

Důvěryhodný elektronický archiv pro dlouhodobou archivaci, požadavky a jejich řešení

Petr Hönig*

petr.honig@i.cz

Abstrakt: Článek formuluje požadavky kladené na důvěryhodný elektronický archiv (DEA). Definuje důvěryhodnost elektronických dokumentů (ED). Jeho těžiště je v popisu informačních technologií (IT) používaných k vytvoření DEA. Zabývá se elektronickým podpisem, časovými razítky, linkováním hashů, emulací a migrací. Popisuje některá úspěšně realizovaná řešení ve světě na úrovni Národních archivů. Článek si klade za cíl srozumitelnou formou informovat o IT technologiích, které pomohou vyřešit problémy s důvěryhodností ED, a to jak během aktivní části jejich životního cyklu, tak při dlouhodobé archivaci. Je určen archivářům, knihovníkům, ale i zájemců ve velkých institucích i malých firmách.

Klíčová slova: dlouhodobá čitelnost, důvěryhodný elektronický archiv, elektronický podpis, emulace, hashovací funkce migrace, linkovaný hash

1 Úvod

Elektronické dokumenty (ED) mají před papírovými mnoho výhod. Jejich vyjmenovávání na tomto fóru by bylo nošením sov to Athén. Jsou známé již více než 10 let a byly shrnuty do termínu „bezpapírová kancelář“. Další impuls dostala vize digitalizace při zavedení I.Certifikační Autority v r. 2002 a tím i reálné možnosti podepisovat e-podpisem.

Ve prospěch ED bych uvedl jen málo známou, a přesto velmi zajímavou zkušenost Národního archivu UK, že doručení objednávky ED přes WEB je 46 x levnější než doručení dokumentu papírového [6].

Není bez zajímavosti, že ve zveřejněném průzkumu provedeném na Magistrátu hlavního města Prahy je konstatováno, že v případě elektronické komunikace je dvakrát větší úspěšnost reakce a několikanásobně kratší doba reakce oproti komunikaci papírové.

Dalo by se očekávat, že usnadnění a zrychlení manipulace s ED, ruku v ruce se silným ekonomickým tlakem, způsobí během několika málo let, že na chodbách všech podniků i úřadů zcela zmizí sekretářky s haldami papírových dokumentů. Z kanceláří zmizí skříně šanonů a mladí budou nechápavě kroutit hlavou nad pošetilostí lidí, kteří kvůli manipulaci s informacemi, často pomíjivými, káceli drahocenné stromy rostoucí desítky let.

Jaká je však skutečnost po deseti letech? Opak té růžové vize je pravdou. Podíl ED na úřadech dokonce rok od roku klesá. Zejména u smluv, faktur a jiných důležitých dokumentů se úředníci bojí spoléhat na ED a je vůbec úspěch, když používají alespoň hybridní dokumenty: pro interní potřebu ED, pro externí komunikaci (zejména pro soudy) dokumenty papírové. Digitalizace v ČR zřejmě bude muset projít mezietaou hybridních dokumentů, kterou nebude možno přeskochit.

I já jsem však byl velmi překvapen následujícím výrokem Billa Gatese : Spotřeba papíru se zdvojnásobuje každé čtyři roky a 95 % všech informací v USA zůstává na papíře, jen 1% se ukládá elektronicky. Papírů přibývá rychleji, než je digitální technologie schopná adsorbovat.

*Analytik-konzultant divize Správa a řízení dokumentů, ICZ a. s.

Tedy ani vyspělé USA a všemocný Microsoft si s digitalizací dokumentů neporadily. Otázkou je proč. Jsou zde jistě důvody subjektivní : konzervativnost, neznalost, nechť učít se novým věcem, ale i strach, že zavedení nových technologií přinese ztrátu zaměstnání. K překonání těchto překážek lze nabídnout jen knižecí rady : osvětu, poskytování výhod (např. rychlejší vyřízení žádosti a nižší poplatky), jistotu, že úředník nebude propuštěn.

Existují však i velmi vážné objektivní důvody. Právní nejistota, jak se soudy budou chovat k ED, které nemají odpovídající papírový originál. Nedůvěra soudů není bezdůvodná: s paděláním papírových dokumentů již mají několik set let zkušeností. Odhalení papírového padělku je usnadněno vazbou informace na médium. Padělek se dá odhalit i po velmi dlouhé době, vzpomeňme na Rukopisy. U ED je však dodatečné odhalení padělku, pokud proti padělaní není ED ošetřen předem, dodatečně nemožné. Nechráněný ED může hacker i na dálku zcela smazat, pozměnit nebo jej nahradit jiným. Ještě větší nebezpečí hrozí ED uvnitř institucí, které je zpracovávají. Celosvětová statistika říká, že 90 % útoků na banky, finanční instituce a archivy byla provedena vlastními zaměstnanci.

V neposlední řadě existuje i oprávněná obava, že ED nebude za krátký čas čitelný v důsledku zastarání technologie. Je ironií osudu, že formáty používané před deseti lety, jako neformátovaný ASCII nebo unicode UT-8 budou asi čitelné i za mnoho set let, kdežto obrovské množství formátů vznikajících jako houby po dešti jen pro potřeby jedné aplikace asi nepřežijí ani pár let. Potřeby rychlého přístupu k datům jsou v rozporu s požadavky dlouhodobé čitelnosti.

2 DEA – řešením objektivních překážek

Elektronickým dokumentem (ED) budeme zde rozumět nestrukturovanou informaci uloženou v elektronických souborech. Strukturovanými daty v relačních databázích se tedy zabývat nebudeme. Prakticky tedy ED rozumíme soubor libovolného formátu (s libovolnou příponou) a libovolného obsahu (textové, obrázky, video, audio, výkresy, chemické vzorce, databáze atd.). Výjimkou je soubor XML, kde je informace strukturována a přesto je ED. Elektronickým archivem rozumíme úložiště těchto ED, které jsou již v poslední, tj. archivační fázi svého životního cyklu. To se již ED, až na výjimky, nemění.

Důvěryhodným elektronickým archivem (DEA) pak rozumíme takový elektronický archiv, který obsahuje právně nezpochybnitelné ED. Za DEA v širším slova smyslu lze považovat i archivy, které plní podstatné (a explicitně deklarované) požadavky vybrané z úplného souboru požadavků, kladených na plně legální DEA, viz dále. Pak jde o důvěryhodnost opírající se autoritu a důvěryhodnost instituce, která DEA provozuje. Postačuje-li to té instituci i jejím partnerům, pak je takovýto „podnikový“ DEA velmi užitečný a je i významnou konkurenční výhodou.

Tato definice je prostá a jednoznačná. Problém nastane, když se zeptáme, jaký soud a podle jakých kritérií bude legálnost ED posuzovat. V EU je situace nejednoznačná. Nejprísnejší je legislativa v nástupnických zemích bývalého Rakouska–Uherska. Např. rakouský notář papírovou smlouvu podepsanou smluvními stranami podepíše sám, naskenuje ji a ED elektronicky podepíše svým e-podpisem certifikovaným státní autoritou. V ČR soud vyžaduje originály papírových dokumentů. Ve správním řízení je nepřípustná kopie papírového dokumentu, avšak ED s elektronickým podpisem je dle rozhodnutí Nejvyššího soudu akceptován, ačkoliv není řečeno, co se stane po skončení platnosti jeho certifikátu.

Situace v ČR bude snad vyřešena již příští rok, kdy má vstoupit v platnost nový zákon o pravidlech konverze papírového dokumentu na ED a o použití e-podpisu obecně. Pokusme se na základě diskuse probíhající mezi odbornou veřejností v EU definovat vlastnosti, jaké by

DEA měl mít, aby jeho ED byly právně nezpochybnitelné. Skutečná budoucí legislativa zemí EU se od této definice nemůže příliš odchylovat.

3 Požadavky kladené na funkčnost DEA a vlastnosti jeho ED

1. Integrita obsahu ED
2. Autenticita obsahu ED
3. Autenticita osob, spojených s ED
4. Identifikovatelnost archivačních metadat ED
5. Trvalá čitelnost ED
6. Důvěryhodnost kopie ED z DEA, pořízené pro manipulaci mimo něj
7. Validací vazby mezi ED v DEA i mimo něj
8. Možnost legální úpravy ED
9. Možnost rychlého vyhledání ED v DEA, tvorba vyhledávacích metadat
10. Přístupová práva k ED v DEA
11. Audit událostí spojených s manipulací s ED v DEA

Všechny tyto požadavky musí být splněny po celou dobu uložení. To může být bez omezení. Po té může být ED skartován nebo odlit“ do „statického archivu“ na nepřepisovatelné médium, kde na něj již nejsou kladeny přísné požadavky DEA.

3.1 Integrita obsahu ED

Integritou obsahu chápeme to, že ED nesmí být z DEA vyřazen, a to ani záměrně ani technickou poruchou. Nesmí být nelegálně pozměněn a musí být naprosto identický s tím ED, který byl do DEA vložen. Jednotlivé DEA se mohou lišit tím, zda vyžadují jen neměnnost samotné informace nebo i formy. Té se někdy poeticky říká „chuť a vůně“ (barvy, fonty, grafika, obrázky, způsob obtékání obrázků textem).

3.2 Autenticita obsahu ED

Autenticitou se rozumí to, že identita ED je opravdu taková, za jakou je prohlašována. ED byl vložen do archivu opravdu v době, která je v něm a jeho metadatech uvedena. Časové údaje uvedené v ED i metadatech jsou pravdivé. Archiv neručí za změny provedené **před** vložením do DEA. Vstupní kontrola může kontrolovat jen pravidla, definovaná politikou samotného DEA. Testy autenticity jsou popsány v [15].

3.3 Autenticita osob spojených s ED

Tato funkčnost se týká původce ED, jeho autora, schvalovatelů, odesilatele, archiváře, atd. Autenticita se dá prokazovat e-podpisem. V DEA mohou být i ED bez e-podpisu. Jestliže je však e-podpis uveden, musí být jeho platnost udržena i po skončení platnosti jeho certifikátu.

3.4 Identifikovatelnost archivačních metadat

Identifikovatelnost má dva dva účely:

- specifikovat režim zacházení v DEA
 - zařazení do kategorie, určující stupeň požadované důvěryhodnosti
 - formát, kódování
 - popis, zda má být uchovávána jen vlastní informace nebo i „chuť a vůně“

- zvýšit důvěryhodnost ED a celého DEA uchováním dalších údajů souvisejících se vznikem ED, zasláním do archivu a testováním při ukládání. Důkladnost archivačních metadat je důležitým kritériem pro posouzení důvěryhodnosti DEA. Tu ovlivňuje i dobrá pověst instituce, která DEA provozuje [15]. Sem patří celý komplex personálních a technických opatření včetně bezpečnostních certifikací, ale to je již na jiný článek.

3.5 Trvalá čitelnost ED

ED musejí být čitelné, a to i poté, kdy technologie, kterou byly vytvořeny a aplikace, kterými jsou prohlíženy, nejsou již pro svoji zastaralost běžně dostupné. To je tvrdý oříšek.

3.6 Důvěryhodnost kopie ED z DEA, pořízené pro manipulaci mimo něj

U elektronických i papírových kopií musí být zajištěno, že to jsou kopie originálního ED z DEA, nebo dokonce přímé kopie. Technologie použité k tomuto důkazu lze úspěšně využít i v aktivní fázi životního cyklu ED.

Na trhu jsou dvě „krabicová“ řešení: „digitální vodoznak“ a „deformovaná rozptylová mřížka“. V první z nich je do „vodoznaku“ vytištěného tiskárnou formou shluku teček a čárek na vytištěné papírové kopii zapsána nějaká kontrolní informace. Vodoznak může vytvářet logo firmy nebo ilustrační obrázky. Písmenům kontrolního textu odpovídají 2D kombinace teček a čárek. Tento text není okem viditelný. Naskenováním kopie skenerem vybaveným příslušným SW lze zašifrovanou informaci zobrazit. Tato technika nebrání vytvoření kopie z kopie, ale zajišťuje, že je vytvořena z originálu v DEA.

Druhá technika využívá při tisku rozptylu světla na dvourozměrné nepravidelné mřížce. Kontrolní slovo vytištěno tiskárnou do nějakého obrázku je deformováno rozptylem na této mřížce a tím je okem neviditelné. Přiložíme-li sklíčko se stejnou mřížkou na zakódovaný text, kontrolní slovo se zobrazí. Při naskenování nebo překopírování dokumentu dojde ke ztrátě čitelnosti kontrolního textu a lze tak objevit, že nejde o přímou kopii ED z DEA, ale o kopii kopie.

3.7 Validační vazby mezi ED v DEA i mimo něj

Jsou to odkazy na jiné ED, které potvrzují pravdivost uvedených údajů. Např.

- u smlouvy jsou to výpisy z obchodního rejstříku obou partnerů, posudky znalců, údaje z katastru nemovitostí, technické výkresy,
- u faktury to jsou odkazy na smlouvu, objednávku, přejímku nebo předávací protokol.

3.8 Možnost legální úpravy ED

Některé DEA mohou u některých kategorií ED povolovat úpravy a to updatem nebo verzováním. V takovém případě nutno zajistit, aby úprava byla:

- provedena oprávněnou osobou,
- auditována zápisem do logovacího souboru,
- zdokumentována (kdo, kdy, jak, co, proč, jakým pověřením), a to nezpochybnitelně.

Legální úprava je riziková operace, a musí být proto velice pečlivě ošetřena.

3.9 Možnost rychlého vyhledání ED v DEA, tvorba vyhledávacích metadat

DEA musí umět vyhledávat dle atributů vyhledávacích metadat, klíčových slov i fulltextově. Ve vyhledaných souborech rekordů metadat musí být možnost vybírat jednotlivé rekordy a kliknutím zobrazovat jednotlivé ED. Je žádoucí, aby základní vyhledávací atributy metadat byly mezinárodně sjednoceny. Jako neformální standard se ujala sada atributů nazývaná dle místa vzniku (Dublin, Ohio) „Dublin Core“ [4, 5].

Jsou to: název (jméno, pod nímž je ED znám nebo je mu přiděleno), tvůrce (entita tj.osoba, organizace nebo služba zodpovědná za obsah ED), předmět (klíčová slova a fráze, klasifikační kódy), popis (obsahu ED), vydavatel (entita zodpovědná za zveřejnění ED), přispěvatel (entita, která přispěla k obsahu ED), datum (vytvoření ED), typ (obecná kategorie, povaha, žánr), formát (typ média, SW), identifikátor (unikátní v použitém DEA), zdroj (odkud byl dokument získán), jazyk, vazby (na další související ED), pokrytí (geografické, časové, rozsah působnosti obsahu ED), práva (vlastnická, autorská).

3.10 Přístupová práva k ED v DEA

Přístup k ED musí být řízen přístupovými právy přidělenými jednotlivým rolím.

3.11 Audit událostí spojených s manipulací s ED v DEA

Evidovány musí být zejména operace s ED a to včetně osob, které si ED přečetly nebo pořídily (v souladu s přístupovými právy) jeho kopii.

Při vkládání do DEA by ED měly být rozděleny dle typu, důležitosti, doby uložení atd. do kategorií. Na vstupu nelze kontrolovat „správnost“ ED (ta se ostatně neověřuje ani u papírových dokumentů). Lze jen zkontrolovat, zda jsou pro danou kategorii ED splněny podmínky, kladené bezpečnostní politikou DEA. Necht' se jedná o velký podnik a necht' kategorie ED je „Smlouva o nákupu budovy“, pro kterou je definováno, že musí mít e-podpis ředitele. Pak musí být zkontrolováno, zda e-podpis je dosud platný a zda podepsaná osoba byla skutečně v době podpisu smlouvy ředitelem. Za věcnou správnost smlouvy neručí DEA, ale ředitel, který ji podepsal.

4 Implementace DEA

DEA je obvykle implementován na nějakém standardním datovém úložišti, ke kterému je přidána řada procesů probíhajících při vstupu ED, periodicky nebo při zvláštních událostech (např. žádost o zpřístupnění ED, ukončení platnosti certifikátu časového razítka nebo čitelnosti formátu ED).

4.1 Standardní úložiště pro DEA

1. souborový systém (tam jsou uloženy ED) v kombinaci s jednoduchou databází obsahující archivační a výběrová metadata,
2. proprietární repozitory (např. Lotus Notes),
3. RDBMS (Oracle, MS SQL). ED v BLOBech, archivační a vyhledávací metadata v tabulkách,
4. DMS (solistikovaný Dokument Management Systém) způsob uložení jako u RDBMS, ale je přidána další aplikační vrstva pro urychlení přístupu k ED a zvýšení uživatelského komfortu při manipulaci s ED.

Důvěryhodnost DEA lze zvýšit, když výběrová a archivační metadata budou duplicitně uložena i spolu se svým ED. Uživatelem může být široká škála uživatelů končící u soukromníka doma na jednom PC, který pro své ED nepožaduje plnou právní legálnost, ale postačí mu splnění jen některých požadavků. I důvěryhodnost lze odstupňovat v závislosti na tom, čemu je uživatel ochoten věřit (technice, kolegům, třetím stranám nebo ničemu).

Jak požadavky tak i jejich řešení lze výběrem vhodných technologií a nastavením jejich parametrů poměrně jemně odstupňovat a každý uživatel může získat přesně to, co potřebuje a na co má prostředky.

5 Jak splnit požadavky kladené na DEA

Na první pohled je vidět, že mnoho požadavků je splněno automaticky, nebo je lze snadno splnit použitím datových úložišť 3. a 4. typu. Zabývejme se jen těmi zbývajícími:

- Zajištění dlouhodobé integrity a autenticity dokumentu včetně e-podpisu.
- Zajištění dlouhodobé čitelnosti ED.

Objasněme si několik důležitých pojmů: elektronický podpis, hashovací funkce, časové razítko a linkovaný hash.

5.1 Elektronický podpis

Technika je všeobecně známá a proto jen velmi stručně: e-podpis je založen na asymetrické šifrovací technice RSA. Je to metoda založená na skutečnosti, že číslo n , vzniklé součinem dvou prvočísel, není možno na tato dvě prvočísla zpětně rozložit jinak než postupným dělením jednotlivými prvočísly. Je-li je n dostatečně veliké (64 až 128 bitů), pak toto dělení může trvat až 5 let na nejvýkonnějších počítačích. Se vzrůstající rychlostí výpočetní techniky se však tyto intervaly zkracují. Proto je nutno buď zkracovat dobu platnosti certifikátu nebo prodlužovat počet cifer čísla n .

Pomocí jednoduchého algoritmu se vytvoří vzájemně zaměnitelný pár klíčů **soukromý a veřejný**. Ty slouží k šifrování a dešifrování, což jsou dva jednoduché aritmetické algoritmy. Soukromým klíčem se podepisuje (šifrování), veřejným se podpis ověřuje (dešifrování). Veřejným klíčem se nějaká zpráva šifrováním kryptuje (tj. činí ji běžně nečitelnou), dešifrováním soukromým klíčem se odkryptuje (převádí zakryptovanou zprávu zpět na běžně čitelný text). Podepisuje, ověřuje, šifruje a dešifruje jsou jen synonyma pro dva jednoduché algoritmy využívající veřejného nebo soukromého klíče. Ověření podpisu nebo odkryptování zprávy pomocí komplementárního klíče je velmi jednoduché. Jestliže se to někomu povede bez znalosti tohoto klíče, mluvíme o rozluštění šifry nebo o „rozbití kódu“.

V případě RSA je k tomu nutno najít metodou pokusů a omylů soukromý klíč. E-podpis nějaké osoby je nějaký identifikační text (zpravidla jméno, r.č, datum, značka CA) zašifrovaný soukromým klíčem přiděleným této osobě (spolu s komplementárním veřejným klíčem) nějakou certifikační autoritou. Ta se může opírat o autoritu státu, nebo stát může tuto autoritu delegovat na nějaký podnik nebo instituci pro její interní potřebu a konec konců, když to té instituci stačí, může si pro svou potřebu soukromý a veřejný klíč vytvářet sama.

5.2 Hashovací funkce

Je to substitučně transpoziční kryptovací metoda, kterou lze aplikovat na nějaký text nebo celý dokument tak, že výsledkem je asi desetiznakový alfanumerický řetězec pevné délky (bez ohledu na délku vstupního textu) zvaný hash, což je jakýsi „otisk prstu“ ED. Její základní vlastností je to, že nepatrná změna zdrojového textu způsobí dramatickou změnu hashe. Hashovacích funkcí je celá řada, jsou veřejně přístupné. Po prolomení MD5 a objevu chyb v SHA-1 je asi nejlepší volba SHA-2, používaná americkou vládou. Prolomením hashovací funkce je nalezení algoritmu, který umožňuje vytvořit k danému ED dokument jiný, se stejným hashem. Prolomení je i taková změna ED, která nezmění hash. Čas k tomu potřebný se odhaduje na dvacetinásobek času výpočtu soukromého klíče.

Jestliže je dokument opatřen hashem, lze pomocí veřejné hashovací funkce najít jeho hash. Je-li výsledek shodný, mám jistotu, že dokument nebyl pozměněn. Hash se zašifruje soukromým klíčem organizace, autora ED nebo archiváře a to je považováno za e-podpis ED.

5.3 Časové razítko

Předpokládejme, že doba platnosti ED v DEA je 20 let a platnost e-podpisu 1 rok. Pak je možnost fakticky „prodlužovat platnost“ e-podpisu tak zvaným časovým razítkem. To poskytuje tak zvaná TSA (time stamping autorita, u nás opět I.CA), která e-podpis (nebo předchozí časové razítko) doplní o aktuální datum a čas, který garantuje. Tento řetězec pak zašifruje svým privátním klíčem.

Spokojíme-li se s důvěryhodností vlastní firmy, pak lze časové razítko získat pomocí HW modulu umístěného na serveru [13] organizace. Spokojíme-li se navíc s přesností systémového času, pak jej lze nahradit i SW modulem.

Časové razítko nás zbavilo závislosti na autorovi ED, ale nezbavilo povinnosti neustálého obnovování, a to v době platnosti starého razítka. Jde pořád jen o šifry RSA.

5.4 Linkovaný hash [12]

Použitím časového razítka lze platnost e-podpisu prodlužovat donekonečna. Prakticky je však neúnosné časovým razítkem opatřovat každoročně miliony dokumentů po celou dobu jejich uložení, když za tu dobu je vyžádán jen zlomek z nich.

Dosud jsme vycházeli z předpokladu, že ED jsou volně přístupné široké veřejnosti. To však není pravda. V DEA je přístup kontrolován přístupovými právy RDBMS a dokonce je prováděn jeho audit. Pak je možno využít souvislosti časové posloupnosti, se kterou jsou ED vkládány do DEA. Doba vložení ED_n musí být mezi dobou vložení předcházejícího ED_{n-1} a následujícího ED_{n+1} . Spokojíme-li se s přesností udání času 1 den a do archivu přichází denně 100 ED, postačí dávat časové razítko na každý stý ED.

Časovou posloupnost ED lze snadno získat „ukotvením“ ED k přecházejícímu a následujícímu ED pomocí linkovaného hashe. Ten se získá tak, že k hash ED_n se přidá zleva hash ED_{n-1} a zprava hash ED_{n+1} , a na výsledný řetězec se aplikují hashovací funkce. Nově vniklý „linkovaný hash“, se zašifruje primárním klíčem a запиše do ED_n . Tato procedura se aplikuje na každý ED.

Důvěryhodnost lze zvyšovat počtem časových řad, ke kterým je každý ED zakotven. Lze použít časovou řadu ED jedné kategorie nebo časovou řadu ED od jednoho původce. Další zvýšení bezpečnosti lze získat zkracováním intervalu, se kterým jsou vkládána časová razítka. V mezním případě lze kombinovat souvislou řadu časových razítek spolu s ukotvením ED pomocí 6 kotev ke třem časovým řadám. To již odradí i zkušeného narušitele. Ukotvení ED v časové řadě navíc brání ztrátě ED nebo jeho podvodné náhradě.

5.5 Dlouhodobá čitelnost ED

Čitelnost ED může být ztracena v důsledku technologického vývoje a zastarávání HW platformy, OS, aplikací, změny kódování a zejména souborových formátů.

Čitelnost lze obnovovat v podstatě dvěma způsoby migrací a emulací. Každou z nich lze virtualizovat. Ostatní techniky nenašly praktické použití. Za zmínku snad stojí jen zápis v obrazové podobě na nehořlavé filmy.

5.6 Migrace

Migrace je převod ED z jedné HW konfigurace nebo SW aplikace na jinou konfiguraci či aplikaci neboli transformace ze zastaralého formátu na formát aktuální. Je to patrně nejspolehlivější metoda, ale má mnoho úskalí ve ztrátě „chuti a vůně dokumentů“ a přináší vážné problémy se zachováním autenticity. Z hlediska zajištění dlouhodobé čitelnosti ED je výhodné při jejich ukládání do DEA provést jejich migraci do časově stálých formátů

(PDF/A, XML, TIFF, WAV, MPEG2, plain text s ASCII nebo unikódem). Při migraci je důležité rozhodnout, zda se migrace provede při uložení ED do archivu nebo při požadavku na jeho zobrazení nebo zkopírování.

Každá z metod má své výhody. Přikláním se k tlaku na původce, aby při archivaci transformoval ED do jednoho z několika málo formátů vytipovaných pro každou oblast. To by mělo postačit pro krátkou (5-10 let) a v mnoha případech i střední (10-30 let) dobu uložení. K migraci pak přistupovat až v okamžiku požadavku na čtení ED. Tento postup má celou řadu výhod. Omezí se počet formátů, které je nutno migrovat. Při změně formátu není třeba migrovat všechny ED, ale jen ty, které již byly vyžádány a to jen jednou.

5.7 Formát PDF/A [1]

Zcela výjimečnou pozici pro archivaci má formát PDF a zejména jeho archivační modifikace PDF/A. Svědčí o tom i informace z Německa, kde bylo přerušeno skenování do formátu TIFF a pokračuje se do formátu PDF [16]. Užívá se již 15 let po celém světě vládami i největšími firmami. Ačkoliv jej vytvořila firma Adobe, je to otevřený formát splňující normu ISO. Lze jej zobrazovat i tisknout na platformách MS Windows, Macintosh, Unix a i na řadě mobilních platforem. Používá ho 1800 dodavatelů SW produktů na celém světě. Na WEBu je umístěno více než 200 mil. dokumentů v PDF. Při konverzi z libovolného formátu zachovává text, obrázky, barvy, 3D grafiku, foto a dokonce business logiku bez ohledu na aplikaci, která ji vytvořila. Zachovává e-podpis. Podporuje fulltextové vyhledávání, dokumentové značky i datová pole.

5.8 Emulace

Emulace je simulace původního HW prostředí a OS na aktuální platformě. ED je možno zobrazit v originálním SW. Je nutno uchovávat i instalační soubory původní aplikace. Rizikem je ztráta schopnosti emulovat některou vlastnost původní platformy. Výhodou je možnost uchovávat i vektorovou grafiku a multimediální ED. Emulace je obecně náročnější a riskantnější, ale dává lepší výsledky. Má vyšší počáteční náklady, ale není třeba vždy manipulovat s každým ED. Upravuje se jen s prostředím. Výhodnost emulace oproti migraci vzrůstá s počtem dokumentů a se vzrůstajícími požadavky na přesnost zobrazení ED.

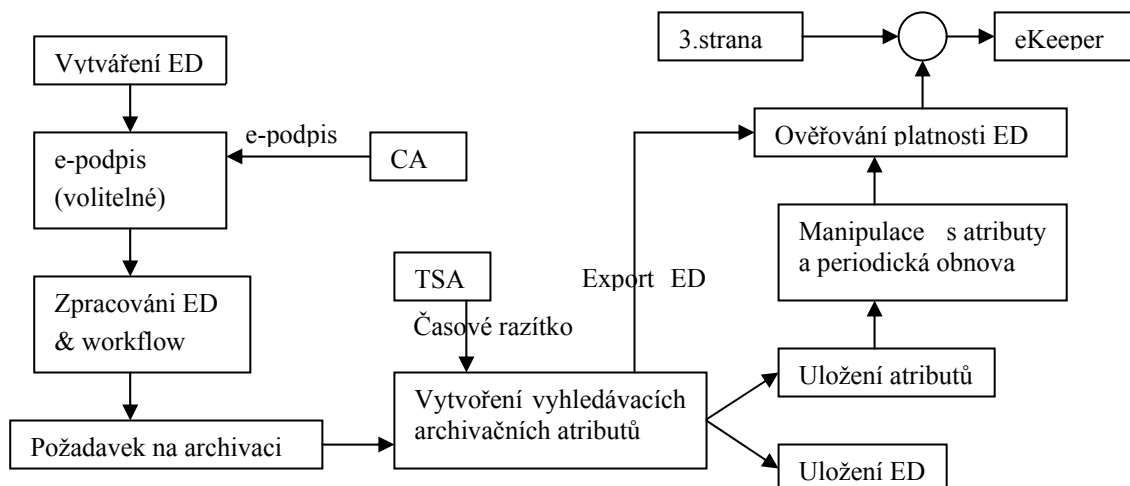
5.9 Virtualizace

Principem je uchování ED v původním formátu spolu s programem, který umožní jeho interpretaci ve strojovém jazyku „Univerzálního virtuálního počítače“. Jazyk UVC musí být jednoduchý otevřený a všeobecný. Pak postačí, aby jakákoliv nová platforma obsahovala interpret jazyka UVC. Dále bude popsána vedle této virtualizace „migrační“ i virtualizaci „emulační“. Rozhodnutí, která z metod je lepší, není jednoduché. O tom svědčí rozdílné cesty, kterými se vydaly Národní archiv UK a Národní knihovna Holandska, viz dále.

Věnujme pozornost třem naprosto rozdílným archivům, které byly realizovány.

6 Popis realizovaných řešení

6.1 e-Keeper (e-Archivář) lublaňské firmy SETCCE [2, 3]



Z obrázku je patrné využití časového razítka. Výrobce uvádí přednosti svého komerčně dodávaného produktu : Shoda s průmyslovými standardy, škálovatelnost, spolehlivost a bezpečnost, schopnost zpracovávat ED libovolného formátu s i bez e-podpisu, periodická obnova kryptovacích metod, nízká cena, nezatěžuje koncového uživatele, snadná implementace na libovolný systém založený na JAVA nebo C++, nezávislý na platformě : souborový systém, DMS, ERP.

Toto řešení uvádím jako důkaz toho, že čas již skutečně dozrál pro pružné přizpůsobení se libovolným požadavkům zákazníka. Systém zřejmě není určen pro dlouhodobé uložení ED, nemá totiž zpracovanou dlouhodobou čitelnost ED.

6.2 Národní archiv UK

Archivace 176 km regálů dokumentů stála jen v r. 2002 14.3 mil. liber a stala se neúnosnou. Proto bylo rozhodnuto od r.2004 přejít zcela na ED bez krytí papírovými originály. Obrovským problémem je požadavek na dlouhodobé uchování obrovského množství formátů. Zkoumání mnoha metod se zredukovalo na emulaci a migraci. Výhodou emulace bylo zachování „chuti a vůně“ ED, ale SW pro emulaci je složitější a jeho vývoj dražší a delší. Migrace je technicky jednodušší, ale dochází k částečné ztrátě informace. Na rozdíl od emulace je nutno při každém zastarání formátu konvertovat všechny ED. Rozhodování bylo velmi těžké (vývoj však pokračuje oběma směry). Všechny ED budou uchovávány ve svých původních formátech, ale pro jejich publikování bude využito migrace.

Pro migraci je nutno uchovávat obrovské množství úplných informací o HW, SW, OS a zejména o formátech, které nejsou vždy otevřené. Firmy se brání jejich zveřejnění. Proto bylo rozhodnuto o založení volně přístupné WEB-ové databáze PRONOM, která bude uchovávat informace o všech souborových formátech. Uchovává se kompletní technické informace o SW potřebném pro zobrazení ED, o SW nutném pro jeho migraci do aktuálního formátu, o firmě, která formát vytvořila. Bylo dokonce nutno vytvořit vlastní systém koncovek názvů souborů. Ukázalo se totiž, že ve světě Windows jsou nejednoznačné a neunikátní.

PRONOM by umožňoval vytvořit pro každý dokument uložený v originálním formátu cestu pro automatizovanou migraci od zdrojového formátu k aktuálnímu cílovému, která by proběhla v okamžiku vyžádání dokumentu. DROID [7] [8] je SW balík, který byl vytvořen v r. 2006, který využívá informace z PRONOMu ke zdokonalené automatizované migraci libovolného zdrojového formátu, do aktuálního formátu.

Během roku se podařilo vyvinout celý digitální archivní systém postavený na platformě Sun Solaris, Oracle, Java, XML a knihovnách s magnetickými páskami. ED i metadata jsou uchovávány na vysokorychlostních páskových jednotkách s robotickou obsluhou. Aktuální kapacita je 2 TB, ale je rozšiřitelná až na 100TB.

Z bezpečnostních důvodů je systém rozdělen na dva naprosto oddělené systémy. Hlavní, který obsahuje všechny dokumenty ve zdrojovém formátu, ale není přístupný veřejnosti. Provozní systém je otevřený a obsahuje dokumenty, které už někdy byly migrovány.

6.3 Národní knihovna Holandska [9]

Použitý archivační systém e-Depot na rozdíl od Britů vsadil na emulaci. Základní handicap emulace – vysoké počáteční náklady na vývoj emulačního SW se podařilo velmi elegantně odstranit použitím virtuální emulační mezivrstvy, která svou filozofií připomíná mezijazyk ve MS Visual Studiu. Je to vlastně emulační virtualizace.

Byl vyvinut SW pro emulaci libovolného formátu na virtuálním platformu[10]. Vývoj této spodní emulační vrstvy je sice náročný, ale provede se jen jednou provždy. V průběhu času pak postačí v souladu s rozvojem technologie upravovat jen horní prezentační vrstvu pro zobrazení neměnné virtuální platformy aktuálním zobrazovacím SW. Podrobným teoretickým výpočtem bylo dokázáno, že tato cesta je levnější než migrace. Zachovává snadněji autenticitu, neztrácí se informace, zejména „chuť a vůně“. Přesto se uznává, že pro skupinu uživatelů, která potřebují se získaným ED intenzivně pracovat v aktuálním prostředí, je migrace nezastupitelná.

Z detailního výpočtu plyne, že celkové náklady jsou po 50 letech u emulace 2.5 x nižší, což představuje úsporu 23 mil. \$. Při zobrazení pro 1 milion ED (což je příznivější pro migraci) je pouze prvních 5 let migrace výhodnější. Po 50 letech je cena emulace nižší 1.8 x, což je stále rozdíl oproti migraci úctyhodných 3.3 mil. \$.

Tyto dvě poslední reference jsou pro mne osobně velmi významné a inspirující neboť ICZ a.s., jejímž jsem zaměstnancem, zvítězila v prestižní veřejné zakázce na „Projekt pracoviště pro dlouhodobé ukládání a zpřístupňování dokumentů v digitální podobě“ pro Národní archiv ČR. ICZ si tohoto vítězství v těžké konkurenci nejvýznamnějších firem v ČR velmi cení a považuje ho za výzvu k vyřešení velmi obtížného a zodpovědného úkolu. Úspěšnost jeho splnění budou posuzovat naši potomci ještě za mnoho set let, což pro SW firmu je zcela mimořádné.

7 Závěr

Doba dozněla pro tvorbu DEA pro široké spektrum uživatelů, od firmy o 3 uživateli po národní archivy. DEA může mít všechny vlastnosti 100% právní legálnosti až po výběr jen některých vlastností, garantovaných jen důvěryhodností vlastní firmy. Pro velké DEA typu Národní archiv nebo knihovna je největším problémem dlouhodobá čitelnost. Zde je potřeba velké počáteční práce na vytvoření SW, ale lze se již opřít o výsledky Národního archivu UK a Národní knihovny Holandska.

Použitá literatura a WWW odkazy

1. <http://www.adobe.com/products/acrobat/adobepdf.html>
2. http://www.setce.si/eng/download/Trusted_Electronic_Archives_-_White_Paper.pdf
3. <http://www.hs-soft.com>
4. http://en.wikipedia.org./Dublin_Core
5. <http://dublincore.org/>
6. Adrian Brown : Preserving the digital heritage: building a digital archive for UK, Government records, Online Information 2003 Proceedings
7. <http://droid.sourceforge.net/wiki/index.php/Introduction>
8. Adrian Brown : Automating Preservation : New Developments in the PRONOM Service, <http://www.rlg.org.en/page.php>
9. Erik Oltmans : A Comparison Between Migration and Emulation in Terms of Costs, <http://www.rlg.org.en/page.php>
10. <http://www.alphaworks.ibm.com./tech/uvc>
11. Adrian Brown : Selecting File Formats for Long-Term Preservation, Digital Preservation, Guidance Note 1, The National Archives 2003, document Reference DPGN-01
12. L. Dostálek, M. Vohnoutová, PVT a. s. Long-term Archive Architecture. Internet Draft draft-ietf-ltans-arch-00.pdf
13. <http://www.deltax.cz/dtx2004>
14. Eun G.Park, University of Kalifornia : Understanding „Authenticity“ in Records Management, <http://www2.sis.pitt.edu/~gaeconf/park.doc>
15. http://www.inter pares.org/book/inter pares_book_k_app02.pdf
16. T. Kalina, M. Kunt, PVT: DOMEA Koncept, Zpráva ze služební cesty Berlín-Koblenz