Metodika bitové ochrany digitálních dat

Růžička, M.
2019

Dostupný z http://www.nusl.cz/ntk/nusl-393240

Metodika bitové ochrany digitálních dat

Revize 1.0, 2018-09-13

Metodika byla vytvořena v rámci *Programu na podporu aplikovaného výzkumu a experimentálního vývoje národní a kulturní identity na léta 2016 až 2022 (NAKI II)*, pro projekt *ARCLib – komplexní řešení pro dlouhodobou archivaci digitálních (knihovních) sbírek*, Id-kód projektu *DG16P02R044*.

Autoři: Michal Růžička, Andrea Miranda, Lukáš Hejtmánek
Ve spolupráci s: Zdeněk Vašek, Vlastimil Krejčíř, Miroslav Bartošek

Cíle metodiky

Cílem metodiky bitové ochrany digitálních dat jsou doporučení pro implementaci úložiště vhodného pro dlouhodobé uchovávání dat (Long Term Preservation, LTP). Metodika poskytuje návod, jak k problému bitové ochrany digitálních dat v prostředí českých knihoven přistoupit na základě poznatků dobré praxe v této oblasti dostupných v ČR i ve světě. Autorský tým přitom mimo jiné zúročuje zkušenosti získané při řešení projektu CESNET LTP Pilot[1]. Pro širokou dostupnost knihovnám ČR i jinde je celé doporučení postaveno na volně dostupných technologiích s otevřeným zdrojovým textem a licencí umožňující další rozvoj a přizpůsobení potřebám uživatele.

Principem bitové ochrany digitálních dat je zachovat, tedy ochránit, digitální objekt ve stavu, ve kterém byl uložen. Metodika definuje postupy, jakými dlouhodobě uchovávat digitální data v systémech využívajících software ARCLib. Je zaměřena zejména na zachování bitové integrity uložených objektů v souladu s postupy doporučenými mezinárodními normami ČSN ISO 14721 a ČSN ISO 16363 a dalšími doporučeními z oblasti bitové ochrany. Metodika vychází z těchto norem a doporučení a představuje jejich aplikaci na konkrétním systému dlouhodobého uchovávání. Součástí metodiky je popis základních nároků na úložiště, které lze za účelem dlouhodobého ukládání dat spolu s bitovou ochranou využít.

Na rozdíl od postupů pro logickou ochranu dat je cílem bitové ochrany zajistit bezpečné a neměnné uložení binárních streamů, kterými jsou uchovávaná data reprezentována. A to i při velkém objemu dat. Na základě analýzy rizik ohrožujících fyzické uložení dat popisuje metodika jak technické aspekty, tj. doporučené úložné technologie a způsob jejich napojení na informační systém, tak administrativní aspekty, tj. výběr lokalit pro fyzické umístění úložišť, politiky kontroly integrity dat a plány zotavení dat při jejich poškození. Jednotlivé postupy včetně správy většího množství dat jsou již v praxi vyzkoušené. Pro dosažení účelu dlouhodobého uchovávání jsou však nezbytné komplexní postupy propojující jednotlivé dílčí aspekty. Právě ony jsou popsány v této metodice s ohledem na konkrétní nástroj. Procesy k uchování binárních streamů vykonává aktuálně v ČR více organizací spravujících velké objemy dat, nicméně dokument metodické povahy s aplikovanými postupy v českém prostředí dosud chybí[2]. Některé z obecných principů jsou však zmíněny v legislativě určené pro uchovávání archiválií.

Obdobně lze konstatovat, že ani výše popsané normy ČSN ISO 14721 a ČSN ISO 16363 neurčují konkrétní postupy, jak zajistit dlouhodobé uchování. Pouze definují soubor doporučení, kterými je třeba se řídit. Existuje však mnoho technických postupů, jak doporučení realizovat. Cílem této metodiky je tedy shrnout obecná doporučení a strategie, jakými se řídit při praktické správě velkého množství dat a to nejen s pomocí výše zmíněných norem, ale též dalších dostupných dokumentů, které vznikly v mezinárodní komunitě a na jejichž základě jsou pak spravována konkrétní řešení (např. soustava založená na principech LOCKSS).
V obecné části metodika shrnuje nároky na systémy, které jsou nebo budou využívány pro správu digitálních dat a jejich ochranu na úrovni bitstreamu. Všímá si také způsobů, jakými lze kvalitu systému dokázat a jaké jsou hrozby pro důvěryhodnost a správnou funkci těchto systémů.

V praktické části pak metodika popisuje, jak jsou výše popsaná obecná kritéria provedena v softwarovém nástroji ARCLib, jakým způsobem komunikuje ARCLib s okolním hardwarem a jak konkrétně zajišťuje nároky na ochranu bitstreamu uložených dat. Metodika zároveň dokumentuje konkrétní funkce nástroje a je tedy nezbytným podkladem, jak aplikovat výsledky postupů, které byly navrženy během vývoje softwarového řešení ARCLib. Z obecného hlediska nejde o nové postupy, avšak v každém nástroji, který je vyvinut za účelem reálné aplikace doporučených postupů, se jejich konkrétní aplikace odlišuje na základě rozdílného charakteru softwarového řešení, zvolených hardwarových řešení a metod pro ukládání dat. V tomto ohledu

[1] http://ltp-portal.mzk.cz/ltp-pilot/

[2] Snad nejblíže tomuto dokumentu je Metodika pro vytváření bezpečnostních kopií archiválií v digitální podobě. Primárně je určena k využití archivům, které plní povinnosti dle archivního zákona. Dostupná je na stránkách Národního archivu ČR: http://cesarch.cz/wp-content/uploads/2015/06/metodika-pro-bezpecnostni-digitalizaci_v1.pdf.

jde o nové řešení. Právě metody ukládání dat jsou jedním z podstatných momentů celé metodiky, protože společně se způsoby kontroly tvoří základ dlouhodobého uchovávání dat na úrovni bitové ochrany. Metodika by měla být užívána společně s Metodikou logické ochrany digitálních dat [MLO][3]. Toto doporučení platí, pokud chtějí uživatelé uplatnit všechny funkce systému ARCLib. Lze ji využít i samostatně, pokud uživatel usiluje pouze o ochranu bitstreamu.

Pravidla a postupy metodiky byly ověřeny jejími autory jak v souvislosti s jinými systémy (např. systém LTP Národní knihovny ČR, digitální repositář Univerzity Karlovy a další realizované systémy, se kterými se autoři během své kariéry seznámili), tak s jejich vlastními badatelskými postupy a při vývoji řešení ARCLib.

Pro koho je metodika určena

Metodika bude uplatněna zejména u uživatelů systému ARCLib z řad knihoven spravujících větší objemy digitálních dat, ale umožňuje i nezávislé využití. Typickými uživateli jsou krajské knihovny a odborné knihovny, např. Národní technická knihovna, Národní lékařská knihovna, Knihovna AV ČR, které mají své digitalizační projekty, rozšiřují své fondy o elektronické dokumenty a musí tedy řešit alespoň střednědobou archivaci digitálních dat. Tato metodika je rozdělena do dvou hlavních částí, které jsou spolu provázány, ale přesto mohou do určité míry fungovat samostatně. Z tohoto rozdělení vyplývá i dvojí určení metodiky.

V obecné rovině (teoretická část) je metodika bitové ochrany digitálních dokumentů určena všem institucím a jejich odborným pracovníkům, kteří mají dlouhodobé uchovávání těchto dokumentů jako svůj úkol. Popisuje principy péče o fyzické uchovávání, nutné požadavky na dlouhodobá úložiště a postupy pro jejich ověření. Metodika souhrnným způsobem popisuje v současnosti hlavní principy ochrany bitstreamu. V tomto ohledu může být metodika základním dokumentem v oboru v ČR a poskytovat pracovníkům paměťových institucí (nejen knihoven) metodickou podporu při plánování a správě systémů dlouhodobé ochrany digitálních dokumentů bez ohledu na konkrétní technické řešení nebo vymezení na jednotlivé druhy dokumentů. Tyto aktivity jsou již v ČR vykonávány a v mnoha případech mají některé instituce již vypracované vlastní postupy, které jsou však omezeny na jejich konkrétní situaci. Vedle toho existují i další instituce, které tyto postupy nemají nebo je nemají zpracované do potřebné hloubky. Průběžně se také zapojují další a další instituce, pro které může být předložená metodika přínosná i v obecné rovině.

Z hlediska samotné metodiky je však důležitější druhá část speciálně určená pro uživatele systému ARCLib. V případě ochrany bitstreamu nejde dlouhodobé uchovávání digitálních dat redukovat jen na využívání určitého hardwaru (technického vybavení). Jádro úkolů spojených se zajištěním bitové ochrany představují procesy výběru vhodného hardwaru a na něm využívaného softwarového nástroje se všemi funkcemi, které jsou nutné pro správu digitálních dat. Konkrétní část metodiky definuje postupy, jak provádět tyto úkony v systému ARCLib. Popisuje jednotlivé funkcionality systému a jeho mapování na funkční prvky referenčního modelu OAIS a dalších doporučených standardů, které vyplynuly z první části metodiky. Metodika dále představuje podrobný návrh a funkcionalitu modulů ARCLib nutných pro realizaci všech doporučených procesů bitové ochrany. Tvoří proto nedílnou součást užívání systému ARCLib, který ve své komplexnosti nabízí specifický způsob péče o uložené dokumenty, což je vlastnost všech systémů s tímto určením. Vzhledem k open-source charakteru celého systému je nutné, aby jeho uživatelé měli k dispozici veřejně dokumentovanou metodiku. Lze navíc očekávat, že uživateli systému budou často krajské knihovny a specializované knihovny, které potřebují ukládat vědecká a výzkumná data. U knihoven tohoto typu nelze očekávat, že by měly k dispozici plné spektrum odborníků, kteří by byli schopni sami definovat všechny procesy pro ochranu bitstreamu. Systém ARCLib a postupy doporučené metodikou budou v rámci projektu „ARCLib – komplexní řešení pro dlouhodobou archivaci digitálních (knihovních) sbírek“ využity v Knihovně Akademie věd ČR, v. v. i. Vzhledem k veřejnému určení metodiky není třeba uzavírat další smlouvy o jejím využívání. Software ARCLib v Knihovně AV ČR je referenční implementací této metodiky.

[3] http://hdl.handle.net/11104/0282107

Obsah

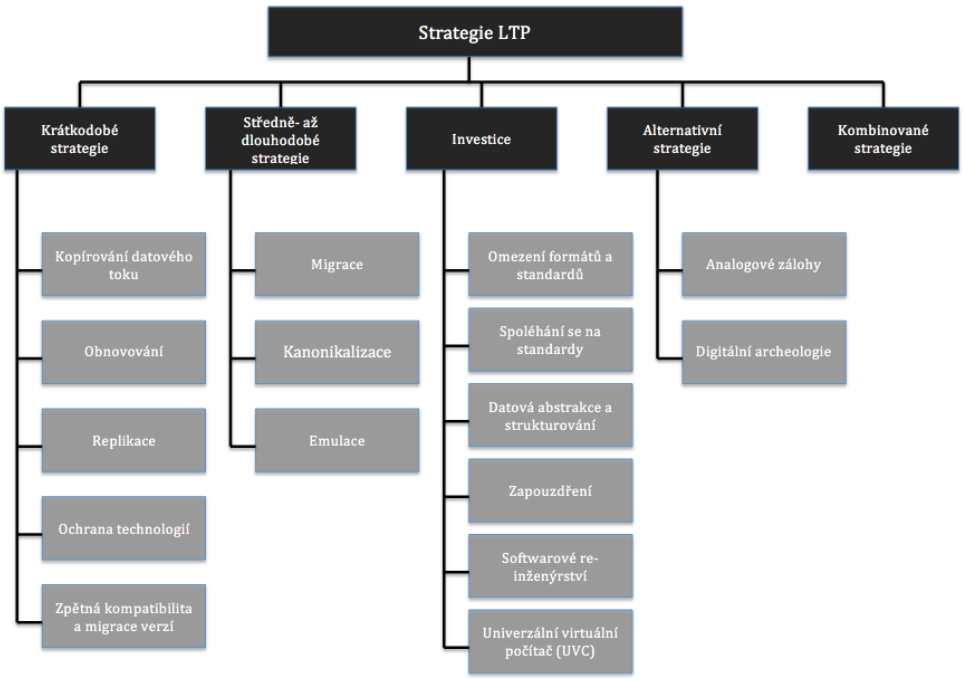
1 Teoretická část

Při plánování dlouhodobého uchovávání dat je nutno zvážit a eliminovat rizika, která ohrožují datová úložiště a stanovit vhodné postupy pro minimalizaci škod způsobených nežádoucími událostmi. Mezi rizika patří selhání hardware, neúmyslná chyba obsluhy, úmyslný útok obsluhy nebo jiného subjektu, přírodní katastrofy, ozbrojené konflikty, legislativní omezení nakládání s daty uchovávanými na určitém území apod., ale také netechnická rizika jako selhání financování, nedostatečný management projektu, chybějící dokumentace či standardizace procesů. Postupy pro minimalizaci škod zahrnují zejména ukládání identických kopií dat ve více geograficky oddělených lokalitách na různých typech úložišť spravovaných různými skupinami osob, a to při zajištění pravidelných kontrol dostupnosti a integrity dat.

Strategie postupu pro každý výše uvedený případ nazýváme politikou zacházení s daty. Nedílnou součástí politiky musí být její pravidelné revize a úpravy dle změněných okolností v průběhu času. Zvolené technické řešení musí být v souladu s politikou definovanými požadavky na exit strategie (export všech dat ve vhodném tvaru pro přenos do jiných/novějších systémů).

Navržené řešení musí být dobře škálovatelné (použitelné jak pro menší, tak i velmi velké objemy dat a pro rozvoj systému s ohledem na počet zapojených účastníků), s dobrou propustností (technicky řešitelné např. hierarchickým uložením dat s rychlým on-line přístupem často využívaného menšího objemu dat versus off-line uložení velkého množství dat, které však znamená velké latence při jejich vybavování) a musí umožňovat použití i více nezávislých řešení podle specifických potřeb jednotlivých institucí. Řešení se dále musí vypořádat s některými základními omezeními, která jsou specifická pro řadu typů úložišť, kupříkladu obtíže a omezení při ukládání velkého množství malých souborů aj.

1.1 Strategie dlouhodobé ochrany



Obrázek 1: Strategie LTP podle [NCDL2014].

Obecně rozlišujeme několik strategií dlouhodobé ochrany, viz Obrázek 1, [NCDL2014], [Guidelines2003], [DPM2012]. Z hlediska archivního úložiště dat jsou relevantní zejména:

1. **Krátkodobé strategie** (nejlépe fungující na krátké časové období).

    1.1. **Kopírování datového toku** (*bitstream copying*, známé jako zálohování) odkazuje na vytváření duplikátů digitálních objektů; podchycuje pouze ztrátu dat v případě selhání médií nebo hardware (ať už lidskou chybou, nebo jinými způsoby).

    1.2. **Obnovování** (*refreshing*) – přesun dat mezi dvěma typy stejného úložiště (úložného média) bez změn a *bitrot* (poškození bitů uvnitř souboru). Tuto strategii je možné kombinovat i s migrací, kdy např. data na CD nosičích již není vhodné přesouvat na nová CD média, ale na zcela nový typ úložiště (např. do cloudu).

    1.3. **Replikace** (*replication*) – má několik možných významů; vytváření duplicitních kopií dat na jednom nebo vícero systémech. Replikovaná data s sebou ale přináší i komplikace v podobě nutného obnovování, migrací, verzování apod. Např. kopírování datového toku (*bitstream*) představuje určitou formu replikace. OAIS standard [OAIS] považuje replikaci za formu migrace. Zřejmě nejznámější realizací této strategie dlouhodobé ochrany je konsorciální projekt LOCKSS. Jedná se o open-source platformu dlouhodobé ochrany vytvořenou na Stanfordské univerzitě, která je postavená na principu *Lots Of Copies Keep Stuff Safe* (LOCKSS), tedy mnoho kopií udrží věci v bezpečí. Klíčovou myšlenkou je vytvoření redundantních dat, kdy je každý digitální dokument uložen v několika separátních repositářích (každý repositář je uchovávaný jinou institucí). Z pohledu dlouhodobé ochrany se LOCKSS řídí těmito hlavními principy [LOCKSS]:

        - Decentralizovaný a distribuovaný princip dlouhodobé ochrany (mnoho kopií udrží věci v bezpečí),
        - lokální opatrovnictví a kontrola nad obsahem v rukách samotných knihoven,
        - uchování původní vydavatelské verze dokumentů,
        - trvalý přístup – garantovaný a bezproblémový,
        - cenová dostupnost a udržitelnost.

2. **Střednědobé až dlouhodobé strategie.**

    2.1. **Migrace** (*migration*) – představuje primární strategii LTP mnohých organizací. Může být chápána dvěma způsoby, jako migrace nosičů na nová média nebo migrace samotných dokumentů do jiných formátů. Migrace formátů souvisí s konceptem zastarávání a *bitrot* souborových formátů[4]. Jedná se tedy o proces transformace digitálních formátů, včetně ochrany datového toku (*bitstream*) a schopnosti zobrazení obsahu reprezentovaného daným datovým tokem, jelikož je digitální dokument neoddělitelně provázaný se svým prostředím (SW, HW, formát), čímž může docházet ke ztrátě samotných informací.

3. **Alternativní strategie** – ne zcela standardní přístupy k dlouhodobé ochraně.

    3.1. **Analogové zálohy** (*analogue backups*) – je zálohování digitálních dokumentů formou jejich převodu do analogové podoby (např. tištěním textových dokumentů na papír nebo

[4] Migrace formátů souborů je nejčastěji realizována v případě zastaralých formátů. Lze ji ale využít i v případech, kdy např. instituci uchovávající dané dokumenty vypršela softwarová licence nebo uchování konkrétních formátů je finančně náročné a dlouhodobě neudržitelné (proprietární formáty).

mikrofilmování[5]), čímž se ale ztrácí výhody digitálních dokumentů[6] (např. sdílení a bezztrátový přenos). Za nejvhodnější typy dokumentů pro analogové zálohování jsou považovány text a černobílé obrazové dokumenty, které nejméně „trpí“ při naplňování této strategie. Jelikož jde o poměrně nákladnou a náročnou formu ochrany s omezeným počtem typů dokumentů, je vhodná pro obsah, který vyžaduje nejvyšší úroveň redundance a ochranu před ztrátou.

3.2. **Digitální archeologie** (*digital archeology*) – zastupuje metodu získávání dat[7] ze zastaralých softwarových a hardwarových prostředí a zastaralých či poškozených médií (např. děrné štítky, 8" diskety apod.). Tato metoda je ale časově, finančně i technicky náročná a nemá stoprocentní úspěšnost (např. v případě chybějících metadat, značně poškozených médií apod.).

4. **Kombinované strategie** výše uvedených přístupů a postupů.

Tato metodika pokrývá zejména strategie z bodu 1. a 2. a jejich kombinace.

1.2 Zákonné požadavky ČR na uchovávání digitálních archiválií

Cílem této metodiky není navrhnout datové úložiště odpovídající specifickým zákonným regulacím vztahujícím se např. na archivy, neboť cílovými uživateli jsou knihovny s jejich specifickými potřebami ukládání knihovních dat. Přesto je však vhodné uvést vybrané zákonné požadavky, jejichž dodržování při budování systému dlouhodobého ukládání dat může být často vhodné i u subjektů, na které se přímo nevztahují.

Obecně pro uchovávání digitálních archiválií stanovuje zákon č. 499/2004 Sb. sadu pravidel. Konkrétně jde o paragraf 61, část 2, kde jsou požadavky zejména na zajištění, že:

a. budova archivu není umístěna v záplavových územích a v ochranných pásmech vzletových a přistávacích drah letišť,
b. budova archivu je situována mimo oblasti plynného a prašného znečištění,
c. prostory pro ukládání archiválií jsou zabezpečeny proti škodlivému působení přírodních vlivů a jevů vyvolaných činností člověka, a to zejména proti průniku vody, páry, dešťové a splaškové kanalizace, nebezpečných chemických a biologických látek nebo působení fyzikálních jevů a proti nadměrné prašnosti, které by mohly vést k poškození nebo zničení archiválií,
d. prostory pro ukládání archiválií jsou umístěny nad hladinou spodní vody, je zajištěno přirozené nebo umělé větrání k udržování stanovené teploty a relativní vlhkosti vzduchu a prostory jsou opatřeny přístroji k měření těchto hodnot,
e. prostory pro ukládání archiválií s magnetickým záznamem jsou chráněny před účinky elektromagnetického pole,
f. archiv, který je současně digitálním archivem, disponuje nejméně dvěma plnohodnotnými úložišti archiválií v digitální podobě vzdálenými od sebe vzdušnou čarou nejméně 50 km, umístěnými v lokalitách, které svým geografickým charakterem vylučují současné nebo následné škodlivé působení přírodních vlivů nebo jevů vyvolaných činností člověka vedoucích k poškození nebo zničení archiválií anebo vyžadujících provedení záchranných prací.

[5] Převod digitálního záznamu na mikrofilm, při němž se vytváří archivní mikrofilmy a digitalizace mikrofilmů druhé generace určené ke zpřístupnění. Tento systém, pod zkratkou COM (*computer-output microfilm*), charakterizuje zápis datového toku přímo z počítače na mikrofilm. Poté se promítá do miniaturizovaných papírových dokumentů. Některé společnosti (např. Kodak, Zeutschel a jiné) nabízejí přístroje Archive Writers, které dokáží zpracovat kvalitní digitální obraz a poté jej zapsat na 16 mm nebo 35 mm mikrofilm.

[6] [Vojtášek2000] nabízí přehled charakteristických vlastností digitálního a analogového dokumentu.

[7] K získávání dat v rámci digitální archeologie se často využívá specializovaný hardware pro digitální forenzní analýzu (pro soudní znalce, forenzní experty, policii a státní instituce).

Dále pak část 4, kde jsou požadavky zejména na zajištění, že:

a. budovy archivů mají zpracovánu bezpečnostní dokumentaci, jejíž součástí musí být opatření proti vniknutí nepovolaných osob do archivních prostor, proti krádeži archiválií a proti teroristickým útokům; u specializovaných archivů a bezpečnostních archivů též opatření zajišťující objektovou bezpečnost,
b. budovy archivů mají zpracovánu požární dokumentaci, jsou vybaveny elektronickou požární signalizací a ručními hasicími přístroji; v prostorách pro uložení archiválií jsou pouze práškové hasicí přístroje,
c. mechanická a elektronická zabezpečovací zařízení umísťovaná na okna a dveře jsou na plášti budovy do výše druhého nadzemního podlaží nebo vyššího podlaží, do něhož by bylo možno vniknout z vodorovných prvků konstrukce budovy, a uvnitř budovy na všech místech, kde se stýkají prostory veřejnosti přístupné a veřejnosti nepřístupné,
d. archivní prostory bez přístupu veřejnosti jsou zajištěny ochranným mechanickým a elektronickým zabezpečovacím zařízením proti přístupu nepovolaných osob a proti násilnému vniknutí,
e. klíče od všech vstupů do archivních prostor jsou uloženy u pověřeného zaměstnance archivu, který je povinen vést evidenci jejich výdeje a vracení, a aby v případech, kdy je vstup do archivních prostor ovládán elektronicky, byla stanovena přístupová práva jednotlivých zaměstnanců archivu,
f. archivní prostory, v nichž jsou uloženy národní kulturní památky, jsou nepřetržitě střeženy.

Paragraf 60 dále přidává podmínky, za kterých vzniká nebo zaniká akreditace k provozování digitálního archivu, což je již nad rámec této metodiky.

Zákon č. 499/2004 Sb. dále odkazuje na prováděcí vyhlášku č. 259/2012 Sb. Tato vyhláška odkazuje na národní standard pro elektronické systémy spisové služby, který požaduje:

a. kontinuální udržování relační a datové integrity,
b. transakční protokol pro auditování,
c. automatické zálohy zachovávající plnou integritu a jejich plnou dokumentaci,
d. zajištění bezpečnosti informačních systémů a jejich komunikace,
e. sledování výskytu chyb na paměťových médiích,
f. schopnost doložit zachování integrity podepsaných dokumentů.

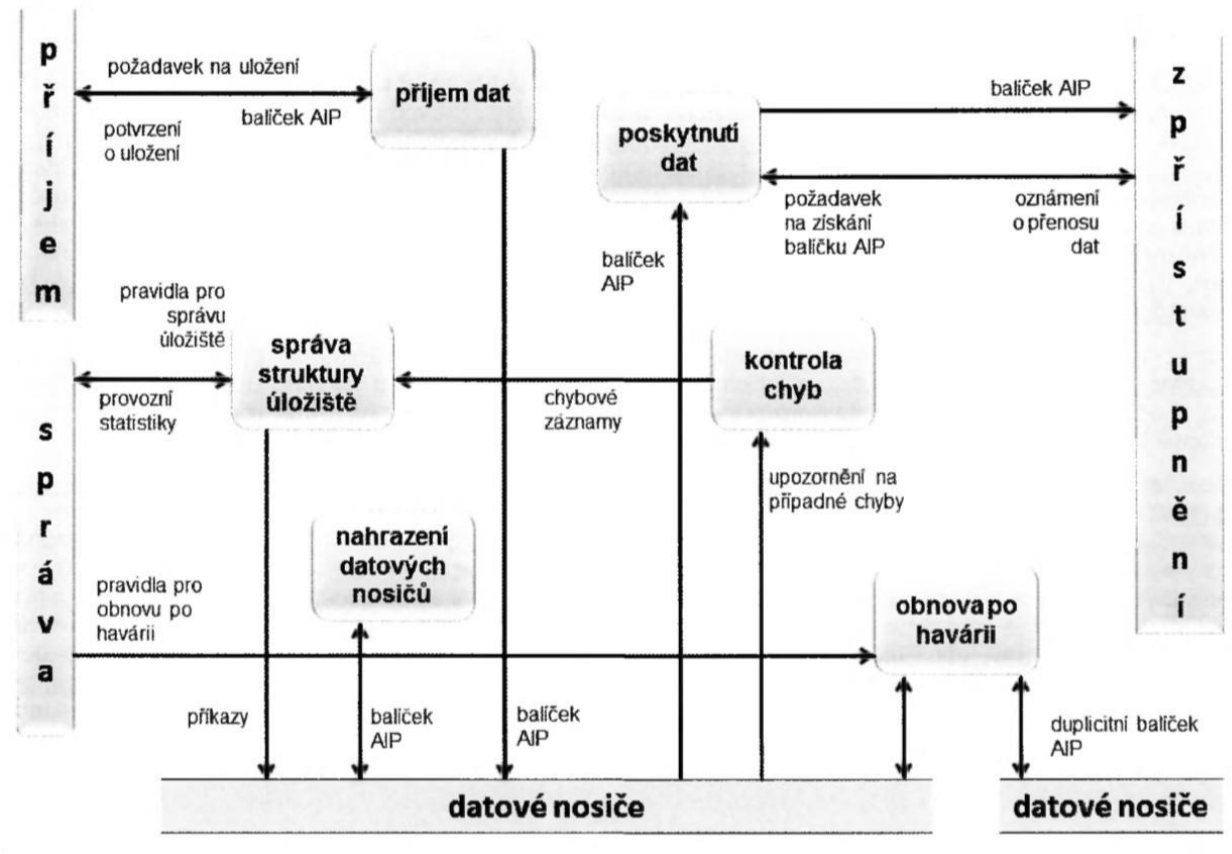
Příloha národního standardu specifikuje datový balíček SIP, který zahrnuje kontrolní součty HAVAL, MD5, SHA-1, SHA-256, SHA-384, SHA-512, TIGER, WHIRLPOOL, CRC32, Adler-32 a MNP.

Věstník ministerstva vnitra čá. 65/2012 (část II) – vzorový provozní řád archivu oprávněného k ukládání archiválií v digitální podobě požaduje:

a. část úložiště – úložiště jsou funkčními celky ukládající AIP balíčky, poskytující balíčky a umožňují jejich administraci; musí existovat popis úložiště včetně mapové lokalizace (pro naplnění požadavku zákona o dislokaci replik), popis musí zahrnovat organizaci úložiště, ze které plyne, jak je garantována integrita ukládaných dat,
b. plánování ochrany – musí existovat popis činností, které zajistí komplexní sledování, uchovávání a vyhodnocování informací, které mohou mít dopad na autenticitu a čitelnost dat. Tento požadavek se dotýká i logické ochrany (sledování trendů použitelných formátů dat jejich převodu),
c. zajištění autenticity archiválií v digitální podobě – musí existovat popis postupu pro zajištění autenticity archiválií v digitální podobě, popřípadě jejich reprodukcí a zároveň pro ověření autenticity archiválií.
d. uložení archiválií – musí existovat popis postupů ukládání AIP balíčků do úložišť a na technické nosiče dat a dále postupů pro ověřování bezpečnosti a neměnnosti takto uložených dat
e. strategie řešení krizových situací – musí existovat popis možných rizik a postupů jejich eliminace nebo řešení.

1.3 Požadavky referenčního rámce OAIS

Referenční rámec [OAIS] je základním a klíčovým dokumentem a mezinárodním standardem v oblasti *Long-term Digital Preservation*. Shrnuje obecné principy logické dlouhodobé ochrany, avšak nedefinuje přesné způsoby technického řešení.



Obrázek 2: Funkční entita Archivní uložení dle OAIS (převzato z [OAIS, Figure 4-3]).

Požadavky OAIS na *Archivní uložení* shrnuje Obrázek 2. Definované funkce se dají stručně shrnout takto:

1. **Příjem dat** (*Receive Data*) – Přijímá *Archival Information Package* (AIP), což je z hlediska *Archival Storage* nejmenší vstupní datová jednotka. AIP jsou následně uloženy na úložné médium a po dokončení přenosu je odeslána volající straně potvrzující zpráva o uložení.
2. **Správa úložné hierarchie** (*Manage Storage Hierarchy*) – Zajišťuje umístění AIP na konkrétní úložná média dle požadavků na AIP aplikované úložné politiky, tj. z hlediska provozních charakteristik média, požadavků na redundanci, zabezpečení apod. Důležitou funkcí je také sledování integrity AIP v rámci přenosů a reporting provozních statistik archivního úložiště administračnímu modulu systému.
3. **Nahrazení datových nosičů** (*Replace Media*) – Funkce zajišťující možnosti výměny úložného média bez poškození, ztráty nebo změny chráněného obsahu uložených AIP. Tj. zajišťuje možnost změny úložných technologií bez ztráty chráněných dat.
4. **Kontrola chyb** (*Error Checking*) – Funkce zajišťující statisticky přijatelnou jistotu, že žádná součást AIP není v *Archival Storage* v průběhu času poškozena ani při uložení, ani při interním přenosu. Nutné je proto shromažďovat informace z kontrolních mechanismů všech zapojených částí systému a výsledky reportovat obsluze systému. Dostupné by měly být i funkce pro selektivní kontrolu integrity datových objektů.
5. **Obnova po havárii** (*Disaster Recovery*) – Funkce poskytující mechanismus duplikace obsahu archivu pro uložení na jiném médiu / v jiné úložné lokalitě.

6. **Poskytnutí dat** (*Provide Data*) – Funkce přijímá žádosti o poskytnutí AIP a na jejich základě dané AIP zpřístupní na žádaném typu média nebo zkopírováním na dočasné úložiště. Když je kopie AIP připravena, je o tom volající strana zpravena potvrzovací zprávou.

Pro podrobnější popis viz [ČSNISO14721].

1.4 NDSA Levels of Digital Preservation

National Digital Stewardship Alliance (NDSA, viz [NDSA]) byla založena jako součást archivačního programu americké *Library of Congress* [LoC]. V roce 2012 vydala dokument *NDSA Levels of Digital Preservation: Release Candidate One* [NDSAlevelsRC1], který má organizacím sloužit jako základní pomůcka pro zvládnutí a předcházení rizik dlouhodobého uchovávání digitálních dat. Nedlouho poté byl vydán také doprovodný materiál [NDSAlevelsV1], který zasazuje jednotlivé požadavky do širšího kontextu a poskytuje jejich podrobnější vysvětlení.

NDSA Levels of Digital Preservation je živý dokument, na jehož úpravách se stále průběžně pracuje. V roce 2016 došlo k návrhu na rozšíření o oblast *Access* [NDSAlevels2016]. Tato podoba *NDSA Levels of Digital Preservation* je přiložena v příloze B.

Tento dokument je významný tím, že vzniká pod vedením organizace založené *Library of Congress* (Kongresovou knihovnou USA) [LoC], která sdružuje několik stovek organizací[8] napříč celým územím USA. Jedná se tak o významnou instituci v oblasti *digital preservation*.

Na rozdíl od referenčního modelu [OAIS] popisujícího problém uchovávání digitálních dat v plné šíři a v obecnosti, se *NDSA Levels of Digital Preservation* soustředí jen na několik technických oblastí. Pro každou z nich pak definuje postupy pro dosažení jedné ze čtyř úrovní ochrany – od základní ochrany dat až po schopnost v datech detekovat výskyt chyb a být schopen chyby opravit. Jedná se tedy o krok od obecného rámce OAIS k metodice praktické implementace archivačního systému.

Při budování dlouhodobého archivního úložiště je proto vhodné definovat, na kterou z uvedených úrovní ochrany systém cílí, a následně při návrhu úložiště zohlednit doporučení, která dané úrovni odpovídají.

1.5 Preservation Storage Criteria

V rámci workshopu konference iPRES 2016 [WiPS] vznikla první verze dokumentu *Preservation Storage Criteria* [PSCv1], která byla dále prezentována na konferenci [PSC] hostované *Library of Congress*. Na základě veřejné diskuse byla později aktualizována do druhé verze dokumentu *Preservation Storage Criteria* [PSCv2].

V rámci zájmové skupiny, která kolem dokumentu vznikla (veřejná diskuse je možná v rámci [DPSG]), je vyvíjena zajímavá sada požadavků na implementaci dlouhodobého úložiště. Sadu doporučení převzatou z [PSCv2] naleznete v příloze A.

Současná sada 58 doporučení z [PSCv2] je zajímavá zejména proto, že

1. je výsledkem delší diskuse a praktických zkušeností v rámci dané zájmové komunity,
2. je stále ve vývoji, jedná se o živý dokument přizpůsobovaný měnícím se požadavkům na dlouhodobou archivaci,
3. jedná se o praktickou pomůcku směřující konkrétně k implementaci archivního úložiště pro dlouhodobou ochranu dle požadavků OAIS.

Na rozdíl od obecného [OAIS] je dokument *Preservation Storage Criteria* zaměřen specificky na budování archivního úložiště. Stejně jako v případě dokumentu *NDSA Levels of Digital Preservation* se jedná o dokument revidovaný komunitou s praktickým zaměřením na danou problematiku v oblasti archivního

[8] http://ndsa.org/members-list/

úložiště. Pokládá tedy konkrétní otázky, jejichž zodpovězení je nutné při budování archivního úložiště brát v úvahu.

Vzhledem k zaměření specificky na OAIS entitu *Archival Storage* jde do větší míry detailu než samotný OAIS, přitom je ale stále dostatečně stručný, aby neodváděl pozornost ke zbytečným detailům a mohl být použit jako užitečná metodická pomůcka při návrhu dlouhodobého archivního úložiště. Při budování dlouhodobého datového úložiště je proto vhodné seznámit se s aktuální podobou těchto pravidel a pro každý z uvedených požadavků zkusit popsat, jak jej budovaný systém naplňuje.

1.6 Dobrá praxe bitové ochrany

Bitovou ochranou se rozumí ochrana datového toku (*bitstream*), tj. binární reprezentace („řetězec jedniček a nul“) dat bez ambicí na interpretaci jejich významu. Cílem je pouze uchovat daný datový tok (datový soubor) v naprosto nezměněné podobě, tj. zachovat jeho integritu.

Obecně se bitová ochrana zakládá na správě více kopií a kontrole jejich integrity. Pro naplnění zákonných požadavků na ukládání archiválií je dále nutné uchovávat auditní informace o manipulaci s daty.

Je tedy nezbytné:

1. Udržovat více kopií (zálohování),
2. využívat systémy a postupy architektonicky navržené jako odolné proti chybám a výpadkům (např. úložiště s vnitřní redundancí, např. RAID pole apod.; datové propojení systémů přes více nezávislých cest; potvrzování kritických akcí – např. mazání – více nezávislými osobami apod.),
3. udržovat kopie v nezávislých lokalitách (ochrana proti živelným katastrofám apod.),
4. využívat více různých úložných technologií (ochrana proti chybám návrhu dané technologie, řešení problémů způsobených mechanickými/chemickými/fyzikálními vlastnostmi úložných médií),
5. zajistit fyzickou bezpečnost (ochrana proti krádeži datových nosičů, odposlechu/modifikaci dat při přenosu, živelným pohromám jako je požár či zátopa),
6. periodicky kontrolovat integritu nezávislých kopií a obnovitelnost záloh (ochrana proti degradaci média, skrytým chybám systému / pracovních postupů apod.),
7. mít celý systém dostatečně dimenzovaný z hlediska kapacity, propustnosti a budoucí rozšiřitelnosti, a to pro všechny funkce, které poskytuje,
8. uchovávat historické verze dat (ochrana proti chybám, např. neúmyslnému smazání, architekturou systému),
9. zajistit správu kopií – vhodným způsobem kopie organizovat a evidovat,
10. mít transakční log pro audit,
11. mít definované a dokumentované návrhy systémů a pracovních postupů (ochrana proti chybám lidského faktoru, zvládání mimořádných situací),
12. mít dlouhodobý plán udržitelného provozu (organizační postupy, lidské zdroje, uchování a předávání znalostí, udržitelné financování, plánovaná pravidelná obnova hardware a software).

1.7 ISO 16363

1.7.1 Část 5 – Infrastruktura a řízení rizik

Norma ISO 16363 „Audit a certifikace důvěryhodných digitálních úložišť“ [ČSNISO16363] nenařizuje institucím konkrétní postupy či nástroje pro dlouhodobou ochranu digitálních dokumentů v rámci digitálních úložišť. Pomáhá ale podchytit veškeré možné oblasti, které mají vliv na plnohodnotné uchovávání digitálních dokumentů. Kapitolu, která řeší převážně technické aspekty spojené s dlouhodobou ochranou (se zaměřením na bitovou ochranu), představuje část 5 – Infrastruktura a řízení rizik. Cílem je:

- porozumět výpočetním a úložným technologiím (*storage technology*), dle kterých je možné identifikovat a ohodnotit základní technické změny pro dopady nebo rizika AIP balíčků,
- dostatečně znát postupy řízení rizik a validovat rizika, která jsou relevantní pro technickou infrastrukturu, stejně tak rizika infrastruktury, která byla identifikována na organizační úrovni,
- identifikovat třídy rizik, která jsou specifická pro úložiště a procesní technologie, které mohou verifikovat plány na zmírnění identifikovaných tříd rizik,
- identifikovat faktory a hrozby bezpečnostních rizik pro správu dat, systémů, personální a fyzická zařízení,
- identifikovat prvky a role při implementaci změn faktorů bezpečnostních rizik,
- vyhodnotit vývoj a implementaci bezpečnosti a plánů managementu rizik.

1.7.2 Vazba na jiné standardy

Série ČSN ISO 27000: Standardy z oblasti bezpečnosti informací. Všechny standardy „rodiny 27k“ mají definovanou jednotnou strukturu a pravidla pro začlenění specifických požadavků.

ČSN ISO 31000: Označení standardu pro Management rizik. Poskytuje návody (obsahuje Principy a směrnice), jak řídit systematickým, transparentním a spolehlivým způsobem různé formy rizik a jak harmonizovat systému řízení rizik do organizace a všech jejích procesů, rozhodování, produktů, služeb a aktiv. Není určena pro účely certifikace.

Výše zmíněné standardy se ale více zaměřují na organizaci jako takovou než na to, co daná organizace spravuje – tedy AIP balíčky jako reprezentace digitálně zakódované informace a podstata každého důvěryhodného digitálního úložiště. Zároveň ISO 16363 zmiňuje rizika a bezpečnost do takové míry, aby důvěryhodná úložiště nebyla nucena podstupovat náročnou a nákladnou certifikaci podle ISO 27000.

1.7.3 Požadavky ISO 16363, část 5 – Infrastruktura a řízení rizik

Požadavky	Možné důkazy	Možné problémy
5.1 Řízení rizik technické infrastruktury		
5.1.1 Úložiště musí zjišťovat a řídit rizika infrastruktury systému vztahující se k provozu a cílům v oblasti uchovávání	● inventář systémových komponentů infrastruktury ● export reálných digitálních objektů na nové nosiče ● opětovné vytvoření archivních kopií ze záloh ● využití některého z podporovaných systémů v rámci určené skupiny (např. Apache, Fedora, iRODS apod.)	● nedostatečné plánování změn infrastruktury ● finanční tíseň ● nedostatečné povědomí o technologiích používaných určenou skupinou (jaké jiné služby by určená skupina ráda využívala?) ● důležité pravidelné získávání a vyhodnocování zpětné vazby
5.1.1.1 Úložiště musí zjišťovat a řídit rizika infrastruktury systému vztahující se k provozu a cílům v oblasti uchovávání (včetně 5.1.1.1.1–5.1.1.1.8)	● pravidelné posudky technologií ● odhad životnosti systémových komponentů ● správa pravidelných zpráv o hodnocení technologií ● srovnání stávajících technologií a jednotlivých nových hodnocení	● zastarávání technologií (HW, SW, formáty; jaký je pro úložiště využíván nejstarší HW, proč je potřebný, je možné jej nahradit, jak rychle? apod.) ● identifikace rizik je subjektivní – auditor může identifikovat jiná rizika než ta, které identifikoval personál úložiště ● výchozí bod hodnocení je OAIS – funkce a samotné AIP balíčky nesmí být ohroženy
5.1.1.2 Úložiště musí mít patřičnou hardwarovou a softwarovou podporu pro zálohování, která bude dostatečná pro uchovávání obsahu úložiště i záznam jeho provozu	● dokumentace toho, co je zálohováno a jak často (pokud určitý výběr častěji nebo ve více kopiích, proč právě daný výběr apod.) ● auditní logy/inventář záloh ● validace kompletnosti záloh ● plán obnovy po katastrofě, politiky a dokumentace ● protipožární cvičení ● testování záloh ● smlouvy o podpoře hardwaru a softwaru pro záložní mechanismy ● schopnost prokázat uchovávání systémových dat (např. kontrola přístupů, lokace replik, auditní stopy, hodnoty kontrolních součtů apod.)	● vytváření nekompletních setů záloh (jaký proces záloh?) ■ ne u všech systémů ■ SW nebo systémová konfigurace není zálohována ■ načasování, „*rotation of media*“ ● nedostatek paměťových médií ● obnovení ze záloh netestováno (jakým způsobem by probíhala obnova digitálního objektu, havárie systému/zařízení a jak dlouho by to trvalo?) ● žádné zálohy mimo pracoviště (kolik kopií záloh je vytvářeno a na jaká média?) ● zabezpečení paměťových médií

5.1.1.3 Úložiště musí mít proces pro zaznamenávání a reagování na dostupnost nových bezpečnostních aktualizací a tento proces musí být založen na posouzení rizik a přínosů (včetně 5.1.1.3.1)	● dokumentace, ve které je specifikována detekce chyb bitů a popis využitých nápravných opatření ● analýza rizik a hrozeb ● reporty chyb ● pravidelná analýza integrity obsahu repositáře ■ postupy při hlášení událostí administrátorům, příp. vedení ■ záznamy o metadatech o uchovávání (např. o PDI) ■ srovnání chybových záznamů vůči zaslaným hlášením administrátorům ■ eskalační postupy při ztrátě dat ■ sledování zdrojů událostí ■ opatření k odstranění zdrojů událostí	● neexistující nebo nevhodné načasování spouštění kontroly pro detekci chyb ■ žádná systematická pravidelná kontrola ● pokrytí a výběr dat pro kontrolu (primárně obsah, metadata, dokumentace) ● typy/funkčnosti mechanismů zjišťování chyb (jaké způsoby kontroly chyb úložiště využívá) ● reportování/řešení chyb (co se stane, pokud je nalezena chyba?, provádějí se kontroly i na dalších kopiích?, jsou vytvářeny příslušné logy z těchto kontrol?, kdo tyto logy prochází a kontroluje? a co konkrétně se v logu/reportech hledá?) ● zabezpečení skriptů zjišťování chyb
5.1.1.4 Úložiště musí zaznamenávat všechny případy poškození nebo ztráty dat, ohlašovat je vedení a podnikat kroky k opravě/náhradě poškozených nebo ztracených dat	● registr rizik (seznam všech dostupných softwarových záplat a analýza dokumentace rizik) ● důkazy provádění pravidelných aktualizací (např. program pro správu aktualizací) ● dokumentace vztahující se k instalacím aktualizací	● žádná služba/program k identifikaci a poskytování aktualizací ● chybějící identifikace kritického hardwaru, softwaru, procesů ● neexistující postupy k aktualizacím, které mají být nainstalovány ■ žádný postup/sledování, které zabezpečí, že aktualizace jsou instalovány podle plánu (kterých systémů/SW se aktualizace týkají?, co se stane pokud by aktualizace nebyly nainstalovány?, kdy jsou aktualizace považovány za úspěšně nainstalované?)

5.1.1.5 Úložiště by mělo mít stanoveny procesy pro výměnu datových nosičů a/nebo hardwaru (např. obnovu, přesun)	● dokumentace postupů při migracích a přesouvání dat (HW, SW, média, formáty) ● pravidla a politiky vztahující se k hardwarové podpoře, údržbě a náhradě ● dokumentace popisující očekávanou dobu podpory ze strany výrobce hardware ● politiky související s přesunem dat do jiných hardwarových systémů	● povědomí o životnosti HW, SW, formátů a médií (existují záznamy kontrol, které by životnost potvrdily nebo vyvrátily?) ● existence „*orphaned media*“ (např. soubory s nulovou velikostí, neplatné odkazy apod.) ● slabé povědomí o typech potřebných aktualizací (kopírování, obnovení, emulace, migrace) ● zničení/ztráta originálů před ověřením nových kopií/verzí ■ nutné ověření nových kopií/verzí
5.1.1.6 Úložiště musí mít stanovené a zdokumentované klíčové procesy, které ovlivňují jeho schopnost dodržovat závazné povinnosti (včetně 5.1.1.6.1–5.1.1.6.2)	● matice vzájemných vztahů („*traceability matrix*“) mezi procesy a závaznými požadavky, která zajišťuje zpětnou dohledatelnost ■ dokumentace procesu řízení změn ■ posouzení rizika spojeného se změnou procesu ■ analýza očekávaného dopadu změny procesu ■ porovnání záznamů (logů) skutečných změn procesů se souvisejícími analýzami jejich dopadu a důležitosti ● dokumentace testovacích postupů ● dokumentace výsledků předchozích testů a dokumentace změn provedených na základě těchto testů ● analýza dopadů u změněných procesů	● možné problémy při identifikaci kritických procesů (jakým způsobem jsou identifikovány?, jaké funkce úložiště by byly ztraceny, pokud by fungovaly pouze tyto procesy?) ● identifikace veškerých procesů po změnách, aktualizacích a jaký efekt budou mít ● řízení, sledování, vyhodnocování, reportování implementovaných změn (jakým způsobem byly plánovány, testovány, implementovány, ověřeny, reportovány změny a aktualizace kritických procesů? apod.) ● nízká schopnost obnovy po špatně provedených aktualizacích

5.1.2 Úložiště musí spravovat počet a umístění kopií všech digitálních objektů (včetně 5.1.2.1)	● testy nahodilého vyhledávání a získání digitálních objektů ● ověření výskytu objektů pro každou zaznamenanou lokaci ● ověření zaznamenané lokace pro každý objekt v systémech úložiště ● seznam/záznam umístění digitálních objektů srovnaný s očekávaným počtem a umístěním kopií konkrétních digitálních objektů ● pracovní postupy pro synchronizaci („synchronizační workflows“) ■ systémová analýza doby trvání synchronizace kopií ■ postupy/dokumentace synchronizačních procesů	● nedostatečný počet potřebných kopií ● chybějící umístění/úložiště pro každou z kopií ● nejasný, problematický přístup/zabezpečení ke každé z těchto kopií ● žádné ověření existence, synchronizace či úplnosti všech kopií ● rozdílná pravidla pro rozdílné sbírky/typy informací, které mohou ztížit pracovní postupy pro synchronizaci
5.2 Řízení bezpečnostních rizik		
5.2.1 Úložiště musí zajišťovat systematickou analýzu bezpečnostních rizik, která souvisí s daty, systémy, zaměstnanci a fyzickým zařízením	● analýza faktorů bezpečnostních rizik spojených s daty, systémem, personálem a fyzickým prostředím ● repositář má zavedený kód praxe („*codes of practice*“) podle normy řady ISO 27000 ● seznam pro řízení systému ● analýza rizik, hrozeb a způsobu řízení ● finanční audity ● přidání kontrol založených na průběžném zjišťování a hodnocení rizik (spoléhá úložiště na jiné instituce ve věci dlouhodobé ochrany?, mají k dispozici plán zmírňování, který ochrání vaše AIP balíčky?)	● neexistence lokálního registru rizik ● žádné řazení rizik v registru ■ hodnocení rizik je subjektivní (viz 5.1.1.1) ● žádný plán řešení závažných rizik ● žádné pravidelné / žádné revize a aktualizace registru rizik
5.2.2 Úložiště musí mít zaveden způsob řízení, který patřičně řeší každé z vymezených bezpečnostních rizik	● úložiště využívá postupy podle norem z řady ISO 27000 (resp. ISO 27002) ● úložiště provádí certifikaci podle ISO 27000 ● seznam pro řízení systému ● analýza rizik, hrozeb a způsobu řízení	● viz 5.2.1

5.2.3 Zaměstnanci úložiště musí mít vymezené role, odpovědnosti a oprávnění vztahující se k provádění změn v systému	● úložiště využívá postupy podle norem z řady ISO 27000 ● organizační schéma (v centru stojí AIP – je důležité znát jejich proces tvorby/uchování a zálohování; kdo může provádět jednotlivé kroky v rámci těchto procesů?, existuje nějaká dokumentace těchto kroků?, může tyto kroky provádět i někdo jiný?, jak dlouho se tato dokumentace uchovává?, kdo má přístup k této dokumentaci?, kdo zodpovídá za jejich revizi?) ● dokumentace oprávnění systému (kdo z personálu je v pozici, že by mohl ohrozit proces ochrany obsahu repositáře?) ● úložiště provádí certifikaci podle ISO 27000	● nejasně definované role, oprávnění a procesy v rámci digitálního úložiště ■ role vs. konkrétní zaměstnanci ● chybějící akční protokoly („*action protocols*“) a zabezpečení těchto protokolů
5.2.4 Úložiště musí mít vhodné písemně zaznamenaný plán (plány) připravenosti na živelnou pohromu a na obnovu po havárii, a to včetně alespoň jedné zálohy všech uchovávaných informací mimo pracoviště a jedné kopie plánu (plánů) na obnovu po havárii uložené mimo pracoviště	● písemné plány připravenosti a plány na obnovu, včetně alespoň jedné zálohy veškerého obsahu plánů na obnovu po havárii mimo pracoviště ■ informace a důkaz o alespoň jedné kopii veškerého uchovávaného obsahu mimo pracoviště ● úložiště využívá postupy podle norem z řady ISO 27000 ● úložiště provádí certifikaci podle ISO 27000 ● plán nepřetržité služby ● dokumentace rolí a jejich činností	● nepřipravenost na možné hrozby a neexistence plánu na obnovu po havárii (potřeby úložiště vs. neadresovaná rizika) ● závažná rizika nejsou řešena v plánech ● žádné revize a aktualizace plánů ● personál není vyškolen k implementaci daných plánů ● nedostatečné zabezpečení plánu

2 Implementace v systému ARCLib

ARCLib je systém pro logickou a bitovou ochranu digitálních dat navržený v souladu s požadavky odvozenými z normy ČSN ISO 14721 (OAIS, [OAIS]). Jejich naplňování je popsáno dále v této sekci. Systém ARCLib umožňuje institucím implementovat funkční moduly OAIS a jeho informační modul. Systém ARCLib byl navržen jako *temný archiv* (*dark archive*) – není určen ke zpřístupňování dokumentů koncovým uživatelům. ARCLib byl projektován s ohledem na digitalizovaná data a potřeby knihoven v oblasti dlouhodobé digitální archivace knihovních dat.

Fyzické ukládání dat v rámci systému ARCLib obstarává komponenta *ARCLib Archival Storage*.

2.1 ARCLib Archival Storage

Komponenta *ARCLib Archival Storage* je komplexní služba pro zajištění bitové ochrany umožňující využití replikace dat do více geografických lokalit a využití více technologií ukládání dat. *Archival Storage* poskytuje směrem k ARCLib objektové úložiště (*object storage*) použitelné přes jednoduché REST rozhraní.

REST rozhraní poskytuje mimo jiné i funkce pro řízení přístupu k datům, tj. poskytuje např. rozhraní pro autentizaci klienta služby, na základě čehož je možné rozlišovat různé uživatele a řídit tak přístup k datům. *ARCLib Archival Storage* také vede podrobné protokoly o přístupu k datům a manipulaci s nimi.

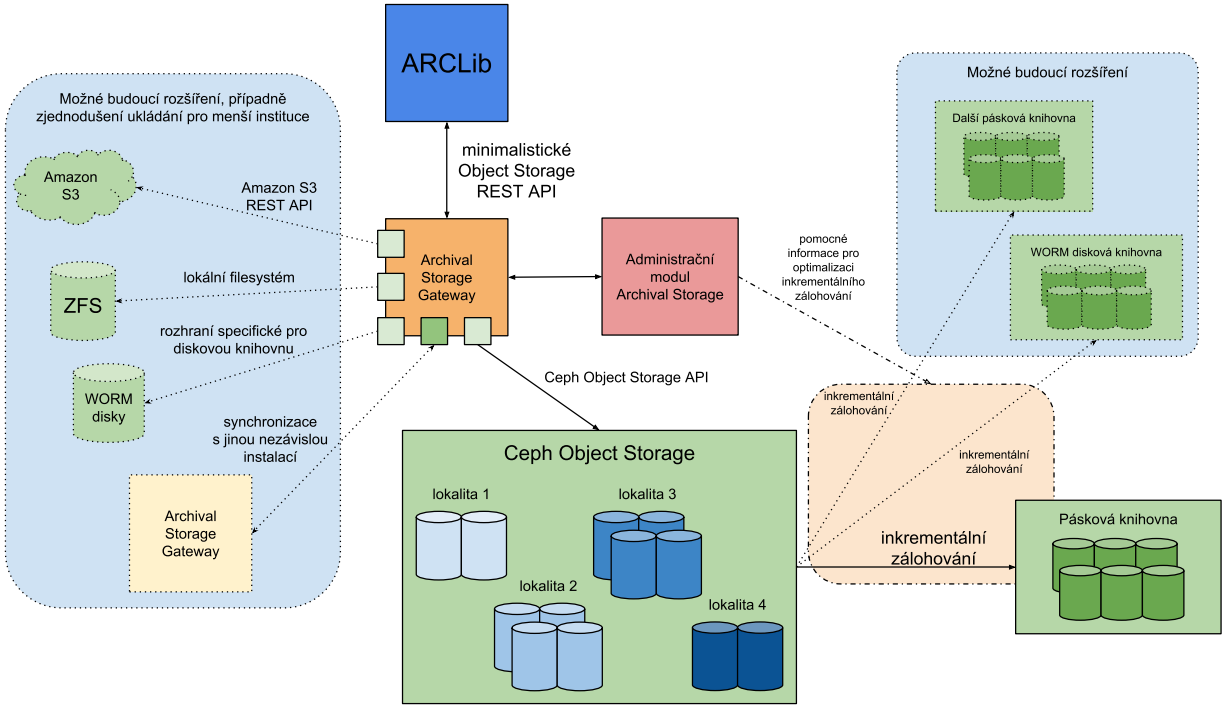
Směrem k úložišti ukládá tento modul data v definované struktuře na různé typy úložišť, např. distribuovaný souborový systém, do *Ceph Object Storage* nebo jiné vhodné technologie, která umožňuje dosažení cílů *Archival Storage* z hlediska bezpečnosti uložení dat a manipulace s nimi.

Důležitou vlastností *ARCLib Archival Storage* je návrh s ohledem na výkon a jeho škálování při masivně paralelním přístupu k úložišti.

2.2 Koncepce ARCLib Archival Storage

ARCLib Archival Storage je modul zastřešující jednotný způsob ukládání balíčků AIP na fyzická úložiště různých typů a technologií a přístup k těmto uloženým datům. Modul přijme od systému ARCLib balíček AIP a potvrdí jeho úspěšné uložení až poté, co se jeho kopie bezpečně uloží na každé z připojených úložišť, která modul obsluhuje.

Modul *ARCLib Archival Storage* je vhodný k dalšímu použití i mimo projekt ARCLib, je od zbytku systému oddělen a přistupuje se k němu jako k samostatně spustitelné aplikaci. Tato aplikace je schopna poskytovat zejména ukládání a vydávání balíčků AIP, reportování o jejich stavu a stavu všech fyzických úložišť, která jsou k ní připojena.



Obrázek 3: Schéma ARCLib Archival Storage.

Vysokoúrovňovou architekturu *ARCLib Archival Storage* znázorňuje Obrázek 3. V kontextu projektu ARCLib je podporováno několik technologií fyzických úložišť, kromě běžného souborového systému (zejména pro testovací účely) především *ZFS* a *Ceph*. Logická úložiště jsou na Obrázku 3 vyznačena zeleně.

2.2.1 Storage Service Gateway

Storage Service Gateway představuje obecné abstraktní rozhraní komponenty *ARCLib Archival Storage* pro ukládání balíčků AIP do datového úložiště a pro komunikaci s ním. Pro jednotlivé typy datových úložišť pak existuje implementace jednotlivých metod tohoto rozhraní, především pro:

- Uložení AIP na úložiště,
- načítání AIP z úložiště,
- zjištění stavu AIP na úložišti (kontrola existence a neporušenosti, *fixity*),
- zjištění stavu úložiště a všech jeho uzlů a
- automatické spuštění procesu pro smazání nekompletních AIP z úložiště, např. po pádu a restartu *ARCLib Archival Storage*.

Mazání nekompletních AIP je řešeno ukládáním identifikátorů zpracovávaných balíčků v databázi a jejich mazáním po úspěšném uložení. Identifikátory, které budou po startu *ARCLib Archival Storage* uloženy v databázi, nebyly úspěšně uloženy a je potřeba je ve všech úložištích smazat.

Storage Service Gateway komunikuje přes REST API rozhraní, které poskytne služby pro:

- Uložení AIP na úložiště,
- načítání AIP z úložiště,
- zjištění stavu AIP na úložišti (kontrola existence a *fixity*) a
- zjištění stavu úložiště a všech jeho uzlů.

Kromě toho API poskytne služby pro řízení přístupu k datům. To je zabezpečeno specifikováním autentizačních údajů klienta služby (minimálně na úrovni původců dat) např. ve formě API key a API secret,

na základě kterých je možné autentizovat a posléze autorizovat uživatele *ARCLib Archival Storage* k práci s daty.

ARCLib Archival Storage podrobně loguje záznamy o přístupech k datům a o manipulaci s nimi do auditního logu.

2.2.2 Adaptéry

Pro každý typ datového úložiště je implementován tzv. *Adaptér*, který implementuje jednotlivé metody rozhraní definované v rámci *Storage Service Gateway*; především jsou to funkce pro:

- Uložení AIP na úložiště,
- načítání AIP z úložiště,
- zjištění stavu AIP na úložišti,
- zjištění stavu úložiště a všech jeho uzlů,
- automatické spuštění rutiny pro smazání nekompletních AIP z úložiště, např. po pádu a restartu *ARCLib Archival Storage*.

2.2.3 Databáze

V databázi *ARCLib Archival Storage* jsou uloženy především transakční informace o aktuálně ukládaných balíčcích AIP.

2.3 Naplnění požadavků OAIS

2.3.1 Příjem dat

Příjem dat je zajišťován komponentou *ARCLib Archival Storage Gateway*, která představuje jediný komunikační bod mezi archivním úložištěm a jeho uživateli. Komunikace, tj. zejména příjem balíčků AIP a zpětná notifikace o jejich uložení, probíhá prostřednictvím minimalistického rozhraní REST API, které uživatele archivního úložiště kompletně odstiňuje od použitých konkrétních úložných technologií. Přes toto rozhraní je možné poskytovat libovolné informace o stavu úložiště v definovaném tvaru bez ohledu na komunikační protokoly a stavové informace specifické pro používaná fyzická úložiště.

2.3.2 Správa úložné hierarchie

O správu celé úložné infrastruktury *ARCLib Archival Storage* se stará komponenta *ARCLib Archival Storage Gateway* ve spolupráci s administračním modulem *ARCLib Archival Storage*, které rozhodují o umístění AIP na jednotlivá fyzická úložná média dle zvolené úložné politiky. Zde jsou také řízeny pravidelné a mimořádné kontroly integrity uložených dat a sledování stavu všech komponent úložiště.

2.3.3 Náhrada médií

Náhrada médií se děje na dvou úrovních:

1. Výměna komponent v rámci jednoho logického úložiště, tj. na „hardwarové vrstvě“ (např. výměna fyzického disku v rámci ZFS úložiště).
2. Výměna logického úložiště.

Výměna komponent v rámci logického úložiště na hardwarové úrovni je záležitostí vnitřního fungování daného úložiště a je z hlediska *ARCLib Archival Storage* transparentní. Použitá úložiště musí poskytovat dostatečnou míru vnitřní redundance (např. uspořádáním více fyzických disků do nějaké formy RAID apod.), aby výpadek jednoho fyzického komponentu (např. disku) nevedl ke ztrátě základní funkčnosti úložiště.

Výměna logického úložiště je transparentní z hlediska uživatelů *ARCLib Archival Storage*. Při migraci je nejdříve přidáno nové logické úložiště, kam je synchronizován obsah původního logického úložiště. Původní logické úložiště je následně odstraněno. V průběhu převodu jsou uložená data vybavována z původního logického úložiště a nová data jsou zapisována na původní i nové logické úložiště.

2.3.4 Kontrola chyb

Kontrola chyb se skládá z jejich detekce a případné opravy. Pro detekci chyb se používají kontrolní součty[9]. Opravy se provádí obnovou z jiné kopie daného datového objektu.

Kontrola integrity AIP se provádí ověřením kontrolního součtu udaného vkladatelem, který je zároveň zaznamenán v samotném AIP. Kontrolní součet v AIP může být využit také pro selektivní kontrolu libovolného AIP v archivním úložišti.

Aby se předcházelo nutnosti provádět časté kontroly integrity dat vyčítáním AIP z archivního úložiště, což by bylo při očekávaných objemech dat nepraktické (pomalé a s vysokými nároky na přenosovou kapacitu i výpočetní výkon), počítá se s použitím logických úložišť, která sama o sobě dokáží provádět pravidelnou, průběžnou a distribuovanou kontrolu integrity přenosu a uložení dat (*ZFS/Ceph + scrubbing* apod.).

Výsledky kontrol jsou definovaným rozhraním reportovány uživatelům *ARCLib Archival Storage*.

2.3.5 Obnova po katastrofě

ARCLib Archival Storage počítá s užitím minimálně dvou logických úložišť v různých lokalitách obsahujících kompletní sadu ukládaných AIP. Mechanismem zotavení po katastrofě je pak použití kopií AIP z logických úložišť, která nebyla katastrofou zasažena.

2.3.6 Poskytnutí dat

Data jsou poskytována prostřednictvím komunikačního rozhraní komponenty *ARCLib Archival Storage Gateway*, obdobným způsobem, kterým jsou na archivní úložiště ukládána.

2.4 Naplnění požadavků Preservation Storage Criteria

Jak je uvedeno v teoretické části metodiky v sekci *Preservation Storage Criteria*, je sada doporučení *Preservation Storage Criteria* [PSCv2] komunitním živým doporučením pro praktickou implementaci archivního úložiště dle požadavků OAIS. Na rozdíl od obecného OAIS se přitom zaměřuje právě na archivní úložiště a rozpracovává požadavky na ně do vyšší míry detailů. Ve zbytku této sekce metodiky je proto popsáno, jak jsou požadavky *Preservation Storage Criteria* naplňovány v implementaci systému ARCLib.

2.4.1 Integrita obsahu (Content integrity)

2.4.1.1 Provádí se kontrola integrity (Provides integrity checks)

Kontrola integrity dat v archivním úložišti se provádí:

1. Při vstupu AIP do archivního úložiště – po předání dat od vkladatele na vstup *ARCLib Archival Storage Gateway* je ověřen kontrolní součet doručeného AIP balíku oproti očekávanému kontrolnímu součtu, který je vkladatelem předán společně s AIP balíkem. S balíkem se dále nepracuje, pokud kontrolní součet vytvořený systémem neodpovídá součtu poskytnutým vkladatelem.
2. Po předání na každé z logických úložišť – obdobně jako na vstupu do archivního úložiště, tak i na vstupu každého logického úložiště je po přenosu dat ověřena integrita doručeného balíku před jeho uložením na každém z logických úložišť. V případě výskytu chyby je vyžádán nový přenos dat.
3. Před vydáním AIP z archivního úložiště – těsně před odesláním AIP balíku žadateli je ověřen kontrolní součet AIP, aby se ověřilo, že jsou data na začátku transportu v pořádku. V případě výskytu chyby jsou data získána z jiné kopie (jiného logického úložiště).
4. Na vyžádání – pokynem operátora je možné vyvolat kontrolu integrity libovolného AIP v libovolném logickém úložišti. Kontrola se provádí vyčtením AIP z daného úložiště a ověřením jeho kontrolního

[9] Základním cílem kontrolního součtu je detekování chyb v datech. K samotným datům se připojuje doplňková informace (výsledek určité operace), která je vypočítána ze vstupních dat. Tu poté příjemce dat může přepočítat a zkontrolovat tak správnost dat a úplnost dat.

součtu. V případě nalezení chyby je defektní kopie nahrazena korektní kopií získanou z jiného logického úložiště.

5. Automaticky průběžně v rámci jednotlivých úložišť – jednotlivá logická úložiště jsou budována pomocí úložných technologií, které umožňují provádět pravidelnou, průběžnou a distribuovanou kontrolu integrity přenosu a uložení dat.
   Konkrétně předpokládáme použití souborového systému *ZFS*, respektive *Ceph*, které samy o sobě drží více kopií všech dat a automatizovaně zajišťují opravu poškozených kopií z kopií korektních při zjištění chyby. Je přitom snadné provádět automatizovanou, pravidelnou a častou kontrolu na výskyt chyb, neboť tyto úložné technologie integrují kontrolní funkce (*scrubbing*), které takovou kontrolu automatizovaně provedou. A to distribuovaně přes jednotlivé hardwarové prvky úložiště, tj. maximálně efektivně.

Automatické distribuované kontroly přímo na logických úložištích jsou velmi důležité, neboť výrazně zvyšují bezpečnost dat bez nutnosti vyčítání AIP z úložiště za účelem kontroly integrity, což by mělo vysoké nároky na přenosovou i výpočetní kapacitu na řídícím bodu úložiště, a v případě uložení většího množství dat by tedy nebylo možné provádět kontroly s potřebnou frekvencí.

O všech výše uvedených činnostech jsou vedeny auditní logy ve všech komponentách archivního úložiště, které se činnosti účastní. Logy se následně definovaným rozhraním předávají do modulu *ARCLib Administration* pro jejich centrální evidenci.

2.4.1.2 Je podporována nezávislá kontrola integrity (Supports independent integrity checks)

Všechny AIP balíčky v úložišti mají integrovány kontrolní součty. Na pokyn operátora je možné vyvolat kontrolu integrity libovolného AIP v libovolném logickém úložišti. Kontrola může být vyvolána i na požadavek třetí strany.

O činnost systému ARCLib jsou vedeny auditní logy, které mohou být ověřeny třetí stranou.

O všech operacích s AIP balíčkem jsou vedeny záznamy (viz [MLO, str. 43]), které mohou být třetí stranou ověřeny.

V případě potřeby je možné do AIP integrovat digitální podpis jeho obsahu, kde by podepisující stranou byl např. přímo vlastník/dodavatel AIP a provozovatel instalace ARCLib by neměl přístup k příslušnému soukromému klíči. Kontrola integrity obsahu by pak byla možná ověřením platnosti daného digitální podpisu obsahu AIP.

2.4.1.3 Je podporováno provádění akcí pro uchování dat (Provides preservation actions)

Problematika je řešena v modulu *ARCLib Preservation Planning*, viz Metodika logické ochrany [MLO].

2.4.2 Zohlednění nákladů (Cost considerations)

2.4.2.1 Ekonomicky efektivní (Cost-efficient)

Architektura systému ARCLib je od prvotního návrhu stavěna velmi modulárně. I samotné úložiště *ARCLib Archival Storage* je navrženo modulárním způsobem a nezávisle na konkrétní úložné technologii. Náklady je proto možné snadno optimalizovat použitím různého počtu logických úložišť a výběrem technologií jejich implementace. Je možné použít vše od lokálního souborového systému po ukládání dat do *cloudu* u libovolného poskytovatele. Pro zapojení nového typu úložné technologie stačí do systému doprogramovat adaptér pro dané úložiště implementující rozhraní definované *ARCLib Archival Storage Gateway*.

2.4.2.2 Šetrný k životnímu prostředí (Energy-efficient)

Viz předchozí bod *Ekonomicky efektivní*.

2.4.2.3 Hmotnost úložiště (Storage weight)

Viz bod *Ekonomicky efektivní*.

2.4.3 Přizpůsobitelnost a odolnost (Flexibility & resilience)

2.4.3.1 Vysoká odolnost (High resilience)

Architektura *ARCLib Archival Storage* je založena na použití více logických úložišť, kde každé z těchto logických úložišť obsahuje kompletní sadu uložených AIP balíků, tj. každé z logických úložišť obsahuje všechna chráněná data. Dostupnost libovolného z logických úložišť tak zachovává dostupnost všech dat.

Každé z logických úložišť může být realizováno jinou úložnou technologií, a tento přístup je doporučován. Případné systémové problémy jedné konkrétní technologie tak neohrožují ostatní kopie dat a jejich dostupnost.

V rámci ARCLib se předpokládá, že logická úložiště jsou budována pomocí úložných technologií, které samy o sobě poskytují redundantní uložení dat a umožňují provádět pravidelnou, průběžnou a distribuovanou kontrolu integrity přenosu a uložení dat a automatické zotavení z automaticky detekovaných chyb. Příkladem může být použití souborového systému *ZFS*, či distribuovaného objektového úložiště systému *Ceph*. *ZFS* poskytuje redundanci dat ukládáním více kopií na více fyzických disků. *Ceph* je pak úložiště distribuované přes více disků či serverů, možná je i varianta geografické replikace mezi více *Ceph* clustery, které jsou jako celek považovány systémem ARCLib za jediné logické úložiště.

Výše uvedené poskytuje redundanci na mnoha úrovních. Velmi podstatný je také aspekt automatické průběžné detekce chyb a zotavení se z nich přímo na úrovni automatické práce úložiště bez nutnosti zásahu operátora systému. Jako celek se proto jedná o velmi robustní řešení odolné proti selhání jednotlivých částí.

2.4.3.2 Vysoká dostupnost (High availability)

Architektura systému ARCLib je od prvotního návrhu stavěna velmi modulárně s důrazem na paralelní zpracování a eliminaci společného bodu selhání. Většina komponent je navržena jako bezstavová, tj. v případě selhání je možné rychle spustit jejich jinou instanci, která převezme její funkce s minimálním dopadem na probíhající operace.

Samotná data jsou ukládána na logická úložiště, která je možné implementovat pomocí různých technologií. Jednou z očekávaných implementací je objektové úložiště technologie *Ceph*, které je samo o sobě navrženo s cílem zajistit vysokou dostupnost, tj. nedostupnost některých uzlů *Ceph* clusteru neohrožuje funkčnost úložiště jako celku.

Pro prostředí s nižšími nároky na dostupnost je možné použít i jednodušší typy úložných technologií (např. lokální souborový systém), respektive nenasazovat duplicitní instance ostatních komponent systému.

2.4.3.3 Zotavení (Recovery)

ARCLib poskytuje redundanci na úrovni logických úložišť i v rámci logických úložišť. Chyby v rámci logického úložiště jsou zpravidla detekovány a řešeny automaticky bez zásahu operátora, viz sekce *Přizpůsobitelnost a odolnost | Vysoká dostupnost* výše.

V ostatních případech bude postup obnovy dokumentován v rámci dokumentace software ARCLib (pokud se jedná o použití standardních funkcí systému), nebo musí být dokumentován přímo provozovatelem ARCLib v konkrétním místě nasazení, pokud závisí na lokální konfiguraci logických úložišť a nastavení politik jejich použití.

2.4.3.4 Navrženo pro nulovou ztrátu dat (Designed for zero data loss)

Viz sekce *Přizpůsobitelnost a odolnost | Vysoká dostupnost* výše.

2.4.3.5 Nezávislost (Independence)

Systém ARCLib je od počátku navržen jako vysoce modulární. Nezávislosti komponenty *ARCLib Archival Storage* byla věnována zvláštní péče, neboť se předpokládá její nasazení jako datového úložiště i pro jiné systémy než samotný ARCLib.

S *ARCLib Archival Storage* jeho uživatel komunikuje výhradně přes definované REST rozhraní objektového úložiště s minimální nutnou množinou operací (ulož soubor, získej soubor, …), čímž je uživatel plně odstíněn od detailů ovládání použitých fyzických úložišť.

Naopak *ARCLib Archival Storage* nevyžaduje žádnou znalost o funkcích systému, který jej používá. Fyzicky jsou v rámci *ARCLib Archival Storage* data ukládána na nezávislá logická úložiště, jež je také snadné zaměnit (viz sekce *Koncepce ARCLib Archival Storage*, *Zohlednění nákladů | Ekonomicky efektivní* a *Přizpůsobitelnost a odolnost | Vysoká dostupnost* výše).

2.4.3.6 Bezvýpadkové migrace úložiště (Nondisruptive storage migrations)

Modulární architektura ARCLib předpokládá výměnu různých svých komponent. Výměna úložiště je možná i bez přerušení provozu, viz sekce *Design ARCLib Archival Storage* a *Naplnění požadavků OAIS | Náhrada médií* výše.

2.4.3.7 Integrovatelné (Integratable)

Návrh *ARCLib Archival Storage* od počátku počítá s použitím této komponenty nezávisle od zbytku ARCLib i s jiným systémy, viz sekce *Přizpůsobitelnost a odolnost | Nezávislost* výše.

2.4.3.8 S otevřeným zdrojovým textem (Open source)

Použití open-source technologií a implementace všech komponent ARCLib pod open-source licencí bylo z důvodu udržitelnosti a možnosti širokého nasazení výsledného software jedním ze základní požadavků na vývoj ARCLib.

2.4.3.9 Podpora více protokolů pro přístup k souborovým systémům (Supports multiple file system protocols)

Úložiště ARCLib Archival Storage je plně modulární a použití různých úložných technologií je jedním z jeho základních rysů, viz sekce *Zohlednění nákladů | Ekonomicky efektivní* výše.

2.4.3.10 Podpora různých úložných médií (Diverse storage media types)

Použití různých úložných technologií je jedním ze základních rysů architektury systému ARCLib, viz sekce *Zohlednění nákladů | Ekonomicky efektivní* a *Přizpůsobitelnost a odolnost | Vysoká odolnost* výše.

2.4.3.11 Správa napříč více úrovněmi odbavení uložených dat (Management across multiple storage availability levels)

Management se provádí přes *Administrační modul Archival Storage*, ovšem vzhledem k očekávanému množství dat u cílových uživatelů systému ARCLib je hlavní důraz kladen na on-line média. Off-line úložiště jsou zvažována zejména jako záloha primárních logických úložišť z důvodu diverzifikace úložných technologií a management se omezuje jen na funkce nutné pro tento účel (tj. evidence, která data jsou zálohována na off-line média apod.).

2.4.3.12 Kvalita komponent úložných médií (Quality of storage media components)

Provozní charakteristiky všech úložišť jsou sledovány a shromažďovány v modulu *ARCLib Administration*, viz sekce *Integrita obsahu | Provádí se kontrola integrity* výše.

2.4.4 Informační bezpečnost (Information security)

2.4.4.1 Zabezpečení (Secure)

Bezpečnostní strategie ARCLib vychází z několika konceptů:

- ARCLib je temný archiv s velmi omezeným počtem osob, které k němu přistupují. Nepočítá se s veřejným zpřístupněním nějakého rozhraní široké skupině uživatelů, a to ani v režimu „jen pro čtení“, což samo o sobě zmenšuje potenciální možnosti útoků.
- Veškerá komunikace mezi jednotlivými moduly ARCLib se děje přes dokumentovaná rozhraní, při komunikaci přes síť šifrovaným spojením s parametry odpovídající soudobé nejlepší praxi.
- Dokumentace ARCLib se v rámci každé komponenty explicitně věnuje potenciálním bezpečnostním rizikům a zavedeným protiopatřením.
- Speciální péče musí být věnována bezpečnosti konkrétních implementací logických úložišť.
    - Jelikož ARCLib počítá s použitím různých variant logických úložišť, musí být dbáno na zpracování analýzy rizik a zavedení a dokumentaci protiopatření každé implementace obdobně, jako u zbytku systému ARCLib.
    - Použití více logických úložišť by mělo být při instalaci ARCLib využito pro ochranu dat před úmyslnou i neúmyslnou závadnou činnosti operátorů ARCLib instalace:
        - Mazání nebo nevratný přepis dat v rámci jednoho logického úložiště musí být možný jen osobám s administračním oprávněním pro dané úložiště.
        - Neměla by existovat žádná osoba, která by měla administrační práva ke všem používaným logickým úložištím.
- Fyzické zabezpečení přístupu k serverům a fyzickým úložištím musí být řešeno v dokumentaci a politikách přístupné instalace systému ARCLib.
- ARCLib nevynucuje šifrování ukládaných AIP.
    - ARCLib se primárně zaměřuje na dlouhodobou ochranu dat knihoven, která jsou všeobecně veřejná.
    - Šifrování přidává velmi komplexní vrstvu ohrožující dostupnost dat v případě ztráty šifrovacích klíčů, ztráty znalosti použitého šifrovacího algoritmu apod., která v dlouhodobém horizontu hrozí.
    - Ostatní bezpečnostní opatření a omezení přístupu k datům se proto zdají dostatečné pro ochranu tohoto typu dat bez nutnosti zvyšování rizika ztráty dostupnosti dat v důsledku ztráty schopnosti jejich dešifrování.
    - ARCLib ale je schopen zpracovat a následně dlouhodobě archivovat balíčky SIP, které obsahují šifrovaná data, pokud obsahují pro ARCLib čitelné metadatové informace dostatečné pro jejich zpracování ve shodě s požadavky Metodiky logické ochrany [MLO].
        - Správa šifrovacích klíčů apod. pak už ovšem je na vlastníkovi dat, který je vkládá do systému ARCLib. ARCLib k tomu neposkytuje speciální podpůrné prostředky.
    - Je možné zvážit použití transparentního šifrování na úrovni jednotlivých logických úložišť, kde jsou šifrovací klíče pro celé úložiště ve správě administrátora daného úložiště.
        - Rozdělení odpovědností za jednotlivá úložiště doporučovaná výše by měla zabránit ztrátě dat z ostatních úložišť, kde je případné šifrování ve správě jiného týmu osob.
        - Transparentní šifrování chrání před únikem dat např. při vyřazení a likvidaci použitých pevných disků, kazet s páskami apod., když se tato média dostávají mimo vliv provozovatele instalace ARCLib.

2.4.4.2 Řízení přístupu (Access controls)

Systém ARCLib je budovaný jako temný archiv, přístup k němu bude mít je velmi omezený počet uživatelů. Řízení přístupu v ARCLib je založeno na systému rolí:

- administrátor – spravuje a konfiguruje systém,
- dodavatel – dodává data (spouští ingest),
- analytik/editor – řeší problémy vznikající při operaci ingest (chyby při validací XML a identifikaci metadat), spravuje data.

Přístup k datům pak mohou mít ještě pověřené osoby (administrátoři), které se starají o jednotlivá logická úložiště či servery, na kterých běží součásti ARCLib.

2.4.4.3 Propojení s autentizací (Integration with authentication)

Autentizace uživatelů v systému ARCLib je postavena na systému LDAP. Implicitně se předpokládá použití vlastní instance LDAP vyhrazené pro potřeby ARCLib. Návrh systému ale explicitně počítá s možností použití externího LDAP serveru pro snadnou integraci s centralizovanou správou identit instituce.

2.4.4.4 Mazání (Deletion)

Mazání AIP umožňuje ARCLib ve dvou úrovních:

1. logické mazání – AIP je označen za smazaný, ale fyzické kopie zůstávají v archivu, aby mohly být v případě potřeby obnoveny,
2. fyzické mazání – AIP je fyzicky skutečně odstraněn z úložiště bez možnosti obnovy dat.

Fyzické mazání mohou provádět jen osoby se zvláštním oprávněním v systému a na základě politiky archivu, která upravuje podmínky, za kterých mohou být data fyzicky smazána.

Pokud ARCLib používá logická úložiště, která fyzicky neumožňují mazání dat (např. úložiště využívající WORM disky), musí být jejich použití explicitně řešeno v politice archivu, kdy musí být dodavatelé dat seznámeni s možností pouze logického mazání dat apod.

V případě logického i fyzického mazání dat jsou o akci provedeny podrobné záznamy v auditním logu, které nemohou být za žádných okolností změněny nebo smazány.

2.4.4.5 Šifrování na straně serveru se spravovanými klíči (At-rest server-side encryption with managed keys)

Jak je podrobněji popsáno v sekci *Informační bezpečnost | Zabezpečení*, systém ARCLib umožňuje použití transparentního šifrování na úrovni jednotlivých logických úložišť, kde jsou šifrovací klíče pro celé úložiště ve správě administrátora daného úložiště. Primárním cílem je ochrana dat při vyřazení a likvidaci použitých úložných médií (pevných disků, kazet s páskami apod.), kdy se média dostávají mimo vliv provozovatele instalace ARCLib, neboť šifrovací klíče jsou v držení provozovatele instalace ARCLib, a takové šifrování je tedy před samotným provozovatelem ARCLib nechrání.

2.4.4.6 Šifrování na straně serveru s klíči ve vlastní správě (At-rest server-side encryption with self-managing keys)

Jak je podrobněji zdůvodněno v sekci *Informační bezpečnost | Zabezpečení*, systém ARCLib nevynucuje šifrování dat s šifrovacími klíči v držení majitelů dat. Podpora takového šifrování na úrovni logických úložišť není plánována. Pokud chce vlastník data ochraňovat data před provozovatelem instalace ARCLib, může využít šifrování dat na úrovni AIP. Správa šifrovacích klíčů je pak plně v jeho kompetenci.

2.4.4.7 Šifrování přenosu (Encrypted transfer)

Veškerá komunikace mezi jednotlivými moduly ARCLib se děje přes dokumentovaná rozhraní, při komunikaci po síti šifrovaně. Pro komunikaci po síti se typicky využije protokol HTTPS, tj. HTTP zabezpečený technologií TLS s parametry odpovídající soudobé nejlepší praxi.

2.4.4.8 Víceklientskost (Multi-tenancy)

Veškerá komunikace uživatelů *ARCLib Archival Storage* se děje výhradně prostřednictvím REST rozhraní poskytovaného komponentou *ARCLib Archival Storage Gateway*, která také řídí další práci s daty. Tato architektura umožňuje zavedení libovolných politik nakládání s daty vzhledem k jejich vlastníkovi, typu, roli uživatele, se kterým *ARCLib Archival Storage Gateway* právě komunikuje apod.

První verze návrhu ARCLib počítá pouze s řízením přístupu k datům dle rolí uživatelů. Nepředpokládá se, že by se různé typy dat ukládaly odlišným způsobem (např. v odlišném počtu kopií apod.). Architektura systému je však na tuto možnost připravená.

2.4.4.9 Odhalování virů / škodlivého software (Virus/malware detection)

Systém ARCLib poskytuje standardizované rozhraní pro připojení antivirových programů, které umožňuje zapojení antivirových kontrol jako součást operace ingest. Standardní součástí implementace ARCLib je integrace s open-source antivirovým řešením ClamAV [ClamAV]. Rozhraní však umožňuje připojení jiného antivirového řešení a takové připojení je z hlediska ostatních komponent *ARCLib Archival Storage* transparentní.

V případě detekce škodlivého software (malware) v balíčku SIP je další průběh operace ingest řízen stejně, jako v případě výskytu jiných typů chyb, tj. je možné operaci ingest zrušit, pozastavit, chybu ignorovat apod.

Průběh operace ingest (Ingest workflow) může být v rámci antivirové kontroly rozšířeno také o karanténu, která vstupní balíčky SIP pozastaví na definovanou dobu (typicky jeden až dva týdny) z dalšího zpracování v karanténním úložišti, než jsou předány antivirové kontrole a do dalšího zpracování v rámci ingestu. Cílem odloženého zpracování je poskytnou prostor pro vývoj a aktualizaci detekčních definic použitého antivirového software pro nové typy škodlivého software.

Zpětné antivirové kontroly AIP v archivním úložišti nejsou plánovány. V případě potřeby je možné je realizovat mimo systém ARCLib vyčtením daného AIP z archivního úložiště a antivirovou kontrolou jeho obsahu mimo systém ARCLib. Antivirová kontrola tedy spadá mimo odpovědnost komponenty *ARCLib Archival Storage.*

2.4.4.10 Nápravná opatření při výskytu virů / škodlivého software (Virus/malware remediation)

Jak je uvedeno v sekci *Informační bezpečnost | Odhalování virů / škodlivého software*, antivirovou kontrolu je možné realizovat v systému ARCLib jako součást operace ingest, a to včetně karantény. Výskyt škodlivého software je pak oznamován stejně, jako jiné chyby zjištěné v průběhu ingestu a standardními prostředky pro řízení ingestu je na ně pak možné také reagovat. Cílem je standardizace a využití stejný postupů pro zpracování malware jako v případě ošetření jiných typů chyb dat vstupujících do systému ARCLib.

Antivirová kontrola tedy spadá mimo odpovědnost komponenty *ARCLib Archival Storage.*

2.4.4.11 Hlášení systémových chyb (System error reporting)

O všech operacích prováděných systémem ARCLib, stejně jako o významných událostech (jako jsou výskyty chyb apod.) jsou vedeny logy, a to ve všech komponentách archivního úložiště a s nastavitelnou úrovní podrobností. Logy se následně definovaným rozhraním předávají do modulu *ARCLib Administration* pro jejich centrální evidenci, kde je zajištěno i jejich neměnné uložení pro auditní účely.

2.4.5 Škálovatelnost a výkon (Scalability & performance)

2.4.5.1 Podpora rozšíření (Supports expansion)

Jak je uvedeno v části *Přizpůsobitelnost a odolnost | Vysoká odolnost*, architektura *ARCLib Archival Storage* je postavena modulárně – plná sada AIP balíků je uložena na každém z logických úložišť, které *ARCLib Archival Storage* využívá. Zvýšit bezpečnost dat je možné přidáním dalšího logického úložiště do *ARCLib Archival Storage*. To může být provedeno za plného provozu systému – nově ukládané AIP jsou průběžně zapisovány i na nově připojené logické úložiště, již dříve uložené AIP jsou na zkopírovány z některého

z ostatních logických úložišť. Nové logické úložiště je označeno za plně nasazené do provozu až v okamžiku, kdy obsahuje plnou sadu AIP.

Implementace logického úložiště může být realizována různými technologiemi (viz *Zohlednění nákladů | Ekonomicky efektivní* a *Přizpůsobitelnost a odolnost | Vysoká dostupnost* výše), pro produkční provoz se předpokládá nasazení pokročilých úložných technologií typu *Ceph* a *ZFS*, které jsou navrženy jako vysoce škálovatelné a rozšiřitelné za běhu bez přerušení provozu.

2.4.5.2 Podpora zmenšení (Supports reduction)

Obdobně jako je možné úložiště bez přerušení provozu zvětšovat, viz *Škálovatelnost a výkon | Podpora rozšíření* výše, je možné úložiště také zmenšovat s možnými omezeními danými použitou technologií pro realizaci logických úložišť.

V případě potřeby je možné odebrat ze systému celé logické úložiště, pokud tím neklesne redundance dat pod úroveň požadovanou úložnou politikou provozovatele ARCLib.

Implementace každého logického úložiště může být realizována různými technologiemi (viz *Zohlednění nákladů | Ekonomicky efektivní* a *Přizpůsobitelnost a odolnost | Vysoká dostupnost* výše), pro produkční provoz se předpokládá nasazení pokročilých úložných technologií typu *Ceph* a *ZFS*, které jsou navrženy jako vysoce škálovatelné. Ceph umožňuje snadnou redukci počtu úložných uzlů bez přerušení provozu, pokud jejich celková kapacita neklesne pod hranici umožňující uložení všech dat s požadovanou úrovní redundance. Redukce ZFS úložiště je možná jen v omezené míře – v současné implementaci ZFS umožňuje pouze odebírání disků zrcadlící data, tj. umožňuje odebírat jen disky za účelem snížení redundance uložení dat, nikoliv pro snížení dostupné úložné kapacity ZFS úložiště.

2.4.5.3 Podpora globálního jmenného prostoru (Supports a global namespace)

Pro koncového uživatele je *ARCLib Archival Storage* implementováno jako objektové úložiště s globálním jmenným prostorem všech uložených objektů. Koncový uživatel tedy vždy má přehled o uložení všech dat.

Aby nedocházelo ke kolizi identifikátorů ukládaných objektů, používají se jako identifikátory bezvýznamová ID, generovaná např. dle UUID standardu [UUID].

2.4.5.4 Využití více úrovní odbavení uložených dat (Use of multiple storage availability levels)

Jak je uvedeno v sekci *Zohlednění nákladů | Ekonomicky efektivní*, úložiště *ARCLib Archival Storage* je navrženo modulárním způsobem a umožňuje použití prakticky libovolné úložné technologie. Koncový uživatel *ARCLib Archival Storage* však je od konkrétní implementace kompletně odstíněn a k datům přistupuje přes jednotné rozhraní API *ARCLib Archival Storage Gateway*.

Jak je uvedeno v sekci *Informační bezpečnost | Víceklientskost*, komunikace prostřednictvím komponenty *ARCLib Archival Storage Gateway* umožňuje zavedení libovolných politik pro ukládání dat, např. i v závislosti na typu úložiště z hlediska dostupnosti (rychlosti odbavení požadavku na uložení/získání dat na/z úložiště). Nicméně v návrhu první verze ARCLib se nepočítá s implementací této funkcionality, neboť uložení očekávaných objemů dat vystačí s použitím on-line úložišť. Near-line a off-line média budou využita jen pro zálohování těchto primárních úložišť.

2.4.5.5 Odstupňovaný výkon (Tiered performance)

Pro možnost rychlého vyhledávání v obsahu archivního úložiště jsou v rámci systému ARCLib shromažďována a indexována vybraná metadata z AIP balíčků (viz [MLO]) v rámci komponenty *ARCLib Data management*. Tím je zajištěn rychlý přístup k základním údajům o ukládaných datech bez nutnosti přístupu k datům z archivního úložiště.

Samotné archivní úložiště umožňuje použití různých úložných technologií s různým výkonem, viz *Škálovatelnost a výkon | Využití více úrovní odbavení uložených dat* výše.

2.4.5.6 Škálovatelné na velké objemy dat (Scalable to large data sizes)

Škálovatelnost archivního úložiště je v konečném důsledku limitována implementací logických úložišť. Pro produkční provoz se předpokládá nasazení pokročilých úložných technologií typu *Ceph* a *ZFS*, které jsou navrženy jako vysoce škálovatelné.

Např. Dutch National Archive využilo Ceph pro uložení více než 2 PB dat s trojitou redundancí, tj. hrubá úložná kapacita je kolem 8 PB [CephDNA]. ZFS byl od počátku projektován jako extrémně škálovatelný, samotný název *Zettabyte File System* (ZFS) odkazuje na teoretický horní limit úložné kapacity jednoho ZFS svazku, který je 256 kvadrilionů zettabytů, tj. 2^{128} bytů [ScaleZFS].

V případě implementace *ARCLib Archival Storage* logického úložiště pomocí souborového systému, jakým je např. ZFS, je při uložení velkého množství jednotlivých souborů důležité zajistit jejich vhodné uspořádání do adresářové struktury. Ta je odvozena z UUID identifikátoru [UUID] daného objektu (viz *Škálovatelnost a výkon | Podpora globálního jmenného prostoru*). Např. objekt s ID *331c32e6-4e10-11e7-b114-b2f933d5fe66* bude uložen v adresářové struktuře takto: *331c/32e6/4e10/11e7/b114/331c32e6-4e10-11e7-b114-b2f933d5fe66* Tímto způsobem je zajištěno rovnoměrné rozprostření všech dat do adresářové struktury, a tedy ukládání jen omezeného množství souborů do jediného adresáře. Zároveň ale tento postup poskytuje jednoznačný způsob odvození cesty k souboru v úložišti z jeho ID, a naopak odvození ID daného objektu pro každý soubor v úložišti.

2.4.5.7 Omezení systému souborů (File system limits)

Konkrétní limity jsou dány použitou implementací logického úložiště, v rámci systému ARCLib jsou plánovány extrémně škálovatelné technologie jako Ceph a ZFS, které představují soudobou špičku v oboru. Viz *Škálovatelnost a výkon | Škálovatelné na velké objemy dat* výše. Modulární architektura *ARCLib Archival Storage* zároveň umožňuje přejít k jiné implementaci logického úložiště, pokud by to bylo zapotřebí, viz sekce *Naplnění požadavků OAIS | Náhrada médií*.

2.4.5.8 Dodání (Delivery)

Návrh úložiště *ARCLib Archival Storage* je zaměřen na paralelní zpracování a vysokou škálovatelnost, viz sekce *Přizpůsobitelnost a odolnost | Vysoká dostupnost* a *Škálovatelnost a výkon* výše, tak, aby bylo možné dosáhnout požadované rychlosti dodání dat z úložiště při požadavku na jejich získání.

2.4.5.9 Úplný export (Complete exports)

Možnosti exportu dat nejsou nijak omezovány, v případě potřeby je možné přistupovat k datům přímou komunikací s konkrétním logickým úložištěm. Možnosti exportu dat jsou pak limitovány pouze možnostmi tohoto úložiště. Technologie jako *Ceph* nebo *ZFS*, které jsou primárně plánovány pro projekt ARCLib, nemají žádné významné limitace. V případě použití cloudového úložiště je třeba zajistit si konkrétní podmínky exportu (přenosová rychlost, poplatky za přenos objemu dat v časovém úseku apod.) smluvně s poskytovatelem cloudového úložiště.

2.4.5.10 V/V výkon (I/O performance)

Návrh úložiště *ARCLib Archival Storage* je zaměřen na paralelní zpracování a vysokou škálovatelnost, viz sekce *Přizpůsobitelnost a odolnost | Vysoká dostupnost* a *Škálovatelnost a výkon* výše. Optimálního výkonu pro přenos dat (vstup/výstup) je možné dosáhnout použitím vhodné implementace použitých logických úložišť. Např. v případně úložiště postaveného na technologii *Ceph* je možné paralelní výkon zvýšit přidáním dalších uzlů do úložného clusteru.

2.4.5.11 Výpočetní výkon (Compute power)

Systém ARCLib je navržen pro paralelní zpracování, aby bylo možné dosáhnout optimálního výpočetního výkonu použitím patřičného počtu zpracujících uzlů. Viz sekce *Škálovatelnost a výkon* výše.

2.4.6 Umístění úložiště (Storage location)

2.4.6.1 Geografické oddělení (Geographic separation)

Geografická distribuce dat v *ARCLib Archival Storage* je plánována primárně použitím více logických úložišť, která data fyzicky ukládají v geograficky odlišných lokalitách. Každé z logických úložišť přitom obsahuje kompletní sadu všech balíčků AIP, viz *Přizpůsobitelnost a odolnost | Vysoká odolnost* výše.

V závislosti na použité technologii logického úložiště může být geografická distribuce dat zajištěna také replikací dat v rámci daného logického úložiště. Např. systém *Ceph* v *Multi-Site* konfiguraci podporuje automatickou asynchronní replikaci dat mezi geograficky vzdálenými *Ceph* clustery [CephMultiSite].

2.4.6.2 Replikace (Replication)

Základním způsobem replikace dat v *ARCLib Archival Storage* je ukládání dat na více logických úložišť distribuovaných do více geograficky vzdálených lokalit. Samotné logické úložiště by typicky mělo být implementováno pomocí technologie, která sama o sobě poskytuje redundanci uložení dat. V rámci ARCLib konkrétně předpokládáme použití systému Ceph a ZFS, vždy s redundantním uložením dat v minimálně dvou kopiích. Viz *Přizpůsobitelnost a odolnost | Vysoká odolnost*.

Data vkládaná uživatelem do *ARCLib Archival Storage* jsou archivním úložištěm uživateli oznámena za uložená až v okamžiku, kdy jsou označena za uložená každým z logických úložišť.

Replikace uvnitř logického úložiště realizovaného jedním ZFS svazkem, respektive jedním Ceph clusterem, probíhá v reálném čase. Pokud je v rámci Ceph logického úložiště nasazena Ceph Multi-Site geografická replikace (viz [CephMultiSite]) do jiného Ceph clusteru, probíhá tato replikace již asynchronně s možnou časovou prodlevou v závislosti na zatížení Ceph clusteru a vytížení síťového spojení k replice.

2.4.6.3 Přizpůsobení replikace dle obsahu (Customizable replication based on content)

Pro první verzi systému ARCLib jsme se z důvodu zachování jednoduchosti systému rozhodli neimplementovat možnost odlišného způsobu ukládání pro různé typy dat. Všem datům je tedy poskytována stejná úroveň ochrany.

Architektura systému však umožňuje implementaci této funkcionality v budoucnu – jak je uvedeno v sekci *Informační bezpečnost | Víceklientskost*, veškerá komunikace uživatele s archivním úložištěm probíhá prostřednictvím *ARCLib Archival Storage Gateway*, což umožňuje zavedení libovolných politik pro distribuci dat na logická úložiště např. i v závislosti na jejich typu. V první verzi systému ARCLib jsou však všechna data zpracovávána identickým způsobem.

2.4.6.4 Známé a omezené umístění (Expose and constraint location)

Jak je uvedeno v oddílu *Umístění úložiště | Přizpůsobení replikace dle obsahu*, současná implementace *ARCLib Archival Storage* ukládá všechna data identickým způsobem na všechna logická úložiště. Úložné lokace jsou tak implicitně známy – odpovídají lokacím fyzických úložišť všech logických úložišť používaných danou instalací *ARCLib Archival Storage*.

2.4.7 Podpora (Support)

2.4.7.1 Závazná podpora (Support commitment)

Podpora použité úložné infrastruktury a souvisejících služeb musí být definována politikou provozovatele systému ARCLib tak, aby byla zajištěna požadovaná úroveň bezpečnosti a dostupnosti dat. Politika musí definovat odpovědnosti za všechny části infrastruktury, úroveň záruk na dostupnost dat, časy reakce na hlášené incidenty na infrastruktuře a požadavky na dobu jejich řešení.

2.4.7.2 Školení (Training)

Provozovatel ARCLib musí stanovit politiky školení všech pracovníků, kteří se systém ARCLib pracují, respektive zajišťují provoz komponent, které ARCLib pro svůj provoz využívá.

2.4.7.3 Přístupnost (Accessibility)

Dokumentace systému a jím produkované reporty jsou ve standardních formátech umožňujících jejich využití i hendikepovaným osobám s použitím odpovídajících pomůcek (hlasové čtečky apod.).

2.4.8 Otevřenost (Transparency)

2.4.8.1 Podpora otevřených úložných formátů (Supports open storage formats)

Systém ARCLib ukládá data v otevřených standardizovaných formátech (BagIt, METS, PREMIS, Dublin Core atd.), viz [MLO, str. 44].

2.4.8.2 Upozornění na chyby v datech (Data error notification)

Systém ARCLib poskytuje nástroje pro podrobné průběžné sledování stavu systému i dat, viz sekce *Přizpůsobitelnost a odolnost | Hlášení systémových chyb*. Politikou provozovatele ARCLib a jeho dohodou s dodavateli dat pak musí být dohodnut způsob předávání těchto informací z ARCLib k vlastníkům dat.

2.4.8.3 Otevřenost při automatickém zotavení (Self-healing transparency)

Viz sekce *Přizpůsobitelnost a odolnost | Vysoká odolnost* a *Informační bezpečnost | Hlášení systémových chyb* výše.

2.4.8.4 Podpora nezávislých ochranných akcí (Supports independent preservation actions)

ARCLib Archival Storage je od počátku navrženo jako datové úložiště neutrální k obsahu dat. Stará se pouze o bezpečné uložení datových souborů a neposkytuje funkce pro migraci datových formátů apod. – návrh *ARCLib Archival Storage* počítá s prováděním ochranných akcí mimo *ARCLib Archival Storage*.

2.4.8.5 Sledování (Monitoring)

Viz sekce *Informační bezpečnost | Hlášení systémových chyb* výše.

2.4.8.6 Poskytuje hlášení o obsahu (Provides content reports)

Pro možnost rychlého vyhledávání v obsahu archivního úložiště jsou v rámci ARCLib shromažďována a indexována vybraná data o AIP balíčcích v rámci komponenty *ARCLib Data management*. Data o úložné infrastruktuře jsou pak dostupná prostřednictvím definovaného rozhraní pro reporting úložiště *ARCLib Archival Storage*.

2.4.8.7 Poskytuje hlášení o aktivitách (Provides activity reports)

Viz sekce *Informační bezpečnost | Hlášení systémových chyb* výše.

2.4.8.8 Přizpůsobitelná hlášení (Custom reports)

Centrálním bodem pro zpracování všech reportů je modul *ARCLib Administration*, jemuž *ARCLib Archival Storage* předává potřebná data. V rámci modulu *ARCLib Administration* je možné parametry hlášení konfigurovat.

2.4.8.9 Dokumentovaná infrastruktura (Documented infrastructure)

Celá použitá infrastruktura, související služby a postupy musí být podrobně dokumentovány. Dokumentace konkrétního nasazení je odpovědností provozovatele dané instance systému ARCLib.

2.4.8.10 Dokumentace přístupů (Documented access)

Viz sekce *Přizpůsobitelnost a odolnost | Hlášení systémových chyb*.

2.4.8.11 Dokumentace původu (Documented provenance)

Systém ARCLib uchovává informace o původu dat i všech změnách v datech provedených, viz [MLO, str. 44] a sekce *Integrita obsahu | Provádí se kontrola integrity* a *Přizpůsobitelnost a odolnost | Hlášení systémových chyb* výše.

Zdůvodnění metodiky

Metodika je reprezentací postupů, které byly na základě mezinárodních doporučení, zkušenosti členů realizačního týmu a jejich výzkumu přijaty jako principy pro vývoj systému ARCLib. Tyto principy se odrazily ve funkčnosti systému. Metodika představuje souhrn doporučení a jejich zdůvodnění, jak zajistit ochranu bitstreamu v nástroji ARCLib. Přístupů, jak zajistit požadovaný výsledek je více, metodika je doporučením konkrétních metod a jejich kombinací. V tomto ohledu představuje originální výsledek, který je novým z hlediska určeného postupu.[10] Svým uživatelům garantuje správnost postupů a jistotu výsledku při dodržení doporučených principů.

Metodika definuje postupy nutné k důvěryhodné realizaci ochrany bitstreamu jako základní a nezbytné podmínky dlouhodobého uchovávání digitálních dat. Správa digitálních dat tak, aby nepodléhala fyzické degradaci, se v současnosti považuje za základ všech aktivit spojených s dlouhodobým uložením. Přes tuto samozřejmost existují doporučení jen obecné povahy, jak požadovaný výsledek zajistit. Konkrétní operace a úkony jsou vázány na konkrétní nástroje k jejich provádění. Mnohdy jsou podrobnosti neveřejného charakteru a jsou spojeny s nákupem komerčního softwaru. Tato metodika proto přináší souhrn doporučených postupů v obecné rovině, návrhy na jejich konkrétní realizaci společně s doporučeními danými zkušenostmi a tzv. dobrou praxí. Pro uživatele systému ARCLib přináší funkční popis systému v oblasti bitové ochrany a doporučení pro konkrétní operace, které lze přímo realizovat – což je ve spojení s mapováním na postupy dobré praxe novým výsledkem založeným na originálním přístupu vyvinutým v rámci řešení projektu. V oblasti využití systému ARCLib tedy představuje metodika zcela nové výsledky a postupy, které jsou logicky spjaté s tímto nástrojem. Ovšem i souhrn dobré praxe představuje v českém prostředí chybějící doporučení, kterého se lze s vysokou mírou důvěryhodnosti držet a realizovat s jeho pomocí odpovědnou péči o digitální dokumenty. Tento výsledek je důležitý zejména pro veřejné instituce, které nedisponují patřičným množstvím expertů.

Metodika vychází z doporučení mezinárodních norem ČSN ISO 14721, ČSN ISO 16363, NDSA Levels of Digital Preservation a Preservation Storage Criteria. Jejich ustanovení lokalizuje do českého prostředí a obecné zásady převádí do nově vytvořených konkrétních postupů a doporučení, jak žádaných cílů dosáhnout v instituci, která bude využívat řešení ARCLib. Metodika vychází z výzkumů provedených v projektu NAKI II ARCLib i v dřívějším působení jednotlivých autorů. Spojuje teoretické postuláty a zásady s konkrétní implementací. Uživatelům softwarového řešení ARCLib poskytuje doporučení, jak výše zmíněné postupy aplikovat, jak provádět správu dat v systému a hodnotit rizika. Metodika vychází z mezinárodních doporučení a normativních dokumentů. Na jejich základě byly postulovány principy pro fungování nástroje ARCLib.

[10] V oblasti knihoven existuje jen obecné doporučení Národní knihovny ČR pro zajištění bitové ochrany digitálních dat http://kramerius-info.nkp.cz/index.php/cinnosti-ve-visk-7/ochranne-reformatovani-ohrozenych-bohemikalnich-dokumentu/, které však kvůli neznalosti konkrétní situace v jednotlivých institucí logicky zůstává pouze v obecné rovině.

Seznam použité literatury

[CephDNA] Wido den Hollander. *Ceph: building the dutch national archive*. 2016 [cit. 2018-09-13] Dostupné z: http://widodh.o.auroraobjects.eu/talks/Ceph_dutch_national_archive_2016.pdf

[CephMultiSite] Red Hat, Inc. *Ceph Documentation » Ceph Object Gateway » Multi-Site*. 2017. [cit. 2017-06-22] Dostupné z: http://docs.Ceph.com/docs/master/radosgw/multisite/

[ClamAV] ClamAV Team. *ClamAV: an open source antivirus engine for detecting trojans, viruses, malware & other malicious threats*. 2018. [cit. 2018-09-13] Dostupné z: https://www.ClamAV.net/

[ČSNISO14721] ČSN ISO 14721. Systémy pro přenos dat a informací z kosmického prostoru – Otevřený archivační informační systém – Referenční model. Praha: Úřad pro technickou normalizaci, metrologii a státní zkušebnictví, 2014.

[ČSNISO16363] ČSN ISO 16363. Systémy pro přenos dat a informací z kosmického prostoru – Audit a certifikace důvěryhodných digitálních úložišť. Praha: Úřad pro technickou normalizaci, metrologii a státní zkušebnictví, 2014.

[DPM2012] *Digital Preservation Management: Implementing Short-term Strategies for Long-term Problems*. MIT Libraries: Massachusetts, 2012. [cit. 2017-03-13] Dostupné z: http://www.dpworkshop.org/

[DPSG] Digital Preservation Storage Group. 2018. [cit. 2018-09-12] Dostupné z: https://groups.google.com/forum/#!forum/dpstorage

[Guidelines2003] *Guidelines for the Preservation of Digital Heritage*. Melbourne: National Library of Australia, 2003, s. 170. [cit. 2017-03-13] Dostupné z: http://unesdoc.unesco.org/images/0013/001300/130071e.pdf

[LoC] Library of Congress. 2017. [cit. 2017-06-04] Dostupné z: https://www.loc.gov/

[LOCKSS] Stanford University. *Preservation Principles | LOCKSS*. [cit. 2018-09-12] Dostupné z: https://www.lockss.org/about/principles/

[MLO] Hutař, Jan; Miranda, Andrea; Pavlásková, Eliška; Vašek, Zdeněk; Hruška, Zdeněk. *Metodika logické ochrany digitálních dat*. Certifikovaná metodika Ministerstva kultury ČR. Knihovna AV ČR, 2018, s. 83. Dostupné z: http://www.nusl.cz/ntk/nusl-371612

[NCDL2014] Mohanty, Rasmita; Kumar, Ranjir, das. *Proceedings of National Conference on Digital Libraries: Reshaping Traditional Libraries into Next Generation Libraries (NCDL-2014)*, 2014. [cit. 2016-12-09] ISBN 9788184249019. Dostupné z: https://books.google.cz/books/about/Proceedings_of_National_Conference_on_Di.html?id=0_EBogEACAAJ&redir_esc=y

[NDSA] Library of Congress. *National Digital Stewardship Alliance*. 2017. [cit. 2017-06-04] Dostupné z: http://www.digitalpreservation.gov/ndsa/

[NDSAlevelsRC1] Owens, Trevor. *NDSA Levels of Digital Preservation: Release Candidate One*. Library of Congress, National Digital Stewardship Alliance. November 20, 2012. [cit. 2017-06-04] Dostupné z: https://blogs.loc.gov/thesignal/2012/11/ndsa-levels-of-digital-preservation-release-candidate-one/

[NDSAlevelsV1] Phillips, Megan; Bailey, Jefferson, Goethals, Andrea; Owens, Trevor. *The NDSA Levels of Digital Preservation: An Explanation and Uses*. Library of Congress, National Digital Stewardship Alliance, 2013. [cit. 2017-06-04] Dostupné z: http://www.digitalpreservation.gov/documents/NDSA_Levels_Archiving_2013.pdf

[NDSAlevels2016] Ashenfelder, Mike. *Expanding NDSA Levels of Preservation*. Library of Congress, National Digital Stewardship Alliance, April 12, 2016. [cit. 2017-06-04] Dostupné z: https://blogs.loc.gov/thesignal/2016/04/expanding-ndsa-levels-of-preservation/

[OAIS] Consultative Committee for Space Data Systems. *Reference Model for an Open Archival Information System (OAIS)*, Recommended Practice, CCSDS 650.0-M-2 (Magenta Book), 2012, Issue 2, s. 135. [cit. 2018-09-12] Dostupné z: http://public.ccsds.org/publications/archive/650x0m2.pdf

[PSC] Mandelbaum, J. *Preservation Storage Criteria*. Presentation at Designing Storage Architectures for Digital Collections, 2016, Library of Congress.

[PSCv1] iPRES 2016. *Preservation Storage Criteria (Short version)*. iPRES 2016, 2016-09-17, s. 5. [cit. 2018-09-12] Dostupné z: https://docs.google.com/document/d/1WByfeonuTae5oOckSgeWeXI1TRntJE9RNMhMQqh4sS8

[PSCv2] Digital Preservation Storage Group. *Preservation Storage Criteria, Version 2*. 2017-05, s. 7. [cit. 2018-09-12] Dostupné z: https://docs.google.com/document/d/1CEkcWskAbph0gQ4ATK_SWYw-6-8zBxnGFfqZiwtfpEI

[ScaleZFS] Oracle Corporation and/or its affiliates. *Oracle Solaris ZFS Administration Guide: What Is ZFS?* 2010. [cit. 2018-09-13] Dostupné z: http://docs.oracle.com/cd/E19253-01/819-5461/ZFSover-2/#6n7ht6qtm

[UUID] Internet Engineering Task Force, Network Working Group. *Request for Comments (RFC) 4122: A Universally Unique IDentifier (UUID) URN Namespace*. Proposed standard, July 2005. [cit. 2018-09-13] Dostupné z: https://tools.ietf.org/html/rfc4122

[Vojtášek2000] VOJTÁŠEK, Filip. *Dlouhodobá archivace digitálních dokumentů*. Ikaros [online], 2000, ročník 4, číslo 10. urn:nbn:cz:ik-10646. ISSN 1212-5075. [cit. 2016-11-25] Dostupné z: http://ikaros.cz/node/10646

[WiPS] Goethals, Andrea; Schaefer, Sibyl; Zwaard, Kate; Truman, Gail; McGovern, Nancy; Mandelbaum, Jane; Knight, Steve. *What is Preservation Storage?* Workshop at iPRES 2016, Swiss National Library, Bern, 2016. [cit. 2018-09-12] Dostupné z: https://phaidra.univie.ac.at/view/o:502819

Seznam publikací předcházejících metodice

Bartošek, Miroslav. *Uchování digitálního dědictví a systém Archivematica.* ITlib: Informačné technológie a knižnice, Bratislava: Centrum vedecko-technických informácií SR, 2016, roč. 20, č. 2, s. 32–35. ISSN 1335-793X. Dostupné z: http://itlib.cvtisr.sk/archiv/2016/2/uchovani-digitalniho-dedictvi-a-system-archivematica-preservation-of-digital-heritage-and-system-archivematica.html?page_id=3171

Melichar, Marek; Hutař, Jan; Růžička, Michal; Hruška, Zdeněk; Vašek, Zdeněk; Bartošek, Miroslav; Lhoták, Martin. *Projekt ARCLib – budování systému pro dlouhodobou archivaci digitálních dat v českých knihovnách.* ITlib: Informačné technológie a knižnice, Bratislava: Centrum vedecko-technických informácií SR, 2016, roč. 20, č. 2, s. 13–17. ISSN 1335-793X. Dostupné z: http://hdl.handle.net/11104/0263716

Bartošek, Miroslav. *Archivematica – open-source systém pro digitální archivaci*. Knihovna, Praha: Národní knihovna ČR, 2015, Ročník 26, č. 2, s. 25–38. ISSN 1801-3252. Dostupné z: http://knihovnarevue.nkp.cz/archiv/2015-2/recenzovane-prispevky/archivematica-2013-open-source-system-pro-digitalni-archivaci

Bartošek, Miroslav. *Dlouhodobé uchování digitálních informací, Archivematica a projekt LTP-pilot*. Duha, Brno: Moravská zemská knihovna v Brně, 2015, roč. 29, č. 3, s. 36–37. ISSN 0862-1985. Dostupné z: http://duha.mzk.cz/clanky/dlouhodobe-uchovani-digitalnich-informaci-archivematica-projekt-ltp-pilot

Hutař, Jan; Melichar, Marek. *The Long Decade of Digital Preservation in Heritage Institutions in the Czech Republic: 2002–2014*. International Journal of Digital Curation, 2015, roč. 10, č. 1. DOI: https://doi.org/10.2218/ijdc.v10i1.324 Dostupné z: http://www.ijdc.net/article/view/10.1.173

Hutař, Jan; Melichar, Marek. *Principy strategie rozvoje knihoven oblasti dlouhodobé archivace digitálních informací v České republice: stav v roce 2014 a výhled do roku 2019*. Duha, Brno: Moravská zemská knihovna v Brně, 2014, roč. 28, č. 1. Dostupné z: https://duha.mzk.cz/clanky/principy-strategie-rozvoje-knihoven-oblasti-dlouhodobe-archivace-digitalnich-informaci-v-cesk

Pavlásková, Eliška. *Techniky posuzování rizik a jejich využití v institucionálních repozitářích – užití v Digitálním repozitáři Univerzity Karlovy v Praze*. ProInflow: Časopis pro informační vědy, 2014, Brno: Masarykova univerzita, roč. 6, č. 1, s. 26–37. ISSN 1804-2406. DOI: https://doi.org/10.5817/proinflow.v6i1.943 Dostupné z: http://www.phil.muni.cz/journals/index.php/proinflow/article/view/943

Hruška, Zdeněk. *Audit digitálních repozitářů*. Duha, Brno: Moravská zemská knihovna v Brně, 2013, roč. 27, č. 4. ISSN 0862-1985. Dostupné z: http://duha.mzk.cz/clanky/audit-digitalnich-repozitaru

Hutař, Jan; Melichar, Marek. *Analýza dostupných technologií a srovnání národních strategií v oblasti dlouhodobé archivace digitálních informací*. Praha, Brno, Wellington: Moravská zemská knihovna v Brně, 2013. Dostupné z: https://drive.google.com/a/mzk.cz/file/d/0B1hyhqrXs6dtT094RWlfY1lVSW8

Hutař, Jan. *Archives New Zealand – budování digitálního archivu pro dlouhodobou ochranu digitálních dokumentů*. Archivní časopis, 2013, Roč. 63, č. 1, s. 5–24. Dostupný z: https://docs.google.com/a/mzk.cz/file/d/0B1hyhqrXs6dtMENvaE5RZ1lFVkE

Hutař, Jan; Melichar, Marek. *Hodnocení kvality v oblasti dlouhodobé ochrany digitálních informací*. Brno: Moravská zemská knihovna v Brně, 2012. Dostupné z: https://docs.google.com/a/mzk.cz/file/d/0B1hyhqrXs6dtVXc1ckh0bGo3eDg

Hutař, Jan; Melichar, Marek. *OAIS: možnosti a limity aplikace*. ITlib: Informačné technológie a knižnice, Bratislava: Centrum vedecko-technických informácií SR, 2012, roč. 16, č. 3, s. 37–44. ISSN 1335-793X. Dostupné z: http://itlib.cvtisr.sk/archiv/2012/3/oais-moznosti-a-limity-aplikacie.html?page_id=492

Hutař, Jan. *Assessing Digital Preservation Strategies*. International Council of Archives Congress. Brisbane, 2012, s. 10. Dostupné z: http://ica2012.ica.org/files/pdf/Full papers upload/ica12Final00155.pdf.

Rosenthal, Colin; Blekinge-Rasmussen, Asger; Jan Hutař a kol. (z angličtiny přeložili Hutař, Jan; Cubr, Ladislav; Melichar, Marek). Průvodce plánem důvěryhodného digitálního repozitáře (PLATTER). Praha: Národní knihovna České republiky, 2009. 1. vydání. Dostupné z: http://www.ndk.cz/platter-cz

Přílohy

A. Preservation Storage Criteria

Převzato z [PSCv2].

Number	Criteria	Category	Description
1	Provides integrity checks	Content integrity	Performs verifiable and/or auditable integrity checking as part of the preservation storage
2	Supports independent integrity checks	Content integrity	Supports fixity checking by other parties, for example the content-owning institution
3	Provides preservation actions	Content integrity	Provides tools and/or services to support digital preservation actions (e.g. fixity checking, migration, auditing processes) as part of the preservation storage
4	Cost-efficient	Cost considerations	Costs relatively less than other more expensive solutions per GB, by being designed with cost efficiencies, for example, has resource pooling and sharing, multi-tenancy (multiple users share the same applications)
5	Energy-efficient	Cost considerations	Designed to conserve energy, for example, requires less cooling, consumes less power, uses less rack space, as in green computing initiatives
6	Storage weight	Cost considerations	The physical weight of the storage should meet certain qualifications, for example, be under a certain amount required for a particular floor.
7	High resilience	Flexibility & resilience	Has high resilience, which is the ability to adapt under stress or faults (e.g. resilient to equipment failures, power outages, attacks, surges in user demand)
8	High availability	Flexibility & resilience	Has a high percentage of uptime, i.e. operational for a long length of time, due to techniques such as eliminating single points of failure by having redundant equipment, load-balanced systems and effective monitoring to detect software or hardware failures
9	Recovery	Flexibility & resilience	Has documented ability to replace any corrupt/bad file, file system, or large-scale set of files in reasonable/expected/negotiated timeframes
10	Designed for zero data loss	Flexibility & resilience	Error detection and correction 24/7/365 (e.g. using RAID, Erasure coding, ZFS, triple copies/rebuild)
11	Independence	Flexibility & resilience	Storage layer is independent from other systems in the digital preservation environment so that it could be separately replaced without affecting the entire infrastructure
12	Nondisruptive storage migrations	Flexibility & resilience	Allows for storage tier changes over time (without disruption to availability)
13	Integratable	Flexibility & resilience	Storage layer is easily integrated with other systems and applications (i.e. plug and play)
14	Open source	Flexibility & resilience	Storage infrastructure can be integrated with open source tools and services in accordance with the organization’s preferences

15	Supports multiple file system protocols	Flexibility & resilience	Infrastructure supports multiple file system protocols, e.g. NFS, CIFS, iSCSI, open or standard APIs etc. enabling vendor-neutral, direct addressability.
16	Diverse storage media types	Flexibility & resilience	Uses different storage media types together (e.g. disk and tape)
17	Management across multiple storage availability levels	Flexibility & resilience	Supports management and monitoring across multiple storage availability levels, e.g. online, near-line, off-line
18	Quality of storage media components	Flexibility & resilience	The known failure rate and technical characteristics of the storage media components is acceptable.
19	Secure	Information security	Includes safeguards, data security and documented procedures to prevent security incidents related to hardware, software, personnel, and physical structures, areas and devices.
20	Access controls	Information security	Provides role-based, access controls for storage infrastructure, e.g. user, staff, admin, to ensure only the appropriate people have the appropriate levels of access
21	Integration with authentication	Information security	Able to integrate with relevant organisational authentication systems to authenticate internal and external users of the system.
22	Deletion	Information security	Supports 'true deletion' (i.e. not just pointers) by authorised user or in accordance with local rules
23	At-rest server-side encryption with managed keys	Information security	Provides encryption at the storage layer, with no keys for customers to manage
24	At-rest server-side encryption with self-managing keys	Information security	Provides encryption at the storage layer, but customers manage encryption keys
25	Encrypted transfer	Information security	An appropriate transport layer encryption is used at all times when moving content
26	Multi-tenancy	Information security	Storage infrastructure supports separate roles/rules/access controls for separate agencies/departments/colleges/faculties etc
27	Virus/malware detection	Information security	Includes software that regularly runs virus checks and malware detection.
28	Virus/malware remediation	Information security	Provides remediation actions for content with viruses and/or malware, e.g. quarantine, notification, etc.
29	System error reporting	Information security	Provides immutable logs and/or reports that show all system errors, failures and other critical system activities
30	Supports expansion	Scalability & performance	Can increase storage over time as needed

31	Supports reduction	Scalability & performance	Can decrease storage over time to support deaccessions, transfer of ownership, etc.
32	Supports a global namespace	Scalability & performance	Nothing limits ability to have a global (i.e. consolidated) view of files
33	Use of multiple storage availability levels	Scalability & performance	Supports use of multiple storage availability levels, e.g. online, near-line, off-line
34	Tiered performance	Scalability & performance	Meets specified/negotiated performance levels appropriate to material being stored, e.g. Tier1 storage for metadata indexing and searching, Tier2 for caching, Tier3 or lower for bulk storage.
35	Scalable to large data sizes	Scalability & performance	Able to support very large amounts of content, e.g. multiple PBs of data, hundreds of millions of files and directories, terabyte size files
36	File system limits	Scalability & performance	Able to support long file, path or directory names; large amount of files in a directory, diverse character encodings
37	Delivery	Scalability & performance	Meets expectations for delivery from the storage layer, e.g. at a reasonable/negotiated rate and supporting concurrent users
38	Complete exports	Scalability & performance	Supports the bulk exporting of content and metadata for any reason, at an acceptable rate, for example, as part of an exit strategy
39	I/O performance	Scalability & performance	The input/output performance of the system or service is at an acceptable rate
40	Compute power	Scalability & performance	Computing power of the system or service is at an acceptable rate and available when needed
41	Geographic separation	Storage location	Ensures multiple redundant copies in geographically-separate locations for protection from catastrophic loss
42	Replication	Storage location	Has documented ability to create redundant, distributed copies of content in reasonable timeframes
43	Customizable replication based on content	Storage location	Storage infrastructure supports content-specific user-defined replication rules, for example less copies of a particular stream of content
44	Expose and constraint location	Storage location	Storage infrastructure exposes the specific storage location of data to meet content-specific requirements (e.g. location constraints, transparency of expectations and requirements)
45	Support commitment	Support	Documented vendor or IT commitment to support storage infrastructure, e.g. through SLAs (addressing for example responsibilities, data assurance, response times, end-of-service exit provisions, etc.)
46	Training	Support	Training provided to appropriate staff across all relevant operational and maintenance tasks
47	Accessibility	Support	Ensures people with disabilities equivalent access to reports, documentation and other content

48	Supports open storage formats	Transparency	Infrastructure supports open, standard, non-proprietary storage formats, e.g. TAR, archive eXchange format (AXF), LTFS
49	Data error notification	Transparency	Notifies content-owners of all data errors, remediation actions and issues in reasonable/expected/negotiated timeframes
50	Self-healing transparency	Transparency	Systems that use mechanisms to correct altered data (like bit corruptions) do so in a transparent, documented manner.
51	Supports independent preservation actions	Transparency	Supports digital preservation actions (e.g. migration, auditing processes) by other parties or external tools, for example a format migration by the content-owning institution running tools that are not part of the storage infrastructure
52	Monitoring	Transparency	Supports ability to observe or check activity in the storage infrastructure (e.g. see activity in real-time, examine logs, observe the performance status, determine the overall status or drill-down into activities)
53	Provides content reports	Transparency	Provides reports about content in the storage infrastructure (e.g. number of objects/files/formats, average file size, types of objects, size of storage in use)
54	Provides activity reports	Transparency	Provides reports about activity in the storage infrastructure (e.g. fixity or virus results, corruption, replacement with good copies)
55	Custom reports	Transparency	Supports custom (for example configurable and/or on-demand) reporting of content or activity in the storage infrastructure
56	Documented infrastructure	Transparency	Provides full, complete, current, and available documentation of key processes, services, systems, procedures, known limitations and functions
57	Documented access	Transparency	Provides immutable logs and/or reports that show all file system access
58	Documented provenance	Transparency	Documents audit/provenance information about all changes, for example about integrity check failures, deletions, modifications, additions, preservation actions; and who or what performed the actions

B. NDSA Levels of Preservation

Převzato z [NDSAlevels2016].

	1. úroveň (Ochrana dat)	2. úroveň (Znalost dat)	3. úroveň (Monitorování dat)	4. úroveň (Oprava dat)
Úložiště a geografické umístění	Dvě úplné kopie, které nejsou propojené. Přesun dat z heterogenních médiích (optické disky, pevné disky apod.) na vaše úložiště.	Alespoň tři úplné kopie. Alespoň jedna kopie v odlišném geografickém umístění. Dokumentace úložišť a paměťových médií a soupis předpokladů k jejich používání.	Alespoň jedna kopie v geografickém umístění s rozdílným typem „hrozby katastrofy“. Proces monitorování zastarávání úložišť a médií.	Alespoň 3 kopie v geografickém umístění s rozdílným typem „hrozby katastrofy“. Zavedený komplexní plán pro udržení dat a metadat na aktuálně přístupných médiích nebo úložných systémech.
Neporušenost souboru a integrita dat	Kontrola neporušenosti souboru při příjmu, pokud je k dispozici společně s přijímaným obsahem. Vytvoření informace o neporušenosti v případě, že nebyla poskytnuta společně s přijímaným obsahem.	Kontrola neporušenosti pro všechna přijímaná data. Použití blokátorů zápisu při práci s originálními médii. Antivirová kontrola u vysoce rizikového obsahu.	Kontrola neporušenosti ve stanovených intervalech. Správa logů s informacemi o neporušenosti; audit na vyžádání. Schopnost rozpoznat poškozená data. Antivirová kontrola veškerého obsahu.	Kontrola neporušenosti veškerého obsahu ve vztahu ke konkrétním událostem nebo aktivitám. Schopnost nahradit/opravit poškozená data. Ujištění se, že žádná osoba nemá právo zápisu u všech kopií.
Informační bezpečnost	Identifikace oprávněných osob ke čtení, zápisu, přesunu a mazání jednotlivých souborů. Omezení přístupu k jednotlivým souborům.	Dokumentace omezení přístupu k obsahu.	Správa logů uživatelů a jejich provedených akcích na souborech, včetně mazání a činností uchovávání.	Audit logů.
Metadata	Katalog obsahu a úložných lokalit dat. Zabezpečení záloh katalogu a uložení v lokalitě nezávislé úložiště dat.	Uchovávání administrativních metadat. Uchovávání transformačních metadat a logování událostí.	Uchovávání standardních technických a popisných metadat.	Uchovávání standardních metadat o uchovávání.

Souborové formáty	V případě možnosti ovlivnění vytvářených souborů na vstupu se doporučuje používat omezenou množinu zavedených otevřených souborových formátů a kodeků.	Soupis používaných souborových formátů.	Monitorování problémů při zastarávání souborů.	Provádění přesunu souborů (*migration*), napodobení (*emulation*) a obdobných potřebných aktivit.
Přístup	Definování určené skupiny.[11] Schopnost zabezpečení materiálu během zpřístupnění. Může se jednat o opatření k fyzickému zabezpečení (např. personální obsazení v čítárně) a/nebo k elektronickému zabezpečení (např. uzamčená pracovní stanice, omezení pro stahování materiálů, omezení přístupu na IP adresy atd.). Schopnost identifikace a úpravy osobních údajů a jiných citlivých informací.	Veřejně přístupné katalogy, vyhledávací nástroje, inventáře, nebo popisy sbírek pro vyhledávání materiálů badateli. Vytvoření vstupního informačního balíčku (SIP) a archivního informačního balíčku (AIP) po přijetí.[12]	Schopnost vygenerovat výstupní informační balíček (DIP) na příjmu.[13] Uchovávání vysvětlujících informací a informací o uchovávání (PDI).[14] Veřejně dostupná politika přístupů.	Schopnost zabezpečení přístupu ke zastaralým médiím v jejich nativním prostředí a/nebo napodobením.

[11] **Určená skupina** v podstatě odkazuje na „uživatele“; termín byl převzat z referenčního modelu pro Otevřený archivační informační systém (OAIS).

[12] **Vstupní informační balíček** (SIP) představuje obsah a metadata přijaté od tvůrce dat úložištěm pro dlouhodobou ochranu. **Archivní informační balíček** (AIP) je souborem obsahu a metadat spravovaných úložištěm pro dlouhodobou ochranu a uspořádaný způsobem, který umožňuje repositáři provádět ochranné akce.

[13] **Výstupní informační balíček** (DIP) je zaslaný koncovým uživatelům jako odpověď na jejich požadavek a může obsahovat jeden nebo více AIP balíčků.

[14] **Vysvětlující informace** na jakýkoliv software, algoritmy, standardy a další informace, které jsou nezbytné pro korektní přístup k archivovaným digitálním souborům. Nebo, jak to definují „Preservation Metadata“ a OAIS: „Digitální objekt je složený z řady posloupností bitů; vysvětlující informace přiřazují význam těmto bitům.“ **Informace o uchovávání** odkazuje na informace potřebné k dostatečnému uchování digitální objektu. Například, informace o původu, informace o identifikátorech, neporušenost, informace o souvislostech a informace o přístupových právech.