PADI
(Preserving Access to Digital Information)

Recommended Practices for Digital Preservation
http://www.nla.gov.au/preserve/digipres/digirespractices.html

Jaroslav Kortus, 7. 1. 2006

Projekt PADI spadá pod National Library of Australia. Zabývá se aktivitami souvisejícími s dlouhodobým uchovávním digitálních dat. S přibýváním digitálních dat logicky vyvstává stále větší potřeba jejich archivace. Jejich dlouhodobé skladování však přináší i problémy, kterým je třeba čelit.

Hlavním problémem je zastarávání. Zastarává jak hardwarové vybavení, tak software. Při budování archivační politiky je tedy třeba myslet dopředu a vybírat pokud možno takové datové formáty a takové technické vybavení, které přečká delší časový horizont.

Hlavní potřeby jsou:
- zajistit neporušenost dat
- používat takové způsoby zobrazení a materiály, aby neutrpěla autentičnost díla

Související potřeby:
- zaznamenání dostatečného množství metadat
- vyřešit související právní otázky
- monitorovat změny v technologiích, které by způsobily nepřístupnost

Projekt se snaží identifikovat klíčové body, kterých bychom se měli držet, pokud budeme v naší organizaci takovou archivační politiku zavádět. Organizuje ji do třech částí.

Jako první zmiňuje stanovení cílů projektu. Měli bychom zjistit, co a pro koho budeme archivovat, zjistit využití těchto materiálů a zajistit tak jejich plnohodnotné využití i v budoucnosti. V druhém bodě se zmiňuje o bližším prozkoumání sbírek - jejich identifikaci a technických charakteristikách. Měli bychom pojmenovat priority a zvážit rizika (hardwarová, různé typy reprezentace dat, komprese...) a zvolit vhodný formát metadat. Ve třetím kroku bychom měli zajistit vhodnou ochranu takto shromážděných dat. Měli bychom vyřešit právní otázky kopie a archivace děl, umístit je na spolehlivý hardware a definovat rozumnou zálohovací politiku.

Zbývá vyřešit problém dostupnosti. Největší problém je, že operační prostředí i datové formáty v důsledku technologických změn zastarají. Zmíním jen několik z důležitých opatření, která projekt zmiňuje jako dlouhodobě vhodná:
- přimět autory k produkci děl v takových formátech a softwaru, který má naději na delší životnost (hlavně otevřené a podporované standardy)
- minimalizovat množství použitých formátů, zvážit konverzi proprietárních do otevřených
- uschovat originální data pro možnou budoucí lepší metodu digitalizace.

To byl stručný přehled cílů projektu PADI a nyní prozkoumáme některé jeho části podrobněji.

Podívejme se podrobněji na problém zastrávání, který je z hlediska uchovávání dat jedním z kritických bodů. První problém je problém softwarový a datový. Data mohou být buď ve formátu, který bude mít krátkou životnost (většinou nějaký proprietární) nebo vázaný přímo na software, bez něhož jsou data jen shluk bitů. Pro krátkodobé použití to jistě není problém, ale pro archivaci je to problém, který je nutné vyřešit. Řešení se nabízí hned několik. Můžeme provést archivaci tak jak je, spolu s pomocným softwarem. Toto řešení ale není vůbec vhodné, protože nám opět nic nezaručí, že tento software půjde vůbec kdy v budoucnosti spustit. Produkty v počítačovém odvětví zkrátka zastarávají příliš rychle. Dalším řešením je provádět emulaci prostředí, ve kterém by takový software běžel. To klade opět další nároky, hlavně se musí i tato mezivrstva pravidelně kontrolvat, zda ještě vyhovuje současnému trendu, abychom o data nepřišli úplně. Všechny tyto postupy ale mají společné nevýhody. Ke všem se mohou tvořit separátní metadata vztahující se ke zpracování takových dat v archivu, režie na správu různorodých dat se zvyšuje. A mnohem větší problém představuje například indexace takových dat pro snadné prohledávání. Pro každý takový nekompatibilní datový formát musíme vymýšlet další metody indexace a tyto nadále udržovat. Protože je toto vše velmi nevýhodné, existuje další metoda - převod na trvanlivější formát. Trvanlivější ve smyslu definovaného standardu a otevřených a podporovaných specifikací. Nejlepší samozřejmě je data přímo v takových formátech vytvářet, jak naznačovaly body nahoře.

Další problém je zastarávání hardwaru. Je třeba zvolit takové médium, které bude mít příznivé parametry trvanlivosti, ceny a kapacity. Pokud bude trvanlivost malá, budou se muset data relativně často přenášet na nová média téhož typu nebo migrovat celý archiv na nějaké budoucí médium. Velmi důležitá je také obsluha média, čímž rozumíme to, jak snadno se médium při manipulaci (skladování) poškodí a jak dobře se s takovým poškozením vyrovnává. Zajímavý je rovněž pohled na nechtěný přepis média. Doporučují se tedy buď write-once média nebo takové nosiče, které poskytují dobrou zápisovou ochranu. Anglický Národní archiv provedl průzkum různých medií z hlediska klíčových parametrů pro archivaci dat a dospěl k následující tabulce (1=nesplňuje, 3=plně splňuje):

Médium	***CD-R***	***DVD-R***	***Zip Disk***	***3.5" Floppy***	***DLT***	***DAT***
Trvanlivost	3	3	1	1	2	1
Kapacita	2	2	1	1	3	3
Odolnost	2	2	1	1	3	3
Vyspělost a nezávislost	3	2	2	3	2	2
Cena	3	2	1	1	3	3
Citlivost	3	3	1	1	3	2
Celkem	16	14	7	8	16	14

Zdroj: http://www.nationalarchives.gov.uk/preservation/advice/pdf/selecting_storage_media.pdf

Škoda, že v přehledu nejsou zahrnutá vysokokapacitní (nebo dokonce distribuovaná) datová úložiště.

Standardy
Jak již bylo řečeno, doporučuje se směrovat vývoj ke standardům, protože se věří, že otevřené a používané standardy pomohohou k delší životnosti digitálních dat a jejich snadnější migraci na případný další formát či platformu.

Můžeme rozlišit tři druhy standardů

- de-facto standard - standard obecně používaný, rozšířený a známý
- veřejná specifikace - většinou standard nějakého konsorica
- de-jure standard - standard definovaný standardizační organizací, například ISO normy

Hned z toho vidíme, že orientace na standardy nemusí být vůbec tak jednoduchý úkol, jak se zprvu zdálo. Aby toho nebylo málo, můžeme k tomu přidat další zmatek. Ten se týká toho, že ne každý implementuje do svých produktů standard úplně ideálně. Někteří se odchylují od specifikací, například tím, že implementují jen "vhodné" části standardu (třeba z časových nebo finančních důvodů) nebo naopak standard po svém "obohatí". Standard také není neco navěky neměnného, vyvíjí se. Musíme se pak potýkat s různými verzemi téhož standardu, přičemž nejstarší s nejnovější ani nemusí být kompatibilní.

Dneska již existuje standard skoro pro všechny oblasti (jmenujme třeba datové formáty (GIF, JPEG, DjVu), formáty identifikačních dat, formáty výměny dat (XML) a další). To situaci také komplikuje a znepřehledňuje orientaci. Rovněž to, že něco je standard, ještě nezaručuje, že to bude za 50 let stále aktuální (pořád to ale usnadní migraci).

Zvláštním standardem, který je v oblasti uchovávání digitálních dat velmi důležitý, je standard pro jednoznačnou identifikaci digitálního objektu. Přesněji bychom asi měli říci standardy, protože je jich opět několik. Dnešním asi nejpoužívanějším identifikátorem je URL (Uniform Resource Locator). Jeho problém ale je, že spíše než na objekt ukazuje na jeho umístění. To se ukazuje jako problém zvlášt při přesunu takových dokumentů (nebo dat obecně), protože při změně jejich umístění se mění URL a jeho aktualizace všude, kde je na něj odkazováno, je nemožný úkol. Vzniká tak bezpočet odkazů, které ukazují "do prázdna". To je při archivaci digitálních dat rovněž nežádoucí.

Bylo tedy definováno několik dalších standardů. Jedním z nich je URN, který obsahuje jednak identifikaci jmenného prostoru a potom ještě identifikaci v rámci tohoto prostoru (http://www.ietf.org/rfc/rfc2141.txt). Od tohoto schematu je odvozen Handles Systém, který je dále rozšiřuje o funkcionalitu potřebnou pro správu takových dat (např. duševní vlastnictví). Společným rysem těchto identifikátorů je, že již neukazují na žádné místo. Migrace dokumentu tedy není problém. Problém ale je takový dokument lokalizovat. Je tedy potřeba k takovým identifikátorům postavit překladače (buď na principu DNS nebo pluginů do prohlížečů). Taková služba nám pak umí vyhledat po zadání jednoznačného identifikátoru aktuální umístění požadovaných dat (většinou HTTP redirect). Na podobném principu funguje i PURL (Persistent URL). Důvodem je usnadnění vyhledávání takto indentifikovaných zdrojů, protože současné běžné webové prohlížeče neumožňují takové identifikátory překládat.

Projekt PADI shrnuje mnoho takových myšlenek ohledně všech aspektů týkajících se archivace digitálních dat v dlouhodobém měřítku. Soustřeďuje se na hlavní a společné aspekty, které by měly všechny projekty chystající se úkol archivace digitálních dat řešit vzít v potaz. Nedává sice jasné a přesné direktivy jak archivovat texty, obrázky, neklade přímé požadavky na jejich kvalitu a formáty, ale spíše směruje k tomu jak se co nejlépe rozhodnout. Ukazuje nám, jak podobné věci řeší současné knihovny různě ve světě, i jaké kvůli tomu prosazují zásahy do současného práva.

DC Metadata:

```
<link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" />
<meta name="DC.Title" content="Recommended Practices for Digital Preservation" />
<meta name="DC.Creator" content="Jaroslav Kortus" />
<meta name="DC.Subject" content="PADI" />
<meta name="DC.Subject" content="digital" />
<meta name="DC.Subject" content="preservation" />
<meta name="DC.Description.abstract" content="Esej do kurzu Digitalni knihovny, FI MU, Brno 2006."
/>
<meta name="DC.Publisher" content="Jaroslav Kortus" />
<meta name="DC.Date.created" content="7.1.2006" />
<meta name="DC.Format" scheme="IMT" content="application/pdf" />
<meta name="DC.Format.medium" content="computerFile" />
<meta name="DC.Format.extent" content="100KB" />
<meta name="DC.Identifier" content="http://www.fi.muni.cz/~xkortus/esej.pdf" />
<meta name="DC.Source" scheme="URL"
content="http://www.nla.gov.au/preserve/digipres/digiprespractices.html" />
<meta name="DC.Language" scheme="RFC1766" content="cs" />
```