

Posuzování souborových formátů z hlediska dlouhodobého uchovávání a návrh metodiky pro Národní knihovnu České republiky¹

File formats assessment from the digital preservation viewpoint and proposal of a methodology for the National Library of the Czech Republic

Mgr. Natalie Ostráková, Mgr. Vojtěch Kopský, Ph.D. / Oddělení pro standardy, Národní knihovna České republiky (Digital Preservation Standards Department, National Library of the Czech Republic), Klementinum 190, 110 00 Praha

Resumé:

Článek pojednává o posuzování vhodnosti souborových formátů pro dlouhodobé uchovávání. Prezentuje postupy v několika zahraničních institucích a v Národní knihovně ČR a směřuje k návrhu systému hodnocení pro Národní knihovnu ČR, tak, aby hodnocení formátů bylo co nejvíce objektivní, konzistentní.

Klíčová slova: souborové formáty, dlouhodobé uchovávání, digitální archivace, archivační formáty

Summary:

The article deals with the process of file formats assessments from the point of view of digital preservation. Practices of several foreign institutions and of the National Library of the Czech Republic are presented. The aim is to design a system for file format assessments for the NL CR so that the assessments are as objective and consistent as possible.

Keywords: file formats, digital preservation, archival formats

Úvod

Výběr vhodných archivačních formátů je jednou ze základních problematik oblasti digitální archivace dat. Volba vhodného formátu přímo ovlivní využitelnost digitálních dat v budoucnu. Nevhodně zvolený formát povede v nejhorším případě ke ztrátě uchovávaného obsahu, v lepším případě bude obsah zachráněn, ale obvykle to bude s vynaložením dalších finančních prostředků a lidské práce. Zvolený archivační formát sice není archivačním formátem „navždy“, je-li však vhodně zvolen, pak ochranné aktivity jsou účinné v delším časovém intervalu, a v případě, že je zvolen široce rozšířený formát, zde vždy bude existovat zájem formát udržet. V případě zastarání takového formátu bude hodně zájemců o jeho migraci do formátu jiného.

Při volbě vhodného archivačního formátu mohou v současnosti paměťové instituce vycházet z doporučení významných institucí, především knihoven a archivů. Tyto instituce provádějí vlastní hodnocení souborových formátů z hlediska digitální archivace. Výsledkem jejich aktivity bývá doporučení vhodných archivačních formátů pro různé typy dat. Jedním z nejvýznamnějších zdrojů pro volbu archivačních formátů je každoročně

1 Článek vznikl na základě institucionální podpory dlouhodobého koncepčního rozvoje výzkumné organizace poskytované Ministerstvem kultury ČR.

aktualizované doporučení Kongresové knihovny (*Recommended Formats Statements, RFS*), které knihovna vydává od roku 2014. Toto doporučení obsahuje seznam vhodných archivačních formátů pro digitální i fyzické objekty – např. pro digitální textové materiály, digitální fotografie, digitalizované zvukové dokumenty, databáze apod.

V případě, že instituce volí nový archivační formát, může tato existující doporučení plně následovat (tj. přímo na jejich základě formát vybrat) nebo z nich může vycházet a považovat je za jakýsi „předvýběr“ vhodných formátů, na kterých pak provede svoje vlastní zhodnocení, případně instituce může provést kompletně své vlastní hodnocení nezávisle na existujících doporučeních. Instituce pravidelně hodnotí nejen zvažované nové formáty, ale i formáty, které již spravují, s cílem zjistit aktuální míru jejich rizika. Z toho důvodu je vhodné zabývat se i samotným procesem posuzování formátů a nastavit jej v institucích tak, aby bylo posuzování co nejvíce konzistentní a co nejvíce objektivní. Některé instituce si vytvořily proces či metodiku posuzování formátů, jejich přístupy budou ilustrovány níže.

Cílem této studie je prověřit praxi jiných institucí a vybrat postupy, které lze aplikovat při posuzování formátů v Národní knihovně ČR (dále také NK ČR). V závěru textu bude prezentován návrh postupu při posuzování formátů, který bude výhledově součástí interní metodiky pro hodnocení souborových formátů v NK ČR.

1 Kritéria pro výběr a hodnocení formátu

Problematika hodnocení souborových formátů digitálních dokumentů z hlediska jejich dlouhodobého uchovávání je významným tématem digitální archivace. Věnují se mu odborné texty a prakticky i jednotlivé instituce uchovávající obsah. Výběru vhodného archivačního formátu předchází stanovení vlastností, které by takový formát měl mít. Pro tyto účely se stanovují hodnotící kritéria, případně i význam (váha) jednotlivých kritérií. Předtím, než budou postupy vybraných institucí podrobněji zkoumány, uvádíme pro představu kritéria, která se objevují ve vybraných informačních zdrojích.

Významný informační pramen pro oblast digitální archivace, online publikace *Digital Preservation Handbook*, zpracovaná organizací *Digital Preservation Coalition* doporučuje u souborových formátů zvažovaných pro archivační účely sledovat: dostupnost a standardizaci specifikace formátu, rozšíření formátu v uživatelské komunitě, vlastnictví formátu (proprietární vs. neproprietární), ztrátovost/bezeztrátovost, podporu metadat a schopnost zachytit významné vlastnosti předlohy (*Digital Preservation Coalition c2015*).

Definici kritérií pro hodnocení souborových formátů se v roce 2008 zabývaly i autorky Rog a Wijk z nizozemské Královské knihovny. Za důležitá kritéria považují: otevřenost formátu (resp. dostupnost informací o něm), rozšíření formátu, komplexnost, mechanismy technické ochrany, autodokumentace formátu, robustnost, závislost na softwaru a hardwaru (Rog a Wijk 2008).

Výběrem vhodného formátu pro archivaci digitálních dat se v roce 2008 zabýval také Brown z Národních archivů Velké Británie (*The National Archives*). Producentům dat doporučuje při výběru souborového formátu sledovat tato kritéria: rozšířenost formátu mezi uživateli (*ubiquity*), softwarová podpora, míra zveřejnění formátu (*disclosure*), kvalita dokumentace, stabilita², snadnost identifikace formátu, související práva duševního vlastnictví, podpora metadat, podpora vlastností a funkcionalit odpovídajícího typu dat

2 Tj. aby specifikace formátu nebyla předmětem častých změn a aby nové verze formátu byly vždy zpětně kompatibilní.

(komplexita formátu), interoperabilita (tj. závislost na externím HW a SW), životaschopnost (tj. robustnost formátu), opakovaná použitelnost³ (Brown 2008).

Z výše uvedených výčtů je například patrné, že důležitými kritérii pro posuzování formátů jsou míra rozšířenosti formátu mezi uživateli a stav a dostupnost dokumentace. Tato kritéria uvádějí všechny výše uvedené zdroje.

1.1 Posuzování souborových formátů ve vybraných institucích

Jak bylo výše uvedeno, některé paměťové instituce provádějí pro své potřeby vlastní hodnocení souborových formátů. V uplynulém období byly zkoumány postupy hodnocení u významných zahraničních institucí. Vybrány byly instituce, o nichž je známo, že řeší úkoly v oblasti dlouhodobého uchování⁴, jasně definují postup hodnocení formátů a tento postup je veřejně dostupný. V následujících částech budou představeny postupy vybraných pěti institucí, uvedena kritéria a indikátory naplnění těchto kritérií a případně váhy jednotlivých kritérií, pokud s nimi instituce pracuje. Pro ilustraci bude u některých institucí uvedeno, jak vybrané formáty ohodnotily.

1.1.1 Kongresová knihovna

Jednou z nejvýznamnějších paměťových institucí nejen v oblasti dlouhodobé archivace je Kongresová knihovna (dále též jako LOC). Podílí se na vývoji některých otevřených formátů, na vývoji metadatových schémat archivace, digitalizuje a uchovává informační zdroje, sbírá informace o souborových formátech a vytváří formátový registr.⁵ Pravidelně také sleduje stav formátů vhodných pro archivační účely. Pro hodnocení formátů Kongresová knihovna definovala tzv. faktory udržitelnosti (orig. *Sustainability factors*), tj. vlastnosti, které jsou z jejího pohledu pro archivační formáty (bez ohledu na typ dat) důležité. U formátu tedy sleduje (Library of Congress 2017b):

- A. míru „zveřejnění“ formátu (angl. *disclosure*) – je ověřena existence a dostupnost kompletní specifikace a dostupnost validačních nástrojů
- B. rozšíření, užívanost formátu (angl. *adoption*) mezi uživateli, profesionály; LOC definuje i indikátory dostatečné rozšířenosti formátu, jimiž jsou:
 - nástroje pro práci s formátem jsou součástí softwarových balíčků pro PC
 - nativní podpora formátu v internetových prohlížečích
 - nativní podpora formátu na trhu významných softwarových nástrojů
 - existence softwarových nástrojů od více konkurenčních firem
 - formát figuruje v doporučeních jiných paměťových institucí.

3 Dle Brownova komentáře k tomuto kritériu (s. 8) je možné si tuto vlastnost vysvětlit zřejmě jako zachování funkcionality souboru. Jako příklad Brown uvádí soubor z tabulkového procesoru (např. Excel), který je konvertován do PDF, čímž ztratí vlastnost, resp. schopnost být zpracováván a upravován.

4 Kongresová knihovna každoročně vydává doporučení (RFS) týkající se souborových formátů a spravuje a vyvíjí metadatová schémata používaná v archivačních balíčcích (METS, PREMIS, MIX...). Knihovna Harvardovy univerzity vyvíjí validační a charakterizační nástroj FITS a v minulosti vyvíjela nástroj JHOVE.

5 Zde mimo jiné shromažďuje informace o formátech, které jsou z jejího hlediska preferované či akceptovatelné. V současnosti registr obsahuje cca 600 záznamů, jedná se o popisy souborových formátů, jejich verzí, tříd formátů, bitových sekvencí, kompresí apod.

- C. transparentnost formátu (*transparency*), kde je posuzováno, nakolik je formát analyzovatelný dostupnými nástroji a nakolik je formát jednoduchý. Transparentní formát je jednodušší konvertovat do jiných formátů.
- D. autodokumentace (*self-documentation*), kde je hodnoceno, zda formát umožňuje uložení metadat.
- E. závislost na externím prostředí (*external dependencies*), tj. zda je formát závislý na konkrétním hardwaru a softwaru.
- F. právní situace formátu (*impact of patents*), tj. licence a patenty a související možná omezení užití formátu.
- G. mechanismy technické ochrany (*technical protection mechanisms*), resp. jde o posouzení, zda případné budoucí migraci formátu budou tyto mechanismy bránit, tj. soubor nebude chráněn proti změně ani přístupu.

Vedle výše uvedených kritérií, která sledují obecně vhodnost formátu z pohledu digitální archivace a která jsou společná pro všechny formáty nezávisle na typu dat, sleduje Kongresová knihovna při hodnocení formátů i vlastnosti specifické pro konkrétní typ dat, jež mohou mít dopad na kvalitu a funkcionality konkrétního uchovávaného obsahu. Tato kritéria označuje jako faktory kvality a funkcionality (*Quality and functionality factors*); zjišťují, nakolik formát dokáže uchovat vlastnosti a funkce objektu. U obrazových dat se tak sleduje, zda hodnocený formát umožňuje základní zobrazení obsahu, zda podporuje vysoké rozlišení, správu barev, grafické efekty a typografii, multispektrální pásma (*multispectral bands*), a další funkce nad rámec běžného zobrazení (Library of Congress 2017a).

Posuzování formátů v Kongresové knihovně neužívá kvantitativně vyjádřenou významnost kritérií, tj. předem se neurčují význam a tedy váha jednotlivých vlastností. Zda formát bude vyhodnocen jako preferovaný či akceptovatelný, je dáno spíše rovnováhou mezi jednotlivými kritérii, tedy že formát významně nezaostává v některé z důležitých vlastností. Případně, je-li předloha specifická a vyžaduje podporu nějaké funkcionality, může mít podpora konkrétní funkcionality u formátu vyšší váhu, než některý z faktorů udržitelnosti (Library of Congress 2017a).

1.1.2 Hodnocení formátů v knihovně Harvardovy univerzity

Kritéria Kongresové knihovny užívá i hodnotící systém Harvardovy univerzity.

Knihovna Harvardovy univerzity provozuje vlastní úložiště digitálních dat a před přijetím nového formátu, případně nového typu dat, provádí hodnocení souborových formátů za účelem stanovení preferovaných a akceptovatelných formátů.

Pro hodnocení vlastností formátů vytvořili hodnotící matici ve formě excelové tabulky s třiceti devíti kritérii a dílčími indikátory (Goethals 2016b), které mají určit, do jaké míry je kritérium splněno. Míra naplnění kritéria se nevyjadřuje numericky (srov. KOST-CECO a NARA), ale sledovaná vlastnost je v matici slovně popsána nebo vyjádřena výrazy z kontrolovaného slovníku. Zda je vlastnost hodnocena pozitivně či negativně, je vizuálně znázorněno podbarvením buňky v tabulce, kladná vlastnost je podbarvena zeleně, negativní vlastnost červeně a oranžově stav mezi nimi (viz obr. 1).

Každý indikátor má doporučenou váhu při hodnocení, hodnotitel může ale také navrhnout vlastní váhu kritérií.

Codec Format	Cost to maintain environment for access or processing	Dependency on a single organization or company	Ease of accurate identification	Format expertise available
JPEG 2000 (Lossless) (JP2)	High	No	Valid files are recognized by a JP2 signature box (a fixed-length 12-byte string which shall have the value: 0x0000 000C 6A50 2020 0D0A 870A) at start of file.	Medium
TIFF/EP	Creating and managing TIFF files is currently low-cost and well-supported.	No	Valid files are recognized by initial byte-order "MM" or "II" flag and following magic number "42" or higher. Difficult to differentiate between TIFF versions.	High

Obr. 1 Barevně vyjádřené naplnění kritéria v harvardské hodnoticí matici – výřez z originální tabulky (zdroj: Harvard, 2016)

Jak bylo výše uvedeno, hodnoticí kritéria vycházejí z kritérií navržených Kongresovou knihovnou a jejich obsah je víceméně shodný, proto nejsou samotná kritéria v následujícím výčtu blíže popsána, popis uvádíme v částech, kde se kritéria od Kongresové knihovny liší nebo kde oproti Kongresové knihovně sledují, dle našeho názoru, vlastnost navíc, kterou Kongresová knihovna explicitně neuvádí apod. Pro ilustraci je uvedeno, co bylo během hodnocení zaznamenáno k formátu JPEG 2000 (dále také zkráceně JP2).

Knihovna Harvardovy univerzity používá následující hodnoticí kritéria (Goethals 2016b):

- Míra „zveřejnění“ formátu
K formátu JPEG 2000 v knihovně Harvardu zaznamenali, že se jedná o standard ISO, je dostupný, dokumentace formátu je kompletní a srozumitelná.
- Rozšíření, užívání formátu
Formát JPEG 2000 byl popsán jako nezávislý na jedné instituci nebo firmě, málo rozšířený mezi uživateli, středně rozšířený mezi profesionály, agenturou NARA (viz. dále) označený za akceptovatelný a celosvětově rozšířený.
- Transparentnost formátu
Formát JPEG 2000 byl označen za vysoce komplexní, s rozpoznáním kompresním algoritmem (JPEG 2000), je možné jej identifikovat dle bitové sekvence na začátku souboru.
- Autodokumentace
Knihovna Harvardovy univerzity sleduje v rámci tohoto kritéria i to, zda je formát odolný proti chybám.
V případě formátu JPEG 2000 bylo uvedeno, že podporuje uložení metadat, a krátce popsáno, jakým způsobem. Formát byl také označen za odolný proti chybám a bylo uvedeno krátké vysvětlení, proč tomu tak je.
- Závislost na externím prostředí
Formát JPEG 2000 byl popsán jako nezávislý na konkrétním hardwaru

a softwaru, s uspokojivým množstvím a dostupností potřebného softwaru a s omezenou podporou v internetových prohlížečích.

- Související práva duševního vlastnictví (licence a patenty)
Licence pro první část standardu JPEG 2000, jež se v archivech využívá, je bezplatná.
- Mechanismy technické ochrany
Formát JPEG 2000 nepodporuje mechanismy technické ochrany obsahu, umožňuje ale do metadat zapsat údaje o vlastnických právech k obsahu.
- Kvalita a funkcionalita
Pro formát JPEG 2000 jsou v tabulce vyplněny požadované hodnoty, tj. podporuje 48 bitů na barevný kanál, podporuje 16 kanálů, podporované rozlišení není nijak omezeno, podporuje skutečně bezeztrátovou kompresi.
- Finanční náklady
Jedná se o vlastní kritérium Harvardovy univerzity, Kongresová knihovna toto kritérium v rámci hodnocení vhodnosti formátu pro archivační účely explicitně nesleduje.⁶ Hodnotitel odhaduje finanční náročnost správy formátu a potřebného prostředí pro zpřístupňování a zpracování, náklady na uložení formátu v nekomprimované podobě a při maximální kompresi souborů. Správa souvisejícího workflow u formátu JPEG 2000 byla označena za finančně nákladnou.

Hodnocení knihovny Harvardovy univerzity, na rozdíl od Kongresové knihovny, doporučuje i váhu pro jednotlivé indikátory. Nejvyšší váhu mají dle doporučení univerzitní knihovny následující kritéria a indikátory:

1. Finanční náročnost implementace formátu, správy prostředí pro zpřístupnění a zpracování
2. Závislost formátu na jedné organizaci nebo firmě
3. Snadnost a správnost validace formátu
4. Náročnost implementace formátu, existence zkušenosti s formátem a souvisejících hlubších znalostí formátu
5. Právní situace formátu
6. Závislost na konkrétním HW/SW
7. Počet a kvalita prezentačního softwaru
8. Dostupnost specifikace
9. Rozšíření mezi uživateli
10. Rozšíření mezi profesionály
11. Míra podpory barevných prostorů
12. Počet bitů na kanál
13. Počet podporovaných kanálů
14. Bitová hloubka
15. Maximální možné rozlišení
16. Podpora skutečné bezeztrátové komprese
17. Podvzorkování⁷ barvonosných složek
18. Kompresní poměr

6 Nelze ovšem tvrdit, že toto kritérium nesleduje vůbec. Při volbě formátu pro svoje vlastní potřeby se může mezi vhodnými archivačními formáty dále rozhodnout dle finanční náročnosti jejich implementace apod.

7 Redukce informace o barevných složkách ve skupinách pixelů rastrového obrázku při zachování informace o jas provedená po převedení do barevného prostoru YCbCr. Dochází tedy ke ztrátě informací, tato ztráta však není lidským okem pozorovatelná. (Podvzorkování barvonosných složek 2016).

Nízkou váhu mají například tato kritéria: užití formátu pro archivační účely u jiné instituce, podpora popisných a technických metadat, odolnost proti chybám, geografické rozšíření, datum poslední revize specifikace, periodičita revize specifikace formátu.

Pro formáty, které projdou tímto hodnocením jako preferované a akceptovatelné, jsou vytvořeny profily. Každý profil obsahuje základní informace o formátu, seznam existujících nástrojů, identifikaci subjektu, který formát vytvořil, a subjektu, který jej nadále spravuje, odkazy na specifikaci a další významné zdroje informací, informace o právní situaci formátu, seznam známých rizik apod. (Goethal 2016a).

1.1.3 Hodnocení formátů podle NARA

Americká instituce NARA (*National Archives and Record Administration*, česky Národní úřad pro archivaci a dokumentaci) provádí hodnocení formátů, jež aktuálně spravuje, a formátů, které se spravovat chystá. Výsledkem jejich hodnocení je číselné skóre vypočítané z bodového ohodnocení kritérií a vah/důležitostí těchto kritérií (podobně jako u hodnocení KOST-CECO viz dále). Na základě bodového výsledku NARA určí, nakolik je formát rizikový.

Stejně jako knihovna Harvardovy univerzity, vychází NARA z kritérií Kongresové knihovny. Navíc sleduje i stáří formátu (včetně jeho revizí a aktualizací). Specifickým kritériem je pak kritérium sledující možnosti konverze do jiného formátu.

Jednotlivá kritéria, resp. jejich indikátory, jsou formulována jako otázky, na které hodnotitel odpovídá. Odpovědi se bodují, výše bodů je odvislá od váhy kritéria, rozsah bodů je od 2 do -4, kde 2 má nejvyšší pozitivní váhu a -4 má nejvyšší negativní váhu (s výjimkou posledního kritéria). Hodnocení NARA rozlišuje pozitivní a negativní váhu kritéria, a to podle dopadu vlastnosti formátu na instituci. Nejvyšší pozitivní váhu má například široké rozšíření formátu, existence a dostupnost specifikace, žádná nebo minimální závislost na softwaru. Nejvyšší negativní váhu má pokročilé stáří formátu a jeho závislost na konkrétním hardwaru a softwaru. Obsah kritérií je víceméně srovnatelný s tím, jaký jim přisuzuje Kongresová knihovna, pro ilustraci uvádíme ke každému z kritérií výběr otázek, které hodnotící metodika předepisuje.

NARA sleduje tato kritéria při hodnocení:

- Míra „zveřejnění“ formátu
Je formát proprietární?
Existují nástroje, které umí validovat formát oproti specifikaci?
- Rozšíření, užívání formátu
Je formát běžně používán při vytváření, správě a uchovávání záznamů v institucích federální vlády?
Je formát aktivně spravován a aktualizován nějakou organizací, jedincem nebo komunitou?
- Transparentnost formátu
Je specifikace formátu dostatečně detailní, aby na jejím základě byla možná základní analýza formátu s pomocí nástrojů jako je „hex editor“⁸ apod.?
Je zdrojový kód nástroje, jenž se používá pro vytváření souborů, dostupný zdarma nebo za malý poplatek?
- Autodokumentace
Podporuje formát ukládání popisných metadat?
Jsou vložena metadata v souladu s mezinárodními standardy?

8 Hexadecimální prohlížeče/editory, které zobrazí počítačové soubory v binární podobě, tj. číselnými hodnotami.

- Závislost na externím prostředí
Je pro přesun obsahu pod správu NARA potřebný speciální přehrávač (př. Blu-Ray, Audio CD)?
Jsou pro zobrazení formátu potřeba pluginy nebo skripty?
- Související práva duševního vlastnictví (licence a patenty)
Podléhá formát patentovým nárokům, které mohou bránit vývoji open source nástrojů pro práci se soubory?
Umožňují licenční podmínky formátu využití v open source?
- Mechanismy technické ochrany
Podporuje formát šifrování části nebo celého obsahu souboru?
Podporuje formát vkládání informací, například vložení vodoznaku?
- Stáří formátu
Kdy byla vytvořena specifikace formátu?
Kdy byl formát naposledy aktualizován?
- Možnosti konverze do jiného formátu⁹
 Hodnotitel dále posuzuje možnosti konverze formátu, resp. potřebnost takové konverze. Jedním bodovým skóre má vyjádřit, zda existují nástroje pro konverzi, které obsah nezmění, a NARA tak dokáže provést akceptovatelnou migraci. Pokud formát takovou akci nepotřebuje, tj. není nutné jej migrovat, je uděleno nejvyšší ohodnocení, tj. 5 bodů; pokud takové nástroje vůbec neexistují, získá formát nejmenší možné ohodnocení, tj. -5 bodů, pokud nástroje existují, ale NARA je nemá, je ohodnocení -3 body, a pokud již byla migrace v NARA provedena, jsou přiděleny 3 body.

Pro určení rizikovosti držení formátu hodnotitel při tomto posuzování bere v potaz i to, kolik souborů v tomto formátu se v NARA vyskytuje. Čím více takových souborů NARA spravuje, tím více záporných bodů dostane a tím vyšší je i priorita pro ochranné aktivity nad tímto formátem (Johnston 2019).

1.1.4 Hodnocení dle KOST-CECO

Švýcarské Koordinační centrum pro dlouhodobou archivaci elektronických dokumentů Švýcarského federálního archivu (dále KOST-CECO) si klade za cíl poskytovat podporu archivům při archivaci digitálních dokumentů. Mimo jiné se podílí i na specifikaci archivační verze formátu TIFF a je autorem minimálně dvou nástrojů digitální archivace: KOST-Val a KOST-Simy. Skupina vypracovala postup pro hodnocení formátů, stanovila kritéria, jejich bodové ohodnocení a numerickou váhu jednotlivých kritérií. Tento postup v praxi aplikují a výsledkem je katalog padesáti dvou důležitých a rozšířených formátů s hodnocením jejich vhodnosti pro digitální archivaci. Katalog obsahuje formáty pro tyto typy dat: text, obraz, audio, video, tabulková zobrazení (spreadsheet), databáze a hypertext, data GIS a CAD/CAM.¹⁰

Hodnotících kritérií je dvanáct a míra, do jaké formát určité kritérium naplňuje, se stanoví bodovým ohodnocením (jedním až čtyřmi body). Přidělené body se následně vynásobí vahou kritéria (1 nebo 0,5) a sečtou se. Výsledkem je bodové ohodnocení formátu, podle něž jsou v katalogu formáty zařazeny mezi archivační formáty vhodné,

9 V dokumentu NARA uvedeno jako „feasibility“, v kontextu dokumentu možno přeložit také jako proveditelnost konverze.

10 Formát pro digitální prototypy, simulace, 3D objekty.

nevhodné a akceptovatelné za určitých podmínek (Koordinationsstelle für die Daueshafte Archivierung elektronischer Unterlagen 2019b).

V KOST-CECO sledují tato kritéria při hodnocení (pro ilustraci doplněno o stav u JP2 a TIFF) (Koordinationsstelle für die Daueshafte Archivierung elektronischer Unterlagen 2019b):

- Otevřenost

Kritérium sleduje stav specifikace formátu. Specifikace může být standardizovaná a publikovaná standardizačním institutem nebo být široce uznávaným doporučením (např. RFC) (nejvyšší bodové ohodnocení, tj. 4 body), může být publikována vlastníkem práv (např. PDF, PSD; 3 body), případně je dokumentace formátu výsledkem zpětné analýzy formátu¹¹ (2 body), a nebo specifikace není veřejně dostupná (1 bod). Nevyžaduje se, aby specifikace byla dostupná zdarma.

Formát JPEG 2000 získal za toto kritérium 4 body, protože se jedná o standard ISO, formát TIFF získal 3 body, protože se jedná o volně dostupnou specifikaci, vydanou vlastníkem práv.

- Práva duševního vlastnictví

Kritérium sleduje, zda je použití formátu nebo jeho částí (kódování, komprese) omezeno licencí či jinými právními nároky. Ideálním stavem je, když je formát či jeho součásti dostupný pod volnou licencí (volná licence/užití, např. Creative Commons, GPL). Tento stav má garantovat, že se licenční stav formátu nezmění neočekávaně. Přijatelný je i stav, kdy licence neexistuje nebo již expirovala. Z hlediska dlouhodobého uchovávání je nevhodné, je-li formát chráněn licencí a za jeho využití se platí. Formát JPEG 2000 získal 3 body. Patenty na standard JPEG 2000 existují, ale dle KOST-CECO se vlastníci patentu zavázali, že standard může být užíván bezplatně. Dle jejich názoru se však nedá vyloučit, že se nějaké nároky ještě nevyskytnou. Formát TIFF získal také 3 body, protože autorská práva na formát patří společnosti Adobe, ta platbu za licence dosud nepožaduje.

- Prevalence (distribuce, rozšíření)

Z pohledu dlouhodobého uchovávání je ideálním stavem široké rozšíření formátu, jež zároveň znamená, že:

1. je možné očekávat dlouhý migrační cyklus, tj. je dost času na případnou migraci formátu v budoucnu
2. existence velkého množství souborů v tomto formátu znamená, že bude zájem udržovat a vytvářet migrační nástroje pro tento formát
3. je malá pravděpodobnost náhlé ztráty podpory formátu v aplikacích
4. existuje dostatečné množství aplikací a nástrojů pro práci s formátem

Dle názoru skupiny KOST-CECO není formát JPEG 2000 mimo paměťové instituce zatím příliš užíván, proto byl tento formát ohodnocen třemi body. Očekávají však nárůst popularity. Formát TIFF byl ohodnocen čtyřmi body.

- Funkcionalita formátu

Toto kritérium sleduje, zda hodnocený formát dokáže uchovat a reprezentovat významné vlastnosti, jež jsou pro konkrétní typ dat klíčové, tj. zda nedochází ke ztrátě významných vlastností při migraci do tohoto formátu. Minimálním požadavkem je, že formát musí podporovat obvyklé funkcionality své kategorie, které jsou nezbytné pro archivaci.

V KOST-CECO ohodnotili formáty JPEG 2000 i TIFF stejně čtyřmi body, oba dokáží zachytit a uchovat významné vlastnosti obrazových dat.

¹¹ Není to plně legální postup, nicméně někdy nelze zajistit „archivovatelnost“ dat jinak (např. u zastaralých formátů apod.).

- Implementace

Z hlediska dlouhodobé archivace je optimální, existuje-li více nezávislých implementací formátu (tj. nástrojů pro tvorbu, editaci, prohlížení, konverzi...) pocházejících od různých tvůrců (indikují, že je možné formát bez problémů implementovat) a alespoň jeden nástroj je dostupný volně jako open source.

Formáty JPEG 2000 i TIFF získaly za toto kritérium 4 body. Pro oba formáty existuje několik implementací včetně open source. Formát JPEG 2000 je nativně podporován jen v některých aplikacích, pro některé existuje plugin a některé programy jej zpracovat zatím neumí.

- Hustota uložení, zápisu

Skupina KOST-CECO sleduje i hustotu uložení, zápisu dat, přičemž ideálním stavem z jejich pohledu je vysoká hustota uložení.¹² U jednotlivých formátů se pak sleduje, zda data komprimují a kolik úložného místa se tím ušetří.

Formát JPEG 2000 byl ohodnocen třemi body, protože podporuje vysokou kompresi, ta se však v paměťových institucích nevyužívá. Formát TIFF získal za toto kritérium dva body, protože se v paměťových institucích používá jen v nekomprimované nebo beze-ztrátově komprimované podobě.

- Možnost kontroly formátu (ověřitelnost formátu)¹³

Důležitým krokem při dlouhodobém uchovávání souborů je validace souborových formátů, tj. kontrola, zda formát odpovídá normě, standardu. Ideálním stavem pro dlouhodobou archivaci pak je, existuje-li více nezávislých validátorů formátu (4 body).

Formáty JPEG 2000 a TIFF získaly 4 body, pro oba existuje více validačních nástrojů (jhove, jpylyzer, DPF Manager apod.).

- Komplikovanost, složitost formátu

Kritérium je v pojetí KOST-CECO aplikovatelné zatím pouze na videoformáty. Ideálním stavem je nízká složitost formátu, tj. jednoduchá datová struktura bez komprese. Takový formát je pak dle KOST-CECO snazší konvertovat do jiného formátu.

- Autodokumentace

Toto kritérium KOST-CECO aktuálně aplikuje taktéž pouze na videoformáty, tj. pro ostatní formáty zřejmě nepovažuje tuto vlastnost za důležitou (viz dále). Kritérium sleduje, zda formát může nést metadata a zda je možné tato metadata extrahovat softwarovými nástroji.

- Osvědčená praxe

Toto kritérium sleduje stav užití formátu v archivech a zkušenosti, jaké s ním archivy při správě mají. Kritérium má oproti ostatním kritériím nižší váhu, protože posouzení je subjektivní, nelze jej plně expertně podložit.

Formáty JPEG 2000 i TIFF získaly shodně tři body za toto kritérium. Oba formáty jsou archivy využívány, vyskytují se v doporučeních. Formát TIFF je využíván častěji, ale ne všechny jeho podoby jsou vhodné pro archivaci, takové by měly být konvertovány.

- Perspektiva formátu

Skupina KOST-CECO posuzuje u formátu také jeho vyhlídky do budoucna. Z hlediska dlouhodobého uchovávání jsou ideální formáty, u nichž existuje pravděpodobnost, že budou v budoucnu dále používány. Kritérium má oproti ostatním kritériím také nižší váhu, protože posouzení tohoto kritéria je subjektivní, jedná se spíše o odhad.

12 Více viz oddíl 1.1.6.

13 Kritéria *Možnost kontroly* a *Implementace* jsou zřejmě nejproměnlivějšími sledovanými vlastnostmi formátu a situace se pro formát může změnit i v relativně krátké době (rok, dva roky apod.). Z toho důvodu je vhodné aktualizovat celkové hodnocení formátu častěji, resp. vybrané vlastnosti monitorovat v kratším časovém intervalu než vlastnosti jiné.

Formát JPEG 2000 byl za toto kritérium ohodnocen čtyřmi body, dle KOST-CECO má velký potenciál. Formát TIFF získal dva body, vznik a používání nových, modernějších formátů bude dle KOST-CECO mít za následek ztrátu monopolního postavení formátu TIFF.

- Formátová třída

Skupina KOST-CECO třídí formáty do šesti tříd podle toho, jak jsou formáty v komunitě používány, známy a co se od nich očekává.

Třídy:

A známý formát s širokým rozšířením, existuje velké množství souborů v tomto formátu
B používaný formát

C budoucí formát – je nový a zatím ne příliš rozšířený, ale použití narůstá

D potenciální formát – nedá se předpovědět budoucnost formátu kvůli nedostatku rozšíření a stavu formátu

E zastaralý formát – byl nahrazen jiným formátem

F formát je nevhodný nebo nemohl být klasifikován

Formát JPEG 2000 byl zařazen do třídy B, formát TIFF do třídy A.

Za zajímavé považujeme uvést i hodnotící kritéria, která skupina KOST-CECO z hodnocení vyřadila, a jak tento krok zdůvodnila. Některá z těchto vyřazených kritérií jsou totiž použita v hodnocení Kongresové knihovny, knihovny Harvardovy univerzity, NARA a u dalších institucí. Skupina KOST-CECO explicitně uvádí a vysvětluje vyřazení těchto kritérií z hodnocení formátů:

- Podpora ukládání metadat

Toto kritérium sledovalo možnosti ukládání metadat a jejich následnou extrakci. Kritérium zůstalo v hodnocení KOST-CECO jen pro videa, pro ostatní formáty se k hodnocení nepoužívá, protože vložená metadata, dle názoru KOST-CECO, nepředstavují žádný výrazný benefit z hlediska digitální archivace. Externě uložená metadata jsou dle jejich názoru dostatečná, jsou také obvykle kompletnější, kvalitnější, než metadata ukládaná v souborech.

- Interoperabilita

Tým KOST-CECO vyřadil z hodnocení kritérium interoperability formátu, tj. hodnocení, nakolik je formát nezávislý na specifickém hardwaru a softwaru. Je to vlastnost, kterou je třeba sledovat u všech formátů, které v KOST-CECO sledují, resp. je to jejich základní požadavek na formáty, které mají v plánu hodnotit.¹⁴

- Autenticita

Požadavkem kritéria bylo, že formát musí garantovat autenticitu dokumentu. Kritérium bylo vyřazeno, protože autenticita formátu má být deklarována jinými mechanismy (např. metadaty, hašováním) než prostřednictvím souborového formátu.

- Detekce a oprava chyb

Toto kritérium sledovalo, zda formát podporuje detekci chyb a jejich automatickou korekci (jako např. u formátu PNG). V KOST-CECO tuto funkcionalitu neočekávají u souborových formátů, ale u úložných technologií, které tuto funkci dnes poskytují, proto nebylo kritérium zahrnuto pro hodnocení souborového formátu (Koordinationsstelle für die Dauershafte Archivierung elektronischer Unterlagen 2019d).

1.1.5 Hodnocení formátů v Britské knihovně

Britská knihovna provádí hodnocení souborových formátů ve spolupráci s externími pracovníky. Výsledkem je písemná zpráva o každém formátu, kde jsou podrobně kome-

¹⁴ Je tedy možné říci, že toto kritérium sledují, leč není součástí explicitně vyjádřené metodologie hodnocení.

tována jednotlivá hodnotící kritéria. Hodnocení vychází ze zaznamenaných zkušeností a znalostí o formátu nejen v Britské knihovně, ale i ve světě, včetně příspěvků na blozích a v diskuzích.

Při hodnocení v Britské knihovně sledují následující vlastnosti formátu (British Library 2015):

- Stav vývoje formátu

Je sledována historie formátu, vlastníci formátu a aktuální stav, ve kterém se formát nachází.

- Rozšíření, používání

Je sledováno, zda se formát používá a jaké jsou s ním zkušenosti. Především je brán ohled na jeho rozšíření v paměťových institucích.

- Softwarová podpora

Je hodnocena softwarová podpora v nástrojích pro zobrazení a práci s formátem. Konkrétně se zjišťuje, zda existují desktopové aplikace, jaká je podpora na běžných počítačových sestavách apod. Je rovněž sledována podpora v nástrojích využívaných při digitální archivaci, tedy, zda existují nástroje, které dokáží formát spolehlivě identifikovat, validovat, detekovat rizika a poškození, extrahovat metadata a formát migrovat.

- Dokumentace a dostupnost pokynů

Je hodnoceno, zda je dostupná dokumentace formátu, zda existují rady a doporučení, jak s formátem pracovat, jak nastavit prostředí pro práci.

- Složitost formátu

Z důvodu zjištění dopadu na správu souborů v tomto formátu v organizaci posuzují složitost formátu. Čím vyšší složitost, tím jsou vyšší nároky na znalosti pracovníků.

- Vložený obsah, připojený obsah

Jsou hodnoceny možnosti vkládání obsahu či připojování obsahu k souborům a dopad na dlouhodobé uchovávání takových činností.

- Závislost na externím prostředí

Hodnotitelé sledují závislost formátu na hardwaru a softwaru.

- Právní stav

Hodnotitelé popisují, zda existují právní překážky pro používání, správu nebo uchovávání souborů v tomto formátu.

- Mechanismy technické ochrany

Je hodnoceno, zda formát podporuje šifrování, DRM a další ochranné mechanismy a zda tyto představují komplikaci pro digitální archivaci.

- Další rizika pro archivaci

Dále hodnotitelé uvádějí rizika formátu, jež nebyla popsána v žádné z výše uvedených kategorií, zpravidla rizika specifická pro konkrétní formát. Pro formát JPEG 2000 tak v roce 2015 byla uvedena například následující rizika: problémy s identifikací (JPX byl občas identifikován jako JP2), nedostatečná podpora softwaru, malé rozšíření a tedy riziko objevení dalších problémů při postupném rozšiřování formátu apod.

Závěr zprávy obsahuje doporučení pro správu formátu, včetně konkrétních aktivit, které by instituce měly provádět při správě hodnoceného formátu. (British Library 2015).

1.1.6 Porovnání a vyhodnocení

Pro přehlednost byla vytvořena tabulka srovnávající výskyt jednotlivých kritérií u sledovaných institucí, resp. kritérií, která jsou explicitně uvedena v jejich metodikách či příručkách pro hodnocení.

Tab. 1 Přehled užití kritérií ve vybraných institucích

Instituce		Kongresová knihovna	Harvardova univerzita	NARA	KOST-CECO	Britská knihovna
Kritérium						
Disclosure	Zveřejnění	ano	ano	ano	ano	ano
Adoption	Rozšíření, uživatelská přijatelnost formátu	ano	ano	ano	ano	ano
Transparency / complexity ¹	Transparentnost formátu/ složitost formátu	ano	ano	ano	pouze u videoformátů	ano
Self-documentation	Autodokumentace	ano	ano	ano	pouze u videoformátů	ne
External dependencies	Závislost na externím prostředí	ano	ano	ano	ano/ne ²	ano
Impact of patents	Právní situace formátu	ano	ano	ano	ano	ano
Technical protection mechanisms	Mechanismy technické ochrany	ano	ano	ano	ne	ano
Quality and functionality	Kvalita a funkcionálnost ³	ano	ano	ne	ano	ne
Cost factors	Finanční náklady	ne	ano	ne	ne	ne
Format age	Stav vývoje a stáří formátu	ne ⁴	ano	ano	ne	ano
Feasibility	Možnost konverze do jiného formátu	ne	ne	ano	ne	ano
Embedded or Attached Content	Možnost vložení či připojení dalších souborů	ano/ne ⁵	ne	ne	ne	ano

1 Pojmy Transparentnost a komplexita, složitost formátu nejsou vzájemně zaměnitelné, nejsou synonyma, ale zčásti se překrývají, sledují složitost formátu. Platí, že vyšší komplexnost formátu (složitost) spíše snižuje jeho transparentnost a naopak vysoká transparentnost formátu obvykle znamená jeho menší složitost.

2 V KOST-CECO tuto vlastnost očekávají u všech hodnocených formátů, proto ji v přímém hodnocení nesledují, je součástí spíše jejich předvýběru formátů. Explicitně tedy není v dokumentaci k hodnocení uvedena, ale sleduje se.

3 Kongresová knihovna a Harvardova knihovna toto kritérium rozpracovaly do dílčích charakteristik, skupina KOST-CECO ne, ta hodnotí, zda formát dokáže zachytit signifikantní vlastnosti předlohy/předchozího formátu.

4 Kongresová knihovna stáří formátu, resp. jeho specifikace a její aktualizace, nehodnotí, explicitně ji v dokumentaci k hodnocení neuvádí, ale tuto vlastnost u formátů sleduje a zapisuje ji do formátového registru.

5 Explicitně tuto vlastnost v hodnotící dokumentaci neuvádí, ale může být popsána v rámci kritéria Kvalita a funkcionálnost.

Sledované instituce používají k hodnocení formátů podobná kritéria. Všechny instituce sledují u formátů následující vlastnosti: vnitřní složitost formátu (tj. komplexnost, transparentnost formátu¹⁵), stav jeho dokumentace, rozšíření mezi uživateli, právní situaci a existenci nástrojů pro validaci. Tři instituce explicitně uvádějí, že u souborových formátů sledují i závislost formátu na externím prostředí, tj. na konkrétním hardwaru a softwaru. Tuto vlastnost v přímém procesu hodnocení nesleduje skupina KOST-CECO, resp. tuto vlastnost přímo nehodnotí. Neznamená to však, že tato vlastnost pro ni není u souborových formátů důležitá, naopak je důležitá velmi. Dostupné zdroje naznačují, že tato vlastnost určuje, které formáty jsou vůbec vhodné pro to, aby se jimi skupina KOST-CECO dále zabývala, tj. použila je v procesu předvýběru formátů pro další hodnocení (Koordinationsstelle für die Dauerschafte Archivierung elektronischer Unterlagen 2019d). Hodnocení KOST-CECO jako jediné z popsanych nesleduje u formátů mechanismy technické ochrany a možnosti vkládání metadat (autodokumentace), resp. KOST-CECO sleduje podporu vložených metadat jen u videoformátů. Finanční náklady spojené s formátem explicitně sleduje jen knihovna Harvardovy univerzity. Není však vyloučeno, že ostatní instituce finanční náklady u souborových formátů také posuzují, jen tuto vlastnost nezahrnují do hodnocení vlastností formátu pro archivační účely. Náklady se však musí zvažovat vždy, jelikož významně ovlivňují udržitelnost zvoleného řešení, jen je diskutabilní, nakolik je to vlastnost formátu.

Kritérium „hustota uložení“ (*Storage density*) v hodnocení KOST-CECO sleduje, zda a do jaké míry formát data komprimuje a kolik úložného místa tedy ušetří bez ohledu na to, zda je komprese ztrátová. Nejlépe jsou ohodnoceny formáty, které data komprimují a zařazení kritéria je dle našeho názoru tedy spíše v rozporu s aktuálními trendy nároků na archivační formáty. Obvyklým požadavkem na archivační formát bývá spíše opak, tj. žádná nebo bezeztrátová komprese, což znamená, že nedochází ke ztrátě dat. Respektovaná doporučení také jako nejvhodnější archivační formáty nezdědka doporučují na prvním místě formáty nekomprimované, minimálně v případě obrazových (TIFF) a zvukových dat (WAVE). Nicméně pro některé organizace může být i nadále otázka úložných kapacit významná, a to jednak z důvodu finančních nákladů na úložné prostory, a také z důvodu správy dat a jejich přenosů.

Rizikem průzkumu, resp. zhodnocení postupů jednotlivých institucí je, že ne všechny kroky hodnocení a volby formátu jsou v dostupných zdrojích explicitně vyjádřeny. Proto nelze průzkum jednoduše uzavřít zhodnocením, že některé instituce některá kritéria používají a jiné ne a je tedy na tuto situaci třeba brát ohled i při vytváření lokálního postupu hodnocení.

Co se týče postupu při hodnocení a následném porovnávání hodnocených formátů, působí „výhodněji“ použití bodovacích škál při hodnocení naplnění kritérií. Škály umožňují exaktnější srovnání formátů, tj. je zřejmé, které formáty si vedou lépe, a je jasně odlišena míra naplnění jednotlivých kritérií napříč formáty.

Z praktického hlediska se nám jeví jako výhodnější také formulace konkrétních hodnotících dotazů, případně indikátorů kritérií, které, dle našeho názoru, jednak zvyšují objektivitu hodnocení a jednak umožňují sledovat vývoj jednotlivých sledovaných vlastností v čase.

15 Kritéria *Transparentnost* a *Komplexnost* formátu nejsou navzájem zaměnitelná a jejich obsah se mírně liší, nicméně obě kritéria dle našeho názoru sledují vnitřní složitost formátu, sledují, zda formát používá kompresi.

1.2 Hodnocení formátů v NK ČR

Pro hodnocení formátů v Národní knihovně ČR doposud nebyl vytvořen žádný nástroj ani metodika. Průzkum praxí jiných institucí ale směřuje k vytvoření podobné metodiky i pro NK ČR. Standardním postupem při hodnocení vhodnosti formátu pro archivační účely dosud bylo:

• Studium uznávaných doporučení ohledně souborových formátů

Tento krok obvykle znamená studium následujících dokumentů:

- o *Recommended Formats Statement*¹⁶ (Library of Congress 2019)
Doporučení vydávané Kongresovou knihovnou USA již od roku 2014, je každoročně aktualizováno. Kongresová knihovna v tomto dokumentu doporučuje vhodné archivační formáty pro digitální i analogové dokumenty.
- o *Doporučení americké organizace FADGI*¹⁷ (*Federal Agencies Digital Guidelines Initiative*) (Federal Agencies Digitization Initiative 2016)
Tato organizace se zabývá tvorbou a propagací standardů v oblasti digitalizace a digitální archivace. Publikuje doporučení pro práci s digitálními daty – obrazy, audio, video. Vydala doporučení týkající se vhodných archivačních formátů pro obrazová data.
- o *Doporučení švýcarské expertní skupiny KOST-CECO*, jež na svém webu zpřístupnila své hodnocení a srovnání vybraných souborových formátů¹⁸ (obrazové, zvukové, textové, video formáty apod.). Toto hodnocení pravidelně aktualizuje.
- o *Hodnoticí zprávy souborových formátů od Britské knihovny*¹⁹
Britská knihovna dosud vypracovala hodnoticí zprávy pro 15 souborových formátů. Na základě těchto zpráv je možné provádět informovaná rozhodnutí ohledně užití formátů. Hodnoticí zprávy jsou průběžně aktualizovány.

• Vlastní průzkum užití formátu

Na základě studia výše uvedených dokumentů obvykle získáme malou skupinu formátů, které jsou uznávanými subjekty považovány za vhodné pro archivační účely. Tyto formáty pak hledáme v doporučení paměťových institucí. Paměťové instituce někdy zveřejňují svoji formátovou politiku (tj. formáty, které přijímají) a v některých případech i uvádějí, nakolik těmto formátům důvěřují – resp. deklarují to zařazením formátu mezi preferovaný či akceptovatelný²⁰ nebo vyjádřením míry důvěry, příp. pravděpodobnosti dlouhodobého uchování dat²¹. Tímto průzkumem zjišťujeme rozšířenost formátu mezi paměťovými institucemi.

16 <https://www.loc.gov/preservation/resources/rfs/>

17 <http://www.digitizationguidelines.gov/>

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

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

20 Viz např. zmíněné doporučení Kongresové knihovny (Library of Congress 2019) nebo instituce UK Data Service (UK Data Service c2012–2020).

21 Viz např. stránka věnující se formátům knihovny univerzity ve Washingtonu (University of Washington, nedatováno) a pokyny pro vkladatele obsahu do univerzitního repozitáře Cornellovy univerzity (Cornell University 2020).

• Testování softwarových nástrojů

Výše uvedené instituce a jejich hodnoticí nástroje uvádějí mezi hlavními kritérii podporu formátů v softwarovém prostředí. Také v Národní knihovně ČR je toto kritérium považováno za důležité. Zjišťujeme existenci níže uvedených typů softwaru a ty dostupné pak přímo testujeme.

Sledujeme nástroje:

- Pro tvorbu (kodeky, editační software)
- Pro zobrazování, přehrávání, otevíření
- Pro migraci do jiného formátu
- Pro identifikaci
- Pro validaci
- Pro extrakci metadat

Většinou se nám potvrdí, že pokud formát při hodnocení projde úspěšně prvními dvěma kroky (tj. existuje ve významných doporučeních a je používán v paměťových institucích), výše uvedené nástroje existují v uspokojivé míře (platí pro JPEG 2000, WAV, MP3).

• Ověření dostupnosti dokumentace

Výše uvedené kroky obvykle vedou k volbě jednoho či několika málo vhodných formátů. Tyto formáty jsou osvědčené, rozšířené a nástroje existují. Zjišťujeme tedy, jaká je dostupnost dokumentace formátu. Optimální je uložit specifikaci formátu v úložišti spolu s archivovanými daty. Dostupnost dokumentace ověřujeme také proto, abychom zjistili, že producenti a dodavatelé dat mohou získat k dokumentaci přístup bez větších obtíží.

Výše uvedené kroky a jejich pořadí na první pohled nemusí odpovídat postupům dříve uvedených a respektovaných subjektů v této oblasti. Přesto se domníváme, že i tímto postupem se nám daří získat dostatečné množství informací a na jejich základě rozhodovat o vhodnosti formátů. I tak ale vnímáme, že hodnocení formátů by mělo probíhat exaktněji a konzistentněji, z toho důvodu je jedním z cílů vytvoření hodnoticího nástroje – podrobnější metodiky a přehledné hodnoticí tabulky.

1.2.1 Návrh kritérií pro hodnocení formátů v NK ČR

Na základě dosavadní praxe NK ČR a výše uvedených institucí byly pro NK ČR navrženy kritéria a indikátory jejich naplnění, a také váhy indikátorů (zatím pro pilotní provoz). Přehled je uveden v tabulce včetně zdůvodnění volby jednotlivých indikátorů.

Tab. 2 Návrh kritérií a indikátorů pro NK ČR

Kritérium	Indikátor	Zdůvodnění	Váha
Zveřejnění			
	Specifikace formátu je dostupná	Podoba formátu je popsána a budoucí uživatel (tvůrce softwaru apod.) má šanci tuto informaci nalézt.	1
	Specifikace je kompletní a srozumitelná	Specifikace je dostatečným zdrojem pro porozumění formátu, budoucí vývojáři softwaru dokáží podle ní vytvořit odpovídající software.	1
	Specifikace je standardizovaná	Specifikace prošla revizí, externím posouzením důvěryhodným standardizačním institutem a musí být někde uchována, tj. zvyšuje pravděpodobnost pozdější dostupnosti.	0,5
	Formát je popsán ve formátovém rejstříku Kongresové knihovny	Formát byl podroben zkoumání touto významnou autoritou v oblasti hodnocení formátů, o stavu formátu byly zaznamenány informace včetně odkazů na další informační zdroje. Rejstřík je průběžně aktualizován a předpokládáme, že bude dlouhodobě udržován. (1)	0,5
Rozšíření, užití			
	Formát je široce rozšířený mezi uživateli	Používá-li formát více institucí a zároveň existuje větší množství souborů v tomto formátu, bude zájem o vývoj nástrojů pro formát nejen v současnosti, ale i v budoucnu. Mezi uživateli zřejmě bude formátu někdo také dobře rozumět. Situace formátu se nezmění náhle, neočekávaně a pokud se jednou změní a bude nutná nějaká akce, mohou se uživatelé formátu spojit a např. zajistit finance na vývoj nástroje apod. Pro současnost to znamená, že bude zřejmě existovat dost nástrojů pro práci s formátem.	1
	Používá se pro archivační účely	Formát byl jinou institucí vybrán pro archivační účely, tj. zřejmě jeho vhodnost pro archivační účely již někdo hodnotil a formát na základě toho zvolil. V komunitě také existuje zkušenost s formátem při archivaci a bude zájem na vývoji migračních nástrojů.	1

	Formát figuruje v uznávaném doporučení	Formát prošel důkladnějším posouzením významným subjektem v oblasti digitální archivace a je tedy vyšší šance, že formát bude skutečně vhodný pro archivaci.	1
	Formát figuruje ve formátových politikách více institucí	Formát zřejmě prošel lokálním posouzením, instituce jej dokáže archivovat a s tím úmyslem jej do úložiště přijímá. Sleduje průběžně stav formátu a v budoucnu bude také řešit jeho migraci.	0,5
Implementace			
	Existuje více nástrojů pro vytváření souborů v tomto formátu	S daným formátem je možné běžně pracovat. Dle dostupných informací o formátu je možné vytvořit odpovídající nástroj a zřejmě se nejedná o obtížný vývoj, je-li nástrojů více. Zároveň nejsou uživatelé závislí na jednom nástroji a tedy na jedné instituci, která jej spravuje.	1
	Zdrojový kód nástroje pro tvorbu souborů je dostupný jako open source	Používání nástroje může být pro instituce a její uživatele relativně levné, jedna open source implementace může umožnit vznik dalších implementací.	1
	Existuje více nástrojů pro zobrazení formátu	S daným formátem je možné běžně pracovat. Dle dostupných informací o formátu je možné vytvořit odpovídající nástroj a zřejmě se nejedná o obtížný vývoj, je-li nástrojů více. Zároveň nejsou uživatelé závislí na jednom nástroji a současně na jedné instituci, která jej spravuje.	1
	Zdrojový kód nástroje zobrazení formátu je dostupný jako open source	Používání nástroje může být pro instituce a její uživatele relativně levné, jedna open source implementace může umožnit vznik dalších implementací.	0,5
	Je dostupný nástroj pro validaci formátu	Validace (tj. kontrola, zda formát odpovídá své specifikaci) je považována za jeden z důležitých kroků při dlouhodobém uchovávání obsahu. Budoucí vývojáři softwaru se mohou tedy spolehnout na specifikaci při vytváření aplikací, validní soubory by tak mělo být vždy možné použít.	1
	Je dostupný nástroj pro migraci formátu do jiného formátu	V případě rizika je možné formát migrovat ihned.	1

Transparent- nost, složi- tost formátu			
	Formát lze spolehlivě identifikovat	Správně identifikovaný formát lze průběžně sledovat, posuzovat, extrahovat jej z úložiště a provádět s ním odpovídající akce.	1
	Formát lze spolehlivě validovat	Lze provést kontrolu, zda formát odpovídá své specifikaci a je tedy možné formát spravovat s nižší mírou obav ohledně výskytu neočekávaných komplikací. Váha je snížena z důvodu, že validační nástroje se postupně zdokonalují, příp. nové vyvíjejí. V době posuzování formátu může být validační nástroj teprve vyvíjen, zdokonalován, je třeba tuto situaci sledovat, ale formát i tak může být vhodný pro archivační účely.	0,5
	Není použita komprese / Je-li použita komprese, je tento kompresní mechanismus znám a užíván	Nesnižuje porozumění formátu, kompresní mechanismus bude srozumitelný i budoucím vývojářům softwaru.	1
Patenty, licence			
	Právní stav formátu je znám	Na základě informací je možné s jistotou posoudit, zda je právní situace formátu pro instituci akceptovatelná.	1
	Formát lze užít volně	Podporuje rozšíření formátu.	
Externí závislosti			
	Formát může být užit na více operačních systémech	Formát je možné implementovat v různých workflow, uživatelé mohou s formátem pracovat, resp. zobrazit jej, aniž by museli instalovat jiný OS. Formát je zřejmě implementovatelný na různé platformy současné i budoucí.	0,5
	Formát je nezávislý na konkrétním HW	Není třeba uchovávat konkrétní hardware, není třeba pořizovat konkrétní hardware.	1
	Formát je použitelný ve více SW aplikacích	Užití formátu nezávisí na jednom softwaru, nemusí se sledovat stav potřebného softwaru (jeho situace na trhu, potřeba aktualizace)	1

	Užití formátu nevyžaduje pluginy ani skripty	Stav pluginů bývá nejistý, mohou téměř ze dne na den zmizet. Například u internetových prohlížečů, ať už u nového prohlížeče, nebo u nové verze stávajícího prohlížeče, nemusí existovat zájem pluginy „dovytvářet“. Existence pluginů v případě internetových prohlížečů také znamená, že formát není podporován přímo prostředím a není o to ani zájem; tento stav může být indikátorem, že formát pomalu zastarává (viz formát DJVU ale srov. s JPEG 2000).	1
Vlastnosti dat			
	Umožňuje uložení bez komprese	Nekomprimované formáty mají jednodušší strukturu. Nedochází ke ztrátě dat.	0,5
	Podporuje skutečně bezztrátovou kompresi	Nedochází ke ztrátě dat, uchová signifikantní vlastnosti.	0,5
	Uchová významné vlastnosti předlohy	Nezbytné pro odpovídající prezentaci obsahu, tak jak jej tvůrce zamýšlel, tj. zachování vzhledu, funkcionality předlohy. Konkrétní indikátory se liší dle typu dat. U obrazových formátů je důležitá hlavně podpora správy barev a možnost uložit obraz ve vysokém rozlišení apod.	1
1 Předpokládáme, že v budoucnu nebude toto kritérium omezeno pouze na rejstřík Kongresové knihovny. Již nyní existují důvěryhodné zdroje (viz text článku), které by mohly být u kritéria uvedeny jako součást výčtu. Nicméně dosud je rejstřík Kongresové knihovny co do počtu popsanych formátů zřejmě nejkompletnější (nemusí platit o rozsahu informací u jednotlivých formátů) a předpokládáme, či spíše doufáme, v jeho ne-li trvalou, tak alespoň dlouhodobou dostupnost.			

Záměrem tedy je sledovat u formátu výše uvedené vlastnosti a zaznamenat jejich stav v době hodnocení. Následovat bude otestování kritérií a jejich bodového ohodnocení na formátech, jež aktuálně Národní knihovna ČR spravuje, doporučuje nebo o nich do budoucna uvažuje.

2 Frekvence revizí a dialektika formátů

Míra shody formátů se sledovanými kritérii není neměnná a proto je nutné hodnocení periodicky opakovat, aby byly zachyceny případy, kdy se tato míra změnila v důsledku změn příslušných vlastností. Proměnlivost vlastností formátů se liší v závislosti na tom, do jaké míry jsou to vlastnosti formátu vlastní²² a do jaké míry jsou naopak dány vztahem

22 Například transparentnost/komplexnost formátu, sebedopisnost, funkcionality formátu.

formátu a prostředí.²³ Čím více jsou sledované vlastnosti závislé na prostředí, tím více se mohou měnit.

Například rozšíření formátu je vlastností, která je silně závislá na uživateli, jejichž rozhodování ovlivňují vlastnosti formátu, a to všechny vlastnosti včetně rozšíření samotného, a potom vnější okolnosti. Na příkladu JPEG 2000 si můžeme ukázat, jak důležitou z těchto vnějších okolností je hardware. V době, kdy JPEG 2000 vznikl, byly jeho nároky na RAM limitujícím faktorem (Archambault 2015), takže až v následující dekádě mohlo dojít k jeho plošnému rozšíření, které samozřejmě podpořilo další rozvoj nástrojů a dokumentace, až nakonec začal splňovat kritéria archivačního formátu.

Současně se ale zvedala i kapacita hard disků, která snížila význam komprese a dnes řada fotografií dokonce i archivuje ve formátech RAW, které navíc nabízejí vyšší bitovou hloubku. Symbolickou tečkou za přijetím JPEG 2000 ve fotografické komunitě je fakt, že i po 20 letech od vzniku tohoto nového formátu zůstává kompresním formátem implementovaným do digitálních fotoaparátů původní JPG. Na druhou stranu však JPEG 2000 je úspěšným standardem v medicínských a astronomických zobrazovacích technikách (Otemuywa 2017) a samozřejmě v archivnictví, kde jsou nároky na paměť řádově větší, a to, že technické zastarání JPEG 2000 je v tuto chvíli za hranicemi naší představitosti, podporuje i jeho úspěch v oblasti streamovacích videokodeků.

Sledované vlastnosti spolu s faktory, které působí změny těchto vlastností, se navzájem ovlivňují v provázané síti. Hardware v tomto stojí poněkud stranou. Poptávka ze strany uživatelů sice samozřejmě tlačí na jeho vývoj, ale ten je limitován celkovým vývojem našeho poznání. V oblasti hardwaru lze dělat předpovědi, jako např. Moorův zákon (Polesný 2019). Při úvahách o stanovení doby, po které by se měla opakovat revize vhodnosti archivačních formátů, proto zvažujeme možnost vycházet z predikcí vývoje hardwaru.

V současnosti se uvádí, že průměrná doba obměny hardwaru je šest let. Vzhledem k tomu, že nelze určit, v jaké fázi cyklu vývoje hardwaru se nacházíme, rozhodli jsme se stanovit periodu revizí na polovinu této doby, tedy tři roky. Zároveň ale uvažujeme o tom, že bychom pro určité formáty určili prahové hodnoty poměru *cena výkon*, při jejichž překročení bychom přistoupili k hodnocení *ad hoc*. Pro JPEG 2000 byla v minulosti limitujícím faktorem kapacita RAM, ale do budoucna bude hlavním sledovaným faktorem nepochybně cena paměťových médií, která by v kombinaci s příchodem dedikovaného²⁴ archivačního formátu pro rastrové obrázky analogického k PDF/A mohla být signálem k opuštění komprese.

Závěr

V uplynulém období byly sledovány postupy významných institucí při hodnocení formátů z hlediska digitální archivace s cílem navrhnout exaktnější postup hodnocení pro praxi v Národní knihovně ČR. V textu byl tento návrh představen, dalším krokem bude jeho nasazení do pilotního provozu. V následujícím období bude prezentovaný návrh rozpracován do lokální interní metodiky. Postupně by také měly vzniknout hodnotící zprávy pro souborové formáty, jež Národní knihovna ČR aktuálně spravuje či se spravovat chystá, a pro další formáty, které budou v průběhu let podrobeny hodnocení.

V dlouhodobějším výhledu očekáváme v oblasti formátových politik poměrně dynamický vývoj, který nastane s příchodem dedikovaných archivačních formátů. Mo-

23 Například existence, kvalita a množství nástrojů pro tvorbu, prezentaci, validaci formátu; stav standardizace specifikace, rozšíření formátu mezi uživateli a profesionály.

24 Formát navržený pro specifický účel, v tomto případě archivaci.

mentálně je v této oblasti zavedeným formátem pouze PDF/A, zatímco iniciativa TIFF/A vyhlášená v roce 2015 v tuto chvíli o sobě nedává vědět. Nicméně s růstem archivační komunity lze očekávat, že tato komunita postupně vyvine formáty vlastní. Současné formátové politiky se pro tento trend mohou stát významným zdrojem informací o potřebách dlouhodobého uložení. Zároveň budou možná tímto vývojem do jisté míry potlačeny, ale jistě ne zcela, protože i dedikované formáty je třeba hodnotit. V každém případě se pozice archiváře ve vztahu k formátům posune od pasivního hodnotitele k roli aktivního činitele při vývoji formátů.

Zdroje:

ARCHAMBAULT, Michael, 2015. JPEG 2000: The Better Alternative to JPEG That Never Made it Big. In: *PetaPixel* [cit. 2020-03-20]. Dostupné z: <https://petapixel.com/2015/09/12/jpeg-2000-the-better-alternative-to-jpeg-that-never-made-it-big/>.

BRITISH LIBRARY, 2015. Format Assessment Factors. DPC [online]. Datum poslední revize 16. března 2015 [cit. 2020-03-19]. Dostupné z: https://wiki.dpconline.org/index.php?title=File_Format_Assessment_Factors.

BROWN, Adrian, 2008. *Selecting file formats for long-term preservation. The National Archives (UK). Digital preservation guidance note 1.* 2008 [cit. 2020-06-08]. Dostupné z: <http://www.nationalarchives.gov.uk/documents/selecting-file-formats.pdf>.

CORNELL UNIVERSITY, 2020. Recommended File Formats [online]. Seattle: Cornell University Library, Last Updated: May 28, 2020 [cit. 2020-06-08]. Dostupné z: <https://guides.library.cornell.edu/ecommmons/formats>.

DIGITAL PRESERVATION COALITION, 2015. File formats and standards. *Digital Preservation Handbook* [online]. 2nd ed. Glasgow: Digital Preservation Coalition, c2015 [cit. 2020-03-20]. Dostupné z: <https://www.dpconline.org/handbook/technical-solutions-and-tools/file-formats-and-standards>.

FEDERAL AGENCIES DIGITIZATION INITIATIVE, 2016. Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Master Files [online]. Washington (DC): FADGI, September 2016 [cit. 2020-03-15]. Dostupné z: http://www.digitizationguidelines.gov/guidelines/FADGI%20Federal%20Agencies%20Digital%20Guidelines%20Initiative-2016%20Final_rev1.pdf.

GOETHALS, Andrea, 2016a. Format Assessments. *Harvard Wiki* [online]. Harvard College, aktualizováno 15. dubna 2016 [cit. 2020-03-05]. Dostupné z: <https://wiki.harvard.edu/confluence/display/digitalpreservation/Format+Assessments>.

GOETHALS, Andrea, 2016b. Format matrix tool. *Harvard Wiki* [online]. Harvard College, aktualizováno 16. dubna 2016 [cit. 2020-03-05]. Dostupné z: https://docs.google.com/spreadsheets/d/1buM2XZtkc09kUtUo0W5s0lt4IK_6LALF6VooCJDdQZ0/edit.

JOHNSTON, Leslie, 2019. SOP for NARA *Digital Preservation Framework* [online]. Verze 0.3. NARA, [cit. 2020-03-21]. Dostupné z: https://github.com/usnationalarchives/digital-preservation/blob/master/Digital%20Preservation%20Risk%20Matrix/NARA_Risk_Framework_SOP_20190827.pdf ²⁵.

KOORDINATIONSSTELLE FÜR DIE DAUERHAFT ARCHIVIERUNG ELEKTRONISCHER UNTERLAGEN, 2019a. *JPEG2000* [online]. KOST-CECO Version 6.0, Juli 2019 [cit. 2019-11-05]. Dostupné z: <https://kost-ceco.ch/cms/jpeg2000.html>.

25 Stávající odkaz nefunguje. V době tisku tohoto příspěvku bude dostupná nová verze dokumentu.

KOORDINATIONSSTELLE FÜR DIE DAUERHAFT ARCHIVIERUNG ELEKTRONISCHER UNTERLAGEN, 2019b. Kriterienkatalog [online]. KOST-CECO, Version 6.0, Juli 2019 [cit. 2020-06-08]. Dostupné z: <https://kost-ceco.ch/cms/Kriterienkatalog.html>.

KOORDINATIONSSTELLE FÜR DIE DAUERHAFT ARCHIVIERUNG ELEKTRONISCHER UNTERLAGEN, 2019c. TIFF [online]. KOST-CECO Version 6.0, Juli 2019 [cit. 2019-11-05]. Dostupné z: <https://kost-ceco.ch/cms/tiff.html>.

KOORDINATIONSSTELLE FÜR DIE DAUERHAFT ARCHIVIERUNG ELEKTRONISCHER UNTERLAGEN, 2019d. Verworfen Kriterien [online]. KOST-CECO, Version 6.0, Juli 2019 [cit. 2020-06-08]. Dostupné z: <https://kost-ceco.ch/cms/Verworfen-Kriterien.html>.

LIBRARY OF CONGRESS, 2017a. Formats, Evaluation Factors, and Relationships. *Sustainability of Digital Formats: Planning for Library of Congress Collections* [online]. Washington (DC): The Library of Congress, Last Updated: 03/2/2017 [cit. 2020-06-08]. Dostupné z: https://www.loc.gov/preservation/digital/formats/intro/format_eval_rel.shtml.

LIBRARY OF CONGRESS, 2017b. Sustainability Factors. *Sustainability of Digital Formats: Planning for Library of Congress Collections* [online]. Washington (DC): The Library of Congress, Last Updated: 01/ 5/2017 [cit. 2020-06-08]. Dostupné z: <https://www.loc.gov/preservation/digital/formats/sustain/sustain.shtml>.

LIBRARY OF CONGRESS, 2019. *Recommended Formats Statement 2019–2020* [online]. Washington (DC): The Library of Congress, [cit. 2019-10-16]. Dostupné z: <http://www.loc.gov/preservation/resources/rfs/RFS%202016-2017.pdf>.

OTEMUYIWA, Prosper, 2017. The Great JPEG 2000 Debate: Analyzing the Pros and Cons to Widespread Adoption. In: *Cloudinary blog*. 2017 [cit. 2020-06-08]. Dostupné z: https://cloudinary.com/blog/the_great_jpeg_2000_debate_analyzing_the_pros_and_cons_to_widespread_adoption.

Podvzorkování barvosných složek. In: *Wikipedia: the free encyclopedia* [online]. San Francisco (CA): Wikimedia Foundation, 2001–, datum poslední revize 2. 4. 2016, 18:43 [cit. 2020-10-08]. Dostupné z: https://cs.wikipedia.org/w/index.php?title=Podvzorkov%C3%A1n%C3%AD_barvosn%C3%BDch_slo%C5%BEEK&oldid=13520901.

POLESNÝ, David, 2019. Vizualizace: Jak Moorův zákon předpověděl vývoj čipů na padesát let dopředu. *Živě* [online]. 2019 [cit. 2020-06-08]. Dostupné z: <https://www.zive.cz/clanky/vizualizace-jak-mooruv-zakon-predpovedel-vyvoj-cipu-na-padesat-let-dopredu/sc-3-a-200157/default.aspx>.

ROG, Judith a Caroline van WIJK, 2008. *Evaluating File Formats for Long-term Preservation* [online]. 2008 [cit. 2019-10-16]. Dostupné z: http://www.kb.nl/hrd/dd/dd_links_en_publicaties/publicaties/KB_file_format_evaluation_method_27022008.pdf.

UK DATA SERVICE, c2012–2020. *Recommended formats* [online]. University of Essex, University of Manchester and Jisc, c2012–2020 [cit. 2020-06-08]. Dostupné z: <https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats>.

UNIVERSITY OF WASHINGTON. *Preferred File Formats* [online]. Seattle: University of Washington Libraries, [b.r.], [cit. 2020-06-08]. Dostupné z: https://www.lib.washington.edu/preservation/preservation_services/digitization-and-digital-preservation/preferred-file-formats.

OSTRÁKOVÁ, Natalie a Vojtěch KOPSKÝ. Posuzování souborových formátů z hlediska dlouhodobého uchovávání a návrh metodiky pro Národní knihovnu České republiky. *Knihovna: knihovnická revue*. 2020, **31**(2), 83–105. ISSN 1801-3252.