# Dlouhodobá ochrana digitálních dat: co může vaše instituce udělat již dnes?

Mgr. Jan Hutař
Odbor digitální ochrany
NK ČR

AKM 2.12.2010

# Témata

- co řeší dlouhodobá ochrana digitálních dat a nač ji potřebujeme?
- v čem dlouhodobá ochrana digitálních dat spočívá? na co se soustředit?
- volně dostupné nástroje pro dlouhodobou ochranu
- dlouhodobá ochrana digitálních dat a projekt NDK
- VISK7 a nové postupy tvorby metadat a dat
- přehled dostupných nástrojů

# UNESCO o digitálních datech

Listina o ochraně digitálního dědictví, 15. říjen 2003

článek 1:

“*Digitální dědictví sestává z unikátních zdrojů lidského vědění. Mnohé z těchto zdrojů mají trvalou hodnotu a podstatu, čímž tvoří odkaz, který musí být opatrován a ochráněn pro současné i budoucí generace*.”

# Co řeší dlouhodobá ochrana digitálních dat?

- ochranu dat? ochranu vložených financí? ochranu vědění? ochranu děl lidského umu?
  - dostupnost, použitelnost, srozumitelnost v budoucnu
- jak „dlouhá“ je budoucnost?
  - dokud budou mít uchovávané digitální objekty význam pro uživatele
  - 5 let nebo déle x časový úsek, ve kterém dojde k podstatné změně technologií (podpora medií a formátů)

“Pokud mluvíme o ochraně digitálních zdrojů, **termín „dlouhodobá“ neznamená garantované uchování na 5 nebo 50 let**, jako spíše **odpovědný vývoj strategií, které se dokáží vyrovnat s neustálými změnami**, které přináší informační trh.”

Ute Schwens / Hans Liegmann (DNB/nestor)

# Co má dlouhodobá ochrana dat zajistit?

- základní ochrana dat nyní i v budoucnu (**ochrana bit streamu**)
- ochrana použitelnosti dat v budoucnu (**ochrana logická**)
- logická ochrana = procesy zajišťující, aby digitální objekty zůstaly v budoucnosti:
  - vyhledatelné,
  - přístupné (zobrazitelné),
  - využitelné znovu a znovu a
  - pochopitelné (obsah a smysl)
  - autentické

OAIS (ISO 14721:2003 – Open Archival Information System)

– s jednotlivými digitálními objekty musí být uchován nejen **informační obsah uchovávaných objektů**, ale také **další informace o původu a historii změn dokumentu, o jeho kontextu a zdrojích potřebných k porozumění**

# V čem ochrana spočívá?

- **digitalizace** – vytváříme to co chceme? jsou formáty standardní máme kompletní dokument? máme kompletní metadata? máme kontrolní součty?
- **přesuny dat** – kontrolní součty, kompletnost, plán
- **uložení** – kontrola HW, kontrola integrity, přehled co kde je (data a metadata), přehled o změnách, práva přístupu, opatření dlouhodobé ochrany
- **opravy dat a metadat** – kdo, co, kdy, proč a s jakým výsledkem; úpravy metadat
- **zpřístupnění** – autenticita, použitelnost, vyhledávání

# Digitalizace – co promyslet?

- strategická příprava – standardy + jasná strategie projektu/instituce
- víme opravdu co a proč budeme skenovat? kvalita předlohy, existence metadat
- víme jak to budeme uchovávat a zpřístupňovat? kolik to bude stát?
- máme kapacity na uložení dat? zkušenosti a lidi pro manipulaci s daty?
- NEZAČÍNAT bez plánu - MÁME NA TO (TEĎ) PENÍZE… raději naopak
- dělat kontroly integrity (kontrolní součty)
- validace metadat – formáty, kompletnost

co hrozí?

- zdržení, hromadění nehotových dat, ztráty dat, vícepráce/náklady

# Jak pomohou metadata?

většina metadat vzniká během digitalizace x bezprostředně po ní – mohou uchovat vše důležité pro budoucnost – pokud to podstatné do nich dáme

- **metadata popisná (MARC, MODS, DC, EAD)**
  - popis intelektuální entity
  - ideálně použít bibl. záznam z katalogu (konzistence katalog a dig. knihovna)
  - obohacení – popis vnitřních částí – nejsou v katalogu apod.
- **administrativní metadata (PREMIS, MIX, METS)**
  - technická m. – údaje o formátech,  validacích, skeneru, SW, identifikátory …
  - metadata práv – údaje o copyrightu, licencích aj.
  - metadata o provenienci – vztahy mezi objekty, událostí, agenti
- **strukturální metadata (METS)**
  - logická a fyzická struktura dokumentu

- **volné nástroje PRONOM, JHOVE aj.**

# Přesuny dat a metadat

do repozitáře; do aplikace zpřístupnění; na nový HW

- jakýkoliv přesun je riskantní

**rizika přesunů**

- ztráty existujících vazeb, narušení zaběhaných procesů, časově náročný proces, ztráta integrity

**nutno provádět**

- kontroly integrity a kompletnosti dat
- antivirová kontrola
- validace struktury balíku dat a metadat
- identifikace/ validace formátů před a po

# Metadata a data v repozitáři

- monitorování základních vlastností a metadat vkládaného materiálu > hodnocení risků
- obohacení metadat
- doplnění metadat dokumentujících životní cyklus v archivu (použití, exporty, ochranné operace, atd.)
- neustálá kontrola integrity dat
- preservation planning
- jednoduché vyhledávání, filtrování – víme co kde je

- konzistence metadat mezi archivem, zpřístupněním a katalogem? ano x ne

# S čím můžete začít již dnes?

- za všech okolností následovat aktuální standardy
- mít strategii ochrany digitálních dokumentů pro vaši instituci
- dokumentace procesů
- mít spolehlivý systém na správu repozitáře a dat
- provádět kontroly při jakémkoliv přenosu dat i metadat
- využívat volně dostupné nástroje
  - lze využívat okamžitě
  - zvláště pokud máte velké objemy dat, často s nimi manipulujete
  - výrazně zlepší vyhlídky do budoucna
  - není třeba čekat na nějaké řešení „shůry“
- připravit se již během digitalizace (metadata)

# Strategie ochrany aneb víme co děláme

obecná strategie institucionální > strategie ochrany dlouhodobá > střednědobá > **projektová**

- co chceme ochraňovat? a proč? všechno stejně?
- máme kapacity? personál? znalosti?
- víme kolik bude stát uložení? dlouhodobé uložení?
- spolupráce/koordinace = sdílení financí > sdílení technologií > sdílení znalostí
- sdílení NENÍ využívání a spoléhání se na druhého

# Volně dostupné nástroje

- identifikace, validace a charakterizace  formátů - PRONOM, UDFR
- metadata extraktory - JHOVE, NZME aj.
- nástroje na vytváření metadat, migrace, validace metadat
- open source SW na repozitáře
  - správa dat a metadat, událostí, práv (přístup i copyright) aj.
  - vyhledávání + hromadné operace s daty/metadaty
  - zajištění integrity
  - preservation planning ano x ne
- preservation planning
  - PLATO nebo PLANETS testbed

# Externí služby

- využití externích služeb je klíčovou vlastností SW repozitáře
  - z nich čerpají podstatné informace např. o formátech dokumentů (identifikace, charakterizace)
  - validují je oproti těmto službám
  - získávají z nich informace o doporučovaných možnostech migrací i stavu zastarání formátů

- registry formátů - PRONOM (DROID), UDFR
- extraktory/validároty Jhove2, New Zealand Metadata Extractor a
- open source i komerční SW repozitáře s nimi musí pracovat

# Registry formátů – use case

- **Identifikace**
  mám digitální objekt, co je to za formát?
- **Validace**
  mám objekt, který říká, že je to formát F – je to opravdu ono?
- **Transformace**
  mám objekt ve formátu F, ale potřebuji formát G - jak ho mohu vytvořit?
- **Charakterizace**
  mám objekt ve formátu F, jaké má vlastnosti?
- **Odhad risků**
  mám objekt ve formátu F, je s ním spojen risk? např. zastarávání?
- **Zobrazení**
  mám objekt ve formátu F, jak a čím ho mohu zobrazit?

(Abrams, Seaman: Towards a global digital format registry. IFLA 2003)

# Dlouhodobá ochrana digitálních dat a projekt NDK

- jeden ze tří hlavních cílů projektu
- ochrana pro stávající data, nová zdigitalizovaná i digital born
- návrh nových procesů a standardů digitalizace
  - východisko pro VISK7, krajské digitalizace
  - pořízení SW na kompletní workflow digitalizace
- pořízení komerčního LTP systému (Rosetta x SDB)
  - strategie ochrany
  - velké objemy dat z digitalizace nelze uchovávat ve file systému
  - trend ve světě (NZ, Austrálie, Německo, Holandsko, Finsko, Norsko, Polsko, Maďarsko)
  - důvod – ochrana dat, investic

# VISK7 a plánované změny

- nové standardy dat i metadat
  - návaznost na NDK
  - využití standardů běžných v okolních zemích
- VISK7 v roce 2011 = 2 možnosti tvorby dat
  1. stávající formáty dat i metadat (DTD monografie a periodika)
  2. nové formáty dat (JPEG2000 a ALTO XML)
- VISK7 v roce 2012
  - stávající postup?
  - nové metadatové formáty dle NDK (METS, PREMIS, MIX, MODS)
  - nové formáty dat – JPEG2000, ALTO XML
  - nová struktura balíků
- co je smyslem změn? interoperabilita, flexibilita, komunitní podpora namísto vlastního vývoje …

# Dlouhodobá ochrana digitálních dat ve světě

řešena převážně v knihovnách a archivech

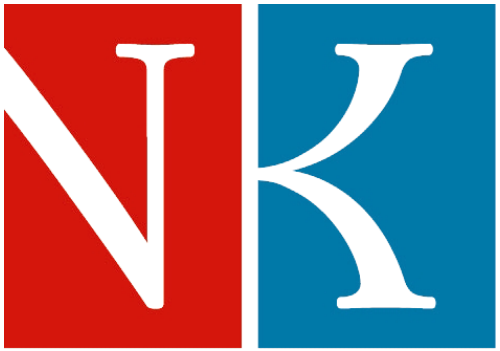
- Nový Zéland (NK a NA)
- Německo (NK)
- Velká Británie (NK, NA a další knihovny a nadace)
- Nizozemí (NK a NA)
- USA (NK, NA, univerzitní knihovny, nadace aj.)
- přidávají se ostatní evropské a světové knihovny (Finsko, Norsko, Singapur, Austrálie, Francie, Estonsko, Polsko, SK, Maďarsko atd. atd) – probíhající tendry

**projektová podpora EU**

- DigitalPreservationEurope
- CASPAR
- PLANETS
- SHAMAN
- KEEP aj.

# Dostupné produkty

- komerční
  - Safety Deposit Box (fa Tessella UK)
  - Rosetta (fa ExLibris, Izrael)
  - DIAS (fa IBM) – systém nemá další vývoj v oblasti knihoven
- open source
  - Fedora a její nadstavby
  - XENA (NA Austrálie) http://xena.sourceforge.net/
  - RODA (Portugalsko, Uni of Minho) http://tinyurl.com/3ynyzs6
  - CRIB (Portugalsko, předchůdce RODA)
  - ARCHIVEMATICA http://archivematica.org/ (Unesco + Kanada); tool pack
  - MOPSEUS – založeno na Fedoře, Řecko
  - HOPPLA – vývoj TUW Vídeň, pro malé instituce nebo domácnosti
  - ePRINTS – Univerzita Southampton
- nástroje na preservation planning
  - PLANETS testbed, PLATO aj.



# Děkuji za pozornost

# Otázky?

jan.hutar@nkp.cz

www.ndk.cz