK je to Komplexní validátor NDK, kterému budeme věnovat samostatnou část, protože rozsah jeho činnosti přesahuje jen metadatovou část SIP. V archivní síti existují dva validátory, které kontrolují splnění nároků na SIP dle standardu.

Národní archiv ČR jako provozovatel Národního digitálního archivu (a Archivního portálu) provozuje svůj validátor SIP ze spisových služeb, který mohou původci používat jednak v testovacím prostředí a jednak i pro jednotlivé SIP pomocí webového rozhraní. Kód validátoru však není uvolněn pro užití v jiných systémech. Druhý validátor vznikl z potřeb některých archivů vybudovat vlastní digitální archiv. Své síly spojili Archiv hlavního města Prahy a Archiv Univerzity Karlovy se společností LightComp v.o.s., která je dodavatelem části informačního systému Archivu UK. Společnými silami vytvořili validační nástroj ZAF, který byl na počátku roku 2022 uvolněn jako open source (licence Apache 2.0) k dalšímu využití.[72] Výhodou nástroje ZAF je, že jde o samostatnou aplikaci, kterou lze zapojit do dalších systémů, ale zároveň je schopen fungovat i nezávisle (pomocí CLI), a to i v režimu dávkového zpracování. Tato vlastnost je nutná kvůli testování a analytickému zkoumání vytvořených SIP mimo komplexní systémy. Validátor tak může být přímo zapojen do systému spisové služby a ověřovat validitu SIPů již u původce. Nástroj poskytuje strojově i lidsky čitelný výsledek validace, vysvětlení chyby a odkaz na předpis, podle kterého rozhodl, a lokalizaci chyby v rámci SIP XML. ZAF má více než 110 validačních pravidel, provádí kontrolu obsahu škodlivého kódu, kontrolu datové struktury, znakových sad, správnosti XML, kontrolu jmenných prostorů, kontrolu proti schématu a logickou kontrolu obsahu. Pravidla jsou aplikována podle typu SIP. Validátor ZAF usiluje o plné dodržování pravidel definovaných Národním archivem ČR, aplikuje je však rozdílným způsobem. Nejde o oficiálně uznaný nástroj, nemůže proto garantovat shodu s řešením Národního archivu. V praxi je však

72 https://validatorzaf.github.io/zaf/

shoda ve výsledcích validace velmi vysoká a blíží se 100 %. Výhodou zvoleného řešení je možnost jeho uplatnění v různých systémech. Validátor ZAF je i nadále průběžně aktualizován a zlepšován, možnost se zapojit do vývoje mají případně i další uživatelé.

7.6.2 Komplexní validátor NDK

Pro digitální objekty vznikající dle Standardu NDK je aktuálně k dispozici aplikace nazvaná Komplexní validátor NDK, softwarový nástroj, určený pro validaci digitalizačních balíčků (PSP/SIP) podle Definic metadatových formátů (DMF). Komplexní validátor je aplikací, vyvíjený Národní Knihovnou, konkrétně Oddělením pro standardy digitálních sbírek (dříve Oddělení pro standardy). Nástroj existuje jako grafická aplikace (GUI) ve verzích pro MacOS, Windows a Linux a také jako aplikace pro příkazový řádek (CLI). Validovat lze buď jeden balíček jako adresář či zip archiv, nebo skupinu balíčků v adresáři/zipu hromadně. Volitelně lze validaci omezit pouze na kontrolu metadatové části balíčku nebo kontrolu datové části balíčku včetně souladu profilu archivního obrazu ve formátu JPEG2000 vůči požadovaným parametrům dle Standardu NDK. V první verzi nástroj vznikl v roce 2017 a přinesl validaci podle aktuálních DMF pro monografie a periodika. Dále v sobě integroval nezávislé validační nástroje jako JHOVE, Jpylyzer, ImageMagick a Kakadu. Uživateli tak dává jistotu, že je SIP vytvořen správně jak z hlediska struktury, popisu i technických metadat, tak validnosti souborových formátů. Validátor je dále průběžně vyvíjen a doplňován o další formáty přijímané NDK. Od roku 2022 je ve vývoji nová verze, která by v podobě webové služby poskytovala validační službu bez nutnosti lokální instalace a s garancí dostupnosti vždy aktuálních validačních šablon. Ostrý provoz této nové služby je plánován na rok 2024. Od roku 2021 je Komplexní validátor ve zkušebním režimu využíván též jako součást ingestu LTP úložiště Národní knihovny, kde by měl do budoucna fungovat jako hlavní validační služba.

7.6.3 Validace PDF

Z hlediska archivů je primárním úkolem ochrana archiválií v digitální podobě. V současnosti jsou těmito archiváliemi z většiny výstupy ze

systémů spisových služeb, které jsou ve formátu PDF/A. Pro archivy je tak stěžejní validace tohoto souborového formátu, respektive různých standardů vzešlých z této rodiny formátů. Vedle souborů ve formátu PDF/A mohou do archivu vstupovat dokumenty v dalších, tzv. výstupních formátech, které jsou specifikovány legislativou (patří sem např. soubory s příponou *.png, *.tiff, *.jpeg, *.mp3, *.xml a další). Digitalizáty analogových archiválií jsou chápány jen jako kopie a jejich ochrana není formalizovaná a legislativně zakotvená. Také v případě validace souborů ve formátu PDF/A se odrazila možnost dosáhnout cíle více způsoby. I v tomto případě byla vyvolána podobou implementace řešení Národního archivu, které uživatelům nabízí webové prostředí pro testování jednotlivých souborů, ale neumožňuje dávkové zpracování. To probíhá až v ostrém prostředí, kde však nalezená chyba může vést k přerušení přejímky. V řešení Národního archivu je užívána kombinace několika nástroj pro validaci, zejména jde o různé verze validátoru 3Heights (a pak také validátor Callas PDFaPilot). Další možností, kterou některé archivy využívají, je aplikace VeraPDF,[73] která má za sebou širokou komunitu, poskytuje spolehlivé výsledky a je postavena na nekomerční bázi. Stejně jako výše zmíněný DPF Manager, také ona vzešla z výzkumu podporovaného Evropskou komisí v rámci projektu Preforma. Z několikaletého výzkumu realizovaného od poloviny druhého decennia 21. století vzešel nástroj s velkým množstvím funkcí. Na rozdíl od DPF Manageru však tato aplikace získala podporu od PDF Association, Open Preservation Foundation a Digital Preservation Coalition, tedy od hlavních profesních sdružení v oblasti LTP. Díky tomu je i po konci projektu dále rozvíjena a podporována. VeraPDF má jak webovou podobu, tak variantu pro dávkové zpracování zapojenou do dalších systémů. Univerzita Karlova nenasadila nástroj VeraPDF jen ve svém Archivním informačním systému, ale také pro kontrolu závěrečných kvalifikačních prací, které jsou od roku 2017 povinně odevzdávány v elektronické podobě ve formátu PDF/A. Dále je validátor integrován do studijního systému STAG, který využívá více českých vysokých škol. Zajišťují tím validnost kvalifikačních prací před uložením k trvalému uchování.

73 https://verapdf.org/

Výhodou validátoru je možnost nastavení vlastních parametrů, tak aby odpovídaly potřebám instituce a nastavené strategii dlouhodobého uchování digitálních dat.

7.7 Extraktory metadat

Dalším krokem v životním cyklu uchovávaných digitálních objektů je extrakce technických metadat. Většina digitálních objektů v sobě nese informace o svém vzniku, technických parametrech a další informace. Ty také patří k údajům, které potřebují kurátoři repozitářů znát, aby s jejich pomocí mohli nastavit ochranná opatření a následně je realizovat. Mnohé z technických informací o vlastnostech digitálního objektu patří k signifikantním (klíčovým) vlastnostem objektu, které je potřeba uchovat i po případných ochranných opatření jako je migrace. Pro kurátory je tedy nutné je znát a nastavit proces migrace tak, aby tyto vlastnosti nebo jejich interpretace v zobrazovacím software byla věrná. Cílem extrakce tak je získat ke každému objektu informace, pokud možno o všech jeho vlastnostech.

Vytěžení metadat může proběhnout již při produkci digitálních dokumentů nebo při vstupu do repozitáře nebo později, při jejich správě. Pokud repozitář dostane vytěžený popis vlastností již z produkce, je vhodné, aby ho přesto ověřil, pokud není zdroj zcela spolehlivý. Vytěžené výsledky jsou zapsány jako metadata do AIPů (repozitář je musí nějak konvertovat nebo zasadit do svého metadatového modelu). Z povahy potřebných informací je zřejmé, že pro extrakci technických metadat nejsou nutné zcela speciální nástroje, které by neměly uplatnění mimo oblast LTP. Běžně se tak užívají základní nástroje jako je ExifTool, který je populární i mimo oblast dlouhodobé archivace, nebo Metadata Extraction Tool, MediaInfo a další. Jejich funkce je však omezená na prostou extrakci. Funkce extrakce metadat byla proto často spojována s dalšími funkcemi užitečnými při dlouhodobém uchovávání a vytvořené nástroje jsou tak multifunkční a extrahování je jen jednou z více funkcí.

Nejrozšířenějším nástrojem s tímto určením, je již zmíněný JHOVE *(JSTOR/Harvard Object Validation Environment)*.[74] Nástroj dosáhl velkého rozšíření a je využíván v mnoha LTP systémech. Byl vyvinutý na Harvardské univerzitě ve spolupráci s organizací JSTOR (nástroj je funkčně dostupný od roku 2008). Cílem bylo automatizovat identifikaci, validaci a extrakci metadat digitálních objektů. Nástroj pracuje s několika datovými formáty, ale kromě *.pdf není schopen identifikovat a validovat kancelářské formáty, což je problém zejména v archivech. Výhodou jsou široké možnosti nastavení, např. podoba výstupu (délka, formát, obsah záznamu), způsob práce s objekty. Silnou stránkou je schopnost validovat soubory ve formátu *.tiff. Úspěšnost a rychlost identifikace a validace formátů je různorodá, a proto tento nástroj nenahradil speciální nástroje pro validaci a identifikaci. Dalším široce využitelným a již zmíněným je nástroj FITS *(The File Information Tool Set)*. Jde o wrapper software, který spojuje různé nástroje, mj. právě JHOVE, ExifTool a DROID. Pro zajímavost lze zmínit nástroj NZME *(New Zealand Metadata Extraction Tool)*[75] – jde o jeden z prvních nástrojů pro dlouhodobou ochranu vůbec. Vytvořen byl v Národní knihovně Nového Zélandu v roce 2003 se záměrem mít nástroj, který dokáže vyextrahovat ochranná metadata. Aktuálně není rozvíjen, byť je stále dostupný a lze ho využívat. V ČR aktivně využíván nebyl, ale v zahraničí měl poměrně široké pokrytí a svou funkci je u vybraných formátu stále schopen plnit.

7.8 Migrace a emulace

Migrace digitálního objektu z jednoho formátu do jiného je jednou ze základních strategií dlouhodobé archivace. Vedle toho se můžeme se stejným termínem potkat jako s označením přesunu dat z jednoho úložiště na jiné nebo replikace dat do další lokality apod. Migrace formátu je z technického hlediska pouhým převodem jednoho formátu na druhý, ale za tímto popisem se skrývá velmi komplikované

74 https://jhove.sourceforge.net/

75 https://meta-extractor.sourceforge.net/

workflow. Obvykle každý formát nebo i varianta formátu vyžadují specializovaný nástroj. Často je nutné takový nástroj nalézt nebo v případě raritnějších formátů nechat vytvořit. Pokud není už v době příjmu do archivu dokument ohrožený, není dopředu zřejmé, kdy a jak proběhne jeho migrace. Klíčovou vstupní informaci představuje správná formátová identifikace a vytěžená technická metadata. Kromě volby nástroje musí kurátoři posuzovat i signifikantní vlastnosti objektu a při realizaci migrace musí zohlednit takový postup, který tyto vlastnosti uchová (MCKINNEY a GATTUSO, 2014). Toto vše musí být ověřeno před tím, než je proces migrace spuštěn. K rozpoznání potřeby ochranného opatření slouží nástroje LTP systémů pro správu dat a plánování ochrany, v některých systémech (např. Rosetta) jsou implementované moduly umožňující simulaci migrace pomocí více nástrojů a její vyhodnocení. Na základě této simulace pak kurátoři rozhodují. V méně vybavených systémech musí simulace proběhnout mimo systém, což je uživatelsky náročnější, ale také v souladu s požadavky na provoz LTP systému.

Jedná se tedy o proces transformace digitálních formátů a schopnosti zobrazení obsahu vytvořeným datovým tokem tak, aby odpovídal původnímu dokumentu. Schopnost porozumět obsahu digitálního dokumentu do jisté míry závisí na způsobu zobrazení, na vnímání uživatelské komunity a samozřejmě na vlastnostech zobrazovacího software. Migrace formátů proto musí obsahovat kontrolu a posouzení změn obsahu předtím, než je provedena. Migrace je jen jedním z řešení zjištěného zastarávání formátu nebo jiného rizika (např. i licenčního). Vedle migrace byla vždy jednou ze základních strategií dlouhodobé archivace také emulace. V řadě oblastí je emulace řešením, které má vzhledem ke komplexnosti uložených dat smysl (archivace počítačových her a softwaru, archivace webu). Díky emulaci lze zajistit zobrazení všech vlastností, které byly určující pro dokument při vzniku. Emulace využívá schopnosti prezentovat datový objekt v prostředí, které simuluje prostředky v době vzniku dokumentu. Není tak třeba měnit samotný objekt, ale jen prostředky, kterými je zobrazován (nejedná se jen o software, někdy je nutné simulovat i funkce hardwaru). V praxi paměťové instituce naráží na praktické potíže s dostupností SW a na právní potíže s licencemi k nim. Z výše

řečeného plyne, že uvádět konkrétní nástroje pro migrace nebo emulace nemá smysl, jejich výběr probíhá vždy ad hoc. U běžných formátů jsou migrace schopny obvykle provést běžně dostupné nástroje (např. software Adobe v případě dokumentů ve formátech rodiny *.pdf). Obvykle se předpokládá, že tyto nástroje budou napojené na workflow systémů LTP jako celé aplikace, případně jako pluginu do těchto systémů. Jako nyní často používané nebo zvažované nástroje, které připadají v úvahu při migraci běžných formátů dat, lze uvést nástroje ImageMagick, ffmpeg, VLC nebo LibreOffice.

7.9 Long Term Preservation komplexní systémy v ČR

Paměťové instituce v ČR používají některé z komplexních systémů na zajištění dlouhodobého uchování digitálních dokumentů dle konceptu OAIS. Repozitářů vědeckých a dalších typů dat existuje poměrně značné množství, ale jen málo z nich naplňuje všechny prvky OAIS. Do workflow LTP systémů jsou zapojeny výše prezentované nástroje a napojení na znalostní báze. Není cílem této části podrobně představit jednotlivé softwary užívané pro správu a dlouhodobé uchovávání digitálních dat. Všechny fungují v rámci konceptu OAIS (viz kapitola ke konceptu OAIS a architektuře úložiště dle tohoto konceptu), dílčí technické a administrativní úkoly řeší různými způsoby, jsou uživatelsky více či méně komfortnější, některé mají více funkcí než jiné, ale ve způsobu fungování se zásadně neliší. U všech lze konstatovat, že využívají běžné metadatové standardy (METS, PREMIS a další) a ukládají data do informačních balíčků.

V odborné komunitě je zřejmě nejznámější implementací LTP systému v ČR LTP systém Národní knihovny vyvinutý a provozovaný v rámci Národní digitální knihovny. Systém byl vyvinutý společností AiP Safe a software nese název LTP Safe. Původně (tedy od roku 2012) využíval pro správu úložišť řešení IBM Information Archive, nyní přešel na řešení iRods. Systém původně vznikl jako přímo napojený na produkční linku NDK. Z obecného hlediska architektury systému bylo toto propojení v rozporu s doporučeními koncep-

tu OAIS, jednalo se hlavně o absence validací na vstupu do úložiště a spoléhání se na validační nástroje digitalizační linky. Tyto výtky se však v průběhu dalšího rozvoje podařilo odstranit. Za zmínku dále stojí poměrně rigidní vnitřní formát metadat, který na jednu stranu přináší vskutku komfortní dostupnost podrobných údajů o každé spravované intelektuálně entitě a jejích digitálních objektech, což usnadňuje práci kurátorů, na druhé straně každá změna a zavedení nového typu dokumentu znamenají velké nároky na vývoj. Systém je postavený na kombinaci ukládacích technologií v podobě diskových polí a magnetických pásek, což je v současnosti doporučený standard. Systém spravuje nižší desítky milionů digitálních objektů.

V návaznosti na fungování LTP systému NDK, který nebyl otevřen pro běžnou digitalizaci dalších knihoven, byl v rámci dotačního programu NAKI II realizován projekt ARCLib – komplexní řešení pro dlouhodobou archivaci digitálních (knihovních) sbírek, jehož cílem bylo vytvořit open source LTP systém, který by mohly užívat knihovny, jež se věnují digitalizaci (následný plánovaný rozvoj pro další typy paměťových institucí nebyl grantově podpořen). Konsorcium tvořily Knihovna Akademie věd, v.v.i., Národní knihovna České republiky, Moravská zemská knihovna v Brně a Masarykova univerzita. V roce 2020 byl vývoj dokončen a systém je užíván v Knihovně Akademie věd, plánuje se jeho využití v Muzeu umění Olomouc a je testován v MZK. Software pro správu úložiště ARCLib Archival Storage je nasazen v rámci budovaného Digitálního archivu Univerzity Karlovy.

Světově nejrozšířenějším opensource řešením pro zajištění LTP funkcí je software Archivematica. V rámci ČR ji jako jádro svého systému nasadil Národní archiv ČR jako součást Národního digitálního archivu. V NDK je Archivematica využívána především pro účely normalizace vstupních dat a jejich uložení. Není tak nasazena jako ucelené LTP řešení. Archivematica je vyvíjena kanadskou firmou Artefactual Systems. Na vývoji se podílí i další instituce – UNESCO, City of Vancouver Archives, Mezinárodní měnový fond a další významné knihovny a univerzity z celého světa. Archivematica je typická svým systémem mikroslužeb *(microservices)*, což jsou v podstatě programy, které vykonávají jednotlivé funkce příjmu a zpracování dat. Uživatelé mohou naprogramovat vlastní, pokud jim ty původní nevyhovují.

Svou strukturou se neliší od jiných LTP systémů. Disponuje lokální databází formátů a jejich rizik, spravuje uložená data a verzuje uložená data, takže umožňuje v budoucnu přistoupit rovněž k emulaci. Neumí však spravovat více úložišť v nezávislých lokacích.

Posledním LTP systémem, který je v ČR využíván, je systém RODA.[76] Stejně jako Archivematica ani on není použit v plném rozsahu, ale pouze jeho část pro správu úložišť a souborových formátů. Nasazen je v budovaném Digitálním archivu hlavního města Prahy. RODA je software původně vyvinutý pro potřeby portugalských archivů. Dostalo se mu však širšího uplatnění. Je úplným digitálním repozitářem poskytující funkcionality všech hlavních jednotek referenčního modelu OAIS. Za zmínku stojí nativní podpora metadatových standardů pro archivnictví (EAD, e-ARK), samozřejmě vedle jiných běžných standardů.

76 https://www.roda-community.org/

8 Dlouhodobé uchovávání digitalizátů v českých knihovnách

Mezi hlavní povinnosti knihoven, obdobně jako i jiných typů paměťových institucí, patří ochrana jejich fondů před odcizením, poškozením či nepříznivými vlivy prostředí. Jako prostředek ochrany knihovního fondu pak knihovní zákon vedle preventivních a restorativních opatření přímo uvádí „*jejich převedení na jiný druh nosiče, je-li to třeba k jejich trvalému uchování*". (Zákon č. 257/2001 Sb., § 18 písm. c) Za touto formulací se mj. skrývá i digitalizace fondů. Kromě ochrany tradičních fyzických fondů se tak knihovny musí zabývat i ochranou fondů uchovávaných v digitální podobě.

České knihovny patřily v mezinárodním srovnání k průkopníkům digitalizace knihovních fondů, velké objemy digitálních dat v jejich fondech proto dle očekávání pochází z aktivit zaměřených právě na ochranné reformátování. Vedle digitalizátů tištěných dokumentů ale tvoří digitální kulturní dědictví uchovávané v českých knihovnách také velké množství e-born dokumentů. Jedná se například o digitální data odevzdávaná na pevných nosičích (nejčastěji na optických discích) v rámci povinného výtisku, data z archivace webu prováděné od roku 2001 NK ČR nebo e-born textové dokumenty předávané do knihoven na základě individuálních dohod s vydavateli.[77] V následujících odstavcích se pro zjednodušení budeme věnovat zejména datům vzniklým digitalizací fyzických fondů, základní principy digitální ochrany jsou obdobně uplatnitelné též pro e-born dokumenty.

77 Na tomto místě je vhodné podotknout, že vzhledem k chybějící legislativě, která by upravovala povinné odevzdávání elektronických dokumentů do knihoven, se naprostá většina těchto dokumentů do knihoven vůbec nedostane a pro budoucí uživatele je tak zřejmě z velké části ztracena. Individuální smlouvy s vydavateli proto můžeme v současné době očekávat zejména u akademických knihoven.

8.1 Počátky digitalizace a digitálních dat v knihovnách

První projekty zaměřené na digitalizaci fondů a produkci digitálních dat se v českých knihovnách datují do první poloviny devadesátých let. Již v roce 1992 přistoupila Národní knihovna ČR k programu UNESCO Paměť světa (Memoriae Mundi), v dubnu 1993 pak představila vůbec první kompletně digitalizovaný rukopis na CD-ROM na světě. Provoz vlastního digitalizačního pracoviště byl v NK ČR zahájen v roce 1996 (KNOLL, 1999). V první fázi se digitalizační aktivity zaměřovaly na nejvzácnější rukopisy, později také na staré tisky a kartografické historické dokumenty. Na konci devadesátých let byla v rámci programu Kramerius zahájena rovněž digitalizace novodobých knihovních fondů ohrožených degradací kyselého papíru. Reformátování mělo nejprve podobu digitalizace mikrofilmů, posléze šlo také o přímou digitalizaci z fyzických předloh. Samotná digitalizace však byla v tomto období ještě považována za vhodnou zejména pro zpřístupňování dokumentů, nikoliv za plnohodnotný způsob reformátování dokumentů, sloužící jako trvalá náhrada za originální dokumenty (MELICHAR a HUTAŘ, 2013).

Digitalizované dokumenty byly zprvu ukládány nejčastěji na optické disky, které na konci devadesátých let doplnily magnetické pásky, pevné disky nebo disková pole. Ochrana dat se potom v lepším případě omezovala na neporušenost bitstreamu jednotlivých souborů a vícenásobné zálohování. Základní strategií NK ČR ve sledovaném období tak byla zejména *kombinace vhodných mechanik, režimu záznamu a médií* (KNOLL a PSOHLAVEC, 2002), přičemž digitální archiv na CD-R discích byl tvořen ze dvou archivních kopií uložených ve dvou lokalitách a jedné uživatelské kopie pro rutinní použití. Pokud vezmeme v úvahu pečlivý výběr a kontrolu disků včetně měření vzorků jednotlivých šarží a systém měření kvality vypálených disků společně s vhodně nastavenými klimatickými podmínkami pro jejich uchovávání (KNOLL a PSOHLAVEC, 2002), jednalo se v dobovém kontextu o vyspělé a funkční řešení.

Nárůst dat, společně s klesající kvalitou optických disků dostupných na trhu, vedl NK ČR v roce 1999 k pořízení robotické magne-

topáskové knihovny (MELICHAR a HUTAŘ, 2013). Data byla ukládána vždy na třech kopiích pásek, přičemž dvě byly dostupné online a jedna offline, uložená v jiné lokalitě. Páskový robot pak do roku 2003, kdy došlo k oddělení archivních a zpřístupňujících dat a vytvoření digitálních knihoven Kramerius (pro novodobé dokumenty) a Manuscriptorium (pro historické dokumenty), sloužil vedle archivace také ke zpřístupňování dat uživatelům. Motivací pro oddělení archivních a uživatelských kopií bylo vedle pomalé odezvy magnetopáskového robota také ohrožení archivních dat hackerskými útoky, jimž byl systém opakovaně vystaven (KNOLL et al., 2004).

8.2 Rozvoj digitalizace a digitálních dat v knihovnách po roce 2000

Významným akcelerátorem systematické digitalizace v českých knihovnách se staly povodně v roce 2002. První léta nového tisíciletí byla rovněž počátkem celosvětové změny pohledu na digitalizaci jako na prostředek pro zpřístupňování dokumentů a digitalizace začala být přijímána jako plnohodnotná metoda reformátování ohrožených dokumentů. V tomto období začaly své fondy digitalizovat další české knihovny, například MZK, KNAV nebo Moravskoslezská vědecká knihovna v Ostravě. Významným impulzem pro digitalizační aktivity knihoven byla finanční podpora dotačních programů VISK vytvořených v roce 2000, zejména podprogramu VISK 7 (Národní program mikrofilmování a digitálního zpřístupňování dokumentů ohrožených degradací kyselého papíru Kramerius) a VISK 6 (Národní program digitálního zpřístupnění vzácných dokumentů Memoriae Mundi Series Bohemica). V souvislosti se vznikem open source digitální knihovny Kramerius v roce 2003 byly vytvořeny nové národní metadatové standardy pro digitalizaci novodobých dokumentů na bázi XML, tzv. DTD, pro monografie a periodika (MELICHAR a HUTAŘ, 2014). Povinné využití uvedených DTD pro digitalizaci v rámci VISK 7 bylo prvním významným krokem k standardizaci digitálních dat v českých knihovnách.

V mezinárodní perspektivě byla nultá léta nového tisíciletí v knihovnách obdobím vytváření prvních systémů na logickou ochranu digitálních dat, které by zajistily jejich skutečně dlouhodobou využitelnost. V roce 2003 spustila Nizozemská královská knihovna ve spolupráci se společností IBM systém e-Depot, vůbec první knihovní dlouhodobé úložiště, založené na referenčním modelu OAIS (VAN WIJNGAARDEN, 2010). Po roce 2005 docházelo také k rozvoji mezinárodních metadatových standardů jakými jsou METS, PREMIS, MIX, MODS, TEI P5 apod., které se později začaly využívat také v ČR.

Potřeba systematicky řešit dlouhodobou ochranu digitálních dat v NK ČR vedla v roce 2005 k vytvoření Referátu pro digitální knihovnu NFS (*Novodobé fondy a sbírky*) (HUTAŘ, 2012, s. 48). Z útvaru o jednom úvazku vznikl v roce 2008 Odbor digitální ochrany,[78] jenž se významným způsobem podílel na zavádění zásad dlouhodobé ochrany v prostředí NK ČR a zároveň se stal de facto metodickým centrem pro oblast digitální ochrany pro české knihovny.

V reakci na mezinárodní rozvoj standardů bylo v českém prostředí v roce 2007 zavedeno v omezené míře využívání administrativních a technických metadat MIX a PREMIS. (HUTAŘ, 2008) Ve stejné době došlo také k rutinnímu využívání kontrolních součtů MD5, umožňujících automatickou kontrolu integrity dat, jako součásti archivních balíčků vytvářených v rámci programu VISK 7.

V druhé polovině nultých let již ale také začalo na své limity narážet dosavadní řešení archivace dat v NK ČR, založené zejména na bitové ochraně dat a politikách vytváření kopií na vhodných HW technologiích. V důsledku kombinací selhání technologií a lidského faktoru tak NK ČR čelila již přímo konkrétním ztrátám dat. Chyběly zejména nástroje pro efektivní správu dokumentů i nástroje pro podporu logické ochrany (MELICHAR a HUTAŘ, 2014).

V rámci plnění *Koncepce rozvoje knihoven v České republice na léta 2004 až 2010* formulovala v roce 2005 NK ČR ve spolupráci s Ministerstvem kultury *Koncepci trvalého uchování knihovních sbírek tradičních a elektronických dokumentů v knihovnách ČR do roku 2010*, v níž byla

78 V letech 2011–2023 existoval pod názvem *Odbor digitálních fondů*, od června 2023 je označen jako *Odbor novodobých digitálních sbírek*.

poprvé jasně deklarována potřeba dlouhodobé ochrany digitálních dokumentů a zároveň konstatována nutnost „*zakoupit nebo vytvořit dostatečně mohutný a komplexní systém*“ pro ukládání, zpřístupnění a ochranu digitálních dokumentů a příslušných metadat (Koncepce trvalého uchování, 2005, s. 8). Dokument tak poprvé představil koncept pozdější Národní digitální knihovny.

K jeho realizaci došlo v letech 2009–2014, kdy se NK ČR a MZK společně podílely na řešení projektu Vytvoření Národní digitální knihovny, financovaného z Integrovaného operačního programu EU. Vedle vytvoření digitalizačních kapacit v NK ČR a MZK pro masovou digitalizaci knihovních fondů (26 milionů stran v letech 2012–2014) bylo důležitým výstupem projektu důvěryhodné digitální úložiště vytvořené v souladu s parametry ISO 14721 (OAIS). Závazek k digitální archivaci byl v listopadu 2011 vtělen také do zřizovací listiny NK ČR: „*Formuluje strategie a postupy dlouhodobé ochrany elektronických dokumentů a provozuje důvěryhodné digitální úložiště*.“ V rámci přípravných prací na projektu byly v průběhu roku 2012 zavedeny nové standardy dat i metadat, tzv. Standard NDK a vytvořen systém ČIDLO pro trvalou identifikaci digitálních dokumentů (CUBR, 2016, s. 44–45).

8.3 LTP úložiště NK ČR

V průběhu roku 2012 byl v NK ČR zahájen pilotní provoz dlouhodobého úložiště (též LTP úložiště), vytvořeného v souladu s principy normy ISO 14721 (OAIS). Do plného provozu LTP úložiště NK ČR oficiálně přešlo v roce 2013 a zprvu sloužilo zejména k ukládání dat vytvářených v rámci společného digitalizačního projektu NK ČR a MZK. Později začalo úložiště sloužit k ukládání dat vytvořených knihovnami v rámci dotačního programu VISK 7 a ukládání starších dat z provenience NK ČR (KVAŠOVÁ, 2017; 2018).

Dlouhodobé úložiště NK ČR se skládá ze tří komponent. Samotného LTP systému, tvořeného aplikacemi LTP SAFE (systém pro uložení a zobrazení) a LTP WF (systém pro práci s archivními balíčky), Transformačního modulu (výkonná komponenta provádějící akce nad archivními balíčky) a systému IBM Information Archive, řídí-

cího vlastní ukládání dat na pásky. Ukládání dat mělo probíhat dle nastavených politik ve 3 identických kopiích na 3 geograficky oddělených lokalitách – 2 online kopie na lokalitách v Praze a 1 offline kopie v Brně. Jako médium byly zvoleny nepřepisovatelné WORM *(Write once read many)* pásky typu LTO5. V současné době probíhá migrace na novější generaci pásek LTO7. Po roce 2019 zároveň došlo k nahrazení komerčního systému IBM Information Archive (IIA) jako datového zdroje pro LTP SAFE za open source integrativní platformu iRODS, jež nově využívá IIA pouze jako jedno z možných datových úložišť (OSTRÁKOVÁ a VOZÁR, 2019, s. 19–20).

Dlouhodobou politikou LTP úložiště NK ČR je využívání informačních balíčků vytvářených v souladu se Standardem NDK (VAŠEK, 2017). Můžeme mluvit o tzv. homogenním archivu, který má předem definované specifikace vstupních a archivačních balíčků a striktní formátovou politikou. Politika je motivována jednak usnadněním péče o digitální data v dlouhodobém horizontu a zároveň i minimalizací nutnosti normalizovat data během jejich ukládání do úložiště (KVAŠOVÁ, 2017, s. 23–24). Archivní AIP balíčky v LTP úložišti NK ČR odpovídají vstupním SIP balíčkům, obohaceným navíc o administrativní metadata s informacemi o akcích provedenými nad AIP balíčkem v průběhu jeho uložení v dlouhodobém úložišti (informace o validaci, aktualizace apod.). V současné době (2023) podporuje LTP úložiště NK ČR ukládání SIP balíčků vytvořených dle DMF pro tištěné monografie a periodika, v testovacím režimu pak rovněž pro elektronické monografie a elektronická periodika. Vstup není v současné době zatím nastaven pro SIP balíčky dle DMF pro zvukové dokumenty – gramofonové desky ani fonografické válečky. Vedle Standardu NDK jsou podporovány archivní balíčky vytvořené dle standardizace VISK 6,[79] vycházející ze standardu TEI P5.

V roce 2022 prošlo LTP úložiště NK ČR self auditem pomocí metodiky Digital Preservation Coalition Rapid Assessment Model (DPC RAM).[80] Hodnocením v 11 dílčích oblastech zaměřených na organizační zajištění (A. Životaschopnost organizace, B. Politika a strategie,

79 https://www.manuscriptorium.com/sites/default/files/docs/manuscriptorium_visk6_definice.pdf

80 https://www.dpconline.org/digipres/implement-digipres/dpc-ram

C. Právní rámec, D. IT zajištění, E. Stále zlepšování a F. Komunita) a zajištění služeb (G. Akvizice, přenos obsahu a ingest, H. Bitová ochrana, I. Ochrana obsahu, J. Správa metadat a K. Vyhledávání a přístup) prošlo úložiště s průměrným hodnocením 2,36 (na stupnici 0–4),[81] přičemž v sedmi hodnocených oblastech dosáhlo úrovně 2 (základní zabezpečení digitální ochrany) a ve čtyřech oblastech úrovně 3 (vyspělá úroveň digitální ochrany). Významná rizika byla shledána například v oblasti absence trvale alokovaného základního rozpočtu dedikovaného výhradně pro potřeby digitální ochrany, nedostatečném personálním zabezpečení nebo aktuálním omezení počtu geograficky oddělených lokalit úložiště na dvě pražské lokality namísto původních tří (zahrnujících Brno). Na základě výsledků self auditu bylo rozhodnuto o jeho každoročním opakování, které umožní jak pravidelně vyhodnocovat dopady přijatých nápravných opatření, tak i pružně reagovat na nová rizika.

8.4 Standard NDK

Nové Definice metadatových formátů, vytvořené v souvislosti s projektem Vytvoření Národní digitální knihovny, navazovaly na předchozí DTD pro periodika a pro monografie, vycházely však již plně z mezinárodních metadatových standardů vyvíjených a udržovaných Kongresovou knihovnou – využily METS jako kontejnerový formát, PREMIS pro administrativní technická metadata, MIX pro technická metadata u rastrových obrazů, MODS a Dublin Core pro bibliografická metadata. Tzv. národní aplikační profil, zajišťující lokalizaci a optimalizaci použitých mezinárodních standardů pro konkrétní potřebu v českých knihovnách, již oproti předchozím generacím národních standardů pro digitalizaci novodobých fondů kladl velký důraz na administrativní a technická metadata, důležitá pro zachování autenticity a vůbec pro podporu dlouhodobé archivace dokumentů. Klíčové bylo zapojení standardu PREMIS pro popis datových souborů, jejich

81 0 = minimální povědomí o nutnosti digitální ochrany; 1 = Povědomí o nutnosti digitální ochrany; 2 = Základní zabezpečení digitální ochrany; 3 = Vyspělá úroveň digitální ochrany; 4 = Optimální úroveň digitální ochrany.

původu, použitých nástrojů a všech operací provedených se soubory během jejich existence. Specifikace vznikla na základě porovnání nejlepší soudobé zahraniční praxe a v přímé spolupráci s národními knihovnami Nizozemska, Norska a Finska a s Kongresovou knihovnou (CUBR, 2016, s. 52).

Původní dva národní aplikační profily pro digitalizáty tištěných monografií a periodik byly v průběhu následujících let doplněny o DMF pro zvukové dokumenty (gramofonové desky a fonováleč ky), DMF pro elektronické dokumenty (periodika a monografie) a mnohokrát aktualizovány v návaznosti na zkušenosti s jejich využíváním, aktualizace použitých mezinárodních standardů nebo dle nových požadavků uživatelské komunity. Pro potřeby správy a rozvoje Standardu NDK ustanovila Národní knihovna tzv. Formátový výbor, sdružující odborné pracovníky Národní knihovny a zástupce uživatelů Standardu z českých knihoven. Součástí Standardu NDK je mimo jiné rovněž profil JPEG 2000 pro archivní i uživatelská obrazová data, který v porovnání s konkurenčními JPEG nebo TIFF umožňuje lepší kompresní poměry při matematicky bezeztrátové (archivní kopie) či vizuálně bezeztrátové (uživatelská kopie) kompresi a zároveň se jeví jako prokazatelně odolnější vůči bitovým ztrátám než formáty PNG nebo JPEG (HUTAŘ et al., 2016, s. 8).

Standard NDK je od roku 2012 využíván v rámci interní digitalizace NK a MZK, od roku 2013 je jako jediný povolen rovněž v rámci dotačního podprogramu VISK 7 a byl obecně adoptován českými knihovnami v rámci jejich digitalizačních aktivit. Jeho užívání napříč knihovnami je významným úspěchem, který zajišťuje potřebnou konzistenci dat a interoperabilitu mezi jednotlivými digitalizačními projekty a prezentačními systémy (např. mezi open source nástroji ProArc, ARCLib a Kramerius a jejich komerčními alternativami). Důraz na administrativní a technická metadata, stejně jako povinné kontrolní součty MD5 vytvářené v průběhu vzniku balíčku (digitalizace), pokládá solidní základ pro zajištění základní úrovně ochrany u datových balíčků. (HUTAŘ et al., 2016)

Důležitým doplňkem Standardu NDK je *Metodika pro tvorbu balíčků SIP se zaměřením na digitalizáty tištěných dokumentů* (CUBR et al., 2020), vydaná Národní knihovnou a certifikovaná Ministerstvem

kultury České republiky v roce 2020.[82] Metodika doplňuje zejména specifické procedurální postupy pro užití Standardu NDK při digitalizaci a tvorbě SIP balíčků a popisuje práci s konkrétními nástroji pro optimální podobu datových balíčků určených k archivaci. V době práce na této knize prochází recenzním řízením nově aktualizovaná verze této metodiky.

8.5 ČIDLO

Národní systém trvalé identifikace ČIDLO (*Český systém pro IDentifikaci a LOkalizaci dokumentů digitálního kulturního dědictví)* je provozován NK ČR od roku 2012. Systém je založen na mezinárodním standardu URN:NBN a sestává ze softwarového nástroje CZIDLO a systému pravidel, vydaných NK ČR a certifikovaných MK ČR v letech 2015 a 2018 pod názvem *Metodika pro přidělování a správu životního cyklu unikátních perzistentních identifikátorů digitálních dokumentů podle standardu URN:NBN* (VAŠEK et al., 2018). Metodika upravující nejen přidělování, ale také životní cyklus identifikátorů, a tím potažmo i celých identifikovaných digitálních dokumentů, významně napomáhá v naplňování cílů dlouhodobé ochrany a je de facto nedílnou součástí Standardu NDK. Pravidla pro registrátory vedle autenticity mezi jednotlivými digitálními instancemi dokumentu významně přispívají též k zachování konzistence mezi archivními balíčky a uživatelskými balíčky, když předepisují povinnost zapsat identifikátor URN:NBN do metadat uživatelského i archivního balíčku a zároveň brání signifikantním změnám v popisu dokumentu bez přidělení nového identifikátoru.[83] Využití identifikátoru URN:NBN je povinné při

82 Aktualizovaná verze metodiky je připravována k recertifikaci a vydání v roce 2023.

83 *„3) Registrátor musí zajistit, aby tentýž přidělený identifikátor byl zapsán do metadat obou derivátů digitalizovaného dokumentu (archivní / uživatelský balíček). Archivní balíček je následně uložen do digitálního repozitáře (dlouhodobá ochrana) a uživatelský balíček do digitální knihovny (zpřístupňování uživatelům. 4) Registrátor musí zajistit, aby tyto balíčky nebyly nikdy signifikantně změněny. Pokud dojde k jejich signifikantní změně, musí si registrátor nechat přidělit nové URN:NBN. Za signifikantní změny se považuje zejména rozdělení na dva nebo více dokumentů nebo naopak sloučení více dokumentů do jednoho, dále například změna identifikačních údajů (např. názvu), zatímco např. překlep v těchto údajích nikoliv (jednotlivé sporné případy je třeba konzultovat s kurátorem).“ – Zdroj: https://resolver.nkp.cz/web/#tab1*

digitalizaci dle Standardu NDK a v současné době jej využívá více než sto paměťových institucí.

8.6 Komplexní validátor NDK

Klíčovým nástrojem pro kontrolu výstupů digitalizačních projektů vůči Standardu NDK je od roku 2017 Komplexní validátor NDK. Open source aplikace zpřístupněná na platformě GitHub byla NK ČR vyvinuta v letech 2016 až 2017 v podobě grafické aplikace (GUI) i jako aplikace pro příkazový řádek (CLI).

Dostupnost garantované validační služby je dalším důležitým krokem ve standardizaci dat vytvářených v knihovnách, a podobně jako samotná existence Standardu NDK významným způsobem snižuje náklady na dlouhodobou ochranu digitálních dat v dlouhodobém horizontu. Jako jednomu z významných nástrojů LTP je mu věnována samostatná podkapitola (viz 7.6.2) v rámci předchozí kapitoly, věnované nástrojům LTP, používaných v české praxi.

8.7 Národní koncepce dlouhodobé ochrany digitálních dat v knihovnách

V rámci přípravy *Koncepce rozvoje knihoven ČR na léta 2017–2020* byla v roce 2016 na základě zadání Ústřední knihovnické rady zpracována *Národní koncepce dlouhodobé ochrany digitálních dat v knihovnách* pro období 2016–2020. Autoři v dokumentu upozorňují na velké množství projektů zaměřených na digitalizaci a tvorbu digitálních dat i na přetrvávající nedostupnost digitální ochrany ve velké většině českých knihoven a definují kroky potřebné pro její účinné zajištění v národním měřítku. Koncepce navrhovala řadu dodnes nenaplněných opatření, např. dostupnost garantovaných datových center, budování sítě certifikovaných úložišť v jednotlivých institucích nebo zřízení Metodického centra pro dlouhodobou ochranu digitálních dat. Přestože zárodek Metodického centra můžeme spatřovat ve standardizační činnosti Oddělení pro standardy NK ČR, ostatní navrhované činnos-

ti, jako např. zajištění a podpora certifikace digitálních úložišť nebo vytváření národních strategií včetně plánování financování, přípravy a správy grantových výzev, dalece přesahují jeho působnost.

Navzdory své nenaplněnosti zůstává koncepce aktuálním dokumentem, přítomným například v řadě opatření *Koncepce rozvoje knihoven na roky 2021–2027*. Z dosud realizovaných opatření si zaslouží vyzdvihnout kupříkladu podpora open source řešení pro zajištění logické ochrany digitálních dat (viz ARCLib), finanční podpora při instalaci a zavádění LTP systému či konsolidace dříve pořízených dat (VISK 7) nebo zavedení kvalifikačního standardu pro kurátora digitálních dat do Národní soustavy povolání.

8.8 Projekt a systém ARCLib

Příležitost diverzifikovat možnosti dlouhodobého uložení dat v českých knihovnách a zajistit dostupnost nekomerčního LTP řešení pro menší a středně velké knihovny přinesl v letech 2016–2020 projekt *ARCLib – komplexní řešení pro dlouhodobou archivaci digitálních (knihovních) sbírek*.[84] Cílem tohoto NAKI projektu řešeného ve spolupráci NK ČR, MZK, KNAV a MUNI bylo vyvinout nástroj na bázi open source s plnou podporou OAIS modelu, který by přímo poskytoval podporu pro všechna nejpoužívanější softwarová řešení využívaná pro produkci, uložení a zpřístupnění dat v českých knihovnách. Konkrétně se jedná o data vytvářená v souladu se standardizací NDK, data z produkčního systému ProArc a data z digitálních knihoven Kramerius nebo systému DSpace, využívaného ke zpřístupnění digitálních dokumentů především v univerzitním prostředí.[85] Díky podpoře standardizace NDK měl být systém plně interoperabilní s LTP úložištěm NK ČR, pro nějž by tak mohl vystupovat v roli možného nástupnického úložiště v případě ukončení provozu. Díky své otevřenosti by poskytoval možnost integrace i dalších nástrojů a systémů pro potřeby paměťových institucí jiných typů. Pilotní provoz systému

84 https://arclib.cz/

85 https://github.com/LIBCAS/ARCLib/wiki/predefined-profiles

ARCLib v KNAV přešel po skončení projektu v roce 2020 do rutinního provozu, systém je proto i po skončení projektu NAKI nadále rozvíjen a podporován aktivní uživatelskou komunitou a je k dispozici všem zájemcům na platformě GitHub[86] (LHOTÁK et al., 2019).

První zkušenosti s provozem systému ARCLib naznačují, že představa vytvoření open source systému, který by byl dlouhodobě provozovatelný jako plnohodnotný LTP systém v menších knihovnách, byla zřejmě příliš optimistická. Řešení je tak vhodné spíše pro středně velké a velké knihovny se stabilním IT zázemím, stabilním financováním a kvalifikovanými digitálními kurátory.

Projekt NAKI ARCLib přispěl vedle samotné aplikace ARCLib k digitální ochraně v českých knihovnách také dvěma důležitými certifikovanými metodikami zaměřenými na logickou a bitovou ochranu. Jedná se o *Metodiku logické ochrany digitálních dat* z roku 2018 (HUTAŘ et al., 2018) a *Metodiku bitové ochrany digitálních dat* z roku 2018 (RŮŽIČKA et al., 2019). Obě metodiky obsahují vedle praktické části zpracované na míru systému ARCLib také rozsáhlé teoretické části, které jsou obecně platné nezávisle na využitém archivačním systému.

O finanční podporu pro instalaci a implementaci systému ARCLib (či jiného softwarového nástroje) pro zajištění dlouhodobé archivace digitálního obsahu mohou knihovny od roku 2022 žádat v rámci dotačního programu VISK 7. Dotaci je možné využít na instalaci a zprovoznění systému, normalizace dat ze starších digitalizačních projektů nebo školení odborného personálu, zajišťujícího provoz dlouhodobého úložiště, migrace dat a další procesy dlouhodobé archivace. Z uvedeného dotačního programu však nelze hradit každoroční provozní náklady, návaznou hardwarovou infrastrukturu a v případě komerčních nástrojů ani licenční poplatky.[87]

86 https://github.com/LIBCAS/ARCLib

87 https://visk.nkp.cz/visk-7

8.9 Dlouhodobá archivace v knihovnách v současných strategických dokumentech

Klíčovými strategickými dokumenty pro oblast dlouhodobého uchovávání digitálních dat v knihovnách jsou na národní úrovni v současné době zejména *Státní kulturní politika 2021–2025+* a *Koncepce rozvoje knihoven v České republice na léta 2021–2027*. Státní kulturní politika se tématem zabývá v rámci cíle *2) Účinná péče o kulturní dědictví*, v němž definuje specifické cíle, zaměřené v oblasti dlouhodobé ochrany digitálních dat v knihovnách výhradně na rozvoj LTP úložiště NK ČR, podporu provozu Webarchivu, podporu nástrojů pro konverzi a trvalé uchování elektronických dokumentů na pevných nosičích a dokončení procesu legislativy povinného výtisku elektronických publikací, včetně vybudování systému pro jejich příjem, zpracování, dlouhodobé uchovávání a zpřístupnění. *„Odpovídající rozvoj LTP (Long-term Preservation) úložiště Národní knihovny ČR odpovídající objemu digitálních dokumentů, které jsou součástí kulturního dědictví ve správě knihoven. Účelem je komplexní archivace digitálních dokumentů důležitých pro záchranu bohemikálního kulturního dědictví vzniklých digitalizací tištěných knihovních fondů i e-born dokumentů (e-knihy, online časopisy, web – obsah webarchivu,…). Součástí prací bude i rozšíření o podporu uchování obsahu dokumentů na pevných nosičích v souladu se specifikacemi vytvořenými v rámci opatření."* Oproti tomu Koncepce rozvoje knihoven v rámci specifického cíle 13) *Knihovny shromažďují, dlouhodobě uchovávají a přívětivě zpřístupňují různé formy kulturního obsahu v analogové i digitální formě* definuje opatření 51) *Zajistit provoz a rozvoj infrastruktury LTP úložišť v NK a největších knihovnách. Podpora přípravy LTP systému NK ČR na příjem a archivaci digitálních dokumentů ze všech knihoven v ČR včetně jejich certifikace a jejího pravidelného obnovování*.

Porovnání obou dokumentů naznačuje pokračující[88] koncepční rozpor v přístupu k řešení dlouhodobého uchovávání dat v knihovnách. Na jedné straně stojí silné centrální řešení provozované NK ČR, na druhé straně síť více diverzifikovaných certifikovaných dlou-

88 Srov. s (KVAŠOVÁ, 2018, s. 23–24)

hodobých úložišť provozovaných velkými knihovnami, tedy řešení více odpovídající doporučením *Národní koncepce dlouhodobé ochrany digitálních dat v knihovnách* z roku 2016.

8.10 Aktuální stav

S rozvojem digitálních služeb knihoven a nárůstem objemu spravovaných dat se také téma dlouhodobé ochrany digitálních dat v knihovnách dostává do popředí pozornosti nejen strategických dokumentů. Úroveň ochrany digitálních dat v jednotlivých knihovnách zůstává nicméně velmi nevyrovnaná, stejně jako její dostupnost pro jednotlivé knihovny. Kapacity LTP úložiště NK ČR v tuto chvíli dostačují k bezpečnému ukládání dat z provenience NK ČR, dat MZK, vytvořených v rámci společné digitalizační linky a dat institucí digitalizujících v rámci dotačních mechanismů VISK 7. Vývoj a rozvoj open source řešení pro dlouhodobou archivaci ARCLib, podpora implementace systémů pro dlouhodobou archivaci z programu VISK 7 nebo pokračující rozvoj nástrojů na podporu standardizace NDK jsou důležité dílčí kroky pro velké knihovny, které s jejich pomocí dokáží budovat fungující řešení pro dlouhodobou archivaci svých dat (např. KNAV), samy o sobe nicméně komplexní řešení pro celou knihovní síť nepřinášejí. České knihovny tedy v jistém smyslu stále přešlapují na křižovatce, zda zvolí cestu jednoho silného národního řešení sloužícího celé knihovní síti, nebo síť technologicky diverzifikovaných a vzájemně spolupracujících LTP úložišť ve více knihovnách, pro něž může například zastřešující národní řešení sloužit pouze v roli tzv. archivu poslední záchrany a NK ČR zajišťovat potřebný metodický dohled.

9 Identifikace dokumentu a perzistentní identifikátory

Úkolem identifikátorů je zabezpečovat jednoznačnou identifikaci nějakého objektu. Již tradičně pod pojmem *identifikátor* rozumíme znakový řetězec tvořený číslicemi (např. číselný kód u ISBN), případně dalšími typy znaků (např. alfanumerický kód). Ladislav Cubr rozlišuje v rámci knihovnictví dva druhy identifikátorů – digitální a analogové (bibliografické). Zatímco ty digitální identifikují internetové informační zdroje, analogové identifikátory označují tištěné dokumenty (knihy, časopisy, hudebniny apod.) (CUBR, 2010b). V praxi kulturních a paměťových institucí je však třeba, aby byla kromě jednoznačné identifikace zajištěna i další důležitá funkce – trvalá (perzistentní) identifikace určitého objektu.

Podle finského informačního vědce Juhy Hakaly by perzistentní identifikátor (a služba, kterou poskytuje) měl fungovat dlouhodobě, resp. minimálně tak dlouho, jako bude existovat objekt, kterému byl identifikátor přidělen (HAKALA, 2010). Použitý identifikátor by měl být navíc jedinečný i v mezinárodním měřítku. Mezinárodní trvalé identifikátory mají standardizovanou syntaxi a přidělují se na základě specifických pravidel, která musí dodržovat všichni uživatelé daného identifikačního systému. Klíčová je tak pro perzistentní identifikátory existence komplexního identifikačního systému pro přidělování a správu těchto identifikátorů (CUBR a VAŠEK, 2013).

9.1 Implementace perzistentních identifikátorů v českých knihovnách

Příkladem jednoho z prvních globálních a perzistentních identifikačních systémů je identifikátor ISBN *(International Standard Book Number)*. Identifikační formát ISBN byl vymyšlen již v roce 1967 a v roce 1970 byl aktualizován a vydán Mezinárodní organizací pro normalizaci (ISO) jako mezinárodní standard ISO 2108 (ISBN, 2023).

Systém přidělování ISBN má důkladně propracovanou infrastrukturu, kterou tvoří síť národních agentur přidělujících ISBN, a rovněž jasně stanovená pravidla pro přidělování identifikátoru, která musí všichni dodržovat. Na základě ISBN je tak možné identifikovat zemi, ve které kniha vyšla, stejně jako vydavatele, který za jejím vydáním stojí. Každý identifikátor ISBN má totiž stejnou délku (skládá se z třinácti číslic) a je rozdělen do pěti částí, které jsou odděleny spojovníky. První část tvoří stanovený prefix ISBN a v současnosti jej tvoří číslo 978 nebo 979. Druhá část se nazývá identifikátor skupiny a označuje zemi nebo jazykovou oblast, kde vydavatel působí. Délka této části je proměnlivá a může ji tvořit jedna až pět číslic. Pro Českou republiku (a rovněž pro Slovensko) bylo v minulosti stanoveno číslo 80, které se používá dodnes. Třetí část identifikuje konkrétního vydavatele a její identifikátor se může skládat ze dvou až sedmi číslic. V rámci přidělování této části platí pravidlo, že vydavatelům s velkým objemem produkce se přiděluje krátký identifikátor a malým vydavatelům naopak delší. Předposlední čtvrtou část tvoří identifikátor titulu a jeho délka závisí od délky předešlých částí. Pro ČR platí, že tato část se může skládat celkově až z pěti číslic. Závěrečnou část pak tvoří kontrolní číslice, která se získá matematickým výpočtem (Národní knihovna, 2019). Celkový identifikátor ISBN tak může mít v ČR např. následující podobu: 978-80-257-0368-7.

Jedinečnost ISBN a dohled nad dodržováním pravidel na národní úrovni zajišťuje od roku 1989 Národní agentura ISBN v České republice. Obdobně se tištěným dokumentům v ČR přidělují i další perzistentní identifikátory: ISMN a ISSN. ISMN *(International Standard Music Number)* se přiděluje českým hudebninám od roku 1996 a identifikátor ISSN *(International Standard Serial Number)* označuje česká periodika již od sedmdesátých let minulého století (CUBR et al., 2020, s. 55).

Od roku 2010 se v ČR používá ještě další identifikátor – čČNB (číslo České národní bibliografie), který stojí na pomezí mezi analogovými a digitálními identifikátory. Důvodem vytvoření čČNB byla snaha o identifikaci všech tištěných dokumentů, které vycházely na území ČR od 19. století do současnosti, z nichž některé s ohledem na datum vydání nemají ISBN ani ISSN. V rámci digitalizace NDK tak identifi-

kátor čČNB slouží jako prvotní řešení, které umožňuje propojit digitalizáty tištěných dokumentů s jejich fyzickými předlohami (CUBR, 2010b).[89]

Dalším důležitým, v tomto případě již čistě digitálním, perzistentním identifikátorem používaným v českém prostředí je URN:NBN *(Uniform Resource Name: National Bibliography Number)*. Identifikátor URN:NBN je určený pro digitální objekty a kromě jejich trvalé a jednoznačné identifikace je jedním z jeho úkolů zprostředkovat jejich zpřístupnění internetovým uživatelům, případně poskytovat metadata k identifikovaným dokumentům (CUBR et al., 2016, s. 18). Přidělování identifikátoru URN:NBN v českém prostředí, stejně jako i jinde ve světě, musí zajišťovat zvlášť navržená technická infrastruktura, jejíž součástí je aplikace resolver. Jedná se o přesměrovávací internetovou službu, která poté, co uživatel zadá perzistentní identifikátor dokumentu, provede okamžité přesměrování na funkční URL adresu, na které je dokument zveřejněn. Přestože bývají URL adresy nespolehlivé a mohou se často měnit, díky dodržování pravidel pro používání perzistentních identifikátorů by aplikace resolver měla být vždy schopná najít aktuální umístění digitálního objektu v dané digitální knihovně.

Kromě identifikátorů perzistentních existují i další identifikátory, které ale nezabezpečují trvalou identifikaci digitálních objektů. Za zmínku stojí např. výše uvedený URL *(Uniform Resource Locator)*, který je v současnosti stále jedním z nejpoužívanějších identifikátorů internetových zdrojů. Identifikátor má však svoje specifika (a v laickém pojetí je často zaměňován s webovou adresou). Zvláštností tohoto identifikátoru je, že nespecifikuje zdroj samotný, ale označuje pouze umístění daného zdroje na internetu (podobně jako např. signatura označuje umístění knihy v knihovním fondu). Díky tomu URL umožňuje uživateli získat okamžitý přístup k danému zdroji. Výrazným nedostatkem z pohledu uživatelů je ale naopak velká nestabilita URL adres, jelikož umístění digitálních objektů na internetu se může měnit a často se tak i děje (CUBR a VAŠEK, 2013).

89 V současné době se identifikátor čČNB přidává i elektronickým dokumentům.

9.2 Vytvoření systému pro trvalou identifikaci digitálních dokumentů

Díky používání perzistentních identifikátorů bylo možné v českém prostředí vytvořit stabilní systém trvalé identifikace digitálních dokumentů. Technický provoz tohoto systému zajišťuje software ČIDLO[90] vytvořený v Odboru digitálních fondů NK ČR v letech 2011–2013. Jeho hlavní funkcí je přidělování trvalých identifikátorů URN:NBN digitálním dokumentům a také zabezpečování komplexního systému pro správu a identifikaci českého digitálního dědictví. Systém ČIDLO slouží zejména českým knihovnám a paměťovým institucím a sestává ze souboru pravidel, která musí všichni účastníci systému dodržovat (pod pojmem účastníci systému se rozumí registrované instituce, které jsou vlastníky digitálních dokumentů a jež mají zájem o přidělení URN:NBN svým dokumentům). Podrobná pravidla celého systému jsou zachycena v aktuální certifikované metodice vydané v roce 2018 (VAŠEK et al., 2018).

Do roku 2018 přidělil systém ČIDLO českým digitalizátům více než 1,5 milionu identifikátorů URN:NBN (CUBR, 2017). Pro český jmenný prostor má identifikátor URN:NBN definovaná vlastní pravidla a syntaxi. Platí, že může být tvořen pouze řetězcem alfanumerických znaků, přičemž povolenými znaky jsou dále ještě dvojtečka a spojovník. Pro český jmenný prostor je identifikátor URN:NBN složen ze tří částí. První část tvoří všeobecný mezinárodní kód ve tvaru „urn:nbn:cz“, přičemž „cz“ označuje dokumenty české provenience. Druhou částí identifikátoru je jedinečný kód knihovny nebo paměťové instituce, která je registrována v systému ČIDLO. Kód paměťové instituce může mít 1–6 alfanumerických znaků a v případě knihovny bývá většinou i její siglou a je zapsán v Adresáři knihoven.[91] Třetí a poslední část generuje aplikace resolver a tvoří ji vždy šestimístný alfanumerický kód. Výsledný identifikátor tak může mít

90 Dále o systému ČIDLO též v předchozí kapitole o dlouhodobém uchovávání digitalizátů v českých knihovnách.

91 Výjimku představuje několik českých knihoven, u kterých se v rámci kódu nezapisuje sigla. Např. pro NK ČR se od roku 2012 používá dvoumístný kód „nk“ a její sigla ABA001 se v systému od té doby nepoužívá.

celkově 19 až 24 znaků, přičemž kvůli lepší čitelnosti se píše vždy malými písmeny (VAŠEK et al., 2018). Dle výše uvedených pravidel může tedy identifikátor URN:NBN vypadat např. takto: urn:nbn:cz:osa001-0003kl. Systém ČIDLO neslouží jen zapojeným institucím, ale i běžným uživatelům. Jedním z jeho technických subsystémů je již zmiňovaná aplikace resolver, která zajišťuje přesměrování daného identifikátoru URN:NBN na aktuální adresu digitální knihovny, ve které je daný dokument zpřístupněn. Pravidla systému ČIDLO totiž od zapojených institucí vyžadují, aby do systému dodávaly aktuální informace o webových adresách, na kterých se identifikované dokumenty nacházejí. Každá změna URL adresy tak má být oznámena v systému ČIDLO, ať již automatizovaně, nebo manuálně. Poskytování aktuálních URL adres i celá administrativa v rámci systému ČIDLO ale od zapojených subjektů vyžaduje určité personální, finanční i technologické zázemí, kterým musí organizace disponovat, aby se systému trvalé ochrany mohly účastnit.

10 Dlouhodobé uchovávání v archivářské praxi

V České republice probíhá systematická debata o uchovávání digitálních dokumentů v paměťových institucích od počátku 21. století. Z paměťových institucí se to týká zejména knihoven a archivů, u ostatních institucí jde zatím spíše o ojedinělé snahy. Samotné zásady dlouhodobého uchovávání jsou dostatečně známé a prověřené, ale jejich realizace v jednotlivých typech institucí se liší podle jejich potřeb. Stejně jako v jiných zemích se také v ČR do popředí aktivit v oblasti uchovávání a zpřístupnění digitálních dokumentů dostaly knihovny. Ačkoliv základní iniciativa u archivů vznikla ve stejné době jako v knihovnách, její realizace se protáhla. V případě knihoven se již na počátku podařilo postoupit od teoretických debat k realizaci kroků pro dlouhodobé uchovávání. Na počátku byly knihovny vedeny snahou zabezpečit výstupy velkých investic, které vložily do digitalizace. Až následně se záběr jejich aktivit rozšířil o born digital dokumenty.

Využívání LTP nástrojů v archivech následovalo se značným odstupem po knihovnách. Neznamená to však, že by v archivní oblasti nebyla trvalé archivaci digitálních dokumentů věnována pozornost. I v prostředí archivů stála na počátku digitalizace analogových dokumentů. Postavení archivů se od knihoven liší, ač vnějšímu pozorovateli nemusí být tyto odlišnosti na první pohled zřejmé. V prvé řadě je třeba zmínit zcela odlišné legislativní zakotvení, které archivy v mnoha ohledech omezuje a svazuje, ale na druhou stranu přesně definuje jejich vztah k producentům dokumentů, které se stanou archiváliemi. Také typy uchovávaných dat jsou mnohem rozmanitější, než s jakými se obvykle potkávají knihovny. Pro archivy byla vždy zásadní dvě témata. Jednak získání digitálních dokumentů, a to ve standardizovaných výstupech, a jednak trvalé a důvěryhodné uložení (a zpřístupnění) těchto dat. Obě tyto oblasti jsou regulované zákonem a dalšími právními předpisy. Dalším tématem se stává způsob a nástroje pro zpracování. Archivní síť v ČR je decentrali-

zovaná, s právní subjektivitou nejen specializovaných archivů, ale i sítě státních archivů. Metodickou roli vykonává Ministerstvo vnitra a částečně Národní archiv ČR. Archivy tak sdílejí metodické dokumenty, ale využívají odlišná softwarová a koncepční řešení. Možností pro srovnání jednotlivých přístupů budou akreditace Digitálních archivů, které jsou možné od roku 2012, ale dosud pro ně nebyly vydány všechny předpisy. Přesto již ke dvěma pokusům o získání akreditace došlo, ty však byly neúspěšné. V následujícím příspěvku se zaměříme na některé z popsaných aspektů spojených s aktuálním stavem digitálního archivnictví a jeho předchozím vývojem.

10.1 Digitální dokumenty v ČR jako archiválie a jejich uchování

Archivy v sobě spojují vědecké, kulturní a úřední funkce. Jejich posláním je uchovávat, zpracovávat a zpřístupňovat archivní dědictví. Předmětem jejich péče jsou archiválie, které jsou zákonem definovány jako „*takové dokumenty, který byly vzhledem k době vzniku, obsahu, původu, vnějším znakům a trvalé hodnotě dané politickým, hospodářským, právním, historickým, kulturním, vědeckým nebo informačním významem vybrány ve veřejném zájmu k trvalému uchování a byly vzaty do evidence archiválií*“. Úloha archivů ochránit kulturní dědictví a dokumenty o vývoji společnosti se rozšířila i na segment digitálních dokumentů.

Česká archivní legislativa (především zákon č. 499/2004 Sb. o archivnictví a spisové službě v platném znění) pracuje s pojmem „archiválie v digitální podobě“. Pod tímto pojmem jsou chápány dokumenty vzniklé jako originál v digitální podobě (born digital dokumenty), které byly vybrány za archiválie a jsou uloženy a spravovány v některém z archivů zřízených dle výše uvedeného zákona (viz § 15, odst. 3.). Dlouhodobé uchování digitálních dokumentů v archivech ovšem nezahrnuje pouze born digital archiválie. Vedle příjmu těchto dokumentů samozřejmě v mnohých archivech probíhá digitalizace archiválií analogových za účelem jejich ochrany a zpřístupnění. V tomto případě zůstávají digitalizáty většinou pouze v pozici kopií – na rozdíl

od born digital dokumentů, které je nutné chápat jako originály a zajistit jim v souladu s legislativou patřičnou ochranu. Za originál však bývá digitální (či analogová) kopie považována v situaci, kdy dojde ke ztrátě či zničení originální analogové archiválie. V této situaci je digitální kopie z legislativního hlediska prohlášena za archiválii nahrazující originál a dostává se jí stejné právní ochrany.

České archivnictví získává digitální dokumenty pro trvalé uchování dvěma způsoby. Tím prvním je výběr z dokumentů spravovaných ve standardizovaných systémech elektronických spisových služeb, druhým pak akvizice veškerých dalších dokumentů (osobní pozůstalosti, informační systémy neintegrované do spisové služby aj.). Dokumenty vznikající ve spisových službách by měly odpovídat tzv. výstupním formátům (nebo do nich být převedeny), tak jak je definuje § 23 vyhlášky č. 259/2012 Sb. o podrobnostech výkonu spisové služby, která specifikuje formáty dat pro textové, obrazové a audiovizuální dokumenty a pro databáze, a vždy jsou součástí informačních balíčků, které mimo samotný dokument obsahují i jeho standardizovaná administrativní a technická metadata. Jako výstupní datové formáty jsou definovány takové, u kterých je předpokládána stabilita a podpora formátu, a tedy jeho dlouhodobá udržitelnost. Vzhledem k legislativnímu zakotvení je však tento výčet málo flexibilní. České archivy se podílely na definici příslušného standardu (Národní standard pro elektronické systémy spisové služby) a ovlivnily tak datovou (formátovou) i metadatovou podobu spravovaných digitálních dokumentů, tak aby je bylo možné v případě výběru za archiválie trvale uložit. Archivy však spravují i další digitální dokumenty, jejichž formát a případná metadata jsou v různých standardech či dokonce nejsou standardizovány vůbec. V tom případě musí přijmout data tak, jak jsou. Nástrojem pro standardizaci těchto dat jsou formátová pravidla jednotlivých archivů, která doporučují původcům převod v budoucnu archivovaných dat do doporučených formátů, případně formátová konverze prováděná při příjmu dokumentů do archivu, či v průběhu jejich uložení. Metadata těchto nestandardizovaných archiválií jsou při příjmu manuálně editována archiváři a, pokud je to možné, strojově vytěžována. Před touto výzvou stojí české archivy právě teď. Dokumenty z elektronických spisových služeb do skartač-

ních řízení v masovém měřítku teprve začínají přicházet (intenzita se však ještě zvýší), starší dokumenty, mnohdy z devadesátých let 20. století, již v archivech jsou nebo se objevují mezi přebíranými datovými soubory.

Důležitou součástí digitálních archivů je samozřejmě správa dat, která zahrnuje mj. tvorbu, správu a aktualizaci popisných metadat uložených digitálních dokumentů. Zejména born digital dokumenty se do archivů dostávají jen s omezeným rozsahem popisných metadat. K vytvoření archivního popisu (tvorbě archivních vyhledávacích pomůcek, což je obdoba katalogizace v knihovnách) tak dochází v rámci archivního zpracování až po přijetí dokumentů do archivu. Archivní zpracování je v České republice standardizováno pomocí tzv. Základních pravidel pro zpracování archiválií (Ministerstvo vnitra ČR, 2022b) a je vykonáváno v tomu odpovídajícím software. Archivní zpracování born digital archiválií je v českém archivnictví zatím spíše výjimkou. Některé specializované a další archivy (například Archiv Univerzity Karlovy) provozují či plánují tuto vrstvu data managementu jako součást informačního systému svého digitálního archivu.

Vzhledem k faktu, že většina českých archivů bude své digitální archiválie ukládat v Národním digitálním archivu (NDA), který nástroje pro zpracování archiválií pečujícím archivům neposkytuje, bude v těchto situacích archivní zpracování realizováno softwarem ve správě pečujících archivů. Vznik potřebných rozhraní (pro poskytování přístupu k zpracovávaným archiváliím a ukládání výsledku archivního zpracování do aktualizovaných archivních informačních balíčků) je předmětem probíhajícího projektu Technologické agentury (TA ČR) *Vytvoření standardů pro komunikaci informačního systému digitálního archivu s jeho okolím*. Po jeho dokončení je možné předpokládat změnu popsaného stavu.

Na nutnost archivace dokumentů v digitální podobě začali čeští archiváři narážet již v polovině devadesátých let, kdy se stále častěji setkávali s dokumenty trvalé historické hodnoty, které existovaly jen v digitální podobě. Od počátku 21. století jsou digitální dokumenty již běžnou součástí života současné společnosti a není proto nijak překvapující, že bylo nutné začít uvažovat o tom, jak zajistit jejich dlouhodobé uchování poté, co se z nich dle platné legislativy

staly archiválie. Na počátku tohoto století (2001–2005) byly za účasti Odboru archivní správy MV ČR, Českého vysokého učení technického a Národního archivu ČR realizovány první výzkumné projekty v oblasti možného technického a procesního řešení digitální archivace. Dílčí zprávy byly publikovány na stránkách Archivního časopisu.

Důležitým milníkem bylo vládní usnesení č. 11/2004 k dlouhodobému uchovávání a zpřístupňování dokumentů v digitální podobě, které uložilo Ministerstvu vnitra vytvořit při NA ČR odborný tým, který zajistí vybudování celostátního digitálního archivu (Usnesení vlády ČR č. 11/2004). V předkládací zprávě usnesení byla nejen konstatována a zdůvodněna nutnost řešit dlouhodobé uchovávání digitálních dokumentů, ale rovněž byl určen postup směřující k vybudování digitálního archivu. V roce 2005 proto v rámci NA ČR vznikl *Projekt dlouhodobého uchovávání a zpřístupňování dokumentů v elektronické podobě* (CHIMERA – Czech History Information Management and Electronic Records Archiving).[92] Původní harmonogram byl optimistický, počítal se zahájením rutinního provozu digitálního archivu od začátku roku 2008. Do tohoto data se však nakonec jen podařilo připravit projekt digitálního archivu.

10.2 Potřeby pro zřizování digitálních archivů v České republice

Archivnictví v České republice jako celek spadá do gesce Ministerstva vnitra, které jej řídí prostřednictvím svého Odboru archivní správy a spisové služby (OASSS MV ČR). Základem archivní sítě je dvoustupňová soustava státních archivů tvořená Národním archivem ČR a státními oblastními archivy, jejichž organizačními součástmi jsou státní okresní archivy. Tyto archivy se starají o archiválie všech původců, kteří si nezřídili vlastní archivy. V oblasti ukládání digitálních archiválií státní archivy musí dle platného zákona ukládat v Národním digitálním archivu. Přestože jsou digitální archiválie státních archivů uloženy v jediném existujícím digitálním archivu,

92 https://www.nacr.cz/vyzkum-publikace-akce/vyzkum/projekty/chimera

jejich správa (v současné době především jejich výběr a evidence) je nadále zajištěna archivy podle jejich územní a věcné příslušnosti. Jednotlivé archivy přitom využívají nástrojů a workflow, které jsou součástí Národního archivního portálu, který je spravován NA ČR.

Státní archivní síť doplňují další typy archivů (specializované, bezpečnostní, soukromé a územních správních celků), které slouží primárně svým zřizovatelům. Ministerstvo vnitra je přímo neřídí, vykonává nad nimi jen metodický dohled. Zpravidla nejde o samostatné instituce, ale o organizační součásti svých zřizovatelů. Ti musí pro zřízení svých archivů splnit řadu personálních, technických a organizačních požadavků, jejichž splnění je prokazováno před ministerstvem vnitra v rámci akreditačního řízení. V oblasti ukládání digitálních archiválií mají nestátní archivy více možností než státní, které musí povinně ukládat v NDA. Legislativní požadavky na nestátní digitální archiv jsou spíše obecného rázu, podrobněji jsou popsány jen technické požadavky na umístění technologií digitálního archivu (mj. platí požadavek na minimální, padesátikilometrovou, vzdálenost hlavního a záložního úložiště), součástí řízení o oprávnění však není žádný audit podle mezinárodně uznávaných norem. Požadavky na získání oprávnění pro ukládání archiválií v digitální podobě jsou shrnuty v § 60a zákona č. 499/2004 Sb. Prováděcí normou je v současné době *Vzorový provozní řád digitálního archivu* (Věstník ministra vnitra, částka 65/2012).

Cíl uchování digitálních dokumentů (digitálních archiválií či born digital dokumentů prohlášených za archiválie) je ve všech českých archivech shodný: zajistit uchování (ve smyslu dlouhodobého uložení, bitové a logické ochrany) pro další generace, tyto dokumenty v souladu s legislativou a určením archivu spravovat a zpřístupňovat je oprávněným uživatelům.

Východiska a konkrétní potřeby se však často liší. Primární potřebou, která má odraz i v archivní legislativě, je potřeba uchovávat ve veřejných (státních) archivech digitální dokumenty veřejné správy a samosprávy vybrané za archiválie s ohledem na jejich trvalou hodnotu. Tyto dokumenty se do archivu dostávají vždy po konci jejich životního cyklu u původce a potřeba jejich uchování je především historická. Primárním archivem tohoto typu je Národní digitální

archiv (součást Národního archivu ČR), který slouží jako úložiště pro všechny státní archivy a ty nestátní archivy, které nezřídily vlastní digitální archiv.

Řada státních i nestátních archivů řešila a řeší potřebu dlouhodobého uchování digitalizátů a dalších dat, vzniklých při rozsáhlých digitalizačních projektech. Na rozdíl od uchovávání born digital archiválií v této oblasti neexistuje dosud centralizované řešení ani legislativní omezení,[93] a jednotlivé archivy budují vlastní repozitáře, které mimo zpřístupnění v různé míře zajišťují i funkce dlouhodobé ochrany a dlouhodobého uložení pořízených dat. Nejpokročilejší je patrně řešení *Digidepot* Státního oblastního archivu v Třeboni. Druhým příkladem je nasazení open source LTP systému Archivematica pro uložení a správu digitalizátů audiovizuálních děl v Národním filmovém archivu (Archivematica je také základem aktuálního řešení Národní digitální archiv II).

Třetí potřeba vzniká zejména ve specializovaných archivech zřizovatelů spravujících množství digitálních dokumentů s dlouhým životním cyklem či specifickým charakterem. Zřizovatelé těchto archivů mají potřebu dlouhodobého uložení svých digitálních dokumentů s historickou hodnotou (prohlášených za archiválie), které z technických důvodů nemohou (či z provozních důvodů nechtějí) předat k uložení v Národním digitálním archivu. Hlavním důvodem je většinou potřeba původce (zřizovatele archivu) dokumenty, které se staly archiváliemi, dále aktivně využívat. Dalším důvodem může být specifický charakter či velký objem uložených dat a neochota zřizovatele předávat své archiválie k uložení mimo svou instituci.

Své digitální archivy tak budují či chtějí vybudovat Český rozhlas a Česká televize, které archivovaná audiovizuální díla mohou stále využívat pro svá vysílání, objem uložených digitálních dat v případě těchto institucí navíc přesahuje technické možnosti infrastruktury Národního digitálního archivu. Vlastní digitální archivy budují či plánují budovat také některé velké české univerzity (včetně Univer-

93 V roce 2015 vznikla Metodika pro vytváření bezpečnostních kopií archiválií v digitální podobě, jejíž součástí je i oddíl věnující se otázkám bezpečného uložení. V zásadě jde o nejkomplexnější soubor doporučení pro dlouhodobé uchovávání v českých archivech (viz DVOŘÁK et al., 2015).

zity Karlovy, v jejímž archivu působí autoři tohoto textu), které mají potřebu rychlého přístupu k archivovaným digitálním dokumentům, což je funkcionalita, kterou NDA zatím nebyl schopen dlouhodobě garantovat. Své digitální archivy budují, či plánují i zřizovatelé archivů z okruhu silových resortů (Ministerstvo obrany, bezpečnostní a zpravodajské organizace), jejichž zřizovatelé nechtějí svá citlivá data předávat do péče Národního archivu či mají potřebu uchovávat archiválie obsahující utajované informace.

Archivace digitálních dokumentů je tedy v českém prostředí již realitou, born digital dokumenty jsou součástí archivních fondů. Přestože v uplynulých patnácti letech byly diskutovány varianty vybudování Národního digitálního archivu jako de facto samostatné instituce nezávislé na stávajících archivech pro ukládání „papírových" archiválií a vydání samostatného zákona o digitální archivaci, nedošlo nakonec k jeho vydělení. Problematika ukládání digitálních archiválií byla legislativně řešena v několika novelách archivního zákona a NDA, který po několika neúspěšných projektech zahájil provoz, je organizační součástí oddělení NA ČR pověřeného péčí o nejnovější archiválie a dozorem nad výkonem spisové služby u centrálních úřadů České republiky.

Ačkoli česká archivní legislativa umožňuje vznik dalších digitálních archivů, zůstává NDA de iure (nikoliv de facto, viz dále) v roce 2023 jediným digitálním archivem v české archivní síti a zajišťuje uložení digitálních archiválií pro všechny státní archivy a část specializovaných archivů. NDA tak funguje jako digitální repozitář pro archiválie ostatních archivů. NDA odpovídá za jejich uložení a dlouhodobou ochranu, samotná správa (obnášející výběr, evidenci, zpracování a v budoucnu i zpřístupnění) digitálních archiválií nadále zůstává v kompetenci jednotlivých archivů podle jejich věcné a územní příslušnosti.

Pro archiváře ostatních archivů je klíčovým nástrojem NDA Národní archivní portál, který poskytuje soubor nástrojů pro výběr dokumentů, které se mají stát archiváliemi, validaci metadat úředních dokumentů, případně tvorbu metadat neúředních dokumentů a uložení vybraných dokumentů jako digitálních archiválií v NDA. Vcelku rutinně jsou realizovány výběry a ukládání úředních digitál-

ních dokumentů a digitálních metadat papírových dokumentů, které jsou evidovány v elektronických systémech spisových služeb. Zatím není možné uložené digitální archiválie archivně zpracovávat a velmi omezená je možnost jejich zpřístupnění (ať už pro archivy, do jejichž péče přísluší, či pro původce).

Jak bylo však výše uvedeno, dnes existuje v České republice relativně široký okruh zřizovatelů specializovaných a bezpečnostních archivů, kteří chtějí vybudovat vlastní digitální archivy a jsou v různých fázích jejich realizace (od prosté deklarace po provozování zatím neakreditovaných digitálních archivů). Zřizovatelé těchto archivů většinou neukládají digitální dokumenty vybrané za archiválie v Národním digitálním archivu, své digitální archiválie ukládají v digitálních repozitářích, které jsou pod správou jejich akreditovaných specializovaných archivů. De facto (vzhledem k neproběhlé akreditaci nikoliv de iure) tak v České republice již funguje (mimo Národní digitální archiv) několik digitálních archivů provozovaných veřejnými archivy. Dva ze zřizovatelů (Ministerstvo obrany jakožto zřizovatel Vojenského ústředního archivu a Masarykova univerzita) v minulých letech podali žádost o udělení oprávnění pro ukládání archiválií v digitální podobě (akreditaci Národního archivu), Národní archiv však Odboru archivní a spisové služby Ministerstva vnitra nedoporučil její kladné vyřízení a žádosti tak byly zamítnuty. Oba zřizovatelé však deklarovali znovupodání této žádosti.

Své digitální archivy v roce 2023 fakticky provozuje v rámci svých akreditovaných archivů např. ještě hlavní město Praha, Český úřad zeměměřičský a katastrální, Česká televize a Univerzita Karlova. Podle dostupných informací každý z těchto archivů provozuje digitální repozitář se zajištěnou bitovou, v některých případech i logickou ochranou. Úložiště LTP a archivně správní funkcionality jsou teprve dokončovány. Zřizovatelé dalších archivů (Akademie věd ČR, zřizovatelé bezpečnostních archivů) zájem o zřízení digitálních archivů v nedávné době deklarovali a svá řešení v současné době začínají realizovat.

Dosud však k vybudování žádného „nestátního“ digitální archivu nedošlo (viz dále). Zřizovatel nestátního archivu v České republice má tedy v oblasti ukládání digitálních archiválií možnost volby mezi uklá-

dáním v NDA nebo zřízením vlastního digitálního archivu. Legislativa připouští i třetí, zatím jen teoretickou, možnost, kdy by zřizovatel nestátního archivu ukládal své digitální archiválie na základě smlouvy v digitálním archivu, který je součástí jiného nestátního archivu. Tato možnost je zatím prakticky nerealizovatelná.

Veřejné vysoké školy v České republice patří k jednomu z největších producentů digitálních dokumentů. V minulých desetiletích úspěšně digitalizovaly většinu svých úředních a správních agend a díky tomu zpracovávají velké množství digitálních dokumentů, z nichž řada má trvalou historickou hodnotu a je tedy nutné zajistit jejich budoucí archivaci.

Patří k nim například i agenda závěrečných kvalifikačních prací, která vytváří desítky tisíc digitálních dokumentů (avšak ne všechny VŠ je trvale archivují). Zhruba třetina českých vysokých škol zřídila své specializované archivy, a proto se bude v budoucnu aktivně podílet na správě svých digitálních archiválií; pro ostatní vysoké školy archivaci zajišťují státní oblastní archivy. Důležitým zdrojem zkušeností v oblasti digitální archivace v prostředí vysokých škol byl centralizovaný rozvojový projekt Ministerstva mládeže, školství a tělovýchovy probíhající v letech 2015 až 2017, jehož tématem bylo prohloubení správy a zajištění budoucí digitální archivace dokumentů veřejných vysokých škol. Na projektu se podílela většina českých veřejných vysokých škol (hlavním řešitelem byla Masarykova univerzita). Důležitou částí projektu však bylo vyjasnění podmínek pro zřízení vlastního digitálního archivu, vysoké školy v této věci komunikovaly s Ministerstvem vnitra a NA ČR. V rámci projektu vznikl metodický materiál řešící možnosti digitální archivace v oblasti předarchivní péče i vzniku samotných digitálních archivů (PICHL et al., 2015). Jedním z výstupů byla metodika pro vyřazování a archivaci úředních dokumentů ve skartačním řízení (CAJTHAML, 2017).

10.3 Akreditace digitálního archivu dle legislativy ČR

V úvodní kapitole nastiňující vývoj oboru Digital Preservation byl zmíněn i vývoj způsobů certifikace digitálních repozitářů, aby mohly být v souladu s teoretickými doporučeními prohlášeny za důvěryhodné. Dlouho připravovaná a podrobná metodika TRAC, která vedle ke vzniku ISO normy 16363, se ukázala jako v praxi příliš složitá, komplikovaná a tedy nákladná. Aktuálně disponuje certifikací dle této metodiky jen několik jednotlivých repozitářů. Samotný důraz na důvěryhodné repozitáře tím však nezmizel. Stále jde o jeden z klíčových požadavků a představuje jedno z důležitých témat, kterému se odborná komunita intenzivně věnuje. Zejména pro archivy, které by měly garantovat uchování archivního dědictví, je nutností doložit, že jsou schopny se o uložená data postarat. V zásadě lze konstatovat, že v oblasti uložení digitálních dat se s ohledem na praxi ustupuje od velmi podrobných, administrativně náročných auditů k pružnějším formám kontroly s důrazem na self audity. Ty však současně prohlubují rozsah požadovaných informací a jejich výpovědní hodnota narůstá (např. self audit DPC RAM, který již poskytuje širokou škálu informací o repozitáři).

Ve snaze zpřehlednit situaci vznikl již v roce 2010 *European Framework for Audit and Certification of Digital Repositories* (Evropský rámec pro audit a certifikaci), který stanovil tři stupně auditu, přičemž certifikovaný veřejný audit představuje stupeň číslo tři. Debaty o způsobu certifikace představovaly klíčové téma odborné komunity věnující se Digital Preservation zejména v první polovině druhého decennia 21. století. V současnosti toto téma částečně ustoupilo do pozadí, neztratilo však na důležitosti. Přes příklon k jednodušším formám certifikace (to se týká zejména knihoven) zůstává externí audit podle zdokumentované metodiky stále cílovým stavem. Na kombinaci plnění kontrolovatelných kritérií a sebeevaluace sází české archivy. Podobně jako i v jiných oblastech, tak i v případě certifikace jsou možnosti regulované legislativou. Ta předpokládá, že archiv, který chce být Digitálním archivem, musí získat „oprávnění k ukládání archiválií v digitální podobě“. Požadavky na získání oprávnění pro

ukládání archiválií v digitální podobě jsou shrnuty v § 60a zákona č. 499/2004 Sb. Aby toto oprávnění archiv získal, musí projít posouzením, které zkoumá jeho schopnosti pečovat o uchovávané digitální dokumenty. Jak bylo řečeno výše, aktuálně je Digitálním archivem v ČR pouze Národní archiv ČR, který tento status získal ze zákona, bez nutnosti sám absolvovat audit. Je však poradním orgánem při certifikaci dalších archivů a jeho stanovisko je rozhodující. Proces získání oprávnění pro ukládání archiválií v elektronické podobě a podmínky pro získání se v čase proměňovaly. Samotná možnost akreditace digitálního archivu se objevila v roce 2012, kdy byl novelizován archivní zákon zákonem č. 167/2012 Sb. V archivním zákonu se tak objevila nebo byla upravena řada ustanovení, na které zde bylo odkazováno. Nově přibyly odst. 3 v § 15, § 18b, § 18c a § 60a až § 60c. Změněny byly podmínky upravené v § 61 odst. 2 a 4, které upravují stavební podmínky pro umístění pracoviště Digitálního archivu.

Samotný zákon obsahuje jen několik základních podmínek pro získání povolení k ukládání elektronických archiválií. Vyžaduje především zpracování koncepce dlouhodobého uchovávání a ochrany dokumentů a seznam metadat týkajících se popisu archiválií v digitální podobě, popisu a evidence archivních souborů a popisu původců. Dále je podmínkou úspěšné odeslání archiválií z příslušného digitálního archivu do Národního digitálního archivu. V souladu s obecnými nároky na důvěryhodné repozitáře pak zákon žádá, aby záložní úložiště bylo umístěno alespoň 50 km od primárního úložiště. Požádat o udělení akreditace může jen ten, kdo již má akreditovaný archiv (případně lze akreditovat nový archiv současně s žádostí o vytvoření digitálního). V zásadě tedy zákon předpokládá, že akreditovaný digitální archiv musí disponovat informačním systémem, který pokryje spektrum povinných činností, bude provozovat LTP řešení, které zajistí dlouhodobé uchovávání digitálních dat a toto řešení budou spravovat pracovníci s dostatečnou praxí a vzděláním.

Jak je zřejmé, zákon pouze vymezuje okruh povinností, ale již nestanovuje, jak je přesně naplnit (kromě fyzické vzdálenosti obou úložišť). V kontextu zásad Digital Preservation lze sice dovodit, jak správným postupem požadavky zákona naplnit, ale samotná textace zákona by umožnila i takový výklad, který by ji formálně splnil, ale

výsledný archiv by neodpovídal zásadám OAIS. Na zpřesnění podmínek pomocí rozšiřujících materiálů se zaměřil Národní archiv, který za kontrolu dodržování zákona společně s OASSS MV ČR odpovídá. Ze zkušeností s realizací certifikace dle ISO 16363 vyvodil, že využít tuto normu není pro české archivy akceptovatelné (kromě velmi striktních požadavků sehrávají roli samozřejmě i finanční možnosti archivů). Na základě zkušeností evropských institucí i historicky daného napojení českého archivnictví na Německo vyhodnotil Národní archiv jako nejvhodnější využít normu *DIN 31644:2012 – Kriterien für vertrauenswürdige digitale Langzeitarchive*.

Tato norma vznikla v prostředí německých archivů jako reakce na komplexní způsob hodnocení pomocí ISO 16363. V souladu s *European Framework for Audit and Certification of Digital Repositories* umožňuje certifikovat repozitář na stupeň 2 (ve formě veřejného self auditu) nebo stupeň 3 (jako externí hodnocení). V evropském prostředí je vnímána jako rovnocenný ekvivalent k normě ISO 16363, její struktura je však jednodušší a její kritéria lze snáze doložit. Formální certifikace (st. 3) dle DIN zatím není řešena. Konsorcium nestor se zaměřilo na tzv. rozšířenou certifikaci (st. 2) a nabízí tzv. Pečeť nestoru *(nestor-Siegel)*. Jde o pokročilou sebeevaluaci.

K první certifikaci podle této normy se přistoupilo v roce 2015 a další instituce postupně následovaly, byť stále jde jen o jednotlivé instituce. Tým Národního digitálního archivu v roce 2018 tuto normu přeložil do češtiny a oficiálně ji vydal. Zároveň v tomto roce uspořádal mezinárodní workshop AiDA 2018: Self-Audit and Certification of Digital Archives in Central European Perspectives, kde byly otázky certifikace digitálních archivů v ČR obsáhle diskutovány. Díky překladu a adaptaci normy nestor byly konečně vyjasněny podmínky pro akreditaci Digitálního archivu dle archivního zákona. V roce 2019 tak byl vydán Metodický návod č. 2/2022 odboru archivní správy a spisové služby Ministerstva vnitra pro akreditaci digitálního archivu. Po zohlednění některých poznatků z praxe byl v roce 2022 vydán nový metodický návod, který je aktuálně platnou normou pro získání povolení pro ukládání digitálních archiválií (Metodický návod č. 2/2022 odboru archivní správy a spisové služby Ministerstva vnitra pro akreditaci digitálního archivu) (Ministerstvo vnitra, 2022a). Ani

tento dokument však zatím nelze považovat za definitivní nastavení podmínek. Z požadavků zákona stále nelze realizovat testovací předání dat do Národního digitálního archivu.

Tato podmínka je vyžadována kvůli tomu, aby byla zajištěna čitelnost dat Národním archivem, jenž v případě, že digitální archiv zanikne, data převezme. V Metodickém pokynu je sice definována základní struktura výměnného balíčku, chybí však stanovení rozsahu a obsahu předávaných metadat. Tento problém by měl vyřešit projekt Technologické agentury ČR *Vytvoření standardů pro komunikaci informačního systému digitálního archivu s jeho okolím*, který začal být realizován v roce 2023. Jedním z výstupů by mělo být stanovení výměnného formátu, který by umožnil realizovat výměnu dat mezi Digitálními archivy. Lze konstatovat, že postup akreditace Digitálních archivů podle české archivní legislativy tak aktuálně představuje hledání mezi různými mechanismy pro certifikaci digitálních repozitářů. Cílem je věrohodně potvrdit důvěryhodnost úložiště, a přitom nevyčerpat dostupné lidské i finanční zdroje instituce. Jak bylo výše uvedeno, Národní digitální archiv získal oprávnění k ukládání archiválií v elektronické podobě z definice. Další výjimky náleží bezpečnostním archivům. Všechny ostatní archivy musí, pokud chtějí oprávnění získat, projít procesem popsaným v zákoně a v příslušném Metodickém návodu. Dosud se o to pokusily dva archivy: před rokem 2019 Archiv Masarykovy univerzity, v roce 2021 Vojenský historický archiv. Ani jeden z nich neuspěl (BERNAS et al., 2019).

V závěrečné pasáži této kapitoly se podrobněji seznámíme se strukturou Metodického návodu. Vlastní text návodu je doplněn třemi přílohami, v nichž jsou obsažena technická kritéria pro stavební a bezpečnostní nároky na budovy, ve kterých jsou umístěna úložiště s daty, popis základní struktury informačního balíčku pro přenos do NDA a český překlad nestoru. V rámci popisu balíčku jsou využity specifikace z projektu E-ARK (European Archival Records and Knowledge Preservation). Specifikace CSIP *(Common Specification for Information Packages)* upřesňuje společné vlastnosti balíčků SIP, AIP i DIP. Struktura informačního balíčku je zde navržena poměrně volně. V případě Pečeti nestoru jsou do metodického návodu přebrána všechna její kritéria (celkem 34). Ve vlastní žádosti o získání oprá-

není k ukládání, která je specifikována právě v metodickém návodu, musí být všechna aplikována a doložena. Kritéria jsou rozložena k různým požadavkům zákona, metodický návod připouští, že část podkladů pro akreditaci digitálního archivu lze nahradit předložením výsledků formální certifikace digitálního repozitáře dle ISO 16363 nebo DIN 31644.

I kdyby však archiv získal certifikaci podle těchto norem, nevyjímá ho to z povinnosti podstoupit akreditační proces dle Metodického návodu. Z popsaného je zřejmé, že klíčovým pro celý proces je doložení kritérií Pečeti nestoru. Podle ní je digitální archiv definován jako organizace sestávající z osob a technických systémů, která převzala odpovědnost za dlouhodobé uchovávání, dlouhodobou dostupnost informací a jejich zpřístupnění určené cílové skupině. Objektem certifikace tak vždy nutně musí být jak softwarové nástroje, tak celkové workflow digitálního archivu, jeho lidské zdroje a politiky dlouhodobého uchovávání digitálních dokumentů. Za klíčové považuje Metodický návod naplnění kritérií K1-13 a K34. K rozhodnutí o udělení či neudělení souhlasu k ukládání digitálních archiválií má Národní archiv a OASSS jeden rok.

Po dokončení zmíněného projektu TAČR lze očekávat, že některé specializované archivy se pokusí akreditovat svůj digitální archiv. Pokud k tomu dojde, vznikne v ČR unikátní situace, kdy se de facto poměrně velký počet institucí pokusí získat Pečeť nestoru. Počet žádajících organizací převýší aktuální počet držitelů této certifikace. V případě úspěšných žádostí bude v ČR jedna z nejvyšších koncentrací certifikovaných digitálních archivů splňujících kritéria dle druhého stupně Jednotného rámce pro audit a certifikaci. Sledování procesu umožní srovnávat plány dlouhodobého uchovávání digitálních dokumentů ve větším počtu institucí, srovnávat jejich personální zázemí a další kritéria. I z globálního hlediska půjde o pozoruhodný stav, který poskytne prostor pro teoretické prohloubení certifikačních procesů.

Seznam kritérií Pečeti nestor

- K1 Výběr informačních objektů a jejich reprezentací
- K2 Odpovědnost za uchování
- K3 Cílové skupiny
- K4 Přístup
- K5 Interpretovatelnost
- K6 Právní a smluvní základ
- K7 Právní shoda
- K8 Financování
- K9 Lidské zdroje
- K10 Organizace a procesy
- K11 Činnosti související s uchováváním
- K12 Krizové řízení a exit plán
- K13 Významné vlastnosti
- K14 Integrita: rozhraní pro příjem
- K15 Integrita: Funkcionality archivního úložiště
- K16 Integrita: uživatelské rozhraní
- K17 Autenticita: Příjem
- K18 Autenticita: činnosti související s uchováváním
- K19 Autenticita: Přístup
- K20 Technická oprávnění
- K21 Vstupní informační balíčky
- K22 Převod vstupních informačních balíčků do archivních informačních balíčků
- K23 Archivní informační balíčky
- K24 Interpretovatelnost archivních informačních balíčků
- K25 Převod archivních informačních balíčků do výstupních informačních balíčků
- K 26 Výstupní informační balíčky
- K27 Identifikace
- K28 Popisná metadata
- K29 Strukturální metadata
- K30 Technická metadata
- K31 Zaznamenání postupů při činnostech související s uchováváním
- K32 Administrativní metadata
- K33 IT infrastruktura
- K34 Bezpečnost

Závěr

Za posledních třicet let zároveň s masivní digitalizací společnosti získalo lidstvo nový druh kulturního dědictví v podobě nejrůznějších digitálních objektů. Paměťové instituce velmi rychle identifikovaly potřebu digitální objekty ukládat a chránit, což vyústilo v poměrně překotný vývoj na poli standardizace postupů, požadavků na repozitáře a vývoje jednotných standardů a souborových formátů. V současné době tak má velká část zemí na planetě digitální knihovny a archivy, které se starají jak o dokumenty, které byly do digitálního prostředí transformovány, tak o dokumenty, které v digitálním prostředí vznikly.

Konkrétně v České republice archivy a knihovny v současné době uchovávají tři skupiny dokumentů: textové, zvukové a e-born, přičemž v těchto množinách nalezneme cokoliv od monografií a dokumentů, vzniklých z činnosti státní správy, přes kroniky, matriky, mapy, grafiky až po gramofonové desky a fonografické válečky. V knihovním prostředí lze s dalším rozvojem programů na ochranné reformátování co nejširšího množství typů dokumentů v blízké budoucnosti očekávat výsledek standardizačního úsilí pro digitalizaci dalších typů písemných a obrazových dokumentů či jejich souborů, datových CD a dalších nosičů audiovizuálního obsahu.

Digitalizace knihovních fondů tak v České republice již patří mezi standardní činnosti velkých knihoven. Národní knihovna společně s Moravskou zemskou knihovnou v Brně mají sdílenou digitalizační linku a LTP úložiště, jež jsou financovány z rozpočtu zřizovatele těchto knihoven, Ministerstva kultury. Dále jsou knihovní fondy digitalizovány na krajských digitalizačních jednotkách, jež byly zřízeny buď přímo v krajských knihovnách, nebo na tamních úřadech.

Nákladnou digitalizační činnost knihovny nehradí pouze z prostředků vlastního rozpočtu, ale z velké části za pomoci různých dotačních iniciativ. Dlouhodobě fungují programy NAKI (program na podporu aplikovaného výzkumu v oblasti národní a kulturní identity 2023–2030). Současný běh NAKI III (2023–2027) mimo jiné poskytuje řešitelům prostředky pro vývoj nástrojů, které pomocí strojového

učení zefektivní práci s velkým počtem digitalizovaných dokumentů, jenž se nacházejí v digitálních knihovnách, archivech a repozitářích. Na samotnou digitalizaci je rovněž možné žádat dotaci prostřednictvím jiného dlouhodobě podporovaného dotačního mechanismu, a to Veřejné informační služby knihoven (VISK), konkrétně VISK 6 (digitalizace historických dokumentů) a VISK 7 (digitalizace novodobých tištěných a zvukových dokumentů). V posledních dvou letech specifické financování umožňuje také iniciativa Digitalizace kulturního a kreativního sektoru, spadající pod Národní plán obnovy, která by měla umožnit obměnu infrastruktury jednotlivých digitalizačních linek a jednotek.

Co se týče aktivity v mezinárodním společenství, v době, kdy vzniká tato monografie, se chystají změny hned v několika zásadních standardech. Kongresová knihovna oznámila změny v katalogizačním standardu MARC21, které se jistě promítnou i do popisných metadat. Dále se počítá se změnami ve formátu METS ve variantě 2.0, která by oproti původní verzi měla být mnohem jednodušší a variabilnější v možnostech popisu. Zároveň s tím pozorujeme, jak legislativní změny ohledně archivování dokumentů ve státních organizacích Spojených států amerických vyvolávají potřebu diskuse o všeobecné nutnosti zjednodušovat mezinárodně platné standardy pro digitální archivaci, tak aby s dokumenty zároveň bylo nakládáno podle pravidel dobré praxe, a zároveň bylo jednodušší tato pravidla dodržovat. Změny ve způsobu nakládání s dokumenty vzniklými z činností státního aparátu ovšem nastávají celosvětově a paměťové instituce, které mají garantovat jejich uchování, na ně musejí reagovat.

Je tedy jisté, že ve spojení se stále více digitalizovanou společností, státní správou a přesunem výrazné většiny mezilidské komunikace do elektronického prostředí bude obor digitální archivace nevyhnutelně konfrontován s potřebou vytvářet teoretická východiska a následně standardy a nástroje k popisu a uchování nových typů dokumentů, digitálních objektů a médií. Zároveň je ale nucen co nejvíce se otevřít společnosti mimo současnou odbornou komunitu, protože digitální archivace se stane všeobecným zájmem napříč obory lidské činnosti. To může přinést další interoperabilitu mezi

standardy a jejich zjednodušení, a zároveň paradoxně potřebu o co nejdetailnější možnosti popisu.

Větší a větší roli bude hrát snaha o propojení dat napříč katalogy, databázemi, jednotlivými digitálními archivy, paměťovými institucemi a zdroji na internetu. Lze předpokládat masivní rozvoj již nyní probíhajícího trendu agregátorů dat z digitálních knihoven, jako jsou například Europeana, české Manuscriptorium či třeba Google Books.

Snad největší výzvu ale nyní představuje oblast vzdělávání budoucích informačních pracovníků. Obory a školy, na kterých se v současné době připravují knihovníci, archiváři, muzejní pracovníci a spřízněné profesionálové, stojí před nelehkým úkolem: ve spolupráci s kolegy z cílových paměťových institucí transformovat obsah studia tak, aby v době neustále se vyvíjejících technologií a měnících se požadavků zůstalo relevantní a sloužilo potřebám společnosti, která žije zároveň ve dvou světech – hmotném a digitálním.

Shrnutí

Publikace se zaměřuje na problematiku dlouhodobé archivace digitálních dokumentů v paměťových institucích. Stručně seznamuje s historií oboru péče o digitalizované a digitální dokumenty ve světě i v České republice a nastiňuje vznik odborných společností a postupnou standardizaci na poli formátů a fungování digitálních archivů, zejména datového archivu OAIS. Dále uvádí problematiku digitálního objektu, intelektuální entity a digitalizačních balíčků, souborových a metadatových formátů. V teoretické i praktické rovině se pak věnuje repozitářům, nástrojům LTP úložišť, dlouhodobé archivaci a bitové ochraně a identifikaci digitálních dokumentů. V poslední části publikace uvádí praktickou stránku soudobého dlouhodobého uchovávání v českých paměťových institucích na příkladu Národní knihovny České republiky a v tuzemských archivech.

Resume

This work is centered around long-term preservation of digital documents in memory institutions. The first part briefly discusses the history of the field of digital preservation, founding of digital-preservation initiatives and beginning of standardization both in the international and Czech environment. In this first part it also describes the concept of OAIS archive and its application, theory of digital object, intellectual entity and preservation packages. It also contains a chapter on internationally used file and metadata formats. Next part of the book is dedicated to repositories, LTP storages and its tools, long-term preservation theory and bit-level preservation of digital objects, also discussing its identification. Last part of this publication describes current state-of--affairs in Czech libraries and public archives.

O autorech

PhDr. Zdeněk Vašek, Ph.D.
je absolvent oborů historie a politologie na FF UK, archivář, knihovník a historik se specializací na starší české dějiny. V letech 2012–2017 zaměstnanec NK ČR, od roku 2014 jako vedoucí Oddělení pro standardy dlouhodobého uchovávání digitálních dokumentů, od roku 2017 zaměstnanec Ústavu dějin a archivu Univerzity Karlovy, spoluautor funkčního řešení Archivního informačního systému (Digitálního archivu) Univerzity Karlovy, který byl oceněn cenou CNZ 2021 v oblasti dlouhodobého uchovávání dokumentů a informací, člen řešitelského týmu ARCLib. Spoluautor metodik „Metodika logické ochrany digitálních dat v systému ARCLib" a „Metodika pro přidělování a správu životního cyklu unikátních perzistentních identifikátorů digitálních dokumentů podle standardu URN:NBN". Člen Vědecké rady Národního archivu ČR, nositel Prémie Ceny Miroslava Ivanova za literaturu faktu 2019, externí vyučující ÚISK FF UK. Specializuje se na oblast Digital Preservation, identifikátory digitálních dat a procesní postupy v dlouhodobých repozitářích.

Mgr. Pavlína Kočišová
vystudovala obor Informační věda a knihovnictví na Filosoficko-přírodovědecké fakultě Slezské univerzity v Opavě. Od roku 2017 působí jako metadatová specialistka na Oddělení standardů digitálních sbírek Národní knihovny ČR. Profesně se zaměřuje zejména na popisná metadata a metodickou činnost v knihovní digitalizaci. V minulých letech se zapojila do řešitelských týmů projektů ArcLib a Nový Fonograf, v současnosti je členkou pracovních skupin projektů PerMonik a ReČek. Do této publikace přispěla kapitolami o referenčním modelu OAIS, digitálních objektech a metadatových formátech pro dlouhodobé uložení. V roli editorky se podílela na přípravě této knihy k vydání.

Bc. Václav Jiroušek
vystudoval obor Informační vědy a knihovnictví na FF UK. Od roku 2012 zaměstnán v Národní knihovně ČR na pozicích spojených s dlouhodobou archivací a digitalizací, v letech 2014–2018 vedoucí Oddělení LTP úložiště, 2020–2021 vedoucí Oddělení výběru knihovního fondu pro reformátování, v letech 2021–2023 vedoucí Oddělení standardů. Specializuje se na digitalizaci, souborové formáty, dlouhodobou ochranu digitálních dat a popularizaci digitalizačních aktivit NK ČR. Do této publikace přispěl zejména kapitolou o dlouhodobé archivaci digitálních dat v knihovnách.

Vojtěch Kopský, Ph.D.
vystudoval obor vývojová biologie na Přírodovědecké fakultě Univerzity Karlovy. Poté se v Endokrinologickém ústavu podílel na výzkumu vlivu neuropeptidů na příjem potravy na modelech anorexie a obezity. Dále pracoval jako produktový manager pro firmy Serva a Duchefa u jejich českého distributora, firmy BioTech. V současnosti působí jako formátový specialista na Oddělení standardů digitálních sbírek Národní knihovny ČR. Do této publikace přispěl kapitolou o souborových formátech pro dlouhodobé uložení.

Mgr. Jan Bilwachs
působí od října roku 2017 v Národní knihovně ČR na pozici obsahového správce dlouhodobého úložiště dat (LTP). Kromě jeho hlavní agendy, jíž je příjem, validace a ukládání digitálních dokumentů od externích producentů, se v současnosti podílí na vývoji aplikace Komplexní validátor. V minulosti byl rovněž členem řešitelského týmu projektu ARCLib.

Mgr. Filip Pavčík, PhD.
vystudoval na Katedře historie Filozofické fakulty Univerzity Konštantína Filozofa v Nitře. Doktorské studium absolvoval v letech 2015–2019 na Historickém ústavu SAV v Bratislavě, obor Slovenské dějiny. Během studií několik let pracoval v neziskové organizaci Post Bellum SK, kde působil jako koordinátor slovenské části databáze vzpomínek pamětníků. Od roku 2020 pracuje v Národní knihovně ČR na Oddělení

standardů digitálních sbírek, přičemž profesně se zabývá zejména identifikací digitálních dokumentů a popisnými metadaty. Do této publikace přispěl kapitolou o perzistentních identifikátorech.

Mgr. Petr Cajthaml
absolvoval obor historie na FF UK. V letech 2000–2007 zaměstnanec Úřadu dokumentace a vyšetřování zločinů komunismu (historik). Od roku 2008 zaměstnanec Ústavu dějin a archivu Univerzity Karlovy, od roku 2013 vedoucí Archivu Univerzity Karlovy. Hlavní řešitel digitalizačního projektu Archiv UK „Studenti pražských univerzit". Autor koncepce archivace digitálních dokumentů Univerzity Karlovy a spoluautor funkčního řešení Archivního informačního systému (Digitálního archivu) Univerzity Karlovy, který byl oceněn cenou CNZ 2021 v oblasti dlouhodobého uchovávání dokumentů a informací. Předseda Odborné skupiny vysokoškolských a vědeckých archivů České archivní společnosti. Specializuje se na archivaci digitálních dokumentů a informační systémy pro správu archiválií v prostředí nestátních veřejných archivů.

Seznam zkratek

AACR2	Anglo-American Cataloguing Rules, 2nd Edition
AES	Audio Engineering Society
AHMP	Archiv hlavního města Prahy
AIP	Archival Information Package
ALR	Association of Research Libraries
ALTO	Analyzed Layout and Text Object
ANSI	American National Standards Institute
API	Application Programming Interface
AWS	Amazon Web Services
BL	British Library
CCSDS	Consultative Committee for Space Data Systems
CD	Compact Disc
CD-R	Compact Disc Recordable
CD-ROM	Compact Disc Read-Only Memory
CDO	Content Data Object
CHIMERA	Czech History Information Management and electronic Records Archiving
CIFS	Common Internet File System
CLI	Command-line Interface (příkazový řádek)
CLOCKSS	Controlled LOCKSS
CNA	Converged Network Adapter
CNZ	Co po nás zbude (občanské sdružení)
CPA	Commission on Preservation and Access
CRC	Cyclical Redundancy Check
CRUSH	Controlled Replication Under Scalable Hashing
CSIP	Common Specification for Information Packages
CSV	Comma-separated Values
CZIDLO	Czech Identification and Localization Tool
čČNB	číslo České národní bibliografie
ČIDLO	Český identifikační a lokalizační systém
ČNB	Česká národní bibliografie
ČR	Česká republika
ČSN	Česká technická norma (původně Československá státní norma)

DA	digitální archiv
DC	Dublin Core
DCC	Digital Curation Centre
DCMI	Dublin Core Metadata Initiative
DIN	Deutsche Industrie-Norm
DIP	Dissemination Information Package
DKRVO	Dlouhodobý koncepční rozvoj výzkumných organizací
DMF	Definice metadatových formátů
DPC	Digital Preservation Coalition
DPC RAM	DPC Rapid Assessment Model
DPTR	Digital Preservation Technical Registry
DRM	Digital Rights Management
DROID	Digital Record Object Identification
DTD	Document Type Definition
E-ARK	European Archival Records and Knowledge Preservation
EAD	Encoded Archival Description
EBU	European Broadcast Union
EPUB	electronic publication
EU	Evropská unie
EXIF	Exchangeable Image File Format
FADGI	Federal Agencies Digital Guidelines Initiative
FC	Fibre Channel
FCoE	Fibre Channel over Ethernet
FDD	Format Description Document
FEC	Forward Error Correction
FF UK	Filozofická fakulta Univerzity Karlovy
FIDO	Format Identification for Digital Objects
FITS	File Information Tools Set
FRBR	Functional Requirements for Bibliographic Records
GDFR	Global Digital Format Registry
GIF	Graphics Interchange Format
GUI	Graphical User Interface (grafické uživatelské rozhraní)
HDFS	Hadoop Distributed File System

HW	hardware
IANA	Internet Assigned Numbers Authority
IASA	International Association of Sound and Audiovisual Archives
IBM	International Business Machines Inc.
IEC	International Electrotechnical Commission
IETF	Internet Engineering Task Force
IIA	IBM Information Archive
IP	Internet Protocol
iPRES	International Conference on Digital Preservation
iRODS	Integrated Rule-Oriented Data System
iSCSI	Internet Small Computer System Interface
ISBN	International Standard Book Number
ISMN	International Standard Music Number
ISO	International Organisation for Standardization
ISSN	International Standard Serial Number
IT	informační technologie
ITIL	Information Technology Infrastructure Library
JHOVE	JSTOR/Harvard Object Validation Environment
JP2	JPEG 2000
JPEG	Joint Photographic Experts Group
JPL	Jet Propulsion Laboratory
JRE	Java Runtime Environment
JSTOR	Journal Storage
KNAV	Knihovna Akademie věd ČR
KOST-CECO	Koordinační centrum pro dlouhodobou archivaci elektronických dokumentů Švýcarského federálního archivu
LAN	Local Area Network
LOC	Library of Congress
LOCKSS	Lots of Copies Keep Stuff Safe
LPCM	Linear Pulse Code Modulation
LTO	Linear Tape Open
LTP	Long-term Preservation
LUN	Logical Unit Number
LVM	Logical Volume Management

LZW	Lempel-Ziv-Welch (kompresní algoritmus)
MARC	Machine-readable Cataloging
MD5	Message Digest Algorithm 5
METS	Metadata Encoding and Transmission Standard
MIME	Multipurpose Internet Mail Extension
MIX	Metadata for Images in XML Standard
MK ČR	Ministerstvo kultury ČR
MLC	multi-level cells
MODS	Metadata Object Description Schema
MOSFET	Metal Oxide Semiconductor Field Effect Transistor
MPEG	Moving Pictures Experts Group
MUNI	Masarykova univerzita
MV ČR	Ministerstvo vnitra ČR
MXF	Material Exchange Format
MZK	Moravská zemská knihovna v Brně
NA ČR	Národní archiv ČR
NAKI	Program aplikovaného výzkumu a vývoje národní kultury a identity MK ČR
NARA	National Archives and Records Administration
NAS	Network Attached Storage
NASA	Národní úřad pro letectví a vesmír
NCSA	National Center for Supercomputing Applications
NDA	Národní digitální archiv
NDK	Národní digitální knihovna
NDSA	National Digital Stewardship Alliance
NFS	Network File System
NFS	Novodobé fondy a sbírky
NISO	National Information Standards Organization
NK ČR	Národní knihovna České republiky
NSESSS	Národní standard pro elektronické systémy spisové služby
NSLA	National and State Libraries Australasia
NZME	New Zealand Metadata Extraction Tool
OAIS	Open Archival Information System
OASSS MV ČR	Odbor archivní správy a spisové služby Ministerstva vnitra ČR
OCLC	Online Computer Library Center

OCR	Optical Character Recognition
ODIF	Odbor digitálních fondů
ONDS	Odbor novodobých digitálních sbírek
OPF	Open Preservation Foundation
OS	operační systém
OSD	Object Storage Daemon
PASIG	Preservation and Archiving Special Interest Group
PDF	Portable Document Format
PDF/A	PDF pro Archivaci
PDI	Preservation Descriptive Information
PLANETS	Preservation and Long-Term Access through Networked Services
PLATTER	Planning Tool for Trusted Electronic Repositories
PNG	Portable Network Graphics
PREMIS	Preservation Metadata Implementation Strategies
PRONOM	Public Record Office and Nôm (formátový registr)
PSC	Preservation Storage Criteria
PSP	Producer Submission Package
PUID	PRONOM's Persistent Unique Identifier
RADOS	Reliable Autonomous Distributed Object Storage
RAID	Redundant Array of Inexpensive Disks
RDA	Resource Description and Access
REST	Representational State Transfer
RFC	Request for Comments
RGB	red-green-blue
RIFF	Resource Interchange File Format
RLG	Research Libraries Group
S3	Simple Storage Service
SAN	Storage Area Network
SAV	Slovenská akademie věd
SCSI	Small Computer System Interface
SDB	Safety Deposit Box
SSD	Solid State Drive
SDK	Software Development Kit
SDS	Software Defined Storage
SHA	Secure Hash Algorithm

SIP	Submission Information Package
SLC	single-level cells
SMB	Server Message Block
TA ČR	Technologická agentura ČR
TCP	Transmission Control Protocol
TEI	Text Encoding Initiative
TIFF	Tagged Image File Format
TLC	triple-level cells
TRAC	Trustworthy Repositories Audit and Certification
UDFR	Unified Digital Format Registry
ÚISK	Ústav informačních studií a knihovnictví
UK	Univerzita Karlova
UNESCO	Organizace OSN pro vzdělání, vědu a kulturu
URL	Universal Resource Locator
URN:NBN	Uniform Resource Name: National Bibliography Number
USA	Spojené státy americké
VISK	Veřejné informační služby knihoven
VRA	Visual Resources Association
VŠ	vysoká škola
W3C	Word Wide Web Consortium
WORM	Write Once Read Many
XML	eXtensible Markup Language
ZAF	Zpracování archivních formátů

Použitá literatura

- ADAMS, Margaret O. a BROWN, Thomas E., 2000. Myths and Realities About the 1960 Census. In: *Genealogy Notes* [online]. 2000, **32**(4). Dostupné z: https://www.archives.gov/publications/prologue/2000/winter/1960-census.html [cit. 2023-09-30].
- ADDIS, Matthew, 2020. *Which checksum algorithm should I use?* Digital Preservation Coalition. DOI: 10.7207/twgn20-12.
- ALTMAN, Micah a LANDAU, Richard, 2020. Selecting Efficient and Reliable Preservation Strategies. In: *International Journal of Digital Curation*, **15**(1), s. 18. ISSN 1746-8256. DOI: 10.2218/ijdc.v15i1.727.
- *ALTO Principles*, 2016. The Library of Congress [online]. Washington (DC): The Library of Congress. Dostupné z: https://www.loc.gov/standards/alto/description.html [cit. 2023-06-22].
- *ALTO: Technical Metadata for Layout and Text Objects*, 2022. The Library of Congress [online]. Washington (DC): The Library of Congress. Dostupné z: https://www.loc.gov/standards/alto/ [cit. 2023-06-22].
- ANSI/NISO Z39.87-2006, 2006. *Data Dictionary – Technical Metadata for Digital Still Images.* Bethesda (MD): NISO Press, xiv, 107 s. ISBN 978-1-937522-37-7. ISSN 1041-5653.
- APACHE HADOOP, 2023a. *HDFS Architecture* [online]. In: Apache Hadoop 3.3.6. Dostupné z: https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HdfsDesign.html [cit. 2023-08-21].
- APACHE HADOOP, 2023b. *HDFS Erasure Coding* [online]. In: Apache Hadoop 3.3.6. Dostupné z: https://hadoop.apache.org/docs/stable/hadoop-project-dist/hadoop-hdfs/HDFSErasureCoding.html [cit. 2023-08-21].
- APACHE TIKA, 2022a. In: *Community Owned digital Preservation Tool Registry (COPTR)* [online]. last modified on 18 January 2022. Dostupné z: https://coptr.digipres.org/index.php/Apache_Tika [cit. 2023-09-22].

- APACHE TIKA, 2022b. In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 2001–, last edited on 29 April 2022. Dostupné z: https://en.wikipedia.org/wiki/Apache_Tika [cit. 2023-09-22].
- ARAVINDAN, Arunjith, 2014. *How to avoid hash collisions when using MySQL's CRC32 function* [online]. Dostupné z: https://www.percona.com/blog/how-to-avoid-hash-collisions-when-using-mysqls-crc32-function/ [cit. 2023-08-25].
- ARMS, Caroline a FLEISCHHAUER, Carl, 2005. Digital Formats: Factors for Sustainability, Functionality, and Quality [online]. In: *IS&T Archiving 2005 Conference*, Washington, D.C. Dostupné z: https://memory.loc.gov/ammem/techdocs/digform/Formats_IST05_paper.pdf [cit. 2023-08-16].
- ARSLAN, Suayb S. et al., 2022. *LTO-9 technology and user data reliability analysis* [online]. The LTO Program. Dostupné z: https://www.lto.org/wp-content/uploads/2022/08/LTO-UBER-Technical-Paper-August-2022.pdf [cit. 2023-09-13].
- BARNESS, Sarah et al., 2018. *Results of fixity survey* [online]. NDSA. An NDSA report. Dostupné z: https://osf.io/https://osf.io/grfpa [cit. 2023-08-25].
- BAUCOM, Erin, 2019. A Brief History of Digital Preservation. In: *Mansfield Library Faculty Publications.* 31. Dostupné z: https://scholarworks.umt.edu/ml_pubs/31 [cit. 2023-09-20].
- BÁRTA, Stanislav, BRZOBOHATÁ, Hana, ČERVENÁ, Radana, JELÍNEK, Jiří, STODŮLKA, Zbyšek a ZEMÁNKOVÁ, Michaela, 2019. *Digitální archivnictví.* 1. elektronické vyd. Brno: Masarykova univerzita. 134 s. ISBN 978-80-210-9450-5. doi:10.5817/CZ.MUNI.M210-9450-2019; dostupné z: https://munispace.muni.cz/library/catalog/view/1407/3886/1725-1/0#preview [cit. 2023-09-29].

- BEŇAČKOVÁ, Miroslava, KOČIŠOVÁ, Pavlína, KOPSKÝ, Vojtěch a OSTRÁKOVÁ, Natalie, 2020a. Dlouhodobé uchování digitálních dat vzniklých digitalizací zvukových záznamů na fonografických válečcích a šelakových deskách. In: *Knihovna: knihovnická revue.* 2020, **31**(2), 45–61. ISSN 1801-3252. Dostupné z: https://knihovnarevue.nkp.cz/archiv/2020-2/recenzovane-prispevky/dlouhodobe-uchovavani-digitalnich-dat-vzniklych-digitalizaci-zvukovych-zaznamu-na-fonografickych-valeccich-a-selakovych-deskach [cit. 2023-09-29].
- BEŇAČKOVÁ, Miroslava, KOČIŠOVÁ, Pavlína, KOPSKÝ, Vojtěch, MALLY, Richard a OSTRÁKOVÁ, Natálie, 2020b. Signifikantní vlastnosti: příspěvek ke kolektivnímu nevědomí. In: *ProInflow,* **12**(2). https://doi.org/10.5817/ProIn2020-2-3. Dostupné z: https://journals.phil.muni.cz/proinflow/article/view/2020-2-3/15518 [cit. 2023-09-29].
- BERNAS, Jiří, STODŮLKA, Zbyšek a VOJÁČEK, Milan, 2019. Certifikace Nestor. in: *LTP 2019. Nové trendy a východiská pri budovaní LTP archívov. Zborník príspevkov zo 4. medzinárodnej konferencie o dlhodobej archivácii.* Bratislava: Univerzitná knižnica, s. 106–114. ISBN 978-80-89303-77-9.
- BLAKESLEE, Sandra, 1990. Lost on Earth: Wealth of Data Found in Space. In: *New York Times* [online]. 1990, 140, March 2020, 1990. Dostupné z: https://web.archive.org/web/20180829074209/https://www.nytimes.com/1990/03/20/science/lost-on-earth-wealth-of-data-f [cit. 2023-09-30].
- BLOOD, George, 2011. *Refining Conversion Contract Specifications: Determining Suitable Digital Video Formats for Medium-term Storage* [online]. FADGI. Dostupné z: https://www.digitizationguidelines.gov/audio-visual/documents/IntrmMastVidFormatRecs_20111001.pdf [cit. 2023-08-16].
- BRITISH LIBRARY DIGITAL PRESERVATION TEAM, 2015. *TIFF Format Preservation Assessment. Version 1.3* [online]. British Library. Dostupné z: https://wiki.dpconline.org/images/6/64/TIFF_Assessment_v1.3.pdf [cit. 2023-09-14].

- BRITISH LIBRARY DIGITAL PRESERVATION TEAM, 2016. *WAV Format Preservation Assessment. Version 1.0* [online]. British Library. Dostupné z: https://wiki.dpconline.org/images/4/46/WAV_Assessment_v1.0.pdf [cit. 2023-09-29].
- BRITISH LIBRARY DIGITAL PRESERVATION TEAM, 2018. *FLAC Format Preservation Assessment. Version 1.00* [online]. British Library. Dostupné z: https://wiki.dpconline.org/images/f/fe/FLAC_Assessment_v1.0.pdf [cit. 2023-09-29].
- BRITISH LIBRARY DIGITAL PRESERVATION TEAM, 2019. *PDF Format Preservation Assessment Part 1: PDF. Version 1.5* [online]. British Library. Dostupné z: https://wiki.dpconline.org/images/f/ff/PDF_Assessment_v1.5.pdf [cit. 2023-09-29].
- BRITISH LIBRARY DIGITAL PRESERVATION TEAM, 2019. *EPUB Format Preservation Assessment [online]. Version 1.4* [online]. British Library. Dostupné z: https://wiki.dpconline.org/images/7/73/EPUB_Assessment_v1.4a.pdf [cit. 2023-09-30].
- BROWN, Adrian, 2008. *Selecting File Formats for Long-Term Preservation* [online]. Dostupné z: https://cdn.nationalarchives.gov.uk/documents/selecting-file-formats.pdf. [cit. 2023-09-12].
- BUCKLEY, Robert, 2013. *Using Lossy JPEG 2000 Compression For Archival Master File* [online]. The Library of Congress. Dostupné z: https://www.digitizationguidelines.gov/still-image/documents/JP2LossyCompression.pdf [cit. 2023-08-16].
- BUITENHUIS, Derek, 2019. Wrote an FFV1 Decoder in Go for Fun. In: *No Time To Wait.* Dostupné z YouTube: https://www.youtube.com/watch?v=4HB7v7dItWE [cit. 2023-08-22].
- CAJTHAML, Petr, 2017. *Metodika pro skartaci dokumentů evidovaných v systémech elektronické spisové služby v prostředí vysokých škol* [online]. Praha: Ústav dějin a archiv Univerzity Karlovy. Dostupné z: https://is.muni.cz/digitalniuniverzitaMetodika_skartace_UK.pdf [cit. 2023-09-23].

- CAJTHAML, Petr a PAVLÁSKOVÁ, Eliška, 2018. Evolution of Electronic Management Systems, Digital Archiving and Czech Universities: From Student Information Systems to Digital Records Management and Long Term Preservation. In: BLECHER, Jens, Sabine HAPP a Juliane MIKOLETZKY. *Normen und Ethos: Schreiben Archivarinnen und Archivare Geschichte?* Leipzig: Leipziger Universitätsverlag, 2018, s. 193–202. ISBN 978-3-96023-188-2.
- CAPLAN, Priscilla, 2018. PREMIS a jak mu porozumět. Praha: Univerzita Karlova, 26 s.
- CEPH DOCUMENTATION, 2016a. *Architecture*. In: Ceph Documentation [online]. Dostupné z: https://docs.ceph.com/en/latest/architecture/ [cit. 2023-08-21].
- CEPH DOCUMENTATION, 2016b. *Erasure code.* In: Ceph Documentation [online]. Dostupné z: https://docs.ceph.com/en/latest/rados/operations/erasure-code/ [cit. 2023-08-22].
- CEPH DOCUMENTATION, 2016c. *Ceph Object Gateway*. In: Ceph Documentation [online]. Dostupné z: https://docs.ceph.com/en/latest/radosgw/ [cit. 2023-08-28].
- CUBR, Ladislav, 2010a. *Dlouhodobá ochrana digitálních dokumentů.* 1. vyd. Praha: Národní knihovna České republiky. 154 s.
- CUBR, Ladislav, 2010b. Budování důvěryhodného systému trvalé identifikace digitálních dokumentů. In: *Knihovna* [online], **21**(1). Dostupné z: https://oldknihovna.nkp.cz/knihovna101/10123.htm [cit. 2023-08-25].
- CUBR, Ladislav, [2015]. *Využití perzistentních identifikátorů pro elektronické publikace.* Praha: Národní knihovna ČR. Interní studie pro projekt „správa elektronických publikací v síti knihoven ČR".
- CUBR, Ladislav, 2016. Formátová strategie LTP úložiště NK ČR. In: KELEMENOVÁ, Lucia, (ed.). *CDA 2016: Formátové výzvy LTP: Zborník prispevkov z 1. mezinárodnej konferencie o dlhodobej archivácii.* Bratislava: Univerzitná knižnica v Bratislave, 2016, s. 44–57. ISBN 978-80-89303-51-9. ISSN 2453-9309.

- CUBR, Ladislav, 2017. *Autenticita a digitální informace.* Disertační práce. Praha: Univerzita Karlova, Filozofická fakulta, Ústav informačních studií a knihovnictví. Vedoucí práce RNDr. Jiří Ivánek, CsC.
- CUBR, Ladislav a VAŠEK, Zdeněk, 2013. Identifikátory digitálních dokumentů se zaměřením na systém URN:NBN v ČR. In: *Čtenář* [online]. **65**(6). Dostupné z: https://svkkl.cz/en/ctenar/clanek/1254 [cit. 2023-08-25].
- CUBR, Ladislav, LODROVÁ, Iveta, ŘEHÁNEK, Martin a VAŠEK, Zdeněk, 2016. Srovnání vybraných národních identifikačních systémů užívajících identifikátory URN:NBN. In: *ProInFlow* [online], **8**(1), s. 14-53. DOI: https://doi.org/10.5817/ProIn2016-1-3. Dostupné z: https://journals.phil.muni.cz/proinflow/article/view/2016-1-3/15441 [cit. 2023-08-25].
- CUBR, Ladislav, OSTRÁKOVÁ, Natalie a KOČIŠOVÁ, Pavlína, 2020. *Metodika pro tvorbu balíčků SIP se zaměřením na digitalizáty tištěných dokumentů*, 91 s. [online]. Praha: Národní knihovna ČR. Dostupné z: https://invenio.nusl.cz/record/432324 [cit. 2023-08-25].
- CUBR, Ladislav, Natalie OSTRÁKOVÁ, Pavlína KOČIŠOVÁ, Vojtěch KOPSKÝ, Václav JIROUŠEK, Filip PAVČÍK, Ivo MILÁČEK a Květa FREMROVÁ, 2023. *Metodika pro tvorbu balíčků SIP se zaměřením na digitalizáty tištěných dokumentů. Verze 2.0* (rukopis). Praha: Národní knihovna České republiky.
- ČSN ISO 14721, 2014. *Systémy pro přenos dat a informací z kosmického prostoru – Otevřený archivační informační systém – Referenční model.* Praha: Úřad pro technickou normalizaci, metrologii a státní zkušebnictví, 111 s.
- DELJANIN, Sandra, 2012. Digital Obsolescence. In: *INFOtheca*, **13**(1), 43-53. Dostupné z: http://infoteka.bg.ac.rs/index.php/en/archives/2012/1/infoteka-13-3-2012-47-58 [2023-09-30].
- DjVu, 2023 [online], poslední aktualizace 23. července 2023, *Wikipedia.* Dostupné z: https://en.wikipedia.org/wiki/DjVu [cit. 2023-09-14].

- DVOŘÁK, Tomáš, KOUCKÝ, Karel, ŠULC, Jaroslav, VICHTA, Jiří a VOJÁČEK, Milan, 2015. *Metodika pro vytváření bezpečnostních kopií archiválií v digitální podobě.* Praha: Národní archiv [online]. Dostupné z:https://www.nacr.cz/wp-content/uploads/2019/05/metodika2015.pdf [cit. 2023-09-29].
- EASY INNOVA, 2017. *DPFManager* [online]. Dostupné z: https://github.com/EasyinnovaSL/DPFManager [cit. 2023-09-14].
- EUROPEAN BROADCASTING UNION, 2001. *BWF – a format for audio data files in broadcasting. Technical Specification, Version 1*. July 2001. Dostupné z: https://web.archive.org/web/20091229093941/http://tech.ebu.ch/docs/tech/tech3285.pdf [cit. 2023-09-29].
- FEDERAL AGENCIES DIGITIZATION GUIDELINES INITIATIVE, 2023. *Technical Guidelines for Digitizing Cultural Heritage Materials, 3rd. edition* [online]. Washington (DC): FADGI, May 2023. Dostupné z: https://www.digitizationguidelines.gov/guidelines/FADGI%20Technical%20Guidelines%20for%20Digitizing%20Cultural%20Heritage%20Materials_3rd%20Edition_05092023.pdf [cit. 2023-06-24].
- GIARETTA, David, 2011. *Advanced Digital Preservation.* Berlin: Springer-Verlag, 510 s. ISBN 978-3-642-16808-6.
- GILLIS, Alexander S. a KRANZ, Garry, 2021. What is an SSD (Solid-State Drive)? In: *TechTarget* [online]. Dostupné z: https://www.techtarget.com/searchstorage/definition/SSD-solid-state-drive [cit. 2023-08-26].
- GLUSTER DOCS, 2023. *Architecture*. In: Gluster Docs [online]. Dostupné z: https://docs.gluster.org/en/latest/Quick-Start-Guide/Architecture/#what-is-gluster-without-making-me-learn-an-extra-glossary-of-terminology [cit. 2023-08-27].
- HAKALA, Juha, 2010. *Persistent identifier - an overview* [online]. Dostupné z: https://www.academia.edu/77589367/Persistent_identifiers_an_overview [cit. 2023-08-25].
- HANNAN, Ed, 2016. *What is file storage?* In: TechTarget [online]. Dostupné z: https://www.techtarget.com/searchstorage/definition/file-storage [cit. 2023-08-27].

- HRZINOVÁ, JANA a JIROUŠEK, Václav, 2022. Výběr archivačních formátů pro povinný depozit e-publikací v ČR: EPUB a PDF/A jako řešení? In *IT-Lib*. Dostupné z: https://itlib.cvtisr.sk/wp-content/uploads/2022/07/1_2_2022_Hrzinova_Jirousek.pdf [cit. 2023-08-16].
- HUTAŘ, Jan, 2008. *Plnění Administrativní metadat: Určeno spolupracujícím knihovnám a dodavatelům* [online]. Verze 1.0. Praha: Národní knihovna ČR, 2008. Dostupné z: https://wayback.webarchiv.cz/wayback/20100428092629/http://digit.nkp.cz/Kramerius/AdminMetaData/ADMnarodniStandardVerze1_Zapis.pdf [cit. 2023-09-13].
- HUTAŘ, Jan, 2012. *Digitalizace, popis pomocí metadat a jejich formáty* [online]. Disertační práce. Praha: Univerzita Karlova, Filozofická fakulta, Ústav informačních studií a knihovnictví. Vedoucí práce Stanislav Kalkus. Dostupné z: https://dspace.cuni.cz/bitstream/handle/20.500.11956/44181/140015545.pdf?sequence=1&isAllowed=y [cit. 2023-09-07].
- HUTAŘ, Jan, 2016. Identifikace formátů – jednorázový nebo opakovaný proces? In *CDA 2016. Formátové výzvy LTP.* Bratislava: Univerzitná knižnica v Bratislavě, 2016, s. 35–43. ISBN 978-80-89303-51-9.
- HUTAŘ, Jan a MELICHAR, Marek, 2015a. *Nástroje pro digital preservation. In:* LTP-portál.cz. Web o digitální archivaci [online]. Brno: Moravská zemská knihovna. Dostupné z: https://ltp-portal.mzk.cz/digital-preservation/nastroje [cit. 2023-14-09].
- HUTAŘ, Jan a Marek MELICHAR, 2015b. Nástroje pro digitální archivaci. In: *Knihovna: knihovnická revue,* **26**(2), s. 69-82. ISSN 1801-3252. Dostupné také z: http://knihovnarevue.nkp.cz/aktualni-cislo/knihovny-a-informace/nastroje-pro-digitalni-archivaci [cit. 2023-09-09].
- HUTAŘ, Jan, Marek MELICHAR a Tomáš GEC, 2016. *Národní koncepce dlouhodobé ochrany digitálních dat v knihovnách* [online]. Dostupné z: https://ipk.nkp.cz/docs/Narodni_koncepce_dlouhodobe_ochrany_digitalnich_dat.pdf [cit. 2023-09-09].

- HUTAŘ, Jan, MIRANDA, Andrea, PAVLÁSKOVÁ, Eliška, VAŠEK, Zdeněk a HRUŠKA, Zdeněk, 2018. *Metodika logické ochrany digitálních dat* [online]. Praha: Knihovna AV ČR, 2018. Dostupné z: http://www.nusl.cz/ntk/nusl-371612 [cit. 2023-09-07].
- ISBN, 2023 [online], poslední aktualizace 16. června 2023, *Wikipedia.* Dostupné z: https://en.wikipedia.org/wiki/ISBN [cit. 2023-08-25].
- ITIL Foundation, 2019. *ITIL 4 Edition.* London: Axelos. ISBN 978-0-11-331607-6.
- INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2005. *ISO 19005-1:2005 Document management — Electronic document file format for long-term preservation — Part 1: Use of PDF 1.4 (PDF/A-1).* 1st edition. Geneva: ISO, 29 s.
- INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2007. *ISO/IEC 15444-3:2007. Information technology — JPEG 2000 image coding system: Motion JPEG 2000 — Part 3.* 2nd ed. Geneva: ISO, 29 s.
- INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2012. *ISO 14721:2012. Space data and information transfer systems – Open archival information system (OAIS) - Reference model.* 2nd ed. Geneva: ISO, 126 s.
- INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2019. *ISO/IEC 15444-1:2019 Information technology — JPEG 2000 image coding system — Part 1: Core coding system.* 4th. ed. Geneva: ISO, 234 s.
- INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2020. *ISO/IEC 23736-1:2020 Information technology — Digital publishing — EPUB 3.0.1 — Part 1: Overview.* 2020.1st edition. Geneva: ISO, 22 s.
- JASTRAB, Ben, 2008. *Introduction to Fibre Channel over Ethernet (FCoE)* [online]. EMC Corporation. Dostupné z: https://www.cisco.com/c/dam/global/no_no/assets/expo2009/pdf/EMC_Intro_to_FCoE.pdf [cit. 2023-09-23].
- *JHOVE – JSTOR/Harvard Object Validation Environment*, 2009. [online]. JSTOR and Harvard University Library, 2003–2009. Dostupné z: https://jhove.sourceforge.net/ [cit. 2023-09-30].

- JOSHI, Vaidehi, 2019. *Redundancy and Replication: Duplicating in a Distributed System* [online]. Dostupné z: https://medium.com/baseds/redundancy-and-replication-duplicating-in-a-distributed-system-7ab4322d7378 [cit. 2023-08-27].
- KEITEL Christian, SCHOGER, Astrid (ed.), 2013. *Vertrauenswürdige digitale Langzeitarchivierung nach DIN 31644.* Berlin: Beuth Verlag, 120 s. ISBN 978-3410234999.
- KEJSER, Ulla Bøgvad, NIELSEN, Anders a THIRIFAYS, Alex, 2011. Cost Aspects of Ingest and Normalization. In: *IPRES 2011 – 8th International Conference on Preservation of Digital Objects* [online]. Singapur: National Library Board Singapore, s. 107–115. ISBN 978-981-07-0441-4. Dostupný z: https://phaidra.univie.ac.at/open/o:294293 [cit. 2022-09-14].
- KINNAMAN, Alex a MUNSHOWER, Alan, 2022. *Green Goes with Anything. Decreasing Environmental Impact of Digital Libraries at Virginia Tech.* Glasgow, Scotland. DOI: http://doi.org/10.7207/ipres2022-recordings.
- KLEIN, Andy, 2022. *The Cost of Hard Drives Over Time* [online]. Dostupné z: https://www.backblaze.com/blog/hard-drive-cost-per-gigabyte/ [cit. 2023-08-26].
- KNOLL, Adolf, 1999. *Memoriae Mundi Series Bohemica – digitální zpřístupnění vzácných dokumentů.* Národní knihovna: knihovnická revue [online]. Národní knihovna ČR, 1999, (3), 105-109 [cit. 2023-09-07]. ISSN ISSN 1214-0678. Dostupné z: https://full.nkp.cz/nkkr/NKKR9903/9903105.html
- KNOLL, Adolf a PSOHLAVEC, Stanislav, 2002. *Zpráva o řešení projektu výzkumu a vývoje Optimalizace archivace a zpřístupnění digitálních dat: Závěrečná zpráva za léta 2001–2002.* Praha: Národní knihovna ČR, 2002. Dostupné také z: https://wayback.webarchiv.cz/wayback/20100415000000*/http://digit.nkp.cz/knihcin/digit/vav23/zpravaweb.doc [cit. 2023-09-29].

- KNOLL, Adolf, Jiří POLIŠENSKÝ a Stanislav PSOHLAVEC, 2004. *Závěrečná zpráva o řešení výzkumného záměru Digitální knihovna – produkce, ochrana a zpřístupnění digitálních dokumentů: 1999–2003.* Praha: Národní knihovna ČR, 2004. Dostupné také z: https://wayback.webarchiv.cz/wayback/20100428092737/http://digit.nkp.cz/PROJECTS_WEB/CEZFinal1999-2003.pdf [cit. 2023-09-29].
- KNIJF, J, 2014. *Quattro Pro for DOS:an obsolete format at last?* In: Open Preservation Foundation. Dostupné z: https://openpreservation.org/blogs/quattro-pro-dos-obsolete-format-last/ [cit. 2023-09-29].
- *Koncepce rozvoje knihoven v České republice na léta 2021–2027 s výhledem do roku 2030: knihovny – pilíře občanské společnosti, vzdělanosti a kultury,* 2020. [online]. Praha: Národní knihovna České republiky – Knihovnický institut. ISBN 978-80-7050-735-3. Dostupné z: https://ipk.nkp.cz/docs/koncepce-rozvoje-knihoven-2021-2027 [cit. 2023-09-07].
- *Koncepce trvalého uchování knihovních sbírek tradičních a elektronických dokumentů v knihovnách ČR do roku 2010*, 2005. [online]. Dostupné z: https://ipk.nkp.cz/odborne-cinnosti/docs/Koncepcetext.doc [cit. 2023-09-09].
- KOPSKÝ, Vojtěch, 2022. Výběr video formátů v paměťových institucích. In: *ProInflow*, **14**(1-2). Dostupné z: https://journals.phil.muni.cz/proinflow/article/view/2022-2-7 [cit. 2023-08-16].
- KUZNETSOV, Anton A, 2014. *An algorithm for MD5 single-block collision attack using high-performance computing cluster* [online]. Program Systems Institute of Russian Academy of Sciences. Dostupné z: https://eprint.iacr.org/2014/871 [cit. 2023-08-24].
- KVASNICA, Jaroslav, 2015. *Budoucnost českého webového archivu* [online]. Praha. Dostupné z: https://docplayer.cz/1001491-Budoucnost-ceskeho-weboveho-archivu.html [cit. 2023-08-21].

- KVAŠOVÁ, Zuzana, 2017. LTP úložiště NJ ČR a zkušenosti s jeho provozem. In: STRNISKO, Juraj, (ed.). *CDA 2017: Výmena skúseností z prevádzky a budovania LTP archívov: Zborník príspevkov z 2. mezinárodnej konferencie o dlhodobej archivácii.* Bratislava: Univerzitná knižnica v Bratislave, 2017, s. 22–28. ISBN 978-80-89303-57-1. ISSN 2453-9309.
- KVAŠOVÁ, Zuzana, 2018. Stav dlouhodobé archivace v NK ČR. In: TOMKOVÁ, Katarína, (ed.). *CDA 2018: Trvalá udržateľnosť a perspektívy ďalšieho rozvoja LTP archívov: Zborník príspevkov z 3. mezinárodnej konferencie o dlhodobej archivácii.* Bratislava: Univerzitná knižnica v Bratislave, 2018, s. 19–25. ISBN 978-80-89303-67-0. ISSN 2453-9309.
- LAWRENCE, Gregory W., KEHOE, William R., RIEGER, Oya Y., WALTERS, William H. a KENNEY, Anne R, 2000. *Risk management of digital information: a file format investigation.* Washington (DC): Council on Library and Information Resources, ISBN 18-873-3478-5. Dostupné z: https://www.clir.org/pubs/reports/pub93/pub93.pdf [cit. 2023-08-18].
- LHOTÁK, Martin, Zdeněk VAŠEK a Michal RŮŽIČKA, 2019. Projekt ARCLib – vývoj open-source řešení pro dlouhodobou archivaci digitálních dokumentů pro knihovny a další paměťové instituce – aktuální stav. In: TOMKOVÁ, Katarína (ed.). *CDA 2019: Nové trendy a východiská pri budování LTP archívov. Zborník prispevkov zo 4. mezinárodnej konferencie o dlhodobej archivácii.* Bratislava: Univerzitná knižnica v Bratislave, 2019, s. 53–63. ISBN 978-80-89303-76-2. ISSN 2644-6286.
- LIBRARY OF CONGRESS, 2017. *Sustainability Factors* [online]. Washington D.C.: Library of Congress. Dostupné z: https://www.loc.gov/preservation/digital/formats/sustain/sustain.shtml. [cit. 2023-09-11].
- LIBRARY OF CONGRESS, 2020. *PDF/A Family. PDF for Long-Term Preservation.* [online]. Washington D.C.: Library of Congress. Dostupné z: https://www.loc.gov/preservation/digital/formats/fdd/fdd000318.shtml [cit. 2023-09-29].

- LIBRARY OF CONGRESS, 2021. *Motion JPEG File Format* [online]. Washington D.C.: Library of Congress. Dostupné z: https://www.loc.gov/preservation/digital/formats/fdd/fdd000127.shtml [cit. 2023-09-29].
- LIBRARY OF CONGRESS, 2022a. *Bit Level Preservation and Long Term Usability* [online]. Washington D.C.: Library of Congress. Dostupné z: https://www.loc.gov/programs/digital-collections-management/digital-formats/bit-level-preservation-and-long-term-usability/ [cit. 2023-09-30].
- LIBRARY OF CONGRESS, 2022b. *Formats Descriptions.* [online]. Washington D.C.: Library of Congress. Dostupné z: https://www.loc.gov/preservation/digital/formats/fdd/descriptions.shtml [cit. 2023-09-29].
- LIBRARY OF CONGRESS, 2023a. *Formats, Evaluation Factors, and Relationships* [online]. Washington D.C.: Library of Congress. Dostupné z: https://www.loc.gov/preservation/digital/formats/intro/format_eval_rel.shtml. [cit. 2023-09-11].
- LIBRARY OF CONGRESS, 2023b. *Formats Descriptions. Browse alphabetical list* [online]. Washington D.C.: Library of Congress. Dostupné z: https://www.loc.gov/preservation/digital/formats/fdd/browse_list.shtml [cit. 2023-09-14].
- LIBRARY OF CONGRESS, 2023c. *Formats Descriptions. Mapping FDDs to PRONOM and Wikidata Unique Identifiers* [online]. Washington D.C.: Library of Congress. Dostupné z: https://www.loc.gov/preservation/digital/formats/fdd/fdd_puid_qid.shtml [cit. 2023-09-14].
- LIBRARY OF CONGRESS, 2023d. *Recommended Formats Statement.* [online]. Washington D.C.: Library of Congress. Dostupné z: https://www.loc.gov/preservation/resources/rfs/ [cit. 2023-09-29].
- LOCKSS, 2018a. *Networks.* In: LOCKSS Program [online]. Dostupné z: https://www.lockss.org/join-lockss/networks [cit. 2023-09-29].
- LOCKSS, 2018b. *Preservation Principles.* In: LOCKSS Program [online]. Dostupné z: https://www.lockss.org/about/preservation-principles [cit. 2023-09-29].

- *LTO-9: LTO Generation 9 Technology.* In: Ultrium LTO [online]. Dostupné z: https://www.lto.org/lto-9/ [cit. 2023-09-29].
- McKINNEY, Peter, Steve KNIGHT, Jay GATTUSO, David PEARSON, Libor COUFAL, David ANDERSON, Janet DELVE, Kevin De VORSEY, Ross SPENCER a Jan HUTAŘ, 2014. Reimagining the Format Model: Introducing the Work of the NSLA Digital Preservation Technical Registry. In: *New Review of Information Networking,* 19:2, s. 96–123, DOI: 10.1080/13614576.2014.972718.
- MCKINNEY, Peter a GATTUSO, Jay, 2014. Converting WordStar to HTML4. In: COATES, Serena et al. (eds.). *iPRES 2014 – Proceedings of the 11th International Conference on Digital*
- *Preservation*, 6-10. 10. 2014, Melbourne. Melbourne: State Library of Victoria, 2014, s. 149–159. ISBN 978-0-642-27881-4. Dostupné z: http://www.ipres-conference.org/ipres14/sites/default/files/upload/iPres-Proceedings-final.pdf
- MEGGYESI, Zoltán, 1994. *Fibre Channel Overview*. In: Fibre Channel Overview [online]. Dostupné z: https://hsi.web.cern.ch/fcs/spec/overview.htm [cit. 2023-08-25].
- MELICHAR, Marek a Jan HUTAŘ, 2013. České paměťové instituce a digitální data – historický exkurz, současný stav a předpokládaný vývoj I. In: *Duha: Informace o knihách a knihovnách* [online], **27**(4) ISSN 1804-4255. Dostupné z: https://duha.mzk.cz/clanky/ceske-pametove-instituce-digitalni-data-historicky-exkurz-soucasny-stav-predpokladany-vyvoj-2 [cit. 2023-09-07].
- MELICHAR, Marek a Jan HUTAŘ, 2014. České paměťové instituce a digitální data – historický exkurz, současný stav a předpokládaný vývoj II. In: *Duha: Informace o knihách a knihovnách* [online]. 2014, **28**(1) ISSN 1804-4255. Dostupné z: https://duha.mzk.cz/clanky/ceske-pametove-instituce-digitalni-data-historicky-exkurz-soucasny-stav-predpokladany-vyvoj-0 [cit. 2023-09-07].
- MELLOR, Chris, 2023. *50TB IBM tape drive more than doubles LTO-9 capacity* [online]. Dostupné z: https://blocksandfiles.com/2023/08/23/50tb-ibm-tape/ [cit. 2023-09-01].
- MILLER, Steven, 2022. *Metadata for Digital Collections.* 2nd ed. London: Facet Publishing, 505 s. ISBN 978-1-78330-616-9.

- MINISTERSTVO VNITRA ČR, 2022a. *Metodický návod č. 2/2022 odboru archivní správy a spisové služby Ministerstva vnitra pro akreditaci digitálního archivu,* 2022. Praha: Ministerstvo vnitra ČR.
- MINISTERSTVO VNITRA ČR, 2022b. *Metodický pokyn č. 4/2022 odboru archivní správy a spisové služby, kterým se vydávají Základní pravidla pro zpracování archiválií ver. 3.1.* [online]. Praha: Ministerstvo vnitra ČR. Dostupné z: https://www.mvcr.cz/clanek/metodiky.aspx?q=Y2hudW09Mw%3d%3d [cit. 2023-09-20].
- MINISTERSTVO VNITRA ČR, 2023. *Národní standard pro elektronické spisové služby* [online]. Praha: Ministerstvo vnitra ČR. Dostupné z: https://www.mvcr.cz/clanek/narodni-standard-pro-elektronicke-systemy-spisove-sluzby.aspx [cit. 2023-09-14].
- MIRANDA, Andrea, 2015. Důvěryhodná digitální úložiště, jejich audit a certifikace. *Knihovna: knihovnická revue,* **26**(2), s. 49-57. ISSN 1801-3252. Dostupné z: http://knihovnarevue.nkp.cz/aktualni-cislo/knihovny-a-informace/duveryhodna-digitalni-uloziste-jejich-audit-a-certifikace [cit. 2023-09-23].
- THE NATIONAL ARCHIVES, 2020a. *DROID: User Guide,* Richmond: The National Archives [online], 23 s. Dostupné z: https://cdn.nationalarchives.gov.uk/documents/information-management/droid-user-guide.pdf [cit. 2022-09-22].
- THE NATIONAL ARCHIVES, 2020b. *ZFO (Proof of delivery) File* [online]. Richmond: The National Archives, 23 s. Dostupné z: https://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=detailReport&id=2063 [cit. 2022-09-22].
- THE NATIONAL ARCHIVES, 2021a. *ISDOC Information System Document.* Richmond: The National Archives [online]. Dostupné z: https://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=detailReport&id=2395 [cit. 2023-09-14].
- THE NATIONAL ARCHIVES, 2021b. *ISDOCX Information System Document.* Richmond: The National Archives [online]. Dostupné z: https://www.nationalarchives.gov.uk/PRONOM/Format/proFormatSearch.aspx?status=detailReport&id=2396 [cit. 2023-09-14].

- NÁRODNÍ ARCHIV ČR, 2022a. *Národní standard formátů pro archivaci* [online]. Praha: Národní archiv ČR. Dostupný z: https://archi.gov.cz/_media/dokumenty:narodni_standard_formatu_pro_archivaci.pdf [cit.: 2022-09-22].
- NÁRODNÍ ARCHIV ČR, 2022b. *Národní archiv spolupracuje se světem* [online]. Praha: Národní archiv ČR. Dostupné z: https://www.nacr.cz/mezinarodni-spoluprace/narodni-archiv-spolupracuje-se-svetem [cit. 2023-09-15].
- NÁRODNÍ KNIHOVNA ČR, 2019. *Struktura čísla ISBN a jeho rozsah v České republice, přidělování identifikátorů vydavatele* [online]. Praha: Národní knihovna ČR. Dostupné z: https://text.nkp.cz/sluzby/sluzby-pro/isbn-ismn-issn/isbn/isbn-5 [cit. 2023-09-04].
- NÁRODNÍ KNIHOVNA ČR, 2021. *Standardy pro obrazová data.* In: Národní digitální knihovna [online]. Praha: Národní knihovna ČR, poslední aktualizace 27.10.2021. Dostupné z: https://standardy.ndk.cz/ndk/standardy-digitalizace/standardy-pro-obrazova-data [cit. 2023-06-23].
- NÁRODNÍ KNIHOVNA ČR, 2021. *Standardy pro zvuková data.* Národní digitální knihovna [online]. Praha: Národní knihovna ČR, poslední aktualizace 25.07.2023. Dostupné z: https://standardy.ndk.cz/ndk/standardy-digitalizace/standardy-pro-zvukova-data [cit. 2023-09-23].
- NÁRODNÍ KNIHOVNA ČR, 2021. *Standardy pro elektronické publikace.* Národní digitální knihovna [online]. Praha: Národní knihovna ČR, poslední aktualizace 25.07.2023. Dostupné z: https://standardy.ndk.cz/ndk/standardy-digitalizace/standardy-pro-elektronicke-dokumenty [cit. 2023-09-23].
- NDSA, 2019. *Using the Levels of Digital Preservation: An Overview for V2.0. In:* Open Science Framework, 2019. DOI: DOI10.17605/OSF.IO/QGZ98.
- NESTOR working group long-term preservation standards, 2009. *Catalogue of criteria for assessing the trustworthiness of PI systems. Draft for public comment* [online]. Göttingen: Niedersächsische Staats und Universitätsbibliothek. Dostupné z: https://d-nb.info/1047610442/34 [cit. 2023-08-20].

- NESTOR, 2014. *Leitfaden zur Erstellung einer institutionellen Policy zur digitalen Langzeitarchivierung* [online]. In: Nestor-Materialien 18. Dostupné z: http://nbn-resolving.de/urn:nbn:de:0008-2014052004 [cit. 2023-08-18].
- NEUROTH, Heike, OSSWALD, Achim. a SCHEFFEL, Regine (eds.), 2010. *Eine kleine Enzyklopädie der digitalen Langzeitarchivierung* [online]. Göttingen: nestor. Dostupné z: http://nbn-resolving.de/urn/resolver.pl?urn:nbn:de:0008-2010071949 [cit. 2023-06-23].
- NISO FRAMEWORK WORKING GROUP, 2007. *A framework of guidance for building good digital collections: a NISO recommended practice* [online]. 3rd ed. Baltimore (MD): National Information Standards Organization (NISO), iii, 95 s. ISBN 978-1-880124-74-1. Dostupné z: https://www.niso.org/sites/default/files/2017-08/framework3.pdf [cit. 2023-06-23].
- *NISO: Metadata for Images in XML Schema,* 2006. The Library of Congress [online]. Washington, D. C.: The Library of Congress. Dostupné z: https://groups.niso.org/higherlogic/ws/public/download/17937/ANSI-NISO%20Z39.87-2006%20%28R2017%29%2C%20Data%20Dictionary%20-%20Technical%20Metadata%20for%20Digital%20Still%20Images.pdf [cit. 2023-06-23].
- NOVAK, Audrey, 2006. *Fixity Checks: Checksums, Message Digests and Digital Signatures* [online]. Dostupné z: https://docplayer.net/9057558-Fixity-checks-checksums-message-digests-and-digital-signatures-audrey-novak-ilts-digital-preservation-committee-november-2006.html [cit. 2023-06-23].
- OAIS, 2012. *Reference Model for an Open Archival Information System (OAIS)* [online]. Magenta Book. Dostupné z: https://public.ccsds.org/pubs/650x0m2.pdf [cit. 2023-09-29].
- OPEN PRESERVATION FOUNDATION, 2019-20. *Digital Preservation Community Survey: Raw data* [online]. Leeds: Open Preservation Foundation, [2020]. Dostupné z: https://openpreservation.org/resources/surveys [cit. 2023-09-22].

- OPEN PRESERVATION FOUNDATION, 2022. *International Comparison of Recommended File Formats.* Dostupné z: https://openpreservation.org/resources/member-groups/international-comparison-of-recommended-file-formats/ [cit. 2023-09-29].
- ORACLE, 2010. *Detecting the MIME Type for a File* [online]. Oracle. Dostupné z: https://docs.oracle.com/cd/E19754-01/806-6878/6jfpqt2ts/index.html#:~:text=A%20file%20content%20sniffer%20associates,application%20can%20check%20the%20filename [cit. 2022-09-22].
- ORACLE, 2023. *What Is Block Storage?* In: Oracle [online]. Dostupné z: https://www.oracle.com/cz/cloud/storage/block-volumes/what-is-block-storage/ [cit. 27.08.2023].
- OSTRÁKOVÁ, Natalie, 2018. JPEG 2000 jako archivní formát obrazových dat. In: *Knihovna: knihovnická revue*, **29**(1), 5–26. ISSN 1801-3252.
- OSTRÁKOVÁ, Natalie a ŠÍR, Filip, 2017. Zvukové dokumenty ve fondech paměťových institucí v kontextu dlouhodobého uchovávání v ČR. Přípravná studie NK ČR k možnosti dlouhodobého uložení digitalizovaných dat. *Knihovna: knihovnická revue*, **28**(1), 5–19. ISSN 1801-3252.
- OSTRÁKOVÁ, Natálie a VOZÁR, Zdenko, 2019. Aktuální stav dlouhodobé archivace v Národní knihovně ČR: Zborník príspevkov zo 4. mezinárodnej konferencie o dlhodobej archivácii. In: TOMKOVÁ, Katarína (ed.). *CDA 2019: Nové trendy a východiská pri budování LTP archívov.* Bratislava: Univerzitná knižnica v Bratislave, 2019, s. 15–22. ISBN 978-80-89303-76-2. ISSN 2644-6286.
- OSTRÁKOVÁ, Natálie, KOČIŠOVÁ, Pavlína a BEŇAČKOVÁ, Miroslava, 2019. Vývoj standardu PREMIS a možnosti jeho dalšího využití ve standardech NDK. In: *ProInflow*, **11**(2). https://doi.org/10.5817/ProIn2019-2-6. Dostupné z: https://journals.phil.muni.cz/proinflow/article/view/2019-2-6/15505 [cit. 2023-09-29].
- OSTRÁKOVÁ, Natalie a KOPSKÝ, Vojtěch, 2020. Posuzování souborových formátů z hlediska dlouhodobého uchovávání a návrh metodiky pro Národní knihovnu České republiky. In: *Knihovna: knihovnická revue*, **31**(2), 83–105. ISSN 1801-3252.

- PALMER, William, MAY, Peter a CLIFF, Peter, 2013. *An Analysis of Contemporary JPEG2000 Codecs for Image Format Migration. British Library*. Dostupné z: https://purl.pt/24107/1/iPres2013_PDF/An%20Analysis%20of%20Contemporary%20JPEG2000%20Codecs%20for%20Image%20Format%20Migration.pdf [cit. 2023-08-16].
- PAVLÁSKOVÁ, Eliška, 2014. Techniky posuzování rizik a jejich využití v institucionálních repozitářích – užití v Digitálním repozitáři Univerzity Karlovy v Praze. In: *ProInflow: Časopis pro informační vědy* [online], **6**(1). ISSN 1804–2406. Dostupné z: http://www.phil.muni.cz/journals/index.php/proinflow/article/view/943 [cit. 2023-06-01].
- PAVLÁSKOVÁ, Eliška, 2017. From the Dissemination of Electronic Theses and Dissertations to Their Long-term Archiving. In: *10th Conference on Grey Literature and Repositories: proceedings* [online]. Prague: National Library of Technology, s. 33–44. ISSN 2336-5021. Dostupné z: http://nrgl.techlib.cz/conference/conference-proceedings [cit. 2023-09-29].
- PICHL, Marek, KŘIPAČ, Miroslav, BRANDEJSOVÁ, Jitka, ZEMANOVÁ, Růžena a BRANDEJS, Michal, 2015. *Metodika dlouhodobého ukládání a archivace digitálních dokumentů* [online]. Brno: Fakulta informatiky Masarykovy univerzity. ISBN 978-80-210-8113-0. Dostupné z: https://is.muni.cz/repo/1322181/Metodika_dlouhodobeho_ukladani_a_archivace_digitalnich_dokumentu.pdf [cit. 2023-09-25].
- *PNG* [online]. 2023, Poslední aktualizace 2023-06-26. Dostupné z: http://www.libpng.org/pub/png/ [cit. 2023-09-30].
- *PNG Specification (Third Edition)* [online]. 2023. Dostupné z: https://www.w3.org/TR/png/ [cit. 2023-09-30].
- POMERANTZ, Jeffrey, 2015. *Metadata.* Cambridge: MIT Press, 239 s. ISBN 978-0-262-52851-1.
- PREMIS EDITORIAL COMMITTEE, 2015. *PREMIS Data Dictionary for Preservation Metadata: version 3.0* [online]. Washington, D. C.: The Library of Congress. Dostupné z: https://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf [cit. 2023-09-29].

- PRENTICE, Will a GAUSTAD, Lars (ed.), 2017. *Uchování audiovizuálního dědictví: etika, principy a strategie uchovávání* [online]. IASA. Dostupné z: https://www.iasa-web.org/sites/default/files/downloads/publications/TC03_Czech.pdf [cit. 2023-09-29].
- RAFFO, Dave, 2019. What is iSCSI and How Does it Work? In: *TechTarget* [online]. Dostupné z: https://www.techtarget.com/searchstorage/definition/iSCSI [cit. 2023-09-30].
- Red Hat Customer Portal, 2023. *Red Hat Gluster Storage Life Cycle.* [online]. Dostupné z: https://access.redhat.com/support/policy/updates/rhs [cit. 2023-09-30].
- REESE, Terry a BANERJEE, Kyle, 2008. *Building Digital Libraries.* New York: Neal-Schuman Publishers, 275 s. ISBN 978-1-55570-617-3.
- RIMKUS, Kyle. PADILLA, Thomas. POPP, Tracy a Greer, Martin, 2014. Digital Preservation File Format Policies of ARL Member Libraries: An Analysis [online]. In: *D-Lib Magazine.* Dostupné z: https://www.dlib.org/dlib/march14/rimkus/03rimkus.html [cit. 2023-09-07].
- ROG, Judith, 2007. Compression and digital preservation: do they go together? [online]. In: *Archiving 2007 Final Program and Proceedings, Society for Imaging Science and Technology*. Dostupné z: https://library.imaging.org/admin/apis/public/api/ist/website/downloadArticle/archiving/4/1/art00020 [cit. 2023-09-30].
- ROSENTHAL, David, 2017. *SHA1 is dead* [online]. Dostupné z: https://blog.dshr.org/2017/03/sha1-is-dead.html [cit. 2023-09-30].
- ROSENTHAL, David S. H. et al., 2005. Requirements for Digital Preservation Systems: A Bottom-Up Approach. In: *D-Lib Magazine,* **11**(11). ISSN 1082-9873. DOI: 10.1045/november2005-rosenthal.
- RŮŽIČKA, Michal, Andrea MIRANDA, Lukáš HEJTMÁNEK, Zdeněk VAŠEK, Vlastimil KREJČÍŘ a Miloslav BARTOŠEK, 2019. *Metodika bitové ochrany digitálních dat* [online]. Praha: Knihovna AV ČR. Dostupné z: http://www.nusl.cz/ntk/nusl-393240 [cit. 2023-09-07].

- ŘEHÁNEK, Martin, 2012. *Identifikátory URN:NBN v prostředí českých knihoven a systém pro jejich správu* [online]. Brno. Diplomová práce. Masarykova univerzita, Fakulta informatiky. Vedoucí práce Miroslav BARTOŠEK. Dostupné z: https://is.muni.cz/th/u33ng/ [cit. 2023-09-30].
- SCHAEFER, Sibyl et al, 2018. *Digital Preservation Storage Criteria.* OSF. DOI: 10.17605/OSF.IO/SJC6U.
- SCOTT, Tamara, 2019. *Big Data Storage Wars: Ceph vs Gluster.* In: Perma.cc [online]. Dostupné z: https://perma.cc/2YY2-BBXG [cit. 2023-09-30].
- SHEINWALD, Dafna et al., 2002. *Internet Protocol Small Computer System Interface (iSCSI) Cyclic Redundancy Check (CRC)/Checksum Considerations.* Internet Engineering Task Force. DOI: 10.17487/RFC3385.
- SHENOY, Abhijith, [2020]. *The Pros and Cons of Erasure Coding & Replication vs. RAID in Next-Gen Storage Platforms.* In: Perma.cc [online]. Dostupné z: https://perma.cc/YFS5-KXKK [cit. 2023-09-30].
- SHIRKY, Clay, 2005. *Library of Congress Archive Ingest and Handling Test (AIHT) Final Report.* Dostupné z: https://www.digitalpreservation.gov/documents/ndiipp_aiht_final_report.pdf [cit. 2023-09-30].
- SNEYERS, Jon, 2022. *The Case for JPEG XL*. In: Cloudinary Dostupné z: https://cloudinary.com/blog/the-case-for-jpeg-xl [cit. 2023-09-11].
- SPEAKS, Scott, 2005. *Reliability and MTBF Overview* [online]. Vicor Reliability Engineering, [s.a.]. Dostupné z: https://www.vicorpower.com/documents/quality/Rel_MTBF.pdf [cit. 2023-09-11].
- SPENCER, Ross, 2022. Fractal in detail: What information is in a file format identification report?. In: *Code4Lib* [online], (**53**). ISSN ISSN 1940-5758. Dostupné z: https://journal.code4lib.org/articles/16351 [cit. 2023-09-22].
- De STEFANO, Paula et al., 2014. *Checking your digital content. What is fixity, and when should I be checking it?* [online]. NDSA. Dostupné z: http://hdl.loc.gov/loc.gdc/lcpub.2013655117.1 [cit. 2023-09-22].

- *Státní kulturní politika 2021–2025+* [online]. Praha: Ministerstvo kultury, 2021. ISBN 978-80-87546-41-3. Dostupné z: https://www.mkcr.cz/statni-kulturni-politika-cs-69 [cit. 2023-09-07].
- STIGGE, Martin et al., 2006. *Reversing CRC – Theory and Practice* [online]. Berlin: Humboldt-University. Dostupné z: https://sar.informatik.hu-berlin.de/research/publications/SAR-PR-2006-05/SAR-PR-2006-05_.pdf [cit. 2023-09-29].
- ŠURDA, Daniel, 2022. *Audit a certifikace digitálních úložišť v rámci archivnictví.* Bakalářská práce. Univerzita Karlova. Filozofická fakulta. Vedoucí práce Milan Vojáček. Dostupné z: https://dodo.is.cuni.cz/bitstream/handle/20.500.11956/178185/130341789.pdf?sequence=1&isAllowed=y [cit. 2023-09-29].
- SVATOŠ, Jonáš, 2022. *Popis poloprovozu vytváření a zpřístupnění digitálních archivních balíčků se zaměřením na umělecká díla s audiovizuálním obsahem nekinematografické povahy* [online]. Národní filmový archiv. Dostupné z: https://videoarchiv-nfa.cz/wp-content/uploads/2022/06/NFA_poloprovoz-vytvareni-a-zpristupneni.pdf [cit. 2023-09-29].
- TALLMAN, Nathan, 2021. A 21st Century Technical Infrastructure for Digital Preservation. In: *American Library Association*, s. 1–20. DOI: 10.6017/ital.v40i4.13355.
- TALLMAN, Nathan a WANG, Hannah, 2022. *Seeking sustanability. Developing a modern distributed digital preservation system.* Glasgow, Scotland. DOI: http://doi.org/10.7207/ipres2022-recordings [cit. 2023-09-29].
- THIBODEAU, Kenneth, 2002. *Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years* [online]. Council on Library and Information Resources. ISBN 1-887334-92-0. Dostupné z: https://www.clir.org/pubs/reports/pub107/thibodeau/ [cit. 2023-09-29].
- *TIFF. Revision 6.0*, 1992. [online] Adobe Systems Incorporated. Dostupné z: https://developer.adobe.com/content/dam/udp/en/open/standards/tiff/TIFF6.pdf [cit. 2023-09-30].
- TOMÁŠEK, Jan, 2018. *Design and implementation of Archival Storage component of OAIS Reference Model*. Diplomová práce. Brno: Masarykova univerzita.

- *Validation*, 2021. Community Owned digital Preservation Tool Registry (COPTR) [online]. Dostupné z: https://coptr.digipres.org/index.php/Validation 2021-20-04 [cit. 2023-09-30].
- *Validátor ZAF*, 2017 [online]. Dostupné z: https://validatorzaf.github.io/zaf/ [cit. 2023-09-14].
- VAN WIJNGAARDEN, Hilde, 2010. The seven year itch: Developing a next generation e-Depot at the KB. In: *World Library and Information Congress: 76th IFLA General Conference and Assembly* [online]. Gothenburg, Sweden: IFLA, 10–15 August 2010, s. 1–8. Dostupné z: https://www.ifla.org/past-wlic/2010/157-wijngaarden-en.pdf [cit. 2023-09-07].
- VAŠEK, Zdeněk, 2017. Standardizace Národní digitální knihovny. In: STRNISKO, Juraj, (ed.). *CDA 2017: Výmena skúseností z prevádzky a budovania LTP archívov: Zborník prispevkov z 2. mezinárodnej konferencie o dlhodobej archivácii.* Bratislava: Univerzitná knižnica v Bratislave, 2017, s. 113–123. ISBN 978-80-89303-57-1. ISSN 2453-9309.
- VAŠEK, Zdeněk, ŘEHÁNEK, Martin a CUBR, Ladislav, 2018. *Metodika pro přidělování a správu životního cyklu unikátních perzistentních identifikátorů digitálních dokumentů podle standardu URN:NBN, Verze 2.0.* In: Národní digitální knihovna [online]. Praha: Národní knihovna ČR. Dostupné z: https://standardy.ndk.cz/ndk/archivace/Certifik_metodika_urnnbn_2018.pdf [cit. 2023-08-17].
- VÄTTÖ, Kristian, 2015. *The Truth About SSD Data Retention.* In: AnandTech [online]. Dostupné z: https://www.anandtech.com/show/9248/the-truth-about-ssd-data-retention [cit. 2023-27-08].
- veraPDF CONSORTIUM, 2015-2023. [online]. Dostupné z: https://verapdf.org/ [cit. 2023-09-03].
- VOJÁČEK, Milan a KUNT, Miroslav, 2019. Stav a perspektivy digitálního archivnictví v ČR. In: *Paginae historiae: sborník Státního ústředního archivu v Praze. K životnímu jubileu PhDr. Evy Drašarové, CSc.* Praha: Národní archiv, s. 758–771. ISSN 1211-9768.

- Vzorový provozní řád archivu oprávněného k ukládání archiválií v digitální podobě (pouze vybraná ustanovení, 2012. In: *Věstník Ministerstva vnitra*, částka 65/2012. Dostupné z: http://www.mvcr.cz/soubor/65-vmv-pdf.aspx [cit. 2023-09-29].
- WATERS, Donald a GARRETT, John, 1996. *Preserving Digital Information. Report of the Task Force on Archiving of Digital Information.* In: Commission on Preservation and Access, Washington, DC. [online] 68 s. Dostupné z: https://eric.ed.gov/?id=ED395602 [cit. 2023-09-30].
- WEIL, Sage A. et al., 2007. *RADOS: a scalable, reliable storage service for petabyte-scale storage clusters.* Reno Nevada: ACM. ISBN 978-1-59593-899-2. DOI: 10.1145/1374596.1374606.
- WHEATLEY, Paul, 2022. *File format recommendations - I wouldn't say they are unacceptable, but I wouldn't recommend them either.* In: Digital Preservation Coalition. Dostupné z: https://www.dpconline.org/blog/file-format-recommendations [cit. 2023-09-14].
- WITTIG, Michael a WITTIG, Andreas, 2019. *Amazon web services in action.* Second edition. vyd. Shelter Island: Manning. ISBN 978-1-61729-511-9.
- *Zákon č. 257/2001 Sb. o knihovnách a podmínkách provozování veřejných knihovnických a informačních služeb (knihovní zákon)*, 2001. In: Sbírka zákonů, číslo 98. Dostupné také z: https://www.zakonyprolidi.cz/cs/2001-257 [cit. 2023-09-14].
- *Zákon č. 499/2004 Sb., o archivnictví a spisové službě v platném znění. In:* Zákony pro lidi. AION CS, © 2010–2023. Dostupné z: https://www.zakonyprolidi.cz/cs/2004-499 [cit. 2023-09-29].
- ZENG, Marcia a Jian QIN, 2016. *Metadata.* 2nd ed. Chicago: Neal-Schuman Publishers, 555 s. ISBN 978-1-55570-965-5.
- *Zřizovací listina Národní knihovny České republiky*, 2011. Praha: Ministerstvo kultury ČR, vydaná 30. listopadu 2011. Dostupné z: https://text.nkp.cz/soubory/ostatni/zrizovaci-listina-nk.pdf [cit. 2023-09-23].

Pavlína Kočišová — Zdeněk Vašek — Václav Jiroušek —
Vojtěch Kopský — Jan Bilwachs — Filip Pavčík — Petr Cajthaml

**Zachováno navěky? Teorie a praxe
dlouhodobého uchování digitálních dokumentů**

K vydání připravila Mgr. Pavlína Kočišová
Jazyková korektura a redakce: Mgr. Květa Fremrová
a Mgr. et Bc. Michaela Bežová
Grafická úprava a sazba: MgA. Kateřina Šuterová
Obrázek na obálce byl vygenerován neuronovou sítí DALL-E

Vydala Národní knihovna ČR
1. vydání
Praha 2023