Metodika logické ochrany digitálních dat

Hutař, Jan
2018

Dostupný z http://www.nusl.cz/ntk/nusl-371612

Metodika logické ochrany digitálních dat

Autoři: Jan Hutař, Andrea Miranda, Eliška Pavlásková, Zdeněk Vašek, Zdeněk Hruška

Předmluva

Metodika pro logickou ochranu digitálních dat předkládá postupy pro dlouhodobé uchovávání těchto dat zejména z knihovních sbírek pomocí softwaru ARCLib.[1] Cílem metodiky je popsat životní cyklus uložených dat, definovat proces péče o jejich ochranu a předložit postupy, které lze na ochranu dat uplatnit. Definuje potřebné kroky k naplnění cíle dlouhodobého uchování digitálních dat na úrovni logické ochrany[2] v souladu s postupy doporučenými mezinárodními normami ČSN ISO 14721 a ČSN ISO 16363. Metodika vychází z těchto norem a představuje jejich aplikaci na konkrétním systému pro dlouhodobé uchovávání. Podobný dokument metodické povahy s aplikovanými postupy v českém prostředí dosud chyběl.[3]

Logická ochrana dat je v této metodice vnímána jako souhrn aktivit, opatření, procesů a nástrojů, které přispívají k zachování dlouhodobé použitelnosti a srozumitelnosti digitálního obsahu. Je třeba ji odlišovat od bitové ochrany, jejímž účelem je zajistit bezpečné a neměnné uložení binárních streamů, kterými jsou uchovávaná data reprezentována. Logická ochrana představuje stále ještě poměrně nový pojem, i když její význam radikálně roste. Přináší zcela jiné a koncepčně odlišné postupy, než jaké byly dosud obvyklé. Ochrana není automaticky zaručena využitím konkrétního ověřeného nástroje a patřičným hardwarovým vybavením a není ani možné na základě měřitelných parametrů přesně odhadnout dobu, po kterou jsou dokumenty chráněné. Jádrem dlouhodobé ochrany proto zůstávají kvalifikovaní pracovníci využívající mezinárodně osvědčené a zdokumentované postupy ideálně ve spojení s diverzifikovanými technologiemi. Účelem této metodiky je nejen poskytnutí návodu, jak provádět správu a ochranu digitálních dat v rámci specifického softwarového řešení ARCLib, ale i popis obecných postupů k provádění kroků logické ochrany a definování nároků na personální zajištění chodu úložiště. Mezi klíčové vlastnosti systémů poskytujících logickou ochranu patří důvěryhodnost, kterou lze zajistit mj. dodržováním norem a transparentním chodem systému (ve smyslu odborně zajištěné obsluhy a zdokumentovaných procesů i používaného softwaru). Toto je ověřováno řadou certifikačních nástrojů. Součástí předkládané metodiky jsou tedy i doporučené postupy, jak věrohodným způsobem provádět certifikaci digitálních repozitářů.

Metodika přináší uživatelům systému ARCLib pro dlouhodobé uchovávání dat definici postupů, jak výše zmíněné principy aplikovat, jak provádět správu dat v systému a hodnotit rizika. Zároveň dokumentuje konkrétní funkce řešení, popisuje způsob uložení dat, jejich identifikaci a strukturu. Právě tato

[1] ARCLib využívá v maximální míře existující volně dostupné nástroje, jako jsou ProArc a Archivematica, a to především pro tvorbu vstupních SIP. Archivematica je open source systém pro dlouhodobou digitální archivaci vyvíjený společností Artefactual Systems Inc. – https://www.archivematica.org/en/. Vzhledem k překotnému vývoji, řadě nedostatků a současnému stavu vývoje systému Archivematica je v současné době riskantní stavět toto řešení do jádra plánovaného systému ARCLib.

[2] Logická ochrana digitálních dat je ochrana jejich intelektuálního obsahu. Protipólem je ochrana bitová, kde cílem je zachovat, tzn. ochránit digitální objekt ve stavu, ve kterém byl uložen. Uchovávají se bity. Logická ochrana neklade důraz na formu objektu, tedy ani na jeho formát, který se může v průběhu doby měnit, nýbrž na intelektuální obsah dokumentu, který musí zůstat neměnný, použitelný a pochopitelný i navzdory technologickým změnám nebo změnám vnímání tzv. cílové skupiny, které je obsah určen.

[3] 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 (Dvořák et al., 2015).

dokumentace je nezbytná pro uznání důvěryhodnosti repozitářů využívajících systém ARCLib. Metodika tak vychází z mezinárodně definovaných požadavků na systémy logické ochrany digitálních dokumentů a převádí je na konkrétní postupy dostupné v rámci vyvinutého řešení. Využívání návodů této metodiky ve všech zmíněných oblastech (provoz řešení, personální a finanční zajištění a příprava na certifikaci) představuje nezbytnou podmínku pro využití softwarového řešení ARCLib. I když jsou postupy logické ochrany obecně popsány, v každém konkrétním případě se jejich aplikace liší na základě rozdílného charakteru softwarového systému.

Pravidla a postupy metodiky byly ověřeny jejími autory jak v souvislosti s jinými systémy, tak s jejich vlastními badatelskými postupy a při vývoji řešení ARCLib.

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.

1. Metodika

1.1 Obecná/teoretická část

1.1.1 Plánování vybudování systému pro dlouhodobou digitální archivaci

Vybudování funkčního systému pro dlouhodobou archivaci je úkol, před který může být postaveno relativně mnoho institucí, nejen paměťových. Nejčastěji se s tímto zadáním setkávají knihovny a archivy. Může se jednat o knihovny oborové, univerzitní nebo krajské až po knihovny národní. Podobná situace nastává v archivech. Se systémy pro dlouhodobou archivaci se ale také v posledních letech můžeme setkat jak v soukromých (např. banky, velké korporace), tak vládních organizacích (např. ministerstva).

Vybudování takového systému může mít několik podob odvíjejících se od toho, zda se instituce rozhodne systém pro dlouhodobou ochranu digitálních dat nakoupit již jako hotový produkt, nebo se rozhodne takový systém vytvořit vlastní. Oba přístupy budou mít shodné části, např. vybudování IT infrastruktury nebo vytvoření procesní a jiné dokumentace Odlišovat se ale mohou ve způsobu údržby a rozvoje systému, náročnosti implementace apod. Hotové systémy (komerční nebo open source) mají výhodu ve fungující komunitě, která systémy dále směruje, vyvíjí a poskytuje cenné rady. Tato komunita samozřejmě neexistuje v případě systému vyvíjeného vlastními silami. Oba přístupy se také odlišují nároky na zaměstnance. Pro vývoj je potřeba mít stálý tým vývojářů, pro hotový systém nic takového není nutné. U hotových systémů je nutné si dát pozor, aby nedošlo k tzv. vendor-lock-in, tedy že se instituce díky uzavřenosti systému stane doslova závislá na jeho producentovi nebo prodejci. Systémy musí podporovat běžné standardy a koncepty, instituce sama si musí vytvořit proveditelný exit plán, tedy scénář, kdy bude nutné opustit dosavadní systém a data migrovat do systému nového.

Scestná je představa, že pouhé pořízení nebo vývoj konkrétního systému vyřeší problém dlouhodobého uchování v konkrétní organizaci. Systém a jeho technické zázemí je jen jednou vrstvou, kterou je potřeba vyřešit. Druhým a možná ještě důležitějším aspektem budování systému na dlouhodobou ochranu je stránka organizační a procesní. Ta hraje velkou roli jak před samotným započetím budování systému, tak i poté, co je systém na místě a v provozu.

Z pohledu ukotvení celé problematiky digitální archivace v organizaci je nutné, aby tato byla součástí priorit a zaměření instituce. Ideálně by měla být také takto popsána ve statutu organizace a vyjádřena v organizační struktuře s odpovídajícím personálním zabezpečením. Budování systému v určitém okamžiku zákonitě přejde do jeho rutinního provozu, na což je třeba myslet už během vývoje a plánování, zvláště pokud je systém na digitální archivaci budován v rámci časově a finančně omezeného projektu. Je třeba uvažovat nad tím, jaké procesy a jakým způsobem budou probíhat po ukončení projektu. Do těchto úvah musí spadat provozní náklady, další vývoj a rozšiřování systému, obnovování digitálního úložiště a vybavení, personální náklady a rozvoj, případně pořízení následnického systému v budoucnu. S touto otázkou se pak pojí otázka migrace ze stávajícího do nového systému.

Budování nového systému na dlouhodobou archivaci digitálních dat v konkrétní instituci bude vždy probíhat podle zvyklostí, procesů a postupů dané instituci vlastních. V každém případě by ale proces měl následovat a odpovídat zavedeným normám (IT infrastruktura, standardy pro bezpečnost apod.).

Základním konceptuálním rámcem pro jakýkoliv digitální archiv je ISO standard pro Otevřený archivační informační systém (ČSN ISO 14721). Tento standard popisuje digitální archiv, jeho části, funkcionalitu

i datový model; a je klíčovým standardem, se kterým pracují, i se z něj odvíjejí další systémy a standardy, jako např. metadatové schéma PREMIS (PREMIS: Preservation metadata maintenance activity, 2017). Druhou oblastí možné inspirace je proces certifikace (případně auditu) digitálních archivů. V posledních letech vzniklo několik návodů pro audit a certifikaci, některé z nich jsou platné jako ISO standardy. Většina z nich obsahuje výčet požadavků, vlastností a funkcionalit, které má digitální archiv ucházející se o certifikaci nebo úspěšný audit plnit. Právě tyto požadavky lze velmi dobře využít již při plánování a následném budování systému na dlouhodobou ochranu digitálních dat. Tímto způsobem lze zajistit, že výsledek, tedy nový digitální archiv, bude odpovídat požadavkům na tyto systémy kladeným. A také bude připraven na certifikaci nebo audit, které jsou často požadovány za účelem potvrzení kvality a funkcionality digitálních archivů.

Jako příklady nástrojů a metodik na provedení auditu lze uvést např. DRAMBORA (DCC, DPE, 2008), DataSeal of Approval,[4] Nestor Seal (Nestor Seal for Trustworthy Digital Archives, 2017), SPOT (Simple Property-Oriented Threat) (Vermaaten et al., 2012), NDSA Levels of Digital Preservation (Phillips et al., 2013). Jeden z prvních dokumentů, který se certifikací zabýval a obsahoval seznam požadavků, vyšel v roce 2007 pod názvem Trusted Repository Audit and Certification: Criteria and Checklist (OCLC and CRL, 2007). Během doby prošel mnoha úpravami, v roce 2012 byl publikován jako ISO 16363:2012 (dostupná v českém překladu jako ČSN ISO 16363 (319621)). Pro ISO 16363 existují nástroje, které mají pomoci provádění jak externího, tak interního auditu. Norma ISO 16363 je velmi často používána pro interní audit, nástroje v podobě excelových tabulek[5] nebo pracovních sešitů v podobě wiki jsou tak velmi užitečné.[6]

Oblasti, na které se tyto certifikační procesy zaměřují, jsou následující:[7]

- organizační stránka,
- správa digitálních objektů/dat,
- technologie, infrastruktura a bezpečnost.

Je tedy zřejmé, že samotné technické řešení není dostačující, je nutno jej doplnit odpovídajícími procesy a postupy, spolu s podporou v rámci organizace. ISO 16363 je dále podrobněji popsána v kapitole 1.2.6.1 ČSN ISO 16363.

V červenci 2010 vznikl Evropský rámec pro audit a certifikaci digitálních repozitářů (European Framework for Audit and Certification of Digital Repositories). Dohodu podepsali zástupci CCSDS/ISO Repository and Audit Certification Working Group/RAC), Data Seal of Approval Board a německé DIN Working Group "Trustworthy Archives – Certification". Tento rámec stanovil tři základní úrovně auditů a certifikace digitálních repozitářů (Data Seal of Approval: Evropský rámec pro audit a certifikaci):

- Základní certifikace – udělena repozitářům, které dostojí směrnicím Data Seal of Approval (DSA);
- Rozšířená certifikace – udělena repozitářům, které již splnily Základní certifikaci, a navíc provedly za externího dohledu strukturovaný a veřejně dostupný "self-audit" podle ISO 16363 nebo DIN 31644;

[4] (Data Seal of Approval) v českém překladu (Data Seal of Approval: Český překlad Směrnic DSA, 2017).
[5] Nástroj v podobě tabulky MS Excel je dostupný např. zde: https://github.com/databrary/curation/blob/master/spec/projects/certifications/dsa/reference/Self-AssessmentTemplateforISO16363.xls.
[6] Příklad dobré praxe, viz CLOCKSS Archive (CLOCKSS Archive: documentation Wiki, 2014).
[7] Jednotlivé body uvedeny dle TRAC.

- Formální certifikace – po dosažení Základní certifikace repozitář nechal provést plně externí, nezávislý audit a certifikaci dle ISO 16363 nebo DIN 31644.

Na proces budování digitálních archivů se přímo zaměřuje jen velmi málo dokumentů. Starší (2007), ale stále relevantní je seznam Deseti zásad tzv. důvěryhodného repozitáře. Tyto zásady lze vztáhnout na jakýkoliv digitální repozitář nebo archiv (Ten principles). Principy říkají, že repozitář (Rosenthal et al., 2009):

1. trvale spravuje digitální objekty pro *určenou komunitu/komunity*;
2. musí prokázat organizační způsobilost pro tento úkol (např. financování, personální zajištění a řízení);
3. dostojí smluvním a právním požadavkům a splní povinnosti z nich vyplývající;
4. má vypracované účelné a účinné metody, strategie a zásady;
5. získává a zpracovává digitální objekty podle stanovených kritérií, která odpovídají jeho cílům a možnostem;
6. udržuje a zajišťuje dlouhodobou integritu, autenticitu a použitelnost uložených digitálních objektů;
7. uchovává potřebná metadata o všech akcích, které byly s digitálními objekty v průběhu jejich uložení provedeny; také shromažďuje související informace o vzniku, podpoře dostupnosti a využívání objektů před jejich uložením v repozitáři;
8. naplňuje potřebná kritéria pro zpřístupňování;
9. má strategický program pro plánování ochrany a ochranné aktivity;
10. má odpovídající technické zázemí, potřebné k trvalému udržování a zabezpečení uložených digitálních objektů.

Přímo jako pomůcka pro plánování a budování digitálních archivů byl vytvořen Planning Tool for Trusted Electronic Repositories (PLATTER) (DigitalPreservationEurope, 2008). Vznikl v rámci evropského projektu DigitalPreservationEurope a je dostupný i v českém překladu (Rosenthal et al., 2009). Zaměřuje se na definici cílů digitálního archivu pro jednotlivé tematické okruhy, včetně uvedených návodných příkladů. Aspekty managementu a realizace těchto cílů nebyly do PLATTERu záměrně zahrnuty. Níže jsou uvedeny podrobnosti z PLATTERu, relevantní pro plánování a budování digitálního archivu jakéhokoliv typu.

1.1.1.1 Klasifikace repozitářů

Všechny digitální archivy nemají stejné cíle. Je klíčové určit, jaký typ digitálního archivu bude budován, na co se bude zaměřovat a co naopak nebude až tak důležité.[8] PLATTER se zaměřuje na tyto čtyři oblasti: 1) účel a funkce repozitáře; 2) jeho velikost; 3) provozní otázky, 4) technická řešení a implementace.

1. Účel a funkce repozitáře
    - Mandát a pověření – existuje mandát digitálního archivu v rámci instituce? Kdo je jeho původcem?
    - Status – digitální archiv by měl mít status – je budován pro komerční nebo jiné cíle?
    - Právní podmínky získávání obsahu – tedy způsoby akvizice. Vyplývají z konkrétního zákona nebo z jiných předpisů? Nebo jsou data získávána na základě smluv, či dobrovolně? Od koho?

[8] Některé archivy kladou důraz na zpřístupnění obsahu, jiné od počátku budují tzv. dark archive, tedy bez zpřístupnění.

2. Velikost repozitáře
    - Množství dat – jaký je předpokládaný objem dat pro uložení? Jaký je plánovaný roční nárůst objemu? Kolik jednotlivých souborů bude do archivu proudit v konkrétním časovém úseku? V jakém řádu se budou pohybovat počty uložených souborů?
    - Lidské zdroje – jaký bude počet zaměstnanců určených pro provoz archivu? Kdo se na provozu archivu bude podílet zvenčí, tedy mimo organizaci? Kolik bude pravidelných uživatelů?
3. Provoz
    - Metody akvizice – jsou data pro archivaci získávána zvenčí nebo jsou vytvářena organizací provozující archiv? Pokud zvenčí, jsou data aktivně vkládána nebo archiv data sám aktivně získává?
    - Komplexnost dat – jaká data budou do archivu akceptována či zasílána? Jednoduchá (např. jednoduché textové formáty, obrázky či video); středně komplexní (složené dokumenty s množstvím vazeb mezi jednotlivými částmi); velmi komplexní data (např. software, texty s vloženými tabulkami, celé internetové weby, databáze aj.).
    - Specializace dat – do jaké míry je k použití a interpretaci dat v archivu nutné mít expertní znalosti? Jak to bude ovlivňovat jejich dlouhodobou ochranu?
    - Citlivost dat – jaké etické a právní normy jsou relevantní pro ukládaná data, které bude nutno dodržovat?
    - Oprávnění přístupu – jaká data budou převládat, otevřená a volně dostupná pro každého, nebo s omezeným přístupem?
4. Technická řešení a možnosti implementace
    - Zdroj metadat – odkud přicházejí a jak jsou získávána potřebná popisná, administrativní, ochranná a jiná metadata? Od producentů, nebo automatizovanou extrakcí pomocí konkrétních nástrojů?
    - Interoperabilita – jak bude digitální archiv komunikovat s ostatními systémy? Jak bude interoperabilita podporována z pohledu použitých formátů dat a metadat?
    - Strategie ukládání – bude archiv mít vlastní datové úložiště, nebo v něm budou data ukládána formou nákupu služby od externího dodavatele? Je údržba povinností provozovatele archivu, nebo dodavatele služby? Jak často bude nutno úložiště obměňovat za nové?
    - Správa softwaru – jak jsou do archivu získávány, udržovány a provozovány nástroje nezbytné pro chod archivu i správu a ochranu v něm uložených dat? Je systém vlastního archivu podporován externím dodavatelem, nebo vlastním provozovatelem archivu, případně externí komunitou?

1.1.1.2 Plánovací cyklus PLATTER

V okamžiku, kdy je definován digitální archiv, lze použít metodiku PLATTER a začít plánovat další detaily, specifikovat cíle archivu a jeho činnosti. PLATTER tento proces plánování vidí jako opakovatelný proces (Rosenthal et al., 2009, s. 17), kdy specifikované cíle jsou realizovány, hodnoceny a případně přeformulovány a provedeny znovu. Pro budování digitálního archivu jsou důležité následující oblasti:

- strategické plánování,
- specifikace cílů nebo účelu – operační plánování,
- vypracování plánů,
- realizace plánu, hodnocení/úprava a implementace.

1.1.1.3 Plány strategických cílů

PLATTER spojuje soubor plánů strategických cílů s výše uvedenými Deseti principy důvěryhodného repozitáře. Při plánování digitálního archivu lze vytvářet vlastní plány strategických cílů, pokud oblast není obsažena v doporučených cílech PLATTER, viz níže (Rosenthal et al., 2009, s. 21).

Plán strategických cílů	**Odpovědnost**	**Odpovídající základní principy**
Finanční plán	Zabývá se plánováním, monitorováním a vykazováním financí.	2
Plán řízení lidských zdrojů	Zabývá se získáváním a udržováním souboru dovedností, které jsou relevantní pro správu repozitáře.	2
Datový plán	Specifikuje datové a metadatové objekty, formáty, struktury pro vkládání, uchovávání a zpřístupňování dat a související transformace a mapování dat.	5, 6, 7, 8
Akviziční plán	Zabývá se vztahy s depozitory a dalšími poskytovateli dat.	3, 5
Plán zpřístupňování	Zabývá se vztahy s koncovými uživateli a pravidly pro zpřístupňování.	1, 8
Plán ochrany	Zajišťuje zpřístupňování a použitelnost dokumentů bez ohledu na zastarávání nebo změny technologií.	9
Technický plán	Specifikuje požadavky na hardware, software a síťové systémy.	10
Plán zajištění kontinuity	Zabývá se povinností a zajištěním ochrany dokumentů i po zániku repozitáře.	1
Krizový plán	Reaguje na náhlé změny v prostředí repozitáře.	1, 6

Tabulka 1 – Plán strategických cílů PLATTER

Níže jsou uvedeny jednotlivé druhy Plánů strategických cílů s příklady:

- **Finanční plán**
  - Pravidelně sledovat a revidovat finanční plán.
  - Udržet financování na úrovni, kterou vyžaduje běžný provoz archivu.
    - Jaké jsou fixní náklady na provoz, jaké finance jsou nezbytně nutné na splnění závazků archivu (viz kapitola 2.3.3 Doporučení pro odhad finančních nákladů na provoz systému pro dlouhodobé ukládání).
  - Vytvořit nouzové plány pro případ finančních omezení nebo krizí tak, aby byla data za každých okolností ochráněna.
    - Jaké oblasti a činnosti budou v krizi upřednostněny a jaké naopak potlačeny, jak zajistit zpřístupňování a použitelnost dat.
  - Stanovit a naplňovat marketingové plány spolu s externí komunikací tak, aby odpovídaly potřebám archivu.
    - Komunikace s depozitory, uživateli nebo externími stranami.

- **Akviziční plán**
  - Získat relevantní dokumenty.
    - Určit kolik dokumentů a jaké bude získáváno za určité období (dle mandátuaj.), analýza trhu, analýza depozitorů a zájmu uživatelů, analýza nákladů a případných výnosů.
  - Vyjednat dohody o uložení.
    - Smluvní zajištění získávání dokumentů od depozitorů, pokud není přímo uloženo zákonem, specifikace množství dokumentů, formátu, způsobu předání, přiložených metadat, intelektuálních práv apod.
  - Získat fyzickou kontrolu nad dokumenty.
    - Mít pracovní postup získávání dat.
  - Monitorovat akvizici.
    - Kontrola plnění dohod o uložení a procesu ukládání.
  - Zajišťovat aktuálnost smluv o uložení.

- **Plán řízení lidských zdrojů**
  - Definovat zaměstnanecké pozice, odpovědnosti a pravomoci pracovníků archivu (viz kap. 1.3.2 Doporučení pro organizační a personální zajištění projektu).
  - Získat a udržet zaměstnance pro práci na specifických pozicích.
    - Zajištění kvalifikovaných zaměstnanců tak, aby nebyly porušovány závazky archivu.
  - Rozvíjetkvalifikaci zaměstnanců.
    - Rozvíjení a udržování kvalifikace musí být prioritou archivu.

- **Plán zpřístupňování**

  Ne každý archiv má za cíl zpřístupňovat obsah, ale pokud je zpřístupnění jedním z cílů, je třeba mít jasno v tom, jakým způsobem, kdo je určená skupina a jaké má nároky a očekávání, jak budou zajištěna autorská a jiná práva apod.

  - Formulovat, udržovat a aktualizovat programové prohlášení odpovídající mandátu digitálního archivu.

        - Vytvořit a schválit programové prohlášení archivu a revidovat jej po určité době, definuje cíle a závazky archivu.
        - Měl by obsahovat formulaci cílů archivu, závazek dlouhodobého uchovávání, správy a zpřístupnění dokumentů.
    - Definovat komunitu nebo komunity uživatelů archivu, porozumět jejich potřebám a dovednostem.
        - Na webové stránce by měla být určená komunita popsána včetně možností a limitů získání dokumentů.
    - Formulovat a implementovat politiku zpřístupňování obsahu archivu.
        - Bude ovlivněno mj. autorským právem, obsahem citlivých dat a další legislativou, zahrnuje způsob autorizace a minimální požadavky na metadata.
    - Specifikovat a realizovat technologické požadavky distribuce a zpřístupňování.
        - Vytyčení cílů pro podporu zpřístupňování, jaká jsou minimální metadata akceptovatelná archivem, jak bude probíhat vyhledávání a distribuce dokumentů, systémové nároky na zpřístupnění.

- **Technický plán**
    - IT infrastruktura si musí umět poradit s takovým rozsahem ukládání, zpracování a přenosu dat, který odpovídá potřebám daného archivu.
        - Archiv musí udržovat infrastrukturu na takové úrovni, která umožní dostát závazkům, úkolům a procesům (zátěži) v archivu a musí být schopna reagovat na pozvolné i velmi rychlé změny nároků.
        - Jsou hardwarové, softwarové a síťové systémy dostatečné pro to, čeho chce archiv dosáhnout?
    - Infrastruktura IT musí garantovat integritu a bezpečnost uložených dat.
        - Plány záloh, plány pro případy havárie a přírodních nebo jiných katastrof, plán a strategie bezpečnosti proti útokům zvenčí i zevnitř, plán auditu.
    - Infrastruktura IT musí garantovat dostupnost daných služeb pro uživatele.
        - Plán řešení problémů při dostupnosti archivu a zpřístupnění, smlouvy s poskytovateli podpory HW, SW, připojení, SLA smlouvy o délce trvání možných výpadků, architektura pro předcházení problémům (např. duplicitní virtuální prostředí, testovací prostředí apod.).
- **Datový plán**
    - Určit, jaké formáty digitálních objektů bude archiv akceptovat a přijímat (SIP).
        - Doporučení může být obecné pro veškeré příchozí dokumenty, nebo specifické pro konkrétního producenta dat nebo typ producenta či typ dat.
        - Specifikace musí obsahovat jak formáty dat, tak detailní nároky na jejich podobu.
        - Je vhodné si určit, zda, jak a jaké formáty budou v rámci archivu transformovány do jiných formátů.
    - Specifikovat zdroje a formáty bibliografických a popisných metadat pro SIP.
        - Popisná metadata jsou klíčová pro kontext, vyhledávání, popis toho, čím objekt je a co obsahuje.
    - Specifikovat technická metadata pro SIP.
        - Určit, jaká technická metadata budou očekávána od producenta, jaká metadata vytvoří archiv automaticky, jaké budou probíhat validace (formátů) a pomocí jakých nástrojů.

- Specifikovat formát dat a obsah metadat pro archivaci digitálních objektů (AIP).
    - Určit kupříkladu, v jakých formátech se budou ukládat zvukové dokumenty, v jakých textové apod.
- Specifikovat metadata pro AIP.
    - Podobně musí být jasné, jak přesně vypadají a co obsahují metadata AIP, v jakém formátu (např. XML) a kde budou uložena, jaké události budou zapisovány do AIP metadat apod.
- Specifikovat formáty dat pro digitální objekty distribuované uživatelům (DIP).
    - Stanovit nejvhodnější formáty dat pro koncové uživatele. Tyto nemusí nutně být stejné jako formáty pro AIP, mohou to být jiné formáty, např. s kompresí apod. Cílem je zjednodušit přístup k obsahu archivu pro uživatele.
- Specifikovat metadata pro DIP.
    - Metadata pro koncového uživatele jsou často pouze podmnožinou AIP metadat a často mohou být důležitější než samotná data.
- Specifikovat transformaci SIP do AIP.
    - Definovat postupy vytváření AIP ze SIP, popsat, jaká metadata obsažená v SIP budou použita a přenesena do AIP, jaká metadata budou nově vytvořena, co se bude dít s daty (migrace např.), validace dat, identifikace formátů, antivirová kontrola aj.
- Specifikovat transformaci AIP do DIP.
    - Jak vytvořit DIP z AIP– může jít o podmnožinu dat a metadat, případně DIP v některých případech může odpovídat AIP jedna k jedné.
    - Jak bude zajištěna autenticita obsahu DIP i přes případné migrace z původního AIP.

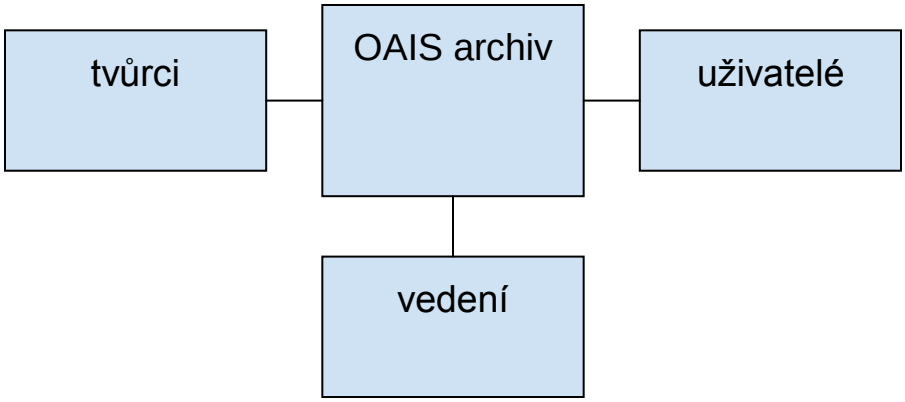
- **Plán zajištění kontinuity**
    - Ochrana digitálních dat by měla být zajištěna po dobu, která překračuje existenci digitálního archivu.
        - Mělo by být jasné, co se stane s obsahem archivu ve chvíli, kdy instituce archiv provozující např. přestane existovat. Dohoda s jinou institucí, obsahující specifikaci licenčních podmínek, závazky nového archivu, případné kompenzace pro nový archiv.
- **Krizový plán**
    - Digitální archiv včas reaguje na podstatné změny prostředí.
        - Vypracovat konkrétní analýzu rizik a strategii managementu rizik (přírodní katastrofy, katastrofy způsobené lidmi (záměrně nebo nechtěně), technologický otřes, ztráta důvěry uživatelů (např. únik citlivých dat), ztráta zaměstnanců a tedy know-how apod.
- **Plán ochrany**
    - Archiv musí mít přehled o současném a vznikajícím hardwaru, softwaru a technologiích ukládání dat.
        - Archiv by měl mít plán na posuzování použitelnosti dat na konkrétním hardwaru a softwaru, aby byl schopen reagovat a případně provádět akce ochrany (např. migrace na jiný hardware, do jiného formátu, nasazení nového softwaru apod.).
        - Dokumentace používaného hardwaru a softwaru musí být k dispozici.

- Archiv by měl udržovat srozumitelné informace o strukturálních standardech (např. kódování souborů) a formátech.
    - Analýza a identifikace formátů je nutnou součástí dlouhodobé ochrany dat, archiv musí mít seznam všech přijímaných a ukládaných formátů s detaily o nich.
- Archiv musí udržovat znalosti o komunitě uživatelů, jejich kompetencích a znalostní bázi.
    - Určená komunita uživatelů musí být identifikována, popsána včetně jejich očekávání, technologických nároků a dovedností. Tyto údaje musí být po určitém čase ověřovány a případně upravovány.
- Archiv by měl vědět, jaké nároky na dlouhodobou ochranu má každý typ uloženého informačního zdroje nebo třída dat.
    - Dokumentovat požadavky na obsah, chování, vzhled, kontext, srozumitelnost a interoperabilitu dat definováním minimálních a maximálních nároků přijaté strategie ochrany dat.
    - Specifikovat například, jak rozšířené a podporované mají být použité formáty dat, požadavky na infrastrukturu, zaměstnance, procesy (míra automatizace, míra chybovosti apod.).
- Archiv musí udržovat, provádět a hodnotit takové strategie dlouhodobé ochrany dat, které vyhovují konkrétním cílům dlouhodobé ochrany.
    - Vytvářet strategie ochrany v závislosti na konkrétních potřebách dlouhodobé ochrany. Tyto by měly být obecného rázu a připravené na okamžik, kdy dojde ke konkrétní události, kdy je strategii potřeba aktivovat a provést ochrannou akci (migrace, emulace, výměna hardwaru nebo softwaru).
- Archiv by měl na základě vypracované strategie pravidelně hodnotit, které informace mají být dále uchovávány.
    - Vytvořit kritéria, podle kterých bude možné posoudit nutnost ochrany pro konkrétní dokumenty, např. podle aspektů jako přínos pro organizaci a její cíl, původnost obsahu, autenticita obsahu apod.

1.1.2 Popis procesů a funkčních celků vyplývajících z OAIS

Otevřený archivační informační systém (OAIS) je výstupem snahy o vývoj formálních standardů pro dlouhodobé ukládání digitálních dat generovaných v kosmickém výzkumu. Consultative Committee for Space Data Systems (CCSDS) tento vývoj inicioval pro své vlastní potřeby. Cílem bylo mít standard, který umožní konzistentně popsat problematiku dlouhodobého uchovávání, procesy a koncepty, které zahrnuje. Referenční model OAIS byl publikován v roce 1999 pro veřejné připomínkování a poté jako interní dokument CCSDS (Lavoie, 2000). Model se brzy rozšířil i do dalších komunit, včetně knihovnické a archivářské, které jej vzaly za svůj pro potřeby digitálních repozitářů. V roce 2002 byl referenční model OAIS uznán jako mezinárodní norma ISO 14721. Revidované a doplněné vydání normy bylo publikováno v roce 2012 (Sierman, 2012). Český překlad byl vydán v srpnu 2014 Úřadem pro technickou normalizaci, metrologii a státní zkušebnictví jako česká technická norma ČSN ISO 14721.

Archiv OAIS obecně funguje v prostředí tří typů subjektů (Lavoie, 2000, s. 9) – vedení (Management), tvůrci (Producers), koncoví uživatelé (Consumers). Zvláštním typem koncového uživatele je určená skupina (Designated Community). Její členové by měli být schopni nezávisle porozumět archivované informaci v takové podobě, v jaké je archivována a zpřístupňována archivem OAIS.



Obrázek 1 – Prostředí OAIS (Lavoie, 2000, s. 9)

Vedení (Management)

Vedení by mělo formulovat, revidovat a v některých případech také vynucovat obecný rámec řídící činnosti OAIS archivu. Činností prováděnou vedením může být mj. strategické plánování, financování, strategický dohled nebo vyhodnocování pracovních postupů, výkonu a spojených rizik. Vedení zaručuje ochranu svěřených dokumentů, není ale zodpovědné za každodenní provoz archivu (tuto odpovědnost má jiná funkční komponenta uvnitř archivu).

Tvůrci (Producers)

Tvůrci jsou producenti obsahu. Mohou to být jednotlivci, organizace nebo systémy. Tito předávají obsah archivu OAIS k dlouhodobému uchování. Archiv OAIS a tvůrci tedy mj. vyjednávají o tom, jaký obsah a jaká související metadata budou do archivu OAIS dodána. Komunikace mezi OAIS a tvůrci se řídí tzv. dohodou o dodávání dat (Submission Agreement), která upravuje konkrétní podmínky předání dat (typ dodávaných informací, rozsah metadat dodávaných tvůrcem, informace o přesunu správy dat od tvůrce na archiv, nebo informace o omezení přístupu). V minulosti vznikly standardy PAIMAS (ISO 20652:2006) a PAIS (Smith, 2013), jako pokus popsat interakce mezi tvůrci a archivy OAIS.

Koncoví uživatelé (Consumers) a určená skupina (Designated Community)

Koncoví uživatelé jsou jednotlivci, organizace nebo systémy, používající informace uchovávané v archivu OAIS, a to několika způsoby – žádají například o asistenci, vyhledávají v archivu nebo požadují zpřístupnění archivovaných informačních objektů. Určená skupina je zvláštní typ koncových uživatelů. Členové určené skupiny by měli být schopni porozumět archivovaným informacím v OAIS v té podobě, v jaké jsou zpřístupňovány, tedy pouze za pomoci kontextu (metadat) poskytnutých OAIS archivem. Na druhou stranu OAIS archiv musí tedy uchovávat kontextové informace v podobě metadat tak, aby toto pochopení určené skupině umožnil. Je to jeden z jeho hlavních cílů.

1.1.2.1 Funkční model OAIS

Referenční model OAIS je kromě datového modelu také modelem funkčním, specifikuje a popisuje entity OAIS archivu a mimo něj a procesy s nimi spojené. Tyto procesy naplňují hlavní cíl archivu – tedy ochraňovat informace v dlouhodobém horizontu a zpřístupňovat je členům určené skupiny. Funkční model OAIS popisuje šest obecných služeb neboli funkčních celků (Functional Entities), které společně

definují fungování OAIS při ochraně a zpřístupňování: příjem (Ingest), archivní uložení (Archival Storage), správa dat (Data Management), plánování uchovávání (Preservation Planning), zpřístupnění (Access), správa (Administration). Těchto šest celků (často v literatuře označovaných jako *moduly*), je doplněno základními službami (Common Services), kterými jsou základní výpočetní kapacity a nástroje pro správu souborů, síťové služby (např. datové komunikační mechanismy), bezpečnostní služby (např. autentizační/autorizační služby), bez kterých se OAIS archiv neobejde a jsou nutnou podmínkou jeho fungování (Lavoie, 2000, s. 14).

Šest funkčních celků OAIS – dle (Lavoie, 2000, s. 12-13):

Příjem (Ingest)

Funkční celek příjem je rozhraním mezi systémem OAIS a tvůrci, řídí celý proces přejímání vkládaných informací do správy a přípravu na jejich archivní uložení. Jedná se např. o přijetí informací předaných OAIS tvůrcem; ověření, zda přijímané informace jsou kompletní a nepoškozené; transformaci dodaných informací do podoby vhodné pro uložení a správu; extrakci a/nebo vytvoření popisných metadat a přesun vkládaných informací a souvisejících metadat do archivního skladu.

Archivní uložení (Archival Storage)

Funkční celek archivní uložení nemá žádné přímé vnější rozhraní, interakce s archivním uložením je omezena na vnitřní služby archivu OAIS. Řídí dlouhodobé úložiště a vykonává správu digitálního materiálu svěřeného OAIS archivu. Tento funkční celek je odpovědný za zajištění toho, že se archivovaný obsah nachází na odpovídajícím typu úložiště – například online, near line, offline, a že sekvence bitů tvořící ochraňované informace je i po dlouhé době úplná a zobrazitelná. Provádí pravidelné procesy, jako je obnova médií nebo migrace souborových formátů. Také spouští ochranné mechanismy pro kontroly výskytu chyb, hodnocení výsledků ochranných akcí, anebo obnovu po havárii ke zmírnění dopadů katastrofických událostí. Kromě toho archivní uložení plní požadavky koncových uživatelů na zpřístupnění a načítá objekty z úložných systémů.

Správa dat (Data Management)

Funkční celek správa dat udržuje databáze obsahující popisná metadata identifikující a popisující archivované informace a podporující vyhledávací mechanismy archivu OAIS. Spravuje (vytváří i validuje) také mj. administrativní metadata podporující operace interních systémů OAIS, jako jsou informace o výkonu systému nebo např. statistiky zpřístupňování. Hlavní funkce a odpovědnosti správy dat zahrnují správu databází, pokládání dotazů do těchto databází a generování odpovědí na základě požadavků jiných funkčních celků OAIS. Dále provádí pravidelné aktualizace databází po vložení nových informací do archivu nebo po změnách a smazání již archivovaných informací, tedy vlastně vytváření a správa administrativních metadat. Udržováním databází podporuje správa dat vyhledávání a dodávání obsahu uloženého v OAIS archivu a řízení vnitřního provozu OAIS.

Plánování uchovávání (Preservation Planning)

Plánování uchovávání je funkční celek zabývající se vytvářením strategií uchovávání a jejich aktualizací. Klíčové pro plánování je monitorování změn ve vnějším prostředí a identifikace rizik, která by mohla mít dopad na schopnost archivu OAIS uchovat a zpřístupnit uložené informace. Takové změny mohou zahrnovat např. inovace v ukládacích a zpřístupňovacích technologiích nebo také změnu očekávání a rozsahu určené skupiny. Plánování uchovávání vytváří doporučení pro aktualizace strategií a pracovních postupů archivu OAIS tak, aby byl archiv schopen se těmto změnám přizpůsobit.

Zpřístupnění (Access)

Zpřístupnění je vnějším rozhraním archivu OAIS směrem ke koncovým uživatelům (a určené skupině). Funkční celek zpřístupnění je odpovědný za řízení procesů a služeb, které používají koncoví uživatelé, především členové určené skupiny, k nalezení a podání žádosti o dodání a získání objektů uložených v archivním skladu. Služby poskytované zpřístupněním zahrnují typicky zpracování požadavků na archivované objekty, konkrétně zprostředkování dotazu do správy dat a dodání odpovědi (např. výsledného souboru) koncovému uživateli, koordinaci dodání požadovaného obsahu a předání požadavku do archivního uložení. Součásti procesu může být také provedení nezbytných transformací, které je třeba udělat před dodáním koncovému uživateli (změny souborových formátů na vhodnější pro zpřístupnění nebo odstranění nepotřebných metadat). Funkční celek zpřístupnění je také odpovědný za implementaci bezpečnostních mechanismů a přístupových omezení k archivovanému obsahu.

Správa (Administration)

Funkční celek správa je klíčový pro každodenní provoz OAIS. Je centrálním místem komunikace uvnitř OAIS a navenek komunikuje přímo s pěti dalšími funkčními celky – příjem, archivní uložení, správa dat, plánování uchovávání a zpřístupnění a také s externími subjekty – s tvůrci, koncovými uživateli a vedením. Dále odpovídá za komunikaci s tvůrci (např. vyjednává dohody o dodávání dat), s koncovými uživateli (tj. poskytuje zákaznickou podporu), s vedením (např. implementuje a udržuje politiky a standardy archivu). Funkční celek správa dohlíží na provoz archivních a zpřístupňovacích systémů, monitorování výkonu systému a podle potřeby na koordinaci aktualizace systému.

1.1.3 Informační balíčky v OAIS

Referenční model OAIS popisuje informační objekty spravované OAIS archivem. Informační model OAIS pracuje s konceptem informačního balíčku, který sestává z objektu,[9], který je předmětem ochrany, a z metadat, která jsou nezbytná pro zabezpečení dlouhodobé ochrany, zpřístupnění a zajištění srozumitelnosti konkrétního objektu (ČSN ISO 14721, 4.2.1.1). OAIS rozlišuje tři varianty informačního balíčku:

- Vstupní informační balíček (Submission Information Package – SIP),
- Archivní informační balíček (Archival Information Package – AIP),
- Výstupní informační balíček (Dissemination Information Package – DIP).

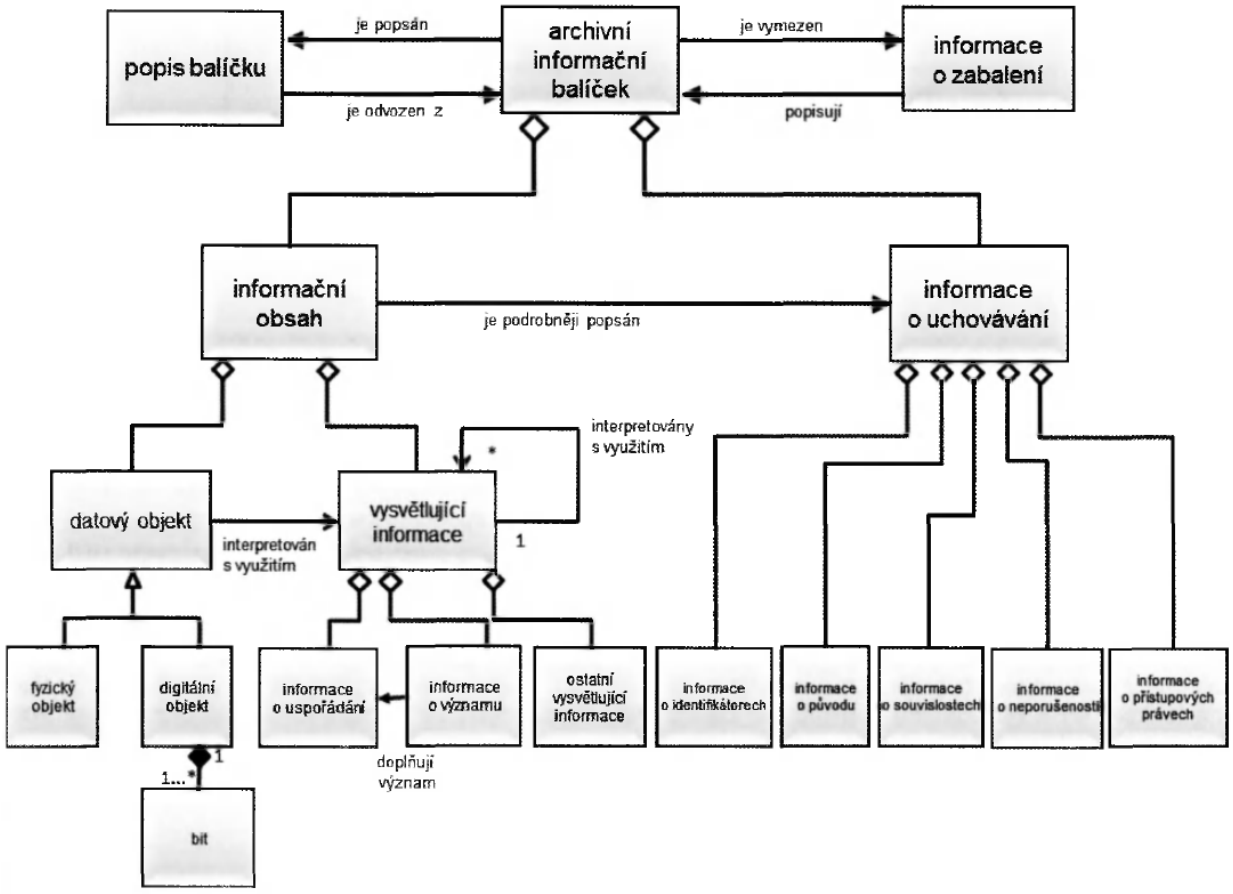
Hlavním předmětem zájmu a ochrany je AIP. Tento balíček by měl mít všechny vlastnosti nutné pro trvalé nebo časově neomezené dlouhodobé uchování informačního objektu (ČSN ISO 14721, 4.2.2.3). Obsah AIP tvoří datový objekt s obsahem (Content Data Object), což je informace, která je předmětem ochrany. Datový objekt s obsahem doprovází vysvětlující informace (Representation Information), tj. informace potřebná k zobrazení nebo pochopení sekvencí bitů, které tvoří datový objekt s obsahem. Datový objekt s obsahem a vysvětlující informace se společně nazývají informační obsah (Content Information). Dlouhodobá archivace informačního obsahu vyžaduje další metadata podporující a dokumentující ochranné procesy OAIS. Tato metadata se nazývají informace o uchovávání (Preservation Description Information (PDI)). PDI se skládají z pěti částí:

1. Informace o identifikátorech (Reference Information),
2. Informace o souvislostech (Context Information),

[9] Tento objekt může být dle OAIS fyzický i digitální.

3. Informace o původu (Provenance Information),
4. Informace o neporušenosti (Fixity Information),
5. Informace o přístupových právech (Access Rights Information).

Informace o zabalení (Packaging Information) spojuje informační obsah a informace o uchovávání do jednoho logického balíčku. Popisná informace (Descriptive Information) pak podporuje nalezení a získání informačního obsahu koncovými uživateli OAIS (Lavoie, 2000, s. 2).



Obrázek 2 – Archivní informační balíček (ČSN ISO 14721, Obrázek 4.18)

Vstupní informační balíček, SIP, je informace zabalená tvůrcem, který ji zasílá za účelem ochrany do archivu OAIS. Konkrétní podoba SIP by měla být výsledkem vyjednávání mezi tvůrcem a archivem OAIS, ale SIP může být vytvořen ad hoc, například podle stávajících možností Tvůrce. Koncept SIP zdůrazňuje skutečnost, že informace nemusí být uchována přesně v takové podobě, v jaké ji archiv přijal. Uchovávaný objekt může být například reprezentován obsahem vloženým ve více SIP, nebo tvůrce může poskytnout informace ve formátu, který archiv OAIS nepodporuje, což vyžaduje formátovou migraci před vložením do archivního skladu. Jinými slovy, SIP nemusí nutně odpovídat jedné logické entitě určené k ochraně. Také se může stát, že metadata poskytovaná tvůrcem jsou neúplná nebo nedostatečná a musí být doplněna během procesů vkládání dat.

Archivní informační balíček, AIP, je verzí informačního balíčku, kterou OAIS ukládá a uchovává.

Obsahem AIP je informace, která je předmětem uchovávání. Vedle toho je v něm obsažený úplný soubor metadat potřebných pro služby a procesy uchovávání a zpřístupňování systémem OAIS. Archivovaná informace a související metadata tvoří uvnitř archivačního systému logický balíček: není však vyžadováno, aby mezi metadaty samotnými a ochraňovaným objektem existovala fyzická vazba. Volba konkrétního způsobu uložení archivované informace a jejích metadat je na těch, kdo OAIS archiv implementují. Možná je jak plná fyzická integrace archivované informace a k ní se vztahujících metadat, tak uložení v oddělených, ale logicky propojených databázích.

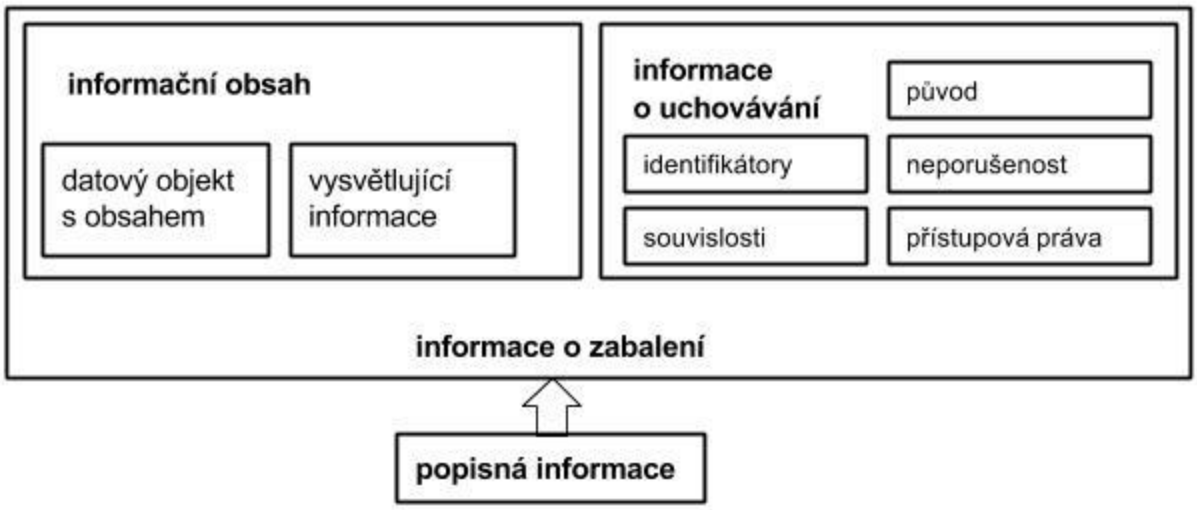
Referenční model definuje dvě „specializace“ AIP: archivní informační jednotku (AIU) a archivní informační sbírku (AIC). AIU ukládá obsah a metadata jednoho „atomického“ objektu (např. jednoho digitálního filmu nebo elektronické knihy), AIC se skládá z více AIU spojených do samostatné sbírky. Jednotka tvořící AIU je jen konceptuální, ve skutečnosti tento objekt také může existovat ve více fyzických nebo digitálních částech (například každá kapitola elektronické knihy může být uložena v samostatném souboru). V případě AIC má jak každá vložená AIU, tak i AIC samotná vlastní metadata. Jedna AIU může být součástí více AIC a kromě toho samotná AIC může být členem jiné širší AIC. AIC mohou spojovat dohromady AIU na základě vlastností, jako jsou předmět, téma, původ, nebo na základě jakéhokoli jiného kritéria, které vyhovuje archivu OAIS, jenž je spravuje. Referenční model uvádí, že AIC mohou zefektivnit proces vyhledávání (a tedy zpřístupnění určené skupině), pokud jsou AIU obsažené v archivu organizovány do smysluplné hierarchické struktury. AIC jsou nástrojem k organizaci obsahu na konceptuální úrovni (tj. metadata na úrovni AIC) sedícím nad jednotlivými AIU uvnitř archivu odpovídajícího OAIS (Lavoie, 2000, s. 15 a ČSN ISO 14721, 4.2.2.4).

Posledním třetím typem informačního balíčku je výstupní informační balíček, DIP, který koncový uživatel dostává jako odpověď na požadavek zpřístupnění. Informační balíček dodávaný archivem OAIS koncovému uživateli se velmi často liší od toho, co je uloženo v archivním skladu. Rozdíly mezi DIP a AIP mohou být například ve formátu obsahu (např. obrazový soubor může být konvertován do jiného formátu pro rychlejší zpřístupnění), v množství obsahu (DIP může odpovídat jednomu AIP, více AIP, nebo jen nějaké části AIP), nebo v metadatech dodaných spolu s obsahem (DIP často neobsahuje úplná metadata dostupná pro AIP, protože většina z nich nemá pro koncového uživatele žádný význam).

AIP je pro OAIS archiv klíčový – právě on, respektive informace, kterou obsahuje, je předmětem dlouhodobého uchovávání.

1.1.3.1 Složení Archivního informačního balíčku (AIP)

AIP musí obsahovat kompletní soubor metadat nutných k zajištění dlouhodobého uchování a zpřístupňování obsahu určené skupině. Referenční model popisuje jednotlivé typy metadat, která by měla být součástí archivované informace. Obrázek 3 znázorňuje informační komponenty AIP.



Obrázek 3 – Archivní informační balíček, dle (Lavoie, 2000, s. 16)

Popis jednotlivých částí níže je vytvořen volně dle (Lavoie, 2000, s. 15–18) s přihlédnutím k ČSN ISO 14721.

Datový objekt s obsahem (Content Data Object)

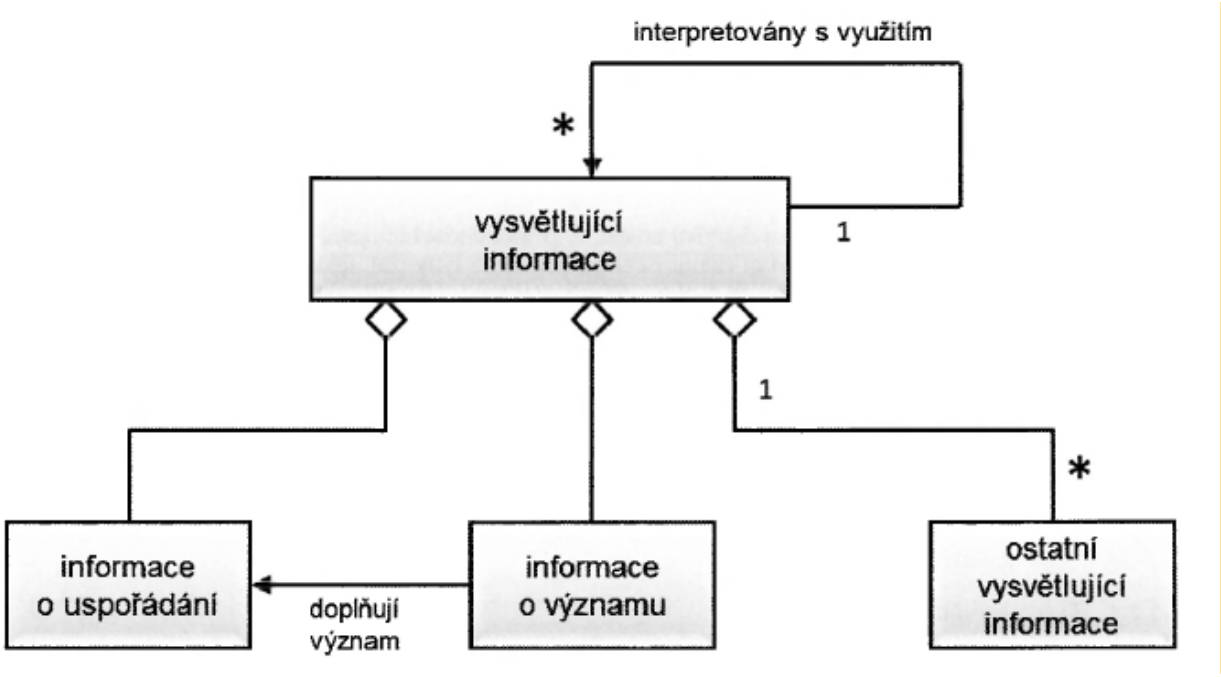
AIP obsahuje v první řadě datový objekt s obsahem, tj. informaci, která je vlastním předmětem uchovávání. Datový objekt s obsahem může mít podobu jakéhokoli materiálu: může to být text, obrázek, video, databáze, počítačový program, nebo dokonce fyzický materiál. Datový objekt s obsahem může být jediným soběstačným objektem, jako je například dokument ve formátu PDF, nebo může být složený. Tedy může obsahovat více datových objektů, jako například webová stránka skládající se z textu (HTML soubory) a statických obrázků (GIF, JPEG soubory) (ČSN ISO 14721, 4.2.1.3.2). Archiv OAIS je odpovědný za dlouhodobé uchování datového objektu s obsahem a za to, že ho uchová v nezávisle srozumitelné podobě pro určenou skupinu. K tomu napomáhá vysvětlující informace – viz níže.

Vysvětlující informace (Representation Information)

Aby mohl archiv OAIS dosáhnout uchování datového objektu s obsahem v takové podobě, že bude nezávisle srozumitelný určené skupině, musí datový objekt s obsahem doprovázet odpovídající vysvětlující informace. Ta je nezbytná pro zobrazení a porozumění bitům, které tvoří datový objekt s obsahem. Vysvětlující informace může obsahovat jakékoliv informace potřebné pro zobrazení (správné využití) datového objektu s obsahem nebo k zpřístupnění jeho obsahu. Může také shrnovat postup interpretace datového objektu s obsahem. Pokud je například datový objekt s obsahem ASCII soubor s čísly, vysvětlující informace může vysvětlovat, že čísla odpovídají průměrným denním teplotám vzduchu měřeným v Klementinu ve stupních Celsia v období 1972 až 2000.

Vysvětlující informaci lze rozdělit na dva typy: informace o uspořádání (Structure Information) a informace o významu (Semantic Information). Informace o uspořádání je v kontextu digitálních objektů snadno pochopitelná, odkazuje na mapování mezi digitálními bity a různými koncepty a datovými strukturami, které umožňují načíst bity do srozumitelné informace nebo formy – například obrazový soubor, text nebo interaktivní program. Informace o uspořádání popisuje formát digitálního objektu. Informace o významu na druhou stranu objasňuje význam nebo poskytuje odpovídající interpretaci

datového objektu s obsahem. Příkladem informací o významu, které mohou být spojeny s datovým objektem s obsahem, jako součást jeho vysvětlující informace, jsou glosář, datový slovník nebo dokumentace k softwarové aplikaci. Referenční model také definuje zbytkovou kategorii ostatní vysvětlující informace (Other Representation Information), která zahrnuje vysvětlující informace, jež nejsou snadno definovatelné jako informace o uspořádání nebo informace o významu. Referenční model uvádí, že do této kategorie by spadala například informace, která spojuje informace o významu a informace o uspořádání (ČSN ISO 14721, 4.2.1.3.1).



Obrázek 4 – Objekt s Vysvětlujícími informacemi (ČSN ISO 14721, 4.2.1.3.1)

Struktura vysvětlující informace bývá ve skutečnosti často velmi komplexní. Konkrétní soubor vysvětlujících informací může vyžadovat další vysvětlující informace k tomu, aby je mohla určená skupina správně zobrazit, interpretovat nebo pochopit. I tento druhý soubor vysvětlujících informací může vyžadovat další vysvětlující informace. Taková regrese by mohla pokračovat v libovolném počtu kroků. Jako příklad lze použit digitální objekt ve formátu METS (Metadata Encoding and Transmission Standard). K zajištění srozumitelnosti dokumentu METS může OAIS archiv potřebovat jako součást vysvětlujících informací kopii XSD[10] schématu METS. Nicméně, schéma METS je vyjádřeno v XML (Extensible Markup Language). Proto k pochopení METS schématu (a tedy nepřímo také k pochopení původního METS dokumentu) mohou uživatelé potřebovat specifikaci XML. Samotné XML je profilem SGML (Standard Generalized Markup Language) ISO Standard 8879:1986, tedy k plnému pochopení XML je potřeba popis standardu SGML jako součást vysvětlujících informací. METS schéma, XML specifikace, SGML standard – to vše tvoří síť vysvětlujících informací (Representation Network) spojenou s datovým objektem s obsahem (METS dokumentem). V praxi pochopitelně OAIS archiv sítě vysvětlujících informací v nějakém bodě zastaví. A to v bodě, ve kterém lze předpokládat rozumnou míru znalosti u určené skupiny. Referenční model OAIS popisuje tuto předpokládanou znalost určené skupiny jako znalostní základnu (Knowledge Base). Tento příklad zřetelně ilustruje, že vymezení určené skupiny

[10] XML schema – jazyk popisující XML.

má dopad na metadata, která je potřeba vytvářet a uchovávat v rámci AIP. Nelze zapomínat na fakt, že určená skupina se časem mění, rozvíjí, rozšiřuje. Podobně musí být revidována metadata pro tuto skupinu určená.

Datový objekt s obsahem a s ním spojené vysvětlující informace (nebo jejich síť) se společně označují jako informační obsah (Content Information). OAIS archiv tedy musí dlouhodobě uchovávat právě informační obsah, tj. informace, které jsou předmětem uchovávání spolu s dostatečnými metadaty k zajištění zobrazitelnosti a srozumitelnosti určenou skupinou. Pro dlouhodobé uchovávání jsou ale potřeba další informace v podobě informací o uchovávání.

Informace o uchovávání (Preservation Description Information)

Informace o uchovávání, PDI, se podle referenčního modelu „zaměřují na popis dřívějších a stávajících stavů informačního obsahu. Přičemž zajišťují, že informační obsah lze jednoznačně identifikovat a že nebyl nevědomky změněn“ (ČSN ISO 14721, 4.2.1.4.2).

PDI se skládá z pěti komponent[11] (ČSN ISO 14721, 4.2.1.4.2):

- *Informace o identifikátorech (Reference Information)* jednoznačně identifikuje informační obsah v interních systémech OAIS, nebo vzhledem k entitám vně OAIS. Příkladem může být systémem generovaný jedinečný identifikátor či ISBN nebo URN:NBN.
- *Informace o souvislostech (Context Information)* popisuje vazby informačního obsahu na jiný informační obsah, například na tematicky související informační obsah (předmětem vymezené sbírky), nebo na ten, který představuje další verze stejného obsahu v alternativních formátech.
- *Informace o původu (Provenance Information)* dokumentuje historii informačního obsahu, včetně popisu jeho vzniku, změn v obsahu nebo formátu během času, změn v předávání správy objektu. Dále jakékoli kroky, které byly učiněny s cílem uchovat informační obsah (jako jsou normalizace nebo formátová migrace) a výsledek těchto kroků.
- *Informace o neporušenosti (Fixity Information)* zajišťuje, aby informační obsah nebyl změněn nezdokumentovaným způsobem. K zajištění neporušenosti se používají mechanismy validace autenticity nebo integrity – jako jsou kontrolní součty, digitální podpisy nebo digitální vodoznaky. Mechanismy samotné nejsou informací o neporušenosti, tou jsou hodnoty jejich výstupů (tedy metadata).
- *Informace o přístupových právech (Access Rights Information)* dokumentuje podmínky a omezení spojená s informačním obsahem, týkající se jak uchovávání, tak zpřístupňování. Může též zahrnovat popis mechanismů zajišťujících dodržení těchto podmínek.

Informace o zabalení (Packaging Information) balí do jednoho celku informační obsah (datový objekt s obsahem a vysvětlující informace) a informace o uchovávání (informace o identifikátorech, informace o souvislostech, informace o původu, informace o neporušenosti, informace o přístupových právech). Informace o zabalení tak (logicky) svazuje všechny informační komponenty AIP a zajišťuje, aby se daly vyhledat jako objekty v archivačním systému. Informace o zabalení může být velmi stručná, například jen název souboru a jeho lokace, nebo podrobnější – jako například METS.

Popisné informace spolu s vyhledávacími pomůckami podporují vyhledávání a dodávání informačního obsahu koncovému uživateli archivu OAIS.

[11] S výjimkou informací o přístupových právech jsou komponenty PDI založeny na diskuzi v klíčové zprávě Waters and Garrett 1996; s..11–19.

Informační model archivu OAIS

Výše popsané informační komponenty, informační obsah (datový objekt s obsahem a vysvětlující informace), informace o uchovávání (informace o identifikátorech, informace o souvislostech, informace o původu, informace o neporušenosti, informace o přístupových právech), informace o zabalení a popisná informace tvoří společně informační model archivu OAIS. Tedy, informační obsah a informace o uchovávání tvoří archivní informační balíček, informace o zabalení pak AIP identifikuje a lokalizuje jako jednu logickou jednotku. Popisná informace podporuje vyhledávání a dodávání AIP. OAIS referenční model nevyžaduje žádný konkrétní způsob implementace jednotlivých komponent informačního modelu.

1.1.3.2 Životní cyklus balíčku

Důležitou podstatou odborné práce správců úložiště je porozumění životnímu cyklu uložených dokumentů. Jen tak jsou schopni mu poskytnout plnohodnotnou ochranu na logické úrovni, tedy aby byla zachována nejen data, ale i jimi nesená informace ve srozumitelné podobě. U uchovávaných dat by proto mělo mít úložiště sestavený model životního cyklu pro jednotlivé typy dokumentů. Životní cyklus dlouhodobé ochrany by měl být zaveden již na počátku sestavování plánu dlouhodobé ochrany.

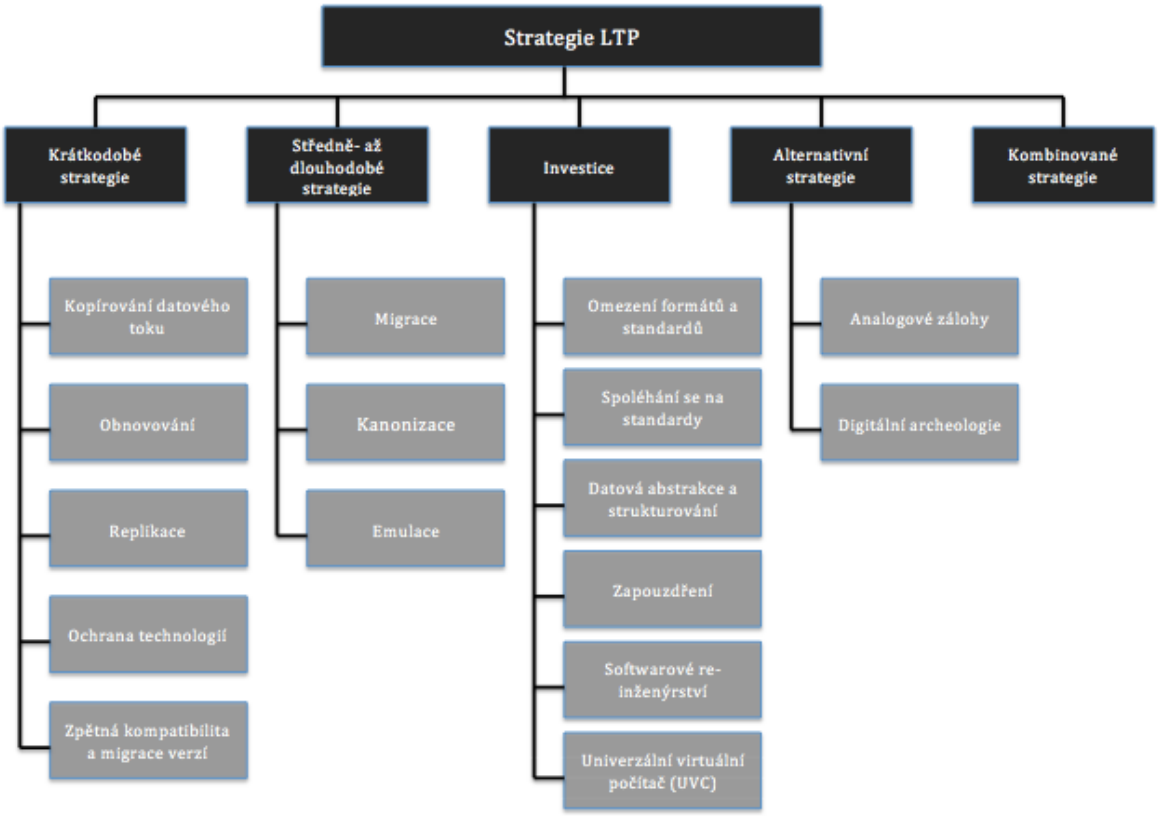
V konkrétním repozitáři porozumění životnímu cyklu znamená znalost obsahu a způsobů vytvoření SIP. To je obvykle první krok v rámci životního cyklu. Následuje vytvoření AIP. V této fázi životního cyklu jsou nad balíčkem prováděny ochranné aktivity, je aktualizován a doplňován. Závěr životního cyklu tvoří export mimo úložiště či případná tranformace. Pracovníci úložiště tedy musí znát i ty části životního cyklu, které proběhly před vstupem do úložiště, aby byli schopni provádět potřebné ochranné aktivity.

Mezi nejdůležitější součásti životního cyklu balíčku obecně i v rámci v systému ARCLib patří: tvorba či příjem (Create Or Receive), import (Ingest), ochranné aktivity (Preservation Action), uložení (Store) a případná transformace (Transformation).

1.1.4 Strategie plánování dlouhodobé ochrany digitálních dat

Dlouhodobá ochrana digitálních dokumentů představuje komplexní přístup k řízení, rozhodování a provádění příslušných opatření. Rozhodnutí pro jednotlivá opatření jsou strategické nebo manažerské povahy a týkají se jak organizačních hledisek, tak odborných a technických problémů, jako např. rozhodování o konkrétních krocích spojených s ochranou daného obrazového souboru (LTP Portál.cz, cca 2015).

Obecně jsou rozlišovány typy strategie dlouhodobé ochrany, které popisuje Obrázek 5 (Mohanty, 2014; Guidelines, 2003; Digital, 2012) a které jsou blíže popsány v následujících kapitolách.



Obrázek 5 – Strategie dlouhodobé ochrany, podle (Mohanty, 2014)

1.1.4.1 Krátkodobé strategie (nejlépe fungující v krátkém časovém období)

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

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

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. Kupříkladu kopírování datového toku (*bistream*) představuje určitou formu replikace. OAIS standard považuje replikaci za formu migrace. Zřejmě nejznámější realizací této strategie dlouhodobé ochrany je konsorciální projekt LOCKSS.[12] Klíčovou myšlenkou je vytvoření redundantních

[12] Jedná se o open-source, knihovní systém dlouhodobé ochrany vytvořený na Stanfordské univerzitě, který je postavený na základním principu "Lots Of Copies Keep Stuff Safe" (LOCKSS), tedy „mnoho kopií udrží věci

dat, každý digitální dokument je uložen v několika separátních repozitářích.

Na podobném principu funguje i uzavřený systém CLOKSS,[13] který uchovává elektronický obsah vydavatelů v *dark archive.*[14] Aby CLOCKSS prokázal svou důvěryhodnost v uchovávání dat, byl Centrem pro vědecké knihovny (The Centre for Research Libraries – CRL) proveden audit podle kritérií ISO 16363.[15]

Ochrana technologií (*technology preservation*) – strategie založená na udržování a spravování technického prostředí, které slouží pro vytváření obsahu včetně operačního systému, originální softwarové aplikace, médií. Může se zdát, že zachování původní technologie je nejúčinnější a nejjednodušší způsob, jak zachovat funkcionalitu, vzhled a dojem z digitálního prostředí; rozhodně se ovšem nejedná o životaschopnou strategii pro dlouhodobé uchovávání digitálních objektů. Ochrana technologií by měla být spíše strategií obnovy po havárii (*disaster recovery*) pro digitální objekty, které nepodléhají žádné jiné vhodné strategii ochrany.

Zpětná kompatibilita a migrace verzí (*backwards compatibility and version migration*) – spoléhá na schopnost nejnovějších verzí softwaru interpretovat a zobrazit digitální materiál vytvořený v předchozích verzích stejného softwaru. V případě zpětné kompatibility může být zobrazení formátu časově omezené, migrace do nové verze permanentně zkonvertuje dokumenty do formátu, který je možné zobrazit v současné verzi softwaru, ovšem s rizikem možné ztráty informací.

1.1.4.2 Střednědobé až dlouhodobé strategie

Migrace (*migration*) – představuje primární strategii dlouhodobé ochrany dat 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.[16] Jedná se tedy o proces transformace digitálních formátů, včetně ochrany datového toku (*bit stream*) a schopnosti zobrazení obsahu reprezentovaného daným datovým tokem. Digitální dokument je neoddělitelně provázaný se svým prostředím (software, hardware, formát) a změnou formátu může dojít ke ztrátě samotných informací (obsahu). Migrace formátů proto musí obsahovat kontrolu a posouzení změn obsahu předtím, než je provedena.

Emulace (*emulation*) – proces, kterým počítačový program nebo zařízení umožňuje jednomu systému imitovat jiný. V oblasti digitální archivace jde o ochrannou aktivitu, která využívá speciální software (emulátor), který překládá instrukce z originálně uchovaného programu do novější platformy, čímž obchází nutnost uchovávat zastaralý hardware nebo systémový software.

1.1.4.3 Strategie ad-hoc výzkumu a kontroly pomocí standardů a omezení

Omezení formátů a standardů (*restricting formats and standards*) – digitální archivy se mohou

---

v bezpečí".

[13] *Controlled LOCKSS.*

[14] Podrobněji o "dark archive" např. v učebním materiálu Úvod do ochrany digitálních dat od Jana Hutaře (2008, s. 3).

[15] Zpráva je dostupná na Center for Research Libraries (Center for Research Libraries, 2014). Výsledkem celého procesu je veřejná dokumentace podle jednotlivých kritérií ISO standardu 16363 (resp. TRAC) na wiki webu (CLOCKSS Archive: documentation Wiki, 2014).

[16] 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).

rozhodnout uchovávat a dlouhodobě ochraňovat pouze omezenou škálu formátů. Z důvodů možné ztráty dat či jejich vlastností je žádoucí stanovit spíše seznam formátů pro vznikající data, než podporovat konverzi do stanovených „archivních" formátů (často označováno termínem normalizace).

Důraz na standardy (*reliance on standards*) – strategie zahrnuje používání otevřených a široce užívaných a podporovaných standardů a formátů. Tyto mají větší šanci na přetrvání a vyhnutí se např. problémům vyplývajících z vývoje počítačového prostředí (operační systémy, softwarové aplikace). Spoléhání se na standardy může snížit bezprostřední ohrožení digitálních dokumentů před zastaráváním, není však trvalým a jediným řešením dlouhodobé ochrany.

Datová abstrakce a strukturování (*data abstraction and structuring*) – někdy označována i jako normalizace. V rámci archivního úložiště, většinou na vstupu do úložiště, jsou všechny digitální objekty určitého typu (např. barevné obrázky) konvertovány do jednoho zvoleného formátu, který má představovat nejlepší možný kompromis mezi vlastnostmi, jakými jsou funkčnost, dlouhá životnost a vysoká odolnost. Podobně jako u spoléhání se na standardy má i normalizace své výhody a nevýhody.

Zapouzdření (*encapsulation*) – představuje klíčovou techniku pro emulaci, případně jiné ochranné aktivity. Původní objekt je zachován formou datového toku, který je zapouzdřen spolu s instrukcemi, případně jinými informacemi, které jsou nezbytné pro zpřístupnění do budoucna (např. prohlížeče, softwarová specifikace, ochranná metadata apod.). Informační balíčky definované OAIS je možné považovat za formu zapouzdření, kdy je digitální objekt zabalený společně s vysvětlující informací, potřebnou k interpretaci bitů a informací o uchovávání (PDI), která zahrnuje informace o obsahu, původu, identifikátorech, neporušenosti, souvislostech a přístupových právech.

Softwarové re-inženýrství (*software re-engineering*) – digitální materiály jsou navázány na aplikační software a jejich funkčnost je tak závislá na specifickém systému nebo platformě. Dlouhodobá ochrana aplikačního softwaru je málokdy nativně podporována, proto se nabízí princip migrace (obdobný u formátů), například úpravou a rekompilací zdrojového kódu na novou platformu, přeložení (*translation*) kompilovaných binárních instrukcí jedné platformy do jiné platformy apod.

UVC (*Universal Virtual Computer*) – představuje jednu z forem emulace. Myšlenka UVC, tedy jakéhosi univerzálního virtuálního počítače, se zrodila počátkem roku 2000 a byla použita v projektu dlouhodobé ochrany JPEG v Národní knihovně Nizozemska. UVC byl postaven na principu dekodérů formátů souborů a programů. Uživatel měl mít možnost vytvořit a uložit digitální soubory pomocí SW aplikace dle svého výběru, ale zároveň se všechny soubory zálohovaly tak, aby je bylo možné číst pomocí UVC.

1.1.4.4 Alternativní přístupy k dlouhodobé ochraně digitálních dat

Analogové zálohy (*analogue backups*) – tedy zálohování dokumentů formou převodu do analogové podoby (např. tištěním textových dokumentů na papír nebo mikrofilmování[17]), čímž se ztrácí výhody digitálních dokumentů[18] (např. sdílení a bezztrátový přenos). Za nejvhodnější typy dokumentů jsou považovány text a černobílé obrazové dokumenty, které nejméně „trpí" při naplňování této strategie.

[17] 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 16mm nebo 35mm mikrofilm.

[18] Vojtášek (2000) nabízí přehled charakteristických vlastností digitálního a analogového dokumentu.

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.

Digitální archeologie (*digital archeology*) – metoda získávání dat[19] ze zastaralých softwarových a hardwarových prostředí a zastaralých či poškozených médií (např. děrné štítky, 8" diskety). Tato metoda je časově, finančně a technicky náročná a nemá stoprocentní úspěšnost (např. v případě chybějících metadat, značně poškozených médií apod.).

1.1.5 Prokazování kvality v oblasti dlouhodobé digitální archivace

1.1.5.1 Manuál pro přípravu auditu podle Data Seal of Approval

Jednou z často využívaných možností na provedení auditu je metodika Data Seal of Approval (DSA), která představuje první stupeň Evropského rámce pro audit a certifikaci digitálních repozitářů. Nabízí relativně jednoduchou formu auditu procesů a postupů repozitářových řešení. Hlavními „nedostatky", které byly předchozím verzím této certifikace vytýkány, byly hodnocení na základě důvěry a obsahově se překrývající směrnice. Druhá výtka je však díky spojení DSA a ICSU World Data System (WDS) komunit již minulostí. Pod záštitou RDA/WDS (RDA/WDS Interest Group on Certification of Digital Repositories) zájmového sdružení k certifikaci digitálních repozitářů vznikla pracovní skupina, jejímž cílem bylo během 18 měsíců definovat skutečné klíčové charakteristiky, zefektivnit hodnocení a zvýšit dopad na cílovou komunitu. DSA-WDS audit není tak detailní jako např. audit podle ISO 16363, a je proto brán jako první stupeň na cestě ke kompletnímu externímu auditu. DSA-WDS zásady představují jakési minimum vyextrahované z různých relevantních zdrojů (Dillo, 2015) a poskytují poměrně jednoduchý způsob auditu a certifikace, zejména pro začínající repozitáře nebo repozitáře menšího rozsahu. I pro rozsáhlejší repozitáře s národním významem nabízí certifikace podle zásad DSA-WDS výhody, zejména konsolidaci vlastní pozice před náročným externím hodnocením. Výhodou této certifikace je i poměrně nízká finanční náročnost, protože zaměstnanci instituce by měli být schopni požadované podklady připravit sami.

Zásady DSA-WDS jsou rozděleny do 16 okruhů a mají 2 doplňková kritéria (DSA-WDS, 2016):

I. Organizační infrastruktura

1. Poslání/Rozsah (Mission/Scope)
2. Licence (Licences)
3. Kontinuita přístupu (Continuity of access)
4. Důvěrnost/Etika (Confidentiality/Ethics)
5. Organizační infrastruktura (Organizational infrastructure)
6. Odborná pomoc (Expert guidance)

II. Management digitálních objektů (Digital Object Management)

[19] 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).

7. Integrita a autenticita dat (Data integrity and authenticity)

8. Posouzení (Appraisal)

9. Dokumentace postupů uchovávání (Documented storage procedures)

10. Plán dlouhodobé ochrany (Preservation plan)

11. Kvalita dat (Data quality)

12. Pracovní postupy (Workflows)

13. Plán dlouhodobé ochrany (Data discovery and identification)

14. Opětovné využití dat (Data reuse)

III. Technologie

15. Technická infrastruktura (Technical infrastructure)

16. Bezpečnost (Security)

Komentáře (Additional information) a Zpětná vazba uchazeče (Applicant feedback) certifikace DSA nepatří mezi hodnotící kritéria repozitářů v pravém smyslu slova. Nabízí možnost dopsání doplňujících informací relevantních pro hodnocení, přičemž není možné nebo vhodné je připsat k jednotlivým zásadám, a zpětnou vazbu celkového průběhu certifikace DSA z pohledu uchazeče.

1.1.5.2 Principy DSA-WDS

Pro Katalog pravidel DSA-WDS je klíčových 5 základních principů, které definují vhodně archivovaná data:

- je možné je najít na internetu,
- jsou zpřístupňována v souladu s platnou a relevantní legislativou a je respektována ochrana duševního vlastnictví,
- jsou dostupná v použitelném formátu pro určenou komunitu,
- jsou spolehlivá,
- je možné na ně odkazovat (persistentní identifikátory).

Pravidla jsou zacílená podle participujících stran:

- producenti dat, kteří zodpovídají za samotnou kvalitu digitálních dat,
- repozitář/archiv, který zodpovídá za kvalitu uchovávaných a (případně) zpřístupňovaných dat,
- uživatelé, kteří zodpovídají za správné využívaní dat v souladu s platnými regulemi.

Průběh auditu a certifikace podle DSA-WDS je možné rozdělit na tři hlavní části:

- sebehodnocení pomocí online webového nástroje,
- prozkoumání odborníky – hodnocení na základě důvěry,
- udělení pečeti DSA-WDS.

Podrobný manuál pro získání pečeti DSA-WDS je uveden v Příloze A, kde jsou rozepsána jednotlivá kritéria hodnocení auditu DSA-WDS, včetně potřebných požadavků, materiálů a dokumentů.

1.2 Praktická část vázaná na systém ARCLib

1.2.1 Návrh praktické implementace systému ARCLib podle modelu OAIS[20]

ARCLib je systém pro logickou a bitovou ochranu digitálních dat navržený v souladu s požadavky odvozenými z ČSN ISO 14721 (dále OAIS). Toto řešení institucím umožňuje implementovat všechny funkční moduly OAIS a jeho informační model. ARCLib je navržen jako tzv. "dark archive", tj. není primárně určen ke zpřístupňování dokumentů koncovým uživatelům. Nedisponuje prostředky pro zobrazení archivovaných dat (image servery, prohlížeče apod.). Uživateli ARCLibu jsou převážně správci archivních digitálních dat.

ARCLib se skládá z několika klíčových funkčních modulů: ARCLib Ingest, ARCLib Data Management a ARCLib Archival Storage; ty jsou doplněny moduly ARCLib Administration, ARCLib Access a ARCLib Preservation Planning, jehož funkce jsou ovšem realizovány převážně mimo systém ARCLib.

ARCLib je schopen pracovat s výstupy nástrojů jako např. ProArc[21] a Archivematica[22] a přijmout SIP jimi vytvořené. SIP je pak dále ukládán ve struktuře BagIt (Kunze et al, 2016) a spolu s ním je udržován XML soubor s metadaty – ARCLib AIP XML. AIP je ukládán jako logický objekt, tj. soubor SIP.BagIt a na jiném místě (jiná storage) soubor AIP XML nebo více XML.

ARCLib umožňuje verzování AIP a jeho mazání. AIP je možné vymazat jak logicky (jen označením za vymazané), tak fyzicky (skutečně odstraněním dat, přičemž jsou uchována metadata balíčku a záznamy v databázi). V obou případech je třeba uchovávat auditní záznam.

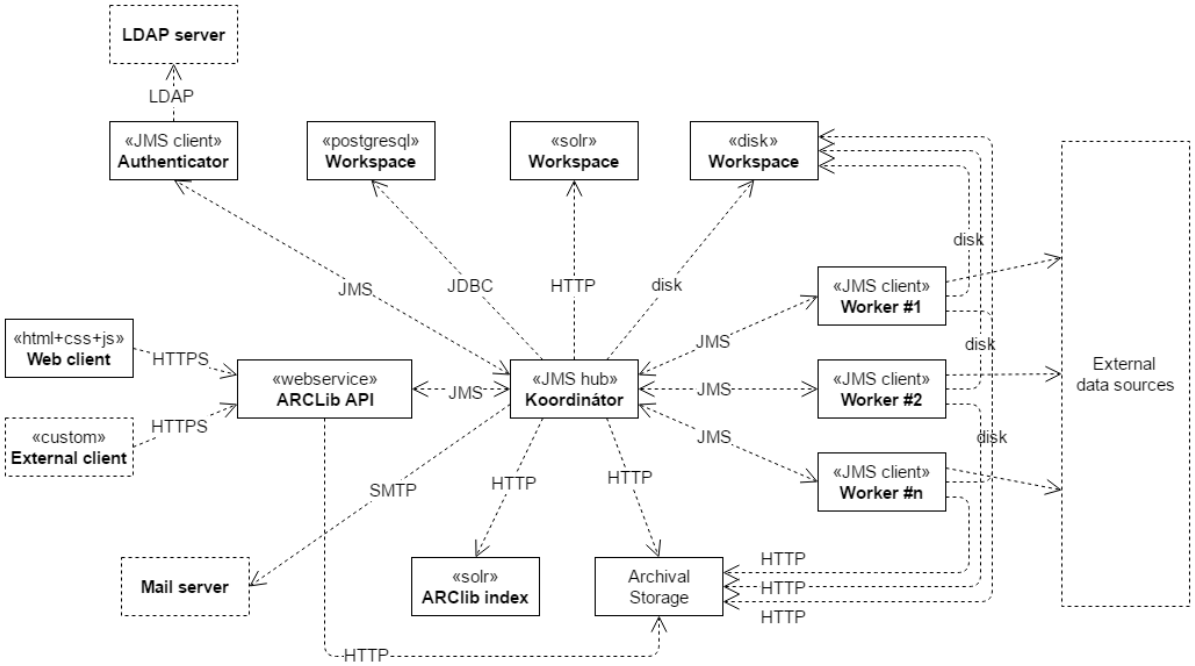
1.2.2 Architektura systému ARCLib

Systém ARCLib respektuje principy SOA (Service-Oriented Architecture), které jsou zabezpečeny s použitím implementace JMS (Java Message Service), jež řeší tzv. problém producenta a spotřebitele. Producent vyrábí určitá data a dává je do fronty o omezené velikosti. Spotřebitel je z fronty odebírá. JMS je komunikační standard, který umožňuje Java aplikacím vytvářet, odesílat, přijímat a zpracovávat zprávy na úrovni komponent, které tak mohou být velice volně provázané. Operace tak probíhají asynchronně, což je vhodné především u distribuovaných systémů. Komunikaci řídí tzv. JMS provider, v tomto případě komponenta Koordinátor. Další komponenty systému jsou přihlášeny k odběru pro ně určených JMS požadavků z fronty a také tyto požadavky produkují.

Komunikace pomocí JMS zpráv je kombinována se standardem JDBC pro přístup k databázi a HTML voláními API rozhraní. Dále je použit protokol SMTP pro komunikaci s mailovým serverem. Transformaci mezi použitými komunikačními formáty zabezpečuje Koordinátor.

Samostatnou vrstvu systému tvoří tenký webový klient, který však nemusí být nutně součástí funkčního systému; může být nahrazen libovolným jiným klientem zabezpečujícím komunikaci s API rozhraním systému ARCLib. Modularita a bezstavovost většiny komponent systému ARCLib otevírá široké možnosti škálování a paralelizace.

[20] Kapitoly 1.2.1, 1.2.2 a 1.2.3 vychází z implementační analýzy zpracované společností InQool, a.s.
[21] https://github.com/proarc/proarc/wiki.
[22] archivematica.org/en/.



Obrázek 6 – Architektura systému ARCLib

1.2.2.1 Popis komponent systému ARCLib

Koordinátor

Koordinátor je JMS provider, tedy komponenta sloužící v rámci standardu JMS ke sběru a správě JMS zpráv ve frontě zpráv. Koordinátor je bezstavový a neudržuje si žádná data o zpracovávaných požadavcích. V případě potřeby je tedy možné tento komunikační uzel duplikovat, například pro zajištění vysoké dostupnosti. Vzhledem k povaze a malé velikosti požadavků se však nepočítá s paralelním provozem více Koordinátorů.

JMS požadavky jsou krátké koordinační zprávy s konkrétním cílovým modulem, které Koordinátor v případě potřeby transformuje do formátu kompatibilního s webovými službami či protokolem cílového modulu nebo komponenty. Objemná data – obsah SIP/AIP– si komunikující moduly vyměňují bez účasti samotného Koordinátora.

Workspace

Workspace je souhrnné označení pro veškerá úložiště pracovních dat systému ARCLib – diskové úložiště, relační databázi a index.

Diskové úložiště je použito pro uložení přijímaného SIP po dobu jeho validace, vytěžení a vygenerování průvodního metadatového souboru ARCLib XML. Jedná se buď o lokální disk, anebo disk namapovaný na virtuální server jako složka. Cesta k úložišti je konfigurovatelná na úrovni systému ARCLib.

Relační databáze obsahuje systémová data, jako jsou data o poskytovatelích, uživatelích a oprávněních, systémová nastavení, definice validačních, importních, exportních a dalších profilů atd. Abstrakci databázových služeb zabezpečuje JDBC API, které umožňuje jednoduché použití téměř jakékoli běžně používané databázové technologie.

Index je využit pro optimalizaci komunikace s relační databází. Abstrakce je zde zabezpečena implementací wrapperu, tedy API webových služeb komunikujících navenek standardizovaným způsobem s ostatními moduly systému ARCLib a poskytujících dovnitř směrem k indexovací technologii sadu metod k implementaci konkrétní použitou technologii.

Worker

Worker je označení pro instanci/vlákno vykonávající operace spojené s data ingestem na základě definovaných jobů/rutin. Worker si vybírá úkol k zpracování z fronty úkolů a spouští workflow. Veškeré informace o workflow, které se má spustit, a jeho konfiguraci jsou součástí JMS zprávy. Rychlost zpracovávání úkolů ve frontě je možné regulovat přidáváním a odebíráním Workerů z množiny všech aktuálně dostupných Workerů, tzv. Worker pool.

Externí zdroje dat jsou sdílená datová úložiště, která musí být dostupná všem Workerům. Může se jednat o filesystém, FTP server a jiné.

Každý Worker v sobě obsahuje vlastní workflow engine, který mu umožňuje vykonávat workflow procesy na základě BPMN 2.0 definicí. Workflow engine ukládá svoje transakční a provozní data do databáze systému.

Authenticator

Autentizace je řešena vnitřním LDAP serverem pro autentizaci postaveným na technologii OpenLDAP. Obsluhu tohoto LDAP serveru zabezpečuje komponenta Authenticator, která umožňuje také použít či paralelně připojit k ARCLib i externí LDAP server pro autentizaci. Komunikace s externím LDAP serverem probíhá protokolem LDAP/LDAPS.

ARCLib API

Toto RESTful rozhraní webových služeb zabezpečuje stěžejní logiku systému ARCLib:

- administrace veškerých entit systému
    - původců dat,
    - uživatelů, uživatelských rolí a oprávnění,
    - SIP, validačních a exportních profilů;
- vyhledávání v indexovaných metadatech
- správu vyhledávacích dotazů
- rozesílaní notifikací
- spouštění naplánovaných exportů
- informace o stavu a konfiguraci indexů
- informace o verzích použitých externích nástrojů (např. FIDO, Siegfried, DROID aj.), případně možnost nastavení předem definovaných parametrů nástrojů

ARCLib Index

ARCLib Index slouží pro indexaci metadat vytěžených z ingestovaných SIP pro potřeby efektivního a komplexního vyhledávání v archivovaných datech. Abstrakce je zde zabezpečena implementací wrapperu, tedy API webových služeb komunikujících navenek standardizovaným způsobem s ostatními moduly systému ARCLib a poskytujících dovnitř směrem k indexovací technologii sadu metod k implementaci konkrétní použitou technologii.

Webclient

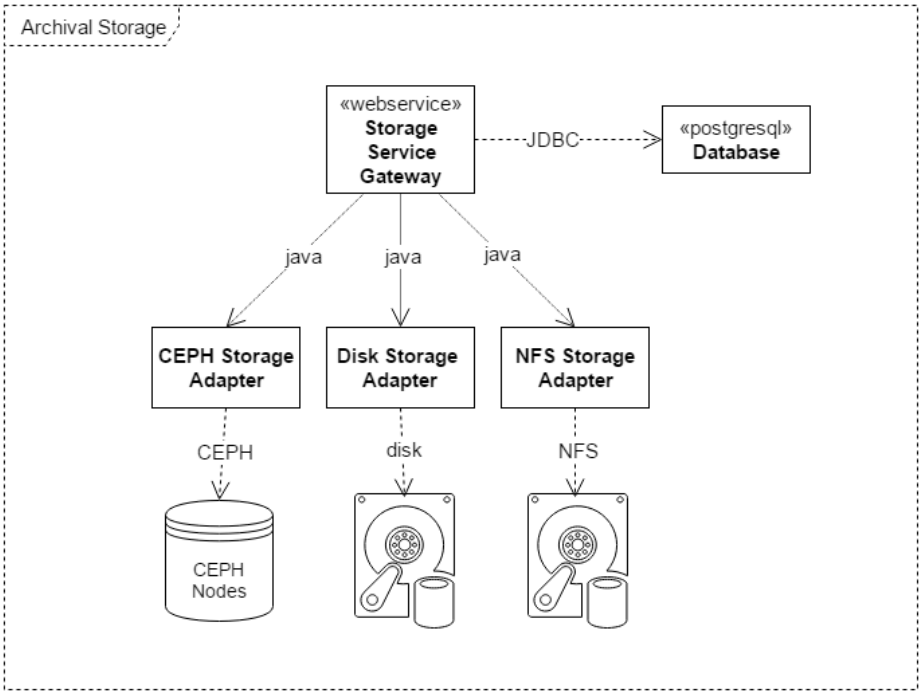
Webclient je tenký webový klient, poskytující především administrační rozhraní a rozhraní pro vyhledávání v indexovaných datech.

1.2.2.2 Popis komponent modulu ARCLib Archival Storage

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

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

V kontextu projektu ARCLib je podporováno několik technologií fyzických úložišť, kromě běžného filesystému (zejména pro testovací účely) především ZFS a Ceph.



Obrázek 7 – Architektura modulu ARCLib Archival Storage

Storage Service Gateway

Storage Service Gateway představuje obecné abstraktní rozhraní komponenty Archival Storage pro ukládání balíčků do datového úložiště a pro komunikaci s ním. Pro jednotlivé typy datových úložišť pak existuje Java třída implementující jednotlivé metody tohoto rozhraní, především pro:

- uložení AIP na úložiště,
- načítání AIP z úložiště,
- zjištění stavu AIP na úložišti – kontrola existence a fixity,
- zjištění stavu úložiště a všech jeho uzlů,
- zutomatické spuštění rutiny 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 Archival Storage uloženy v databázi, nebyly úspěšně uloženy a je potřeba je ve všech úložištích smazat.

Storage Service Gateway obsahuje REST API rozhraní, které poskytne webové služby pro:

- uložení AIP na úložiště,
- načítání AIP z úložiště,
- zjištění stavu AIP na úložišti – kontrola existence a fixity,
- 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.

Adaptéry

Pro každý typ datového úložiště je implementována Java třída, tzv. Adapter, implementující jednotlivé metody rozhraní iArchivalStorageAdapter definované v rámci Storage Service Gateway, především 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.

Database

V této databázi jsou uloženy především transakční informace o aktuálně ukládaných AIP.

1.2.2.3 Integrace ARCLib s externími systémy

Integrace s externími systémy je obecně zabezpečena vystavením ARCLib API. K tomuto API je možné připojit libovolnou klientskou aplikaci. Autentizace klientské aplikace není nutná, autentizuje/autorizuje se vždy uživatelským jménem a heslem uživatele uloženého v Authenticatoru.

Zabezpečení komunikace – TLS certifikáty

ARCLib API poskytuje připojení protokolem HTTPS zabezpečené TLS certifikátem. Pokud bude ARCLib API sloužit pouze pro interní klienty, je dostačující certifikát vygenerovaný vlastní certifikační autoritou. Při vystavení API pro externí klienty (případně i na vlastní doméně) bude třeba zajistit TLS certifikát vydaný globálně uznávanou certifikační autoritou, např. GeoTrust. Stejná pravidla jako pro ARCLib API platí i pro ARCLib Archival Storage a jeho Storage Service Gateway.

Webclient s ARCLib API komunikuje také přes HTTPS protokol. Pokud by měl být Webclient dostupný z internetu (případně i na vlastní doméně), bude i pro něj třeba zajistit TLS certifikát vydaný globálně uznávanou certifikační autoritou, např. GeoTrust.

Postup pro instalaci a konfiguraci TLS certifikátů je popsán v technické dokumentaci systému ARCLib.

1.2.2.4 Požadavky na software

Ačkoli ARCLib je platformově nezávislé řešení, které je schopno provozu na libovolném OS, preferovaným OS je Debian GNU/Linux ve stabilní verzi.

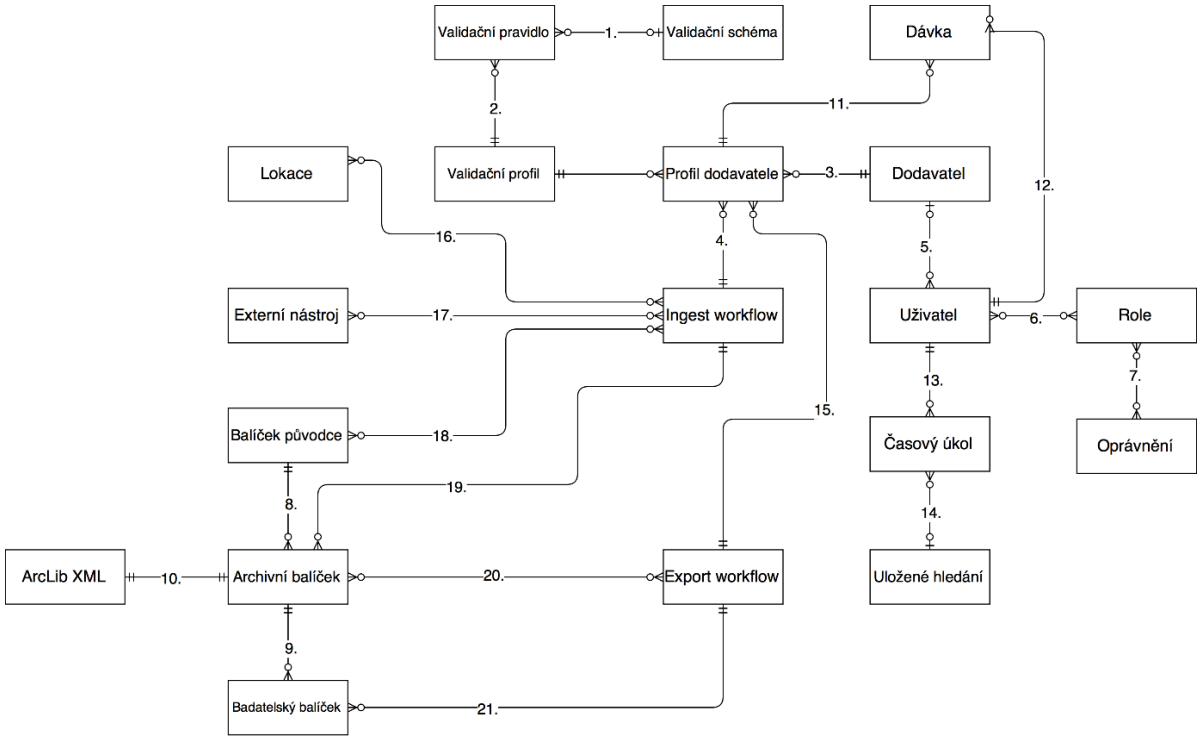
Veškerý další potřebný software je nainstalován prostřednictvím root účtu.

Pro ARCLib, tak pro ARCLib Archival Storage je možno využít databázový server postavený na technologii PostgreSQL. PostgreSQL může být případně nainstalováno i přímo na virtuální servery.

1.2.3 Konceptuální model systému ARCLib

Základem pro efektivní návrh informačního systému je pochopení domény, ve které má nový informační systém fungovat. Jedná se o určení základních entit, jejich vztahů a atributů. Takový model se nazývá konceptuální. V tomto modelu není nutno zachytit každou entitu, protože by model byl příliš složitý (zejména kvůli vztahům). Důležité je zachytit ty podstatné entity, které doménu odlišují od jiných domén, a zakotvení definic entit tak, aby byly jednoznačně srozumitelné a pochopitelné.

Konceptuální model je pak v průběhu implementace podkladem pro tvorbu datového modelu systému.



Obrázek 8 – Konceptuální model systému

Konceptuální model systému ARCLib obsahuje tyto koncepty:

- Location (Lokace)
- Tool (Externí nástroj)
- Validation Scheme (Validační schéma)
- Validation Rule (Validační pravidlo)
- Validation Profile (Validační profil)
- Supplier (Dodavatel)
- Supplier Profile (Profil dodavatele)
- Ingest workflow
- Export workflow
- User (Uživatel)
- Permission (Oprávnění)
- Role (Role)

- SIP (Vstupní informační balíček)
- AIP (Archivní informační balíček)
- DIP (Výstupní informační balíček)
- ARCLib XML
- Batch (Dávka)
- Time Job (Časový úkol)
- Search Query (Uložené hledání)

Konceptuální model dále zachycuje tyto vazby mezi koncepty:

- Validační pravidlo využívá Validační schéma.
- Validační profil obsahuje Validační pravidla.
- Dodavatel má Profily dodavatele.
- Profil dodavatele se řídí Ingest workflow.
- Uživatel patří pod Dodavatele.
- Uživatel má Roli.
- Role má Oprávnění.
- Archivní balíček vychází z Balíčku původce.
- Badatelský balíček vychází z Archivního balíčku.
- Archivní balíček obsahuje popisné ARCLib XML.
- Dávka se řídí Profilem dodavatele.
- Dávka je spuštěna nebo naplánována Uživatelem.
- Časový úkol je definován Uživatelem.
- Časový úkol má vstup definován přes Uložené hledání.
- Profil dodavatele se řídí Export workflow.
- Ingest Workflow používá Lokaci.
- Ingest Workflow používá Externí nástroj.
- Ingest Workflow má na vstupu Balíček původce.
- Ingest Workflow produkuje Archivní balíček.
- Export Workflow má na vstupu Archivní balíček.
- Export Workflow produkuje Badatelský balíček.

Location (Lokace)

Externí lokace, ke které se ARCLib umí připojit. Podporované protokoly jsou:

- lokálně připojené úložiště,
- úložiště připojené přes SFTP protokol,
- úložiště připojené přes FTP protokol,
- úložiště připojené přes CIFS protokol,
- úložiště připojené přes NFS protokol.

Tool (Externí nástroj)

Libovolný konfigurovatelný nástroj, který se dá využít při Ingestu. Nástroj musí být dostupný ve formě skriptu s možností předání parametrů.

Validation Scheme (Validační schéma)

XSD validační schéma evidované v systému.

Validation Rule (Validační pravidlo)

Validační pravidlo, které je možné spustit na definovaném vstupu a jehož výsledkem je správnost/nesprávnost vstupu. Existují tři typy validačních pravidel:

1. kontrola souboru/části souboru vůči Validačnímu schématu,
2. kontrola přítomnosti a hodnoty elementu/atributu pomocí XPath a regulárních výrazů (regexů),
3. kontrola přítomnosti souboru.

Validation Profile (Validační profil)

Soubor Validačních pravidel s definicí vstupních dat pomocí XPath.

Supplier (Dodavatel)

Organizace, která je vedena v ARCLib a má přiřazené uživatele systému.

Supplier Profile (Profil dodavatele)

Přiřazení konkrétního Validačního profilu a Ingest Workflow k Dodavateli.

Ingest workflow

Popis zpracovaní balíčku pri Ingestu. Je definován pomocí BPMN 2.0 notace.

Export workflow

Popis zpracovaní balíčku při Exportu. Je definován pomocí BPMN 2.0 notace.

User (Uživatel)

Uživatel systému patřící pod Dodavatele s definovanými právy.

Permission (Oprávnění)

Oprávnění k vykonávání konkrétní činnosti v systému.

Role (Role)

Zástupní role pro Uživatele, na kterou jsou vázána Oprávnění.

SIP

Balíček na vstupu do archivu, který se po celou dobu archivace nemění.

AIP

Archivní balíček, základní prvek ukládání do archivu, který podporuje verzování.

DIP

Balíček, jenž vzešel z exportu z ARCLibu. V současné době odpovídá AIP.

ARCLib XML

XML s vyextrahovanými a doplněnými metadaty pro AIP.

Batch (Dávka)

Jednotlivé spuštění Ingestu/Exportu. Na vstupu má konfiguraci, včetně definice vstupních dat a Profilu dodavatele.

Time Job (Časový úkol)

Časový úkol, který definuje vytvoření Dávky v definovaný čas s daným vstupem.

Search Query (Uložené hledání)

Uložený filtr pro hledání balíčků.

1.2.4 Moduly ARCLib systému dle OAIS

1.2.4.1 ARCLib Ingest modul

ARCLib Ingest předpokládá vstup dat v pokročilejší fázi zpracování, tedy data, která už mají podobu plnohodnotného SIP definovaného obsahového standardu vytvořeného systémy, jako jsou ProArc nebo Archivematica, zabalené do kontejneru BagIt.

SIP vstupující do ARCLibu musí zahrnovat obsahovou informaci (content information) sestávající z datového objektu a informace o reprezentaci (data object + representation information), a dále Informaci o uchovávání (preservation description information).

ARCLib Ingest modul:

- validuje příchozí SIP a vytvoří metadata o této validaci do ARCLib AIP XML souboru (každý typ vstupujících dat je validován podle šablony, která je vytvořena podle informací poskytnutých producentem o standardu, kterým se řídí výroba SIP; každý typ vstupujících dat tak musí disponovat odlišnou validační šablonou)
- extrahuje ze SIP některá metadata (ID dodavatele do ARCLib, původní lokaci SIP, původní ID SIP, popisná metadata, některá technická a administrativní metadata, údaje o fixity)
- vytvoří nová metadata (záznam o vložení, datum, údaj o kompletnosti balíku, fixity, a výsledek identifikace formátů v ARCLib Ingest workflow),
- SIP BagIt nástroj je zabalí do TAR formátu, vygeneruje/pojmenuje UUID a uloží do struktury generované z UUID.

Systém ARCLib neobsahuje pre-Ingest nebo depositní modul, nepředpokládá se konverze nezpracovaných vstupních dat do formátu SIP. Pokud vkladatel disponuje pouze naskenovanými daty a metadaty (v MARCXML, Dublin Core apod.), musí použít externí aplikaci (ProArc, Archivematica, svoje scripty) pro vytvoření SIP v požadované struktuře, a na straně ARCLibmusí zadefinovat profil vstupujících dat a Ingest workflow (pokud tento již neexistuje).

ARCLib modul Ingest je schopen zpracovat následující struktury vstupujících dat:

- ProArc NDK monografie

- ProArc NDK periodika
- ProArc nativní monografie a periodika
- ProArc zvukové dokumenty
- NDK periodika a monografie
- Archivematica DSpace
- Archivematica Obecný
- NDK elektronické dokumenty

Zpracování vstupujících dat je řízeno konfigurovatelným workflow (BPM), v jehož rámci jsou dostupné následující kroky:

- připojení profilu dodavatele
- přesun balíku do pracovního úložiště
- validace přesunu
- antivirová kontrola
- identifikace a validace formátu balíku (BagIt)
- kontrola MD5
- extrakce balíku
- identifikace formátu všech souborů v balíku (Siegfried, DROID, FIDO)
- validace XML
- extrakce MD5 z informací v balíku
- kontrola MD5 všech souborů
- extrakce popisných metadata a jejich validace
- generování fixity pomocí jiných algoritmů než MD5
- záznam eventů vložení, validace do PREMIS
- složení ARCLib AIP XML

Pro vstupní zpracování dat je využíván sdílený diskový prostor (vstupní úložiště, vlastní pracovní úložiště a archivní úložiště). Systém pro řízení workflow zaznamenává událost přesunu AIP z pracovního do archivního úložiště.

Pro konfiguraci systému řízení zpracování vstupu dat jsou využívány registry (kroků, dostupných nástrojů, profilů SIP, dodavatelů, lokací zdrojových dat a registr XSD).

Administrátor systému může využívat databázi rozpracovaných dat a dat připravených k vložení. Díky ní může pracovat se SIP, které neprošly validací – může je opravit, odmítnout či vrátit dodavateli. Zde je

možné i změnit použitý validační profil. Administrátor i vkladatel dat mohou procházet data připravená k Ingestu.

Ingest podléhá reportingu.

1.2.4.2 ARCLib Data Management modul

Funkce správy dat spočívají především v oblasti udržování databáze AIP, její aktualizace a řízení jejího využívání. Rozsah využívání databáze v ARCLibu ale přesahuje požadavky OAIS funkční entity Správa dat (Data management), databáze je využívána i v dalších částech systému, a to pro evidenci dalších informací, např. v průběhu zpracování dat v Ingestu; pro uchovávání informací o nastavení systému, jeho uživatelích, workflow apod.

ARCLib Data management:

- obsahuje informace o AIP (resp. jejich částech) uložených v trvalém úložišti v modulu Archival storage,
- umožňuje vyhledávání a indexaci AIP
- umožňuje prohlížení obsahu AIP a editaci metadat AIP, a to včetně možnosti vytvořit jeho novou verzi,
- poskytuje reporting.

Vyhledávat lze nad popisnými metadaty, administrativními metadaty a technickými metadaty vytvořenými v ARCLib (tedy v rozsahu obsahu ARCLib AIP XML). Vyhledávací dotazy lze kombinovat a ukládat jako sady výsledků (logický set), které je možno dále použít k exportu dat. Minimální rozsah informací o AIP zahrnuje identifikátory, popisná metadata, administrativní metadata, auditní informace o Ingestu, informace o dodavateli dat, informace o původním SIP. Funkce pro vyhledávání jsou dostupné přes API.

Z vyhledávacího prostředí modulu Data management lze vyvolat událost exportu DIP (předpokládá se, že DIP je shodný s AIP) do pracovního úložiště, jednotlivě nebo hromadně lze prohlížet informace ARCLib AIP XML. Je možno zobrazit strukturu AIP a prohlížet jeho obsah.

Editací metadat AIP se rozumí aktualizace ARCLib AIP XML. Po editaci systém XML validuje a vytvoří novou verzi AIP. Editace metadat je dostupná prostřednictvím API.

Reporty obsahují statistiky zpracování, statistiky AIP podle různých kritérií, a kompletní reporty o obsahu úložiště. Generování reportů je automatizované, časovatelné a výstup je možné ukládat v Excelu, PDF a CSV formátu.

1.2.4.3 ARCLib Administration modul

Modul Administrace umožňuje konfigurovat workflow pro zpracování Ingestu a kontroluje infrastrukturu systému ARCLib. Obsahuje registry uživatelů a jejich rolí, provádí se zde nastavení jejich autentizace, udržuje se registr profilů SIP a registr dodavatelů dat, registr kroků workflow pro Ingest a scriptů v nich používaných, registr validačních profilů, registr storage lokací – pracovních a trvalých storage poolů. Tento modul zajišťuje připojení a komunikaci s databází, kontroluje, zda stačí dostupná storage kapacita. Nastavují se zde některé administrativní úkoly (čistění cache, čistění pracovního úložiště, kontroly po restartu a pádu, spouští se a časují Ingest workflow), kontroluje se stav databáze, index, logy.

V systému jsou dostupné následující uživatelské role nebo jejich kombinace:

- administrátor – konfiguruje systém, nastavuje workflow pro Ingest, pro generování DIP atd.
- dodavatel – spouští Ingest
- analytik – řeší problémy vznikající při Ingestu
- editor – může měnit metadata AIP

1.2.4.4 ARCLib Archival Storage modul

Archival Storage modul je komplexní služba pro zajištění bitové ochrany umožňující využití replikace dat do více geografických lokalit a využití více technologií ukládání dat. Archival Storage směrem k ARCLibu poskytuje Object Storage použitelné přes jednoduché REST rozhraní, které je obsahem Storage Service Gateway.

Funce Archival Storage modulu zahrnují příjem a výdej AIP, udržování informací o lokaci balíků, udržování provozních metadat, aktualizaci metadat, kontrolu integrity, propojení na konkrétní lokaci ukládací technologie, replikaci dat do více lokalit, zálohování, správu úložných technologií a médií a reportování.

REST rozhraní poskytuje mimo jiné i funkce pro řízení přístupu k datům, tj. poskytuje například rozhraní pro specifikaci autentizačních údajů klienta služby (náhodný token apod.), na základě kterých je možné rozlišovat různé uživatele a řídit tak přístup k datům. Archival Storage také vede podrobné protokoly o přístupu k datům a manipulaci s nimi.

Směrem k úložišti pak tento modul ukládá data v definované struktuře na distribuovaný file systém, do Ceph object storage nebo jiné vhodné technologie, která umožní dosažení cílů Archival Storage z hlediska bezpečnosti uložení dat a manipulace s nimi.

Důležitým požadavkem Archival Storage modulu je potřeba dostatečného výkonu a jeho škálování při masivně paralelním přístupu k úložišti.

Archival Storage je od systému abstrahován a přistupuje se k němu jako samostatné aplikaci. Je tak využitelný i mimo projekt ARCLib.

1.2.4.5 ARCLib Access modul

Filosofie přístupu pro systém ARCLib předpokládá, že uživatelé potřebují získat zpět vložená data v původním stavu, respektive po případné migraci musí být zachován jejich informační obsah. Modul tedy v první řadě umožní export AIP jako DIP s tím, že obsah AIP a DIP je identický.

ARCLib je back-end aplikace a není určena pro koncové uživatele. Nevynucuje tedy dodržování konkrétní politiky omezující přístup k datům AIP. Metadata k Access Rights jsou součástí dodaného SIP a jsou kontrolována při Ingestu, ale nejsou konvertována do ARCLib AIP XML.

1.2.4.6 ARCLib Preservation Planning modul

Velká část funkcí OAIS funkční entity Preservation Planning bude realizována mimo systém ARCLib. Definice a monitorování určené komunity a monitorování technologií jsou činnosti především výzkumné a organizační povahy, a jejich výkon je předmětem zájmu řady komunit.

Každý provozovatel ARCLib by měl mít vlastní formulace strategií a výklad standardů pro dlouhodobou archivaci a pro svoje konkrétní data by měl mít plány ochrany a migrační plány. Cílem ARCLib není poskytnout prostředí pro vytváření a testování plánů migrace dat, nicméně prostředky a možnosti systému musí umožnit opakovaný ingest balíčků upravených v externích systémech, efektivní správu dat a důvěryhodné zpracování Ingestu.

V rámci ARCLib je realizována databáze/registr formátů napojený na registr formátů PRONOM (Technical registry PRONOM, 2017). Báze formátů ARCLib obsahuje informace o použitých formátech, nástrojích pro identifikaci formátů a číselníky souvisejících chyb a pravidel pro jejich řešení. Databáze je doplněna skripty spouštějícími validační a identifikační nástroje. Databáze je plně pod kontrolou uživatele.

1.2.5 Informační balíčky v systému ARCLib

Vzhledem k tomu, že systém ARCLib je koncipován jako dark archive určený zejména pro využití v rámci odborných institucí, které mají dlouhodobou ochranu digitálních objektů přímo v náplni své práce, lze předpokládat, že tvůrci informačních balíčků budou zejména odborní uživatelé, kteří v rámci své činnosti produkují strukturované, standardizované a dobře dokumentované informační balíčky. U výstupních balíčků se naopak na straně uživatele ARCLibu (knihovny, archivu) předpokládá další zpracování a úpravy před případným zpřístupněním svému koncovému uživateli (čtenáři, badateli) ve vlastním rozhraní (digitální knihovně). Tyto skutečnosti významně ovlivňují způsob práce i samotný charakter SIP a DIP a nepřímo i strukturu AIP. Systém ARCLib by měl být primárně využíván knihovnami a dalšími paměťovými institucemi. Jeho informační balíčky proto odpovídají standardům využívaným v knihovní síti v ČR, tedy zejména standardům Národní digitální knihovny (Standardy digitalizace, 2017). Pro využívání systému je třeba připravit mapování každého typu vstupujících dat (SIP profil) na strukturu AIP ARCLib. Mapování musí vytvořit producent dat ve spolupráci se správci systému ARCLib. Jen on má dostatečné znalosti svých dat. K splnění nároků na dlouhodobou logickou ochranu dat je třeba, aby všechny SIP profily byly dostupné uživatelům systému a aby zároveň byly archivované jak v samotném systému jako AIP, tak i externě.

Systém ARCLib pracuje s předem danou skupinou různých struktur/typů SIP, nicméně je navržen tak, aby bylo možno tuto skupinu v budoucnu rozšiřovat, a to minimálně v prostoru ohraničeném mezinárodně uznávanými standardy pro digitální repozitáře působící v knihovním prostředí.

V SIP i AIP systému ARCLib je používána celá řada metadatových standardů, jejichž množina je vzhledem k charakteru systému prakticky neomezená. Za klíčové je ovšem možno považovat následující množinu formátů, které jsou buď aktivně využívány v rámci systému ARCLib XML, nebo jsou významnou součástí struktur/profilů předpokládaných v základní verzi.

- **METS** – Metadata Encoding and Transmission Standard[23]
  Schéma využívané pro zachycení administrativních a strukturálních metadat. Jedná se o kontejnerový formát umožňující vnoření dalších metadatových formátů.
- **PREMIS** Data Dictionary for Preservation Metadata[24]
  Schéma určené k zachycení metadat nezbytných pro dlouhodobou archivaci digitálních objektů. Zachycuje jak technické vlastnosti objektů, tak mimo jiné historii jejich zpracování a správy.
- **Dublin Core**[25]
  Schéma popisných metadat. Skládá se z patnácti prvků, které mohou být případně rozšířeny kvalifikátory. Je určen zejména pro základní orientační popis a pro sdílení metadatových záznamů.
- **MODS** – Metadata Object Description Schema[26]
  Schéma popisných metadat umožňující detailní popis digitálních objektů, který se granularitou přibližuje běžně používaným katalogizačním záznamům.
- **copyrightMD** – autorsko-právní metadata[27]
  Schéma využívané k zápisu informací o autorsko-právních souvislostech dat.
- **MIX** – Metadata for Images in XML Schema[28]

  Metadatové schéma pro zápis technických a administrativních metadat pro statické obrazy.
- **ARCLib Schema**

  Jedná se o interní metadatové schéma vlastní systému ARCLib. Bylo vytvořeno pro zápis části provenance metadat, která užitá standardní schémata nebyla schopna zaznamenat. ARCLib schéma vychází ze standardu PREMIS a v metadatovém záznamu ho lze identifikovat podle prefixu ARCLib. Je tvořeno následujícími elementy (viz tab. 2). Všechny jsou opakovatelné.

Název elementu	**Význam elementu**
ARCLib:format	Informace o formátu uložených dat. Agregované údaje z celého AIP
ARCLib:fileCount	Informace o počtu souborů
ARCLib:identificationTool	Informace o nástroji, který provedl identifikaci a charakterizaci AIP
ARCLib:device	Informace o zařízení, na kterém byla data vytvořena, např. typ skeneru

[23] http://www.loc.gov/standards/mets/.
[24] http://www.loc.gov/standards/premis/.
[25] http://dublincore.org/.
[26] http://www.loc.gov/standards/mods/.
[27] http://www.cdlib.org/groups/rmg/.
[28] http://www.loc.gov/standards/mix

ARCLib:deviceId	Identifikátor zařízení, např. sériové číslo
ARCLib:eventAgent	Informace o názvu události
ARCLib:eventType	Informace o typu události
ARCLib:agentName	Informace o názvu agenta
ARCLib:date	Datum události
ARCLib:eventCount	Informace o počtu zaznamenaných událostí

Tabulka 2 – Elementy ARCLib schema

- **ARCLib AIP XML Schema** – detailní popis viz kapitola 2.2.5.2. Liší se od ARCLib schématu, které je jen dílčí součástí ARCLib AIP XML Schema.

1.2.5.1 Vstupní informační balíček – SIP

SIP je obecným pojmenováním formátů dat vstupujících do systému ARCLib. Veškeré podporované SIP formáty budou v systému ARCLib definovány pomocí tzv. SIP profilů.

Systém ARCLib je schopen uložit balíčky z různých repozitářových a produkčních systémů, balíčky pro různé typy dokumentů i balíčky, které se sice vztahují ke stejným druhům dokumentů pocházejícím ze stejných systémů, ale jedná se o odlišné verze implementace metadatového standardu. Klíčové jsou zejména standardy používané v rámci Národní digitální knihovny (NDK) (Standardy digitalizace, 2017). Dále jsou významné i SIP pocházející z repozitářových systémů DSpace a Kramerius (respektive Fedora) a z produkčního systému ProArc.

Vstupní balíčky budou zpracovány na základě SIP profilů definovaných ve spolupráci s tvůrcem SIP. SIP profil obsahuje informace o datech a metadatech obsažených v balíčku a o pravidlech jejich mapování do ARCLib AIP XML. Původní SIP je vždy součástí uloženého AIP. SIP profily slouží kromě mapování na ARC XML schéma též k validaci vstupujících balíčků.

Rozsah verzí lze pozměnit podle přání uživatele; bude nutné jej specifikovat podle předpokládaného rozsahu využívání systému. Systém ARCLib je variabilní, lze ho rozšiřovat o další typy vstupujících dat, která budou vytvořena podle užitých mezinárodních standardů (základním předpokladem je využití standardu METS[29] jako kontejnerového formátu pro ostatní metadata).

1.2.5.2 Archivní informační balíček – AIP

ARCLib AIP je označení datového formátu pro archivaci balíčků v systému ARCLib. ARCLib je systém pro správu archivních balíčků, nejedná se o systém pro koncové uživatele nebo systém pro management popisných metadat. Z toho vychází i návrh struktury ARCLib AIP, jenž je ukládán jako dva samostatné fyzické objekty – původní SIP a ARCLib XML, které jsou spojené pomocí logické vazby.

ARCLib XML je metadatová struktura specifická pro systém ARCLib. ARCLib AIP XML po vytvoření během Ingestu obsahuje informace ze dvou zdrojů – část vzniká v průběhu ingestu na základě informací obsažených v SIP, část je generována systémem ARCLib. Zpracování dat se řídí SIP profilem, který je specifický pro každý typ dat od konkrétního vkladatele. SIP profil obsahuje pravidla pro mapování povinných a nepovinných metadatových údajů do struktury ARCLib AIP XML.

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

SIP profil je přiřazován množině balíčků na základě druhů dat, použitého standardu, zdrojového systému a dalších parametrů vycházejících z dohody mezi tvůrcem SIP a provozovatelem systému ARCLib.

Struktura ARCLib AIP XML vychází ze standardu METS. Dále je využit standard PREMIS a standard Dublin Core a specifické ARCLib schéma vycházející ze standardu PREMIS, které slouží k zachycení části technických metadat a části provenienčních metadat. Většina těchto metadat je agregovaná ze SIP, váží se k celému balíčku a zároveň k jednotlivým souborům.[30] Struktura ARCLib AIP XML se skládá z následujících částí:

- *Kořenový element METS a hlavička* – obsahuje identifikátory balíčku, údaje o jeho vytvoření a případné modifikaci. Identifikuje původce dat a dodavatele (agenta). Všechny údaje v této sekci jsou povinné. Údaje jsou vytvářeny systémem ARCLib a přebírají se ze SIP.
- *Sekce popisných metadat* – obsahuje vnořený metadatový záznam ve formátu Dublin Core. Jednotlivé údaje se přebírají ze SIP na základě SIP profilu. V závislosti na konkrétním SIP profilu může být sekce popisných metadat rozšířena o další záznam definovaný profilem a identifikovaný specifikátorem záznamu popisných metadat v rámci systému ARCLib.
- *Sekce administrativních metadat* – do skupiny administrativních metadat spadá několik sekcí vnořených metadat. Jsou to:
    - *Technická metadata* – zachycují informace o technických vlastnostech objektu. Údaje jsou vytvářeny systémem ARCLib a zároveň se některé údaje přebírají ze SIP. Váží se k celému balíčku a k jednotlivým souborům.
    - *Provenance metadata* – zachycují informace o událostech v historii objektu a o kontrolách provedených při ingestu. Údaje jsou vytvářeny systémem ARCLib a zároveň se některé údaje přebírají ze SIP. Vychází ze standardu PREMIS a mohou používat specifikace jednotlivých událostí, jak jsou popsány v Kontrolovaném slovníku ochranných událostí PREMIS.[31] Váží se k celému balíčku a k jednotlivým souborům.
    - *Sekce práv k digitálnímu objektu* – zachycují informace o právech k digitálnímu objektu. Tato sekce je nepovinná a přebírá se ze SIP.
    - *Další bibliografická metadata* – extrahují se ze SIP dle potřeb uživatele (např. záznam ve formátu MODS). Tato sekce je nepovinná a přebírá se ze SIP.
- *Sekce vazeb na soubory* – zachycuje vazby na soubory, které jsou obsahem AIP. Údaje jsou vytvářeny systémem ARCLib a vycházejí z obsahu SIP.
- *Strukturální mapa digitálního objektu* – zachycuje strukturální mapu obsahu AIP. Údaje jsou vytvářeny systémem ARCLib a vycházejí z obsahu SIP.

[30] Agregací se v tomto kontextu rozumí souhrnný popis vztahující se ke všem souborům v balíčku, které mohou mít např. odlišné souborové formáty. Zápis proto představuje záznam všech hodnot, které se v daném atributu vyskytují u všech souborů v SIP.

[31] https://www.loc.gov/standards/premis/events-announcement.html.

Informace – typ	Zdroj	Povinné	Umístění v METS	Další komentáře
Root element				
mets	Generuje ARCLib	A	mets	
	Hlavička			
Hlavička	Generuje ARCLib	A	metsHdr	Identifikuje původce dat a dodavatele (agenta) a datum vytvoření či modifikace záznamu
Popisná metadata				
Generický Dublin Core	Extrakce ze SIP	A	dmdSec mdWrap	Váže se na celý SIP
Vazba mezi verzemi AIP	Generuje ARCLib	A	dmdSec mdWrap dc:isVersionOf	
Specifické sady Dublin Core	Extrakce ze SIP	N	dmdSec mdWrap	Identifikováno specifikátorem
Externí popisná metadata	Extrakce ze SIP nebo připojení při Ingestu	N	dmdSec mdRef	
Technická a administrativní metadata				
Technická metadata generovaná ARCLib pro celý SIP	Generuje ARCLib	A	techMD premis object	
Technická metadata převzatá z profilu BagIt nebo odjinud k celému SIP BagIt	Extrakce ze SIP	N	techMD PremisObject	
Technická metadata pro soubory	Generuje ARCLib	A	techMD ARCLib Schema PremisObject	Agregováno, váže se na celý objekt SIP
Provenance a kontroly při Ingestu	Generuje ARCLib	A		Zvlášť pro celý SIP i pro jednotlivé soubory v něm
Provenance přebrané ze SIP	Extrakce ze SIP	N	digiprovMD ARCLib Schema premis Event	Sumárně pro celý SIP

Provenance při aktualizaci AIP metadat	Generuje ARCLib	A	digiprovMD premis Event	PREMIS Event – události jsou specifikované v Preservation Event Controlled Vocabulary https://www.loc.gov/standards/premis/events-announcement.html
Provenance při revalidaci novým profilem	Generuje ARCLib	A	digiprovMD premisEvent	Popis průběhu revalidace novým profilem ; PREMIS Event – události jsou specifikované v Preservation Event Controlled Vocabulary http://www.loc.gov/standards/premis/v3/preservation-events.pdf
Provenance při opakování identifikace formátů v SIP	Generuje ARCLib	A	digiprovMD premisEvent + nové verze PremisObject	Popis průběhu opakované identifikace formátů všech souborů v SIP; PREMIS Event – události jsou specifikované v Preservation Event Controlled Vocabulary http://www.loc.gov/standards/premis/v3/preservation-events.pdf
Práva				
Práva	Extrakce ze SIP	N	rightsMD	
Jiné údaje				
Další metadata	Extrakce ze SIP podle potřeb uživatele	N	sourceMD	Uživatel může volně použít pro svoje administrativní informace – v profilu stanoví, odkud extrahovat
Filesec, strucmap a další				
Filesec	Generuje ARCLib	A		Rozsah popisu – obsah SIP
structMap	Generuje ARCLib	A		Rozsah popisu – strukturální mapa obsahu SIP

Tabulka 3 – Struktura ARCLib AIP XML

Příklad SIP profilu

Příklad SIP profilu slouží jen k ilustraci mapování. Vychází z jednoduché monografie zpracované podle NDK standardu.

ARCLib XML	**NDK standard**
METS:mets/ LABEL	mets:label
METS:mets/ TYPE	mets:type
Popisná metadata – generický Dublin Core	

ARCLib XML	**NDK standard**
title	mods:title
description	mods:partNumber
description	mods:partName
creator, contributor	mods:name
type	mods:typeOfResource
type	mods:genre
coverage	mods:place
publisher	mods:publisher
date:issued	mods:dateIssued
date:created	mods:dateCreated
language	mods:language
format	mods:form
format:extent	mods:extent
description:abstract	mods:abstract
description	mods:note
subject	mods:topic
subject	mods:geographic

ARCLib XML	**NDK standard**
subject	mods:temporal
subject	mods:name
identifier	mods:identifier
source	mods:shelfLocator
Technická metadata – agregovaná, fulltext	
/ARCLib:format	premis:formatName
Technická metadata – přebíraná, fulltext	
/ARCLib:fileFormat	premis:formatName
/ARCLib:fileCount	
/ARCLib:identificationTool	premis:formatDesignation
/ARCLib:device	mix:captureDevice
/ARCLib:deviceId	mix:scannerModelSerialNo
/ARCLib:fileCount	
Provenance přebraná ze SIP – sumárně	
/ARCLib:eventAgent	
/ARCLib:eventType	premis:eventType
/ARCLib:agentName	premis:agentName

ARCLib XML	NDK standard
/ARCLib:date	premis:eventDateTime
/ARCLib:eventCount	
mets:rightsMD	mets:rightsMD
mets:sourceMD	mets:sourceMD

Tabulka 4 – Příklad SIP profilu

Verzování AIP

ARCLib umožňuje verzování AIP. Verzování bude řešeno ve dvou úrovních:

- verzování ARCLib XML,
- verzování celého AIP.

V případě, že je verzován pouze ARCLib XML, vzniká nová verze XML souboru a je dodatečně archivována jako součást aktuální verze AIP (tedy AIP = SIP + ARCLib AIP XML + ARCLib AIP XMLv2). Nová verze ARCLib XML vzniká dvěma způsoby:

- změnou ARCLib XML uživatelem v systému ARCLib,
- ingestem již archivovaného SIP, kdy při spuštění ingestu je uživatelem definováno, že se jedná pouze o aktualizaci metadat.

Pokud je ingestován SIP, který již byl archivován (tedy jeho autorské id je již v systému evidováno) a při spouštění ingest není uživatelem definováno, že se jedná pouze o aktualizaci metadat, verzuje se celý AIP. Nová verze AIP je na úrovni metadat (pole dc:isVersionOf) navázána na původní verzi AIP, který se z archivu nemaže.

Verze AIP je tedy v podstatě definována verzí ARCLib XML, které v sobě obsahuje vazbu na korespondující verzi SIP (v METS:fileSec resp. METS:structMap).

Mazání AIP

ARCLib uživatelům umožňuje mazání AIP. Balíček je primárně smazán logicky (označením za vymazaný na úrovni Workspace) a přestává se uživatelům zobrazovat v přehledech.

Zároveň je implementován mechanismus odpadkového koše, ve kterém jsou logicky smazané balíčky evidovány. Tento seznam balíčků je dostupný pouze administrátorům systému, kteří mají možnost smazat balíčky fyzicky (skutečným odstraněním dat z archivních úložišť, přičemž jsou uchována metadata balíku v databázi a indexu), anebo smazání stornovat (zpětnou změnou příznaku na úrovni Workspace).

Fyzické smazání balíčku vyžaduje ještě potvrzení uživatelem s oprávněním schvalovat fyzické mazání. Veškeré mazání dat je podrobně logováno do auditních logů.

1.2.5.3 Výstupní informační balíček – DIP

Vzhledem k charakteru systému ARCLib je DIP totožný s uloženým AIP. Export konkrétní verze ARCLib AIP bude spočívat v získání dané verze ARCLib XML z archivního úložiště, vyčtení korespondující verze SIP, získání tohoto SIP z archivního úložiště a vyexportování této dvojice.

1.2.6 Doporučení pro procesní dokumentaci, změnový management

Mezi další důležité standardy a metodiky, které jsou relevantní především pro certifikaci repozitářů a archivů většího až velkého rozsahu, patří ITIL, ISO 16363, set standardů ISO 27000 a ISO 31000.

1.2.6.1 ČSN ISO 16363

Standard ČSN ISO 16363 (319621) Systémy pro přenos dat a informací z kosmického prostoru – Audit a certifikace důvěryhodných digitálních úložišť stanovuje doporučený postup pro posuzování důvěryhodnosti digitálních úložišť. Je součástí souboru standardů, které odkazují na důvěryhodný a odpovědný management dat a kurátorství. Cílem je prosazování kvality, respektování integrity dat a závazek dlouhodobého uchovávání dat a přístupu k nim. Vzhledem ke své organizační, časové a finanční náročnosti má standard své opodstatnění zejména pro větší až velké digitální knihovny, repozitáře či archivy. Ale i jako auditní nástroj poslouží pro vyhodnocení spolehlivosti, zavázanosti a připravenosti institucí převzít na sebe zodpovědnost za dlouhodobé uchovávání obsahu. Metriky ISO 16363 se zaměřují na organizační strukturu, způsob správy digitálních objektů, správu infrastruktury a bezpečnost (autenticita uložených informací, možnost převodu dat do jiného repozitáře apod.). Norma obsahuje celkem 108 kritérií (normativních metrik), rozdělených do tří základních kategorií:

A. Organizační infrastruktura (Organizational Infrastructure) – kapitola 3 v rámci standardu

1. Řízení a životaschopnost organizace.

2. Organizační struktura a personální zabezpečení.

3. Procedurální zodpovědnost a strategický rámec.

4. Finanční udržitelnost.

5. Smlouvy, licence a závazky.

B. Správa digitálních objektů (Digital Object Management) – kapitola 4 v rámci standardu

1. Ingest: akvizice obsahu.

2. Ingest: tvorba archivních balíčků.

3. Plánování dlouhodobé ochrany.

4. Archivní úložiště & ochrana/správa AIP.

5. Informační management.

6. Správa přístupu.

C. Technologie, technická infrastruktura a bezpečnost (Technologies, Technical Infrastructure, & Security) – kapitola 5 v rámci standardu

1. Systémová infrastruktura.

2. Vhodné technologie.

3. Bezpečnost.

Repozitář, archiv či úložiště na úrovni organizačního celku se dotýká pracovníků všech úrovní. Zatímco vedení a nižší management musí znát alespoň požadavky na důvěryhodné úložiště z části A (Organizační struktura), systémoví administrátoři, síťoví správci a další techničtí pracovníci, kteří zodpovídají za mnohé části infrastruktury, budou pracovat s částí C (technologie, technická infrastruktura a bezpečnost). Producenti a příjemci dat naleznou relevantní informace především v dokumentaci pro část A a B.

1.3 Implementační část – doporučení

1.3.1 Doporučení pro plán ochrany uložených dat

Plánování dlouhodobé ochrany dat uložených pomocí systému ARCLib je součástí obecných principů dlouhodobé ochrany tak, jak jsou popsány v předchozích částech. Tuto činnost musí provádět odborný pracovník repozitáře, respektive se při ní musí řídit doporučeními relevantních institucí na národní či nadnárodní úrovni.

Převažujícím principem využívaným při zajišťování dlouhodobého uložení digitálních dat v současné době zůstává zejména formátová migrace. Úkolem odborných pracovníků tedy musí být důsledné zmapování formátů uložených v systému dlouhodobé ochrany s použitím mezinárodních formátových registrů. Aktuálně lze doporučit využívání registru PRONOM[32]. Důraz je třeba klást na odlišné varianty týchž formátů, které lze odlišit mimo jiné právě pomocí formátových registrů. Na základě vstupních validací by měl být vytvářen registr formátů dat uložených v konkrétní instalaci systému ARCLib. To je primární odpovědnost dané instituce a jejích pracovníků. Nelze ji nahradit žádnou spoluprací ani následováním doporučení jiných institucí. Nezbytnou součástí tohoto opatření je mít přehled výskytu datových formátů v konkrétních balíčcích a samozřejmě také mít možnost vyhledávat podle těchto kritérií.

Na základě fungujícího registru použitých formátů v daném úložišti musí odborní pracovníci sledovat míru jejich rizikovosti, zejména pokud jde o dostupnost nástrojů pro zobrazení dat. Tento požadavek lze naplnit v zásadě dvěma způsoby, které se doplňují a částečně překrývají. Jedna cesta klade důraz na sledování doporučení centrálních institucí. V České republice zatím jde jen o Národní knihovnu ČR, jejíž Odbor digitálních fondů by měl vydávat varování v případě, že některému z formátů doporučených NK ČR hrozí zastarání. Lze očekávat, že v budoucnu se k tomu připojí další instituce, které budou obdobně

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

standardizovat data ze své oblasti působnosti. V mezinárodním prostředí se pak lze obracet zejména na Kongresovou knihovnu v USA, která vydává pravidelné zprávy, ve kterých přináší seznam doporučených formátů pro jednotlivé typy dat z hlediska jejich perspektivy pro dlouhodobé uchovávání.[33] Obdobné služby však poskytují i formátové registry, jejichž sledování tato metodika doporučuje jako primární zdroj informací. Uživatelům systému ARCLib je doporučeno sledovat formátový registr PRONOM a podle jeho návodu přijímat další opatření. Souběžně s tímto je třeba sledovat upozornění vydávaná centrálními institucemi alespoň na národní úrovni. V případě homogenního archivu je toho obvykle daná instituce schopna sama. To je ale jen ojedinělý stav, obvykle bude rozmanitost použitých formátů mnohem rozsáhlejší a využívání doporučení nebo formátových registrů nutné.

Sledování rizik (ať již kontrolou ve formátovém registru nebo ověřování doporučených formátů) je třeba provádět kontinuálně podle písemného plánu, který musí být součástí dokumentace úložiště. V předem určených intervalech musí být aktivně kontrolovány registry formátů nebo ověřována doporučení centrálních institucí. Doporučená frekvence kontroly rizik uložených formátů je alespoň jedenkrát za kalendářní rok, u statických obrazů alespoň jednou za dva roky.

V případě zjištění rizikovosti formátu musí odborný pracovník instituce spustit postup pro eliminaci rizika. Doporučeným řešením je aktuálně provést formátovou migraci, což je nákladný proces, jak po stránce ekonomické, tak technické a personální. Je tedy nutné případná rizika včas rozeznat, aby měli správci úložiště dostatek prostoru na provedení ochranných opatření. Obecně strategie plánování dlouhodobé ochrany digitálních dat popisuje kapitola č. 1.1.4 této metodiky, které také uvádí další možné alternativy k formátové migraci.

Z hlediska úložiště představují riziko tzv. kontejnerové formáty, které v sobě mohou mít zanořena data jiných formátů. Těmto datům je nutné při plánování ochranných opatření věnovat zvláštní pozornost a identifikovat i zanořené formáty. Pro dlouhodobé uložení jsou nejméně rizikové statické obrazy, jejichž plán ochrany může mít relativně delší časové periody mezi jednotlivými kontrolami. Aktivní pozornost je naopak třeba věnovat zejména vědeckým datům, jejichž různorodost je velká. Dalším doporučením pro plánování dlouhodobé ochrany je definování vhodných softwarových nástrojů pro zobrazení uchovávaných dat, které odborní pracovníci v pravidelných, předem daných, intervalech kontrolují. Nicméně tento postup je značně náročný a nelze ho doporučit jako standardní řešení.

1.3.2 Doporučení pro organizační a personální zajištění projektu

Je třeba, aby instituce, která se rozhodne provozovat digitální archiv ARCLib, měla několik klíčových zaměstnanců (nebo přímo oddělení), kteří budou systém aktivně spravovat a používat. Tito zaměstnanci musí mít potřebné znalosti a schopnosti pro využití všech dostupných vlastností ARCLib systému tak, aby zabezpečoval dlouhodobou ochranu digitálních dat v konkrétní instituci. Provoz systému nemá být omezen pouze na jedno oddělení; mělo by to být právě naopak. Existence systému na dlouhodobou ochranu ovlivní organizaci jako celek, tedy i ostatní oddělení mimo samotný systém. Důležitou roli, nikoli však klíčovou, hraje IT oddělení. Rozvoj a používání systému by mělo stát mimo IT, respektive mělo by být v gesci odborných pracovníků se znalostí teorie dlouhodobé ochrany digitálních dat.

[33] http://www.loc.gov/preservation/resources/rfs/TOC.html.

Oblast	**Aktivita**	**Potřebné znalosti a schopnosti**	**Role**
Správa obsahu a metadat	- Podpora ingestu - Příprava dat na ingest - Řešení problémů při ingestu (formáty dat, virus, metadata) - Plánování ochrany – vytváření plánů a jejich provádění - Vytváření reportů - Editace dat a metadat - Opakovaná validace formátů - Mazání dat a metadat - Vytváření potřebných procesů a jejich údržba - Analýzy obsahu a metadat - Export obsahu - Vývoj a správa metadatových schémat - Hromadné opravy dat a metadat vně systému - Využívání externích nástrojů a aplikací - Sledování trendů dlouhodobé ochrany dat a dostupných nástrojů - Sledování Cílové skupiny - Vytváření a údržba potřebných politik	- znalost využívaných metadatových standardů a schémat (analýza, vývoj, editace, mazání) - znalost práce s nástroji na validace formátů a extrakci metadat - znalost konceptů spojených s logickou dlouhodobou ochranou dat - reportování - zkušenost se správou dat nebo systému na to určeného - sledování trendů v oblasti dlouhodobé ochrany - komunikace s cílovou skupinou (různé kanály), analýza a vyhodnocení zpětné vazby - metodické usměrňování (politiky)	Analytik digitální archivace Správce obsahu
Vyšší management	- Vyjednávání s vedením organizace - Vyjednávání s externími uživateli - Dlouhodobé plánování - Politiky - Provádění auditů nebo jiných kontrol kvality - Propagace dlouhodobé ochrany v rámci organizace, zavádění potřebných procesů - Dlouhodobé plánování potřeb pro storage (nárůst obsahu) - Finanční zajištění a správa	- vedení a usměrňování týmu - změnový management, projektový management - vytváření střednědobých až dlouhodobých plánů - zajišťování dostatečného počtu personálních, technických a finančních kapacit - komunikace s vedením organizace, případné reportování na úrovni repozitáře/archivu jako celku	Manažer týmu

Oblast	**Aktivita**	**Potřebné znalosti a schopnosti**	**Role**
		- spolupráce při komunikaci a propagaci navenek organizace	
Komunikace	- Komunikace navenek – propagace aktivit (konference) - Komunikace uvnitř organizace - Externí konzultace k systému nebo obecně k dlouhodobé ochraně dat - Informování externích a interních uživatelů o změnách, odstávkách, nové funkcionalitě, nových verzích apod. - Podpora uživatelů	- podpora vyššího managementu při propagaci - spolupráce s oblastí Správy dat a metadat (komunikace s cílovou skupinou, podpora uživatelů)	Manažer týmu, případně odpovídající role v instituci již existující
Provoz a údržba	- Komunikace s IT - Účast na plánování rozvoje storage a infrastruktury - Správa ARCLib systému a jeho infrastruktury - Procesní a znalostní dokumentace - Provádění updatů ARCLib a OS serverů - Monitoring HW a SW - Prioritizace nutných oprav a komunikace s dodavatelem ARCLib systému; případně opravy chyb - Technické nastavení a úpravy ARCLib – skripty apod. - Podpora uživatelů	- znalost potřebných technologií, jejich silných a slabých stránek - management rizik - procesní, technická a systémová podpora - spolupráce s oblastí Komunikace (podpora uživatelů)	IT odborník Analytik digitální archivace Správce obsahu
Plánování rozvoje a vylepšení	- Udržování plánu rozvoje a oprav - Komunikace s ostatními uživateli systému - Plánování nasazení nových verzí do ostrého provozu - Testování nových verzí systému	- spolupráce se všemi oblastmi repozitáře/archivu (zejména Technická podpora a Provoz a údržba) - implementace nových verzí systému, údržba	IT odborník Analytik digitální archivace

Oblast	Aktivita	Potřebné znalosti a schopnosti	Role
	- Udržování registru vylepšení a oprav (bug/enhancement tracking)	registru (úpravy, vylepšení atd.)	
Technická podpora	- Údržba infrastruktury - Monitoring - Správa procesní dokumentace	- znalost potřebných HW technologií, jejich silných a slabých stránek - správa procesní dokumentace a řešení rizik - aktivní monitoring a správa infrastruktury	IT odborník

Tabulka 5 – Aktivity spojené s provozem systému ARCLib a jejich personální zajištění[34]

1.3.3 Doporučení pro odhad finančních nákladů na provoz systému pro dlouhodobé ukládání

Jedním z kritérií posuzovaných při hodnocení důvěryhodnosti digitálního repozitáře je oblast zahrnující finanční plánování, analýzu investic, výdajů, rizik a přínosů a také finanční předpovědi s různými alternativami vývoje (viz ČSN ISO 16363, kapitola 3.4 Ekonomická udržitelnost, body 3.4.1 a 3.4.3).

V rámci projektu 4C[35] proběhl průzkum (Stokes, 2014) mezi institucemi, které se zabývají digitální archivací. Mezi nejčastější důvody pro zjišťování informací finanční povahy spojených s digitální archivací patří odhad budoucích nákladů, tvorba rozpočtů, podpora rozhodování a hodnocení alternativ.

Znalost nákladů a navázaných nákladových elementů je zásadní pro obhajobu požadavků na finance poskytnuté zřizovatelem. Tyto informace lze také použít jako důkaz, že repozitář spravuje své zdroje efektivně. Pro udržení a další rozvoj digitální archivace v instituci je nezbytné přesvědčit ty, kteří mají rozhodovací pravomoci, o důležitosti a přínosech této aktivity.

Zpráva organizace The Blue Ribbon Task Force (Blue Ribbon Task Force on Sustainable Digital Preservation and Access, 2008) uvádí, že budoucí přístup k digitálním objektům závisí na ochranných opatřeních, která jsou provedena v současnosti. Funkční strategie dlouhodobé ochrany musí brát v potaz nejen technické, právní a sociální aspekty, ale nesmí zapomínat ani na ty ekonomické.

Z ekonomického hlediska představují nejvýznamnější rizika, kterým čelí ekonomicky udržitelná dlouhodobá ochrana digitálních dokumentů:

Nedostatečnost jednorázového financování (jako jsou projekty a granty) pro zajištění trvalého uchování digitálních dat.

[34] Inspirováno tabulkou vytvořenou v Harvard Library pro Digital Preservation Repository (Goethals, 2017).
[35] Projekt 4C je zaměřen an lepší porozumění nákladům na dlouhodobou archivaci digitálních objektů.

Špatné sladění mezi zainteresovanými stranami, jejich rolemi, povinnostmi a ekonomickými modely. Producenti, správci a uživatelé dat jsou různé skupiny s odlišnými zájmy a rozdílným financováním.

Nedostatek iniciativ na podporu spolupráce, jejímž cílem by bylo zajistit ekonomicky udržitelnou dlouhodobou ochranu digitálních dokumentů.

- Neopodstatněná spokojenost se současným stavem v instituci, i když nebyly použity žádné nástroje na hodnocení kvality a důvěryhodnosti (např. audit dle ISO 16363 nebo DSA).
- Přílišný strach z komplikovanosti digitální archivace, kvůli kterému nebudou prováděny žádné akce.

Použití modelů pro výpočet nákladů na dlouhodobou ochranu (DPC modely)[36] může repozitáři a nadřízené organizaci pomoci ve zvládnutí finančního managementu. Využití těchto modelů je vhodné pro lepší porozumění a řízení aktivit digitální archivace všude tam, kde jsou tyto aktivity provozovány, nebo se s nimi do budoucna počítá.

Kirnn KAUR a kol. definují DPC modely jako:

„*…reprezentaci aktivity dlouhodobé ochrany [digitálních dokumentů], která může být sdílena, podrobena kontrole a kritice a jejímž úkolem je osvětlit náklady na aktivity spojené s dlouhodobou ochranou digitálních dokumentů*“. (Kaur et al., 2013, s. 11)

Z definice plyne, že DPC model nemusí nutně náklady vyčíslit v penězích – stačí, když ukáže, na které aktivity se náklady vážou. Zároveň by však mělo jít o formální a strukturovaný přístup, který je ale dost otevřený na to, aby mohl být kontrolován, kritizován a případně upraven.

Většina DPC modelů je založena na rozdělení workflow na jednotlivé aktivity. Dílčí náklady se po své identifikaci a vyčíslení sečtou do celkových nákladů na provoz repozitáře.

Ty se mohou skládat z následujících dílčích nákladů:

- náklady na pre-ingest (včetně nákladů na výběr dokumentů a jejich digitalizaci),
- náklady na ingest dat do repozitáře,
- náklady na formátovou migraci (může jít o průběžnou migraci při vstupu do repozitáře, nebo o formátovou migraci části repozitáře kvůli zastarávání formátů),
- náklady na updatování dat a metadat,
- náklady na pravidelnou obnovu hardwaru a softwaru (např. výměna úložných médií nebo serverů, změna softwarových nástrojůapod.),
- náklady na zpřístupnění dat uživatelům (např. vytváření a dodávání DIP),
- náklady na konec životního cyklu digitálního dokumentu.

Repozitář může provádět jen některé z uvedených aktivit – v závislosti na povaze spravovaných dat, typu a poslání repozitáře (Kaur et al., 2013).

Definování a rozdělení nákladových elementů podle aktivit, zdrojů a času:

- Aktivity – Náklady jsou strukturovány podle aktivit, které se mohou skládat z jednoho či více procesů nebo funkcí. Většina DPC modelů používá pro popis aktivit v repozitáři normu ISO 14721 (model OAIS).

- Zdroje – Náklady mohou být děleny také podle zdrojů nákladů:

[36] Digital Preservation Costs (DPC) – náklady na digitální uchovávání.

- Kapitálové (investiční) náklady – např. náklady na pořízení nebo obnovu zařízení, jako jsou servery, úložná média a další.
- Mzdové náklady – mohou se dělit podle úrovně vzdělání zaměstnanců, nebo pracovního zařazení (IT odborník, analytik digitální archivace, správce obsahu, manažer, programátor...).
- Přímé a nepřímé náklady – v případě DPC se za přímé náklady považují např. nákupy úložných médií, nebo práce na konkrétním úkolu (tvorba metadat). Nepřímé náklady pak mohou být např. správa sítě pro celou organizaci, celková administrace a management instituce apod.
- Variabilní a fixní náklady – variabilní jsou ty náklady, u nichž se s růstem objemu výkonů mění jejich celková výše.

- Čas – Náklady mohou být děleny také podle toho, zda zachycují náklady v minulosti (ty jsou využívány pro finanční účetnictví), nebo zda se jedná o odhad budoucích nákladů (ty jsou využívány pro finanční management).
  - Jednorázové a opakované náklady – mezi jednorázové lze zařadit např. nákup páskové knihovny, mezi opakované např. periodický nákup pásek do této knihovny, energie či mzdové náklady.
  - Odpis majetku – fyzické opotřebení majetku, nebo jeho morální zastarávání znamená snížení ekonomického prospěchu, odpis majetku je tedy též považován za druh nákladu.
  - Inflace a úrokové sazby – je nutné vzít v potaz i hodnotu peněz, která může klesat – pak jde o inflaci, nebo stoupat – pokud jde o deflaci. Protože kapitálové a mzdové náklady se mohou v budoucnosti měnit (např. kvůli technologickému vývoji), je při výpočtu budoucích nákladů nutné brát v úvahu různé varianty vývoje inflace a deflace.

	Čas								
	Časový úsek 1								Časový úsek 2
	Zdroj (Kč)								Zdroj
	Jednorázové náklady				Opakované náklady				
	Přímé náklady		Nepřímé náklady		Přímé náklady		Nepřímé náklady		
Aktivita	Kapitálové náklady	Mzdové náklady	Kapitálové náklady	Mzdové náklady					
A	100 000	50 000	10 000	10 000					
B									
C...									

Tabulka 6 – Příklad obecné nákladové šablony (Stokes, 2014, s. 18)

V minulých letech proběhlo několik hodnocení a porovnání existujících DPC modelů. Oba níže zmíněné projekty lze využít jako další zdroje informací o nákladech a nákladových modelech. Jsou také dobrými průvodci pro výběr vhodného DPC modelu pro odhad nákladů v repozitáři.

4C Project

- Summary of Cost Models. 4C Project. Dostupné z: http://4cproject.eu/summary-of-cost-models.
- STOKES, Paul (ed.). D3.1—Evaluation of Cost Models and Needs & Gaps Analysis. 4C Project, 2014. Dostupné také z: http://www.4cproject.eu/documents/D3.1_final_report_10May2014-v1.02.pdf.

APARSEN

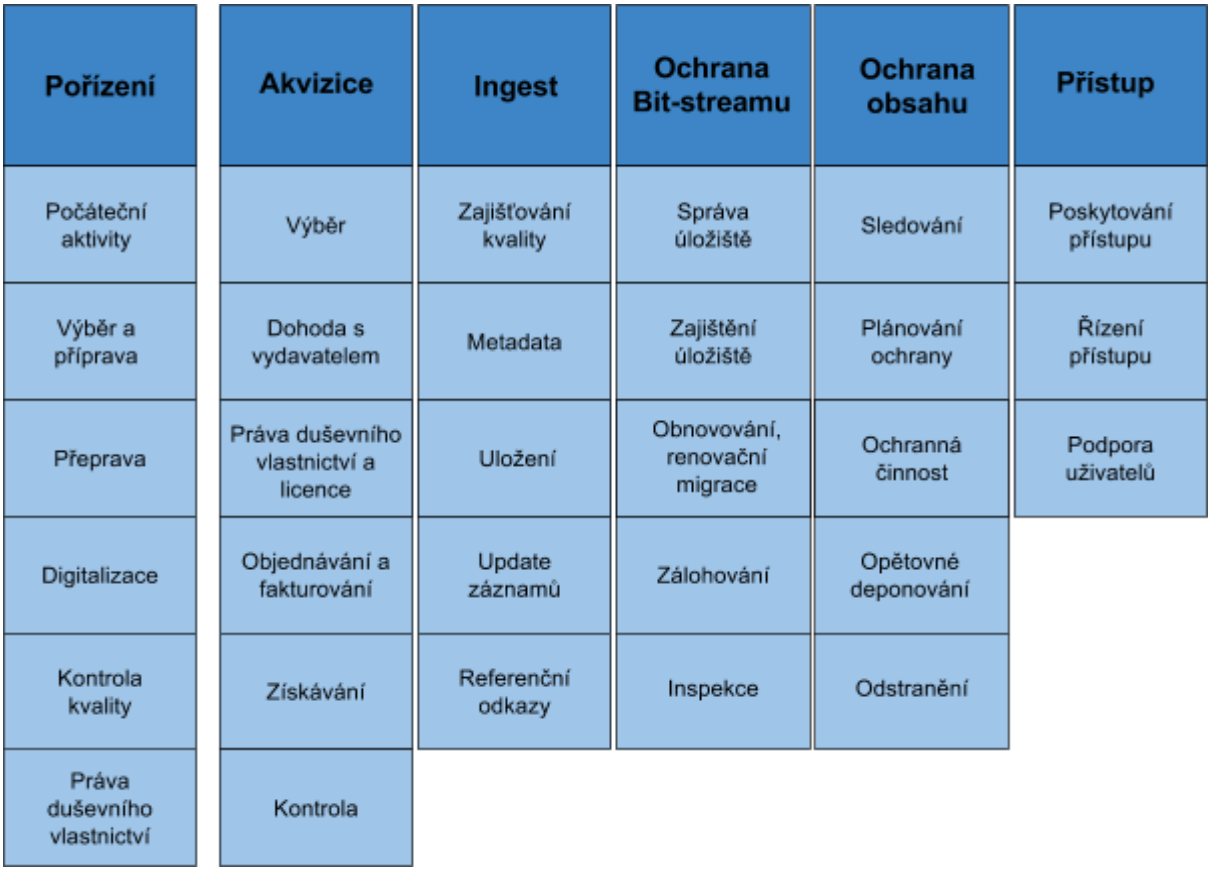
- KAUR, Kirnn et al. APARSEN. D32.1 Report on Cost Parameters for Digital Repositories. 2013. Dostupné také z: http://www.alliancepermanentaccess.org/wp-content/uploads/sites/7/downloads/2014/06/APARSEN-REP-D32_1-01-1_0_incURN.pdf.

Jako nejvhodnější pro knihovní prostředí se jeví model LIFE3 Costing Model (LIFE), za nímž stojí University College London (UCL) a The British Library (BL):

1. Model umožňuje výpočet nákladů v minulosti, v současnosti i do budoucnosti.
2. Pokrývá jak hardwarové, tak i personální náklady.
3. Nabízí široké pokrytí aktivit (včetně digitalizace a pre-ingestu) a pracuje s vysokým množstvím proměnných.
4. Je založen na referenčním modelu OAIS a byl vyvinut v knihovnickém prostředí.
5. Je implementován do podoby tabulky MS Excel; byl vyvinut prototyp webového nástroje, ale ten je nedostupný.[37]

Projekt LIFE3 identifikoval v případových studiích konkrétní činnosti, na nichž jsou náklady navázány. Tento přehled typických aktivit při zpracování dat (včetně digitalizace, nebo jiného způsobu pořízení) může sloužit jako vzor pro mapování nákladů ve vlastním repozitáři:

[37] Prototyp webového nástroje měl být umístěn zde: http://lifedev.hatii.arts.gla.ac.uk/. Nedostupný k 11. 7. 2017.



Obrázek 9 – Celkové náklady na životní cyklus digitálních dokumentů dle modelu LIFE3. [38]

Nástroj LIFE3 na výpočet DPC

Vzniklý model byl implementován do podoby nástroje ve formě sešitu MS Excel a je dostupný on-line.[39] Zde uvedeme krátký popis nástroje:

První list (Basic input) obsahuje pouze několik polí a umožňuje základní odhad nákladů, případně rychlé porovnávání alternativních scénářů. Pro tento základní odhad byla využita data z případových studií v rámci projektu LIFE. Pro české prostředí ale nenabízí odpovídající nastavení, proto je nutné použít pokročilé funkce nástroje, ve kterém lze změnit standardní nastavení a lépe ho přizpůsobit našemu prostředí.

Do tabulky se zadávají následující údaje:

Počáteční rok a koncový rok – zde se určuje počet let, na které je odhad vytvořen.

Typ dat, který bude do repozitář vkládán – na výběr jsou webové stránky a e-časopisy (digital born dokumenty), tištěné materiály (které musí nejprve projít digitalizací), zvukové nahrávky (digital born

[38] Lifecycle Information for E-Literature: An Introduction to the third phase of the LIFE project [online]. Dostupné z: http://www.life.ac.uk/3/docs/life3_report.pdf.
[39] Nástroj je dostupný ve verzi (3).50 ke stažení na webu projektu LIFE: http://www.life.ac.uk/3/docs/life3_ver50.xls.

nebo analogové), vědecká data, mezi které jsou zařazeny např. akademické práce, datasety z výzkumů, tabulky MS Excel a soubory ve formátu PDF. Náklady na správu uvedených typů dat jsou odhadovány na základě zkušeností z reálných repozitářů, které se podílely na projektu LIFE. Kromě výše uvedených typů je tu ještě možnost "ostatní", do které jsou zahrnuty ostatní typy dat, které nebyly zmíněny výše. Při použití "ostatní" pak model uživatele informuje, že pro výpočet nákladů není možné použít podklady ze žádné z provedených studií.

Zdroj dat – data se mohou do repozitáře dostávat z různých zdrojů, z hlediska nákladů a nástroje LIFE je důležité, zda jsou data zakoupena, darována, nebo vytvořena (digitalizována) vlastními silami. Pokud uživatel zvolí analogová média a jako zdroj dat "vytvoření", tak se objeví další možnost a tou je kvalita obrazových dat – na výběr je nízká, střední a vysoká.[40]

Na základě rozsahu dat, vloženého v bodě 1 se vytvoří časová osa, do které se zapíše předpokládaný roční přírůstek dat – pro digitalizované textové dokumenty je to počet stran, pro vědecká data počet souborů, pro zvukové dokumenty počet nosičů a minut.

Velikost instituce – na výběr jsou tyto velikosti malá, střední nebo velká. To je důležité z hlediska systémové architektury repozitáře a typu a množství úložných médií. Za velké instituce jsou považovány organizace na národní úrovni (např. NK ČR). Většina českých institucí tedy bude patřit mezi střední nebo malé organizace.

Po vyplnění všech potřebných údajů je vypočítán základní odhad nákladů. Pro podrobnější volby jsou připraveny další záložky, které umožňují nástroj přizpůsobit specifikům instituce, která nástroj používá:

- upřesnění profilu organizace,
- upřesnění pořízení dat,
- upřesnění akvizice dat,
- upřesnění deponování dat,
- upřesnění ochrany bit-streamu,
- upřesnění ochrany obsahu,
- upřesnění přístupu.

Dalších devět záložek obsahuje finanční model, který je použit pro výpočet v rámci životního cyklu digitálních dokumentů, a poslední záložka generuje z vložených dat jednoduchý graf, který ukazuje rozložení nákladů na jednotlivé procesy.

I přes absenci externího návodu obsahuje nástroj dostatek poznámek a vysvětlivek, takže se v něm lze poměrně rychle zorientovat. Pro plné využití nástroje je ale nutný hlubší průzkum polí, do kterých se vkládají data a která obsahují proměnné a výpočetní vzorce. MS Excel má omezené uživatelské prostředí, které používání nástroje mírně ztěžuje. Citelně zde chybí uživatelsky přívětivější webový nástroj. .

Vzhledem k tomu, že nástroj vznikl ve Velké Británii, je nezbytné upravit většinu původních údajů – především personální náklady jsou třikrát až čtyřikrát vyšší než v ČR. Další nevýhodou je, že nástroj

[40] Jako příklad nízké kvality obrazových dat uvádí nástroj LIFE3 např. skenování novin, jako vysokou úroveň pak např. rukopisy a fotografie. Úrovní kvality je tedy myšlena obrazová kvalita.

počítá náklady v librách, takže je nutné provádět přepočty na koruny a zpět. Nutností je také převádění jednotek množství dat – nástroj počítá s MB, množství dat v dnešní době se ale pohybuje v řádech TB a PB.

Vzhledem k možným nepřesnostem ve výpočtech a limitům, které nástroj LIFE3 má, je nutné brát výsledky s kritickým odstupem. Cílem by nemělo být určení nákladů s přesností na koruny, to žádný nástroj neumožní. I přibližná výše nákladů (zaokrouhlená na stovky tisíc či miliony korun) a jejich rozložení může repozitáři pomoci pochopit náklady na provoz a připravit se na možný budoucí vývoj.

Lze předpokládat, že hlavní položkou nákladů na provoz repozitáře budou personální náklady. Základem kvalitně fungujícího repozitáře jsou motivovaní zaměstnanci, které je nutné ohodnotit odpovídajícími financemi.

S ohledem na velikost přírůstku dat bude také nutné pořizovat nová úložná média, i když cena za MB uložených dat se díky technologickému pokroku neustále snižuje. Úložná média je také nutné měnit po uplynutí jejich životnosti (u pevných disků je udávána doba 5 let, u LTO pásek a optických médií je tato doba vyšší, záleží na doporučení výrobce). S vyšším počtem úložných médií porostou také náklady na elektrickou energii (zvlášť, pokud se s daty uloženými na médiích bude aktivně pracovat a půjde o pevné disky). V pravidelných intervalech bude nutné měnit i klíčová zařízení infrastruktury, jako jsou servery, routery, UPS a další komponenty. Je vhodné budovat finanční zálohu pro případ, že bude nutné tyto prvky vyměnit náhle, z důvodu akutního selhání.

Na každou z činností v repozitáři bude alokována část celkových nákladů – jak v případě automatických činností (vyžadují strojový čas, elektrickou energii a další zdroje), tak především v případě manuálních činností. Pro přehled činností, které je nutné v repozitáři vykonávat a které budou spotřebovávat náklady viz Tabulka 5 – Aktivity spojené s provozem systému ARCLib a jejich personální zajištění.

1.3.4 Nástroje pro provádění konkrétních procesů logické ochrany

V oblasti dlouhodobé ochrany lze využít celou řadu nástrojů, které umožňují automatizaci jednotlivých procesů logické ochrany. Velká část těchto nástrojů je dostupná v režimu open source. Tyto nástroje mohou být v případě, že disponují vhodným rozhraním a že to licenční podmínky umožňují, použity v rámci systému ARCLib, nebo externě pro práci s daty mimo systém.

Při rozhodování o využití určitého nástroje je dobré sledovat zejména, zda se jedná o komerční nebo open source nástroj. Důležitou otázkou je i míra komplexnosti nástroje – některé nástroje jsou schopny agregovat další služby, jiné jsou specifické. Kombinace několika „drobných“ specifických nástrojů může být ideální z hlediska flexibility a možností přizpůsobení požadavkům organizace, komplexní řešení však zřejmě bude méně náročné z hlediska provozu a údržby (např. pravidelných updatů). Obdobné pravidlo je třeba uplatňovat i vzhledem k tvůrci daného nástroje – program vyvíjený větší uživatelskou komunitou (nebo komerční firmou) nebude zřejmě zcela přesně odpovídat všem požadavkům instituce, nicméně je pravděpodobné, že bude vyvíjen a udržován dlouhodobě. Pozornost je třeba věnovat i pozici nástroje v rámci workflow zpracování dat – nástroj by se neměl stát „úzkým hrdlem“, tedy bodem, ve kterém je celé workflow výrazně zpomaleno například kvůli nárokům na výkon serveru. V neposlední řadě by organizace měla sledovat rozvoj v této oblasti, a to jak u nástrojů již používaných, tak nově vznikajících (Digital Preservation Handbook: Tools).

Níže uvedené nástroje se uplatňují zejména v oblastech validace formátů (tedy kontroly toho, zda odpovídají určité normě nebo standardu), identifikace formátů a extrakce metadat (např. zjištění

technických vlastností), plánování ochrany (poměrně komplexní nástroje, které vedou uživatele při vytváření plánu ochrany), správy dat (typická je např. práce s kontrolním součtem) a při migraci dat.

Informace o dostupných nástrojích shromažďují specializované registry – například Community Owned Digital Preservation Tool Registry COPTR[41] nebo Digital Curation Centre (DCC) tools and services list.[42]

1.3.4.1 Identifikace formátů, extrakce metadat a validace formátů

DROID (Digital Record Object Identification)[43]

Nástroj pro automatizovanou hromadnou identifikaci formátů. Pro identifikaci využívá pravidelně aktualizované signature files z registru formátů PRONOM Technical Registry[44] shared-mime-info[45] databázi a Library of Congress's FDD file format signatures[46].

Jedná se o open source nástroj.

FIDO (Format Identification for Digital Objects)[47]

Nástroj pro automatizovanou hromadnou identifikaci formátů. Pro identifikaci využívá pravidelně aktualizované signature files z registru formátů PRONOM Technical Registry. Používá se přes příkazovou řádku.

Jedná se o open source nástroj napsaný v jazyce Python.

Siegfried[48]

Nástroj pro automatizovanou hromadnou identifikaci formátů. Pro identifikaci využívá pravidelně aktualizované signature files z registru formátů PRONOM Technical Registry, shared-mime-info databázi a Library of Congress's FDD file format signatures. Nástroj se používá přes příkazovou řádku.

Jedná se o open source nástroj.

3-Heights(TM) PDF Validator[49]

Validační nástroj pro práci s PDF. Je schopen validovat vůči normám pro PDF a PDF/A (ISO 19005). Nástroj nabízí možnost využití API nebo práce přes příkazovou řádku.

Jedná se o komerční nástroj.

Jhove[50]

[41] http://coptr.digipres.org/Main_Page.
[42] http://www.dcc.ac.uk/resources/external/tools-services.
[43] http://www.nationalarchives.gov.uk/information-management/manage-information/preserving-digital-records/droid/.
[44] http://www.nationalarchives.gov.uk/PRONOM/Default.aspx.
[45] https://freedesktop.org/wiki/Software/shared-mime-info/.
[46] https://www.loc.gov/preservation/digital/formats/fdd/descriptions.shtml.
[47] http://openpreservation.org/technology/products/fido/.
[48] https://www.itforarchivists.com/siegfried.
[49] http://www.pdf-tools.com/pdf20/en/products/pdf-converter-validation/pdf-validator/.

Nástroj pro identifikaci, charakterizaci a validaci formátů. Rozsah zpracování souboru se řídí použitým modulem. Standardně je poskytován s následujícími moduly AIFF, ASCII, BYTESTREAM, GIF, HTML, JPEG, JPEG 2000, PDF, TIFF, UTF-8, WAVE, XML. Oproti níže uvedenému nástroji Jpylyzer, dokáže Jhove validovat formát JP2 (1. část standardu JPEG2000 ISO/EIC 15444-1) i jpx (2. část standardu JPEG2000) Nástroj je možno ovládat přes příkazovou řádku nebo prostřednictvím GUI.

Jedná se o aktivně rozvíjený open source nástroj v jazyce Java.

Jpylyzer[51]

Nástroj pro validaci a extrakci metadat obrázků ve formátu JP2 (1. část standardu JPEG2000 ISO/EIC 15444-1).

Jedná se o aktivně rozvíjený open source nástroj v jazyce Java.

VeraPDF[52]

Nástroj pro validaci a charakterizaci souborů na základě normy PDF/A (ISO 19005). Umožňuje identifikaci a validaci dle jednotlivých verzí a úrovní normy i na základě uživatelsky stanovené množiny pravidel. Nástroj je možno ovládat přes příkazovou řádku nebo prostřednictvím GUI.

Jedná se o aktivně rozvíjený open source nástroj v jazyce Java.

Epubcheck[53]

Nástroj pro validaci a charakterizaci souborů ve formátu Epub. Je možné jej používat přes příkazovou řádku nebo jako součást Java knihovny. GUI rozhraní je dostupné přes nástroj třetí stany pagina Epub-Checker.[54]

DPF Manager[55]

Open source nástroj pro validaci obrazů ve formátu TIFF, vyvinutý v rámci projektu PREFORMA.[56] Dokáže validovat TIFF soubory oproti několika specifikacím a profilům formátu (baseline TIFF, extended TIFF, TIFF/EP apod.) a též je možné vytvořit si vlastní validační profil (přidat vlastní pravidla). Nástroj je možné používat přes příkazovou řádku nebo prostřednictvím GUI rozhraní.

Jedná se o aktivně rozvíjený open source nástroj.

ExifTool[57]

Nástroj pro extrakci a editaci metadat velkého množství souborových formátů (obrazy, audio, video…). Ovládá se přes příkazovou řádku.

Jedná se o aktivně rozvíjený open source nástroj v jazyce Perl.

---

[50] http://jhove.openpreservation.org/.
[51] http://jpylyzer.openpreservation.org/.
[52] http://verapdf.org/.
[53] https://github.com/IDPF/epubcheck/releases.
[54] https://www.pagina.gmbh/produkte/epub-checker/.
[55] http://www.preforma-project.eu/dpf-manager.html.
[56] http://www.preforma-project.eu/.
[57] https://sno.phy.queensu.ca/~phil/exiftool/.

Metadata Extraction Tool[58]

Nástroj vytvořený Národní knihovnou Nového Zélandu v roce 2003. Extrahuje metadata (převážně technická) z různých souborových formátů (např. JPEG, TIFF, MS Word, PDF, WAV, HTML, XML, ARC). Výstupem z nástroje je dokument ve formátu XML. Ovládá se prostřednictvím grafického rozhraní nebo příkazového řádku.

Poslední verze nástroje vyšla v roce 2014. Metadata extraction tool je součástí nástroje FITS.

FITS (File Information Tool Set)[59]

Nástroj, který provádí identifikaci, validaci a extrakci technických metadat z vícero souborových formátů. Obsahuje v sobě nástroje třetích stran (např. DROID, Exiftool, Jhove, MediaInfo) pro různé druhy formátů, výstupem ze všech nástrojů je jediný XML soubor.

Jedná se aktivně rozvíjený open source nástroj v jazyce Java.

Apache Tika[60]

Nástroj pro extrakci metadat a textu z různých souborových formátů (obraz, text, audio, video..). Výstupem může být soubor ve formát XML, HTML, prostý text, JSON. Ovládá se přes příkazovou řádku, protřednictvím GUI rozhraní nebo jako Java knihovnu.

Jedná se aktivně rozvíjený open source nástroj v jazyce Java.

1.3.4.2 Plánování ochrany

DRAMBORA[61]

Jedná se o nástroj pro interní hodnocení repozitářů. Je možné jej používat on-line přes webové rozhraní, či si stáhnou off-line verzi. Nástroj je díky svému zaměření vhodný pro identifikaci, hodnocení a správu rizik, kterým je repozitář vystaven.

PLATO[62]

PLATO je platforma pro podporu plánování dlouhodobé ochrany. Jedná se o webovou aplikaci, jejíž zdrojový kód je volně dostupný na GitHubu.[63] Integrace JHOVE, DROID a FITS umožňuje testovat různé strategie dlouhodobé ochrany (např. emulaci, nebo různé nástroje na migraci formátů) a vyhodnocovat jejich přínosy a rizika. Díky PLATO je možné provádět plánování a testování ve čtyřech krocích:

- stanovení požadavků,
- testování možných alternativ,
- hodnocení výsledků testů,

[58] http://meta-extractor.sourceforge.net/.
[59] https://projects.iq.harvard.edu/fits.
[60] http://tika.apache.org/.
[61] http://www.repositoryaudit.eu/.
[62] http://www.ifs.tuwien.ac.at/dp/plato/intro/.
[63] https://github.com/openpreserve/plato.

- analýza a vytváření dlouhodobých plánů ochrany.

1.3.4.3 Správa dat (kontrolní součty, kopírování, omezení přístupů)

TeraCopy[64]

Jednoduchý nástroj pro kopírování a přesouvání dat v prostředí operačního systému Windows. Úspěšný transfer umí ověřit pomocí kontrolních součtů, využívá tyto algoritmy: CRC32, MD5, SHA-1, SHA-256, SHA-512, Panama, Tiger, RipeMD, Whirlpool a xxHash. Je volně dostupný pro nekomerční využití a jeho používání lze rozhodně doporučit pro bezpečné kopírování a přesouvání souborů.

Bagger[65] **a bagit-java**[66]

Bagit-java je Java knihovna pro tvorbu balíčků ve formátu BagIt dle specifikace Library of Congress. BagIt je využíván v systému ARCLib a také jej využívá Archivematica. Kromě toho jej však lze využít i pro vytvoření kontrolních součtů a zabalení dat do jednoho balíčku a doplnění o vybraná metadata. Nástroj Bagger je opatřen grafickým rozhraním, které umožňuje snadnější práci se soubory a vytváření balíčků BagIt. Díky konfiguračním souborům ve formátu JSON je možné vytvářet i vlastní profily a povinné metadatové položky (např. tvůrce/dodavatel dat, sbírka, identifikátory, apod.).

1.3.4.4 Datové migrace

Apache PDFBox[67]

Open source nástroj pro práci s PDF dokumenty. Mimo jiné umožňuje konverzi textového dokumentu do PDF a konverzi PDF formátu do formátů JPEG a PNG.

Jedná se o aktivně rozvíjený nástroj.

ImageMagick[68]

Nástroj, resp. balík nástrojů, pro tvorbu, úpravu, skládání a konverzi rastrových obrazů. Dokáže pracovat s více než 200 souborovými formáty. Je možné jej použít pro identifikaci, extrakci metadat, identifikaci poškozených obrazů a konverzi mezi formáty. Používá se přes příkazový řádek.

Jedná se o aktivně rozvíjený open source nástroj s širokou uživatelskou komunitou.

Kakadu[69]

Nástroj pro konverzi z formátu TIFF do formátu JP2. Volně dostupné je demo,[70] které neobsahuje některé funkcionality plné verze, ale je možné jej použít na konverzi jednotlivých obrazů např. dle profilu NDK.

Jedná se o aktivně rozvíjený komerční nástroj.

[64] http://www.codesector.com/teracopy.
[65] https://github.com/LibraryOfCongress/bagger.
[66] https://github.com/LibraryOfCongress/bagit-java.
[67] https://pdfbox.apache.org/.
[68] https://www.imagemagick.org/script/index.php.
[69] http://kakadusoftware.com/.
[70] http://kakadusoftware.com/downloads/.

OpenJPEG[71]

Nástroj pro konverzi formátů TIFF, PNG, BMP, RAW a dalších do formátu JP2.

Jedná se o aktivně rozvíjený open source nástroj.

2. Uplatnění metodiky

Metodika bude uplatněna zejména v organizacích využívajících systém ARCLib, ale umožňuje i nezávislé využití. 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 první (obecnější) rovině je metodika pro logickou ochranu digitálních dokumentů určena všem institucím a jejich odborným pracovníkům, kteří mají dlouhodobé uchovávání těchto dokumentů na starosti. Popisuje principy péče o dlouhodobé uchovávání, nutné požadavky na dlouhodobá úložiště, pravidla důvěryhodných repozitářů a nároky na provozující instituce. V tomto ohledu má metodika ambici být základním dokumentem 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. Význam této činnosti je již nyní značný a do budoucna se bude ještě zvyšovat.

Významnější je však druhé speciální určení metodiky pro uživatele systému ARCLib. Jak již bylo výše řečeno, dlouhodobé uchovávání digitálních dokumentů neznamená jen využívání určitého softwaru, případně ověřeného hardwaru. Ty představují jen nutné prostředky pro péči o uložené dokumenty. Jádro úkolů spojených s logickou ochranou digitálních dokumentů spočívá v dodržování zdokumentovaných procesů, užívání osvědčené praxe a v kvalifikovaném personálu. Konkrétní část metodiky definuje postupy, jak provádět úkony logické ochrany v systému ARCLib. Popisuje jednotlivé funkcionality systému, jeho mapování na funkční prvky OAIS, architekturu informačních balíčků, úkoly, které je nutné vykonávat z hlediska vytvoření důvěryhodného systému, a navrhuje základní strukturu dokumentace nutné pro potvrzení důvěryhodnosti. Metodika tedy představuje podrobnou dokumentaci, která definuje postupy, jakými dosáhnout úspěšné aplikace principů logické ochrany na digitální dokumenty uložené v systému ARCLib. Postupy v každém ze systémů jsou částečně odlišné a nelze je jednoduše přebírat. Metodika proto tvoří nedílnou součást užívání systému ARCLib, který nabízí jiný způsob péče o uložené dokumenty než komerční systémy. Vzhledem k open source charakteru celého systému je nutné, aby jeho uživatelé měli k dispozici veřejně dokumentovanou metodiku. Lze navíc předpokládat, že uživateli systému budou zejména krajské knihovny a specializované knihovny, zaměřené především na uchovávání vědeckých a výzkumných dat. U knihoven těchto typů nelze očekávat, že by měly k dispozici plné spektrum odborníků, kteří budou schopni sami definovat celou škálu procesů dlouhodobého uchovávání. Právě tato funkčně vymezená komunita (tj. kurátoři repozitářů) by měla metodiku využívat při správě dat.

Systém ARCLib a postupy doporučené metodikou budou využity v Knihovně Akademie věd ČR, v. v. i. Vzhledem k veřejnému určení metodiky není třeba uzavírat další smlouvy o jejím užívání.

[71] http://www.openjpeg.org/.

3. Seznam použité literatury

BLUE RIBBON TASK FORCE ON SUSTAINABLE DIGITAL PRESERVATION AND ACCESS. Interim Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access. 2008. Dostupné také z: http://blueribbontaskforce.sdsc.edu/biblio/BRTF_Interim_Report.pdf.

CENTER FOR RESEARCH LIBRARIES. *Certification Report on CLOCKSS* [online]. 2014 [cit. 2017-09-09]. Dostupné z: https://www.crl.edu/sites/default/files/reports/CLOCKSS_Report_2014.pdf.

CLOCKSS Archive: documentation Wiki [online]. Stanford: Stanford University, 2014 [cit. 2017-09-09]. Dostupné z: https://documents.clockss.org/index.php?title=Main_Page.

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.

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.

Data Seal of Approval [online]. [cit. 2017-09-09]. Dostupné z: https://www.datasealofapproval.org/en/.

Data Seal of Approval: Český překlad Směrnic DSA [online]. 2015 [cit. 2017-09-09]. Dostupné z: http://dsa.cuni.cz/DSA-3.html?look=new.

Data Seal of Approval: Evropský rámec pro audit a certifikaci [online]. 2017 [cit. 2017-09-09]. Dostupné z: http://dsa.cuni.cz/DSA-8.html?look=new.

DCC, DPE. *DRAMBORA interactive: Digital Repository Audit Method Based on Risk Assessment* [online]. Digital Curation Centre and DigitalPreservationEurope, 2008 [cit. 2017-09-09]. Dostupné z: http://www.repositoryaudit.eu/about/.

Digital Preservation Handbook: Tools [online]. 2nd Edition. Digital Preservation Coalition, 2015 [cit. 2017-09-09]. Dostupné z: http://www.dpconline.org/handbook/technical-solutions-and-tools/tools.

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

DIGITALPRESERVATIONEUROPE (DPE). *Repository Planning Checklist and Guidance* [online]. 2008 [cit. 2017-09-09]. Dostupné z: https://issuu.com/dpe-staff/docs/repository_planning_checklist_and_guidance.

DigitalPreservationEurope. *CORDIS* [online]. Brussel: European Commision, 2016 [cit. 2017-09-09]. Dostupné z: http://cordis.europa.eu/project/rcn/101694_en.html.

Dublin Core Metadata Initiative [online]. DCMI, 2017 [cit. 2017-09-09]. Dostupné z: http://dublincore.org/.

DVOŘÁK, Tomáš, Karel KOUCKÝ, Jaroslav ŠULC, Jiří VICHTA a Milan VOJÁČEK, NÁRODNÍ ARCHIV, STÁTNÍ OBLASTNÍ ARCHIV V PRAZE. *Metodika pro vytváření bezpečnostních kopií archiválií v digitální podobě*. Praha: Národní archiv, Státní oblastní archiv, 2015. Dostupné také z: http://cesarch.cz/wp-content/uploads/2015/06/metodika-pro-bezpecnostni-digitalizaci_v1.pdf.

European Framework for Audit and Certification of Digital Repositories [online]. [cit. 2017-09-09]. Dostupné z: http://www.trusteddigitalrepository.eu/Welcome.html.

GOETHALS, Andrea. Who Does What? Defining the Roles & Responsibilities for Digital Preservation. *Signal* [online]. 2017 [cit. 2017-09-09]. Dostupné z: https://blogs.loc.gov/thesignal/2017/04/who-does-what-defining-the-roles-responsibilities-for-digital-preservation/.

NATIONAL LIBRARY OF AUSTRALIA. *Guidelines for the Preservation of Digital Heritage*. 2003. [online].Canberra: National Library of Australia, 2003 [cit. 2017-03-13]. Dostupné z: http://unesdoc.unesco.org/images/0013/001300/130071e.pdf

HRUŠKA, Zdeněk. *Náklady na dlouhodobou ochranu digitálních dokumentů v českém prostředí na příkladu Moravské zemské knihovny* [online]. Brno, 2017 [cit. 2017-09-10]. Dostupné z: http://is.muni.cz/th/217895/ff_m/. Diplomová práce. Masarykova univerzita, Filozofická fakulta. Vedoucí práce Petr Žabička.

HUTAŘ, Jan. *Úvod do ochrany digitálních dat* [online]. Verze 1.0. Praha: Ústav informačních studií a knihovnictví FF UK, 2008 [cit. 2017-09-09]. Dostupné z: http://uisk.ff.cuni.cz/wp-content/uploads/sites/62/2016/01/%C3%9Avod-do-ochrany-digit%C3%A1ln%C3%ADch-dat_Huta%C5%99.pdf.

Information Standards Quarterly: Special issue: Digital Preservation [online]. 2010, **22**(2) [cit. 2017-03-13]. ISSN 1041-0031. Dostupné z: http://www.niso.org/apps/group_public/download.php/4250/FE_Bishoff_Digital_Preservation_Plan_isqv22no2.pdf.

KAUR, Kirnn et al. APARSEN. D32.1 *Report on Cost Parameters for digital Repositories*. 2013. Dostupné také z: http://www.alliancepermanentaccess.org/wp-content/uploads/sites/7/downloads/2014/06/APARSEN-REP-D32_1-01-1_0_incURN.pdf.

KUNZE, J., J. LITTMAN, L. MADDEN, E. SUMMERS, A. BOYKO a B. VARGAS. *BagIt File Packaging Format (V0.97)* [online]. Network Working Group, 2016 [cit. 2017-09-09]. Dostupné z: https://tools.ietf.org/html/draft-kunze-bagit-14.

LAVOIE, Brian. Meeting the challenges of digital preservation: The OAIS reference model. *OCLC Newsletter* [online]. 2000, **2000**(No. 243), 26–30 [cit. 2017-09-09]. Dostupné z: http://www.oclc.org/research/publications/library/2000/lavoie-oais.html.

LIFE [online]. [cit. 2017-04-03]. Dostupné z: http://www.life.ac.uk/.

Lifecycle Information for E-literature: An Introduction to the third phase of the LIFE project [online]. London, 2010 [cit. 2017-07-11]. ISBN 978-0-7123-5839-2. Dostupné z: http://www.life.ac.uk/3/docs/life3_report.pdf.

LTP Portál.cz. cca 2015 [online]. Brno: Moravská zemská knihovna, cca 2015 [cit. 2017-03-13]. Dostupné z: http://ltp-portal.mzk.cz.

LYNCH, Clifford. 1999. Canonicalization: A Fundamental Tool to Facilitate Preservation and Management of Digital Information. *D-Lib Magazine* [online]. 1999, **5**(9) [cit. 2016-12-09]. DOI: 10.1045/september99-lynch. ISSN 1082-9873. Dostupné z: http://www.dlib.org/dlib/september99/09lynch.html.

MOHANTY, Rasmita a Ranjir KUMAR DAS. 2014. *Proceedings of National Conference on Digital Libraries: Reshaping Traditional Libraries into Next Generation Libraries (NCDL-2014)* [online]. 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.

Nestor Seal for Trustworthy Digital Archives [online]. nestor-Siegel, 2017 [cit. 2017-09-09]. Dostupné z: http://www.dnb.de/Subsites/nestor/EN/Siegel/siegel.html.

'OAIS Introductory Guide (2nd Edition)' DPC Technology Watch Report by Brian Lavoie now available. *OCLC Research* [online]. Leiden: OCLC, 2014 [cit. 2017-09-09]. Dostupné z: http://www.oclc.org/research/news/2014/12-04.html.

OCLC AND CRL. *Trustworthy Repositories Audit & Certification: Criteria and Checklist* [online]. Version 1.0. Chicago (Illinois), Dublin (Ohio), 2007 [cit. 2017-09-09]. Dostupné z: http://www.crl.edu/sites/default/files/d6/attachments/pages/trac_0.pdf.

Paradigm. 2008. [online]. Oxford: University of Oxford, 2008 [cit. 2017-03-13]. Dostupné z: http://www.paradigm.ac.uk/.

PHILLIPS, Megan, et al. The NDSA levels of digital preservation: Explanation and uses. In: *Archiving Conference*. Society for Imaging Science and Technology, 2013. s. 216–222. Dostupné též z: http://ndsa.org/documents/NDSA_Levels_Archiving_2013.pdf.

PREMIS: Preservation metadata maintenance activity [online]. Washington: Library of Congress, 2017 [cit. 2017-09-09]. Dostupné z: https://www.loc.gov/standards/premis/.

ROSENTHAL, Colin, Asger BLEKINGE-RASMUSSEN a Jan HUTAŘ. *Průvodce plánem důvěryhodného digitálního repozitáře (PLATTER).* Praha: Národní knihovna České republiky, 2009. ISBN 978-80-7050-569-4. Dostupné též z: http://www.ndk.cz/platter-cz/Platter.pdf.

SIERMAN, Barbara. OAIS 2012 update. *Digital Preservation Seeds* [online]. 2012 [cit. 2017-09-09]. Dostupné z: http://digitalpreservation.nl/seeds/oais-2012-update/.

SMITH, Kari R. Visualizing PAIMAS and OAIS. *Engineering the Future of the Past: Blog about MIT Libraries' Digital Archives Work* [online]. Cambridge (Massachussets): MIT Libraries, 2013 [cit. 2017-09-09]. Dostupné z: https://libraries.mit.edu/digital-archives/visualizing-paimas-and-oais/.

Standardy digitalizace [online]. Praha: Národní knihovna, 2017 [cit. 2017-09-09]. Dostupné z: http://www.ndk.cz/standardy-digitalizace.

STOKES, Paul, ed. *D3.1—Evaluation of Cost Models and Needs & Gaps Analysis*. 4C Project, 2014. Dostupné také z: http://www.4cproject.eu/documents/D3.1_final_report_10May2014-v1.02.pdf.

Technical registry PRONOM [online]. Richmond: National Archives, 2017 [cit. 2017-09-09]. Dostupné z: https://www.nationalarchives.gov.uk/PRONOM/Default.aspx.

Ten Principles. *Center for Research Libraries* [online]. [cit. 2017-09-09]. Dostupné z: https://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/core-re.

THE BRITISH LIBRARY. Life3_ver50. [List Microsoft Excel 97–2003]. Ver. 50. London, 2009. Dostupné také z: http://www.life.ac.uk/3/docs/life3_ver50.xls.

VERMAATEN, Sally, Brian LAVOIE a Priscilla CAPLAN. Identifying Threats to Successful Digital Preservation: the SPOT Model for Risk Assessment. *D-Lib Magazine* [online]. 2012, **18**(9/10), - [cit. 2017-09-09]. DOI: 10.1045/september2012-vermaaten. ISSN 1082-9873. Dostupné z: http://www.dlib.org/dlib/september12/vermaaten/09vermaaten.html.

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

VYCHODIL, Bedřich a Zuzana KVAŠOVÁ. Projekt LIFE: Návrh metodiky pro sledování nákladů životního cyklu v rámci Národní knihovny v Praze za využití softwaru vyvinutého v rámci projektu LIFE. Praha, 2011.

WATERS, Donald a John GARRETT. *Preserving digital information: report of the Task Force on Archiving of Digital Information* [online]. Washington, D.C.: Commission on Preservation and Access, 1996 [cit. 2017-09-10]. ISBN 1-88733450-5. Dostupné z: https://www.clir.org/pubs/reports/pub63.

4. Seznam publikací, které předcházely metodice

CUBR, Ladislav, Iveta LODROVÁ a Zdeněk VAŠEK. Srovnání vybraných národních identifikačních systémů užívajících identifikátory URN:NBN. *ProInflow* [online]. 2016, **8**(1), 13-53 [cit. 2017-09-12]. ISSN 1804–2406. Dostupné z: http://www.phil.muni.cz/journals/index.php/proinflow/article/view/1220.

FOJTŮ, A. *Dôvera ako kľúčový koncept digitálnych repozitárov*. FAI Knižničná a informačná veda 25 = Library and information science 25 / ed. Jela Steinerová. Bratislava : Univerzita Komenského, 2014.

HUTAŘ, Jan a Marek MELICHAR. The long decade of digital preservation in heritage institutions in the Czech Republic: 2002–2014. *International Journal of Digital Curation* [online]. 2015, **10**(1), -. DOI: 10.2218/ijdc.v10i1.324. ISSN 1746-8256. Dostupné z: http://www.ijdc.net/index.php/ijdc/article/view/324.

HUTAŘ, Jan. Archives New Zeland: budování digitálního archivu pro dlouhodobou ochranu digitálních dokumentů. *Archivní časopis*. 2013, **63**(1), 5-24. ISSN 0004-0398.

HUTAŘ, Jan. Assessing Digital Preservation Strategies. In: *International Council of Archives Congress 2012* [online]. Brisbane, 2012, s. 10 [cit. 2017-09-10]. Dostupné z: http://ica2012.ica.org/files/pdf/Full%20papers%20upload/ica12Final00155.pdf.

MELICHAR, Marek a Jan HUTAŘ. Principy strategie rozvoje knihoven oblasti dlouhodobé archivace digitálních informací v České republice: stav v roce 2014 a výhled do roku 2019. *Duha* [online]. 2014, **28**(1) [cit. 2015-01-19]. ISSN 1804-4255. Dostupné z: http://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* [online]. 2014, **Vol 6**(No 1), 26-37 [cit. 2017-09-13]. ISSN 1804–2406. Dostupné z: http://www.phil.muni.cz/journals/index.php/proinflow/article/view/943.

ROSENTHAL, Colin; BLEKINGE-RASMUSSEN, Asger, HUTAŘ, Jan. *Průvodce plánem důvěryhodného digitálního repozitáře* (PLATTER). 1. vyd. Praha : Národní knihovna ČR, 2009. 65 s. ISBN 978-80-7050--569-4.

5. Terminologický slovník

Terminologický slovník vychází z českého překladu normy OAIS (ISO 14721) a je doplněn o některé další výrazy a termíny specifické pro systém ARCLib. Zbylé definice jsou převzaty z českého překladu normy.

Archiv (archive)

Organizace, jejímž záměrem je uchovávat informace tak, aby mohly být zpřístupněny určené skupině a aby je tato určená skupina mohla využívat.

Archivní informační balíček (archival information package) AIP

Informační balíček, který je složen z informačního obsahu a přidružených informací o uchovávání a je uchováván v archivu OAIS.

Archivní informační jednotka (archival information unit) AIU

Archivní informační balíček, u kterého se archiv rozhodl nerozdělovat informační obsah na další archivní informační balíčky; AIU se může skládat z více digitálních objektů (např. z více souborů).

Archivní informační sbírka (archival information collection) AIC

Archivní informační balíček, jehož informační obsah je souhrnem dalších archivních informačních balíčků,

Celkový popis (overview description)

Zvláštní případ popisu sbírky, který popisuje sbírku jako celek.

Data (data)

Opakovaně interpretovatelná vyjádření informací ve formalizované podobě vhodné pro komunikaci, interpretaci nebo zpracování; mezi příklady dat patří posloupnost bitů, tabulka s čísly, znaky na stránce, nahrávka zvuků pořízená mluvčím nebo vzorek měsíční horniny.

Data o správě dat (data management data)

Data vytvořená a uložená v trvalém úložišti funkčního celku správy dat, která se vztahují k provozu archivu; příklady těchto dat jsou data určená k fakturaci koncovým uživatelům a k ověření jejich přístupu, data o pravidlech, data o rámcových objednávkách (o předplatném) pro opakující se požadavky, data o historii procesu uchovávání a statistická data pro vytváření přehledů pro vedení archivu.

Datový objekt (data object)

Fyzický nebo digitální objekt.

Datový objekt s obsahem (content data object) CDO

Datový objekt, který spolu s přidruženými vysvětlujícími informacemi tvoří informační obsah.

Datový slovník (data dictionary)

Formální zdroj termínů používaných k popisu dat.

Digitální objekt (digital object)

Objekt složený z řady posloupností bitů.

Dlouhodobé uchovávání (long term preservation)

Dlouhodobé udržování informací – v podobě, která je určené skupině srozumitelná sama o sobě –, a dokladů o jejich hodnověrnosti.

Dohoda o dodávání dat (submission agreement)

Dohoda uzavřená mezi archivem OAIS a tvůrcem, která stanovuje datový model a další potřebná nastavení pro spojení pro dodávání dat; datový model určuje formát/obsah a pojmy užívané tvůrcem a způsob, jakým jsou vyjádřeny na všech dodaných datových nosičích nebo při všech telekomunikačních spojeních.

Funkční celek archivního uložení (archival storage functional entity)

Funkční celek archivu OAIS, který zahrnuje služby a funkce využívané k ukládání a získávání archivních informačních balíčků.

Funkční celek plánování uchovávání (preservation planning functional entity)

Funkční celek archivu OAIS, který poskytuje služby a funkce pro sledování okolí archivu OAIS a který stanovuje doporučení a plány uchovávání k zajištění dlouhodobé přístupnosti a srozumitelnosti a dostatečné využitelnosti informací uložených v archivu OAIS pro určenou skupinu, a to i v případě zastarání původního počítačového prostředí.

Funkční celek příjmu (ingest functional entity)

Funkční celek archivu OAIS, který zahrnuje služby a funkce, jež od tvůrců přijímají vstupní informační balíčky, připravují archivní informační balíčky určené k uložení a zajišťují, aby archivní informační balíčky a k nim náležející podpůrné popisné informace byly zařazeny do archivu OAIS.

Funkční celek správy (administration functional entity)

Funkční celek archivu OAIS, který zahrnuje služby a funkce potřebné pro řízení běžného provozu ostatních funkčních celků archivu OAIS.

Funkční celek správy dat (data management functional entity)

Funkční celek archivu OAIS, který zahrnuje služby a funkce pro vkládání, údržbu a zpřístupnění různorodých informací; příklady těchto informací jsou katalogy a seznamy položek, které lze získat z funkčního celku archivního uložení, algoritmy, pomocí kterých mohou být zpracována získaná data, statistické údaje o přístupech koncových uživatelů, faktury pro koncové uživatele, rámcové objednávky, bezpečnostní opatření a harmonogramy, pravidla a postupy archivu OAIS.

Funkční celek zpřístupnění (access functional entity)

Funkční celek archivu OAIS, který zahrnuje služby a funkce, jež činí archivované informační jednotky a příbuzné služby dostupné pro koncové uživatele.

Fyzický objekt (physical object)

Objekt (například měsíční hornina, biologický vzorek, mikroskopické sklíčko) s fyzicky pozorovatelnými vlastnostmi, které představují informace, jež je pro účely uchovávání, šíření a samostatného využívání vhodné patřičně zaznamenat.

Hodnověrnost (authenticity)

Míra, do níž osoba (nebo systém) pokládá objekt za ten, za který se tento objekt vydává; hodnověrnost se posuzuje podle důkazů.

Informace (information)

Jakékoliv znalosti, které mohou být předmětem výměny; informace má při výměně podobu dat; příkladem je řetězec bitů (data) doprovázený popisem, jak tyto řetězce bitů převést na čísla představující záznamy teplot změřených ve stupních Celsia (vysvětlující informace).

Informace o identifikátorech (reference information)

Informace, která je využívána jako identifikátor informačního obsahu; patří mezi ně i takové identifikátory, které vnějším systémům umožňují jednoznačně odkazovat na konkrétní informační obsah; informací o identifikátorech je například ISBN.

Informace o neporušenosti (fixity information)

Informace, která udává, jak je zajištěno, aby objekt s informačním obsahem nebyl nezdokumentovaným způsobem změněn; příkladem je kód cyklické redundantní kontroly souboru.

Informace o přístupových právech (access rights information)

Informace, která udává omezení týkající se přístupu k informačnímu obsahu, a to včetně právního rámce, licenčních podmínek a řízení přístupu; informace o přístupových právech zahrnují přístupové podmínky a podmínky šíření uvedené v dohodě o dodávání dat, vztahující se jak k uchovávání prostřednictvím archivu OAIS, tak k vlastnímu využití koncovým uživatelem; upřesňují také, jak uplatňovat opatření pro vymáhání práv.

Informace o původu (provenance information)

Informace, které dokumentují historii informačního obsahu; tyto informace vypovídají o původu nebo zdroji informačního obsahu, o veškerých změnách, které mohly od doby jeho vzniku nastat, a o tom, kdo o něj od doby jeho vzniku pečoval; archiv nese odpovědnost za vytvoření a uchovávání informací o původu od okamžiku příjmu; nicméně informace o původu z dřívější doby by měl poskytnout tvůrce; informace o původu jsou součástí důkazů o hodnověrnosti.

Informace o souvislostech (context information)

Informace, které dokládají vztah informačního obsahu k jeho okolí; patří mezi ně důvod vytvoření informačního obsahu a jeho vztah k dalším objektům s informačním obsahem.

Informace o uchovávání (preservation description information) PDI

Informace, která je nutná k dostatečnému uchování informačního obsahu a která může být rozdělena na informaci o původu, informaci o identifikátorech, informaci o neporušenosti, informaci o souvislostech

a informaci o přístupových právech.

Informace o uspořádání (structure information)

Vysvětlující informace, které udávají, jak jsou další informace složeny; mohou mimo jiné přiřazovat základní typy dat, např. znaky, čísla a pixely, a seskupení těchto typů dat, např. znakové řetězce pole, k tokům bitů.

Informace o významu (semantic information)

Vysvětlující informace, které podrobněji popisují význam nesený informacemi o uspořádání.

Informace o zabalení (packaging information)

Informace, která slouží k propojení a popisu součástí informačního balíčku; může se jednat například o svazkové a adresářové informace podle ISO 9660 užité na CD-ROM k poskytnutí obsahu několika souborů zahrnujících informační obsah a informace o uchovávání.

Informační balíček (information package)

Pojmová schránka, která může obsahovat informační obsah a přidružené informace o uchovávání; k tomuto informačnímu balíčku jsou připojeny informace o zabalení, které vymezují a určují informační obsah, a informace o popisu balíčku, které usnadňují vyhledání informačního obsahu.

Informační objekt (information object)

Datový objekt s vysvětlujícími informacemi.

Informační obsah (content information)

Množina informací, která je určena k uchovávání nebo která obsahuje část těchto informací nebo všechny tyto informace; informační obsah je informační objekt složený z datového objektu s obsahem a z vysvětlujících informací.

Koncový uživatel (consumer)

Úloha vykonávaná osobami nebo klientskými systémy, které využívají služeb archivu OAIS za účelem nalezení a vlastního zpřístupnění uchovávaných informací; tuto úlohu mohou vykonávat další archivy OAIS nebo též osoby nebo systémy z daného archivu OAIS.

Kopírování (replication)

Přesun digitálního obsahu, při němž nedochází k žádným změnám informací o zabalení, informačního obsahu ani PDI; bity užité k vyjádření těchto informačních objektů jsou při přesunu na stejný nebo nový datový nosič uchovány.

Metadata (metadata)

Data o jiných datech.

Místní skupina (local community)

Skupina, která je obsluhována archivem mimo prostředí sdružených archivů.

Nákladové modely pro digitální archivaci (digital preservation costs models)

Modely a nástroje, které pomáhají repozitářům v alokaci a určování výše nákladů na digitální archivaci. Výsledkem nemusí být nutně výše nákladů vyjádřená konkrétní hodnotou, i samotná alokace nákladů na jednotlivé aktivity je validním výstupem takového modelu.

Nevratný převod (non-reversible transformation)

Převod, u něhož nemůže být zaručeno, že se jedná o vratný převod.

Obnova (refreshment)

Přesun digitálního obsahu, jehož výsledkem je náhrada konkrétního datového nosiče jeho dostatečně přesnou kopií provedená tak, aby byla zachována funkčnost veškerého hardwaru a softwaru používaného ve funkčním celku archivního uložení.

Odvozený AIP (derived AIP)

AIP vytvořený výběrem nebo sloučením informací z jednoho nebo více zdrojových AIP.

Ostatní vysvětlující informace (other representation information)

Vysvětlující informace, které nelze jednoznačně zařadit mezi informace o významu nebo informace o uspořádání; například k porozumění datovému objektu s obsahem lze využít software, algoritmy, šifrování, psané pokyny atd., přičemž všechny tyto informace odpovídají definici vysvětlujících informací, byť nebude zřejmé, zda se vztahují k uspořádání nebo k významu; informace udávající vztah mezi informacemi o uspořádání a informacemi o významu nebo softwaru potřebnému pro zpracování databázového souboru jsou také považovány za ostatní vysvětlující informace.

Otevřený archivační informační systém (open archival information system) OAIS

Archiv, který tvoří uskupení lidí a systémů, jež přijalo odpovědnost za uchovávání informací a jejich zpřístupňování určené skupině, přičemž toto uskupení může být součástí většího celku; plní povinnosti stanovené v kapitole 4, což umožňuje odlišit archiv OAIS od jiných archivů; termín „otevřený“ v případě archivu OAIS vyjadřuje skutečnost, že toto doporučení a budoucí související doporučení a normy vznikají v otevřených fórech, a neznamená, že přístup k archivu je neomezený.

Plán nástupnictví (succession plan)

Plán, jak a kdy budou správa, vlastnictví a/nebo řízení jednotek uložených v archivu OAIS přesunuty do následného archivu OAIS, aby byly tyto jednotky i nadále vhodně uchovávány.

Pomůcka pro získávání (retrieval aid)

Aplikace, která oprávněným uživatelům umožňuje získat informační obsah a informace PDI popsané popisem balíčku.

Popis balíčku (package description)

Informace určené pomůckám pro zpřístupnění.

Popis jednotky (unit description)

Popis balíčku, který se zaměřuje na poskytování informací o archivní informační jednotce a je určen pomůckám pro zpřístupnění.

Popis sbírky (collection description)

Popis balíčku, který se zaměřuje na poskytování informací o archivní informační sbírce a je určen pomůckám pro zpřístupnění.

Popisná informace (descriptive information)

Množina informací, která je složena především z popisů balíčků a je poskytována správě dat za účelem podpory koncových uživatelů při objednávání a získávání informačních jednotek z archivu OAIS.

Přebalení (repackaging)

Přesun digitálního obsahu, při němž dochází ke změně informací o zabalení vztahujících se k AIP.

Přesun digitálního obsahu (digital migration)

Přesun digitálních informací v rámci archivu OAIS, jehož záměrem je tyto informace uchovat; od přesunů obecně jej lze odlišit třemi znaky:

– je zaměřen na uchování celého informačního obsahu, který je třeba uchovat,

– nová podoba informací v archivu nahrazuje podobu předchozí,

– archiv OAIS řídí všechny stránky přesunu a nese za ně plnou odpovědnost.

Převáděná vlastnost informace (transformational information property)

Vlastnost informace, u které je uchování její hodnoty pokládáno za nutné, ale ne dostačující, aby bylo možné ověřit, zda při jakémkoliv nevratném převodu byl dostatečně zachován informační obsah; tato vlastnost může podstatným způsobem dokládat hodnověrnost; taková vlastnost informace je závislá na konkrétních vysvětlujících informacích včetně informací o významu udávajících její kódování a význam; (termín „podstatná vlastnost“, který je v odborné literatuře definován různě, je někdy používán způsobem odpovídajícím užití termínu *převáděná vlastnost informace*).

Převod (transformation)

Přesun digitálního obsahu, při němž dochází ke změně informačního obsahu nebo PDI archivního informačního balíčku; převodem je například změna kódování uchovávaného textového dokumentu z ASCII na UNICODE.

Přidružený popis (associated description)

Informace popisující obsah informačního balíčku z hlediska konkrétní pomůcky pro zpřístupnění.

Referenční model (reference model)

Zásady, které udávají, jak v daném prostředí rozumět vztahům mezi objekty a jak vytvářet odpovídající standardy nebo specifikace; referenční model využívá malý počet sjednocujících pojmů a může být využit k výukovým účelům a k vysvětlení standardů laikům v dané oblasti.

Sbírka pro zpřístupnění (access collection)

Sbírka AIP, která je definována popisem sbírky, ale pro kterou ve funkčním celku archivního uložení neexistují informace o zabalení.

Síť vysvětlujících informací (representation network)

Množina vysvětlujících informací, které úplně popisují význam datového objektu; k vysvětlujícím informacím v digitální podobě je třeba přidat další vysvětlující informace, a to z toho důvodu, aby jejich digitální podoba byla dlouhodobě srozumitelná.

Software pro zobrazení vysvětlujících informací (representation rendering software)

Software, který zobrazuje vysvětlující informace vztahující se k informačnímu objektu v podobě srozumitelné člověku.

Software pro zpřístupnění (access software)

Software, který poskytuje část informačního obsahu nebo celý informační obsah informačního objektu v podobě srozumitelné lidem nebo systémům.

Spojení pro dodávání dat (data submissions session)

Dodání datového nosiče nebo telekomunikační spojení, kterým jsou do archivu OAIS poskytována data; formát/obsah spojení pro dodávání dat využívá datový model, který si mezi sebou archiv OAIS a tvůrce vyjednali v dohodě o dodávání dat; tento datový model určuje pojmy používané tvůrcem a způsob, jakým jsou vyjádřeny na všech doručených datových nosičích nebo při všech telekomunikačních spojeních.

Srozumitelný sám o sobě (independently understandable)

Vlastnost informací, které jsou dostatečně úplné, aby jim určená skupina rozuměla a uměla je využít, aniž by bylo nutné se uchýlit ke specializovaným zdrojům, jež nejsou běžně dostupné, a to včetně konkrétních jednotlivců.

Tvůrce (producer)

Úloha vykonávaná osobami nebo klientskými systémy poskytujícími informace určené k uchovávání; může se jednat o další archivy OAIS nebo také o osoby nebo systémy v daném archivu OAIS.

Určená skupina (designated community)

Stanovená skupina možných koncových uživatelů, kteří by měli být schopni porozumět konkrétní množině informací; určená skupina může být složena z více uživatelských skupin; určenou skupinu si vymezuje archiv a její vymezení se v průběhu času může měnit.

Vedení (management)

Úloha vykonávaná těmi, kdo určují celková pravidla archivu OAIS jako součást širších pravidel, například v rámci větší organizace.

Verze AIP (AIP version)

AIP, jehož informační obsah nebo informace o uchovávání byly získány převodem ze zdrojového AIP; verze AIP může nahradit zdrojový AIP; verze AIP se považuje za výsledek přesunu digitálního obsahu.

Vratný převod (reversible transformation)

Převod, při kterém nové vyjádření stanovuje množinu (nebo podmnožinu) výsledných objektů, které jsou rovnocenné výsledným objektům stanoveným původním vyjádřením; to znamená, že novou množinu

objektů je možné k původnímu vyjádření a jeho množině základních objektů přiřadit jedna ku jedné.

Vstupní informační balíček (submission information package) SIP

Informační balíček, který dodává tvůrce do archivu OAIS tak, aby mohl být využit při sestavení nebo aktualizaci jednoho nebo více AIP a/nebo přidružených popisných informací.

Výstupní informační balíček (dissemination information package) DIP

Informační balíček odvozený z jednoho nebo více AIP a zaslaný archivem OAIS koncovému uživateli jako odpověď na jeho požadavek vůči tomuto archivu.

Vysvětlující informace (representation information)

Informace, které k datovému objektu přiřazují významově bohatší pojmy; v případě posloupnosti bitů, která je souborem FITS, může úlohu vysvětlujících informací hrát standard FITS, který popisuje formát, a slovník, který v souboru s klíčovými slovy, jenž není součástí standardu, popisuje význam; dalším příkladem může být software pro zobrazení souboru ve formátu JPEG; vykreslení souboru ve formátu JPEG jako bitů není pro člověka příliš smysluplné, nicméně software, který je schopen pracovat podle standardu pro formát JPEG, k těmto bitům přiřazuje pixely, které potom mohou být zobrazeny jako obrázek určený pro člověka.

Základní služby (common services)

Podpůrné služby nezbytné pro provoz archivu OAIS; patří mezi ně například komunikace mezi procesy, jmenné služby, přidělení dočasného úložiště, ošetření výjimek, bezpečnost a adresářové služby.

Znalostní základna (knowledge base)

Množina informací, které si osvojila osoba, nebo kterou si osvojil systém, a která této osobě nebo tomuto systému umožňuje porozumět přijímaným informacím.

Přílohy

Příloha A – Pokrytí jednotlivých zásad DSA-WDS

Zásada č. 1 – Poslání/Rozsah

Jedná se o výchozí bod činnosti repozitáře. Repozitář musí mít jasno v tom, jaké má poslání a které cíle má naplňovat. Je nutné zde též upozornit na platnost a váhu dokumentu o poslání (mission statement) tím, že repozitář odkáže na autoritu daného poslání (např. zřizovatel, financující organizace apod.). V případě, že repozitář někdy v budoucnu nebude schopný daný mandát naplnit, je třeba stanovit, co se bude dít s daty. Je proto nutné mít vypracovaný tzv. nástupnický plán (succession plan).

Zásada č. 2 – Licence (návaznost na zásadu č. 4)

Repozitář by měl mít jasno v tom, jaké mají digitální objekty, případně sbírky podmínky přístupu. Pokud některá data vyžadují specifický režim, je nutné jej dostatečně popsat. Lze odkázat i na některou ze speciálních licencí typu Creative Commons, nebo připojit konkrétní licenční podmínky (pokud je

repozitář uplatňuje), která pro uživatele platí. Taktéž je nutné upozornit na to, co se děje v případech, kdy stanovené podmínky nejsou dodržovány, a jaká opatření jsou v platnosti, aby se zneužívání podmínek přístupu předcházelo.

Zásada č. 3 – Kontinuita přístupu (návaznost na zásady č. 1, 10 a 14)

Tato zásada se vztahuje na opatření, která mají zajistit přístup a dostupnost dat v současnosti i do budoucna. Spadá sem i existence krizového plánu, který přehlednou formou dokumentuje identifikovaná rizika a popisuje řešení například při haváriích, živelných pohromách, katastrofách, úmyslných útocích (v rámci organizace/repozitáře, ale i zvenčí), i při nedostatku finančních prostředků aj.

Zásada č. 4 – Důvěrnost/Etika (návaznost na zásady č. 2 a 12)

Transparentnost repozitáře je dána nejen dodržováním interně definované dokumentace, ale i jakýchkoliv právních předpisů (na národní i mezinárodní úrovni), standardů (ISO, de facto standardů atd.), smluv, které ovlivňují chod repozitáře. Je vhodné začít popisem samotného právního/organizačního statutu, v rámci kterého hodnocený repozitář existuje. Pokud je možné zveřejnění smlouvy s producenty dat nebo alespoň vzorů uzavíraných smluv, je to vítáno. Dále je důležité popsat, jakým způsobem repozitář postupuje v případě citlivých a osobních údajů (např. zveřejňování smluv, uchovávání logů o přístupech do repozitáře, zpřístupnění dat důvěrného charakteru). Je využita anonymizace dat, nebo je naopak zabezpečen přístup k nim? Jsou zaměstnanci na řešení takových situací dostatečně proškoleni?

Zásada č. 5 – Organizační infrastruktura

Kritérium důvěryhodnosti podle DSA odkazuje na potřebu důkladné dokumentace, pracovních postupů (workflow), rozhodovacích procesů, výběru dat a přístupu k datům. Jelikož lidský faktor je neoddělitelnou součástí repozitáře, je nutné prokázat dostatečný rozpočet, potřebné technologie a skutečnost, že jeho zaměstnanci mají k dispozici průběžné vzdělávání v oboru i zabezpečený profesní růst.

Zásada č. 6 – Odborná pomoc

Pro hodnotitele je též relevantní, jestli repozitář využívá odbornou pomoc, nebo se zapojuje do národních a mezinárodních projektů, a jestli se zajímá o svou cílovou skupinu (získávání zpětné vazby od uživatelů, například pomocí dotazníkových šetření).

Zásada č. 7 – Datová integrita a autenticita (návaznost na zásady č. 8, 9, 10, 12–14)

Toto kritérium důvěryhodného repozitáře se zaměřuje na zachycení „cesty“ digitálního dokumentu od vzniku do zpřístupnění uživatelům, a to z pohledu spolehlivosti originálu (autenticity) a informací o původu (provenience) dat. Proto je důležité popsat, jakým způsobem data vznikají (digitalizací, povinným elektronickým výtiskem apod.), jak je zabezpečená autenticita, integrita a kvalita, jak repozitář reaguje na změnu dat (strategie změny dat), např. v případě jiné verze objektu, změny formátu souboru po migraci, poškození souboru v důsledku chyby na straně hardwaru nebo softwaru apod. Jakým způsobem repozitář odkazuje na metadata; má k dispozici identifikované signifikantní vlastnosti souborů? Jakou formou repozitář kontroluje identitu vkladatele/producenta dat? Využívá repozitář kontrolní součty pro verifikaci dat? Probíhá pravidelné monitorování integrity dat a metadat? Pracuje repozitář s několika verzemi dat? Pokud ano, jaká je strategie verzování objektů?

Zásada č. 8 – Posouzení (návaznost na zásadu č. 11)

Důvěryhodný repozitář si je vědomý základního omezení dlouhodobé ochrany – tj. skutečnosti, že nelze uchovávat vše. Jelikož stávající technologie a datové (souborové) formáty jsou vystaveny riziku zastarávání, je pro repozitář důležité mít přijímané formáty pod maximální možnou kontrolou. Proto uživatelům a producentům dat poskytuje záruky dlouhodobé ochrany u vybraných formátů. Je dobré určit tyto preferované formáty (ideálně neproprietární), u kterých repozitář dokáže zabezpečit co nejdelší zobrazitelnost (renderability) a použitelnost. Takovýto seznam vybraných formátů by v rámci transparentnosti měl být zveřejněný na stránkách instituce, která repozitář spravuje. Je třeba sepsat i způsob, jak repozitář zabezpečuje kontroly kvality, aby producenti dat poskytovali data jen v preferovaných formátech (např. u tzv. master copies, případně user copies). Zároveň repozitář informuje o tom, jak je nakládáno s digitálními objekty, které jsou v jiných než preferovaných formátech (např. zamítnutí importu, absence garance dlouhodobé ochrany apod.). Je vhodné deklarovat, jestli repozitář požaduje po producentech dat podrobnější informace o formátech souborů a případných nástrojích nebo metodách, které byly při vytváření souborů použity.

Zásada č. 9 – Dokumentace postupů uchovávání (návaznost na zásady č. 15 a 16)

Tento okruh certifikace odkazuje na schopnost repozitáře dlouhodobě uchovávat digitální data. Je důležité mít zdokumentovanou strategii dlouhodobé ochrany, způsob bitové ochrany dat (zálohování – offline a/nebo online, vícenásobné kopie), kontrola celistvosti dat (např. kontrolní součty), způsob obnovy dat (zodpovědnost, postup). Repozitář by měl znát rizika, která mu hrozí (v okolí nebo v uvnitř repozitáře). Ideální je využít některou z technik managementu (např. pomocí DRAMBORA nástroje). Důležité je i monitorování nejnovějších i zastaralých technologií a médií.

Zásada č. 10 – Plán dlouhodobé ochrany

Pokud má repozitář k dispozici určitý způsob sledování zastaralých formátů, je nutné vysvětlit, jakou cestou repozitář půjde v případě jejich zastarání – např. migrace, emulace. Dlouhodobou použitelnost ovlivňuje nejen kompaktnost formátu a objektu jako takového, ale i jeho srozumitelnost a využitelnost určenou komunitou (návaznost na zásadu č. 1). Získává repozitář zpětnou vazbu od uživatelů? Ví, jestli jsou prezentovaná data k užitku uživatelům? Pokud ne, jak tuto situaci řeší?

Zásada č. 11 – Kvalita dat (návaznost na zásady č. 7, 8 a 12)

Zde je nutné poukázat na to, jakým způsobem digitální repozitář podporuje producenty dat při odevzdávání metadat, a to nejen popisného typu. V ideálním případě má repozitář alespoň rámcově stanovené požadavky na:

- popisná metadata (informace potřebné pro jednoznačnou identifikaci a vyhledání, případně i objasnění významu daného dokumentu),
- strukturální metadata (popis vztahů mezi jednotlivými prvky skupiny vzájemné souvisejících dat),
- technická metadata (stanovení specifických vlastností souborů – barevný profil, formát souboru apod.),
- administrativní metadata (popis práv duševního vlastnictví, podmínky využití a přístupu apod.).

Zásada č. 12 – Pracovní postupy (návaznost na zásadu č. 4, 8, 9 a 16)

Zde se zjišťuje, do jaké míry repozitář podporuje producenty dat před a během procesu vkládání dat do repozitáře. Lze popsat workflow a postavení producenta nebo producentů dat v rámci tohoto workflow.

Pokud jde například o zpřístupňování "in-house" digitalizace, producentem dat je v takovém případě samotný repozitář. Je vhodné popsat, případně odkázat na metodiku vkládání dat, definovat kompletní balík informací, které mají být v repozitáři uchovány, kupříkladu informace o zdokumentovaném výběru preferovaných a akceptovatelných formátů souborů, o frekvenci importu dat, způsobu importu (dávkový, ruční apod.) preferovaných popisných, technických a dalších typech metadat. Je namístě přidat i důkazy o tom, že sběr nebo tvorba dat vznikla v souladu s etickými a právními normami (například smlouva s firmou zabezpečující externí digitalizaci, smlouva s producenty dat, odkaz na legislativu v případě povinného sběru dat apod.). Je důležité odkázat na producenty dat a jejich přidružené organizace a prokázat, že data jsou výsledkem činnosti daného producenta obsahu (např. instituce nebo konkrétní osoby). Pokud nejsou odkazované dokumenty dostupné v angličtině, je nutné vytvořit jejich stručný popis (obvykle postačí 10 základních vět) a odkaz na dokumentaci v původním znění.

Zásada č. 13 – Vyhledávání a identifikace dat

Repozitář by měl sepsat, jakým způsobem trvale zpřístupňuje a odkazuje na objekty v repozitáři. Nabízí vyhledávací rozhraní, prvky pokročilého vyhledávání, filtry atd.? Zpřístupňuje data ve vícero formátech – např. přes OAI-PMH protokol? Má k dispozici integrované citační nástroje? Jaké persistentní identifikátory přiřazuje objektům – jen interní, nebo i mezinárodní uznávané persistentní identifikátory (Handle, DOI aj.)?

Zásada č. 14 – Opětovné využití dat (návaznost na zásadu č. 2)

U této zásady je důležité popsat, jakým způsobem repozitář zabezpečuje kontrolu kvality dodaných dat. Jsou k dispozici standardní nebo vlastní nástroje kontroly dat jak na straně producentů dat, tak na straně repozitáře? V případě, že kvalita metadat nedosahuje požadovanou úroveň, je třeba informovat, jak repozitář v daném případě postupuje.

Zásada č. 15 – Technická infrastruktura

Je nutné popsat, do jaké míry repozitář a především jeho technická infrastruktura respektují standardy (ISO či další zaužívané standardy). Samozřejmostí by měl být soulad s referenčním rámcem OAIS (ISO 14721:2012). Další standardy mohou být specifické pro technické řešení (např. ISO 27001:2013), a to zejména pro větší až velké repozitářové/archivní řešení. Výhodou je, pokud má repozitář vypracovaný také plán rozvoje infrastruktury. Má repozitář pro své softwarové řešení (komerčního nebo open-source charakteru) dostatečně silnou komunitu, na kterou se může obracet při sdílení vyvíjených funkcionalit nebo v případě problémů?

Zásada č. 16 – Bezpečnost

Repozitář by měl provést analýzu a vyhodnocení rizik, popsat scénáře škod na základě škodících akcí, lidských chyb nebo technických závad, které mohou představovat hrozbu pro jeho data, produkty, služby, či uživatele. Seznam zavedených postupů s opatřeními k identifikovaným rizikům, závadám, škodám apod. je samozřejmostí. Je potřebné rámcově popsat nástroje analýzy rizik (např. DRAMBORA), mít zavedený bezpečnostní IT systém (včetně popisu rolí dotčených zaměstnanců – např. bezpečnostní technici), dále krizový plán a plán kontinuity. „Plusové body" při hodnocení repozitář získává, pokud má vypracovaný plán rozvoje infrastruktury. Standardy typu ČSN ISO/IEC 27001:2014 Informační technologie – Bezpečnostní techniky – Systémy řízení bezpečnosti informací – Požadavky a ČSN ISO/IEC 27002:2005 Informační technologie – Bezpečnostní techniky – Soubor postupů pro opatření bezpečnosti informací bývají zpravidla aplikovatelné pro technická repozitářová/archivní řešení většího až velkého rozsahu.