Posouzení vlastností elektronických dokumentů z hlediska jejich dlouhodobého uchovávání

Ing. Ivan Zderadička, spolupracovník společnosti Central European Advisory Group

S pokračujícím rozvojem informačních technologií a s jejich nepřetržitým pronikáním do všech oborů lidské činnosti se dostává do popředí otázka jejich dlouhodobého uchování (archivace).

V současné době již řada dokumentů vzniká pouze v elektronické podobě, ať již se jedná o klasické textové dokumenty, či zvukové a obrazové záznamy, které lze také zahrnout do širšího vnímání kategorie elektronických dokumentů. Potřeby jejich dlouhodobého uchování vyplývají jednak z legislativních požadavků, jednak z případné historické hodnoty obsahu elektronického dokumentu.
Dokumenty existující v papírové formě jsou často převáděny do formy elektronické z důvodu jejich snadnějšího zpracování.

Požadovaná doba uchovávání elektronických dokumentů může přesahovat desítky let a pro určité skupiny dokumentů můžeme předpokládat nutnost neomezeného uložení. Dále předpokládáme uložení dokumentů pouze v digitální podobě, která poskytuje značné výhody z hlediska nákladů a zpracovatelnosti uložených dokumentů. Tyto požadavky přináší několik klíčových problémů, které systémy dlouhodobého uchování elektronických dokumentů musí řešit. Jedná se především o faktor stárnutí informačních technologií, který zasahuje jak datová media, na kterých jsou dokumenty uloženy, tak i počítačové systémy (týká se hardware i software), na kterých byly dokumenty vytvořeny. Zde se jedná o tzv. morální stárnutí, kdy jsou stávající systémy nahrazovány novými, které již nemusí být schopny zpracovat starší dokumenty.
Aby se předešlo problémům, kdy elektronické dokumenty nebudeme moci v budoucnosti použít, musí organizace zodpovědné za dlouhodobé uchování dokumentů plánovat vhodné strategie (a na ně navazující praktická opatření), které zabrání znehodnocení elektronických dokumentů v budoucnosti. V současné době v této oblasti existují dva hlavní směry:

- Strategie migrace - kdy dokumenty, u kterých se v průběhu času ukazuje, že jejich formát by nemusel být použitelný v budoucnosti, se převedou (migrují) do nového formátu, který zachová všechny jejich potřebné vlastnosti a přitom bude lépe podporovaný budoucími systémy.
- Strategie emulace – formát dokumentu se nezmění, ale je vytvořeno specifické programové vybavení, které umožní interpretaci elektronických dokumentů na nových platformách.

Tyto strategie nejsou vzájemně se vylučující, například je možné, že organizace použije na nějaké formáty dokumentů jednu strategii a na další jinou. Je však zřejmé, že vhodná spolupráce mezi původcem dokumentu a archivem při tvorbě a uplatnění těchto strategií hraje klíčovou roli. Volbou

vhodných formátů již při vytváření dokumentů lze značně usnadnit dlouhodobé uchování a výrazně snížit náklady s tím spojené. Původce dokumentu by proto měl brát na zřetel vlastnosti elektronických dokumentů z hlediska dlouhodobého uchování a zohlednit je při výběru svého informačního řešení.

Při výběru vhodných formátů je třeba vzít v úvahu dvě hlavní skupiny faktorů:

- faktory vztahující se ke kvalitě nebo speciální funkcionalitě zobrazení obsahu dokumentu,
- faktory, které ovlivňují trvanlivost digitálního formátu.

Jako prvotní při posuzování určitého formátu je nutné zvážit, zda daný formát je schopen uchovat a podat obsah elektronického dokumentu v takové reprezentaci, která bude vyhovovat jak současným, tak i budoucím uživatelům. Je nutné zvážit všechna potenciální možná budoucí využití dokumentu. Tyto faktory se týkají jak kvality reprezentace (jako je např. maximální rozlišení, barevná věrnost u obrázků, minimální zkreslení u zvukových záznamů atd.), tak speciální funkcionality, kdy při reprezentaci je např. potřebná určitá interakce s uživatelem (např. zvětšení a zmenšení u obrázků nebo zastavení a zpětné převinutí u video záznamů). Tyto faktory se mění podle žánru a v daném žánru též v závislosti na formě. Podstatné charakteristiky jsou jiné pro textový dokument, zvuk, obraz nebo video.

Pro statické obrázky budou například důležité následující faktory:

- Normální prohlížení zahrnuje prohlížení na obrazovce a možnost zvětšení. Je podporován standardní tisk;
- Rozlišení – stupeň v jakém lze reprezentovat obsah s velkým rozlišením v daném formátu. U bitových obrázků rozlišení souvisí s počtem bodů na jednotku plochy a počtem barevných odstínů definovaných pro jeden bod;
- Barevná věrnost – míra možnosti řídit rozsah barev (tzv. barevný gamut) reprezentovaný v daném obraze;
- Podpora pro grafické efekty a typografii – např. podpora vrstev;
- Další specifická funkcionalita může zahrnovat vektorovou grafiku, 3-D modely aj.

Při posuzovaní trvanlivosti digitálního formátu dokumentu je vhodné vyjít z definované sady vlastností, které lze snadno posoudit pro různé dostupné formáty, bez závislosti na přesných technických a obsahových detailech. Literatura uvádí sedm faktorů, které mohou být pro toto posouzení použity:

Otevřenost formátu
Uchování obsahu v daném formátu není možné bez detailní znalosti toho, jak jsou ve formátu ukládány informace na úrovni jednotlivých bitů a bytů. Otevřenost znamená, že detailní specifikace formátu je veřejně dostupná (nejde zde ani tak o nutnost standardizace jako spíše

o širokou dostupnost kvalitního a kompletního popisu formátu dokumentu). Tato znalost je velmi důležitá pro možnost ověření integrity formátu ukládaného dokumentu. Pro neproprietární, otevřené formáty jsou většinou validační nástroje a podrobná dokumentace formátu lépe dostupné než pro formáty proprietární, kdy je jejich dostupnost omezena.

Rozšířenost
Odkazuje na stupeň rozšíření využívání formátu zainteresovanými uživateli a tvůrci obsahu. Značná rozšířenost formátu snižuje pravděpodobnost jeho rychlého zastarání. Pro rozšířený formát je větší pravděpodobnost vzniku migračních a emulačních nástrojů v rámci IT průmyslu, bez nutnosti, aby se archivní instituce na vzniku těchto nástrojů přímo podílela. Dobrými kritérii rozšířenosti formátu je standardní dodávka nástrojů pro práci s daným formátem přímo s PC, nativní podpora formátu přímo ve Webových prohlížečích a/nebo existence řady konkurenčních produktů pro tvorbu obsahu, manipulaci s obsahem anebo získávání obsahu z daného formátu.

Transparentnost
Transparentnost je faktor popisující přístupnost obsahu uloženého v daném formátu pro přímou analýzu s použitím základních nástrojů, jako je např. jednoduchý textový editor. Digitální formáty, v nichž je relevantní informace uložena přímo a jednoduše, lze snadněji migrovat, vytvořit pro ně odpovídající prohlížeče a v budoucnu jsou též vhodnější pro tzv. digitální archeologii (jedná se o přístup, kdy lze získat obsah dokumentu neznámého formátu pouze ze studia dokumentu samotného, neboť dodatečné informace již nejsou dostupné) . Čitelnost textových dokumentů je příznivě ovlivněna použitím standardního znakového kódování (např. UNICODE s použitím UTF-8 kódování) a ukládáním v posloupnosti odpovídající přirozenému čtení. U netextových informací jsou základní reprezentace transparentnější než reprezentace optimalizované pro úspornější ukládání, rychlejší zpracování apod.. Jako příklad základní reprezentace uveďme uložení rastrové grafiky jako nekomprimovanou bitovou mapu. Šifrování je samozřejmě v přímém rozporu s transparentností, komprimace též transparentnost potlačuje. Na druhou stranu řada zvukových, obrazových a video záznamů nebude nikdy z praktických důvodů ukládána v nekomprimovaném formátu a to ani v okamžiku vzniku.

Sebedokumentace
Proces dlouhodobého uchovávání elektronických dokumentů se výrazně ulehčuje pokud jsou jejich součástí doplňující metadata (dodatečné informace), obecné, technické i administrativní povahy, popisující účel dokumentu, jeho vznik a následné procesy v rámci jeho životního cyklu. Všechny tyto informace umožňují a ulehčují správné zobrazení a pochopení dokumentu v dlouhodobé perspektivě.
Pro dlouhodobé uchovávání jsou typicky rozlišovány následující kategorie metadat:

- Reprezentace (popisují formát dokumentu, způsoby jeho

zobrazení včetně závislostech na různých programech a platformách);

- Reference (identifikace a popis vlastního obsahu);
- Kontext (důvody pro vytvoření a uchování dokumentu);
- Stabilita (informace umožňující ověření integrity dokumentu);
- Původ (data zachycující vytvoření a vlastnictví dokumentu).

Je vhodné, aby byla metadata uchovávána společně s elektronickým dokumentem (buď přímo vložena do formátu elektronického dokumentu jako „metadata hlavička“, nebo zabalena spolu s dokumentem do vhodné digitální obálky, kdy takto spolu tvoří jeden digitální objekt) a bylo takto zaručeno jejich společné uchování. Nicméně archivní systémy mohou zpětně opět tato metadata extrahovat do specifických indexů (s tím, že samozřejmě zůstanou uložena i společně s dokumentem), které pak slouží pro vyhledávání a třídění uchovávaných elektronických dokumentů.

Vnější závislosti

Tento faktor popisuje stupeň závislosti daného formátu na vnějších faktorech – jako jsou specifický hardware, software, nebo např. specifické služby, které musí být dostupné online. Dále je nutné také zvážit jak bude možné řešit tyto závislosti v budoucnosti (dané závislosti již nemusí fungovat) – zda i bez nich bude formát použitelný, případně jak bude možné nahradit nefunkční závislosti .

Vliv patentů

Možnosti a náklady archivních organizací na uchovávání elektronických dokumentů v určitém formátu mohou být zásadně ovlivněny patenty souvisejícími s daným formátem. Existence patentů může zpomalovat vývoj volně dostupného software pracujícího s daným formátem a ovlivňuje cenu komerčního software. Dále může znamenat dodatečné náklady spojené s uchováváním dokumentů. Problém není v existenci patentu, ale v podmínkách, které mohou držitelé patentu uplatnit. Pokud licenční ujednání zahrnuje poplatky odvozené od používání (jako je např. poplatek odvozený od každého zobrazení elektronického dokumentu), náklady mohou být v dlouhodobém horizontu nepředvídatelné a vysoké.

Ochranné mechanismy

K efektivní správě digitálního obsahu a k poskytování přiměřených služeb budoucím uživatelům musí mít správce archivu možnost replikovat uložený obsah na nová média, migrovat anebo jinak normalizovat tak, aby obsah byl odpovídajícím způsobem využitelný v budoucnu při použití nových technologií. Dlouhodobé uchování obsahu je obtížné anebo může být přímo znemožněné při aplikaci některých ochranných mechanismů. Například mechanismy, které svazují formáty s určitými fyzickými medii jsou zcela nevhodné pro dlouhodobé uchovávání.

Jak z hlediska faktorů kvality a funkčnosti, tak faktorů trvanlivosti elektronických dokumentů je velmi důležitá přesná identifikace formátu dokumentu. V současné době existují tisíce různých formátů, z nichž řada má verze a podverze. Obecně užívané názvy formátů, včetně koncovek souborů jako jpg, pdf, avi, jsou příliš generické a neumožňují rozlišit různé typy a verze. Proto přesná specifikace formátu včetně verzí a podverzí a případných dalších detailů (jako je např. kompresní algoritmus pro obrazová data aj.) musí být připojena k elektronickému dokumentu a předpokládá se, že pro určitý formát budou preferovány pouze určité verze tak, aby se omezila variabilita formátů na této úrovni.

V praxi při výběru vhodných formátů pro dlouhodobé uchování elektronických dokumentů půjde vždy o dosažení rovnováhy mezi výše popsanými faktory kvality, funkčnosti a trvanlivosti daného formátu. Často si budou určité faktory konkurovat, kdy například velmi rozšířené formáty jsou proprietární a nejsou k nim veřejně dostupné jejich popisy. Nebo naopak formát, který je dobře popsaný a popisy jsou veřejně dostupné, používá komplikovaný algoritmus ztrátové komprese dat. V tomto případě tedy otevřenost nahrazuje transparentnost. Pro elektronické dokumenty s vysokou historickou hodnotou, kde hrají významnou roli faktory pro specifickou funkcionalitu formátu (např. řekněme u historických map, kde budou důležité speciální funkce zobrazení a prohlížení), mohou zase tyto převážit nad faktory pro trvanlivost formátu.

Tento článek vznikl v rámci účelové podpory Ministerstva informatiky na projekt č. YA512006002 „Dlouhodobé uchovávání elektronických dokumentů se zaručeným elektronickým podpisem“.

Ministerstvo informatiky České republiky