Systém pravidel pro digitální archivaci NK ČR

1.1 Slovníček pojmů

Digitální archivace

Všechna opatření vykonávaná pro zajištění dlouhodobé ochrany a trvalého zpřístupňování digitálních dokumentů. Tato ochranná opatření zahrnují řešení technických (například hardwarových nebo softwarových), organizačních, finančních a dalších aspektů dlouhodobé ochrany a trvalého zpřístupňování digitálních dokumentů.

Hlavní rizika pro digitální dokumenty:

- selhávání, nekompatibilita a zastarávání hardwarových zařízení,
- výpadky elektrické energie a další rizika spojená s elektrickou energií,
- selhávání, nekompatibilita a zastarávání síťových technologií,
- zastarávání a degradace datových nosičů, na kterých jsou digitální dokumenty uloženy,
- zastarávání softwarových aplikací,
- omezení vyplývající z práv duševního vlastnictví vážících se k softwarovým aplikacím,
- zastarávání formátů digitálních dat,
- omezení vyplývající z práv duševního vlastnictví vážících se k formátům digitálních dat,
- organizační rizika,
- znalostní rizika, např. nedostatek odborných pracovních sil
- finanční rizika, např. nezajištění trvalé udržitelnosti provozu

Dlouhodobý aspekt digitální archivace

Dlouhodobá ochrana není vymezena časovým horizontem, ale výskytem rizik – dlouhodobá ochrana musí trvat tak dlouho, jak dlouho budou existovat technologická a další rizika, zejména ta, která vyplývají ze zastarávání a selhávání informačně-komunikačních technologií, ze zastarávání formátů a z degradace datových nosičů.

Vzhledem k tomu, že se technologie stále mění a

Digitální dokument

Množina počítačových souborů, jejichž reprodukcí získá čtenář přístup k intelektuální entitě. Počítačové soubory z této množiny reprezentují vlastní dokument, ale mohou též reprezentovat k němu přidružená metadata.

Intelektuální entita

Jednotka intelektuálního obsahu (množina informací), která je jako jeden celek popsána bibliografickým záznamem v katalogu nebo v metadatech přidružených k informačnímu balíčku (podle modelu OAIS) v digitálním repozitáři nebo digitální knihovně.

Příkladem digitálního dokumentu je 68 souborů ve formátů JPEG2000 (vlastní dokument) a 1 soubor v XML (metadata), které jsou dohromady potřebné k reprodukci digitálního obrazu tištěného vydání Máchova Máje vydaného Janem Spurným v roce 1836 v Praze. Intelektuální entitou je toto první tištěné vydání Máchova Máje, které má záznam v České národní bibliografii obsahující trvalý identifikátor „cnb001417516“.

Formátová migrace

Převod souboru z aktuálního formátu do nového formátu jako opatření digitální archivace. Jedná se o typ transformační migrace podle OAIS.

Emulace

Vytvoření takového systému, který bude fungovat stejně, jako systém, pro který byl digitální dokument původně vytvořen a ve kterém byl původně užíván.

1.2 Krátkodobý plán ochrany

Následná opatření nejsou specifickými opatřeními digitální archivace – mohou se užívat i pro potřeby krátkodobé ochrany digitálních dokumentů, nicméně jsou nezbytným základem pro vlastní opatření digitální archivace.

1.2.1 Bitová ochrana – základní ochrana digitálních dat

Jako základní podpůrné opatření pro digitální archivaci bude Národní knihovna ČR provádět tzv. bitovou ochranu (ochrana bitového toku /bitstream/). Bitová ochrana znamená pouze zajištění uchovávání digitálních souborů v původní podobě (tzn., že neřeší problém zastarávání formátů digitálních dat). Bitová ochrana je základní podmínkou dlouhodobých opatření digitální archivace. Bez bitové ochrany by hrozilo, že v budoucnosti nebude možné provádět žádné formátové migrace, protože nebudou existovat žádné soubory, např. budou zničeny kvůli selhání konkrétní záložní technologie apod.

Bitová ochrana znamená replikování digitálních dat na několik různých fyzicky oddělených datových nosičů (záloha); pravidelnou kontrolu integrity dat a také systematickou výměnu technologií a datových technologií.

Duplikace (zálohování) dat

- archivní balíček obsahující digitální dokument, bude vždy současně uložen na minimálně dvou (ideálně na třech) fyzicky oddělených datových nosičích (tj. ve dvou, resp. třech totožných instancích)
- minimálně jedna instance bude vždy geograficky významně vzdálena od zbývajících instancí
- např. jedna instance bude na magnetické pásce v NK ČR lokalita Klementinum, druhá na pevném magnetickém disku v NK ČR lokalitě Hostivař
- v případě potřeby může být deklarovaný počet digitálních instancí v budoucnosti zvýšen
- duplikace bude prováděna vždy v rámci prvního uložení archivního balíčku do digitálního repozitáře
- duplikace bude dále prováděna po každé změně v archivním balíčku (formátová migrace, úprava metadat apod.)

Kontrola integrity dat

- v pravidelných intervalech budou všechny soubory v archivních balíčcích kontrolovány z hlediska neporušenosti integrity, zejména užitím aktuálně zvolených kontrolních mechanismů
    - v současnosti na základě kontrolních součtů (tzv. „check sums“)
    - interval – v závislosti na schopnostech LTP systému, ideální stav je kontrola nepřetržitá

Výběr datových nosičů

- Národní knihovna ČR se zavazuje vybírat datové nosiče na základě aktuálních mezinárodních doporučení a zvyklostí v oboru
- Národní knihovna ČR již nepovažuje optické datové nosiče za datové nosiče vhodné pro digitální archivaci (viz výzkumy UNESCO, 2006)
- Jako datové nosiče pro současný systém uložení jsou vybrány magnetické pevné pásky a magnetické pevné disky
- Předpokládaný interval obměny datových nosičů na základě současné vývoje
- V souvislosti s požadavky řešit aktuální problémy změn technologického vývoje a z toho vyplývajících nových doporučení pro výběr nejvhodnějších nových nebo jiných technologií může být tento interval v budoucnosti změněn

Systematická změna hardwarových a síťových technologií

- Předpokládaný interval obměny datových nosičů na základě současné vývoje
- Interval systematické obměny technologií se může v souvislosti s požadavky řešit aktuální problémy změn technologického vývoje a z toho vyplývajících nových doporučení pro výběr nejvhodnějších nových nebo jiných technologií změnit

1.3 Dlouhodobý plán logické ochrany digitálních dat

Z hlediska potřeb dlouhodobé ochrany a trvalé zpřístupnitelnosti digitálních dokumentů nejsou opatření bitové ochrany dostatečná. Musí být doplněno logickou ochranou, která zajistí vyhledatelnost, zobrazitelnost, použitelnost a srozumitelnost archivovaných digitálních dokumentů v budoucnosti navzdory technologickým změnám, které mezitím proběhnou. Proto zavádí NK ČR následující opatření obecné (systémové) a konkrétní roviny (konkrétní opatření).

1.3.1 Systémová rovina

Systémová rovina dlouhodobých opatření logické ochrany specifikuje požadavky na digitální repozitář jako hlavní systém pro dlouhodobou ochranu.

1.3.1.1 Základní povinnosti

Základním systémovým opatřením je vytvoření, provozování a další rozvoj digitálního repozitáře, který bude plnit šest základních funkcí definovaných funkčním modelem normy OAIS:

1) Uzavírat dohody s producenty informací a přijímat od nich patřičné informace.

2) Dosáhnout takového stupně kontroly, která zaručí dlouhodobou ochranu.

3) Samostatně nebo ve spolupráci s dalšími stranami stanovit, jaké uživatelské komunity budou tvořit designovanou komunitu.

4) Zajistit, aby informace v repozitáři byly nezávisle srozumitelné.

5) Řídit se zdokumentovanými pravidly a postupy, které ochrání informace před všemi možnými riziky a umožní je zpřístupňovat ve formě autentifikovaných kopií originálu nebo ve formě, ze které lze originál odvodit.

6) Zpřístupňovat informace designované komunitě.

1.3.1.2 Důvěryhodnost

Digitální repozitář musí splňovat nejen základní požadavky stanovené normou OAIS, ale současně by měl být *důvěryhodným digitální repozitářem*. Tuto skutečnost může zajistit pouze (i opakované) využití nástrojů, případně i externích subjektů pro kontrolu toho, zda repozitář splňuje to, k čemu se zavázal.

Národní knihovna ČR se zavazuje:

- Projít certifikací TRAC provedenou externím subjektem před zahájením provozu digitálního repozitáře
    - http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf

- V případě objevení se nástroje, který bude považován za lepší, dojde k záměně s TRAC
    - NK ČR bude sledovat zejména výstupy aktivity vedenou CCSDS „Audit and Certification of Trustworthy Digital Repositories“
    - http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206521R1/Overview.aspx
- Podpůrné využití dalších nástrojů:
    - DRAMBORA
        - http://www.repositoryaudit.eu/
    - Data Seal of Approval (DSA)
        - http://www.datasealofapproval.org/
    -

1.3.1.3 Plánovací procesy

Funkční model OAIS obsahuje jako jednu ze šesti základních funkčních entit (dále modulů) tzv. modul plánování ochrany („preservation planning“). Modul má za úkol poskytovat služby a funkce pro monitorování vnějšího kontextu repozitáře a dávat doporučení, která zajistí kontinuitu trvalého zpřístupňování informací bez ohledu na zastarávání počítačových technologií a další rizika spojená s plněním cílů digitální archivace.

Mezi základní funkce modulu patří vytváření plánů pro migrace, vývoj softwarových prototypů nebo testování plánů na implementaci migračních cílů. Mezi další funkce patří vytváření doporučení pro institucionální systém pravidel a zavádění standardů a monitorování změn v technologiích a designované komunitě, která je klientem repozitáře.

Modul plánování je mj. tvořen konkrétními zaměstnanci Národní knihovny ČR Důvod je prostý – repozitář OAIS je organizce tvořená nejen technickými systémy, ale též lidmi (obvykle zaměstnanci instituce provozující digitální repozitář). Plánovací modul má za úkol sledovat změny v technologiích, z logiky věci vyplývá, že toto sledování změn a jejich vyhodnocování musejí provádět lidští pracovníci, nelze jej plně automatizovat.

1.3.2 Dílčí specifická opatření

Vzhledem k potřebám digitální archivace NK ČR definuje dvě základní oblasti aktivit: metadatové a formátové strategie. Metadata jsou nezbytnou součástí dlouhodobé ochrany, zejména vzhledem ke skutečnosti, že digitální dokumenty nejsou samy o sobě dostatečně autoreferenční. Formátové strategie jsou nezbytné vzhledem k převažujícímu názoru, že základní problém digitální archivace v současnosti spočívá v problému formátů, jako je zastarávání formátů nebo práva duševního vlastnictví spojená s formáty.

1.3.2.1 Formátové strategie

a) Migrace

Jednou z hlavních strategií pro digitální archivaci v současnosti považované v Národní knihovně ČR jako hlavní řešení pro případy potřeby budoucích formátových změn (v souvislosti s rizikem zastarávání formátů nebo objevením se vhodnějších formátů pro archivaci a zpřístupňování) je formátová migrace.

Zastarávání formátu se udává v řádu 8-20 let.

Výběr nového formátu v budoucnosti se bude zakládat na základě plánování ochrany zmíněného výše. Cílový formát bude vybrán dle budoucí situace.

b) Emulace

Emulace je další možností, jak zajistit použitelnost digitálních objektů v budoucnu. Jedná se o metodu, která není ještě široce probádána a implementována. Principem je znovuvytvoření systému, pro který byl digitální objekt původně vytvořen, a to napodobením (simulace) vlastností hardwaru a softwaru.

Emulace bude primárně využívána pro data WebArchivu, pro zpřístupňování uživatelských verzí.

Pro tyto potřeby bude využita zahraniční aplikace „WayBack machine“.

Na jednu stranu nebude emulace tak náročná na datové toky (operace nad velkou množinou dat jako v případě migrací), na druhou stranu tento nástroj vyžaduje další vývoj.

c) Strategie výběru formátů digitálních objektů

Formáty digitálních objektů musejí být vždy vybrány podle toho, že:

1) *daný formát je mezinárodně uznávaným standardem*
2) *má otevřenou specifikaci – tedy lze získat jeho kompletní dokumenty*
3) *ideálně není proprietární (vázaný právy duševního vlastnictví), v odůvodněných případech (kdy neexistuje neproprietární formát, nebo existuje, ale není široce užíván v mezinárodní knihovnické komunitě) lze akceptovat i (otevřený) proprietární formát*

Výběr formátů pro NDK / standardizace

Formáty pro archivaci

- (AIP, archivní balíčky, archivní kopie, archivní verze)
- JPEG2000 pro produkci digitálních dokumentů prostřednictvím digitalizace
    - naše specifikace: PhDr. Bedřich Vychodil (kandidát PhD.)
- PDF/A – zvažováno pro e-born monografie
- JPEG – historická data zůstanou zachována

- další – bude muset vzniknout seznam preferovaných formátů vhodných z různých hledisek pro archivaci, obecně ale NK ČR musí být schopná archivovat jakýkoliv formát

d) Normalizace dat

Externí data

Převod souborů dodávaných do NK ČR od třetích stran do vybraných formátů. Typicky pro příjem dat z elektronického dobrovolného výtisku. Musí vycházet ze seznamu preferovaných formátů a dohod s třetími stranami.

Data vzniklá činností NK ČR

Strategie pro potřeby převodu dat z CDÚ do připravovaného digitálního repozitáře v rámci NDK.

- kontrola stávajících dat
- jejich následný převod z nevyhovujících formátů na formáty považované v současnosti za nejvhodnější pro potřeby digitální archivace
- oprava dat a převod na formáty považované v současnosti za nejvhodnější pro potřeby digitální archivace

Procesy:
DJVU to JPEG2000
JPEG to JPEG2000 (zvažováno do budoucna)

e) Strategie pro webarchiv

- pro Webarchiv nezaručujeme normalizaci dat
- Řada sklizených dat je uložena ve formátech, které neodpovídají požadavkům dlouhodobé ochrany
    - tato praxe je však natolik běžná ve většině světových webarchivů, že ji lze považovat za jediné současně schůdné řešení
- pro potřeby zpřístupňování bude užita emulace
    - adoptovaná aplikace WayBack machine

1.3.2.2 Metadatové strategie

NK ČR si je vědoma, že metadata jsou jedním ze základních předpokladů úspěšné logické dlouhodobé ochrany digitálních dat. Metadata musí vznikat stále během životního cyklu digitálních objektů (vznik, uložení, opatření ochrany). Metadata jsou východiskem pro systémy uložení i pro jednotlivé metody ochrany, jako je např. migrace formátů.

Je nutno používat výhradně formáty standardní, které jsou široce využívané v knihovnické komunitě a mají tak zajištěnu podporu. Primárně se bráníme vytváření nových schémat metadat a jejich využívání.

Metadatová standardizace

- vytváření současných metadat podle doporučení pro NDK
  - zajišťuje Odbor digitálních fondů v kooperaci s odděleními digitalizace
- doporučované metadatové formáty pro NDK
- strukturální metadata (zabalení všech dalších metadat) – standard METS
  - technická metadata – PREMIS, MIX
  - bibliografická metadata – DC, MODS
  - metadata pro OCR (plnotextové vyhledávání) – METS ALSO

Požadavky pro deponovací entitu repozitáře (Ingest)

a) obrazová data
preferované formáty pro strukturální, popisná, technická i ochranná metadata
METS, MODS, MARCXML, PREMIS, MIX

b) textová data
preferované formáty pro strukturální, popisná, technická i ochranná metadata
METS ALTO, MODS, PREMIS, MASTER TEI

c) webarchiv
preferované formáty pro strukturální, popisná, technická i ochranná metadata
Formáty ARC a WARC

Požadavky pro archivační entitu repozitáře („LTP systém“)

a) obrazová data
preferované formáty pro strukturální, popisná, technická i ochranná metadata
METS, MODS, MARCXML, PREMIS, MIX

b) textová data
preferované formáty pro strukturální, popisná, technická i ochranná metadata
METS ALTO, MODS, PREMIS, MASTER TEI

c) webarchiv
preferované formáty pro strukturální, popisná, technická i ochranná metadata
WARC

Budoucí metadatové migrace

Převod nebo úprava metadat podle aktuálně doporučovaných metadatových formátů v budoucnosti na základě plánovací entity

1.3.2.3 Důvěryhodný systém trvalé identifikace

Strategie implementace a udržování důvěryhodného systému trvalé identifikace.

Význam trvalých identifikátorů pro potřeby dlouhodobé ochrany a trvalého zpřístupňování považujeme za jeden z pilířů digitální archivace. V digitálním světě je požadavek rozšířit funkcionalitu identifikátorů (jednoznačně identifikovat) též o možnost pomocí identifikátoru bezprostředně získat aktuální lokaci dokumentu v internetové síti. Vzhledem k známému problému nefunkčních internetových adres je tento druhý požadavek současně velkým závazkem. Proto v NK ČR zavádíme systém trvalé identifikace založený na de facto standardu URN-NBN.[1] Podrobnosti viz systém pravidel pro trvalou identifikaci

1.3.2.4 Práva duševního vlastnictví

S dlouhodobou ochranou a zpřístupnitelností digitálních dokumentů neodlučně souvisí také otázka přidružených duševních práv. NK ČR se zavazuje nejen dlouhodobě archivovat a zpřístupňovat digitální dokumenty, ale také práva k nim se vážící. Archivace a zpřístupnění bude probíhat výhradně v souladu s platnou legislativou a k jednotlivým digitálním objektům vážícími se omezeními nebo duševními právy.
K tomu je třeba stálá podpora ze strany specializovaného právníka.

1.3.3 Zkušenosti ze zahraničí

Vzhledem k tomu, že výzkum a vývoj v oblasti digitální archivace je teprve ve svém počátku[2] a že problematika je natolik složitá, že ji není možné řešit samostatně na národní úrovni, se Národní knihovna ČR zavazuje sledovat, přejímat, syntetizovat, upravovat a implementovat doporučení, standardy, nejvhodnější postupy („best practices“) mezinárodní doporučení vydaná významnými mezinárodními projekty nebo významnými mezinárodními výzkumnými aktivitami. V současnosti se jedná zejména o tyto instituce a projekty (nikoliv exkluzivně):

Referenční instituce

- Library of Congress / Kongresová knihovna (USA)
    - metadatové standardy (PREMIS, METS, MODS)
    - formátový vývoj (JPEG2000)

[1] http://www.ietf.org/rfc/rfc3188.txt
[2] První významné výzkumné projekty s nadnárodním přesahem byly ukončeny teprve v nedávné minulosti – např. CASPAR v 2010, PLANETS v 2011.

    - http://www.loc.gov/
- National Library of Australia / Australská národní knihovna (Austrálie)
    - archivace webu, digitální repozitář
    - http://www.nla.gov.au/
- National Library of New Zealand
    - http://www.natlib.govt.nz/
    - Návrh – navázat užší česko-novozélandskou spolupráci – tato knihovna patří mezi lídry ve vývoji softwarového řešení pro digitální archivaci
- Koninklijke Bibliotheek / Královská knihovna (Nizozemsko)
    - digitální repozitář, digitalizační worklow, formáty
    - http://www.kb.nl/
- Kansalliskirjasto / Finská národní knihovna (Finsko)
    - digitální repozitář, digitalizační worklow, formáty, trvalá identifikace
    - http://www.nationallibrary.fi/
- Nasjonalbiblioteket / Norská národní knihovna (Norsko)
    - digitální repozitář, digitalizační worklow, formáty
    - http://www.nb.no/

Referenční projekty

- CASPAR (2006-2009)
    - http://www.casparpreserves.eu/
- PLANETS (2006-2010)
    - http://www.planets-project.eu/
- LIFE (2006-2010)
    - http://www.life.ac.uk/
- LIWA (2008-2011)
    - http://www.liwa-project.eu/
- PROTAGE (2007-2010)
    - http://www.protage.eu/
- KEEP (2009-2012)
    - http://www.keep-project.eu/
- SCAPE : Scalable Preservation Environments (2011-2014)
    - http://www.scape-project.eu/

Probíhající standardizační aktivity

- OAIS 2
    - revize normy
    - http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/Overview.aspx
- CCSDS - „Audit and Certification of Trustworthy Digital Repositories“
    - příprava normy pro certifikaci důvěryhodného digitálního repozitáře

- http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206521R1/Overview.aspx

Referenční nástroje

- PLATO
  - softwarový plánovací nástroj
  - http://www.ifs.tuwien.ac.at/dp/plato/
- TRAC
  - certifikační nástroj (kontrolní seznam)
  - http://www.crl.edu/archiving-preservation/digital-archives/certification-and-assessment-digital-repositories
- DRAMBORA
  - softwarový auditní nástroj
  - http://www.repositoryaudit.eu/
- JHOVE / JHOVE 2
  - softwarový nástroj pro formátové operace
  - http://hul.harvard.edu/jhove/
  - https://bitbucket.org/jhove2/main/wiki/Home
- PRONOM
  - online formátový registr
  - http://www.nationalarchives.gov.uk/PRONOM/
- DROID
  - softwarový nástroj pro formátové operace
  - http://droid.sourceforge.net/
- UDFR / GDFR
  - chystaný online formátový registr
  - http://www.udfr.org/
  - http://www.gdfr.info/