Nové směry v dlouhodobém uchovávání digitálních dokumentů v mezinárodním kontextu

Bohdana Stoklasová
Jan Hutař
Národní knihovna ČR
bohdana.stoklasova@nkp.cz
jan.hutar@nkp.cz

Národní digitální knihovna

Obsah

- Potřeby a cíle paměťových a vědeckovýzkumných institucí (výsledky průzkumu DPE)
- Certifikace důvěryhodných úložišť
- Mezinárodní metadatové standardy pro digitální úložiště a vznikající národní standardy
- Hlavní vědeckovýzkumné cíle v oblasti uchovávání digitálních dokumentů (DPE Roadmap)

Národní digitální knihovna

Potřeby a cíle paměťových a vědeckovýzkumných institucí

- **Průzkum DPE – 2006, 2007**
- **NK v rámci Market and Technology Trends Analysis**
- **Evropské NK v úplnosti 2006**
- **Archivy a vědeckovýzkumné instituce – vzorek 2007**

digitalpreservationeurope

Digital Preservation Europe

DPE - cíle

- Vytvořit základnu pro aktivní spolupráci, výměnu a šíření výsledků výzkumu a vývoje a zkušeností
- Přispět k rozšíření služeb vázaných na ochranu dokumentů a zvýšení jejich životaschopnosti a spolehlivosti
- Zvýšit všeobecné povědomí o důležitosti ochrany dokumentů a znalosti i dostupné zdroje pro tuto oblast

Průzkum DPE

Podrobnější verze:

- **Website DPE (publikována celá Market and Technology Trends Analysis) http://www.digitalpreservationeurope.eu**

- **Website NDK (Národní digitální knihovna) http://www.ndk.cz, složka „Publikace“**

- **Knihovna 2/2006**

Průzkum DPE

- ***Otázka 1: Patří dlouhodobé uchovávání digitálních dokumentů (včetně migrace, emulace, ochranných metadat atd.) mezi klíčové strategické priority ve vaší instituci? (možnosti: ano, ne, zatím ne)***
- **Odpovědi: Ano: 82% národních knihoven, 66% archivů a 70% vědeckovýzkumných institucí.**
- **Závěr: Ve všech institucích jsou ochrana digitálních dokumentů a jejich dlouhodobé uchovávání prioritou současnosti i budoucnosti. Rozdíl mezi paměťovými a vědeckovýzkumnými institucemi existuje v poměru mezi dokumenty určenými k dlouhodobému uchovávání a dokumenty určenými k uchovávání středně nebo krátkodobému. Procento dokumentů určených k dlouhodobému uchovávání je u paměťových institucí výrazně vyšší.**

Průzkum DPE

- *Otázka 2: Máte či plánujete vytvořit důvěryhodné digitální úložiště? (tj. splňující kritéria vyjádřená v dokumentu „An Audit Checklist for the Certification of Trusted Digital Repositories“) (možnosti: ano, ne, zatím ne)*
- **Odpovědi: 30% národních knihoven ano, 9% ne a 61% ještě ne. Odpovědi archivů a vědeckovýzkumných institucí - shodné: 32% ano, 40% ne a 27% ještě ne.**
- **Závěr: důvěryhodných úložišť, která by obstála při mezinárodní certifikaci, je ve všech sledovaných skupinách zatím relativně málo. Větší posun v blízké budoucnosti lze zřejmě očekávat v oblasti národních knihoven.**

Průzkum DPE

- ***Otázka 3: Dlouhodobé uchovávání digitálních dokumentů je velmi složitý a náročný proces na to, aby jej realizovala jedna instituce. Vaše instituce v této oblasti spolupracuje/bude spolupracovat s ... (možnosti: paměťové instituce, vědeckovýzkumné instituce, producenti digitálních dokumentů, výrobci a dodavatelé SW, ostatní)***
- **Odpovědi: U všech sledovaných kategorií jsou nejvyhledávanějšími partnery paměťové instituce, mezi něž patří např. knihovny, muzea nebo archivy. Následují vědeckovýzkumné instituce a producenti digitálních dokumentů, pro vědeckovýzkumné instituce jsou vyhledávanou skupinou i producenti a dodavatelé SW.**
- **Závěr: Role paměťových institucí v oblasti dlouhodobého uchovávání digitálních dokumentů je/bude klíčová a vyžaduje důkladnou přípravu.**

Průzkum DPE

- ***Otázka 4: Vytvoření a provoz (správa) důvěryhodného digitálního úložiště je drahou a náročnou záležitostí. Budete vytvářet a spravovat úložiště ... (možnosti: jen pro danou instituci, ve spolupráci s dalšími institucemi).***
- **Odpovědi: Diference: pouze 15% národních knihoven počítá s vybudováním úložiště jen pro vlastní potřeby a 85% počítá s jeho sdílením s dalšími institucemi, 47% archivů počítá s vybudováním úložiště jen pro vlastní potřeby a 53% počítá s jeho sdílením s dalšími institucemi, u vědeckovýzkumných institucí je situace velmi podobná archivům – (48% pro vlastní potřeby a 52% sdílení).**
- **Závěr: Národní knihovny vycházejí z praktických zkušeností, kdy již v rámci uchovávání národního kulturního dědictví (v analogové i digitální podobě) uchovávají dokumenty publikované jinými institucemi.**

Průzkum DPE

- ***Otázka 5: Systém použitý pro vaše digitální úložiště bude... (možnosti: vyvinut ve vaší instituci, na bázi open source, komerční, kombinace, jiné řešení)***
- **Odpovědi: Diference: Národní knihovny, které nedisponují takovým množstvím programátorských a vývojových kapacit jako vědeckovýzkumné instituce, preferují kombinovaná řešení s vysokým podílem komerčních systémů (53%) ; vědeckovýzkumné instituce staví svá řešení převážně na bázi open source (52%). Překvapení: shodné zastoupení vlastního vývoje a komerčních systémů (28%) i open source (38%) u archivů.**
- **Závěr: [Výsledek další části Market and Technology Trends Analysis:] Existuje dobrý výběr řešení na bázi open source a malý výběr komerčních systémů pro správu digitálních úložišť, které lze jako „hotové“ (včetně modulu ochrany dokumentů) zakoupit na trhu. Minimální nabídka a konkurence a prozatím velmi malý trh jsou příčinou astronomických cen těchto komerčních produktů a představují značný problém pro dlouhodobou ochranu a zpřístupnění dokumentů zejména v paměťových institucích klíčového významu.**

Certifikace důvěryhodných úložišť

Prudký vývoj:

Nástroje pro externí audit:

- **Trusted Digital Repositories : Attributes and Responsibilities (RLG, OCLC, 2002) – obecný rámec http://www.rlg.org/legacy/longterm/repositories.pdf**
- **Trustworthy Repositories Audit & Certification : Criteria and Checklist (OCLC, CRL, 2007 – draft OCLC, NARA, 2005) TRAC http://www.crl.edu/content.asp?l1=13&l2=58&l3=162&l4=91**
- **Nestor Catalogue of Criteria for Trusted Digital Repositories (nestor, 2006) http://edoc.hu-berlin.de/series/nestor-materialien/8en/PDF/8en.pdf**
- **ISO ???**

Certifikace důvěryhodných úložišť

Nástroje pro interní audit:

- **Digital Repository Audit Method Based on Risk Assessment (DCC, DPE, 2007 – draft 1, testování) DRAMBORA http://www.repositoryaudit.eu**
- **Risk management, nejen konstatování, ale měřitelnost**
- **Jasný návod, které dokumenty je třeba shromáždit**
- **Hloubková analýza nejen technologického zabezpečení (omyl, že špičkový HW a SW stačí), ale i institucionální způsobilosti a trvalé udržitelnosti**
- **Příklady rizik a možností jejich eliminace**

Certifikace důvěryhodných úložišť

Společné (povinné) desatero pro jakékoli důvěryhodné úložiště (všichni hlavní aktéři, jaro 2007)

1. Zavazuje se k trvalé správě digitálních objektů pro předem vymezenou cílovou skupinu uživatelů.
2. Je schopno prokázat způsobilost splnit tento závazek (včetně prokázání finančního, personálního, strukturálního a procesního zajištění).
3. Shromažďuje a uchovává nezbytné smluvní a další legislativní dokumenty a plní povinnosti z nich plynoucí.
4. Má jasnou a účinnou strategii.
5. Získává a přijímá digitální objekty na základě předem stanovených kriterií, které odpovídají jeho závazkům a možnostem.

Certifikace důvěryhodných úložišť

6. Udržuje/zajišťuje integritu, autenticitu a využitelnost digitálních objektů, které trvale uchovává a udržuje.
7. Vytváří a udržuje potřebná metadata o všech akcích souvisejících s digitálními objekty v průběhu jejich uchování, stejně jako metadata o jejich vzniku, podmínkách zpřístupnění atd. vytvářená před uložením digitálních objektů.
8. Splňuje nutné požadavky související se zpřístupněním uložených digitálních objektů.
9. Má strategii pro plánování ochrany včetně záchranných akcí.
10. Má adekvátní technickou infrastrukturu pro trvalou správu a zabezpečení uložených digitálních objektů.

Certifikace důvěryhodných úložišť

Jak postupovat:

- **Interní audit (finále nebo příprava na externí audit – analýza, dokumentace), DRAMBORA – mezinárodní testování, ČR: NK ČR červen 2007 – verze 1, UK podzim 2007 – verze 2, konzultace, koordinace v rámci NDK, NK bude mít proškolené konzultanty, pasport úložišť v ČR? – lepší spolupráce, sdílení, zálohování, eliminace duplicit/bílých míst**
- **Externí audit – TRAC (???ISO???) – prozatím desetitisíce EUR, většinou velká národní úložiště – požadavek vydavatelů a dalších externích subjektů, NK v roce 2009-10 (interní audit může výrazně usnadnit a zlevnit externí audit)**

Národní digitální knihovna

Mezinárodní metadatové standardy pro digitální úložiště a vznikající národní standardy

Úvod

- **Problematika metadat využívaných v datových úložištích je již několik let velmi aktuální**
- **Závisí na nich, zda budeme schopni naše data využívat v budoucnu, zda budou přístupná, čitelná, autentická ...**
- **Tuto skutečnost si uvědomuje většina paměťových i jiných institucí ...**
- **... vědí, že je to z možných cest ta jednodušší, i když také velmi náročná**
- **Typy uchovávaných metadat budou u většiny institucí vycházet z obecného rámce OAIS a jeho tří „balíčků“ – SIP, AIP a DIP**
- **Většina organizací snažících se o dlouhodobé uchovávání digitálních dat, tento model více či méně implementuje**
- **Všechny dostupné SW systémy pro správu úložišť ho podporují, ať již SW komerční (DIAS, DPS – Digital Preservation System) nebo open source (DSpace, EPrints, Fedora aj.).**

METS + PREMIS

- Velmi populární je v této oblasti formát METS
  - Jde o „kontejner“, do kterého se dají zabalit různé metadatové formáty
  - dá se velmi dobře využít pro všechny tři balíčky (SIP, AIP, DIP) i na prezentaci složených dokumentů.
- Jedním ze zabalených formátů může být např. PREMIS
  - Vše záleží na volbě a potřebách organizace
  - Metadatový formát jehož cílem je vytvoření použitelné sady základních elementů metadat pro ochranu digitálních dokumentů ...
- Ochranná metadata = informace podporující a dokumentující proces ochrany digitálních dokumentů
  - provenienci
  - autenticitu
  - ochranné aktivity
  - technické prostředí
  - management práv

METS obecně

- **Samotný formát METS obsahuje vedle jiných i části, které jsou přímo metadatovým popisem daného objektu.**
  - **část popisných metadat *dmdSec***
  - **část administrativních metadat – *amdSec***
- **Tyto sekce jsou „naplněny“ již hotovým metadatovým popisem v určitém formátu/-tech**
- **Tyto dvě části (dmdSec a amdSec) mohou obsahovat metadata**
  - **popisná**
  - **administrativní**
  - **technická**
  - **ochranná metadata**
- **Formáty použité pro vyjádření těchto metadat jsou stejné jako ty, které je vhodné využít v digitálních úložištích...**

Formáty popisných metadat

- Doporučené metadatové formáty pro sekci popisných metadat „dmdSec“ jsou
  - Dublin Core
  - MARCXML (MARC21 zapsaný v XML podobě)
  - MODS aj.

- Kramerius - MARCXML i DC
- Popisná metadata ve formátu MARCXML se tvoří z konkrétních DTD jednotlivých dokumentů pomocí převodové tabulky.

Formáty administrativních metadat

Sekce administrativních metadat „amdSec“ má sama další 4 části

- **techMD – technická metadata**
- **rightsMD – administrativní případně legislativní práva k objektům**
- **sourceMD – popis původce údajů obsažených v METS dokumentu**
- **digiprovMD – metadata spojená s digitálními zdroji**

Formáty administrativních metadat 2

- **podzim 2006 analýza > řešení**

- **techMD**
  - **formát PREMIS (jeho část object)**
  - **formát MIX (NISO Metadata for Images in XML) specifický formát pro velmi podrobný popis metadat obrazového souboru - typ HW a SW na kterém vznikl, změny které v objektu proběhly, jeho vlastnosti atd.).**

- **rightsMD - vyjadřuje 2 druhy práv potřebných k zacházení s objektem**
  - **administrativní metadata - PREMIS (jeho část PREMIS – Rights)**
  - **legislativní práva - METSRights**

- **digiprovMD - zaznamenání událostí spojených s objektem**
  - **PREMIS – Events**

- **K našemu potěšení toto rozdělení bylo doporučeno jako nejvhodnější na workshopu věnovanému využívání formátu PREMIS ve Stockholmu v březnu 2007 a je podporováno i METS „redakční“ radou.**

Příklad - Kramerius

implementace formátu METS v systému Kramerius - zveřejňování metadat (Qbizm)

- **proběhla analýza formátu METS a dalších metadatových formátů, které nyní jsou (nebo naopak nejsou) využity ve finálním řešení**
  - **PREMIS**
  - **MODS - *Metadata Object Description Schema***
  - **MIX - *Metadata for Images in XML***
  - **MARCXML**
- **řešení bylo vytvářeno s přihlédnutím k podobným projektům ve světě (např. Göttingen nebo francouzský projekt Persee).**

Metadata v Krameriovi

- DTD v Krameriovi obsahuje hlavně popisná metadata
- potřeba do formátu METS pro Krameria doplnit další údaje o digitálních objektech (administrativní a technická metadata)
- proběhl výběr jednotlivých elementů v rámci formátu PREMIS, pro jednotlivé balíčky – SIP, AIP, DIP

- je přihlíženo k nárokům budoucího důvěryhodného úložiště NK a možnostem rozšíření těchto specifikací v případě potřeby
- tímto výběrem de facto vznikle národní formát, který si ale každá instituce bude moci doplnit
- v rámci Krameria je výběr elementů pro SIP a DIP v podstatě ukončen a připraven k připomínkování, AIP se bude finalizovat, až bude jasné jaký SW bude na úložiště nasazen a co dokáže

Příklady specifikace jednotlivých částí administrativních metadat (výběr)

- **technická metadata – techMD:**
  - **PREMIS object**
    - ObjectIdentifier ; PreservationLevel ; ObjectCategory ; ObjectCharacteristics ; CompositionLevel ; Fixity ; MessageDigest ; Size ; Format ; CreatingApplication ; DateCreatedByApplication ; Storage ; StorageMedium ; dependency ; software
- **metadata práv – rightsMD:**
  - **METSRights - legislativní práva (Intelectual Property Rights)**
    - RightsDeclaration ; RightsHolder ; RightsHolderName ; RightsHolderContactAddress ; RightsHolderContactEmail ; Context ; UserName ; Permissions ; Constraints ; ConstraintDescription
  - **PREMIS rights - administrativní práva**
    - permissionStatement ; linkingObject ; grantingAgent ; permissionGranted Act ; termOfGrant
- **metadata spojená s digitálními zdroji- digiprovMD:**
  - **PREMIS events**
    - eventIdentifier ; eventType ; eventDateTime ; eventDetail ; linkingAgentIdentifier ; linkingAgentRole ; linkingObjectIdentifier

Národní digitální knihovna

Příklad z Krameria

```
- <mets:dmdSec ID="DMD_MARC">
  - <mets:mdWrap MDTYPE="OTHER" OTHERMDTYPE="MARC" MIMETYPE="text/xml" LABEL="MARC XML">
    - <mets:xmlData>
      - <marc:collection>
        - <marc:record>
            <marc:leader>-----nas-a22-----uu-4500</marc:leader>
            <marc:controlfield tag="001">dtd20071000024</marc:controlfield>
            <marc:controlfield tag="003">CZ-PrNK</marc:controlfield>
            <marc:controlfield tag="005">070514161954</marc:controlfield>
            <marc:controlfield tag="007">070514b--------xx----------ger--</marc:controlfield>
          - <marc:datafield tag="022" ind1=" " ind2=" ">
              <marc:subfield code="a">1801-3627</marc:subfield>
            </marc:datafield>
          - <marc:datafield tag="245" ind1="1" ind2="0">
              <marc:subfield code="a">Egerer Jahrbuch</marc:subfield>
            </marc:datafield>
          - <marc:datafield tag="850" ind1=" " ind2=" ">
              <marc:subfield code="a">ABA000</marc:subfield>
            </marc:datafield>
          </marc:record>
        </marc:collection>
      </mets:xmlData>
    </mets:mdWrap>
  </mets:dmdSec>
```

Sumarizace hlavních vědecko-výzkumných cílů Digital Preservation

- vytvořena tzv. roadmapa - má ukázat další možné směry vývoje R&D v oblasti ochrany digitálních dokumentů
  - na základě analýz současného stavu
  - existujících projektů na globální úrovni
  - analýz stavu počítačové vědy, informačních a komunikačních technologií.
- Zpráva konstatovala, že „ani po 20ti letech výzkumů nemáme hmatatelný důkaz toho, že se naše znalosti, pozice a východiska v oblasti ochrany digitálních objektů nějak výrazně oproti době před 20ti lety zlepšily.“
- komunita okolo „digital preservation“ se problému ochrany digitálních objektů stále nezhostila odpovídajícím způsobem
- analýza ukázala, že se řešily mnohé okrajové záležitosti, zatímco podstatným problémům nebyla věnována adekvátní pozornost
- komunita neustále rozšiřovala šíři záběru výzkumů
- Jako komunita jsme, obrazně řečeno, zklamali. Potřebujeme ovšem pochopit proč.

Vzniklo hodně strategií, ale široce aplikovatelná řešení k problematice ochrany digitálních dokumentů spíše výjimkou než pravidlem.

Několik důvodů, které mohou být příčinou:

- **nedostatek všeobecného porozumění – nemáme přesnou definici ochrany digitálních dokumentů (digital preservation) a s ní spojených odpovědností**
- **odklon od tématu – zaměření výzkumů se rozšiřuje i do oblastí, které již jsou pokryty jinými obory**
- **nedostatek praktických zkušeností – velmi málo projektů našlo cestu do realizační fáze**
- **roztříštěnost – mnoho projektů se zabývá pouze částí problematiky**
- **zbytečná konkurence – soutěžení mezi institucemi a projekty v národním i mezinárodním měřítku zdržuje rozvoj. Existující řešení jsou ignorována a další proprietární jsou vytvářena**
- **překážky vyplývající z práv duševního vlastnictví**
- **nedostatečný trénink/školení**

Doporučené oblasti budoucího výzkumu 1

Obnova/restoration:

- navzdory dobré péči může být narušena integrita DO
- máme metody záchrany dat z fyz. média
- ALE co když ale není jasné, jakého typu je zachraňovaný objekt? Jak zjistíme, zda bit stream reprezentuje program, soubor nebo několik souborů?
- zobrazení těchto objektů a zpřístupnění jejich obsahu stále zůstává podstatnou výzvou ochrany digitálních objektů.

Uchování/conservation:

- Digitální sbírky potřebují být chráněny, aby byly původní zdroje přístupné i dalším generacím. Pro vypořádání se s problémem zastarávání technologií se objevily metody migrace, emulace a virtualizace. Máme s nimi ovšem velmi malé zkušenosti v reálném provozu.

Doporučené oblasti budoucího výzkumu 2

Management:

- **máme malé znalosti o tom, jak přistupovat k současným digitálním sbírkám s ohledem na zastarávání SW, HW a formátů souborů.**
- **jako nedílnou součást managementu úložiště je potřeba mít efektivní mechanismy, nástroje a podporu pro ně, monitoring a plánování ochrany.**

Riziko:

- **ochrana DO je v podstatě problém tzv. rizikového managementu. Potřeba nástroje, které by pomohly převést nejasnosti provázející ochranu digitálních dokumentů do měřitelných proměnných...**
- **...to by umožnilo přesné měření vyvážení nákladů a přínosů. Na snížení rizika v oblasti digital preservation lze využít výsledky výzkumů z počítačové vědy a business engineeringu, které zavádí pokročilé automatické techniky, jako kategorizace a strojové učení pro odhad momentálního rizika a výběr odpovídající strategie.**

Doporučené oblasti budoucího výzkumu 3

Typické vlastnosti digitálních objektů:

- **chránit DO znamená dobře znát jejich typické vlastnosti a ty zachovat dlouhodobě**
- **vlastnosti částečně závisí na formátu objektu, více na kontextu, ve kterém byl objekt vytvořen a je využíván. Aby byl digitální objekt užitečný pro budoucí uživatele, je třeba znát záměr, s kterým byl vytvořen.**

Interoperabilita:

- **pro ochranu obsahu jsou nyní důležité stálé business modely, standardizované formáty digitálních objektů a jejich metadat a standardizované výměnné protokoly.**
- **toto by mělo přispět k dosažení interoperability mezi různými generacemi systémů.**

Doporučené oblasti budoucího výzkumu 4

Automatizace:

- **úroveň automatizace v oblasti ochrany digitálních dokumentů není ještě dostatečná. Důvodem je fakt, že dosud skoro neexistoval systematický přístup k ochraně digitálních dat, pouze experimenty.**
- **aby bylo možné zpracovat obrovská kvanta objektů, musí být procesy ochrany automatizovány, např. hodnocení, získávání nebo extrakce metadat.**

Kontext:

- **kontext ochrany má velký dopad na způsob, jakým jsou DO organizovány a chráněny.**
- **aby mohl existovat dobrý management ochrany, musí být dobře pochopeno prostředí, kde se vše bude odehrávat (technologie, právní otázky, existující znalosti, nároky uživatelů atd.)**
- **kontext také souvisí se známými vlastnostmi s ním spojenými a s operacemi / událostmi prováděnými na objektu.**

Doporučené oblasti budoucího výzkumu 5

Uložení:

- **oblast ochrany digitálních dokumentů je silně vázána na tradice archivů a knihoven, zároveň si ale přisvojuje metody z blízkých disciplín**
- **např. vývoj ve vědeckých komunitách jako je GRID iniciativa může sloužit jako budoucí základ v distribuované infrastruktuře digitální ochrany. Řeší současné problémy se škálovatelností a kapacitou řešení úložišť.**

Experimentování:

- **pro zajištění budoucí přístupnosti DO bude experimentování hrát hlavní roli pro objektivní hodnocení ochranných strategií a výzkumů.**
- **zavedení testování umožní vývoj tzv. důkazních metod. Experimenty také pomohou pochopit interakci uživatelů s digitálním úložištěm a zhodnotit dopad nově podporovaných médií jako např. technika vizualizace informací pro sbírky.**

Shrnutí a doporučení

digital preservation europe

Stále potřebujeme hledat nový přístup k ochraně digitálních dokumentů, který by:

- **podporoval velké objemy dat**
- **dynamický a nestabilní digitální obsah (web)**
- **sledoval vývoj smyslu a využití kontextu digitálních obsahů**
- **podporoval integritu, autenticitu, dlouhodobou dostupnost a také modelové a svépomocné přístupy k ochraně.**

- **Výzkum a vývoj v oblasti digital preservation by měl zajistit reflexi a využití trendů v další generaci ICT infrastruktur a architektur.**

děkujeme za pozornost

http://www.digitalpreservationeurope.eu

http://www.ndk.cz/