



VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
BRNO UNIVERSITY OF TECHNOLOGY



FAKULTA PODNIKATELSKÁ
ÚSTAV INFORMATIKY

FACULTY OF BUSINESS AND MANAGEMENT
INSTITUTE OF INFORMATICS

ZÁLOHOVÁNÍ DAT A DATOVÁ ÚLOŽIŠTĚ **DATA BACKUP AND STORAGE**

bakalářská práce

bachelor's thesis

AUTOR PRÁCE

AUTHOR

PETR UHLÍŘ

VEDOUcí PRÁCE

SUPERVISOR

BRNO 2013

Ing. Jiří Kříž, Ph.D.

Abstrakt

Tato Bakalářská práce si klade za cíl objasnit vývoj v zálohování, obnovy a dlouhodobém uchovávání dat jak v obecné úrovni, tak v konkrétním projektu NDK. Dále poukazuje na světové trendy v tomto oboru a zabývá se myšlenkou dlouhodobého úložiště, systémem ukládání a životními cykly tohoto úložiště. Okrajovým způsobem popisuje standardizaci používanou v tomto úložišti.

Klíčové slova

Úložiště, standardy formátů, počítačová síť, validace, extrakce, změna formátů, záloha

Abstract

This thesis aims to clarify developments in backup, recovery and long-term storage of data on a general level and in the specific project NDK. Further notes on global trends in this field and discusses the idea of long-term storage, storage system and life cycles of this repository. Marginal manner the standardization used in this store.

Key words

Storage format standards, computer network, validation, extraction, change formats, backup,

Bibliografická citace :

UHLÍŘ, P. *Zálohování dat a datová úložiště*. Brno: Vysoké učení technické v Brně, Fakulta podnikatelská, 2013. 54 s. Vedoucí bakalářské práce Ing. Jiří Kříž, Ph.D..

Čestné prohlášení

Prohlašuji, že předložená bakalářská práce je původní a zpracoval jsem ji samostatně. Prohlašuji, že citace použitých pramenů je úplná, že jsem v práci neporušil autorská práva (ve smyslu zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským, ve znění pozdějších předpisů).

V Brně, dne 15. května 2013

.....
podpis

Poděkování

Tímto bych chtěl poděkovat vedoucímu mé bakalářské práce Ing. Jiřímu Křížovi, Ph.D. za poskytnutí odborných rad, věcné připomínky, ochotu a vstřícný přístup během zpracování této práce. Mé poděkování patří také mým kolegům z Národní knihovny ČR za ochotné jednání, spolupráci a poskytnutí informací o popisovaném tématu.

Obsah

Úvod	8
1 Cíle práce, metody a postupy zpracování	9
2 Teoretická východiska práce	10
2.1 Co je záloha souboru?	10
2.2 Zálohování a archivace jako součást bezpečnosti IT	12
3 Analýza Současného stavu systému	14
3.1 Projekt NDK má tři hlavní cíle:	16
4 Vlastní návrhy řešení	17
4.1 Přehled základních požadavků na systém LTP	17
4.2 Základní funkční požadavky na systém dlouhodobé ochrany digitálních dat	28
4.3 Standardy pro ukládání do LTP	32
4.4 Standardy formátů	37
4.5 Současný vývoj LTP systému	43
4.6 LTP jako systém	46
5 Závěr	48
6 Seznam použitých zdrojů	49
6.1 Informační zdroje:	49
6.2 Internetové zdroje:	49
6.3 Literatura	50
7 Seznam použitých zkratk	51
8 Seznam obrázků	54

Úvod

V současnosti s rozvojem digitálních technologií vzniká v celém světě obrovské množství digitálních dat. Technologie, které se nyní používají ve výpočetní technice, se neustále vyvíjí a to přímo geometrickou řadou. Tento rozvoj se stále zrychluje. Vznikají nové standardy, nové technologie, nároky na výpočetní techniku. Není to tak dávno, co Internet byl spíše z oblasti sci-fi. S rostoucím objemem, typem dat se zvyšuje i jejich cena. Jednotliví uživatelé a to nejen fyzické osoby ale i firmy a instituce přicházejí ke skutečnosti hodnoty digitálních dat. S tímto trendem bohužel vzniká i kriminalita v oblasti digitálního světa. Oběh peněz a obchody se realizují v digitálním světě. Svět na tuto skutečnost reaguje a vznikají nová odvětví, nové příležitosti podnikání.

Svět se zabývá zásadní otázkou co s digitálními daty. Jak je ochránit, jak je bezpečně uložit – archivovat. Vznikají nové a nové systémy pro zálohu dat v období krátkodobém, ale i dlouhodobém. S vývojem technologií vznikají požadavky na uchovávání některých skutečností, jako národního bohatství. Knihovnické instituce přicházejí do digitálního světa a pořizují digitální kopie veškerého materiálu, který se u nich vyskytuje. Jsou to nejen knihy, ale i obrazové, notové, hudební a video záznamy. Vývoj trendu je takový, že vzniká požadavek na dlouhodobé zálohování rozhlasového a televizního vysílání. Toto je pouze zlomek nepřeberného typu dat, na který je kladen důraz, aby byly zachovány

1 Cíle práce, metody a postupy zpracování

Cílem práce je poukázat na problematiku dlouhodobého uchovávání dat a s tím spojenou problematiku. Dále práce pojednává složitostech spojených s touto problematikou. Tato problematika je poměrně nová a neustále se vyvíjí. Vznikají nová pravidla a standardy, dle kterých je nutno se řídit. Jedná se o velice sofistikovaný systém po sobě jdoucích pravidel, která je doslova nutnost dodržovat. Postupy a standardy jsou použity z praxe, kde vznikaly.

Dále k vypracování práce byl využité standardy normy ČSN ISO 690. Jsou zde využity standardy typu „poznámky pod čarou“ citace atd.

2 Teoretická východiska práce

Základem prací s daty je jejich znova – získání pomocí určitých metod. Záloha a obnova dat. Velice dobrým příkladem je firma Microsoft, která velice dobře definuje tento stav:

2.1 Co je záloha souboru?

Záloha souboru je kopie souboru uložená v umístění odděleném od původního souboru. Pokud chcete sledovat změny souboru, můžete mít vytvořeno několik jeho záloh.

2.1.1 Proč je vhodné soubory zálohovat?

Zálohování souborů pomáhá při jejich ochraně před trvalou ztrátou nebo změnou v případě nechtěného odstranění, útoku červa či viru nebo poruchy softwaru nebo hardwaru. Pokud nastane některá z těchto událostí a soubory jsou zálohované, můžete je snadno obnovit. Informace o zálohování souborů naleznete také v tématu Zálohování souborů.

2.1.2 Které soubory by měly být zálohovány?

Měli byste zálohovat vše, co je obtížně nahraditelné nebo nelze nahradit vůbec, a pravidelně byste měli zálohovat soubory, které často měníte. Obrázky, videonahrávky, hudba, projekty a finanční záznamy jsou příklady souborů, které byste měli zálohovat.

Není nutné zálohovat programy, protože k jejich přeinstalování můžete použít originální disky. Programy navíc obvykle zabírají mnoho prostoru na discích.

2.1.3 Jak často je vhodné soubory zálohovat?

Četnost zálohování je závislá na počtu vámi vytvářených souborů a na tom, jak často je vytváříte. Pokud nové soubory vytváříte každý den, měli byste je zálohovat jednou týdně, případně i denně. Pokud příležitostně vytvoříte velký počet souborů (například při ukládání mnoha digitálních fotografií z narozeninového večírku nebo promoce), zálohujte tyto soubory ihned. Nejvhodnější je naplánovat pravidelné automatické zálohování, abyste na tento proces nemuseli neustále myslet. Můžete si vybrat, zda chcete soubory zálohovat denně, týdně nebo měsíčně. Mezi automatickými zálohami můžete také provádět ruční zálohování.¹

Ukládání digitalizovaných dat bylo řešeno více či méně dle aktuálních trendů. Bohužel neexistovala žádná stálá koncepce ukládání dat. Existovaly pouze časově omezené popisové trendy, které se v rámci různého období měnily. Konvence v názvech nebyla nikde popsána. Data se lišila nejen formátem ale i popisná data byla různého charakteru. V některých případech ani neexistovala. Data se ukládala chaoticky. Část dat byla na klasických discích, část dat byla na páskovém úložišti.

Problémem pak bylo, že konvence popisovaných dat byla různá a na základě fluktuace zaměstnanců se nekonceptně měnila. Existovalo mnoho zdrojů umístění a popisu dat.

Dalším a velice závažným tématem je pak otázka propojení zálohování a bezpečnost.

¹ Citace zdroje: <http://windows.microsoft.com/cs-cz/windows-vista/back-up-and-restore-frequently-asked-questions>

2.2 Zálohování a archivace jako součást bezpečnosti IT

Data a informace jsou obvykle nejcennějším majetkem každé společnosti, bez ohledu na velikost či obor podnikání. Data jsou tím pravým základem know-how představující mnohdy zcela zásadní konkurenční výhodu. Není proto žádným překvapením, že bezpečnost informačních technologií velmi úzce souvisí se zálohováním dat. A přeneseně také s jejich archivací.

2.2.1 Smazal jsem si data!“ a mnoho dalších důvodů

Základem datové bezpečnosti je zajištění dat proti ztrátě, zcizení, poškození i kompromitaci. Vedle klasických bezpečnostních prvků, jako je autorizace přístupu k datům lze jako poslední linii využít i zálohování – ve smyslu zabezpečení dat. Jsou data odcizena? Obnovíme je ze zálohy. Došlo k jejich poškození? Ještě že máme zálohu.

Bezpečnostní pohled doprovází také problém výkonnostní. Mnohé servery jsou v důsledku trvale zrychlujícího se nárůstu objemu ukládaných a zpracovávaných dat přetíženy. Situaci navíc komplikuje i celá řada legislativních opatření, která určují, jak dlouho mají být některá data uchovávána. Výjimkou nejsou ani desítky let. Zde může pomoci archivace.²

² Zdroj citace: <http://ciks.vse.cz/citace/proc.aspx>

Tabulka příkladů popisů dat a s tím spojená rizika

Typ záznamu	Rizika s tím spojená
Papírový záznam v sešitu	možnost zničení záznamu, či nečitelnost
Excelová tabulka	různé verze Office, možnost ztráty dat konverzí
Adresářová struktura různých popisových konvencí	Nutná inventura, některé záznamy neidentifikovatelné
Klasická záloha	bez ověření obnovy, nepopsaná zálohovaná dat

Z klasické zálohy vyplývá, že bylo problematické cokoli dohledat. Pouze se „ukládalo“, ale umělo se i obnovit?

Bylo nutné tuto neutěšenou situaci řešit. Řešením bylo pouze komplexní koncepční projekt, který by vše zohlednil a „srovnal“ a nastavil jednotnou koncepci v ukládání dat, jejich management obnovy a formátové správy.

3 Analýza Současného stavu systému

V Národní knihovně ČR započal v únoru roku 2010 projekt Národní digitální knihovna (NDK), z části profinancovaný z dotací z EU. Projekt NDK využívá systém dlouhodobého ukládání digitálních dat do systému dlouhodobého uložení LTP. Národní knihovna si pro tento projekt zvolila za partnera Moravskou zemskou knihovnu. Projekt byl společně podán v rámci Výzvy 07 Integrovaného operačního programu „Elektronizace služeb veřejné správy“. V červnu téhož roku byl projekt schválen. Jednalo se o jeden ze základních stavebních kamenů konceptu eCulture, kterým sektor kultury velice významně přispěl k naplňování cílů Smart Administration.

Projekt NDK je financován společně ze dvou finančních zdrojů a to z Integrovaného operačního programu EU částkou 255 milionů korun a spolufinancován z rozpočtu MK ČR částkou 45 milionů korun. Příspěvek z 85% ze strukturálního fondu ERDF ve výši 254 946 300 Kč byl doplněn o 15% spolufinancován ze státního rozpočtu ve výši 44 990 700 Kč. Celkové veřejné výdaje tedy činily 299 937 000 Kč.

Národní knihovna a Městská zemská knihovna uchovávají ve svých fondech díky právu úplného povinného výtisku většinu vzniklých monografií, periodik a dalších druhů dokumentů publikovaných na našem území (bohemika v užším slova smyslu), obrovské množství těchto dokumentů vztahujících se k České Republice publikovaných v zahraničí (bohemika v širším slova smyslu) a spravují bohaté historické fondy. Od roku 2000 spolupracují i na „sklizení“ českého webu. Obě instituce proto disponují velice rozsáhlým a zároveň unikátním materiálem jedinečné kulturní, ale s ohledem na kontext Smart Administration především faktografické hodnoty.

Důležitým faktem pak je, že se jedná o největší páskové úložiště ve střední Evropě, rozložené celkem ve třech lokalitách. Tento systém je tak unikátní, že se ho pokusím popsat v této práci. Dalším fenoménem je skutečnost, že se systém

stále vyvíjí a konec tohoto rozvoje se nedá předpokládat. Existují pouze určité milníky vyplývající z projektu, z kterého vzniká. Vznikají nové standardy, které se stávají jednoznačně určujícím faktorem pro ostatní zainteresované strany. Předpokladem pak je systém dlouholetého úložiště s časovou otevřenou hranicí. Existuje pouze jeden předpoklad a to je kontinuální vývoj v čase, zavádění nových trendů v technologiích a udržení systému v provozu za předpokladu 1000 let. To vše je dáno skutečností, že se jedná o určitý typ národní památky. Uložené informace v tomto systému mají posloužit pro budoucí generace. Je to prakticky národní bohatství, které je určeno nejen široké veřejnosti, ale i pro vědeckou činnost.

3.1 Projekt NDK má tři hlavní cíle:

1. Digitalizace významné části národní produkce z 19. do 21. století, tj. knih vydaných na území České republiky, napsaných v češtině nebo pojednávajících o Česku. Celkem do konce roku 2019 je cílem zdigitalizovat více než 50 milionů stran, tedy přibližně 300 000 svazků. Dosah projektu není zdaleka omezen jen na dobu jeho trvání, ale bude dále intenzivně pokračovat i po jeho ukončení v roce 2014 – nejen do roku 2019 v rámci povinné udržitelnosti projektu, ale i v dalších desítkách let.
2. Dlouhodobé uložení těchto digitálních dokumentů ve spolehlivém digitálním úložišti. Úložiště poskytne prostor pro bezpečné umístění dosud digitalizovaných dokumentů i digitálních dokumentů vytvořených či získaných v projektu NDK, v rámci dalších projektů z institucí nacházejících se na území České Republiky.
3. Zpřístupnění takto digitalizovaných dokumentů pokud možno volně zdarma, samozřejmě s ohledem na Autorský zákon

Pro dlouhodobé uložení dat se využívá systém LTP (Long Term Preservation).

4 Vlastní návrhy řešení

4.1 Přehled základních požadavků na systém LTP

4.1.1 Základní popis systému

Tento systém má za úkol zajistit dlouhodobě - trvalé (minimálně po dobu, dokud budou mít uchovávané digitální objekty význam pro uživatele) uchování digitálních nebo digitalizovaných dokumentů v takovém stavu, ve kterém je budou moci uživatelé použít. V závislosti na čase se předpokládá nutná konverze pro možnost zobrazení dat. Uchovávané digitální zdroje musí mít možnost vyhledání, takovým způsobem aby uživatel je mohl pro něj běžném technickém prostředí mít možnost zobrazit tak, jak zamýšlel jejich tvůrce. Uživatel musí být schopen jim porozumět, pochopit jejich obsah a jejich smysl. Dosažením těchto cílů se předpokládá uchování nejen bit-streamu reprezentujícího daný digitální objekt, ale také uchování dalších doprovodných informací, které umožní objekt nalézt, správně technicky zobrazit a uživateli objektu i porozumět. V kontextu modelu systémem OAIS to znamená, že spolu s jednotlivými digitalizovanými objekty musí být uchován nejen informační obsah uchovávaných objektů, ale také další doprovodné informace o původu a historii změn těchto dokumentů, o jejich kontextu a dalších zdrojích potřebných k porozumění (identifikační informace, informace o provenienci, kontextuální informace, informace o úplnosti).

4.1.2 Praktická realizace vstupu materiálu do LTP

4.1.3 První fáze - přijetí digitalizovaného materiálu (Ingest I.):

Příprava dokumentů pro vstup do úložiště bude začínat již na pracovišti masové digitalizace fyzických objektů nebo při sklizení ve WebArchivu. Odtud, přes specializované aplikace, které případně zajistí převod a tvorbu metadat do formátu podporovaného LTP systémem, budou tato data přicházet již se základními popisnými metadaty (popisnými, strukturálními, administrativními, po kontrole kvality obrazu, zpracování OCR a vytvoření uživatelské kopie). Pro pracoviště masové digitalizace a WebArchivu musí na straně LTP systému existovat nezávislé specializované aplikace, které budou sloužit pro zpracování příjmu dat a které budou v podstatě plně automatizované a budou se lišit od ručního vkládání do systému. V budoucnu tak může vzniknout důležitá potřeba archivovat i další typy dat (například video data, zvuková data a data ve 3D), což může vyžadovat zapojení dalších specializovaných aplikací pro vstup dat nebo minimálně přizpůsobivost „deposit modulu“ systému, který by si měl umět poradit s jakýmkoliv typem souboru, funkce by měla být plně automatizována. Systém bude v této fázi provádět kontroly úplnosti dat (například kontrolu kontrolním součtem md5), antivirové kontroly, validaci struktury balíčků (metadata a fyzická data), vytvoření intelektuálních entit, přidělení tzv. URN:NBN:CZ nebo i dalších vnitřních vázaných identifikátorů, validaci veškerých metadat, dále kontrolu záznamu v Registru digitalizace, doplnění metadat, vytvoření balíčku pro další fázi, odeslání do další fáze zpracování aj. V každém okamžiku v případě vzniku náhodného problému bude dokument vrácen buď producentovi, nebo zaměstnanci, který rozhodne o jeho dalším zpracování.

4.1.4 Druhá fáze - přijetí materiálu (Ingest II.) :

Druhá postupná fáze vstupu do úložiště by měla být společná pro veškeré typy vkládaných materiálů a pro všechny dodavatele, dle vytvořených standardů. Zde půjde především o provedení plně automatické identifikace a validace veškerých formátů souborů, (obohacení metadat plně automatickou cestou), případně migraci již neplatných standardů do určených standardů - preferovaných formátů (normalizace), opětovou zvlášť důkladnou antivirovou kontrolu. Tento vstupní proces bude probíhat zapojením specializovaných služeb (jako jsou JHOVE, PRONOM/DROID) případně s využitím NZME (New Zealand Metadata Extactor). Průběh pohybu takto zkontrolovaného vstupního balíčku by měl být vidět v každém okamžiku na monitorovacím modulu systému (modul grafického prostředí systému GUI), kde bude informace dostupná správci úložiště i dodavateli dat a tato data se budou dále plně archivovat. V druhé fázi přijetí materiálu, kde jsou stále ještě uživatelské kopie v LTP systému (na pracovním prostoru), po uložení archivních dat do Archivu, se budou předávat do aplikací zpřístupnění a v LTP systému nezůstávají.

4.1.5 Správa archivního modulu

Jedná se o ústřední jádro celého systému dlouhodobé ochrany dat. Základem je systém archivace (archival storage). Je vlastní technologie uchovávání digitálních dokumentů na fyzických úložištích (převážně na páskových jednotkách). Současné metody archivace mají řadu využitelných funkcionalit, které podporují dlouhodobou archivaci. Ačkoli využívají velice levná úložná média, díky pokročilému zdvojení (cluster), inteligentnímu nakládání s daty, monitorování a dalším specifickým funkcím jsou z hlediska dlouhodobé ochrany velice vhodné. Nad základní archivaci musí mít systém komplexní data, metadata management a administrativní rozhraní. To vše bezpodmínečně s odpovídajícím grafickým rozhraním.

4.1.6 Plánování ochrany

Tento systém by měl monitorovat základní vlastnosti (za pomoci externích služeb jako je systém PRONOM) a metadata veškerého vkládaného materiálu a zároveň inteligentně pomáhat správcům úložiště s plánováním dlouhodobé ochrany (musí uchovat a zajistit veškeré informace o vložených formátech a platformách, na kterých fungují, dále o použitých metodách komprese a dalších souvisejících použitých technologiích, které mohou mít potenciálně dopad na možnost použití archivovaného materiálu). Tento systém musí umožňovat, aby oprávněný správce úložiště mohl opětně vy-exportovat z daného archivu různě předem definované množství objektů pro účely například migrace dat nebo jiné ochranné akce mimo úložiště (ve fázi re-testování migračních nástrojů), případně obsahovat základní nástroje pro migraci dat. Další možností je, že tento systém umožní přímo provádět hromadné ochranné akce použitím externích nástrojů. Další vývoj v oblasti dlouhodobé ochrany zcela jistě přinese nové praktické nástroje, jež provádění ochranných akcí podpoří nebo usnadní (typově evropské projekty jako je např. PLANETS a jejich propracovaný nástroj PLATO). Z tohoto důvodu musí být systém plně otevřený, aby tyto nástroje bylo možné v budoucnu plně využívat.

4.1.7 Přístup

Systém úložiště bude muset umožňovat precizní vyhledávání a předání ke zkoumání archivovaných dokumentů a jejich metadat v různě volitelně strukturovaných balíčcích „DIP“ při dodržení základních přístupových práv.

4.1.8 Administrace a monitorování

Systém musí mít propracovaný modul pro veškeré monitorování dění v systému. Musí podporovat detailní sledování pohybu dokumentů v jednotlivých fázích životního cyklu balíčků a tyto informace pečlivě uchovávat k dalšímu možnému zpracování. Dále musí také umět monitorovat a zaznamenávat všechny prováděné akce s použitými formáty a softwary. Musí podporovat monitoring distribuce

dokumentů z úložiště. Bylo by velice vhodné, aby monitoring bylo možné použít v obchodním modelu úložiště pro další využití, tj. sledovat náklady na skladování dokumentů, vykazovat náklady spojené s provozem jednotlivým skupinám uživatelů a dodavatelů dat. Administrace musí být přizpůsobivá, dále musí umožnit nastavení přístupů do systémů, management práv k určitým akcím a nastavení systému samotného, to vše za použití grafického rozhraní.

4.1.9 Integrace s participujícími systémy

- Knihovní katalogy
- Vazba na Resolver URN:NBN což je výměna metadat s resolverem/generátorem URN:NBN, (data z masové digitalizace budou mít URN:NBN vytvořená v procesu digitalizace např. NDK, LTP systém je musí důkladně zkontrolovat v Resolveru, případně se musí vzájemně obohatit. Pro data z ostatních zdrojů by měl být schopen URN:NBN přidělit, nebo tento proces vyvolat)
- Vazba na katalogizační modul ILS – kontrola záznamu v katalogu, případně zpuštění procesu jeho vytvoření při vstup dat (pro data, která nepřícházejí z masové digitalizace, ale z jiných zdrojů)
- Přímá vazba na producenta, dodavatele dat (především automatizované přijímání dat z masové digitalizace v rámci projektu, kromě toho otevřenost dalším potenciálním dodavatelům dat)
- Ochrana - nástroj pro plánování (PLATO - preservation planning tool)
- Autentizace (pro administrátora, pro koncového uživatele přístup nebude)
- Přístupové aplikace (např. Kramerius)
- Oai-pmh propojení (systém Europeana)
- Aplikace pro konverzi, transformaci metadat a uživatelských kopií
- JHOVE, DROID
- Registry formátů (UDFR, PRONOM)
- Softwarový nástroje na migraci dat

- Audit a certifikace
- Registr Digitalizace.cz

4.1.10 Přehled požadavků na systém pro dlouhodobou ochranu digitálních dat v projektu NDK

Systém dlouhodobé ochrany digitálních dat vznikající v rámci projektu NDK musí být založen na referenčním rámci OAIS, které musí obsahovat důsledné plánování ochranných akcí, jejich přesnost provedení a hlavně musí zajišťovat precizní práci s metadaty, která jsou klíčovými aspekty při dlouhodobém uchovávání.

Dalšími standardy, které musí nový systém LTP podporovat, jsou (mimo preferované OAIS):

- Musí mít schopnost propojení certifikací (TRAC, DRAMBORA)
- Musí mít možnost standardizovaného a jednoduchého propojení se systémy jakékoliv nové vzniklých služeb třetích stran
- Metadatové formáty PREMIS, METS, MODS, MIX, které jsou celosvětovými standardy pro digitální objekty
- Musí podporovat služby třetích stran (jako jsou PRONOM, UDFR, Jhove2 apod.)

4.1.11 Základní požadavky:

- Systém musí odpovídat koncepčnímu modelu OAIS
- Informace a digitální dokumenty uložené v systému musí být uloženy na velmi dlouhou dobu, až desítky let a to permanentně, po celou dobu musí být zachována jejich okamžitá použitelnost, čitelnost a srozumitelnost pro uživatele
- Systém musí být připraven pro řádnou certifikaci „důvěryhodného digitálního úložiště“ (TRAC)

- Systém musí odpovídat aktuálním světovým standardům v oboru dlouhodobé ochrany digitálních dat a používat aktuální nástroje pro validaci a charakterizaci používaných formátů
- LTP systém musí obsahovat dostatečné informace o uchovávaných digitálních objektech tak, aby se bylo možné v úložišti vyhledat a aby bylo možné prokázat jejich přesný původ a z dokumentovat veškeré změny na nich provedené
- Hardware a Software systému musí odpovídat existujícím standardům v dané oblasti
- Kontrola všech přístupu
- Integrace s existujícími systémy Národní knihovny a Městské zemské knihovny
- Řízení vzájemných vztahů se všemi dodavateli dat, i mimo rámec IOP projektu
- Vkládání digitálního materiálu (z různých zdrojů)
- LTP systém musí být schopen přijmout data, např. dokumenty v jakémkoli formátu. Je věcí smlouvy s dodavatelem, jak se s daty bude pak nakládat. V některých těchto případech budou data normalizována do přijatelnějších standardních formátů. V jiných případech, pokud to dodavatel nebo formát dat neumožní, bude garantováno pouze uložení a zpřístupnění bitové kopie „bit-stream preservation“.
- Možnost dále rozšiřovat specifikaci metadatových formátů
- Připravenosti pro vstup nepublikovaného materiálu (e-deposit) a potřebné napojení s katalogizací
- Monitorování veškerého digitálního materiálu
- Implementace nových strategií do dlouhodobé ochrany
- Archivní modul musí vždy udržovat minimálně dvě kopie dat na dvou geograficky oddělených lokalitách (v Praze a Brně)
- Monitorování a podpora technologické infrastruktury
- Zpřístupnění veškerého obsahu archivu
- Role základní autorizace
- Protokoly přístupů (FTP, OAI-PMH, sFTP, FTPS, SWORD, SRU/SRW)

- Časové plánování přijímaných dat
- Nutná funkcionality i vložení z fyzického přenosného nosiče (CD, DVD, externí HDD, DAT, BlueRay a dalších)
- Musí být možné přeskočit některé fáze přijímání dat
- Možnost vrácení daného scénáře (Roll back scenario) musí být dostupné pro různé moduly (hlavně ovšem pro přijímání dat)
- Požadavek zvládnutí na vstupu až desítky tisíc dokumentů za den
- Grafické rozhraní (GUI) na všechny moduly

4.1.12 Obecné vlastnosti systému na dlouhodobou ochranu digitálních dat (LTP systému)

- Systém musí být otevřený a vysoce interoperabilní³, tj. musí umožňovat odpojení a připojení různých částí systému bez omezení plné funkcionality jiných, musí umožňovat jednoduchou a otevřenou integraci použitých externích nástrojů (přes rozhraní API a SDK), dále jejich využití v pracovních postupech systému pro možnou a případnou migraci, emulaci, a validaci formátů, ale i pro spolupráci s katalogem, depositem, modulu zpřístupnění, vyhledávacích nástrojů atd.
- Systém musí být velice kvalitně a podrobně zdokumentován
- Systém musí být dále rozšířitelný, tj. musí mít možnost se pružně měnit na základě parametrů systému, jak z hlediska celkové ukládací datové kapacity, tak z hlediska počtu a typů uložených objektů, jejich velikosti a z hlediska datové prostupnosti kritických míst systému
- Systém musí být v mnoha ohledech nastavitelný (Národní digitální knihovna bude například i přes relativní homogenitu materiálů přicházejícího z masové digitalizace potřebovat různé definice informačních balíčků (LTP) SIP, AIP a DIP. Systém musí být dále schopen je zpracovávat paralelně, musí umožňovat snadnou integraci nových formátů a nových projektů)
- Systém musí mít pravidly řízené pracovní postupy
- Řešení musí být zcela nezávislé (universální) na použitých typech Hardware ani na Softwarové platformě (architektura řešení nesmí být svázána s jedním určitým druhem archivního úložiště tj. s jedním výrobcem ...)
- Zajištění trvalého rozvoje systému, migrace na nově typová média, dále využívání nově vzniklých formátů, nového softwaru pro hodnocení a sledování obsahu archivu atd.
- Řešení musí být, pokud možno schopno se plně přizpůsobit automatizovanému, ve smyslu zpracování hromadných dodávek

³ Schopnost různých systémů vzájemně spolupracovat, poskytovat si služby

dokumentů, dále musí být i automatická distribuce a konverze), ovšem při maximálním zachování možnosti manuálního zpracování nastavení

- Nutnost dostupnosti zdrojového kódu v případě potřeby
- Systém musí mít maximálně propracovaný modul pro monitorování a sledování akcí probíhajících na vstupu do úložiště, dále při archivaci, distribuci digitálních dat, správě a administraci
- Systém musí mít naplánován svůj vývoj, který bude nezávislý na zákazníkovi
- Systém by měl umožňovat provádět veškeré ochranné akce hromadně na administrátorem předdefinovaných skupinách objektů. Řešení musí podporovat řízení práv dostupnosti, podporovat systémy typu Onelog, Shibboleth, udržovat aktuální databázi uživatelů i dodavatelů dat, zaznamenávat jejich aktivity v systému
- Systém musí podporovat možnost zapojení nástrojů třetích stran pro plánování dlouhodobé ochrany, realizaci hromadných ochranných aktivit a hodnocení jejich úspěšnosti
- Nutná a nezbytná podmínka je reference na běžící systém s digitalizovanými daty v něm provozované kdekoli na světě.

4.2 Základní funkční požadavky na systém dlouhodobé ochrany digitálních dat

Vložení materiálu - Ingest	
<i>Řízený tok vstupujících dat (podle různých typů sbírek)</i>	<i>Systém poskytuje služby pro kontrolu toku vstupujících dat, s možností nastavení paralelních datových toků pro různé sbírky: např. pro data z digitalizace, z webharvestingu, z povinného výtisku, pro zvukový materiál atd.</i>
<i>Vytvoření vstupního balíčku SIP</i>	<i>Systém umožňuje producentům dat vytvořit vstupní balíček z jakýchkoli digitálních dat s metadaty.</i>
<i>Mechanismus pro předání vstupního balíčku</i>	<i>Systém umožňuje vkládání digitálního materiálu pomocí standardních komunikačních protokolů (např. FTP, HTTP, sdílená síť, atp.) nebo z fyzických elektronických médií (např. CD-R, externí HDD aj.)</i>
<i>Dávkové nebo manuální vkládání</i>	<i>Systém je schopen přijímat vstupující materiál jednotlivě nebo dávkově</i>
<i>Metadata vstupujícího balíčku</i>	<i>Systém je schopen získat data o vstupujícím materiálu od producenta/depozitora, např. popisná metadata, data specifikující práva k zpřístupnění. Systém používá hash ke kontrole integrity dat.</i>
<i>Oznámení o výsledku vstupu</i>	<i>Systém automaticky generuje oznámení o tom, zda vstup materiálu do úložiště proběhl úspěšně nebo ne, toto oznámení zaznamená a pošle producentovi a administrátorovi.</i>
<i>Identifikace producenta</i>	<i>Systém je schopen v okamžiku vkládání dat identifikovat producenta/vkladatele, a využít nastavení, které je pro daného producenta v systému uloženo pro řízení vstupních operací.</i>
<i>Identifikace a sledování PSP⁴/SIP</i>	<i>Systém zajistí, že každý PSP a konečný SIP dostane při vstupu jedinečný identifikátor.</i>
<i>Zajištění integrity PSP</i>	<i>Pro každý soubor v PSP musí systém sledovat hash.</i>
<i>Předvyplnění vstupních metadat</i>	<i>Producent/dodavatel si může na vstupu vybrat z přednastavených šablon s předem vyplněnými metadaty.</i>
Administrace	
<i>Management producentských účtů a nastavení vstupního procesu</i>	<i>V systému je možné vytvářet, udržovat a rušit účty producentů. Účty musí obsahovat kontaktní informace a ujednání o způsobech vkládání materiálů do systému, o zpřístupnění, případně další informace a nastavení.</i>

⁴ PSP – Producent Submission Package – data, která se do LTP systému zasílají a až v LTP systému dojde ke konverzi na SIP balíček (Submission Information Package)

<i>Zaznamenávání aktivity producentů</i>	<i>Systém monitoruje a zaznamenává aktivity producentů, kontroluje, zda jsou v souladu s příslušnými ujednáními, umožňuje prohlížet seznamy vloženého materiálu od jednotlivých producentů atd.</i>
<i>Workflow systému je pružně nastavitelné, automatizované, založené na pravidlech</i>	<i>Systém používá pravidla pro řízení automatického workflow. Workflow je možné měnit (např. přeskokovat některé kroky) a mělo by být možné vrátit se ve workflow o krok zpět (roll back).</i>
<i>Lokální knihovna formátů</i>	<i>Systém udržuje knihovnu formátů, které do něj byly vloženy, a podporuje jejich identifikaci, charakterizaci a validaci.</i>
Management intelektuálních entit	
<i>Jedinečný identifikátor pro každou intelektuální entitu</i>	<i>Systém zajistí, že každá skladovaná logická entita a digitální objekt mají jedinečný trvalý vnitřní identifikátor od okamžiku vstupu do Ingestu (LTP systému).</i>
<i>Automatická kontrola kvality souborů a metadat.</i>	<i>V systému lze nastavit pravidelné a automatizované kontroly kvality dat v archivu. Tyto kontroly mohou obsahovat antivirovou kontrolu (prováděnou v oddělené zabezpečené lokalitě) validaci formátů a kontroly integrity souborů (check-sum).</i>
<i>Rozpoznání formátů souborů a jejich charakterizace</i>	<i>Systém je schopen identifikovat, charakterizovat a validovat formáty souborů a kontroluje, zda je vstupující formát v souladu s nastavenými pravidly workflow.</i>
<i>Bezpečný karanténní úložný prostor</i>	<i>Systém obsahuje oddělenou a zabezpečenou oblast pro skladování digitálního materiálu, který neprojde automatickými kontrolami a poskytuje automatické oznámení správci systému pro další zpracování.</i>
<i>Automatická extrakce metadat z vložených souborů</i>	<i>Systém automaticky získává metadata ze souborů podle nastavených pravidel a spolupracuje s externími službami (pokud je třeba).</i>
<i>Systém umožňuje manuální výběr vložených digitálních objektů</i>	<i>Systém umožňuje lidský zásah za účelem selekce a kontroly obsahu digitálního materiálu. Manuální selekce nebo kontrola digitálních dat je podporována systémem založeným na pravidlech, např. v případech, kdy je digitální materiál na vstupu do úložiště z nestandardního zdroje.</i>
<i>Systém umožňuje uspořádání a popis digitálních objektů</i>	<i>Systém poskytuje možnost manuálně třídit a popisovat vkládaný materiál před tím, než je vložen do archivu.</i>
<i>Mazání souborů, které nebudou součástí SIPu</i>	<i>Systém umožňuje mazání souborů z balíčku PSP, které byly na základě manuálního výběru ze SIPu identifikovány jako odmítnuté. Systém uchovává metadata spojená se soubory vymazanými ze SIPu.</i>

<i>Vyhledávání v repozitáři</i>	<i>Systém poskytuje mechanismy podporující několik metod použití metadat spojených s entitami a jejich reprezentacemi. Umožňuje vyhledávat reprezentace uložené v archivu a poskytuje výsledky vyhledávání ke stažení apod. Archivovaná data by měla být dostupná i bez systému pro dlouhodobou ochranu.</i>
<i>Uchovávání derivátů vložených souborů</i>	<i>Systém umožňuje vložení jiných verzí uchovávaných digitálních objektů v jiných přijatelných formátech a poskytuje nástroje jak spojit tyto deriváty se SIP nebo archivními reprezentacemi, a tato spojení zobrazit, tj. udržuje více verzí jednoho dokumentu vzniklých např. v průběhu let.</i>
<i>Extrakce reprezentace pro manuální údržbu</i>	<i>Systém je schopen komunikovat s nástroji pro autentizaci a poskytuje pravidly řízené workflow pro manuální vytváření nových verzí archivovaných objektů nebo pro mazání reprezentací souborů a/nebo jejich metadat.</i>
<i>Vytvoření dodatečné reprezentace</i>	<i>Systém poskytuje mechanismus pro vytvoření nových reprezentací z vybraných souborů (jako součásti existující intelektuální entity v archivu).</i>
<i>Příprava a vložení metadat (pro jednotlivé soubory nebo pro balíky)</i>	<i>Systém podporuje manuální i automatické vytváření metadat pro jak individuální tak dávkové zpracování.</i>
<i>Aplikace softwarových nástrojů a postupů pro dlouhodobou ochranu</i>	<i>Systém má otevřená API a podporuje integraci externích ochranných nástrojů jako součásti reservačních workflow. Systém umožňuje definovat ochranné workflow a postupy, jejich autorizaci a provedení na nějak vybraných/definovaných souborech reprezentací/digitálních objektů. Nástroje na ochranné akce jsou ideálně součástí LTP systému, za předpokladu zachování možnosti připojit nástroje externí. LTP systém obsahuje modul na plánování a testování ochranných operací.</i>
Plánování dlouhodobé ochrany	
<i>Výběr a vytvoření identických kopií pro účely testování ochranných aktivit</i>	<i>Systém poskytuje mechanismus pro výběr a kopírování reprezentací ven z archivu do odděleného prostředí, ve kterém budou testovány nástroje a procedury pro testování akcí dlouhodobé ochrany. Systém udržuje logy o provedení těchto kopií a jejich exportu.</i>
<i>Identifikace a oznámení ohrožených reprezentací</i>	<i>Systém má mechanismus pro hledání v metadatech objektů a generování reportů o identifikovaných ohrožených datech/formátech.</i>

<i>Výběr souboru reprezentací pro dlouhodobou ochranu</i>	<i>Systém využívá selekční kritéria, která definuje autorizovaný uživatel, a využívá informace z metadat pro vytvoření seznamu úkolů, který definuje různé skupiny reprezentací určených pro konkrétní ochrannou akci.</i>
<i>Hodnocení výsledku ochranných akcí</i>	<i>Systém vybírá a vypíše na parametrech založený soubor aktualizovaných reprezentací, na kterých byla provedena nějaká konkrétní reservační akce, pro hodnocení konkrétním člověkem.</i>
Management archivu	
<i>Management metadat objektů v archivu, která jsou v jiných knihovních systémech</i>	<i>Systém specifikuje, které systémy jsou odpovědné za skladování částí metadat uložených reprezentací</i>
<i>Management lokací souborů v archivu</i>	<i>Systém určuje lokaci reprezentací a jednotlivých souborů v archivu pro účely uložení a přemístění (pokud je to nutné).</i>
<i>Management procesů archivního skladu</i>	<i>Systém poskytuje mechanismus pro management reprezentací v archivním skladu. To znamená, že je schopen uložit, aktualizovat, vyložit data nebo metadata spojená s konkrétní reprezentací (bez nutnosti exportovat celý archivní balíček ven z archivního systému a opětný re-ingest)</i>
<i>Automatické zajišťování kvality souborů a metadat skladovaných objektů</i>	<i>Systém umožňuje pravidelnou a automatickou kontrolu pro zajištění integrity a kvality archivovaného digitálního materiálu, podle nastavených pravidel, která je schopen interpretovat a použít.</i>
<i>Komplexní metadata pro kontrolu přístupu ke skladovaným digitálním objektům</i>	<i>Systém drží, aktualizuje a aplikuje pravidla pro řízení práv a podmínek dostupnosti k jednotlivým reprezentacím jako součást metadat.</i>
Přístup	
<i>Využití existujících nástrojů pro vyhledávání zdrojů</i>	<i>Systém musí spolupracovat s existujícími nástroji NK ČR a MZK pro vyhledávání, musí být schopen zpracovat jejich požadavky na vyhledání digitálního materiálu a metadat. Komplexní vyhledávací mechanismus by měl mít i vlastní samostatný archivní modul LTP systému.</i>
<i>Data pro kontrolu dostupnosti a podmínky použití</i>	<i>Systém drží data o kontrole dostupnosti a podmínkách použití digitálního materiálu pro podporu interpretace práv dostupnosti systémy pro vyhledávání. Systém musí být otevřený jiným systémům pro kontrolu dostupnosti.⁵</i>

⁵ Zdroj citace: www.nkp.cz

4.3 Standardy pro ukládání do LTP

Každá uložená entita v LTP musí obsahovat balíček PSP

(PSP balíček – producer submission package)

- balíček dat a metadat, který přichází od dodavatele dat
- PSP balíček musí obsahovat kompletní intelektuální entitu
- vytváření souborů JPEG2000 se bude používat standardně Software Kakadu
- veškerá metadata musí pro zápis používat kódování UTF-8

Souhrnná tabulka balíčku SIP⁶

Složka	obsahuje	obsahuje
číslo entity	Info.xml	
	masterCopy (složka)	obrazy JPEG2000 lossless
	userCopy	obrazy JPEG2000 lossy
	ALTO(xml)	soubory ALTO.xml
	TXT(složka)	soubory OCR.txt
	amdSec(složka)	AMD_Mets
	hlavní_METZ.xml	
	soubor.md5	

4.3.1 Detailní popis jednotlivých komponent balíčku

4.3.2 Soubor info.xml

Tento popisný soubor formátu xml obsahuje velmi stručné záznamy - údaje o vzniku celého PSP balíčku – kdo, kdy ho vytvořil, jakou měl původně velikost, odkud a kam byl kdy nakopírován atd. Obsahuje informaci o stavu zpracování balíčku. Zaznamenány mohou být údaje o obsahu PSP balíčku – počet a názvy souborů atd. Soubor obsahuje odkaz na soubor s MD5 součtem. Tento soubor info.xml by také mohl být vedle hlavního PSP balíčku.

⁶ Zdroj NK ČR

4.3.3 Údaje a struktura info.xml souboru:

- Vznik balíčku – datum dle normy ISO8601 na úrovni vteřin
- ID balíčku – je použit identifikátor čísla periodika (URN:NBN3. ID titulu - čČNB, ISBN nebo ISSN (opakovatelné))
- Údaje o větším celku (daného projektu), do kterého balíček patří - např. digitalizace pro ANL, ANL+ atd.
- Název instituce, která je zadavatelem pro digitalizace
- Velikost balíčku – v kB
- Tvůrce balíčku – kód instituce (firmy), která balíček vytvořila
- Názvy souborů včetně přesné cesty uložení a koncovky
- Obsah balíčku
- Počet souborů v balíčku celkem
- Odkaz na soubor s MD5 a jeho MD5
- Poznámka – např. o tom, že balíček neobsahuje jednu z povinných komponent (TXT, PDF, JPG2 atd.).

4.3.4 Složka [masterCopy]

Složka obsahuje originální kopie. Obsahuje soubory JPEG2000 v neztrátové kompresi, 1 soubor = 1 stránka, tj. obsahuje všechny naskenované stránky čísla periodika.

4.3.5 Složka [userCopy]

Složka obsahuje uživatelské kopie, Pro každou naskenovanou stránku čísla periodika obsahuje jeden JPEG2000 soubor se ztrátovou kompresí.

4.3.6 Složka [ALTO]

Obsahuje ke každé stránce 1 ALTO XML soubor, tj. tolik ALTO XML souborů kolik je stránek čísla Periodika

4.3.7 Složka [TXT]

Obsahuje ke každé stránce 1 OCR soubor jako čistý text v editovatelné podobě.

4.3.8 Složka [amdSec]

Složka s technickými metadaty – obsahuje pro každou naskenovanou stránku čísla časopisu 1 METS soubor (AMD_METS.xml). Záměrně nejsou tato metadata v hlavním METS záznamu, protože tím by neúměrně narostl počet a velikost. Bylo by velice obtížné s ním pracovat a vznikl by neúměrný nárok na kapacitu úložiště. Je nutné, aby byla z něj přímá adresa na zdroj. Každý METS soubor AMD_METS.xml obsahuje následující části :

- **amdSec** (administrativní metadata) obsahuje části:

technická metadata techMD, která:

ve formátu **PREMISobject** popisují vlastnosti:

- archivní kopie
- ALTO XML
- původního TIFF souboru, ze kterého vznikly archivní kopie

ve formátu **MIX** popisují vlastnosti:

metadat o provenienci digitálních objektů digiprovMD – obsahuje části

PREMISevent

PREMISagent.

- archivní kopie
- původního TIFF souboru

structMap – pouze fyzická strukturální mapa, to je povinná část METS záznamu.

Poukazovat na strukturu souborů k dané stránce, tj. opět archivní kopie, ALTO XML a OCR TXT. Pro další mapování do LTP systému nebude potřeba.

ofileSec – je sekce s odkazy na soubory, je to povinná část METS záznamu - v případě tohoto METS záznamu pro jednu stránku, který vzniká primárně k zachycení technických a administrativních metadat bude odkazovat na soubory, které jsou s tou konkrétní stránkou spojeny, tj. archivní kopie, OCR TXT a ALTO XML.

4.3.9 Soubor Hlavní_METS.xml

Další částí PSP balíčku je hlavní METS dokument. Hlavní METS záznam obsahuje:

- **fileSec** – hlavní část linky na všechny digitální objekty (archivní kopie, ALTO XML a OCR TXT), které se váží k jednomu svazku monografie. Obsahuje také linky na administrativní metadata AMD_METS.xml do složky [amdSec].
- **dmdSec** – bibliografická metadata k svazku monografie včetně popisu nadřazených entit (např. titul) nebo naopak částí (např. kapitola). Základ bude z katalogu, případný další popis částí bude z digitalizace. Formátem hlavním bude MODS, nutná pro LTP je i přítomnost zkráceného záznam v Dublin Core.
- **structMap** – strukturální mapa pro celý dokument, tj. pro jeden svazek monografie. Obsahuje:
 - logickou část – vyjadřuje logickou strukturu svazku s odkazy na ALTO XML
 - fyzickou část obsahující informace o všech reprezentacích konkrétní stránky (archivní kopie, ALTO XML, OCR TXT a AMD_METS.xml) o mapování na ALTO XML areas
- **structLink** – výčet stran jednotlivých úrovní periodika na základě přidání vazeb mezi logickou a fyzickou strukturou

4.3.10 Soubor MD5

Poslední částí PSP balíčku je tzv. soubor s kontrolními součty pro všechny soubory balíčku (kromě info.xml a .md5 souboru samotného). Soubor md5 je vždy jeden pro 1 celý balíček SPS (balíček se svazkem monografie). Tento soubor md5 obsahuje kontrolní součet pro každý soubor obsažený v PSP balíčku. Z tohoto důvodu nejsou samostatné kontrolní součty součástí podsložek balíčku. Kontrolní součty jsou obsaženy také samozřejmě v technických metadatech.

4.4 Standardy formátů

Standardy využívané v rámci celého projektu podléhají neustálému vývoji. Snaží se reagovat na existující moderní trendy. Filozofie těchto standardů je založená na možnosti v rámci časové doby kdykoli získávat data z úložiště LTP, v takové konfiguraci aby je bylo možno prezentovat. Z dlouhodobého úložiště LTP by měl posléze vzniknout archiv národního kulturního bohatství, které by mělo obsahovat studnici vědění celého národa. LTP by mělo obsahovat nejen bibliografie a periodika, ale i notové záznamy, mapy, rozhlasové a televizní vysílání, reklamu a internetový harvesting. Systém LTP by dále měl být flexibilní k různým formátům dat a to takovým způsobem, že se sám dokáže „učit“ a tím i vytvářet specifikace pro ukládání dat. Proto je velice důležitý standard ukládaných dat. Jak jsem již výše zmínil, budeme se nyní bavit o následujících standardech.

- Standard pro metadatové formáty digitalizaci
 - Periodika
 - Monografie
- Standard pro obrazové formáty

4.4.1 Definice metadatových formátů pro digitalizaci

4.4.2 Periodika

Granularita metadatového záznamu

- základní intelektuální entitou periodik je 1 číslo
- každé číslo periodika má svůj vlastní metadatový záznam (=METS)
- tento METS záznam obsahuje údaje o nadřazených entitách čísla, jako jsou ročník, titul periodika, tj. je pro uživatele i pro systém možné spojit jednotlivá čísla do ročníků a titulů

4.4.3 Identifikátory

Identifikátory pro úroveň titulu:

- UUID (vygeneruje dodavatel) - povinné
- Číslo České národní bibliografie – čČNB (povinné, pokud je možné je záznamu přidělit)
- URN:NBN (je přiděleno aplikací „Resolver URN:NBN“) – povinné, pokud lze přidělit
- ISSN (pokud je možné je záznamu přidělit)
- ISBN (pokud je možné je záznamu přidělit)
- Jiný identifikátor – lze využít různé identifikátory, které nějakým způsobem propojí digitální záznam s fyzickou jednotkou – např. čárový kód, systémové číslo nebo pole 001 z knihovního katalogu atd. Tato čísla lze kombinovat např. s pořadím čísla v ročníku, číslem čísla apod.

4.4.4 Identifikátory pro úroveň ročníku:

- URN:NBN (bude přiděleno aplikací „Resolver URN:NBN“) – povinné, pokud lze přidělit
- UUID - povinná část pro dodavatele

4.4.5 Identifikátory pro úroveň čísla:

- UUID - povinná část pro dodavatele
- URN:NBN (je přiděleno aplikací „Resolver URN:NBN“) – povinné, pokud lze přidělit
- ISBN (pokud bylo záznamu v minulosti přiděleno)
- doi, handle apod. (pokud byly záznamu v minulosti přiděleny)
- Jiné identifikátory (čárový kód, systémové číslo, pole 001 apod.)

4.4.6 Identifikátory pro úroveň článku nebo obrázku:

- UUID - povinná část pro dodavatele
- doi, handle apod.
- Jiný identifikátor

4.4.7 Monografie

4.4.8 Granularita metadatového záznamu

- Základní intelektuální entitou pro monografie je 1 svazek
- V knihovních katalozích jsou někdy vícesvazkové monografie katalogizovány jako jeden soubor, tj. mají jeden záznam v katalogu, někdy jsou jednotlivé díly vedeny jako jednotlivé záznamy v katalogu, v obou případech musí vzniknout metadatový popis ke každému svazku jako základní intelektuální entitě a také PSP balíček pro každý takový svazek
- Pokud má monografie pouze jeden svazek, vznikne jeden metadatový popis (=METS záznam) pokud má monografie svazky dva, např. dvousvazkový slovník, jedná se o dvě intelektuální entity (svazek první a svazek druhý) a vzniknou tedy dva systémové metadatové záznamy, ke každému svazku jeden METS záznam a tedy dva PSP balíčky
- každý METS záznam musí obsahovat metadata o nadřazené intelektuální entitě (např. soubor monografie) tak, aby bylo možné obě entity virtuálně spojit a bylo jasné uživateli, že se jedná o jeden soubor/titul, který má dva svazky.

4.4.9 Identifikátory

- UUID - povinná část pro dodavatele

- URN:NBN (bude přiděleno aplikací „Resolver URN:NBN“) – povinné, pokud lze přidělit číslo České národní bibliografie – čČNB (povinné, pokud je možné je záznamu přidělit)
- ISBN (pokud je možné je záznamu přidělit)
- ISSN (pokud je obsaženo v záznamu)
- Jiný identifikátor – lze využít různé identifikátory, které nějakým způsobem propojí digitální záznam s fyzickou jednotkou – např. čárový kód, systémové číslo nebo pole 001 z knihovního katalogu atd. Tato čísla lze kombinovat např. s pořadím čísla v ročníku, číslem čísla apod.

4.4.10 Identifikátory pro úroveň vnitřní části:

- UUID - povinná část pro dodavatele
- URN:NBN (bude přiděleno aplikací „Resolver URN:NBN“) – povinné, pokud lze přidělit
- Jiné identifikátory (čárový kód, systémové číslo, pole 001 apod.)

4.4.11 Identifikátory pro úroveň přílohy:

- UUID (vygeneruje dodavatel) – povinné
- URN:NBN (bude přiděleno aplikací „Resolver URN:NBN“) – povinné, pokud lze přidělit
- doi, handle apod.
- Jiný identifikátor

4.4.12 Základní specifikace JPG2 (JPG2000)

Pro archivní kopie je zvolena tzv. bezeztrátová komprese, matematicky beze ztráty dat v kompresi. Pro kopie, která bude vystavena veřejnosti, je zvolena ztrátová komprese, tzv. vizuálně ztrátová komprese. Nastavení komprese vystavených kopií je optimalizováno vzhledem k zvolenému image serveru, který je naplánován v rámci zpřístupňující aplikace Kramerius 4 (OpenSource⁷). Z vystavených kopií budou pomocí tzv. image serveru automaticky generovány uživatelské kopie v grafickém formátu JPEG v předem nastaveném kompresním poměru, který bude kvalitativně odpovídat vizuální kvalitě vystavených kopií. Tato specifikace vychází z potřeb Národní knihovny České republiky a z porovnání doporučení pro kódování souborů JP2 v rámci zahraničních projektů zabývajících se digitalizací a dlouhodobým uchováváním obrazových dat.

Pro dokumenty (knihy, periodika) počítáme se skenováním v rozlišení minimálně 300 DPI. Barevná hloubka pro tyto dokumenty bude z důvodu urychlení produkce pevně nastavena na 24 bitů, barva (RGB). U dokumentů, kde není barva významová, bude možné v budoucnu snížit bitovou hloubku na 8 bitů, odstíny šedi.

⁷ Jsou to aplikace s otevřeným zdrojovým kódem, zdarma přístupné

Souhrnná tabulka přehledů specifikace JPG2(JPEG2000) pro jednotlivé data

Popis	I. Archivní kopie ¹ (Archival Copy)	II. Archivní kopie ² (Archival Copy)	I. Zpřístupňující kopie (Production Master Copy)	II. Zpřístupňující kopie (Production Master Copy)
Využití (Usage)	Knihy, periodika, mapy, manuskripty	Knihy, periodika, mapy, manuskripty	Knihy, periodika	Mapy, manuskripty
Grafický formát (Image still format)	JPEG 2000	JPEG 2000	JPEG 2000	JPEG 2000
Přípona formátu (File format)	*.jp2	*.jp2	*.jp2	*.jp2
Využitá část specifikace (Specification Part)	1. část JPEG 2000	1. část JPEG 2000	1. část JPEG 2000	1. část JPEG 2000
Migrační software (Migration software)	Kakadu	Kakadu	Kakadu	Kakadu
Druh komprese (Compression)	Bezeztrátová	Ztrátová	Ztrátová	Ztrátová
Transformace (Transformation)	5-3 reversible filter	9-7 irreversible filter	9-7 irreversible filter	9-7 irreversible filter
Kompresní poměr (Compression ratio)	1:2 až 1:3	1:3 až 1:4	1:20 až 1:30	1:8 až 1:10
Dlaždice (Tiling)	4096x4096	4096x4096	1024x1024	1024x1024
Průběh zobrazení (Progression order)	RPCL	RPCL	RPCL	RPCL
Počet dekompozičních úrovní (Decomposition levels)	5 nebo 6	5 nebo 6	5	5 nebo 6
Počet vrstev kvality (Quality layers)	1	1	12 (logaritmicky)	12 (logaritmicky)
Velikost regionů (Precinct size)	256x256 pro první dvě dekompoziční úrovně, 128x128 pro nejnižší úrovně	256x256 pro první dvě dekompoziční úrovně, 128x128 pro nejnižší úrovně	256x256 pro první dvě dekompoziční úrovně, 128x128 pro nejnižší úrovně	256x256 pro první dvě dekompoziční úrovně, 128x128 pro nejnižší úrovně
Zájmové oblasti (Regions of Interests)	Ne	Ne	Ne	Ne
Velikost bloků (Code block size)	64x64	64x64	64x64	64x64
Značka lokalizující dlaždice TLM (Tile Length Markers)	Ano „R“	Ano „R“	Ano „R“	Ano „R“
Přemostění (Bypass)	Ano	Ano	Ano	Ano
ICC profily (ICC profile)	Ano	Ano	-	Ano
Značka začátku hlavičky segmentu paketů SOP (Start Of Packet Header)	Cuse_sop=yes	Cuse_sop=yes	Cuse_sop=yes	Cuse_sop=yes
Značka konce hlavičky segmentu paketů EPH (End Of Packet Header)	Cuse_eph=yes	Cuse_eph=yes	Cuse_eph=yes	Cuse_eph=yes
Vložená metadata (Embedded Metadata)	XML metadata vložená v JP2 XML BOXu	XML metadata vložená v JP2 XML BOXu	XML metadata vložená v JP2 XML BOXu	XML metadata vložená v JP2 XML BOXu
Vytvořeno dne 4. 4. 2012, Bedřich Vychodil Created on 4th April 2012 by Bedřich Vychodil				

4.5 Současný vývoj LTP systému

V současné době je systém v tzv. pilotním provozu, kdy se zatím využívá pouze produkce NDK. Je to svým způsobem test zátěže. LTP systém je nyní v období zrodu. Vstup do tohoto úložiště „kontroluje“ tzv. validátor dat. Tento software je poskytnut zdarma, pro kontrolu externích dodavatelů. Tato před-kontrola napomůže k odstranění nedostatků dat. Dodavatel si již předem dokáže určit, zda jsou jeho data validní a odpovídají standardům používaných pro LTP úložiště

4.5.1 Validátor – vstupní kontrola systému LTP

Účelem této volně šířitelné aplikace je poskytnout jednoduchý nástroj pro kontrolu vytvořených PSP balíčků, zda odpovídají definici metadat pro periodika a monografie NDK .

Aplikace je modifikovatelná. To znamená, že podle nastavení parametrů aplikace je aplikace schopna provést kontrolu jednoho balíčku nebo celé dávky balíčků obsažené v dané složce. Úkolem této aplikace je kontrola struktury PSP balíčku, tj. validovat hlavní metadatový soubor a validovat vedlejší metadatové soubory.

Funkce aplikace

Aplikace provádí validaci kompletního vstupního PSP balíčku podle předdefinovaných standardů pro periodikum, případně monografii a výsledek zapíše formou reportu obsahujícího informaci o tom zda balíček vyhověl pravidlům, popřípadě soupis chyb.

Průběh validace lze rozdělit na několik postupných kroků:

- Kontrola struktury PSP balíčku.
- Kontrola hlavního metadatového souboru (hlavní METS).
- Kontrola vedlejších metadatových souborů (vedlejší METS soubory v amdSec složce).

4.5.2 Validace struktury balíčku

Prvotní kontrola, která vyhodnotí, zda balíček obsahuje všechny povinné soubory a složky a zda jsou správně pojmenované dle jmenné konvence

4.5.3 Struktura balíčku:

- info.xml
- masterCopy (složka) – obsahuje obrazy JPEG 2000 lossless
- userCopy (složka) – obsahuje obrazy JPEG 2000 lossy
- ALTO (složka) – obsahuje soubory ALTO.xml pro každou stranu
- TXT (složka) – obsahuje soubory OCR.TXT pro každou stranu
- Hlavní_METS.xml
- Soubor.md5

4.5.4 Validace hlavního metadatového souboru

Kontrola, zda hlavní „METS“ záznam má správný formát a obsahuje všechny povinné položky.

Tuto validaci lze rozdělit do následujících kroků:

- Validace správnosti xml pomocí mets.xsd a mods-3-4.xsd.
- Validace fileSec části - kontrola zda odkazované soubory (element FLocat atribut xlink:href) opravdu skutečně existují.
- Validace přítomnosti povinných položek a atributů pomocí šablon – na základě typu dokumentu validátor vybere profil, který obsahuje seznam jednotlivých šablon, které mají být použity pro validaci.

4.5.5 Validace vedlejších metadatových souborů

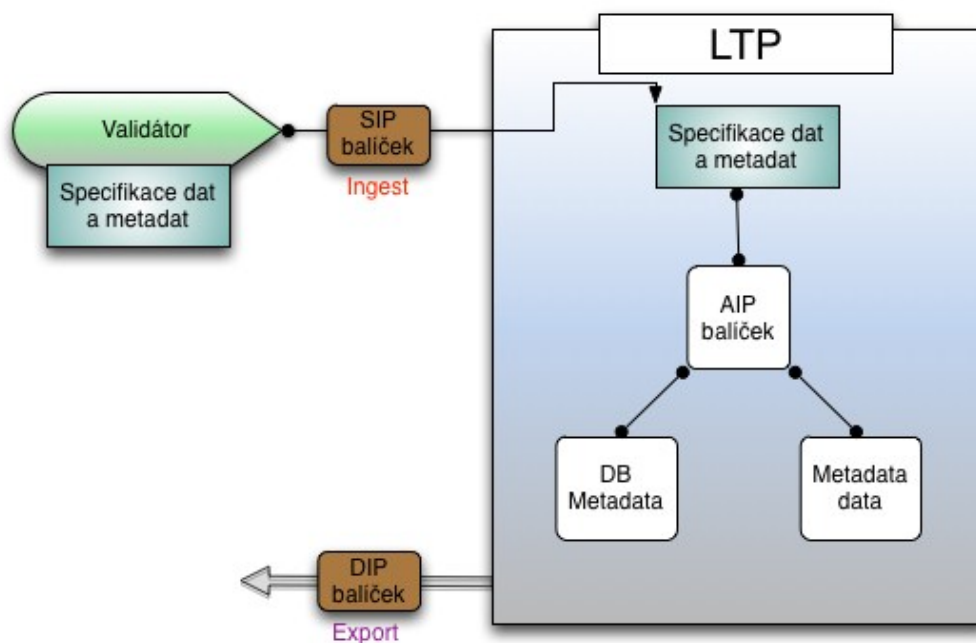
Postupná kontrola vedlejších metadatových souborů (AMD_METS).

Kroky validace pro jednotlivé soubory:

- Validace přítomnosti povinných položek a atributů pomocí šablon.
- Validace správnosti xml jednotlivých událostí (premis:event) pomocí premis.xsd.
- Validace správnosti xml jednotlivých agentů (premis:agent) pomocí premis.xsd.
- Validace správnosti xml technických metadat v elementech MIX001 a MIX002 pomocí mix.xsd.

4.6 LTP jako systém

LTP jako systém je ve fázi přijímání dat pouze z digitalizace a harvestingu webových zdrojů, tzv. webarchiv . Původní myšlenka vytváření automatické specifikace pro zatřídění SIP balíčků zatím neexistuje. Specifikace se vytvářejí ruční metodou a implementaci do LTP. Po té je možnost injestovat balíčky do systému. LTP se chová jako úložiště, jeho předpokládaná vlastnost dynamického úložiště neexistuje. Dále není možnost opravy a konvertování obsahů balíčků SIP do jiných vnitřních formátů. To znamená, že například v případě konverze obrazových materiálů masterCopy a userCopy nelze zpětně získat ze systému data, zkonvertovat je a zpětně je injestovat. Tato zásadní skutečnost omezuje funkčnost celého systému. Tento fakt je kritický i v situaci, kdy by mělo dojít k hromadné změně nejen formátu ale i zápisu do metadat. LTP je nyní ve velikost 1PTb a každým dalším přísunem dat se zvětšuje. Metadata obsahující každou entitu (SIP balíček) prozatím vykazují chybovost.



Obrázek 2 Detailní schéma LTP

Typy vstupních dat, která se plánují pro ingest do LTP, jsou následujících typů:

- Digitalizace staršího typu Monografie a Periodika
- Mapy
- Rozhlasové vysílání
- Radiové vysílání
- Filmový archiv
- 3D archiv kulturních památek

5 Závěr

Závěrem lze říci, že trendy zde popsané vnesly řád do ukládání dat. Systém LTP je prozatím na začátku a čeká ho dlouhý vývoj. Vše stojí v současné době na jenom zásadním rozcestí, kdy je nutné si odpovědět na základní sadu otázek:

Add1:

Jak se chovat k datovým formátům

- Vytvořit statické LTP, kde bude nutné, pro každou formu dat vytvořit speciálně ručně specifikaci
- Vytvořit dynamické úložiště s tím že specifikaci si dokáže úložiště vytvořit samo

Add2:

Jakým způsobem konvertovat data, a tím sledovat vývoj v trendech

- Konvertovat vždy celý datový obsah do nového formátu
- Přizpůsobit pouze emulátor zobrazení

Add3:

Dále zde existuje ekonomická a lidská stránka projektu

- Zda vytvořit plně automatický systém
- Zda vytvořit team techniků a badatelů, kteří se budou o data starat

Takto položených otázek bude spousta. Budou stále vznikat nové a nové a bude nutné na ně hledat odpovědi.

6 Seznam použitých zdrojů

6.1 Informační zdroje:

Jako hlavní zdroj bych uvedl mojí současnou práci Národní knihovnu ČR

6.2 Internetové zdroje:

- www.nkp.cz
- www.ndk.cz
- <http://www.techlib.cz>
- <http://windows.microsoft.com>
- <http://ciks.vse.cz>

6.3 Literatura

- Building and scaling SAP business information warehouse on DB2 UDB ESE [online]. 1st ed. San Jose, CA: IBM, International Technical Support Organization, 2004. IBM redbooks. DB2 information management software [cit. 2013-05-14].
Dostupné z: <http://site.ebrary.com/lib/natl/Doc?id=10112912>.
- The data warehouse lifecycle toolkit [online]. 2nd ed. Indianapolis, Ind.: Wiley Pub., ©2008 [cit. 2013-05-14].
Dostupné z: <http://site.ebrary.com/lib/natl/Doc?id=10521400>.
- Laberge, Robert. Datové sklady: agilní metody a business intelligence. 1. vyd. Brno: Computer Press, 2012. 350 s. ISBN 978-80-251-3729-1.
Implementing Tivoli Data Warehouse 1.2 [online]. 1st ed. White Plains, NY: IBM, International Technical Support Organization, ©2004. IBM redbooks [cit. 2013-05-14]. Dostupné z: <http://site.ebrary.com/lib/natl/Doc?id=10112957>.
- Mundy, Joy. The Microsoft data warehouse toolkit: with SQL server 2008 R2 and the Microsoft Business Intelligence toolset [online]. 2nd ed. Indianapolis, IN: Wiley, ©2011 [cit. 2013-05-14].
Dostupné z: <http://site.ebrary.com/lib/natl/Doc?id=10510406>.
- Buřita, Ladislav et al. Analýza, návrh a implementace datového skladu: prototyp-2/PERSO. Brno: Univerzita obrany, 2006. 99 s. ISBN 80-7231-159-X.
- Šarmanová, Jana. Datové sklady a dolování znalostí z nich: průvodce studiem. Ostrava: Vysoká škola báňská - Technická univerzita Ostrava, Fakulta elektrotechniky a informatiky, 2003. 42 l. ISBN 80-248-0302-X.
- Šimonová, Stanislava. Regionální datový sklad: (teze disertační práce): obor Informatika ve veřejné správě. Vyd. 1. Pecinovský, Josef. Archivace a komprimace dat: jak zálohovat data, jak komprimovat soubory WinRAR, WinZip, WinAce, Windows a nástroje komprese dat, jak archivovat data ve Windows. 1. vyd. Praha: Grada, 2003. 116 s. Snadno a rychle. ISBN 80-247-0659-8.
- Kříž, Libor. Komprimační a archivační programy. Vyd. 1. Praha: Computer Press, 2002. x, 115 s. Utility. ISBN 80-7226-757-4.

- Elektronická archivace [online]. [Praha]: Central European Advisory Group, [2007?] [cit. 2013-05-14]. Dostupné z: <http://digiarchiv.eu>.
- Ernst, Wolfgang a Parikka, Jussi, ed. Digital memory and the archive. Minneapolis: University of Minnesota Press, ©2013. 265 s. Electronic mediations; v. 39. ISBN 978-0-8166-7766-5.
- Olson, Jack E. Database archiving: how to keep lots of data for a very long time. Amsterdam: Elsevier, ©2009. xviii, 291 s. ISBN 978-0-12-374720-4.

7 Seznam použitých zkratk

NDK	Národní digitální knihovna
MZK	Městská zemská knihovna
IOP	Název projektu financování z EU

JHOVE	Název Software
PRONOM	Název Software
DROID	Název Software
NZME	Název Software - Nový Zéland
OAIS	Název Software
PREMIS	Metadatové formáty
METS	Metadatové formáty
MODS	Metadatové formáty
MIX	Metadatové formáty
TRAC	Důvěryhodný digitální repozitář
API, SDK	softwarové propojení systémů
Scheduler	program na časované zpuštění procesů či dalších programů
archival storage	Datové úložiště pro archivy
Ingest	vložení dat
Harvesting	Sklizení dat
Onelog, Shibboleth	Webový portál pro přístup do systému
RegistrDigitalizace	Systém, který eviduje knihy použité v digitalizaci
Roll back scenario	Zpětná obnova scénáře
Shibboleth	Portálová autentizace
GUI	Grafické rozhraní programu
SIP	balíček digitálních dat pro vkládání dat do LTP
Metadata	popisová data k objektu, většinou ve formátu XML
ILS	katalogizační modul
LTP	(Long Term Preservation) dlouhodobé úložiště dat
DPI	DPI (Dots Per Inch) vyjadřuje počet obrazových bodů pixelů na jeden délkový palec, přičemž délkový palec odpovídá 2,54 cm
Resolver URN:NBN	je aplikace, která umožňuje přidělování, evidenci, správu a vyhledávání trvalých identifikátorů URN:NBN a vyhledávání digitálních dokumentů prostřednictvím těchto identifikátorů
čČNB	identifikátor - číslo České národní bibliografie

ISSN	(International Standard Serial Number) je osmimístný číselný kód, kterým se jednoznačně identifikují názvy periodik a ostatních tzv. pokračujících zdrojů vydávaných kdekoliv na světě. Záznamy ISSN jsou uloženy v referenční bázi - mezinárodním Registru ISSN.
ISBN	(International Standard Book Number), mezinárodní standardní číslo knihy) je alfanumerický kód určený pro jednoznačnou identifikaci knižních vydání
JPG2	Obrazový formát JPG2000
FTP, http	protokoly propojení ve světě internetu
Workflow	Pracovní prostředí
check-sum	kontrolní součet
OCR	název programu na rozpoznávání písmen v obrazovém materiálu

8 Seznam obrázků

Obrázek 1 Schéma umístění LTP v systému	18
Obrázek 2 Detailní schéma LTP	46