Plánování dlouhodobé ochrany pomocí nástroje PLATO Preservation Planning Tool

Podklad pro závěrečnou zprávu grantu Výzkum a vývoj

Mgr. Andrea Fojtu

E-mail • andrea.fojtu@nkp.cz

Affiliace • Odbor digitálních fondů (ODF)
Národní knihovna ČR (http://www.nkp.cz/)

Praha, září 2011
1. ÚVOD .................................................................................................................................................. 3
2. DLOUHODOBÁ OCHRANA .............................................................................................................. 5
3. SIGNIFIKANTNÍ VLASTNOSTI ....................................................................................................... 6
  3.1 KATEGORIE .................................................................................................................................. ERROR! BOOKMARK NOT DEFINED.
  3.2 POŽADAVEK .................................................................................................................................. ERROR! BOOKMARK NOT DEFINED.
  3.3 SIGNIFIKANTNOST ...................................................................................................................... ERROR! BOOKMARK NOT DEFINED.
4. POLITIKY A MYŠLENKOVÉ MAPY ................................................................................................ ERROR! BOOKMARK NOT DEFINED.
5. NÁSTROJE PRO DLOUHODOBOU OCHRANU ............................................................................ 12
  5.1 PLATO .............................................................................................................................................. 12
  5.1.1 Nově implementované prvky ve verzi 3.0 .................................................................................. 14
  5.1.2 Práce s nástrojem Plato ............................................................................................................. 15
  5.1.3 Výstupy z nástroje Plato ............................................................................................................. 18
6. NÁVRH ZAPOJENÍ NÁSTROJE PLATO DO PROSTŘEDÍ NK NEBO PODOBNÉ INSTITUCE ERROR! BOOKMARK NOT DEFINED.
7. ZÁVĚR ............................................................................................................................................... 20
8. SEZNAM BIBLIOGRAFICKÝCH ODKAZŮ ....................................................................................... 22
10. PŘÍLOHA Č. 1 - MYŠLENKOVÁ MAPA JPG .................................................................................. 24
12. PŘÍLOHA Č. 2 - MYŠLENKOVÁ MAPA JP2 ................................................................................... 25
13. PŘÍLOHA Č. 3 - MYŠLENKOVÁ MAPA TXT .................................................................................. 26
1. ÚVOD

Není tomu tak dávno, kdy ochrana digitálních dokumentů byla pojímaná čistě z technokratického hlediska, tedy jako “změna” dokumentů pomocí kopírování datových toků, migrace, obnovení (z angl. refreshing), přesunu dat na média (CD, DVD a pod.), analogového zálohování (např. vytíštěním informačních zdrojů na papír), emulace (napodobení systému zpracování údajů nebo jeho části jiným systémem nebo technickými prostředky) a v neposlední řadě UVC (z angl. Universal Virtual Computer)\(^1\).

Často publikovaným názorem je, že dlouhodobou ochranu dokumentů je možné vyředit pouze metadatovým popisem a uložením souborů na (optická či pevná) média. Není tomu však tak. V dnešní době již problematika dlouhodobé ochrany digitálních dat není pouze otázkou správně zvolených technologií. Zahrnuje i celou řadu dalších problematických okruhů - od organizace a řízení, kvalifikovaného personálu až po finanční stránku.


Cílem výzkumu v roce 2011 bylo zaměřit se na jednu z fází životního cyklu dlouhodobé ochrany digitálních dokumntů - migraci formátů - a otestování migrace na vzorových datech nejčastěji zastoupených formátů souborů budoucí Národní digitální knihovny ČR v nástroji PLATO. Samotnému testování však musí předcházet stanovení signifikantních vlastností u testovaných typů objektů. Tyto, společně s myšlenkovými mapami (v příloze), díky kterým lze signifikantní

vlastnosti importovat do nástroje PLATO, budou tvořit teoretickou část podkladu. Praktickou složku výzkumu reprezentují závěry z PLATO. Součástí zprávy jsou zkušenosti z konkrétní práce s nástrojem PLATO, včetně návrhu zapojení PLATO do NDK.
DLOUHODOBÁ OCHRANA

Obecně existují 4 základní přístupy k dlouhodobé ochraně digitálních dokumentů (Wilson, 2007):
1. technokratický (z angl. techno-centric) - ochrana HW a SW po co nejdéle možnou dobu
2. datový (z angl. data-centric) - ochrana dat na úkor originální aplikace, např. konverze dokumentů do PDF
3. procesní (z angl. process-centric) - ochrana originální aplikace a procesů, např. emulace starého HW nebo SW
4. “post hoc”(z angl post hoc) - bez aktivní ochrany; v případě potřeby data “archeologickými” metodami vyfiltrovat.

Je dobré si uvědomit, že cílem dlouhodobé ochrany není uchovávat všechno, nýbrž stanovit pravidla a podmínky za jakých se dokumenty budou uchovávat.

Obr. 1 Politika digitální ochrany

---

Při vytváření plánu, resp. interní politiky dlouhodobé ochrany digitálních dokumentů je vhodné využít postup dle Obr. 1. Na začátku je potřebné zmapovat výchozí situaci repozitáře, pokračovat výběrem nejvhodnějších formátů, postupů, SW, HW, testováním vybraných řešení až po audit a eventuálně certifikaci (např. metodikou TRAC).

2. SIGNIFIKANTNÍ VLASTNOSTI

Termín signifikantní vlastnosti v kontextu dlouhodobé ochrany digitálních dat (z angl. significant properties, significant characteristics, transformational information properties nebo také essence) byl poprvé zaveden v projektu CEDARS (http://www.ukoln.ac.uk/services/elib/projects/cedars/).

Vyjadřuje důležité charakteristiky (komponenty) informačního objektu, které mají být uchovány tak, aby jeho použitelnost, přístupnost a srozumitelnost byla zabezpečena i do budoucna (Wilson, 2007). Správci sbírek můžou využít i formálního vyjádření signifikantních vlastností pro výběr ochranných technik a načasování různých ochranných akcí. Ty jsou důležité pro stanovení klíčových charakteristik souborů, které instituce hodlá uchovávat do budoucna. V nástroji PLATO mají signifikantní vlastnosti podobu myšlenkových map (viz Příloha č.1, 2 a 3).

Signifikantnost má několik základních aspektů (Knight, 2009):
1. relativita (není univerzální a neměnná)
2. závislost (na cílech a aplikovaných kritériích)
3. interpretace (dle cílové skupiny a její znalostní bázi apod.).

Formální vyjádření významných vlastností je rekurzivní; mnoho z vlastností lze dále dělit na dílčí vlastnosti. Vlastnost písmo, například, může být dále rozděleno na dílčí vlastnosti, jako je typ písma, styl, velikosti nebo barva. Pokud některé z těchto dílčích vlastností pozmění vzhled nebo význam digitálního objektu, mohou být považovány za důležité a stojí za to je zachovávat v přesné shodě jako u originálního objektu. Je dobré si uvědomit, že například jen formát konverze může změnit strukturu dokumentu, jeho stylistické rysy, navigaci či vzhled.
Výběr typu objektu pro analýzu

Analýza struktury

Identifikace účelu technických vlastností

Určení očekávaného chování

Klasifikování chování podle funkce

Asociace vlastností ke každé funkci

Vyhodnocení & finalizace
Stanovování signifikantních vlastností však není podchyceno jednotnou metodikou³.
Koncepy signifikantních informací se do určité míry překrývají s reprezentacemi informací. Signifikantní vlastnosti představují charakteristiky abstraktního informačního objektu, naproti tomu reprezentace informací (strukturalní a sémantické informace) indikují charakteristiky objektu z pohledu dat, jakou je např. kódování (Brown, 2008).

Základní požadavky na dlouhodobou ochranu, a vlastně zachování pro nás podstatných (signifikantních) vlastností digitálních objektů, je možné rozdělit do pěti základních okruhů, resp. kategorií: kontext (z angl. context), obsah (z angl. content), chování (angl. behavior), vzhled (z angl. appearance) a struktura (z angl. structure). Z nich se poté “extrahuji” signifikantní vlastnosti (viz myšlenkové mapy v příloze).

Ve výsledku pak například požadavky na dlouhodobou ochranu textového dokumentu a (slovně) vyjádření jeho signifikantních vlastností mohou vypadat následovně:

<table>
<thead>
<tr>
<th>Kontext</th>
<th>Obsah</th>
<th>Struktura</th>
<th>Vzhled</th>
<th>Chování</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organizace a kontext (např. název organizace, datum, vazba na jiné dokumenty apod.)</td>
<td>Vesky objekt musí být uchován včetně záhlaví, zápatí, poznámek pod čarou, obsahu (souhrn) a vlastností dokumentu</td>
<td>Struktura dokumentu musí být uchována tak, aby reprezentovala logické vazby mezi jednotlivými součástmi dokumentu</td>
<td>Vzhled dokumentu a uchovávané verze nemusí být identické, nicméně “nový” vzhled nesmí ovlivnit pochopení originálního významu</td>
<td>Aktivní chování, úprava, zaznamenání původních formách chování musí být zaznamenáno</td>
</tr>
<tr>
<td>Uchování logových souborů s informacemi o preservacích, akcích, o originálů a současném formátu souboru, název a verzi HW, SW, OS</td>
<td>Prostý text musí být vždy čitelný</td>
<td>Pokud kontext, obsah a struktura podle kapitol, ods Shanghai, ale též správná pozice vložených poznámek, poznámek pod čarou a obrázků.</td>
<td>Popis aktivních okázů musí být uchován</td>
<td></td>
</tr>
</tbody>
</table>

Tab. 1. Požadavky a signifikantní vlastnosti textového dokumentu (podle Slats, 2004)

---

Ty je dále možné “přetavit” do podoby myšlenkové mapy (viz Příloha č. 1 až 3) nebo inventářu signifikantních vlastností pro jednotlivé formáty.

Výběru signifikantních vlastností by měl předcházet výběr samotných zdrojů (nebo jejich typů) k uchovávání, pro který se v anglo-amerických kruzích ustálil termín “appraisal”. Do češtiny by se dal volně přeložit jako (kontrolovaný) výběr. Ten představuje proces evaluace aktivit, kterým se určuje jaké dokumenty a na jak dlouho mají být uchovány, při respektování organizačních pravidel a očekávání uživatelské komunity. Základní stavební kameny kontrolovaného výběru představují komplexní analýzu dokumentů, hodnocení aktivit a jejich právních omezení, určení realnosti uchování dokumentů a konečné rozhodnutí.


Nicméně nejlepší řešení se jeví jako kombinace náhodného výběru a kontrolovaného výběru (společně s řízeným opomenutím) pomocí tzv. 3-stupňové strategie. Tato se častečně opírá i o “appraisal” jako jeden z prostředků poskytující autentický a realistický pohled na minulost. Jeho součástí je (Neumeyer, 2007):

1. náhodný výběr
2. manuální/tradiční “appraisal” za účelem zachování kompletnosti
3. přístupová omezení z důvodu ochrany autorských práv.

Analýza užitků (z angl. utility analysis) poskytuje managerům informace pro vyhodnocení finančního dopadu (včetně ROI) v případě určité intervence (Delos, 2005). Tato analýza byla původně vyvinuta pro ekonomickou oblast. V ekonomice je často velmi obtížné stanovit a kvantifikovat přínosy a užitky projektu a následně je vyjádřit v peněžních jednotkách, neboť celá řada efektů plynoucích z investice je nefinanční a někdy dokonce nehmotné povahy. Pro tyto případy se ocenitelné náklady a přínosy

---


převádí na finanční toky pomocí různých metod, které umožňují tyto nefinanční náklady a přínosy ocenit (Ekonomická, cca2010).

Obr. 3 Workflow analýzy užitků (z angl. “utility analysis”) (podle Delos, 2005)

Tento koncept (viz Obr. 3) je však možné aplikovat i na oblast digitální ochrany, kdy obecné strategie (emulace a migrace) je možné rozdělit na kličové charakteristiky (modifikace originálního souboru, náklady při aplikování modifikovaného souboru do praxe apod.) a poté je, ke zvolení optimálního řešení, převést na kvantitativní rozhodovací kritéria.

PLATO Preservation Planning Tool stavit svoji funkcionalitu právě na analýze užitků. Hlavním principem webového nástroje PLATO je vyhodnocování ochranných aktivit podle stanovených cílů, signifikantních vlastností a požadavků na jejich zachování.

Podle doporučení Florida Digital Archive (viz Tabulka 1 – Seznam doporučených formátů souborů pro dlouhodobou ochranu), nejsou proprietární formáty vhodné pro dlouhodobou ochranu dokumentů (např. textové dokumenty ve formě šifrovaných PDF či souborech Microsoft Word).

<table>
<thead>
<tr>
<th>Médium</th>
<th>Nejvyšší stupeň jistoty</th>
<th>Střední stupeň jistoty</th>
<th>Nejnižší stupeň jistoty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text</td>
<td>- Plain text (kódování: US-ASCII, UTF-8, UTF-16 s BOM)</td>
<td>- Cascading Style Sheets (*.css)</td>
<td>- PDF (*.pdf) (zašifrovaný)</td>
</tr>
<tr>
<td></td>
<td>- XML (včetně XSD/XSL/XTML, atd.; s vloženým nebo přístupným schématem a explicitně</td>
<td>- DTD (*.dtd)</td>
<td>- Microsoft Word (*.doc)</td>
</tr>
<tr>
<td></td>
<td>specifikovaným kódováním znaků)</td>
<td>- Plain text (ISO 8859-1 kódování)</td>
<td>- WordPerfect (*.wpd)</td>
</tr>
<tr>
<td></td>
<td>- PDF (*.pdf) (vložené fonty)</td>
<td>- PDF (*.pdf)</td>
<td>- DVI (*.dvi)</td>
</tr>
<tr>
<td></td>
<td>- Rich Text Format 1.x (*.rtf)</td>
<td>- Rich Text Format 1.x (*.rtf)</td>
<td>- ostatní textové formáty, které nejsou specifikovány</td>
</tr>
<tr>
<td></td>
<td>- HTML (včetně deklarace DOCTYPE)</td>
<td>- SGML (*.sgml)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- Open Office (<em>.sxw/</em>.odt)</td>
<td>- ODMXML (ISO/IEC DIS 29500) (*.doc)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>- PDF/A-1 (ISO 19005-1) (*.pdf)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------------------</td>
<td>-----------------------</td>
<td>----------------------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>Prezentace</td>
<td>OpenOffice (* .sxw/* .odp)</td>
<td>OOXML (ISO/IEC DIS 29500) (* .pptx)</td>
<td>PowerPoint (* .ppt)</td>
</tr>
</tbody>
</table>

Tab. 2 Seznam doporučených formátů souborů pro dlouhodobou ochranu (podle Florida Digital Archive)

Díky své netransparentnosti a navázanosti na komerční subjekt přetrvává velké riziko ztráty informací (autentický, srozumitelný obsah). Proto je žádoucí proprietární formáty souborů migrovat do otevřenějších protějších při zachování jejich signifikantních vlastností.

Při testování migrací v nástroji PLATO (The Preservation Planning Tool) jsme identifikovali signifikantní prvky nejčastěji zastoupených objektů budoucí Národní digitální knihovny. Ide o JPG (Příloha č. 1 - Myšlenková mapa JGP), JP2 (Příloha č. 2 - Myšlenková mapa JP2), TXT (Příloha č. 3 - Myšlenková mapa TXT), PDF.

Výstupy z prvních dvou zminěných formátů jsou dostupné v online nástroji PLATO. Formát TXT však nebylo možné vhodně otestovat (viz podrobněji část "Výstupy z PLATO").

Z myšlenkových map pro JPG (Příloha č. 1) a JP2 (Příloha č. 2) je vidět, že jsme signifikantní vlastnosti zaměřili na 7 základních oblasti - charakteristiky objektu (obsah, kontext), formát objektu, čas, procesy, náklady, charakteristiky budoucího formátu a migrační agent. Charakteristiky chování a struktura (tak jak jsou uvedeny v Tab. 1. Požadavky a signifikantní vlastnosti textového dokumentu) nejsou na obrazové formáty JPG a JP2 aplikovatelné. Kritérium, resp. charakteristika vzhledu není v
myšlenkové mapě samostatně zastoupená v uzlu, jelikož identičnost objektů nám v tomto případě zabezpečí uzel s kritérii objektových charakteristik.

Signifikantní vlastnosti ve formě myšlenkové mapy u typů objektů ve formátu TXT (Příloha č. 3) jsou začleněny do 4 výchozích kategorií - charakteristiky objektu (obsah, kontext), náklady, charakteristiky budoucího formátu a migrační agent. Charakteristiky chování, vzhledu a struktury (tak jak jsou uvedeny v Tab. 1 Požadavky a signifikantní vlastnosti textového dokumentu) nejsou pro čistý text (z angl. plain text) využitelné.

3. NÁSTROJE PRO DLOUHODOBOU OCHRANU

V současné době vznikají open source nástroje pro dlouhodobou ochranu digitálních dat doslova jako „houby po dešti“. Některé jsou vyvijeny za účelem jejich integrace do stávajících systémů institucí či cílových skupin (např. CRiB, Hoppla), jiné fungují jako první prototypy open source digitálních repozitářů odpovídajících referenčnímu modelu OAIS a plně podporujících životní cyklus dlouhodobé ochrany digitálních dat (např. Mopseus, Archivematica, RODA). Samostatnou skupinu nástrojů tvoří metodologie (online, offline) sledování, evaluace a hodnocení činnosti repozitáře, včetně plánování dlouhodobé ochrany dat.


3.1 PLATO

http://www.ifstuwien.ac.at/dp/plato

Garant: Department of Software Technology and Interactive Systems, Vienna University of Technology (http://www.ifstuwien.ac.at/)

Instalace: online nástroj

Požadavky: otevřené porty 8080 a 8443

Poslední verze: 3.0.1 (říjen 2010) - drobné úpravy verze 3.0 a doplnění dokumentace
Plány pro další verze: v rámci evropského projektu SCAPE (SCAable Preservation Environments) je plánován další rozvoj nástroje PLATO

Kontakt: Christoph Becker, Hannes Kulovits, Michael Kraxner, Andreas Rauber {becker, kulovits, kraxner, rauber}@ifs.tuwien.ac.at

Implementace pro NK ČR: doporučuje se v celém rozsahu nástroje

PLATO podporuje rozhodování v oblasti plánování dlouhodobé ochrany digitálních dat a výběru nejvhodnějšího formátu pro budoucí migrace. Nejhodnější formát je vybírán na základě vyhodnocení stanovených důležitých vlastností souborů (resp. objektů), které mají být zachovány i do budoucně (např. zalamování řádků a čislování stránek u textových souborů). Tento nástroj vychází z referenčního modelu OAIS, konkrétně z jeho funkcionality "Plánování dlouhodobé ochrany". Jde o online softwarovou aplikaci (licencovanou jako CC-GNU LGPL) pro evaluaci potenciálních ochranných řešení a strategií.

Celý proces evaluace je rozdělen na 3 fáze (podrobněji v části: Práce s nástrojem PLATO):

1. **Definice požadavků** – pozůstává ze specifikace a podrobnějšího popisu sbírky (obsahu dat), která je vybrána pro naplánování ochranných akcí. Tento krok je obdoba první fáze auditu DRAMBORA. Následně jsou vybrána konkrétní vzorová data, u kterých se podrobně stanoví požadavky na dlouhodobou ochranu. Autoři nástroje k tomuto účelu doporučují zejména open source nástroj FreeMind (http://freemind.sourceforge.net/wiki/index.php/Main_Page) pro vytváření myšlenkových map. Jde o aplikaci napsanou v Java jazyce, dostupnou pro všechny nejrozšířenější operační systémy (platformy MS Windows, Mac OS X, Linux, eComStation a j.).

Základním hodnotícím kritériem je obvykle soubor nebo záznam sbírky.

2. **Hodnocení alternativ** – hodnotící kritéria z předchozí fáze jsou podkladem pro provádění experimentálních migračních aktivit, které přicházejí do úvahy.

3. **Posouzení výsledků** - experimentální migrace jsou výstupem pro následnou analýzu a vyhodnocení nejlepšího formátu dané vzorové sbírky digitálních dat.

Výstupem je dokument s objektivním posouzením nejrůznějších migračních alternativ dle specifických požadavků dané instituce. Proto se implementace nástroje do workflow dlouhodobé ochrany digitálních dat doporučuje v plném rozsahu.
3.1.1 NOVĚ IMPLEMENTOVANÉ PRVKY VE VERZI 3

K dispozici je několik drobných vylepšení implementovaných napříč nástrojem PLATO. Často jsou výsledkem zpětné vazby od uživatelů, včetně stahování nových “stromů požadavků” (z angl. requirements tree), které je možné dále upravovat v nástroji Freemind (viz výše 1. Definice požadavků). Tuto možnost jsme zvolili i v NK ČR, zkombinovali jsme tak vlastní znalosti se zkušenostmi jiných uživatelů a odborníků.

**Automatizované měření** - vyvinutý rozšiřitelný rámec pro připojení měřitelných rozhodujících kritérií pro měřitelné vlastnosti a metriky v šesti kategoriích:

1. **výstupní objekt** (z angl. object outcome) a jeho požadované vlastnosti (např. zachování možnosti editace nebo šířky obrazu).
2. **výstupní formát** (z angl. outcome format) spolu s kritérii na formát objektu, ve kterém má být uchováván (např. standardizace).
3. **důsledky** (z angl. outcome effects) vyvolané danou ochrannou akcí (např. náklady)
4. **runtime** (z angl. action runtime) potřebný k provedení akce (např. spotřebovaný čas a paměť)
5. **vlastnosti** (z angl. action static): náklady na licence a kvalita dokumentace konkrétního nástroje.
6. **posouzení** (z angl. action judgement) vlastnosti, jakou je např. použitelnost.

**Rychlé vyhodnocení** - workflow rychlého vyhodnocení, který vyplývá z několika základních předpokladů pro rychlé vyhodnocení možných řešení v pouhých třech krocích.

Vzdálená emulace je integrována do PLATO (v rámci “Planet service registry”⁶) a zprostředkována přes službu GRATE, jež běží na samostatném serveru.

**Integrace P2**⁷ - pro zvýšení počtu alternativních akcí a umožnění automatické evaluace formátů pomocí měřicího rozhraní bylo přidáno 44 tisíc údajů o formátech, které jsou popsány v registru P2. Umožňuje dotazovacím nástrojům (z angl. querying tool) testovací objekty převádět do daného formátu i přesto, že nejsou obsaženy v Service Registry.

---

⁶ Service registry je poměrně nekompletní; kdokoliv, kdo hledá nástroj pro konverzi objektů též požaduje jiné zdroje potenciálních nástrojů. Jedním z řešení bylo přímé prohledávání Planets Service Registry; ten ale bohužel není dostupný na žádné veřejné adrese (Prom, 2010).

Přednastavení pro ePrints (http://www.eprints.org)⁸ - repozitář je schopen realizovat ochranný plán vytvořený v PLATO.
Integrovaný FITS⁹ a vylepšená znalostní báze.

3.1.2 PRÁCE S NÁSTROJEM PLATO

Při přípravě nového plánu dlouhodobé ochrany digitálních dokumentů pomocí nástroje PLATO je nutná posloupnost kroků (viz Obr. 3). Jako konkrétní příklad postupu je vybrán formát JPG, resp. zdigitalizované dokumenty z projektu Kramerius. Po ostatní formáty (např. JP2, PDF, TXT, PDF) by byl postup obdobný.

1.1. Definování požadavků - tj. sesbírání veškeré dostupné dokumentace popisující akce a procedury (včetně nejrůznějších omezení) pro vybrané dokumenty/sbírky. V našem případě se jedná o výběr tří JPG dokumentů ze sbírky zdigitalizovaných novin (od roku 1999).

Cílem je otestovat migraci do vhodnějších formátů (dle doporučení Florida Digital Archive - viz níže Tab. 1) - TIFF, PNG a JPG2000.

1.1.1. Definování báze - definují se institucionální politiky, právní regule a uživatelská kritéria, která mohou ovlivnit plánovaná rozhodnutí pro ochranu dokumentů. Například politiky definující povolené formáty pro “ingest”, zákon o ochraně duševního vlastnictví apod.

1.1.1.1. Identifikace (Obr. 4) - pro pozdější dohledání konkrétního plánu je potřebné všechny ochranné plány náležitě označit a popsat.

1.1.1.2. Status, Popis - představuje další krok v plánování - tedy jestli je konkrétní plán definovaný, čeká na schválení nebo je již aktivní a co vedlo k tomu, že jej bylo nutné nově definovat či pozměnit.

---
⁹ FITS - File Information ToolSet (http://code.google.com/p/fits) zahrnuje a sjednocuje výstupy charakterizačních nástrojů (DROID a JHOVE), včetně extraktoru ExifTool (http://www.sno.phy.queensu.ca/~phil/exiftool) a dalších.
1.1.1.3. **Politiky** - organizace zodpovědné za dlouhodobou ochranu svých dokumentů se v různých ohledech liší. Proto nejen duševní vlastnictví, technické požadavky, ale i strategie, cíle a politiky představují integrální součást plánu dlouhodobé ochrany.

1.2. **Definování příkladů** - jde o obecný popis objektů ve sbírce a sbírky jako celku (velikost sbírky, formáty ve sbírce apod.).

1.2.1. **Profil sbírky** - kromě obecného popisu výběru dokumentů, které nejlépe vystihují sbírku je možné uvést i stupeň rizika pro každý profilový objekt.

1.2.2. **Upload ukázkových souborů**
1.3. **Identifikace požadavků** - představuje nejdůležitější krok při plánování dlouhodobé ochrany. Požadavky jsou definovány ve formě tzv. stromu cílů.

1.3.1. **Strom cílů** - jde o myšlenkovou mapu kritérií popisujících současné charakteristiky daných formátů včetně jejich budoucí podoby. Jednotlivé mapy je možné vytvořit online, přímo v nástroji PLATO, nebo offline, pomocí freewarového nástroje Freemind viz podkapitola 4.1 Plato a poté je nahrát do PLATO.

![Obr. 6 Strom cílů ve formě myšlenkové mapy](image)

1.3.2. **Znalostní báze** - pro definici je možné využít i výsledky jiných institucí.

1.4. **Vyhodnocení alternativ a výsledků** - jde o proces empirické evaluace výsledků, který tvoří základ pro další fáze (zejména transformaci). Některé z hodnocených výsledků je možné provést automaticky (viz těž podkapitola 4.1.3 Výstupy z nástroje PLATO), stále však ještě většinu z nich je nutné ohodnotit ručně. To znamená procházet jednotlivá neautomatizovatelná kritéria a doplňovat hodnotu a poté je finálně odsouhlasit. Poté je možné transformace.

Například u formátu JPEG byla za důležitou signifikantní vlastnost označena kvalita výstupního, přetransformovaného objektu. Je proto nutné všechny možnosti přemigrovaného objektu stažit na lokalní počítač a subjektivní vyhodnotit, zda-li nový formát – výsledek migrace, odpovídá požadavkům instituce.

1.4.1. **Transformace** - hodnoty ve stromě požadavků jsou uvedeny v různých jednotkách (vteřiny, koruny, bity apod.); aby bylo možné tyto rozdílné hodnoty agregovat, je nutné je přetransformovat do jednotné škály. Vývojáři PLATO uvádějí, že k nejrelevantnějším výsledkům se hodnotitelé dopracují při škále celých hodnot.
od 0 do 5, kdy 0 představuje zcela neakceptovatelný výsledek a 5 naopak nejlépší možný rezultát.

3.1.3 VÝSTUPY Z NÁSTROJE PLATO

Podle vývojářů PLATO umožňuje vyhodnocování signifikantních vlastností u téměř všech typů objektů, a to automaticky nebo ručně. Realita však tomu nevždy odpovídá. Ku příkladu TXT soubory není možné pomocí PLATO ani zmigrovat do jiných formátů, nýbrž zhodnotit jejich signifikantní vlastnosti.


Všechny ostatní možné signifikantní vlastnosti je nutné vyhodnotit ručně; což v případě cca 25 kritérií a 3 testovacích souborů přes min. 3 externí služby představuje ruční hodnocení cca 225 bodů. Ruční kontrola musí proběhnout i u takových signifikantních kritérií, u kterých by strojová kontrola neměla být překážkou (např. “image size”, “image height”, “image width” apod.). Jde totiž o porovnávání čiselné shody vstupu a výstupu, nikoliv o vyhodnocení kvality (např. vizuální čitelnost).

Proces hodnocení byl stížen i skutečnost, že jednotlivé přetransformované soubory bylo nutné si stáhnout na lokální počítač (viz Obr. 7 Výsledky migrací externími službami (JPG do TIFF, PNG, BMP a PDF)), zde je i srovnávat a posuzovat poté ručně zanést do PLATO.

![Obr. 7 Výsledky migrací externími službami (JPG do TIFF, PNG, BMP a PDF)](image)

V případě TXT souborů jsou testovací migrace ještě komplikovanější. Externí služby (z angl. external services) nenabízejí žádnou vhodnou alternativu pro migrace TXT souborů. V externích službách...
neexistují žádné vhodné alternativy typu XML. Pro potřeby našeho projektu proběhla zkusmo alespoň migrace do formátu HTML, avšak neúspěšně. Výsledky z testovacích migrací jsou dostupné přímo v nástroji PLATO.

V případě TXT souborů je možné PLATO pro potřeby NK ČR použít pouze jako migrační nástroj, nikoliv jako hodnotící nástroj. Vybrané vzorky formátů se vloží do PLATO, přemigrují do jiných formátů, poté stáhnou z nástroje a mimo něj vyhodnotí.

Při vyhodnocování se kritéria, resp. signifikantní vlastnosti rozdělí na kvantifikovatelné (tedy strojově kontrolovatelné) a kvalitativní (ručně kontrolovatelné) a přiřadí se jim potřebná váha. Všechny výsledky je následně možné sepsat např. ve formě tabulky vytvořené v některém z dostupných tabulkových procesorů.

### 3.1.4 ZAPOJENÍ PLATO DO WORKFLOW INSTITUCE

Stojně jako je PLATO implementováno v open source repozitáři EPrints10, je možné jej integrovat do worflow NDK. Pokryl by tak poslední dvě kritéria plánu dlouhodobé ochrany – plánování a akce11.

---


Postup by byl obdobný vůči uvedenému popisu pro práci s nástrojem PLATO. V rozhraní LTP systému NDK by se pomocí identifikace a charakterizace vybraly rizikové soubory, zaznamenaly jejich signifikantní vlastnosti. V integrovaném nástroji Plato by se soubory přemigrovaly a posoudila by se jejich výstupní kvalita.

4. ZÁVĚR

Odrazovým můstkem k efektivní dlouhodobé ochraně digitálních dokumentů, které daná instituce spravuje, je pochopení, že cílem nemá být uchovávání všeho. Předpokládá se výběr dle stanovených pravidel a podmínek za jakých se budou dokumenty chránit. Zde se jako nejvhodnější metodika pro selekci objektů a jejich typů jeví tzv. 3-stupňová strategie, která kombinuje náhodný výběr a kontrolovaný výběr společně s řízeným opomenutím. Jakmile víme (je specifikováno) co bude instituce uchovávat, je možné se zaměřit na to jakým způsobem. To předpokládá znalost klíčových vlastností (objektů), které je nutné dlouhodobě uchovávat. Správci sbírek tedy potřebují znát signifikantní vlastnosti pro výběr ochranných technik a načasování různých ochranných akcí u objektů určených k dlouhodobé ochraně. Ne všechny digitální objekty byly vytvořeny v otevřených specifikacích. Jelikož jsou proprietární formáty navázány na komerční subjekt, přetrvává u nich velké riziko ztráty informací (autentický, srozumitelný obsah).

Proto je žádoucí takovéto formáty souborů migrovat do otevřenějších protějšků při zachování jejich signifikantních vlastností. Tento proces migrace formátů a rozhodování o vhodném budoucím formátu ulehčují nejrůznější nástroje (většinou open source).

NK ČR se ve svém výzkumu zaměřila právě na open source nástroje, aby se mohly využít případně i další instituce v ČR. Ukázalo se, že ne všechny jsou však vhodné k integraci do systémů Národní knihovny ČR, případně Národní digitální knihovny. Nejpoužitelnějším nástrojem pro plánování dlouhodobé ochrany se zdá být PLATO, který v současné době nemá žádný vhodnější ekvivalent. Hlavním principem webového nástroje PLATO je vyhodnocování ochranných aktivit podle stanovených cílů, signifikantních vlastností a požadavků na jejich zachování.

Reálná práce s nástrojem PLATO ukazuje, že většinu z předem definovaných signifikantních vlastností u JPG, JP2 a TXT není bohužel ve stávající verzi nástroje možné automaticky otestovat. U JPG a JP2 je nutné většinu kritérií ohodnotit ručně. U TXT souborů není možné provést ani samotné migraci. Naději na zlepšení je v tomto ohledu další vývoj PLATO v rámci evropského projektu “SCAPE : Scalable Preservation Environments”.

20
Nástroj PLATO tak poskytuje výborný zdroj pro pochopení a testování vlastností entity OAIS nazvané „preservation planning“. NK ČR počítá v nejbližších letech s tím, že bude mít komerčně dostupný long-term preservation systém, jehož součástí bude funkční modul „preservation planning“. Testování PLATO pomohlo v tom, že víme, jaká je podstata plánování ochrany, a budeme moci LTP systém v tomto směru využít a případně vedle něj PLATO stále používat.

Klíčová je tato zkušenost a návod použití nástroje PLATO pro ostatní knihovny a paměťové instituce, které nebudou mít LTP systém a budou řešit dlouhodobou ochranu digitálních dat z vlastních prostředků cestou minimálních nákladů.


5.
SEZNAM BIBLIOGRAFICKÝCH ODKAZŮ


Blue Ribbon Taskforce (http://www.brtf.sdsc.edu)


