Zajištění dlouhodobé čitelnosti plnotextových souborů, zamezení výskytu duplicitních souborů

SOUHRN : REVIZE 14.10.2011
Dokument je revidován v ročním cyklu.

Tento dokument je průběžně aktualizován na základě nových informací a vývoje v oblasti Digital preservation.

Plán zajištění dostupnosti

Z pohledu rizika zachování digitální informace hrozí v dlouhodobém horizontu zastarání media, případně hardwarového prostředí, ve kterém je médium čitelné. Softwarovému prostředí a formátu dokumentu hrozí ztráta kontextu, ve kterém informace vznikla.

Migrace představuje převedení obsahu do „současného" formátu – nejlépe otevřeného a standardizovaného. Proces migrace může potenciálně přinést rizika: nekonzistence informace, ztráty funkcionality, ohrožení kvality.

Čitelnost souborů

Definujeme-li pro NUŠL čitelnost souboru, jako možnost získat obsah dokumentu, pak zajištění čitelnosti a jeho průběžné ověření může být prováděno indexačním nástrojem FAST a případným výpisem na úrovni logu.

Každý soubor je pro indexaci otevřen a je z něj extrahován čistý text pro samotnou indeaxaci. Z tohoto pohledu je ověření zajištěno udržováním aktuálního indexu vyhledávacího nástroje.

Nečitelný soubor prostřednictvím filtrů vyhledávacího nástroje pravděpodobně nepůjde ani otevřít pomocí asociované aplikace a bude muset být vymazán a nahrazen.

Doporučení

Doporučení jednoznačně v případě NUŠL směřuje k plánovité migraci formátů. Zejména s ohledem na přístupnost prostřednictvím Internetu je emulace nevhodná.

O provedení migrace bude informován administrátorem správce zdrojového úložiště.

Statistika užití formátů

Za nejpodstatnější informace s ohledem na preferované formáty lze považovat budoucí statistiku jejich využití v rámci záznamů v NUŠL.

Ta poskytne okamžitý přehled o využívaných formátech a poskytne správci informace pro rozhodnutí o případném postupu migrace.

Při změně preferovaných formátu, případně jejich verzí, bude provedena úprava programu pro výpis statistiky.

Případné větší množství souborů určených pro migraci bude řešeno operativně.

Formáty dokumentů

Aplikace CDS Invenio je schopna přijmout binární data v libovolném formátu. Také vyhledávací systém FAST ESP je schopen přes 200 formátů indexovat.

Nejpodstatnějším parametrem pro rozhodnutí bude rozšířenost formátu. Bez ohledu na jeho dlouhodobou stabilitu, či komerční původ lze konstatovat, že právě pro rozšířené formáty budou dostupné nástroje na případné migrace.

Z toho pohledu mohou být formáty MS Office (DOC, XLS, PPT) sice patřit mezi formáty s nízkou trvanlivostí avšak to budou formáty snadno migrovatelné. A to ať již do novější verze, tak například do formátu PDF.

PDF jednoznačně patří mezi doporučené formáty, zejména PDF úrovně A. V případě migrace editovatelných formátů, to však může znamenat ztrátu funkcionality.

Preferované a doporučené formáty NUŠL

Mezi preferovanými formáty je vždy uveden základní (neformátovaný), a další rozšířené typy.

Typ dokumentu	Preferované formáty	Doporučené formáty
Textové dokumenty	PDF	TXT, DOC-DOCx, XML (HTML)
Tabulkové dokumenty	PDF	CSV, XLS-XLSx
Prezentace	PDF	(PPT -PPTx)
Grafické formáty	JPG	JP2, PNG, TIFF
Zvuk	MP3	MP3, AVI
Video	MP4	MPEG-1 (-2), AVI

Seznam preferovaných formátů má pouze doporučující charakter. Může být v budoucnu upravován a doplňován dle aktuálního vývoje využívání (na základě statistik).

Preferované formáty zvolené pro uchovávání dokumentů budou využity bez migrace. Ostatní mohou být migrovány do základního formátu – preferovaného (např. PDF) dodatečně v pravidelných cyklech revize úložiště.

Podporované formáty pro indexaci

Fast ESP podporuje pro indexaci formáty uvedené v příloze. Indexace prakticky znamená extrahování čistého textu z dokumentu a zaznamenání výskytu řetězců písmen do indexu.

Z pohledu udržitelnosti informace a čitelnosti souboru lze formáty považovat za vhodné, lze-li z nich získat touto formou textovou informaci.

Indexační služba FAST ESP podporuje následující typy vstupních souborů.

- Textové formáty
- Tabulkové formáty
- Prezentace
- Grafické formáty
- Komprimované formáty
- Ostatní formáty

Existující omezení, kdy není možné data indexovat jsou například heslem chráněné dokumenty, vložená písma, poznámky pod čarou, referenční čísla.

Doporučení zahrnuje průběžné vytváření statistik a vytvoření modul, který by analyzoval logy indexačního systému a vypisoval do uživatelského rozhraní informaci o nečitelnosti souborů (např. formou seznamu identifikátorů).

Formát pro záložní archivaci

FORMAT PDF/A-1a

Část 1 PDF/A ISO standardu [ISO 19005-1:2005] je formou Adobe PDF verze 1.4. Byl vytvořen se záměrem dlouhodobého uchování dokumentů, pro které je používán formát PDF.
Úroveň A (PDF/A-1a) indikuje kompletní shodu s požadavky ISO 19005-1.

Vytváření záložních kopií v archivačním formátu připadá v úvahu až s vyššími počty záznamů v úložišti [řádově v desetitisících]

Použité zdroje

Bernas, Jiří. Národní digitální archiv. Knihovna [online]. 2009, roč. 20, č. 1, s. 22-29 [cit. 2011-10-14]. Dostupný z WWW: <http://knihovna.nkp.cz/knihovna91/bernas.htm>. ISSN 1802-8772.

Planets [online]. 2011 [cit. 2011-10-14]. Planets project EU. Dostupné z WWW: <http://www.planets-project.eu/>.

Planet project EU [online]. 2011 [cit. 2011-10-14]. IntroductiontoDigitalPreservation-TechnicalSummary. Dostupné z WWW: <http://www.planets-project.eu/training-materials/IntroductiontoDigitalPreservation-TechnicalSummary-Final.pdf >.

Wikipedia [online]. 2011 [cit. 2011-10-14]. Digital preservation. Dostupné z WWW: <http://en.wikipedia.org/wiki/Digital_preservation >.

Digital preservation [online]. 2011 [cit. 2011-10-14]. Digital preservation. Dostupné z WWW: <http://www.digitalpreservation.gov/>.