



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

Afiliace • 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	<i>Nově implementované prvky ve verzi 3.0.....</i>	14
5.1.2	<i>Práce s nástrojem Plato.....</i>	15
5.1.3	<i>Výstupy z nástroje Plato.....</i>	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

“Základní výzvou dlouhodobé digitální ochrany je zachování přístupnosti a autenticity digitálních objektů napříč časem, obory a neustále se měnícím technickým prostředím.”
(Wilson, 2008, InSPECT Significant Properties Report)

1. ÚVOD

Není tomu tak dávno, kdy ochrana digitálních dokumentů byla pojímána 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. vytištěním 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 vyřešit 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.

Tento dokument navazuje na podklad pro závěrečnou zprávu grantu Výzkum a vývoj “Open source nástroje pro dlouhodobou ochranu digitálních dat”, který byl sepsán v prosinci 2010. Z něj vyplynulo, že ne všechny hodnocené nástroje jsou vhodné k integraci do Národní digitální knihovny. Nejpoužitelnějším open source nástrojem pro plánování dlouhodobé ochrany byl shledán PLATO Preservation Planning Tool (zkráceně nazýván PLATO), který v současné době nemá žádný jiný vhodný ekvivalent.

Cílem výzkumu v roce 2011 bylo zaměřit se na jednu z fází životního cyklu dlouhodobé ochrany digitálních dokumentů - 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

¹ Myšlenka UVC, tedy jakéhosi univerzálního virtuálního počítače se zrodila počátkem roku 2000 a byla otestována pro potřeby dlouhodobé ochrany JPEG v Národní knihovně Nizozemska. UVC je postaven na principu dekodérů formátů souborů a programů. Podrobněji na: <http://www-935.ibm.com/services/nl/dias/cs/uvc.html>.

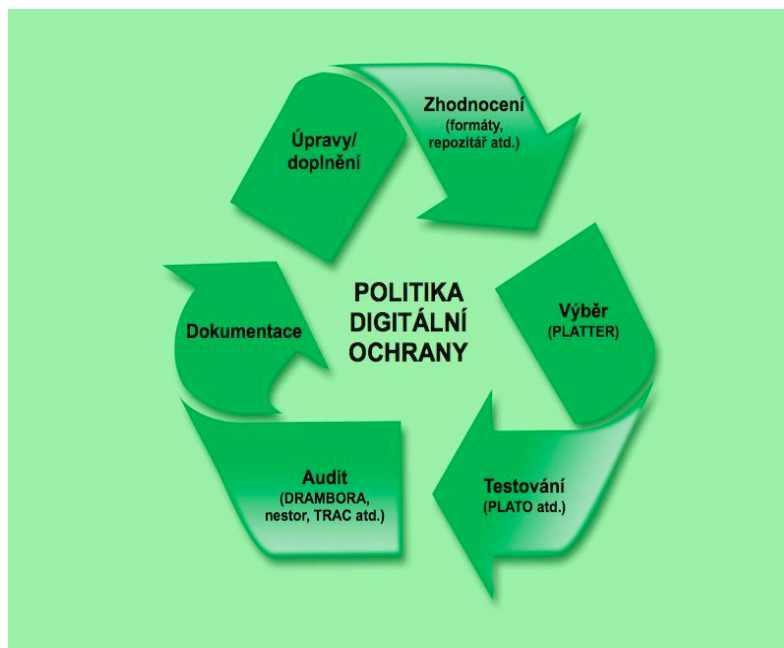
vlastnosti importovat do nástroje PLATO, budou tvořit teoretickou část podkladu. Praktickou složku výzkumu reprezentují závěry z PLATO. Součástí zprávy jsou zkušenosti z konkrétní práce s nástrojem PLATO, včetně návrhu zapojení PLATO do NDK.

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 vyfiltrovat².

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

² Datová archeologie (z angl. data archeology) představuje disciplínu zabývající se obnovou počítačových dat, jež byla vytvořena v dnešních zastaralých formátech nebo uchována na přežitých médiích. Poprvé se termín objevil v roce 1993 v souvislosti s obnovou počítačových záznamů o klimatických podmínkách. Ty byly klíčové pro vyhodnocování teorií klimatických změn. Více na: http://en.wikipedia.org/wiki/Global_Oceanographic_Data_Archaeology_and_Rescue_Project.

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 rozdělena 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.



Stanovování signifikantních vlastností však není podchyceno jednotnou metodikou³.

Koncepty signifikantních informací se do určité míry překrývají s reprezentacemi informací. Signifikantní vlastnosti představují charakteristiky abstraktního informačního objektu, naproti tomu reprezentace informací (strukturní 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 "extrahují" signifikantní vlastnosti (viz myšlenkové mapy v příloze).

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

Kontext	organizační kontext (např. název organizace, datum, vazba na jiné dokumenty apod.)	uchování logových souborů s informacemi o preservačních akcích, o originálu a současném formátu souboru, názvu a verzi HW, SW, OS
Obsah	veškerý obsah musí být uchován, včetně záhlaví, zápatí, poznámek pod čarou, obsahu (souhrn) a vlastností dokumentu	prostý text musí být vzdy citelný
Struktura	struktura dokumentu musí být uchována tak, aby reprezentovala logické vazby mezi jednotlivými součástmi dokumentu	poradí kapitol, odstavců, ale též správná pozice vložených poznámek, poznámek pod čarou a obrázků
Vzhled	vzhled originálu a uchovávané verze nemusí být identické, nicméně "nový" vzhled nesmí ovlivnit pochopení originálního významu	
Chování	aktivní chování, uplate, zaznam o původních formách chování musí být zaznamenán	popis aktivních odkazů musí být uchován

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

³ Dollar navrhuje 8 výchozích kritérií: čitelnost, srozumitelnost, identifikovatelnost, zapouzdřenost, vyhledatelnost, transformovatelnost a autenticitu. Jiný přístup představuje tzv. Model důvěry (z angl. Model of Trust) postavený na autenticitě, aktuálnosti, přesnosti, pokrytí, kredibilitě, objektivitě, validitě, předvídatelnosti a věrohodnosti (Korenkova, 2011). Knight (2009) identifikuje složení digitálního objektu (formu jakou je idea vyjádřena), účel (zamýšlené funkce, typ uživatele), organizační investice (strategické, finanční), potenciál (právní, finanční).

Ty je dále možné “přetavit” do podoby myšlenkové mapy (viz Příloha č. 1 až 3) nebo inventář signifikantních vlastností pro jednotlivé formáty⁴.

Výběru signifikantních vlastností by měl předcházet výběr samotných zdrojů (nebo jejich typů) k uchování, 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 kompletní analýzu dokumentů, hodnocení aktivit a jejich právních omezení, určení reálnosti uchování dokumentů a konečné rozhodnutí.

Podle Neumeyera a Raubera (2007) však kontrolovaný výběr obvykle nahrává hodnotám hlavního proudu, zatímco ty ostatní jsou často potlačovány. Důležitou roli by podle nich mělo sehrávat tzv. “řízené opomenutí” (z angl. planned omission), kdy dokumenty hlavního proudu jsou přijímány, ale neuchovávají se pro příští generace a poté následuje výběr materiálů kulturního dědictví pro **dlouhodobou ochranu na bázi náhodného výběru**. Představuje totiž **nejméně nákladné a zároveň racionální řešení**. **Hlavní výhodou** náhodného výběru oproti “appraisal”, tedy kontrolovanému výběru, **je nestranný pohled na současný život, jednoduchost, finanční přívětivost**.

Nicméně nejlepší řešení se jeví jako kombinace náhodného výběru a kontrolovaného výběru (společně s řízeným opomenutím) pomocí tzv. **3-stupňové strategie**⁵. Tato se částečně opírá i o “appraisal” jakožto 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

⁴ Výbornou ukázkou je např. Florida Digital Archive Format Information - <http://web.archive.org/web/20100527210441/http://www.fcla.edu/digitalArchive/formatInfo.htm>.

⁵ Tuto metodu je dále možné kombinovat např. s rozhodovacím stromem od Digital Preservation Coalition - Decision Tree for Selection of Digital Materials for Long-term Retention. Podrobněji na <http://www.dpconline.org/advice/preservationhandbook/decision-tree>.

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



Obr. 3 Workflow analýzy užitek (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 klíčové 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 staví svoji funkcionalitu právě na analýze užitek. 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).

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

Obraz	- TIFF (nekomprimovaný) - JPEG2000 (bezeztrátový) (*.jp2) - PNG (*.png)	- BMP (*.bmp) - JPEG/JFIF (*.jpg) - JPEG2000 (lossy) (*.jp2) - TIFF (komprimovaný) - GIF (*.gif) - Digital Negative DNG (*.dng)	- MrSID (*.sid) - TIFF (v Planar formátu) - FlashPix (*.fpx) - PhotoShop (*.psd) - RAW - JPEG 2000 Part 2 (*.jpf, *.jpx) - ostatní obrazové formáty, které nejsou specifikovány
Audio	- AIFF (PCM) (*.aif, *.aiff) - WAV (PCM) (*.wav)	- SUN Audio (nekomprimovaný) (*.au) - Standard MIDI (*.mid, *.midi) - Ogg Vorbis (*.ogg) - Free Lossless Audio Codec (*.flac) - Advance Audio Coding (*.mp4, *.m4a, *.aac) - MP3 (MPEG-1/2, Layer 3) (*.mp3)	- AIFC (komprimovaný) (*.aifc) - NeXT SND (*.snd) - RealNetworks 'Real Audio' (*.ra, *.rm, *.ram) - Windows Media Audio (*.wma) - chráněný AAC (*.m4p) - WAV (komprimovaný) (*.wav) - ostatní audio formáty, které nejsou specifikovány
Video	- Motion JPEG 2000 (ISO/IEC 15444-4) (*.mj2) - AVI (nekomprimovaný, motion JPEG) (*.avi) - QuickTime Movie (nekomprimovaný, motion JPEG) (*.mov)	- Ogg Theora (*.ogg) - MPEG-1, MPEG-2 (*.mpg, *.mpeg, zabalený do AVI, MOV) - MPEG-4 (H.263, H.264) (*.mp4, zabalený do AVI, MOV)	- AVI (ostatné) (*.avi) - QuickTime Movie (ostatné) (*.mov) - RealNetworks 'Real Video' (*.rv) - Windows Media Video (*.wmv) - ostatní video formáty, které nejsou specifikovány
Databáze/programy pro zpracování tabulek	- Comma Separated Values (*.csv) - Delimited Text (*.txt) - SQL DDL	- DBF (*.dbf) - OpenOffice (*.sxc/*.ods) - OOXML (ISO/IEC DIS 29500) (*.xlsx)	- Excel (*.xls) - ostatní databázové a tabulkové formáty, které nejsou specifikovány
Prezentace		- OpenOffice (*.sxi/*.odp) - OOXML (ISO/IEC DIS 29500) (*.pptx)	- PowerPoint (*.ppt) - ostatní prezentační formáty, které nejsou specifikovány

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 protějšků 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. Jde o JPG (Příloha č. 1 - Myšlenková mapa JPG), JP2 (Příloha č. 2 - Myšlenková mapa JP2), TXT (Příloha č. 3 - Myšlenková mapa TXT), PDF.

Výstupy z prvních dvou zmíněných 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 oblastí - **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 vyvíjeny 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.

Ne všechny výše uvedené nástroje jsou však vhodné k integraci do systémů Národní knihovny ČR, případně Národní digitální knihovny. Důvodem je profilace na úzkou cílovou uživatelskou skupinu (Hoppla, RODA) nebo "nemodulární" architektura (Mopseus). Integrace metodologií typu AIDA je možná v případě dodatečného vývoje a implementace do stávajících systémů. V kontextu NDK je možnost využití AIDA pro závěrečnou evaluaci projektu. Nejpoužitelnějším nástrojem na plánování dlouhodobé ochrany i nadále zůstává PLATO, který v současné době nemá žádný vhodný ekvivalent.

3.1 PLATO

<http://www.ifs.tuwien.ac.at/dp/plato>

Garant: Department of Software Technology and Interactive Systems, Vienna University of Technology
(<http://www.ifs.tuwien.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** (**SCA**lable **P**reservation **E**nvironments) 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. Nejvhodnější formát je vybírán na základě vyhodnocení stanovených důležitých vlastností souborů (resp. objektů), které mají být zachovány i do budoucna (např. zalamování řádků a číslování stránek u textových souborů). Tento nástroj vychází z referenčního modelu OAIS, konkrétně z jeho funkční entity "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):

- 1.1. **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.

- 1.2. **Hodnocení alternativ** – hodnotící kritéria z předchozí fáze jsou podkladem pro provádění experimentálních migračních aktivit, které přicházejí do úvahy.
- 1.3. **Posouzení výsledků** - experimentální migrace jsou výstupem pro následnou analýzu a vyhodnocení nejlepšího formátu dané vzorové sbírky digitálních dat.

Výstupem je dokument s objektivním posouzením nejrozličnějších 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 3.0

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) vlastností, 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 P2⁷ - 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 by bylo přímé prohledávání Planets Service Registry; ten ale bohužel není dostupný na žádné veřejné adrese (Prom, 2010).

⁷ Podrobněji v TARRANT, D., HITCHCOCK, S., CARR, L. Where the Semantic Web and Web 2.0 Meet Format Risk Management:P2 Registry. *The International Journal of Digital Curation* [online]. Issue 1, Volume 2, 2011 [cit. 2011-05-22]. Dostupné z WWW: <<http://www.ijdc.net/index.php/ijdc/article/viewFile/171/239>>. ISSN: 1746-8256.

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.

⁸ P2 představuje registr formátů, který je sémanticky obohacený a obsahuje informace pro podporu procesů dlouhodobé ochrany. Podrobněji v prezentaci: TARRANT, D., HITCHCOCK, S., CARR, L. A complete preservation workflow in EPrints (+ PLATO) - 10 Minute Summary. *Preservation and Archiving Special Interest Group (PASIG)* [online]. Madrid : 5th July 2010 [cit. 2011-05-22]. Dostupné z WWW: <<http://eprints.ecs.soton.ac.uk/21279/>>.

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.

Plan | Define Requirements | Evaluate Alternatives | Analyse Results | Build Preservation Plan

Define basis

Identification

Status

Description

Policies

[1] Identification

Identification Code:

Document types:

Plan name:

Plan description:

Responsible planners:

Organisation:

Obr. 4 Identifikace nového projektu v PLATO Preservation Planning Tool

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] Sample Records

Description of sample records:

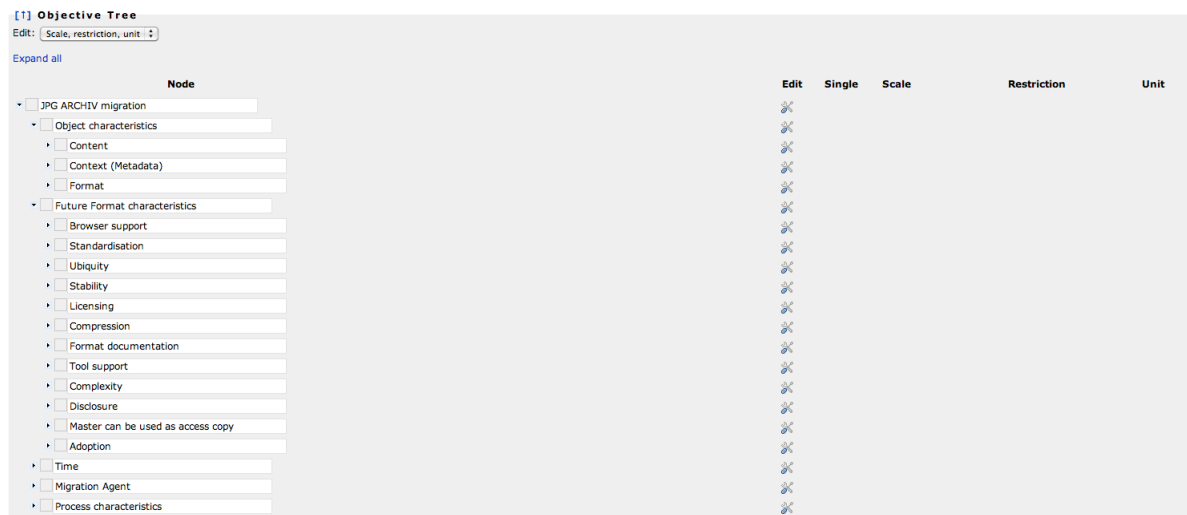
Sample Record	Object Format
<p>Full name: <input type="text" value="1048500202.jpg"/></p> <p>Short name: <input type="text" value="1048500202.jpg"/></p> <p>Has data: <input checked="" type="checkbox"/> 1.33 MB download</p> <p>Original technical environment: <input type="text"/></p> <p>Description: <input type="text"/></p> <p>Remove sample</p>	<p>PUID: <input type="text" value="fmt/43"/></p> <p>Name: <input type="text" value="JPEG File Interchange Format"/></p> <p>Version: <input type="text" value="1.01"/></p> <p>Mime-type: <input type="text" value="image/jpeg"/></p> <p>Extension: <input type="text" value="jpg"/></p> <p>Identify format </p> <p>View Characteristics </p> <p>View FITS Characteristics </p>
<p>Full name: <input type="text" value="1048500338.jpg"/></p> <p>Short name: <input type="text" value="1048500338.jpg"/></p> <p>Has data: <input checked="" type="checkbox"/> 1.41 MB download</p> <p>Original technical environment: <input type="text"/></p> <p>Description: <input type="text"/></p> <p>Remove sample</p>	<p>PUID: <input type="text" value="fmt/43"/></p> <p>Name: <input type="text" value="JPEG File Interchange Format"/></p> <p>Version: <input type="text" value="1.01"/></p> <p>Mime-type: <input type="text" value="image/jpeg"/></p> <p>Extension: <input type="text" value="jpg"/></p> <p>Identify format </p> <p>View Characteristics </p> <p>View FITS Characteristics </p>

XCDL Description:

Obr. 5 Upload ukázkových souborů do nástroje PLATO Preservation Planning Tool

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

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 též 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 možná transformace.

Například u formátu JPEG byla za důležitou signifikantní vlastnost označena kvalita výstupního, přetransformovaného objektu. Je proto nutné všechny možnosti přemigrovaného objektu staáhnout na lokální počítač a subjektivně 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 nejlepší možný rezultat.

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 ne vždy odpovídá. Ku příkladu TXT soubory není možné pomocí PLATO ani zmigrovat do jiných formátů, nýbrž zhodnotit jejich signifikantní vlastnosti.

Automaticky hodnocená kritéria pro obrazové dokumenty typu JPG jsou: “imagequality:AE”, “imagequality:MAE”, “imagequality:MEPP”, “imagequality:MSE”, “imagequality:PAE”, “imagequality:PSNR”, “imagequality:RMSE”, “input:filesize”.

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 číselné shody vstupu a výstupu, nikoliv o vyhodnocení kvality (např. vizuální čitelnost).

Proces hodnocení byl stížen i skutečností, 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 posuzení poté ručně zanést do PLATO.

What is the task here?

Alternatives	Sample		
	Download 2	Download 3	Download 4
JPGtoPNG (JavaImageIO)			
JPGtoTIFF (JavaImageIO)			
JPGtoTIFF (ImageMagick)	Download 	Download 	Download 
JPG2PNG (ImageMagick)	Download 	Download 	Download 
JPG2TIFF (ImageMagick)			
Image2JPG (Acme)			
ImageIO JPG2PNG			
ImageIO JPG2BMP			
Im4JavaImageMagickMigrate @ pl	Download 	Download 	Download 




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

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 zkouška alespoň migrace do formátu HTML, avšak neúspěšně. Výsledky z testovacích migrací jsou dostupné přímo v nástroji PLATO.

Run Experiments

Experiment execution
Resulting objects

Alternative	Description
Im4JavaImageMagickMigrate @ pl	Summary Report
	Record
	autorsky_zakon.txt  Status Edit Report
	jpeg_pdf.txt  Report Edit Report
	i2010.txt  Edit Report

Run all experiments

Save Discard changes Save and proceed

Result Files

Results	Sample Records
Im4JavaImageMagickMigrate @ pl	autorsky_zakon.txt (0 MB) Upload result file
	jpeg_pdf.txt (0 MB) Upload result file
	i2010.txt (0 MB) Upload result file

Describe result files in XCDL

[What is XCDL?](#)

Obr. 8 Neúspěšná migrace TXT souborů do HTML v 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. významné 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

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

¹⁰ Podrobněji v příspěvku Tarrant, D., Hitchcock, S., Carr, L., Kulovits, H. and Rauber, A. (2010) Connecting preservation planning and Plato with digital repository interfaces. In: 7th International Conference on Preservation of Digital Objects (iPRES2010), 19 - 24 September 2010, Vienna, Austria. Dostupné z WWW: <<http://eprints.ecs.soton.ac.uk/21289/>>.

¹¹ Workflow dlouhodobé ochrany dle příspěvku v poznámce pod čarou č. 10: "identification - characterisation - risk assessment - planning - action".

Postup by byl obdobný výše 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 uchování 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 nejrozumnější nástroje (většinou open source).

NK ČR se ve svém výzkumu zaměřila právě na open source nástroje, aby je 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 migrace. Nadějí na zlepšení je v tomto ohledu další vývoj PLATO v rámci evropského projektu "SCAPE : Scalable Preservation Environments".

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.

Z pohledu NK ČR by do budoucna bude nutné podrobně definovat nejružnější přístupy v selekci signifikantních vlastností pro všechny soubory, které budou do budoucí NDK přijímány. Velmi užitečná (nejen pro Národní knihovnu ČR) by byla metodika pro jednotlivé souborové formáty. Jako vzor by skvěle mohl posloužit např. akční plán “Florida Digital Archive Format Information” (<http://web.archive.org/web/20100527210441/http://www.fcla.edu/digitalArchive/formatInfo.htm>). Takový akční plán bude následně připraven pro otestování v případném upraveném nástroji PLATO nebo jiných plánovacích nástrojích dlouhodobé ochrany. Mohl by vycházet z jednotlivých myšlenkových map, vytvořených v tomto výzkumu pro různé typy digitálních objektů.

5.

SEZNAM BIBLIOGRAFICKÝCH ODKAZŮ

ADAM, S. Preserving authenticity in the digital age. *Library Hi Tech* [online]. Vol. 28, no. 4, 2010 [cit. 2011-05-22]. Dostupný z WWW: <<http://www.emeraldinsight.com/journals.htm?articleid=1896431>>. DOI 10.1108/07378831011096259.

Ekonomická analýza [online]. cca 2010 [cit. 2011-05-22]. Dostupné z WWW: <http://www.rrav.cz/sluzby/ekonom_analyzy.html>.

HEDSTROM, M., LEE, C.A. Significant properties of digital objects: definitions, applications, implications. Proceeding of the DLM Forum [online]. 2002 [cit. 2011-05-22]. Dostupné z WWW: <http://www.ils.unc.edu/callee/sigprops_dlm2002.pdf>.

Knowledge Exchange, Danish Agency for Libraries and Media : *Report on the Knowledge Exchange seminar on Persistent Object Identifiers* [online]. The Hague, 14-15 June 2011 [cit. 2011-05-22]. Dostupný z WWW: <http://www.knowledge-exchange.info/Files/Filer/downloads/PersID/Workshop%20POID/Report%20POID%20seminar_14_June_2011.pdf>.

KNIGHT, G., PENNOCK, M. Data Without Meaning: Establishing the Significant Properties of Digital Research. *The International Journal of Digital Curation* [online]. Vol 4, No 1 (2009) [cit. 2011-05-22]. Dostupné z WWW: <www.ijdc.net/index.php/ijdc/article/view/110>. ISSN: 1746-8256.

KNIGHT, G. InSPECT Framework Report [online]. 13 October 2009 [cit. 2011-05-22]. Dostupné z WWW: <<http://www.significantproperties.org.uk/inspect-framework.pdf>>.

A Network of Excellence on Digital Libraries. Deliverable 6.4.1: A Framework for Documenting the Behaviour and Functionality of Digital Objects and Preservation Strategies [online]. 27 June 2005 [cit. 2011-05-22]. Dostupné z WWW: <http://www.dpc.delos.info/private/output/DELOS_WP6_d641_final_vienna.pdf>.

POZO, N.D., LONG, A.S., PEARSON, D. "Land of the Lost": a discussion of what can be preserved through digital preservation. *Library Hi Tech* [online]. Vol. 28, no. 2, 2010 [cit. 2011-05-22]. Dostupný z WWW: <<http://www.emeraldinsight.com/journals.htm?articleid=1864755>>. DOI 10.1108/07378831011047686.

PROM, C. PLATO (Digital Preservation Planning) Software Review [online]. April 25, 2010 [cit. 2011-05-22]. Dostupný z WWW: <<http://e-records.chrisprom.com/?p=1082>>.

UK Data Archive : *UK Data Archive Preservation Policy Public Version* [online]. 26 October 2010 [cit. 2011-05-22]. Dostupný z WWW: <<http://www.data-archive.ac.uk/media/54776/ukda062-dps-preservationpolicy.pdf>>.

"What to preserve? The significant properties of Digital Objects" Conference [online]. London 7 April 2008 [cit. 2011-05-22]. Dostupný z WWW: <<http://alanake.wordpress.com/2008/04/>>.

KORENKOVA, M. HÄGERFORS, A. Quality criteria for digital information in long-term preservation. *Archiving 2011*. Salt Lake City (UT) : Society for Imaging Science and Technology, 2011. 232 p. ISBN / ISSN: 978-0-89208-294-0.

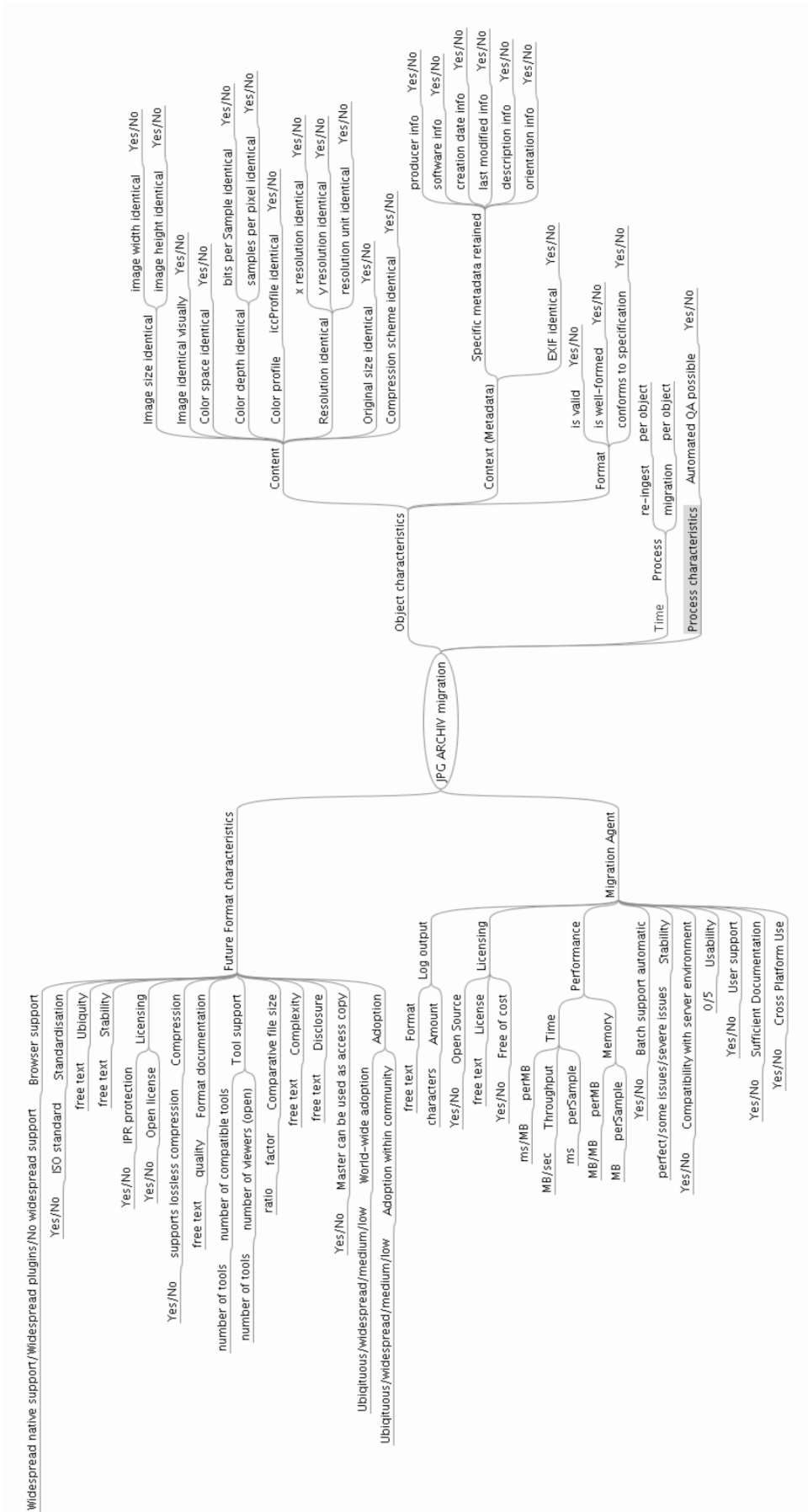
Blue Ribbon Taskforce (<http://www.brtf.sdsc.edu>)

NEUMEYER, R., RAUBER, A. Why appraisal is not 'utterly' useless and why it's not the way to go either : A provocative position paper (PPP) [online]. 12th November 2007 [cit. 2011-05-22]. Dostupné z WWW: <http://www.digitalpreservationeurope.eu/publications/position/appraisal_final.pdf>.

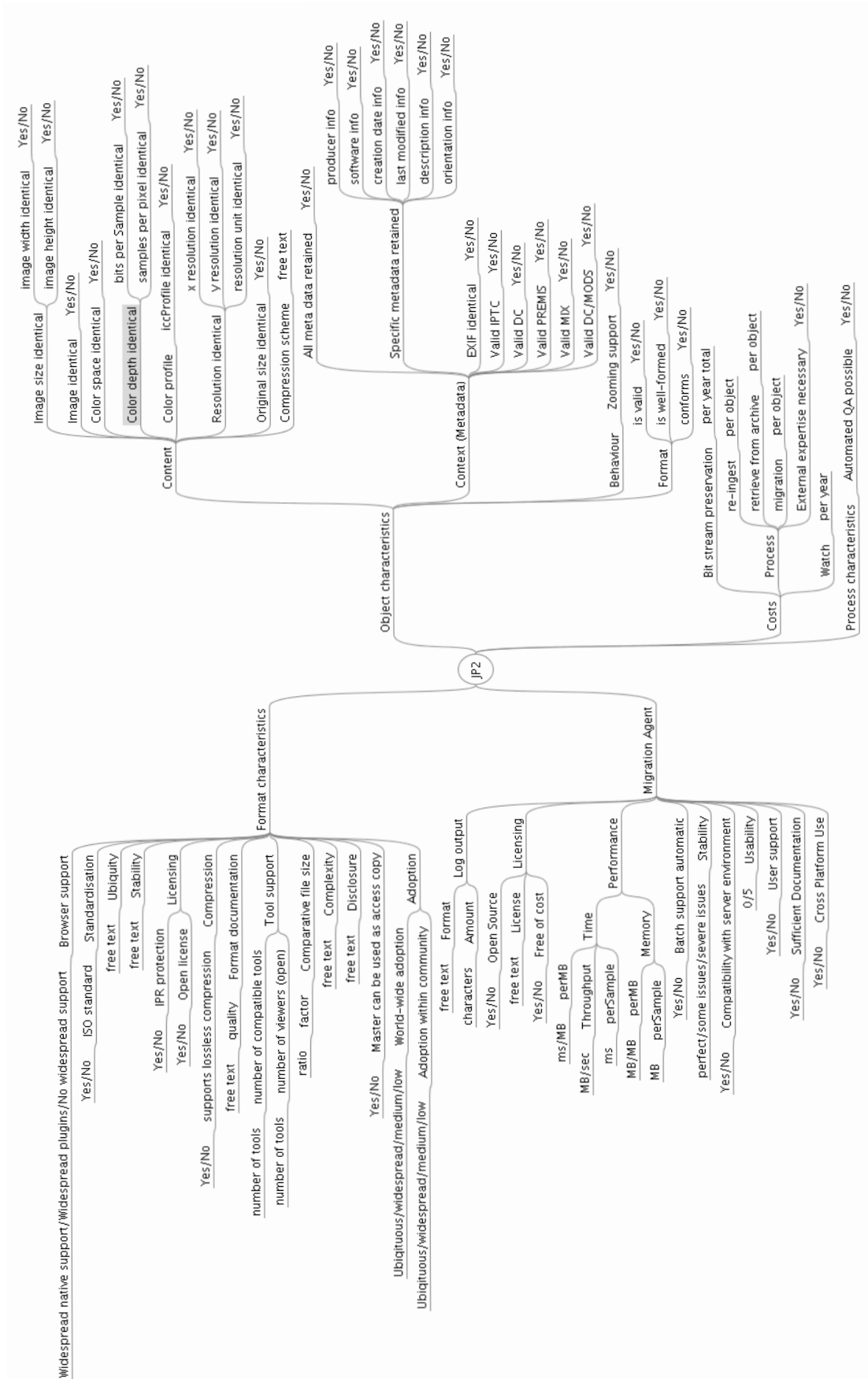
SLATS, J. Practical Experiences of the Digital Preservation Testbed Office formats [online]. May10-11 2004 [cit. 2011-05-22]. Dostupné z WWW: <http://www.erpanet.org/events/2004/vienna/presentations/erpaTrainingVienna_Slats.pdf>.

WILSON, A. *Significant properties report* [online]. 2007 [cit. 2011-05-22]. Dostupné z WWW: <http://www.hf.cz/titindex.php?id=9&zobraz=predplatne&typ_akce=predplatneobnova>.

6. PŘÍLOHA Č. 1 - MYŠLENKOVÁ MAPA JPG



7. PŘÍLOHA Č. 2 - MYŠLENKOVÁ MAPA JP2



8. PŘÍLOHA Č. 3 - MYŠLENKOVÁ MAPA TXT

9.

