Vysoká škola ekonomická v Praze
Fakulta informatiky a statistiky
Katedra systémové analýzy

Řešení integrity digitálního archivu pomocí relativní časové autentizace

DOKTORSKÁ DISERTAČNÍ PRÁCE

Doktorand	:	Ing. Rudolf Vohnout
Školitel	:	doc. Ing. Zora Říhová, CSc.
Obor	:	Informatika

Praha, 2014

Prohlášení

Prohlašuji, že doktorskou práci na téma „Řešení integrity digitálního archivu pomocí relativní časové autentizace“ jsem vypracoval samostatně. Použitou literaturu a podkladové materiály uvádím v přiloženém seznamu literatury.

V Praze dne 1. července 2014

…….
Podpis

ABSTRAKT:

Objektem zkoumání disertační práce je fenomén archivace elektronických dokumentů v dlouhodobém časovém horizontu se zaměřením na jejich zabezpečení a integritu. Předmětné dílo se v úvodu zabývá problematikou rešeršního šetření dosud používaných přístupů a řešení v oblasti stanoveného tématu a soustředí se na zkoumání vztahu dlouhodobé archivace na omezující podmínky dané současně planým legislativním rámcem. Hlavní pozornost je soustředěna na integritu archivu jako celku alternativními metodami nezávislými na absolutním čase archivace. Návrhová pasáž se věnuje tvorbě univerzálního systémového řešení vedoucího k zajištění nepopiratelnosti svěřených, archivem spravovaných, objektů. Tento návrh byl poté empiricky ověřen a jeho výsledky zvolenými metodami statistického aparátu zpracovány a vyhodnoceny. Předmětným návrhem, práce usiluje o příspěvek k posílení důkazní ekvivalence dlouhodobě uchovávaných elektronických dokumentů na stejnou úroveň jejich papírových sourozenců.

KLÍČOVÁ SLOVA:

elektronický dokument, dlouhodobá archivace, integrita, bezpečnost, důvěryhodnost

ABSTRACT:

The research objective of the dissertation is a phenomenon of electronic documents archivation in long-term timeframe with focus on their security and integrity. This work at the begging deals with research issues of current approaches and solutions in the specified topic and focuses on relationship examining of the long-term archivation on restrictive conditions set by currently valid legislation scope. The main focus is on the archive integrity by means of alternative methods not related on absolute archival time. Design part deals with creation of the universal system solution leading to ensure non-repudiation of entrusted, archive managed, objects. This draft was empirically verified and results were processed and evaluated by means of statistics methods. The presented proposal of this work endeavors contribution to strengthen legal and evidence equivalence of long-term archived electronic documents on the same level of their paper siblings.

KEYWORDS:

electronic document, long-term archivation, integrity, security, trustworthiness

Obsah

1 Úvod

V odborných publikacích je klasická forma archivace lidského poznání a vědění definována jako „souhrn činností spojených s řádnou péčí o dokumenty“. Obecný pojem archivnictví je chápáno jako jedna z fází životního cyklu dokumentů. Historický vývoj moderního[1] pojetí archivace se datuje přibližně od doby, kdy bylo vynalezeno písmo. Tendence uchovávat a předávat budoucím generacím vědění a dosavadní poznání je stará jak lidstvo samo. Pro písemné archiválie byly již od počátku vybírány kvalitnější materiály (zprvu papyrus, později papír), tak aby se dochovaly delší dobu než běžné písemnosti. S nástupem starověku se začalo jednoduché systémové třídění archiválií a první knihovny. Jak staletí plynula a stav poznání se neustále rozrůstal, a tím spojené množství písemností.

Archivace elektronických dokumentů je od přelomu tisíciletí na vzestupu (v době psaní této práce například tendr na vytvoření národního digitálního archivu ČR [7]) a v odborných kruzích se stává velmi diskutovaným tématem. Jedná se o logické pokračování digitální éry, kdy budoucím generacím již není záhodno předávat vědění a poznání pouze v tradiční, papírové podobě. Záměrem je minimálně zachovat a pokud možno vylepšit vlastnosti, jako jsou stárnutí, doba vyhledávání a další aspekty. Úkolem vědeckých pracovníků a výzkumných týmu v této oblasti je tedy přicházet na způsoby, jak efektivně, bezpečně a dlouhodobě digitálně archivovat. Pokud nebudou nalezeny vhodný způsoby k dlouhodobé úschově těchto informací, hrozí nebezpečí, že některé nám již známé informace se k budoucím generacím ani nedostanou.

Výše uvedené analogicky vede k celosvětovému exponenciálnímu nárůstu množství existujících a digitálních děl, publikací a dokumentů. Na základě prediktivních algoritmů společnosti Google a jejího produktu Google Books bylo v polovině roku 2010 na světě do té doby vydáno 130 milionů unikátních knih [52]. Knihy z tohoto a

[1] Archivace duševního bohatství, nikoliv úschova předmětů.

dalších repositářů jsou postupně digitalizovány a zpřístupňovány online (z více než 70% se jedná o copyright materiály a monografie)[2]. Pro potřeby této práce bude libovolný titul v digitálním světě považován za elektronický dokument. Velmi dobrým ukazatelem je také počet indexovaných webových stránek (internetu). V době psaní této práce bylo nejvíce webových stránek zaregistrováno v polovině roku 2011 (přes 20 miliard) [45].

Představená fakta se však doposud nestala předmětem intenzivnějšího vědeckého zájmu a to především v oblasti zajištění důvěryhodnosti předmětných objektů bez ohledu na časové omezení doby archivace. Například problematika nepopiratelnosti je v dostupné literatuře zmiňována pouze okrajově. Naopak, hlavní důraz je kladen na standardizaci uchovávaných formátů, která samozřejmě s fenoménem dlouhodobé archivace také úzce souvisí.

V první části předkládané práce je zakotven teoretický výzkum, jenž poslouží ke zmapování současného stavu poznání v předmětné oblasti a k zakotvení problematiky do kontextu širší teorie informatické bezpečnosti. Teoretická pasáž identifikuje specifika dlouhodobého uchování elektronických dokumentů a zakotvuje je v oblasti digitální archivace paralelně s archivací papírovou. Smyslem je zcela objasnit význam postupného omezování fyzické podoby archivace a až k jejímu úplnému nahrazení archivací elektronickou. Je třeba jasně vymezit samotný pojem dlouhodobá archivace elektronických dokumentů, formuluje společný teoretický rámec, provede její kategorizaci a jasně vymezí předmětnou podoblast. Z teoretických metod bude v této části práce užito metody analytické a syntetické založené na studiu odborné literatury, konkrétně obsahové analýzy, komparace a případně reinterpretace jak primárních tak sekundárních pramenů.

V části návrhové a experimentální je využito poznatků zjištěných v části analytické a to především k sestavení návrhu řešícího hlavní a dílčí cíl této práce. Metody empirické jsou užity v rámci realizace výzkumného experimentu, který je

[2] http://www.hathitrust.org/

koncipován jako výzkum kvantitativní. Základem empirického šetření je experiment, jehož cílem je ověřit deklarovaný návrh a v rámci omezujících podmínek verifikovat adekvátnost zvolených empirických výzkumných nástrojů. Výsledky vědeckého experimentu jsou zpracovány a zhodnoceny s použitím vhodných nástrojů moderního statistického aparátu.

Část závěrečná pak prezentuje dosažené výsledky a spolu s jejich hodnocením obsahuje ucelený návrh pro zajištění integrity a zabezpečení jak jednotlivých archivovaných dokumentů tak také dopad tohoto návrhu na celou oblast digitální archivace. Představené řešení tak celého archivu a otevírá směr k širšímu používání v relevantních oblastech jako je elektronická fakturace, účetnictví, daňová problematika a zdravotnictví. Práce jako celek chce přispět k tomu, aby dlouhodobě elektronicky archivované dokumenty byly plně uznávány jako důkazní materiál v soudním řízení.

2 Stanovení výzkumného tématu a struktura práce

Tato kapitola konkretizuje výzkumné téma, vymezuje výzkumný problém, stanovuje výzkumné otázky, definuje cíle a úkoly práce a představuje metody jejich řešení a v závěru rozvádí strukturu předkládané práce.

2.1 Zdůvodnění výběru předmětného tématu a motivace

Autor si zvolil toto výzkumného téma mj. na základě svého zahraničního studijně-výzkumného zahraničního pobytu v rámci doktorského studia ve Slovinské Lublani na „University of Ljubljana“. V rámci své vědecko-výzkumné činnosti v „Jože Štefan Institute“pod vedením A. J. Blažiče ze společnosti SETCCE[3], věnující se vývoji důvěryhodných technologií pro e-business a e-government, se autor spolupodílel na testování a optimalizaci a pilotního nasazení produktu eKeeper. Výsledkem této zkušenosti v rámci výměnného programu Erasmus je publikace „*Thin Client Usage in Long-term Archivation Environment*“ [69].

Motivací autora byl také intuitivního předpoklad, že elektronická archivace v dlouhodobém časovém horizontu vykazuje potenciál značné důvěryhodné a informativní hodnoty a rozšíření jejího využití, ať už ve sféře veřejné či soukromé, by mohlo zvýšit efektivitu specifických oblastí informatických věd. Z prostudované literatury citované v kapitole 3 plyne, že k největšímu rozšíření zabezpečení ED[4] došlo v oblastech, kde je jejich integrita zásadní. Ty jsou jednak reprezentovány doménami, jež jsou ovlivněny zákony a opatřeními jako například elektronické fakturace nebo obecněji oblasti jako eCommerce, eHealth či eGovernment. Obecně se tak jedná o uchování bohatství lidského poznání a vědění, účetnictví či obecně v celé oblasti *DMS*[5].

[3] http://www.setcce.si/index.php?lang=eng
[4] Elektronický dokument
[5] Document Management System

Zvolené téma má vazbu na výzkumnou činnost katedry v oblasti systémové analýzy a dále na výzkumnou činnost fakulty informatiky a statistiky, konkrétně katedry informačních technologií v oblasti *ECM*[6].

2.2 Vyjasnění pojmů

Předkládaná disertační práce věnuje jen a pouze archivaci ED, autor se nebude zabývat archivací písemností. Tato, tradiční forma archivace má ovšem některé principy a přístupy, které byly aplikovány také do prostředí archivace elektronické. Některé tyto zákonitosti jsou svým významem podstatné také pro tuto práci a budou dále objasněny. Je také nutné sjednotit terminologii a jasně definovat, nad jakými objekty jsou vyslovovány hypotézy, stanovovány cíle a předkládány výsledky.

Autenticita

Někdy též nazývána jako synonymum ke slovům pravost, původnost či hodnověrnost. A právě chápání těchto slov stejného či podobného významu hraje pro tuto práci významnou roli. Archiv v žádném případě neřeší správnost archivované informace. Tj. pokud byla informace pozměněna či přímo zfalšována, nebo je jakkoliv nepravdivá před přijetím do archivu, bude v tomto stavu nadále udržována. Archiv pouze zaručuje, že po akceptaci informace, bude uchována v podobě, v jaké byla na vstupu přijata. Jinými slovy, pokud se textu této práce vyskytne slovo autenticita, bude vždy vztahována k autenticitě archivovaného objektu od okamžiku přijetí archivem bez ohledu skutečnou pravdivost obsahu sdělení.

Archiv

První co je nutno vyjasnit, je souvislost mezi pojmem archiv a digitální archiv. Archiv je z historického hlediska pro odborníky z oblasti archivnictví slovo exaktního význam se je nepřípustné jej používat mimo obor. Přesný terminologický význam se vztahuje pouze k uchovávání, ochraňování, evidování a zpřístupňování historicky cenných pramenů. Tento přístup tedy stanovuje, že termín „archiv“ jako

[6] Enterprise Content Management

pojmenování „datového skladu“ není přípustný. Díky těmto omezením je možné se setkat s pojmy digitální či datové úložiště nebo moderním výrazem datové sklady.

Objekt

Základem pro jakoukoliv archivaci, jsou její objekty. V papírovém světě jsou nazývány souhrnným pojmem archiválie, ve světě informačních technologií pak ED, respektive ED obohacené o další informace. Na ty je tak třeba akcentovat zcela odlišné požadavky než na jejich papírovou podobu. Je nutné zajistit po celou dobu archivace integritu, čitelnost a především nezpochybnitelnost a důvěryhodným způsobem archivovat spolu se samotným dokumentem také doplňkové informace (metadata).

Aby došlo ke sjednocení názvosloví pojem „balík“ (package, resp. ED s metadaty, případně dalšími, pro LTA podstatnými informacemi) jakéhokoliv významu bude nahrazen pojmem „objekt“. Objekty budou mít jednoznačně definované indexy a části, ze kterých bude tvořen. Tento pojem je v terminologii této práce od začátku běžně používán.

Archivace

V této fázi je též nutné vyjasnit rozdíl mezi často zaměňovanými pojmy „archivace“ a (operativní) „záloha“. LTA (dlouhodobé uchovávání digitálních dokumentů a informací[7]) dle [66] lze definovat jako „užití digitální technologie pro (dlouhodobé) uchování informačního obsahu“. Tento proces nezahrnuje pouze činnosti technické povahy (ochrana paměťových médií, ochrana dat), ale má i aspekty organizační (zejména výběr dokumentů pro LTA), finanční a právní (týkající se práv autorů a nakladatelů)“. Naproti tomu operativní zálohy slouží především ke krátkodobé duplikaci dat. Operativní zálohou se tedy dle odborných publikací rozumí taková záloha „kdy jsou zálohovaná data vždy rychle a v co nejkratším možném čase přístupná“. Z dlouhodobě archivovaných objektů nelze udělat operativní obnovu dat jako ze zálohy.

[7] http://digital-preservation-cz.blogspot.cz/

Některé zdroje uvádějí [2], že LTA je posloupnost aktivit vedoucích k zajištění, že digitální dokument může být lokalizován (vyhledán), zobrazen, použit a pochopen v budoucnosti. Z toho lze vyvodit dvě publikované a jednu autorovu definici LTA:

- Elektronická archivace kombinuje politiky, strategie a činnosti zajišťující přístup digitálnímu obsahu bez ohledu na selhání medií[8] a technologické změny. [75]
- Zákon zachování informace ve správné a nezávisle srozumitelné podobě, v dlouhodobém časovém horizontu. [58]
- LTA je schopnost udržovat srozumitelnost a použitelnost svěřených dokumentů ve vzdálené budoucnosti bez ohledu na technologické změny pro vyhrazenou skupinu uživatelů.

2.2.1 Základní použitá terminologie

V této části práce jsou podrobněji rozvinuty a definovány základní termíny, které jsou pro tuto práci podstatné a jejichž výskyt v práci je hojný. Použité zkratky jsou definovány v příslušné kapitole na konci této práce.

Archivní objekt

Dlouhodobě uložený, archivem spravovaný ED nikdy není udržován samostatně. Jeho součástí jsou vždy metadata (nepřepisovatelná povinně), případně další informace (LOGy apod.). Na všechny tyto přidružené informace a samotný dokument je pak aplikován prvek zabezpečení daného archivu (např. relativní časové razítko). Jinými slovy, archivní objekt, lze chápat jako jakýsi digitální informační kontejner (bitstream), který je ve finále bezpečně zapečetěn a ochráněn proti neoprávněné modifikaci či smazání. Nejvhodnější příměr z oblasti přírodních věd je komár zalitý v jantaru. V práci se dále vyskytují synonyma, jako jsou digitální či archivační objekt.

[8] Velmi ilustrativní pojednání o vhodnosti použitých archivačních medií s ohledem na délku a typ archivace, včetně jejich vzájemného srovnání, vymezení výhod a nevýhod a cenové kalkulace: http://www.digitisingcontemporaryart.eu/images/uploads/banners/DCA_D62_Best_practices_for_a_digital_storage_infrastructure_20130506_Version1.pdf

Integrita

Zásadní pojem ve spojitosti s archivem. Předmětný termín vymezený pro tuto práci značí, že daný, archivem spravovaný, archivní objekt od doby přijeti do systému (zapečetění kontejneru) nebyl změněn, ani nijak modifikován, ani smazán. Tento pojem ovšem neřeší autenticitu (pravdivost, hodnověrnost) obsažené informace.

Nepopiratelnost

S pojmem integrita úzce souvisí pojem nepopiratelnost. Pokud je objekt, či dokument nepopiratelný, znamená to, že nelze žádným způsobem zpochybnit jeho platnost. Velmi často se vyskytuje v právním prostředí, kde je pravost často napadána. Nepopiratelnost se často zaměňuje s pojmem autenticita, neboli originalita, původní stav dané entity, její pravost. Oba pojmy však znamenají něco jiného.

Zabezpečovací prvek

V kontextu práce se jedná o takový element, který zajišťuje, resp. stará se o bezpečnost a zajišťuje nepopiratelnost svěřených archivních objektů. Jinými slovy, na tomto prvku je postaveno zachování integrity celého archivu. Někdy je o něm v práci též pojednáváno jako o bezpečnostním prvku.

Důvěryhodná archivní autorita

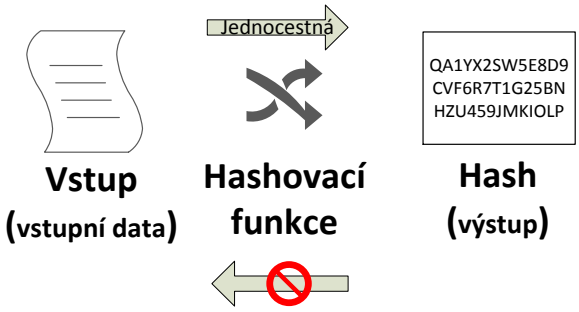
Autorita, které věříme, že se plně postará o svěřené dokumenty. Je reprezentována důvěryhodným archivem, který implikuje všechna dostupná zabezpečení jím spravovaných ED, včetně napojení na autoritu časových razítek, či certifikačních autority. V práci je pro tento pojem používána zkratka *TAA* („Trustworthy Archive Authority").

Metadata

„Data o datech". Představují zásadní element pro rychlé vyhledávání, indexaci atd. Důvěryhodný archiv je zásadním způsobem závislý na jejich správnosti. V kontextu *LTA* a metadat lze konstatovat, že integrita vybraných metadat je v určitých případech zásadnější než integrita samotného archivovaného dokumentu.

Hash

(hash value; message digest, digital fingerprint; česky též nesprávně označován jako otisk) Je představována proměnnou *h*, jehož pomocí hashovací funkce převádí vstupní posloupnost bitů na exaktní bitovou strukturu o konstantní délce *n* bitů. Funkce je jednocestná (z výstupu nelze získat vstup) a jakákoliv změna na vstupu způsobí zásadní změnu na výstupu. Každá hashovací funkce je ze své podstaty konstruována tak, aby malá změna na vstupu (vstupní bitová posloupnost) vyvolala velkou změnu na výstupu. Pokud dva různé dokumenty na vstupu dosáhnou stejné hodnoty na výstupu, dochází k tzv. kolizi a použitá hashovací funkce již není považována za bezpečnou.



Obrázek 1: Princip hashovací funkce Zdroj: [autor]

2.3 Definování předmětné podoblasti a vymezení výzkumného pole

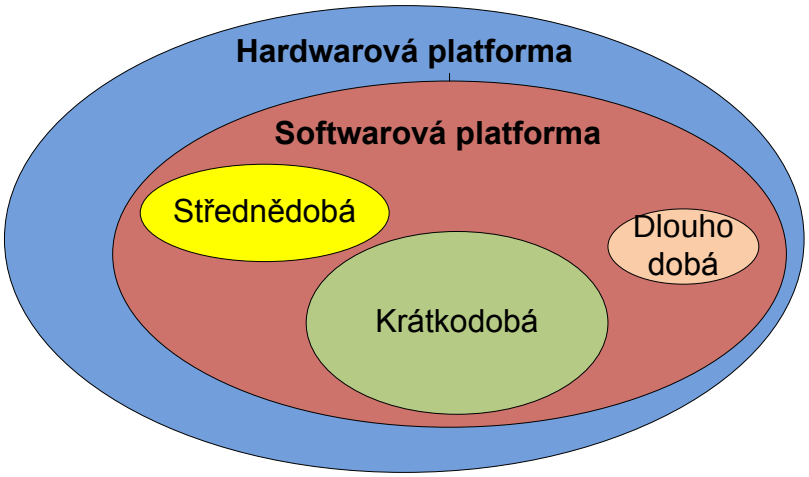
Úkolem této kapitoly je stanovení výzkumného rámce předmětné práce, tak aby na jeho základě bylo možné definovat výzkumný problém. To zahrnuje vymezení předmětné oblasti jejím zúžením a konkretizací. Zkoumaná oblast digitální archivace je doposud nejasně vymezena a její uchopení z hlediska informatických věd se vyznačuje značnou koncepční i terminologickou nekonzistencí. Soustředí se primárně na systémovou problematiku, jejíž řešení mohou být poté aplikovatelná v konkrétních projektech. Na základě prostudované literatury je opodstatněné předpokládat, že skeptický přístup k problematice průkazní hodnoty ED ještě nějaký čas potrvá, než dosáhne úrovně papírové. Tato skutečnost je o to zásadnější ve chvíli kdy by jako důkazní materiál proti sobě stál jak dokument ve formě elektronické tak papírové, avšak oba by měly protichůdnou důkazní hodnotu.

LTA[9] jako takovou je třeba si striktně rozdělit na dvě, spolu sice přímo související, ale z hlediska této disertační práce oddělené části, a to:

1) Hardwarovou část.
2) *Softwarovou část.*

Hardwarová část jako jsou parametry jednotlivých komponent, jejich životnost a čitelnost přenosových (úložných) medií[10], není předmětem této práce. Z hlediska dlouhodobé archivace a vzhledem k době životnosti HW[11] je nutné na HW nahlížet jako na spotřební materiál a mít zcela vyřešenu HW migraci.

Nicméně je nutné podotknout, že softwarová část je přímo závislá na funkčnosti (a tudíž i spolehlivosti) části hardwarové a je tak její podmnožinou (Obrázek 2).



Obrázek 2: Schéma rozdělení archivace **Zdroj:[autor]**

Další hlediska, na základě kterých se ED dělí, jsou:

- Z hlediska účelu.
- Z hlediska hierarchické struktury.
- *Z hlediska času.*

[9] Long Term Archivation = Dlouhodobá archivace

[10] Zde je vhodné poznamenat, že do této chvíle (červenec 2013) není shoda nad optimálním (dokonalým) mediem vhodným pro dlouhodobé uschování digitálních dat. Velmi nadějným se jeví dle posledních výzkumů DNA (viz příslušný článek v časopisu Nature: http://www.nature.com/nature/journal/vaop/ncurrent/full/nature11875.html) nebo speciální nano-strukturované sklo (viz http://www.orc.soton.ac.uk/5dopticalstore.html).

[11] Hardware

Archivaci z hlediska času je nutné chápat vzhledem k diametrálně odlišným přístupům, standardům a technologiím jako na sobě nezávislé entity, které však mají softwarovou platformu jako společnou třídu a dědí její vlastnosti. Velikosti elips reprezentující dobu archivace představují aproximačně objem dokumentů.

Pro tuto práci hlavní je časové hledisko. To se dále dělí dle doby archivace (archivační lhůty) na:

1) Krátkodobou [Short-Term] (v řádu měsíců) – operativní záloha, viz použitá terminologie.
2) Střednědobou [Mid-Term] (v řádu let – například [68] uvádí, že již od pěti let).
3) *Dlouhodobou [Long-Term] (desítky až stovky let).*

Některé prameny [44] uvádějí specifičtější podobu rozdělení archivace na výše uvedené skupiny na základě konkrétních časových údajů doby archivace. Za krátkodobou je považována archivace do tří let, za střednědobou pokud jsou data uschována v ohraničeném časovém intervalu mezi třemi a deseti lety a za dlouhodobou pokud požadovaná doba uložení přesahuje deset let (bez stanovení horní meze). Názory na to, co již je a co ještě není, archivace s přívlastkem dlouhodobá se napříč odbornou komunitou různí. Dlouhodobost se ve vztahu k digitální archivaci dá také vyjádřit bez časového vyjádření:

- Dostatečně dlouhý časový úsek zohledňující dopady měnících se technologií, zahrnující podporu pro nová média, datové formáty a měnící se uživatelskou komunitu uchovávající informaci v repositáři [58].

Jako konkrétní paradigma lze uvést potřebu dlouhodobě archivovat listy důchodového zabezpečení (30 let), kde je klíčové zajistit nemožnost vytvoření duplicitního dokumentu. Níže uvedená čísla vychází z bezpečnosti kryptografických algoritmů dle doporučení ETSI [28]. Následující tabulka rozdělení z hlediska času vyjadřuje:

Tabulka 1: Doba archivace dle hlediska archivační lhůt

Krátkodobá	Střednědobá	Dlouhodobá
$\leq 3 roky$	$< 3\ roky; 10\ let >$	$\geq 10\ let$
Elektronický podpis; Kvalifikované časové razítko	Dlouhodobý elektronický podpis; Archivní kvalifikované časové razítko	Důvěryhodný archiv

Zdroj:[44]

Ve vztahu uživatele, TAA a archivovaných dokumentů je zásadní otázkou nepopiratelnost. To vyplývá z [75], kde se poslání uchovávat data v elektronické podobě dělí na tři úrovně, odstupňované dle výše ambicióznosti (a tudíž také obtížnosti uchovávání) a z toho plynoucí míry bezpečnosti:

1) Nejnižší úroveň.

   Představuje uchování bitový řetězců, bez ohledu na srozumitelnost, čitelnost nebo užitečnost (důležitost) archivovaných dat.

2) Střední úroveň.

   Znamená, že se neuchovají pouze logické stavy jednotlivých bitů ale také jejich význam (sémantika). Tento přístup ovšem vyžaduje uchování dodatečných informací, jakými jsou například metadata.

3) Nejvyšší úroveň.

   Přidává ke střední úrovni takovou metodu, která zajistí provenienci a zároveň nepopiratelnost svěřených dokumentů. Tudíž uživatelé mají důvěru, že předmětný objekt je původní, nezměněný a úplný.

Archiv jako takový má jako své hlavní poslání zajištění nejvyšší úrovně poslání úschovy, tj. integritu archivovaných ED. V rámci konkretizace výzkumného pole je dále třeba jasně specifikovat, kde se bude zajišťovat ochrana proti neoprávněné modifikaci. Proto se tyto mechanismy dále dělí na dvě, na sobě nezávislé skupiny:

1) Mechanismus zajišťující integritu na straně původce[12].

2) Mechanismus zajišťující integritu na straně archivu.

[12] Původní tvůrce dokumentu

V prvním případě jsou to dlouhodobé zaručené elektronické podpisy pro několik podporovaných typů dokumentů. Jedná se o známé podpisy PAdES LTV[13] – „Long Term Validation“ pro dokumenty ve formátu PDF, XAdES-X-L[14] pro dokumenty v XML formátu nebo CAdES-X-L[15] pro všechny ostatní formáty. Úkolem LTA je zajistit bezpečnost svěřených informací. To mělo být možné bez ohledu na formát, v jakém byla informace (dokument) archivu svěřena [34]. Navíc, výše uvedené elektronické podpisy formátů vhodných pro dlouhodobou archivaci (dále jen „*LTA*“), vstoupily k 1. 7. 2012 (díky novelám již citovaných zákonů 227/2004Sb. a 499/2004Sb.) v platnost také v českém právním prostředí. Česká legislativa se tak jednoznačně přihlásila k normám vydaným Evropským institutem pro telekomunikační standardy (ETSI) [55].

Konkretizací předmětné problematiky bylo stanoveno výzkumné pole, či výzkumný rámec práce.

2.3.1 Omezující předpoklady pro stanovení vědeckého úkolu

Níže uvedené předpoklady vychází z obecných principů vztahujících se jak k důvěryhodnosti dlouhodobého archivu, tak k zákonitostem z oblasti informatických věd.

Předpoklad I

V žádném časovém okamžiku kdy se přechází od původního, dále již ne bezpečného prvku k novému, v daném časovém okamžiku dostatečně bezpečnému, nesmí být možné zpochybnit integritu archivu.

Odůvodnění předpokladu I a jeho omezující podmínky

Předpoklad se zakládá na presumpci důvěryhodnosti archivu jako celku. Proto v žádný časový okamžik ji nesmí být možné zpochybnit, stejně tak jako nepopiratelnost spravovaných archivních objektů.

[13] http://www.etsi.org/deliver/etsi_ts/102700_102799/10277801/01.01.01_60/ts_10277801v010101p.pdf
[14] http://www.etsi.org/deliver/etsi_ts/101900_101999/101903/01.04.02_60/ts_101903v010402p.pdf
[15] http://tools.ietf.org/html/rfc5126

Předpoklad II

Lineární nárůst dostupného výpočetního výkonu oscilující kolem průběhu stanoveného na základě tezí Mooreova zákona[16] a formulace Pollackova pravidla[17].

Odůvodnění předpokladu I a jeho omezující podmínky

Tento předpoklad vychází z IT zákonitostí. Skutečný výkon je však obvykle v praxi nižší než prostým součtem těchto pravidel [14]. Další faktor, který je nutné zohlednit (a které nezohledňuje Moore and Pollack), jsou paralelní výpočty[18].

Dalším obecným předpokladem je, že počet ED v archivu bude neustále narůstat a objekty, nebudou také nikdy fyzicky odstraněny (vymazány), aby byla vždy možná jejich zpětná verifikace. Jak bude skartace skutečně řešena, jak detailně popsáno k kapitole 4.1.2).

2.4 Výzkumný problém, výzkumné otázky a cíle řešeného vědeckého úkolu

Ze stanoveného výzkumného pole lze derivovat výzkumný problém, jenž bude základem pro vyvození výzkumných otázek. Ty pak povedou ke stanovení konkrétních cílům a úkolům řešené vědecké práce.

2.4.1 Výzkumný problém

Předmětem zkoumání této disertační práce jsou vybrané bezpečnostní aspekty LTA. Všechny ED a data, jež jsou trvale uloženy v systému takovéhoto archivu, jsou vystaveny entropii stárnutí prvků, které zajišťují jejich nepopiratelnost. To směřuje opětovně k integritě původní uložené informace. V kontextu uvedených myšlenek a dedukcí lze hlavní výzkumný problém práce zformulovat do následující teze:

- Hlavním výzkumný problémem práce je prezentovat důvěryhodný způsob řešení systému zabezpečení LTA, který by přispěl k rozvoji v této a jí příbuzných oblastech.

[16] http://www.intel.com/content/www/us/en/silicon-innovations/moores-law-technology.html
[17] http://www.itrs.net/Links/2001ITRS/Links/design/pollack.html
[18] http://techland.time.com/2012/05/01/the-collapse-of-moores-law-physicist-says-its-already-happening/

Jinými slovy, hlavní problém vychází v rámci určeného výzkumného pole z otázky, zdali je možné jak závisle tak nezávisle na absolutním čase archivace zabezpečit integritu uchovaných archivních objektů a archivu jako celku to za předpokladu, že byly svěřeny ke správě TAA na období určené politikou archivu jako dlouhodobé.

2.4.2 Výzkumné otázky

Na základě takto stanoveného výzkumného problému, vyvstává řada dílčích konkretizací, jež lze formulovat do podoby výzkumných otázek. Na ty se tato práce pokouší přinést odpovědi:

1. Je dlouhodobě archivovaný ED právní alternativou k dokumentům v papírové podobě? Jinými slovy má takovýto dokument s ohledem na současně platnou legislativu dostatečnou oporu v zákonech nejen v jednotlivých členských státech EU, ale také mimo Evropu aby mohl být plnohodnotnou náhradou jeho tištěné formy?
2. Jaké jsou vlastnosti v současné době používaných řešení LTA z hlediska změn použitých technik zabezpečujících nepopiratelnost svěřených objektů?
3. Lze vzhledem k vývoji v oboru informačních technologií, výpočetní kapacity a kryptografických metod považovat digitální objekt v jakémkoliv časovém okamžiku po celou dobu archivace chráněný proti porušení integrity a smazání?
4. Jaké jsou integritní podmínky pro validitu elektronicky archivovaných dokumentů v časově velmi dlouhém horizontu?
5. Lze bezpečně a dlouhodobě uchovávat digitální objekty způsobem, který by nebyl závislý na stárnutí použitých bezpečnostních mechanismů na straně archivu?
6. Je možné nějakým řešením digitální archivace předejít okamžiku zpochybnění nepopiratelnosti archivních objektů?
7. Lze nelézt alternativní způsob zajištění integrity archivovaných objektů v případě, že hlavní bezpečnostní způsob bude v budoucnu zpochybněn?

2.5 Cíle a úkoly práce

Na základě výzkumných otázek formulovaných v předchozí podkapitole je formulován hlavní cíl předkládané disertační práce, z něhož jsou následně derivovány cíle dílčí.

2.5.1 Hlavní cíl

Hlavní cíl má za úkol rámcově stanovit předmět řešení této disertační práce. Jeho konkretizace pak následující v podobě cílů dílčích.

> **Navržení systému zabezpečení, garantujícího nepopiratelnost dlouhodobě uchovaného archivního objektu a zajišťujícího integritu archivu jako celku.**

Konkretizace problematiky hlavního cíle

Problematika hlavního cíle práce spočívá v dlouhodobé udržitelnosti systému bezpečného a důvěryhodného archivu. Navržený způsob zabezpečení, by měl být kombinací relativního a absolutního času a době archivace, po kterou je digitální objekt archivován. Návrh řešení také respektuje tzv. univerzálnost. Ta je vyjádřena svoji připraveností pro možnost implementace tohoto návrhu do široké škály produktů pokrývající problematiku LTA.

2.5.2 Dílčí cíle, jejich popis a definování významu pro disertační práci

1. Stanovit obecný rámec pro tvorbu systému zabezpečení dlouhodobě archivovaného digitálních objektu.
2. Navržení metody zajišťující nepopiratelnost původního digitálního objektu i po případné změně zabezpečovacího prvku v čase na straně archivu.
3. Vytvoření původního identifikátoru dlouhodobě archivovaného ED, který bude plně nezávislý na hlavním zabezpečovacím prvku.
4. Empiricky ověřit a statisticky zpracovat návrh z hlavního cíle.
5. Doporučit či nedoporučit nasazení konkrétních bezpečnostních prvků.

2.5.3 Úkoly práce

Vědecké úkoly, které si tato práce vytýčila a jejichž splnění je nutné k dosažení stanovených cílů se dají shrnout do následujících bodů:

1. Vymezit podoblast LTA včetně její právní legitimity.
2. Analyzovat stávající přístupy k řešení problematiky LTA ED.
3. Specifikovat požadavky na funkční vlastnosti archivu pro zajištění integrity spravovaných objektů a archivu jako celku a identifikovat možnosti zabezpečení archivovaných elektronických dokumentů v dlouhodobém časovém horizontu.
4. Realizovat vědecký experiment ověřující navržený způsob řešení.

Konkretizace problematiky navržených dílčích cílů a stanovených úkolů práce

V rámci provedení rešeršní analýzy si první stanovený dílčí cíl klade za úkol jasně vymezit podmínky pro následné odborné návrhy řešení.

Druhý dílčí cíl předmětného vědeckého úkolu si klade za splnění požadavek na výměnu hlavní kryptografické funkce, na níž je postaven bezpečnostní prvek archivu, v již existující bezpečnostní struktuře. Tato struktura (bez ohledu na její skutečnou podobu) musí zaručit jak integritu posledně archivovaného objektu tak všech předchozích. Jinými slovy druhý dílčí cíl řeší bezpečný a garantovaný přechod z jednoho prvku, jehož bezpečnost bude nedostatečná, na prvek jiný.

Třetí dílčí cíl stanoveného vědeckého úkolu vychází z případné nemožnosti, komplikace či jiné skutečnosti bránící implementaci návrhu druhého řešeného dílčího úkolu. Do jisté míry se tak dá považovat za komplementární řešení při nutnosti aplikace pouze jednoho bezpečnostního prvku a jeho primárním úkolem je zajistit alternativní způsob zachování integrity svěřených archivních objektů.

Poslední dílčí cíl je komplementární k předchozímu, již stanovenému cíli. Bude empiricky ověřovat efektivitu návrhu plynoucího z cíle druhého. Hodnoty vzešlé z empirického ověření spolu s jejich exaktním popisem a provedenou statistickou analýzou lze nalézt v kapitole 7.

2.6 Hypotézy řešeného vědeckého úkolu

Pří testování hypotéz bude využito statistických metod z oblasti regresní analýzy.

2.6.1 Hypotéza I

Zvyšování počtu dokumentu v archivu nemá vliv na rychlost výpočtu u hlavního bezpečnostního prvku na straně archivu.

2.6.1.1 Odůvodnění hypotézy I a její omezující podmínky

Tato hypotéza se zakládá na racionální domněnce vyplývající z obecných principů informatiky. Jedná se o vztah mezi definovanými, závislými a nezávislými proměnnými počet dokumentů v archivu [ks] a výpočetní náročnost [ms].

Výpočet pro každý nový dokument v archivu musí probíhat bez ovlivnění následujícími hodnotami, získanými v čase *n-m*:

- Počet již existujících uzlů existujícího stromu.
- Počet dokumentů v archivu spravovaných archivní entitou.
- Jednoznačné identifikaci použitého způsobu zabezpečení.

Hypotéza bude potvrzována (vyvrácena) na hladině významnosti $p < 0,05$.

2.6.2 Hypotéza II

Nová funkce, na níž bude založen archivní bezpečnostní prvek, bude vždy výpočetně náročnější než předchozí (již nedostatečně bezpečná).

Potvrzení nebo vyvrácení hypotézy bude možné po experimentálním ověření výpočetní náročnosti první hypotézy. Tato hypotéza vychází z dokumentů uvedených v literatuře (například [64]). Představuje vztah mezi definovanými proměnnými:

$h_{id}(Doc_n)$= jednoznačný identifikátor použité bezpečnostní funkce.

t_h = čas potřebný k výpočtu kontrolního součtu dané bezpečnostní funkce.

Doc_n = Objekt nad kterým se bude testovat výpočetní náročnost.

Hladina významnosti bude opět $p < 0,05$.

2.7 Zvolené metody zkoumání k dosažení stanovaných cílů a úkolů práce

Pro hledání odpovědí na výzkumné otázky plynoucí ze stanovených cílů a plnění úkolů předkládané práce, bude užito jako metod teoretických tak empirických. Autor v této části zmiňuje, jak jsou jím zvolené výzkumné metody významné pro naplnění stanovených cílů disertační práce.

V první fázi bude provedena terminologická analýza pojmu dlouhodobá elektronická archivace a její bezpečnostní specifika. Dále bude využito studia primárních a sekundárních pramenů k interpretaci legislativního prostředí, které vymezuje aplikační oblast. Následně bude užito metod klasifikační a vztahové analýzy v oblasti zajištění integrity a zabezpečení ED během jeho životního cyklu. To poté bude podkladem pro návrh syntetického modelu řešícího změnu struktury zajišťujících bezpečnost svěřených archivních objektů.

Takto provedený teoretický výzkum umožní navrhnout a realizovat výzkumný projekt. Ten bude založen na systémovém návrhu a jeho následné experimentální verifikaci. Po získání primární kolekce dat (sekundární kolekce dat bude použita pouze v případě, že nebude možné získat vlastní primární kolekci) budou tato data vyhodnocena statickými nástroji a to z oblasti vícerozměrné domény [33], [51]. Dosažené výsledky a závěry poté budou interpretovány s ohledem na praktické a teoretické výzkumné otázky této práce.

K řešení dílčích cílů se vztahuje inovativní návrhová analýza systémového řešení dlouhodobé bezpečnosti objektů svěřených archivní entitě. Dále pak syntéza vedoucí k tvorbě nezávislého způsobu identifikace objektu, jehož bezpečnost nebude podléhat slábnoucím kryptografických mechanismům.

Provedení experimentu ověřujícího validitu výsledků návrhové analýzy. Budou užity metody vícerozměrné analýzy rozptylu [60] pro zjištění míry závislosti mezi proměnnými. V rámci testování hypotéz a při respektování výchozích předpokladů bude také měřena rychlost výpočtu samotných hashů a její vliv na druhy použité

hashovací funkce a typ objektu. V tomto kroku bude již existovat v systému *n* dokumentů a bude již existovat provázaná důkazní struktura. Na základě těchto informací a naměřených hodnot bude možné při použití metod regresní analýzy [30] predikovat výpočetní náročnost v čase *t+1*. Podrobněji v kapitole 7.

2.8 Struktura práce a její členění

Struktura a členění předkládané disertační práce kopíruje model abstrakce vědeckého modelování. Od analýzy teoretického poznání v dané oblasti vede přes systémový návrh až jeho empirickému ověřování v rámci výzkumného experimentu.

Struktura práce představuje několik úzce souvisejících částí. V úvodu je po vymezení výzkumného pole pozornost věnována zejména základnímu pojmovému aparátu a jeho vyjasnění. Navazující část je věnována teoretickým základům, s nimiž prezentované návrhy a řešení, budou v souladu a budou je respektovat. Další mapuje současná řešení na poli elektronické archivace a to jak v oblasti komerční, tak v oblasti volně dostupné. Následné se text věnuje konkretizaci bezpečnostních aspektů archivace, které tvoří spolu s identifikací výzkumných problémů a specifikací výzkumných otázek výchozí bod pro stanovení cílů práce. Na tuto teoreticky a analyticky orientovanou část práce navazuje část orientovaná empiricky a systémově. Zde jsou jednak popsány návrhy řešení jednotlivých cílů, ale také výzkumný experiment a jeho výsledky. Práce poté dospívá do poslední části, která shrnuje závěry předchozích kapitol.

Práci lze rozdělit do třech logických, spolu vzájemně souvisejících bloků, kde vždy blok následný vychází z předešlých, které tak tvoří jeho podklad. Je zde zcela nezbytné ovšem podotknout, že tato práce se nesnaží kompletně zmapovat všechny přístupy, projekty a formáty vztahující se k dlouhodobému uchovávání digitálních informací. První blok zevrubně mapující současný stav řešené problematiky by ovšem měl tvořit elementární základ a východiska pro bloky následující.

1. *Analytický blok (I).* Tvoří jej analýza teoretických východisek, vztahujících se ke zvolenému výzkumnému tématu. Tvoří nutný teoretický základ pro návrhovou část práce. Zahrnuje kapitoly 3 - 5. Tato část se snaží nalézt odpověď na první čtyři výzkumné otázky, první dílčí cíl práce a objasňuje první tři výzkumné úkoly vyřčené v kapitole 2.4.3
    - Kapitola 3 provádí analýzu současného stavu poznání v předmětné oblasti. V úvodu je prostor věnován stanovení legislativního rámce, který je nezbytným výchozím podkladem pro tento druh práce. Dále pokračuje analýza současných projektů řešení LTA s důrazem na jejich rozdělení na komerční a veřejné projekty. Další podkapitoly se věnují přehledu souvisejících mezinárodních standardů a norem nebo požadavkům na formát ED určeného k LTA.
    - Kapitola 4 stanovuje obecný model funkčnosti archivu a vymezuje obecné požadavky na bezpečnost svěřených objektů spolu s analýzou hashovacích funkcí.
    - Kapitola 5 podává ucelený rozbor existujících mechanismů zajišťujících bezpečnost ED nezávisle na čase, v jakém byly archivovány.
2. *Návrhový blok (II).* Je tvořen kapitolou 6. Jeho nosnou částí jsou řešení druhého a třetího dílčího cíle. To v sobě také implikuje odpovědi na pátou až sedmou výzkumnou otázku.
    - Kapitola 6 prezentuje návrh systému pro zajištění dlouhodobé integrity jak archivních objektů, tak celého archivu. Jako řešení třetího dílčího úkolu předkládá tato kapitola návrh nového identifikátoru archivovaného dokumentu jako komplementární nástroj k hlavnímu bezpečnostnímu prvku archivu.
3. *Experimentální blok (III).* Ověřování empirických dat pro nalezení odpovědí na poslední výzkumnou otázku a na vyřešení dílčího cíle. To bude možné až zpracováním znalostí získaných z předchozích dvou bloků. Tvoří jej kapitola

7. Návrh sestavený v předchozím bloku bude podroben empirickému výzkumnému experimentu dle metod stanovených v kapitole 2.6. Na závěr tohoto bloku dojde statistickému zpracování výsledného datového souboru a k celkovému vyhodnocení experimentu. Tento blok tak odpoví na poslední výzkumnou otázku, dílčí cíl a výzkumný úkol.

Disertační práci završuje zhodnocení naplnění stanovených výzkumných cílů a otázek, včetně jejich teoretického a praktického přínosu.

3 Analýza současného stavu poznání v předmětné oblasti

Stav problematiky v této kapitole může být východiskem pro stanovení rámcové systémové architektury archivu a její následné specifikace. Je nutné jednoznačně deklarovat, že tato práce si ale neklade za cíl být konkrétním projektem ani nabídkou. Teoretickou část provedené analýzy mají oba přístupy ale společnou.

Nejprve je nutné provést analýzu legislativního a právního prostředí, které je pro práce tohoto druhu značně omezující.

3.1 Základní legislativní rámec a relevantní normy

V této kapitole bude stanoven výchozí legislativní rámec v České republice, Evropské unii a dalších vybraných státech. To že je digitálně podepsaný ED je právním ekvivalentem ručně psané formy bude bráno jako dogma, při jehož neexistenci nemůže být elektronický archiv vybudován[19].

Legislativní rámec, zákony, nařízení a direktivy jsou v „e-“ oblastech zásadním, ale také omezujícím elementem. Právní normy se dotýkají jak oblasti obchodu, fakturace nebo celého businessu tak ve veřejné sféře, govermentu, soudnictví nebo zdravotnictví. Cílem této kapitoly, je zmapovat legislativní prostředí, které se jakýmkoliv způsobem vztahuje k problematice uchovávání ED a jejích právní váhy a to jak v České republice, Evropě tak mimo ni. Tato disertační práce bude níže uvedené zákony plně respektovat a bude s nimi v souladu.

3.1.1 Normy ISO

Doporučujícími předpisy jsou standardy v podobě ISO norem (a jejich implementace do národních prostředí, v ČR toto řeší ČSN). Základní ISO normou je ISO 14721:2012[20], která nahrazuje původní ISO 14721:2003. Jedná o aktuální verzi, schválenou v září 2012. Tato norma nebude zde více rozváděna, protože tvoří výchozí podklad této práce.

[19] Vybudován být může, ale bude na něj nahlíženo jako na nedůvěryhodný.
[20] http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=57284

Pro část věnovanou řešení dílčího cíle vědeckého úkolu jsou zásadní dvě, v době psaní této práce velmi čerstvé, normy a to:

- ISO 16919. Toto doporučení představuje „Best Practices“, které mohou být použity při vyhodnocování důvěryhodnosti archivního úložiště. S tím přímo souvisí aktivita „ISO-PTAB“[21], která zajišťuje ISO audity a certifikace digitálních archivů a hraje významnou roli v instruktáži a certifikaci akreditovaných auditorů.
- ISO 16363. Dokument řeší především tři oblasti, které až do doby vydání této normy byly tzv. “ponechány na zvážení příslušné organizaci“. První oblasti je organizační infrastruktura, která zahrnuje dokumentaci historie změn související s archivem. Druhou oblastí je správa digitálních objektů – týká se zajištění přístupu k archivu ke všem prostředkům, které mohou prokázat původ a autenticitu archivovaných objektů. Třetí oblasti risk management, kde jsou definovány procesy, které musí být dodrženy, pokud se zásadním způsobem mění použitý software.

Do této chvíle (leden 2014), všechny dosud provedené audity digitálních archivů jak v USA, tak v Evropě z drtivé většiny nevyhověly[22] požadavkům výše uvedených standardů. Hlavní příčinou jsou špatné nebo chybějící „Archiving Policies“ [59].

Další ISO normy, které by měly být na tomto místě zmíněny, souvisí s doporučením formátu dokumentu určeného k LTA. Jedná se o:

- ISO 19005-1:2005. Standard specifikuje jak použít dokument ve formátu PDF verze 1.4 (známý též jako PDF/A-1) pro LTA.
- ISO 19005-2:2011[23]. Standard specifikuje jak nakládat s dokumentem ve formátu PDF verze 1.7 (ISO 32000-1) pro LTA (známý též jako PDF/A-2).
- ISO 19005-3:2012[24]. Standard publikovaný 17. října 2012 umožňuje vkládání různých formátů (např. XML) do PDF/A jako kompletních objektů k archivaci.

[21] http://www.iso16363.org/
[22] https://www.tsl.state.tx.us/slrm/blog/?p=4812
[23] http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=50655

Dodržování standardu PDF/A zajišťují tzv. validátory. Dostupných jich je celá řada, ovšem ne všechny dokáží být stoprocentní[25].

3.1.2 Přehled stavu ve vybraných členských státech EU

V rámci přehledu relevantních legislativních předpisů již nebudou (až na odůvodněné výjimky) zmíněny implementace elektronického podpisu (tj. Evropské direktivy 1999/93/EC[26]) do příslušných národních zákonů a nařízení jednotlivých členských zemí. Směrnice Evropského parlamentu a rady musí „de jure" implementovat do svých národních legislativ také členské státy. Zmíněné vybrané členské státy EU, se tak specifickým způsobem odlišují od prosté implementace směrnic Evropského parlamentu a rady.

Právní rámec LTA je zakotven v několika předpisech. Na úrovni EU je zásadní dokument SEC(2005) 1578[27] o nakládání s ED a jejich elektronicky podepsaných verzí. Dále jsou to směrnice Evropského parlamentu a rady a to především o zásadách Společenství pro nakládání s elektronickými podpisy ve znění 2008/1137/ES[28] z 22. října 2008. Elektronickou archivaci nepřímo ovlivňují další předpisy jako například směrnice upravující elektronický obchod (2000/31/ES ve znění pozdějších předpisů) a další. Ty obecně definují povinnost, v případě archivace dokladu v elektronické podobě prokázat, kdo je jeho skutečným vystavitelem a zachovat nepopiratelnost tohoto vyhotovení (autenticitu původce). Pro tuto práci nejsou ale zásadní a proto nebudou dále zmiňovány a rozvíjeny.

Další dokumenty jsou již na národních úrovních a definují obecné podmínky implementace dlouhodobých archivů – tj. platí pro všechny státy EU prostřednictvím příslušných národních zákonů. Většinou zahrnují národních knihovny, archivy a archivaci národního bohatství včetně převodu do digitální

[24] http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=57229
[25] http://www.pdflib.com/fileadmin/pdflib/pdf/pdfa/2009-05-04-Bavaria-report-on-PDFA-validation-accuracy.pdf
[26] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:52006DC0120:EN:NOT
[27] http://ec.europa.eu/transparency/regdoc/rep/2/2005/EN/2-2005-1578-EN-1-0.Pdf
[28] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2008:311:0001:0054:CS:PDF

podoby. Na úrovni EU je na místě ještě zmínit aktivitu IP/11/17[29], která si klade za cíl převod Evropského kulturního dědictví do elektronické podoby a jeho dlouhodobé zachování pro budoucí generace.

3.1.2.1 Slovinsko

Jedna z prvních zemí EU, jež adoptovala digitální archivaci do svých legislativních předpisů, bylo Slovinsko a to zákonem č. 30/2006 z 23. 3. 2006[30] a poté nařízením vlády č. 86/2006 z 11. 8. 2006[31] ve znění pozdějších předpisů. Blažič [10] se ve svých publikacích na tuto legislativní normu odkazuje velmi často.

Slovinsko je v oblasti právních norem a vůbec LTA velkým propagátorem a bylo jedním z pionýrů v této oblasti. Projekty na LTA a způsoby jejího řešení jsou rámci této práce mnohokrát zmíněny také proto, že autor se na vývoji těchto metod nepřímo podílel (viz. kapitola 2.1).

3.1.2.2 Německo

Německo má jako zásadní legislativní dokument tzv. Technical Guidelines označený jako BSI TR-03125 (Preservation of Evidence of Cryptographically Signed Document) [16], který zásadním způsobem ovlivňuje způsob uchování důkazních záznamů u archivovaných objektů. Představuje exaktní popis, jak má být nakládáno s elektronicky podepsanými dokumenty pro potřeby důvěryhodné dlouhodobé úschovy.

3.1.2.3 Dánsko

Od 1. 2. 2005 je v Dánsku elektronická archivace povinností. Strategie dlouhodobé ochrany digitálních archiválií[32] z ledna 2013 definuje pravidla pro dlouhodobé uchovávání a tvoří jistou paralelu s českým prostředím. Dánsko má jako jedna z mála zemí funkční model dlouhodobého uchovávání elektronických záznamů. Je uváděno jako modelový příklad v rámci EU, kde se podařilo úspěšně implementovat

[29] http://europa.eu/rapid/pressReleasesAction.do?reference=IP/11/17&format=HTML&language=en
[30] http://www.uradni-list.si/1/content?id=72425
[31] http://www.uradni-list.si/1/content?id=74975
[32] http://www.sa.dk/media(4826,1033)/Strategy_for_archiving_digital_records.pdf

tuto strategii. Nositelem fungujícího systému LTA je Dánsky národní archiv. Základem úspěchu bylo stanovit jednotný standardizovaný formát pro všechny úřady státní správy včetně metadat [63]. Ten je pak ve formě SIP balíčku, přenášen do společného archivu. Archiv je pak zodpovědný za přidělení unikátního identifikátoru (obdoba jednacího čísla).

3.1.3 Česká Republika

Z české legislativy jsou nejvýznamnější zákony č. 101/2000 Sb., o elektronickém podpisu (respektive jeho novela č. 227/2004 Sb.) a o archivnictví a spisové službě č. 499/2004 Sb. Ten dále doplňuje vyhláška 645/2004 Sb. Poslední novela, je v době této práce z 1. 7. 2012 (zákon č. 167/2012 Sb.). Ta mění znění § 3 odst. 4), který je pro tuto práci zásadní na text: „Uchovávání dokumentu v digitální podobě provádí určený původce postupem zaručujícím věrohodnost původu dokumentu, neporušitelnost jeho obsahu a čitelnost dokumentu, a to včetně údajů prokazujících existenci dokumentu v digitální podobě v čase“. Dále doplňuje a konkretizuje § 65 odst. 8 takto: „Veřejnoprávní původci vykonávající spisovou službu v elektronické podobě v elektronickém systému spisové služby podle § 63 odst. 4 mohou ve svých spisových řádech upravit používání zvláštních technologických prostředků, kterými lze výlučně pro potřeby příslušného původce nahradit uznávaných elektronický podpis, uznávanou elektronickou značku nebo kvalifikované časové razítko; tyto zvláštní technologické prostředky musí umožnit zjištění jakékoliv následné změny dat v dokumentu a jednoznačně ověřit identitu osoby, která je k němu připojila. Ustanovení § 69 odst. 5 se na zvláštní technologické prostředky použije obdobě.“ Stejný odstavec pak deklaruje, že dokument je považován za pravý, není li prokázán opak, což je analogicky přeneseno bez použití elektronického podpisu hlavním cílem této práce. Poslední důležitým bodem v tomto zákoně je text v § 60c odst. 2, na jehož konci se píše: „Jednoznačný identifikátor je označení dokumentu zajišťující jeho nezaměnitelnost. Jednoznačný identifikátor musí být s dokumentem spojen. Strukturu a podrobnosti vytváření a nakládaní s jednoznačným identifikátorem

stanový prováděcí předpis“. Tento výklad podporuje stanovený dílčí cíl této práce, jímž je návrh univerzálního jednoznačného identifikátoru dokumentu.

Na závěr této podkapitoly je vhodné zmínit velmi přínosný je článek [55] popisující průlom v elektronické archivaci dokumentů. Je zde zmíněno, že konečně panuje shoda v odkazech na legislativní předpisy napříč kontinenty.

3.1.4 Zákony mimo EU

Tyto země se nemusí řídit směrnicemi a nařízeními EU a mají proto odlišný přístup k této problematice.

3.1.4.1 USA a Kanada

V USA proběhla implementace elektronického podpisu, jeho používání a uznávání federálním zákonem nazvaným UETA[33]. Federální zákon adoptovaly všechny státy kromě Illinois, New Yorku a Washingtonu, které přijaly vlastní zákony. UETA definuje požadavky na plný právní, účinek a důkazní sílu ED. Implikuje E-Sign[34] act z roku 2000[35].

Zákon „Electronic Discovery Act“ o povinné elektronické archivaci byl v roce 2009 byl přijat také nejlidnatějším státem v USA, kalifornií podpisem kalifornského guvernéra[36]. Především zahrnuje elektronické informace mezi důkazní prostředky. Zahrnuje možnost specifikace formátu, ve kterém elektronické informace mají být předloženy.

V Kanadě pak je to zákon S.C. 2004 (An Act to establish the Library and Archives of Canada, to amend the Copyright Act and to amend certain Acts in consequence[37]).

[33] http://www.ncsl.org/issues-research/telecom/uniform-electronic-transactions-acts.aspx
[34] http://www.fdic.gov/regulations/compliance/manual/pdf/X-3.1.pdf
[35] http://www.gpo.gov/fdsys/pkg/PLAW-106publ229/content-detail.html
[36] http://www.weil.com/california-act/
[37] http://laws-lois.justice.gc.ca/eng/acts/L-7.7/

Ten je základní podkladem pro projekt kanadské knihovny a archivů LAC[38] Trusted Digital Repositories [67].

3.2 Přehled vybraných projektů dlouhodobé archivace a jejich analýza

Na téma LTA byla publikována již celá řada prací (například [18]), zabývajících se problematikou jakým způsobem zajistit časovou univerzálnost použitých metod. Na jejich základě vznikla celá řada produktů, jejichž krátký přehled je součástí této podkapitoly. Ty jsou rozděleny do dvou hlavních skupin a to na financované z veřejných zdrojů a na projekty komerční. Důraz bude přitom kladen na analýzu přístupů k zajištění integrity archivovaných objektů, proto výběr projektů zahrnuje výhradně ty, jež ve svém řešení využívají relativní časovou autentizaci.

3.2.1 Veřejné projekty[39]

K několika málo, které jsou dostupné bezplatně a plně otevřeně se řadí především produkty národních archivů a knihoven. Jedná se z velké části o projekty spolufinancované Evropskou Unií. Dokument [40] mapující projekty LTA se spolufinancováním EU dává velmi dobrý přehled v jakém stavu je digitální archivace v jednotlivých členských státech. Další přehled o současně řešených projektech LTA respektujících evropskou legislativu lze nalézt na stránkách RP7[40]. V době psaní této práce je ze sedmého rámcového programu financován projekt[41] kladoucí si za cíl zhodnotit výhodnost úschovy digitálních informací. Projekty, jež deklarují, že respektují OAIS[42] (avšak bez zmínky o využití relativní časové autentizace) jsou například:

- Protage[43] (veřejný) – Estonsko, digitalizace estonských archiválií.
- Preservica[44] (komerční) – produkt fy Tesella, využívá Amazon Cloud, zavádí active preservation[45] přístup, využívá pay-per-use model[46].

[38] http://www.collectionscanada.gc.ca/digital-initiatives/012018-4000.01-e.html
[39] http://wiki.opf-labs.org/display/TR/Digital+Preservation+Tool+Registry
[40] http://cordis.europa.eu/fp7/ict/telearn-digicult/digicult-preservation_en.html
[41] http://4cproject.net/
[42] Open Archival Information System – podrobně kapitola 4.
[43] http://protage.eu

- DRAMBORA[47] (veřejný) – metodika auditu dlouhodobých archivů, včetně softwarového toolkitu.

3.2.1.1 Xena

Za jednu z nejzdařilejších implementací je považován produkt týmu digitální úschovy při australského národním archivu s názvem „Xena“ [77] založený výhradně na tomto standardu. Xena funguje na principu „normalizace“ souborů do tzv. „xena“ formátu. Výsledné soubor je větší než původní a předpokládá, že jej bude možné v budoucnu „de-normalizovat“ na formát, který bude v daný časový okamžik (období) běžný. Původní soubory (v původním formátu) jsou archivovány spolu s jejich XML normalizovanou podobou. To má především účel zaručení integrity, protože umožní uživateli si kdykoliv znovu vytvořit „Xena“ formát a porovnat bitovou strukturu (pomocí hashe) s původním XML formátem.

Xena je pouze jednou z komponent systému LTA australského národního archivu nazvaného „Digital Preservation Software Platform“[48]. Proto nevytváří katalogy a neeviduje soubory, které byly „normalizovány“.

3.2.1.2 Archivematica[49]

Velmi nadějný open-source projekt společnosti Artefactual Systems pro LTA digitálních objektů a pro podporu procesů s LTA spojených. Základem je referenční model OAIS a její vývoj byl financován mimo jiné i z fondů UNESCO. Oproti konkurenčním řešením klade na bezpečnost velký důraz a řeší ji mj. za pomoci relativní časové autentizace. Mezi instituce využívající Archivematicu se řadí zejména City of Vancouver Archives, UBC Library, Rockefeller Archive Center a International Monetary Fund či v rozvojových zemích pod záštitou UNESCO Memory of the World Subcommittee on Technology. Základní charakteristikou řešení jsou

[44] http://preservica.com
[45] O objekt se stará tým odborníků, jež zajišťuje jeho čitelnost (včetně migrace).
[46] http://assets-production.govstore.service.gov.uk/Giii%20Attachments/TESSELLA%20Limited/Bids/Service%20Definition%20for%20Preservica.pdf
[47] Digital Repository Audit Method Based on Risk Assessment, http://www.repositoryaudit.eu/
[48] http://dpsp.sourceforge.net/
[49] http://www.archivematica.org/wiki/Main_Page

tzv. micro-services. Jde o webové služby implementující vždy jednotlivé atomické činnosti, potřebné pro zpracování archivačních požadavků. Tyto micro-služby mohou běžet na dedikovaných serverech a tím se starají o rozložení zátěže. Centrální část tohoto řešení (central micro-services server) deleguje jednotlivé úlohy na pracovní servery. Tímto je dosaženo jednoduché horizontální škálovatelnosti celého řešení. Archivematica byla také zvolena jako základ řešení českého národního archivu.

3.2.2 Komerční projekty

U komerčních projektů nejsou většinou dostupné zdrojové soubory ani podrobná dokumentace, tudíž lze jen stěží určit přesnou implementaci zajištění integrity objektů archivovaných v relativním čase a je nutné se spolehnout pouze na informace a dokumenty publikované v omezené míře autory (například jako tzv. „Whitepapers“ nebo „Technical Whitepapers“). Jako příklad úspěšné implementace bude nakonec zmíněn jeden čistě komerční (základ původně vznikl na Univerzitě v Bellcore [19]).

3.2.2.1 AbsoluteProof

Produkt od společnosti Surety [50] zajišťuje integritu tím, že potvrdí, že dokument byl vytvořen v určitém čase a od té doby nebyl změněn. Jedná se o typický uzavřený produkt typu server-tlustý klient využívající systém, který Surety nazývá „Hash-Chain Linking“. Produkt implementoval „Merkle Tree“, popsaný podrobně v již zmíněné publikaci [18] a kapitole 5.3. Jak sami autoři uvádí, jejich systém je založen na standardech důvěryhodných časových razítek ISO/IEC 18014-3 a ANSI X9.95. Je naprosto zásadní uvést fakt, že jako jedna z hashovacích funkcí je použita „RIPEMD-160“, kterou se ovšem již kvůli krátkému bitovému výstupu již nedoporučuje používat [28]. Surety to řeší tak, že výslednou hodnotu této funkce zřetězí s výslednou hodnotou SHA-256 a vytvoří tak důkazní záznam [29].

3.2.2.2 SecDocs

MiddleWare vytvořený díky spolupráci německých společností Fujitsu a OpenLimit. Je deklarována jeho shoda s požadavky BSI TR-03125 [16] na důvěrný dlouhodobý

archiv ED. Autoři deklarují, že produkt je založen na standardech ERS [32] a elektronickém podepisování dle PKI [26]. Podporuje hashovací algoritmy až do SHA-512. Dokumenty jsou uchovány striktně ve formátu XML.

3.2.2.3 GuardTime[50]

Estonské řešení na bázi technologie bezklíčové infrastruktury (KSI). Jádrem tohoto produktu jsou systémy provázaných hashů, tak jak je v několika publikacích popsal Ahto Buldas ([17],[18],[19],[20]). Toto řešení tedy přímo vychází z univerzitního výzkumu[51]. Produkt, zahrnuje zajištění autenticity jak pro cloud, email tak pro samotné dokumenty (nejen elektronické ale také papírové) či celé adresářové struktury. Pro tuto práci je způsob bezpečnostní řešení jednou z hlavních inspirací.

3.2.2.4 eKeeper[52]

Slovinský projekt vyvíjený v SETCCE při Jože Štefan Institutu v Lublani. Produkt je primárně učen pro archivaci účetních záznamů. V roce 2007 byl produkt pilotně testován ve Slovinské pobočce mobilního operátora Vodafone. Projekt vychází z vědecko-výzkumné činnosti ve spolupráci s „*University of Ljubljana*“ a stejně jako v případě estonského produktu „*GuardTime*“, byl tento výzkum završen vydáním komerčního produktu. Jeho nasazení je podpořeno mj. i velmi pokročilou Slovinskou legislativou.

Autor této práce se během svého výměnného pobytu v rámci doktorského studia podílel na vývoji části tenkého klienta (viz kapitola 2) a spolupracoval při návrhu XML specifikace ERS.

3.3 Taxonomie elektronických dokumentů

Rozdělení do tříd představuje výchozí podmínku pro práci s ED v průběhu LTA. Faktory, ovlivňující dlouhodobé uchování jsou:

- Plná reprodukovatelnost obsahu.
- Integrita obsahu.

[50] http://www.guardtime.com/ksi-technology/
[51] http://www.guardtime.com/ksi-technology/academic-research/
[52] http://www.ekeeper.si

- Autentičnost.

Posledním faktorem je schopnost archivu prokázat, že daný ED byl vytvořen či uložen danou entitou, případně v daný časový okamžik. Zásadním východiskem je fakt, že vlastnosti každého ED jsou vyjádřeny metadaty, jež jsou v archivu jeho nedílnou součástí. V rámci taxonomie objektu ED je vhodné rozdělení do základních typů tříd, kde snahou bude minimalizovat jejich počet, aby rozdělení bylo pouze rámcové:

- Formát.
- Zajištění původu a integrity.
- ~~Doba archivace (historická hodnota).~~

Ostatní klasifikace jako je historická hodnota dokumentu (která je zásadní pro archiváře a badatele), není pro tuto práci podstatná. To především z důvodu, že všechny archivované ED v LTA mají dlouhodobou historickou hodnotu. Za dlouhodobé uchování ED je považováno období 10 a více let (Tabulka 1).

Tabulka 2: Klasifikace elektronických dokumentů

Typ elektronického dokumentu	Popis
Text	Bitový řetězec posloupnosti kódovaných znaků libovolné sady. Zahrnuje také zdrojové kódy binárního formátu.
Binární	Libovolná posloupnost bitů, jejichž RAW formát nemá vypovídající hodnotu.
Otevřený	Dokumentace (popis) formátu je k dispozici jako otevřený a dobře popsaný standard. Tento formát je postačující podmínkou pro čitelnost archivovaného ED v budoucnu. Doporučuje se také spolu s ním archivovat samotný standard.
Uzavřený	Popis formátu není standardizován. Jedná se o tzv. proprietální formát, tj. formát jako uzavřené know-how tvůrce. Není zaručena dlouhodobá čitelnost a často vyžaduje uchování programů, které umožní prezentaci daného typu dokumentu. Jeho používání ve většině případů vyžaduje licencování.

Zdroj: [autor]

Ne stále vyřešenou otázkou je požadavek na sjednocení formátu při potřebě minimalizace ztráty věrnosti a modifikaci obsahu originálu. Ideální by byl konsensus odborné komunity nad jedním typem formátu, který by splňoval všechny požadavky (informační hodnota, důvěryhodnost atd.) na něj kladené. Využití formátů proprietálních, ale také otevřených navázaných na licenční politiku, či patentní ochranu, by v případě jejich použití v prostředí v LTA celý systém výrazně prodražilo. Ochrana duševního vlastnictví bohužel v posledních letech vyústila ve stav, které je pro společnost jako celek spíše ke škodě než užitku [38]. Formát by měl splňovat požadavek, aby jeho důvěryhodnost nebyla závislá na technologickém řešení LTA. Klasifikaci na nejvyšší úrovni představuje Tabulka 2.

Pro tuto práci je podstatná taxonomie na úrovni zajištění autentičnosti dokumentu. Ta je na straně uživatele způsoby uvedenými v následující tabulce:

Tabulka 3: Autentičnost elektronických dokumentů na straně původce

Typ zajištění	Popis
Bez zajištění	Žádný záznam, prokazující autentičnost dokumentu není k dispozici.
Elektronická značka	Především pro právnické osoby. Fyzická osoba je spojena výhradně se (zaručeným) elektronickým podpisem.
Elektronický podpis	Založený na libovolném certifikátu, tj. i na certifikátu vydaném nedůvěryhodnou certifikační autoritou.
Zaručený elektronický podpis	Založený na kvalifikovaném certifikátu vydaným důvěryhodnou certifikační autoritou.

Zdroj: [autor]

3.3.1 Požadavky na formát elektronického dokumentu

Pro odkaz budoucím generacím, pro které je především obsah jednotlivých archivovaných objektů určen je nutné použít pouze kvalitně zdokumentovaný standard. Obsah sdělení je sice důležitější než jeho forma, ovšem sémantika je a bude vždy nutným aspektem. Existence vysokého množství formátů a jejich verzí a podverzí, je pro účely LTA absolutně nevyhovující. Obecně užívané formáty (generické) neposkytují dostatečně přesné rozlišení nutné pro potřeby LTA. Článek [34] deklaruje, že celý systém archivace by měl být založen výhradně na

neproprietálních (otevřených, standardizovaných a kompletně zdokumentovaných) technologiích.

3.3.1.1 Doporučení vhodného formátu pro potřeby dlouhodobé archivace

S ohledem na reprezentaci dokumentů a do tohoto bodu zjištěných skutečností, autor jako optimální formát dokumentů určených k LTA navrhuje:

- XML jako primární formát pro ukládání jak samotných dokumentů tak metadat.
    - ODF (Open Document Format) je založen na XML a jeho verze 1.1 byla standardizována jako ISO/IEC 26300:2006/Amd 1:2012. I z tohoto důvodu je ODF doporučený formát pro LTA.
- PDF/A jako původní formát dokumentu a sekundární (záložní) formát sloužící v případě, že reprezentace primárního formátu nebude možná, nebo bude možná pouze částečně, což by snižovalo jeho vypovídající hodnotu. Tomuto formátu se blíže věnují standardy ISO 19005-1:2005 a 19005-2:2011 (viz kapitola 3.1.1). Zde je nutné poznamenat, že s výhradním použitím tohoto formátu pro LTA, jsou spojena některá rizika[53].
- EPUB („Electronic Publication“). Jedná se o formát vyhovující specifickým potřebám tzv. elektronických knih. Byl vytvořen IDPF konsorciem a v současné době (únor 2013) je verzi 3.0 schválené v říjnu 2012. Strukturou se jedná koncepčně o podobný[54] formát jako OpenXML formát od společnosti Microsoft. Vlastní obsah je reprezentován pomocí XML a HTML5 tagů. Má již širokou podporu (například od roku 2012 jej podporuje společnost Apple („iBook“) na svých zařízeních) mezi výrobci čteček elektronických knih a i v ČR se jedná o neoficiální formát pro e-knihy. V roce 2012 byla publikována studie [42], podle které je to právě EPUB, který je díky své struktuře a použitých standardech nejvhodnějším formátem pro LTA.

[53] http://www.openplanetsfoundation.org/system/files/PDFInventoryPreservationRisks_0_2_0.pdf
[54] ZIP kontejner se specifickou adresářovou a souborovou strukturou.

K výše uvedenému je vhodné podotknout, že existuje celá řada nástrojů (jak offline tak online[55]) na konverzi a migraci formátů. Ke každému archivovanému formátu by měla být vedena (a archivována) příslušná dokumentace, exaktně popisující použité standardy[56]. Dokumentace bude v systému archivu uchována mimo samotný archivovaný ED a budou na ni aplikována stejná bezpečností politika archivu jako na „běžné“ dokumenty s následujícími výjimkami:

- Dokumentace bude v archivu uložena, dokud bude existovat minimálně jeden dokument popisovaného formátu.
- Při archivaci dokumentu bude vždy provedena kontrola existence dokumentace popisující příslušný formát. Pokud dokumentace nebude nalezena, příslušný dokument bude archivem odmítnut a původce bude vyrozuměn.

3.4 Výchozí obecné požadavky pro dlouhodobou archivaci

Pro účely LTA ED jsou stanoveny čtyři klíčové atributy dlouhodobých dokumentů (věrohodnost, integrita, čitelnost a nezávislost na úložišti) [78], definované společností software602, jež se LTA věnuje dlouhodobě. Zajímavé shrnutí požadavků na LTA i TAA je také v článku od p. Höniga [35].

3.4.1 Verifikace platnosti a stavu dlouhodobě archivovaných dokumentů

Archiv podporuje možnost prokazování existence a integrity dat (bez časové vazby i s ní). Jinými slovy archiv poskytuje prostředky pro uchovávání průkazních záznamů pro podepsané i nepodepsané dlouhodobě uchované datové objekty. Další subjekt, kterému musí služba LTA poskytnout tyto informace jsou auditoři, resp. nezávislé auditorské nástroje třetích stran (více v kapitole 4.4). Je tedy periodicky ověřována integrita archivovaného dokumentu. Kontroly jsou prováděny na následujících elementech:

[55] http://www.zamzar.com/

[56] Převod dokumentů a dalších souborů pro (dlouhodobou) archivaci, lze nalézt například na webu Kongresové knihovny (http://www.digitalpreservation.gov/tools/). Většina z nich deklaruje dodržení transformace do archivního objektu dle OAIS.

- Formát

  Kontrola čitelnosti uloženého dokumentu a jeho binární struktura odpovídá standardu a signaturám dle normy.
- Zabezpečení

  Kontrola zda metody zabezpečení uložených objektů jsou stále účinné. Tato aktivita je pro předkládanou práci zásadní.
- Indexy a reference

  Kontrola, že dokument je stále vyhledatelný na základě indexů a referencí, které jsou součástí metadat.

U ED se automaticky předpokládá archivace bezpečnostních mechanismů přicházejících ze strany klienta (pokud existují). Tj. zaručených digitálních podpisů. Jediným vážným problémem dlouhodobé povahy zůstává již zmíněné stárnutí použitých kryptografických algoritmů. Autenticita dokumentu bude v případě navázání pouze na integritu zaručenou původcem přímo vázána na podpisová schémata použitá pro vytvoření zaručeného elektronického podpisu, stejně tak na systémová metadata a hashe. Jinými slovy, po čase bude možné díky zvyšování výpočetního výkonu nalézt duplicitu k již existujícímu hashi z jiného binárního zdroje než z originálního. Na straně uživatele bude tento problém reprezentován skutečností, že bude možné odvodit u asymetrické kryptografie z veřejného klíče klíč soukromý. Tímto způsobem bude narušen vyžadovaný princip a předpoklad nepopiratelnosti. Bližší řešení tohoto problému nastíní kapitola 5.

Jedním z problému, zpomalujících plošné zavedení elektronicky podepsaných dokumentů, je stále ještě velmi malé množství ověřovacích služeb pokrývajících celou EU [55]. Tomu se snaží předcházet aktivity jako například SecuStamp[57].

[57] http://www.secustamp.eu/

3.5 Metadata

Metadata jsou v TAA technicky a procesně tak zásadním prvkem, že je nutné jim věnovat zmínku. Reprezentují jeden z nejdůležitějších elementů LTA, který je pevně svázaný s každým ED a tvoří ták nedílnou součást archivovaného objektu. Jejich význam spočívá především v:

- Vyhledávání dokumentů v archivu (indexy a reference).
- Podklady pro zobrazení a reprezentaci dokumentů (typ dokumentu, formát, verze).
- Prokazování autentičnosti a důvěryhodnosti ED (hashe, certifikáty, časová razítka („TS“), případně dlouhodobé elektronické podpisy [73]).

Data o datech se dělí do třech skupin:

- Původní (získaná od původce).
- Získaná (z vlastností dokumentu a jeho a obsahu).
- Systémová (dodaná archivem).

Jak uvádí [13] metadata jsou dvojího typu:

- Interní (Embedded).

  Metadata uvnitř tvořící spolu s obsahovou stránkou dokumentu jeden bitový celek (jeden soubor). Při konverzi nejsou vždy všechna tato metadata zachována [34].
- Externí.

  Metadata, které spolu s obsahem dokumentů netvoří jeden soubor, ale jsou uchovávána separátně. Primárním klíčem je v tomto případě ID dokumentu.

Dobře definovaný standard pro LTA musí umožňovat metadata v průběhu archivace extrahovat a uložit odděleně do speciálních metadatových úložišť. [48] ovšem poznamenává, že hlavním problém je především sjednocení formátu mezi skupinami produktů spadajících do kategorií DMS a referenčních modelů.

Z hlediska jejich editovatelnosti se „data o datech“ dělí na:

- Přepisovatelná. Nenesou zásadní systémovou informaci a nemohou ovlivnit důvěryhodnost ani bezpečnost dokumentu, ke kterému se vztahují. Je doporučeno, aby byla tzv. „*append-only*“ (po archivaci dokumentu lze další záznamy pouze přidávat).
- Nepřepisovatelná. Zásadní metadata pro bezpečnost archivovaného dokumentu. Jsou povinnou součástí tvořící spolu s dokumentem archivní objekt.

Metadata pro potřeby LTA musí obsahovat:

- Administrativní metadata.

  Slouží ke správě digitálních objektů. Často se dělí na další podskupiny:

  - Identifikační metadata.

    Obsahují jednoznačný identifikátor (reference a vyhledávání objektu a v něm obsaženého dokumentu)[58].
  - Technická metadata.

    Zahrnují především popis formátu. Skládá se z:

    - Formát binární reprezentace dat.
    - Aplikace, která původně formát vytvořila a její verze.
    - Podpůrné informace (operační systém, platforma atd.).
- Popisná metadata.

  Popis obsahu ED pro vyhledávání a kategorizaci dokumentů v LTA systému. Zahrnuje intelektuální popis, tj. obsah, název, kategorie, klíčová slova, autor a další.
- Referenční (strukturní) metadata.

  Integrují všechny odkazy související s dokumentem. To se týká odkazů na:

  - Standard nebo dokumentaci popisující daný formát jeho binární reprezentaci.

[58] Identifikátor musí být unikátní po celou dobu archivace. Volí se globálně jednoznačný identifikátor dle URI. Pro zajištění jednoznačnosti identifikátoru, bude součástí této práce také schémata, odkazující se hash dokumentu, resp. objektu.

    - Další verze dokumentu (jsou-li k dispozici).
    - Odkazy na osoby – původce, vlastníka, autora apod.
    - Odkazy na elektronické podpisy a certifikáty.
    - Licenční podmínky využití (jsou-li nutné).
- Kontext archivovaného dokumentu.
    - Klíčová slova; účel archivace; abstrakt dokumentu; případné další.
- Historie archivovaného dokumentu.
    - Převzetí systémem LTA (čas, kdo).
    - Uložení (kdy byla provedena vlastní archivace).
    - Přístupy (čtení, pokus o modifikaci metadat a další).
    - Skartace.
    - Z hlediska kompatibility informace o nutné migraci.
- Ochrana a bezpečnost archivovaného objektu. Ochrana integrity a autentičnosti archivačního balíčku.
    - Hash archivačního balíčku.
    - ID hashovacího algoritmu.
- Řízení přístupu – odpovídá politice archivu a v této práci nebude řešeno.

Standardy pro reprezentaci metadat jejich formátům a kontejnerům se tato práce nebude věnovat, protože o této problematice byla napsána cela řada publikací (jako jednu z posledních velmi zdařilých je nutné jmenovat [37]).

3.6 Shrnutí výchozích standardů a doporučení

Všechna navrhovaná řešení, naplnění stanovených cílů a výsledné přínosy jsou založeny, nebo vycházejí, z následujících standardů a doporučení technického charakteru. Ty se týkají všech oblastí LTA a dlouhodobého uchovávání ED:

- LTA: OAIS, TRAC/TDR, ISO/TR 18492.
- Reprezentace a uložení metadat: Dublin Core, METS, PREMIS, ISAD(G).
- Správa digitálních záznamů: ISO-15489, MoReq2010, ISO/TR 15801.

V této fázi nelze opomenout, ani české normy a doporučení, týkající se bezpečnosti IS (například ISO/IEC 17799 [24]), případně platný XML standard[59] pro předávání ED a metadat do archivu. Další oblastí, která nepřímo ovlivňuje LTA jsou tzv. navržené (de-facto) standardy, například:

- Bezpečnost digitálních záznamů: XMLERS (RFC-6283), DSSC (RFC-5698) – podrobněji kapitola 5.

V rámci kapitoly 4 byl sestaven analytický přehled definující mantinely předmětné oblasti zkoumání. V úvodu přehled legislativní, jak v ČR, tak ve vybraných zemí EU a mimo ni. V další části na vybrané, již existující projekty dlouhodobé archivace, navázala kapitola věnující se vlastnostem a doporučeným formátům archivovaných dokumentů a stručnému přehledu s nimi souvisejících metadat.

[59] http://www.mvcr.cz/nsesss/v1/nsesss.xsd

4 Výchozí rámec stanovující téma vědního bádání

Rámec je reprezentován zjednodušujícím modelem, který představuje nejvyšší úroveň abstrakce. Ten musí jako systémové řešení dlouhodobé úschovy objektů splňovat podmínky shody se standardy (kapitola 3.6).

4.1 Obecný referenční model

Kunstová [46] ve své publikaci zmiňuje model OAIS. Jedná se o obecný referenční model a má zásadní význam na architekturu archivu a jeho funkce jako jsou autentičnost a bezpečnost, resp. důvěryhodnost. Jeho popisu, významu i rozšíření a aplikacím do konkrétních oblastí informatiky bylo vydáno mnoho publikací (za všechny je nutné zmínit alespoň [49] a [76]). Z veškeré prostudované literatury a deliverables, která se tohoto modelu buď týkala, nebo se snažila o jeho implementaci, budou v následujících odstavcích explicitně zdůrazněny a vysvětleny pouze ty části, které mají pro tuto práci zásadní význam. Každý dlouhodobý archiv v obecné rovině musí zajistit minimálně tyto tři funkcionality:

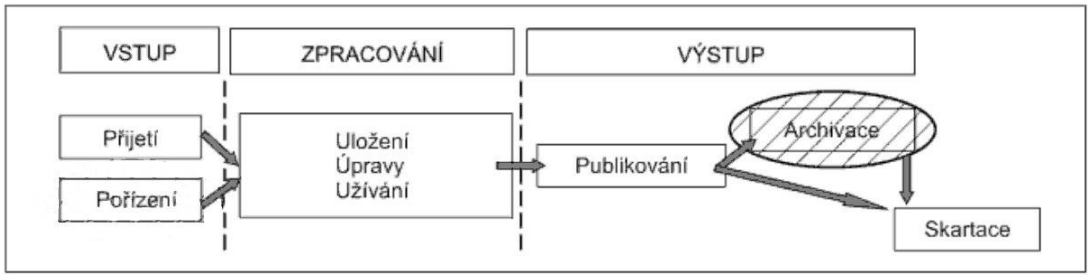
1. Digitální objekt musí být řádně připraven k přijetí archivu, který se o něj bude dlouhodobě starat a udržovat jej.
2. Archiv se bude o objekt, starat takovým způsobem, aby byl přístupný a použitelný bez ohledu na jakékoliv změny v rámci dlouhodobého životního cyklu.
3. Objekt musí být z archivu získatelný skrz jeho popisné informace (metadata) a zkontrolován proti politice přístupových a dalších oprávnění.

4.1.1 Dokument a jeho životní cyklus v archivu

Kunstová [46] pojednává o životním cyklu dokumentu v rámci ECM. Ze schématu (Obrázek 3), lze vypozorovat, že archivace je předposledním krokem. Poslední krokem životního cyklu je skartace, jejímž předkem je kromě publikování také archivace. Tím je potvrzen teorém z [49], že skartace je součástí procesů v rámci LTA.

Příprava pro archivaci a výběr dokumentů, které se budou archivovat, probíhá již v průběhu zpracování dokumentu. Vybrané dokumenty pro předání do archivu, musí mít formu předepsanou jeho politikou. Do archivu musí být předány tři základní součásti, ať již formou balíku nebo jednotlivě, bez kterých není možná bezpečná LTA. Jedná se o:

- Dokument v jednom z podporovaných standardizovaných formátů.
- Vhodná metadata.
- Zabezpečení dokumentu a metadat, které zajistí autentičnost a integritu při převzetí archivem (kapitola 5, [65], [53]).



Obrázek 3: Životní cyklus dokumentu v ECM Zdroj: [46]

Workflow ED a jeho životní cyklus existuje také v rámci samotného procesu archivace (Obrázek 4).

4.1.2 Archivní workflow

U objektů se ověřují tři zásadní prvky:

1) Formát (kontroluje se shoda se standardem a politikou archivu[60]).
2) Metadata
   - Kontrolují se hodnoty polí metadat, jejich shoda s požadavky a politikou archivu.
   - Kontroluje se jejich formát[61].

[60] Existuje několik utilit vyvinutých většinou národními archivy pro automatickou detekci a určení typu formátu elektronického dokumentu. Jedná se například o Digital Record Object Identification (DROID; http://sourceforge.net/apps/mediawiki/droid/) nebo Global Digital Format Registry (GDFR; http://www.gdfr.info/), či FIDO (http://www.openplanetsfoundation.org/software/fido), eventuálně JHOVE (http://sourceforge.net/projects/jhove/), případně FITS, integrující několik utilit (http://www.openplanetsfoundation.org/blogs/2013-04-30-getting-fits-shape).

3) Prvky zabezpečující autentičnost a integritu.
   - Kontrolují se buď u Doc_s nebo zvlášť u jeho prvků.

Procesně se workflow Doc_x v archivu dá chronologicky popsat sledem následujících procesů (Obrázek 4):

1. Příjem příslušným modulem a provedení následných operací:
   1.1. Převzetí objektu Doc_s na vstupu.
   1.2. Kontrola Doc_s (volitelně).
   Zahrnuje kontrolu formátu Doc_n a $Doc_{m(n)}$. Doc_s je pak opatřen systémovými zabezpečovacími prvky a důkazy pravosti (hash, resp. TS).
   1.2.1. Odmítnutí Doc_s.
   Pokud jsou zjištěny závady v předchozím kroku, je Doc_s odmítnut a původce je informován. Po opravě je proces iniciován znovu od kroku 1.1.
   1.3. Vytvoření Doc_a.
   Transformace do podoby (bitové struktury), v němž bude uložen v archivu. Jedná se o původní dokument, deskriptivní metadata a bezpečnostní prvky[62].
2. Uložení indexů a metadat.
   2.1. Vytvoření (extrahování) indexů z metadat.
   Určení primárního klíče a sloupců tabulky pro indexaci, které bude potřebné pro vyhledávání.
   2.2. Uložení metadat do databáze. Tento krok zahrnuje také uložení systémových metadat[63].

[61] OAIS nespecifikuje (velmi vágně) jak mají vypadat (být strukturována) metadata. Za tímto účelem vzniklo několik dalších standardů jako METS, PREMIS a další.

[62] Převod samozřejmě nesmí žádným způsobem ovlivnit autentičnost dokumentu, z čehož je zřejmé, že půjde především o přidávání dodatečných informací, respektive o rozšíření původního Doc_s (požadavky na strukturu objektů uložených v archivu [52]).

[63] Tato metadata tvoří základ pro důvěryhodné uchování a případné pozdější ověření integrity dokumentu.

3. Zápis Doc_a do datového úložiště.
4. Proces dlouhodobé úschovy.
   4.1. Kontrola stavu Doc_a.

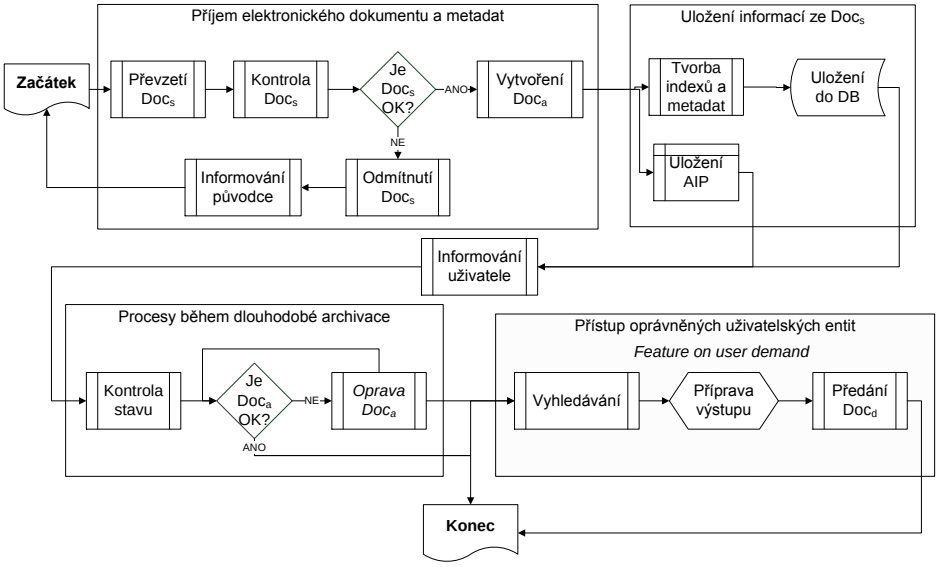
   Periodické ověřování integrity a čitelnosti archivovaných objektů a bezpečnostních mechanismů.

   4.2. Oprava zabezpečení.

   Aktualizace metadat a indexů, aktualizace zabezpečení dokumentů.
5. Přístup uživatelských entit k archivovaným objektům.
   5.1. Vyhledávání ED.
   5.2. Příprava výstupu a předání.

   Konverze na Doc_d. Uživatel si může zvolit, v jaké formě chce data obdržet. Obvykle se volí metoda extrakce a separace metadat a dokumentu.
6. *Likvidace (fyzické vymazání dokumentu).*

   Proces fyzického vymazání v tuto chvíli neaktivní. V LTA neexistuje krok jako je fyzické vymazání záznamu (objektu). V LTA se dokumenty archivují tzv. „navždy“. Dokumentu je při požadavku na smazání pouze změněn příznak na „DELETED“ ([1],[69]) Fyzické vymazání tak přestavuje pouze příznak s binární hodnotou. Proto není na následujícím procesním workflow proces skartace dokumentu.



Obrázek 4: Procesní workflow a životní cyklus dokumentu v archivu.

Zdroj: [autor]

Představuje podmnožinu z obrázku 3

Z hlediska časových operaci lze vyjádřit životní cyklus všech dokumentů v archivu (nebudeme tedy uvažovat časy t_0 který nám reprezentují čas před vlastní archivací). Celková „trvanlivost“ objektů v systému je dána součtem dílčích životních cyklů v jednotlivých modulech. Dá se tedy vyjádřit následujícím paradigmatem:

$$t \forall Doc_{(TA)} = \sum_{i=1}^{j} t_{si} + \sum_{i=1}^{j} (t_{ai} + t_{ri}) + \sum_{i=1}^{j} t_{di} + \sum_{i=1}^{j} t_{Di}$$

2-1: Celková doba archivace

Kde: $t \forall Doc_{(TA)}$= celkový čas pro všechny dokumenty (objekty) v systému TAA [t]

t_{si} = i-tý čas strávený v procesu příjmu [t]

t_{ai} = i-tý čas strávený v procesu úschovy [t]

t_{ri} = i-tý čas strávený v procesu kontroly a případné opravy [t]

t_{di} = i-tý čas strávený v procesu přípravy výstupu [t]

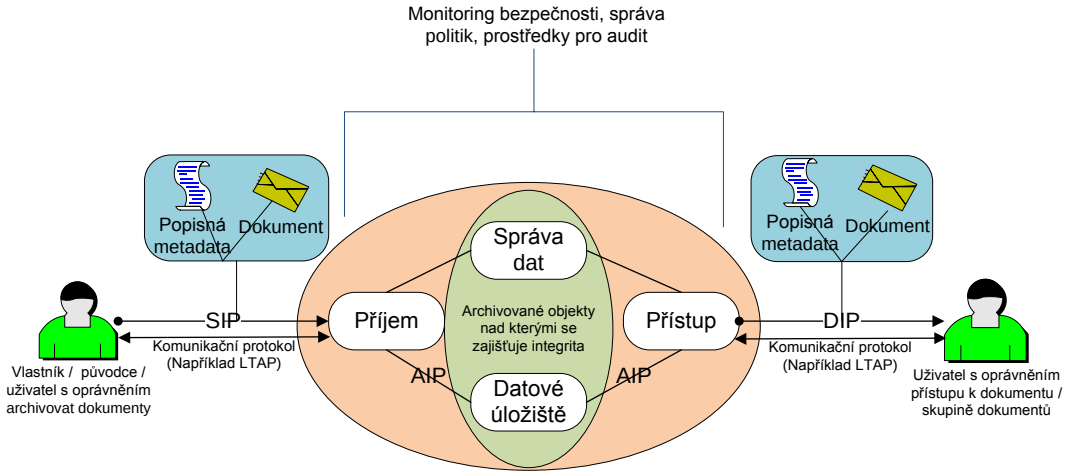
t_{Di} = i-tý čas strávený v procesu skartace/vyřazení [t]

j = počet dokumentů [ks]

4.2 Konceptuální model archivu

Pro potřeby této práce je za referenční model považován OAIS, jehož schéma a workflow jednotlivých objektů je na níže uvedeném obrázku. Obrázek 5 představuje

tzv. funkční model archivu. Ten pouze stanovuje rámec toho, jak by měl být každý archiv designován, nikoliv jeho logický model a všechny funkce. Na každém konkrétním projektu pak spočine, jak tento rámec uchopí a použije. Tento fakt byl prokázán například při tvorbě digitálního archivu pro Amnesty International [76].



Obrázek 5: Referenční konceptuální model OAIS archivu **Zdroj: [76], [autor]**

Výše uvedené schéma představuje systémovou úroveň abstrakce konkrétního modelového řešení. Model zohledňuje stanovený životní cyklus ED v rámci systému LTA tak jak naznačuje kapitola 4.1.2 a jsou vidět jednotlivé moduly archivu a jejich vzájemné interakce. Dle OAIS archiv musí obsahovat další moduly, jako jsou logování, vyhledávání, kontrola formátů dokumentů a další. Ty však nejsou pro tuto práci podstatné.

Každý archiv musí obsahovat archivní úložiště, které je zodpovědné za exaktní, v průběhu času po celou dobu uložení, neměnné bitové reprezentace objektů. Protože tato část je přímo závislá na použitém hardwaru, a ten není předmětem této práce, nebude také v této práci dále rozvíjena.

Komunikační protokol, pomocí kterého se archiv dorozumívá s vnějším světem a reprezentuje archivované objekty navenek. Příkladem takového protokolu postaveného na bázi standardu XML je například LTAP [11]. Blažič [43] deklaruje, že se jedná o dorozumívací nástroj postavený na bázi XML, který zastává nosný

komunikační a interagující prvek mezi uživatelem a archivem. Na straně klienta jsou dvě možnosti využívání LTA služeb a to prostřednictvím tlustého nebo tenkého [69] klienta. LTAP představuje hlavní komunikační protokol zabezpečující obousměrný přenos dokumentu (objektu) z/do archivu. Zajišťuje interakci s uživateli a s dalšími systémy dle definovaných dotazů, odpovědí a potvrzování. Zajišťuje také integritu a autentičnost během přenosu.

Tabulka 4: Zásadní komponenty archivu dle OAIS

Komponenta (modul)	Popis
Metadata	Uložení metadat v tomto modulu je v zásadě dvojí (viz kapitola pojednávajících o požadavcích na metadata). Dostupnost tohoto modulu je pro celý systém kritická, protože poskytuje atributy pro modul vyhledávání. Hlavně ale slouží pro vytváření, čtení kontrolu metadatových struktur. Měl by také obsahovat podrobný popis metadatových schémat. Metadata musí být schopna rozšiřování (i případné redukce) v průběhu času.
Příjem do archivu	Modul příjmu do archivu po provedení nezbytných kontrol vytvoří objekt, který je dále zpracován (kontrola + validace). Validace spočívá především v kontrole formátu, zabezpečení či dostupných metadat. Pokud je validace úspěšná, je balík předán správě dokumentů, kde je vytvořen AIP (LTA objekt) a ten předán správě úložiště.
Datové úložiště	Zabezpečuje, že archivovaný obsah je uložen kompletní a čitelný po požadovanou dobu. Zajišťuje průběžnou kontrolu obsahu („*error-checking*“).
Modul zabezpečení dokumentů	Pro tuto práci zásadní modul. V části věnované řešení vědeckého úkolu bude řešen.
Správa archivovaných dat	Slouží pro identifikaci a ověřování formátů na při předávání. Funkce monitoring a kontrola uložených dokumentů. Periodicky ověřuje integritu svěřených dokumentů a jejich metadat, označuje ty, u kterých je potřeba provést nějakou akci a upozorňuje na ni.
Systémová správa archivu	Zabezpečuje správu celého archivu na administrátorské úrovni. Poskytuje monitoring funkčnosti archivu, správu uživatelů a podporu auditu. Úzce souvisí s modulem logování.

Zdroj: [76], [autor]

Posledním bodem je certifikace samotného archivu. Aby byl archiv certifikován jako důvěryhodný (tj. dle norem ISO 16363, TRAC atd.), musí mít především perfektní dokumentaci. Snahou národních archivů je, aby každá země měla svůj digitální archiv certifikovaný dle těchto norem.

4.2.1 Informační model a klasifikace objektů pro zajištění integrity

V rámci workflow je zásadním krokem přijetí ED do archivu. K tomuto kroku je nutné definovat systém objektů v archivu, aby bylo zřejmé, nad jakými elementy bude integrita zajišťována. Vše opět začíná u referenčního modelu.

OAIS definuje tři základní typy archivačních balíků (tzv. „Information Package“) tj. binárních spojení archivovaných ED a jejich metadat. U metadat se jedná o dva druhy a to popisná a systémová, též někdy nazývaná jako PDI. Tento řetězec může být pak označen časovým razítkem (kapitola 4.3.3) dle platných standardů (implikuje přidání absolutní časové absolutní informace).

- SIP. Představuje ED a metadata ve stavu předávání do archivu. Předán je spolu s popisnými metadaty, resp. s PDI[64].
- AIP. Označuje formu (balíček) dokumentu a příslušných metadat tak jak je uložen v archivu po provedení všech předcházejících nezbytných úkonů nutných pro ověření, kontrolu a potvrzení o přijetí. Skládá se z jednoho nebo více SIP. OAIS standard nespecifikuje v jakém formátu nebo jakým způsobem mají být jednotlivé složky balíčku být propojeny.
- DIP. Označuje výstupní balík. Definuje formu, ve které je poskytnut (předán) uživateli (majícím příslušná oprávnění), který požaduje jeho vydání (stažení, prohlédnutí).

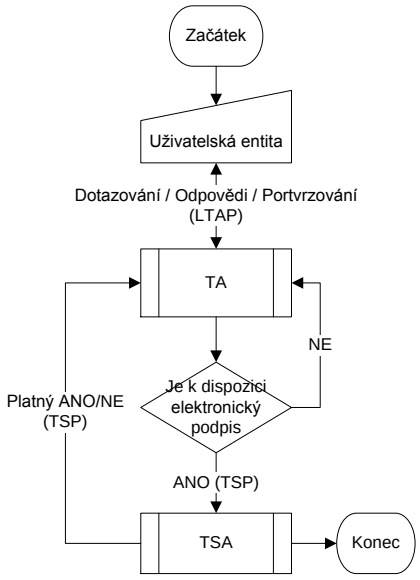
[64] V něm obsažené informace lze kategorizovat jako *průvodní* (historie dokumentu, kdo vytvořil, kdy vytvořil, když změnil (pokud vůbec atd.), *referenční* (obsahuje identifikátor dokumentu (pokud není je přidělen) a odkazy na další dokumenty), *bezpečností* (autenticita, integrita dokumentu – hash, elektronický podpis, vodoznak atd.).

4.3 Dlouhodobá archivace a bezpečnost – výchozí předpoklady

Na základě stavu poznání v předmětné oblasti lze předpokládat, že objem archivovaných dokumentů (počitatelných objektů) bude časem narůstat. V rámci prostudovaných materiálů plynoucích z rešeršní části práce, lze rovněž předpokládat, že objem růstu objektů v archivu bude mít minimálně lineární trend. To se odrazí ve způsobu zabezpečení, které by mělo respektovat univerzální aplikovatelnost bez ohledu na počet objektů, s nimiž bude archiv pracovat. Základní poskytované služby LTA musí splňovat následující:

- *Přijímat datové objekty či skupiny datových objektů (dávka) pro jejich dlouhodobé uchování.*
- *Vytvářet, ukládat, používat a starat se o průkazní záznamy pro převzaté datové objekty.*
- *Sbírat a ukládat další data nezbytná pro ověření průkazních záznamů.*
- *Poskytovat archivované objekty obsahující jak samotné dokumenty, tak metadata či průkazní záznamy (pokud existují).*
- Poskytovat objekty i v případě, že se technologie pro ukládání či technologie pro zpracování změnily během archivního cyklu daného objektu.
- Prokázat, že daný objekt existoval v určitém čase v minulosti a to v případech, kdy uživatelská entita není schopna interpretovat průkazní záznamy.
- *Poskytovat další služby dle politiky důvěryhodného archivu.*

Nad rámec samotné služby je nutné zahrnout také ověřování elektronických podpisů, které mohou být přímo součástí archivovaných dokumentů (tj. ne součástí systémových metadat). To lze řešit například implementací ověřovacích služeb založených na PKI [26].



Obrázek 6: Vývojový diagram způsobu ověřování časového razítka Zdroj: [autor]

4.3.1 Příčiny ztrát průkazních hodnot u archivovaných objektů

V kapitole 2.3 bylo stanoveno, že LTA se rozumí uchovávání dokumentů na deset a více let. Pro potřeby této práce a možnosti co nejširší aplikovatelnosti navržených řešení je LTA považována za ukládání dat bez časového omezení. Průkaznost existence dokumentů v relativním[65] čase v minulosti a jejich nezměněná bitová struktura od tohoto momentu je jedním z vytyčených úkolů.

Nový přístup k zabránění průkazních ztrát musí zajistit alespoň stejné základní bezpečnostní aspekty jako elektronický podpis, které jsou:

- Kdo dokument podepsal (podpisová entita).
- Kdy byl dokument podepsán (čas podpisu).
- Podepsaný dokument nebyl změněn od času a data podepsání (integrita dat).
- Umožnit poskytnutí důkazů pro skupinu objektů.

[65] Relativní časovou autentizací se v kontextu této práce vyjadřuje schopnost posoudit, které razítko (potvrzení existence dokumentu v čase doručení TSA) bylo vydáno jako první pro každý pár těchto časových razítek.

Navíc je nutné jako omezující podmínky stanovující rámec bádání této práce k výše uvedeným bodům přidat následující

- Příznak o odstranění nesmí žádným způsobem ovlivnit průkazní hodnotu objektu ani souvisejících archivovaných objektů.
- Všechny v čase použité hashovací algoritmy musí být jednoznačně identifikovány tak, aby bylo možné jejich jednorázové ověření.
- Měl by obsahovat TS[66] pro možnost ověření existence objektu v absolutním čase[67].

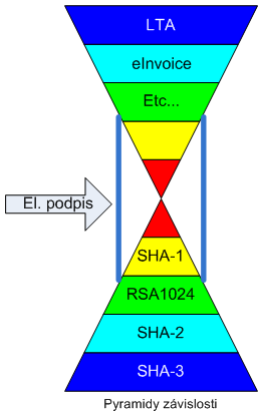
4.3.2 Dlouhodobý elektronický podpis

Po vymezení oblasti systémové, je nutné vymezit oblast bezpečnostní. Její tradiční podoba závisí výhradně na symetrických a asymetrických kryptografických technikách, jejich aplikaci (např. PKI) a kombinacích. Sylvester [27] již v roce 2003 upozorňuje, že elektronický podpis, tak jak si ho společnost 21. století buduje, je základním stavebním kamenem, na kterém stojí (a samozřejmě padá) řada technologií i služeb. Mezi nimi také LTA a to z hlediska pyramidového vyjádření (Obrázek 7) postaveno nejvýše. Pokud by tedy došlo někdy ke zpochybnění bezpečnosti digitálního podpisu (respektive použitých kryptografických funkcí) v průběhu času, všechny tyto, na něm stavící služby, by rázem byly nedůvěryhodné (Obrázek 8). Tuto skutečnost také podporuje diskuse, která vznikla na toto téma v průběhu vědeckého semináře Den doktorandů FIS v roce 2012 [70]. Zazněla při ní věcná poznámka, že ne všechny bezpečnostní mechanismy dlouhodobého charakteru lze stavět pouze na elektronickém podpisu a že by měly být hledány alternativní cesty a způsoby k jejich zajištění. A jednou z nich je právě důkaz o existenci záznamu (objektu), kterým se zabývá tato práce. Daleko zásadnějším argumentem je vyjádření Alexe Stamose [61], že díky pokrokům v matematice budou do čtyř až pěti let vyřešena problematika diskrétních logaritmů, které jsou

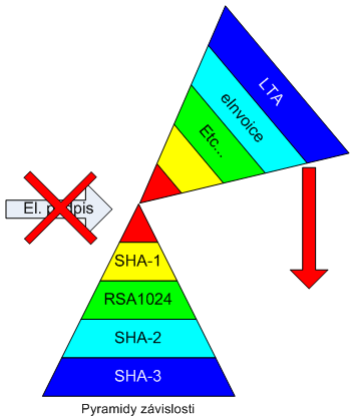
[66] Time Stamp = časové razítko.

[67] Absolutní časová informace není nutnou součástí každého objektu. To jak často (z hlediska posloupnosti archivovaných objektů) bude absolutní časová informace přidána jakou součást nepřepisovatelných metadat by mělo být obsaženo v politice archivu.

podstatou šifry RSA a tím pádem také Diffie-Hellman výměně klíčů, na které staví novodobý elektronický podpis. Jako adekvátní náhradu uvádí šifrování nad eliptickými křivkami, kterou má ovšem patentovanou společnost BlackBerry.



Obrázek 7: Pyramida závislostí Zdroj: [27]



Obrázek 8: Nedůvěryhodný el. podpis Zdroj: [27]

K zajištění bezpečnosti archivovaných objektů je více než vhodné používat tzv. dlouhodobý elektronický podpis a ne standardní, krátkodobý (s platností certifikátu typicky na jeden až tři roky). Ten představuje další způsob zajištění integrity archivovaných objektů v čase (na obrázku 7 je reprezentován „podpěrami“[68]). Zajišťuje fixaci absolutního času na objekt, ale podléhá přirozeně entropii klesající, nebo již dosažené neexistující průkazní hodnoty. Příčiny, kterým je třeba předejít, jsou následující:

- Nedostupná informace o zneplatnění.
- Nedostupný příslušný certifikát pro verifikaci.
- Certifikát přidružený s příslušným podpisem vypršel nebo byl odvolán.
- *Vývoj v oblasti kryptografických technik a výpočetní síly byl rychlejší, než se předpokládalo a lze spočítat duplicity*, či utajované soukromé klíče.

Nejjednodušší možností jak výše uvedeným skutečnostem předejít, je definovat protokol tak, aby spolu s objektem uchovával i příslušné průkazní záznamy

[68] Obecně lze naznačenými podpěrami charakterizovat jakoukoliv alternativní metodu k elektronickému podpisu, tj. i relativní časovou autentizaci.

(certifikáty, CLR, TS atd.), periodicky je obnovoval a přidával další, v průběhu času nezbytné informace (založené na silnějším (tj. bezpečnějším) algoritmu). Jinými slovy, dlouhodobý elektronický podpis narozdíl od krátkodobého, aplikuje metodu „uchování všech prostředků nutných k prokázání platnosti“ zaručující nutnost podepsání pouze jedenkrát a to při archivaci. Jsou však metody, jak zajistit, aby také archivovaný certifikát po vypršení jeho platnosti a/nebo potom, co se stanou použité kryptografické metody zastaralé, bylo stále možné ověřit pravost archivovaného dokument. Jedná se o metody pro uchování certifikátů a jejich revokačních listů (CRL) jako je například SCVP [74].

Je nutné podotknout, že se jedná o metodu zajištění integrity na straně uživatele, a jak bylo uvedeno, ta nebude v této práci zahrnuta. Ze skutečností uvedených v této kapitole i přesto plyne, že stárnutí kryptografických technik je nevyhnutelné a tím pádem je na místě se domnívat, že všechny na nich založené hashovací algoritmy a funkce se stanou zastaralými a tudíž nedůvěryhodnými.

4.3.3 Časové razítko

Platnost elektronického podpisu lze prodloužit časovým razítkem [53], [36]. To reprezentuje datová zpráva, která potvrzuje existenci dokumentu v čase.

1. Slouží jako důkaz, že datový objekt, ke kterému je připojeno, existoval bezprostředně před časovým údajem (v čase *t-1*), uloženým v tomto časovém razítku.
2. Zajišťuje přiřazení aktuálního časového údaje k existujícím datovým objektům, informacím, souborům nebo událostem.

TS obsahuje datum a čas vydání, číslo časového razítka, identifikaci poskytovatele certifikačních služeb, která jej vydala a hash ED, ke kterému je razítko vydáno, a elektronickou značku poskytovatele [53]. Při velkém počtu objektů (resp. požadavků) na „orazítkování“, je pravděpodobné (viz hypotézy), že tento proces může zvýši nároky na dostupné zdroje a výpočetní výkon. Pro takovéto účely je doporučeno používat vyhrazeného prostředku, který bude dedikovaně zajišťovat

službu časových razítek („*TSA*“)[69]. Tato služba, i přesto, že je většinou externí, se zahrnuje do skupiny zabezpečení na straně archivu. TSA obdrží v čase *t* žádost o časové razítko. To je důkazem, že dokument *x* (resp. jeho hash $H_{(x)}$) byl doručen TSA v čase *t*. Časové razítko samo sobě, ale nepředstavuje dostatečnou ochranu [4].

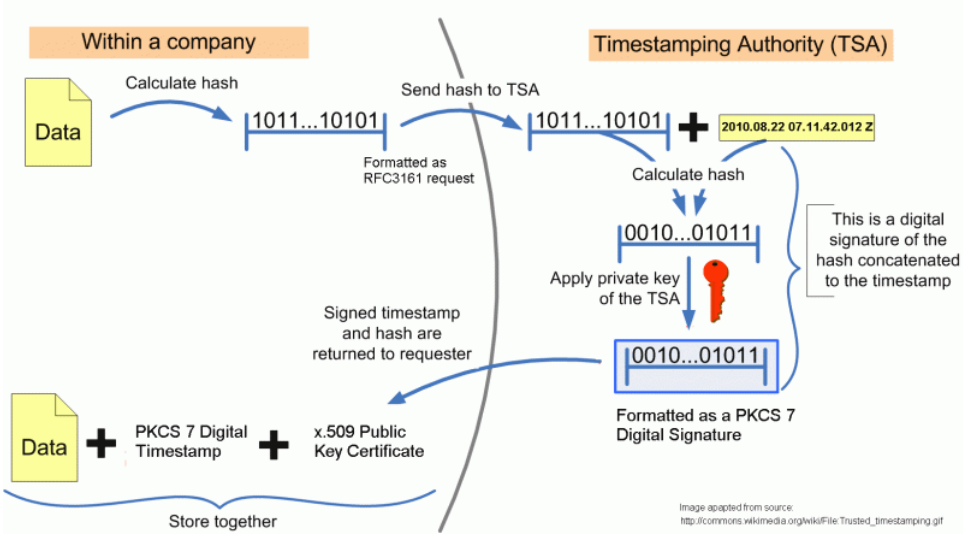
4.3.3.1 Přerazítkování

LTA může zajišťovat systém nepopiratelnosti v kontextu střednědobého časového horizontu prostřednictvím periodického vytváření časových razítek (tzv. systém přerazítkovávání [26], (Obrázek 9), kdy po vypršení platnosti jednoho časového razítka, je dokument „přerazítkován“ razítkem nově platným.

Obrázek 9: Přerazítkování Zdroj: [26]

Tento princip jde použít i při metodách, které nepoužívají pro zajištění původu a integrity primárně elektronický podpis (protože to je zajištění integrity na straně uživatele), ale pouze hashovací funkci z časového razítka, která je jeho součástí. Přerazítkováním tak vlastně jde aplikovat ve chvíli kdy je nalezena kolize výstupní hodnoty hashe (resp. v čase *t-1*, protože v čase *t* není již původní funkce považována za důvěryhodnou). Tento princip volí na straně tlustého klienta například [29] a jak již bylo zmíněno, ve většině případů se pro tento úkon využívá služeb externí TSA (Obrázek 10).

[69] http://www.thales-esecurity.com/EN/Products/Time%20Stamping/Time%20Stamp%20Server.aspx



Obrázek 10: Služba externí TSA Zdroj: [36]

Pro tuto disertační práci v rámci omezujících podmínek je deklarováno, že v LTA není systém přerazítkování uvažován, Jinak řečeno systém LTA v této práci neuvažuje zabezpečení objektů dle PKI[70]. Díky tomuto předpokladu, je jedinou možností vycházet z principu, kterou definoval de-facto standard ERS (RFC 4998), a poté převedl do „řeči“ standardu XML RFC XML ERS [9] (kapitola 4.3.6). Pro úplnost je třeba uvést, že potřebu neustálého „přerazítkování“ prosazuje například „software602“, který konkrétně deklaruje:

„k takto chráněnému dokumentu budou pravidelně připojována časová razítka tak, aby nikdy nebyla přerušena jeho digitální kontinuita a aby byl vždy použit nejsilnější dostupný šifrovací algoritmus“ [78].

4.3.4 Obecná funkce zabezpečení objektů v archivu

Principem zabezpečení objektů za pomoci hash funkcí je:

1. Autentickým způsobem sdružit čas a data.
2. Příslušnou autentickou informaci ověřit v libovolném časovém okamžiku v budoucnosti.

[70] Na straně archivu – na vstupu do archivu být může, ale vůči metodě popisované v této práci je irelevantní. Tj. ES, TS apod. mohou vstupovat do systému relativní časové autentizace.

Z toho vyplývá, že funkci časových zabezpečovacích prvků lze obecně vyjádřit následně:

$$T_n = t(Doc_n)$$

Výše uvedený vztah vyjadřuje, že aplikace funkce zabezpečující časové orazítkování (t) objektu (Doc_n) vytvoří časové razítko T_n.

Význam procesu ověřování platnosti Doc_n lze zapsat jako logickou funkci:

$$If\ A(T_n)\ then\ t(Doc_n)$$

Kde *t* představuje konkrétní časový okamžik archivace zdrojového dokumentu, resp. archivního objektu.

4.3.5 Existence záznamu jako důkaz

Jedná se o nezpochybnitelný záznam, že objekt existoval (byl archivován) v čase. Integrita samotného dokumentu je pak zajištěna hashem vytvořeným (v daném čase) pomocí nejbezpečnějšího hashovacího algoritmu (více v kapitole 4.3.7). U obecné hashovací funkce *H* je třeba si uvědomit, že pokud platí následující předpoklad:

$$H_n = h_a(Doc_n)$$

pak musí platit tři zásadní vlastnosti, které předurčují její vhodnost pro LTA:

- Je jednocestná. To představuje:

$$\mathrm{h_a}^{-1}(H_n) \neq Doc_n$$ 3-2: Jednocestnost

- Hodnota na výstupu je funkcí všech bitů na vstupu. Dá se zapsat následovně:

$$H_n = h_x^1(Doc_n)$$ 3-3: Hodnota na výstupu

- Je bezkolizní[71]. Z toho plyne:

[71] V daném čase, při dané velikost výstupu (snižuje pravděpodobnost výskytu kolize) a dané výpočetní náročnosti použitého matematického algoritmu.

$$h_a(Doc_n) \neq h_a(Doc_m)$$ 3-4: Bezkoliznost

- A má vždy konstantní (bitovou) velikost výstupu.

Kde u výše uvedeného:

H_n = výsledná hodnota hashe n-tého vstupu.

h_a = a-tá použitá hashovací funkce.

Doc_n = n-tý objekt na vstupu (n-tý bitový řetězec na vstupu).

h_a^{-1} = inverzní funkce k funkci h_a.

To samotné ovšem nestačí. Je třeba také zajistit integritu celého systému archivace (archivu), který je představován časovou posloupností archivovaných objektů:

$$\exists\, H_x = h_n\left(\sum_{i=1}^{k} h_{n(i)}\left(\sum_{j=1}^{p} h_n\{Doc_{s(j)}\}\right)\right)$$ 3-5: Integrita systému archivace

Kde: H_x = Hodnota hashe při použití v x-tém kole.

h_n = n-tá hashovací funkce.

$h_{n(i)}$ = n-tá hashovací funkce u i-tého mezilehlého hashe.

$Doc_{s(c)}$ = j-tý přijatý objekt do TA.

Výše uvedený vztah představuje robustní nástroj pro zajištění základního stavebního prvku každého archivu, kterým je nepopiratelnost svěřených objektů. Je nezbytné nalézt takový prostředek, který bude schopen takto stanovený výzkumný koncept naplnit.

4.3.6 Evidence Record Syntax

ERS [32] je pro tuto práci podstatný v tom, že jako první navrhuje jak řešit otázku integrity dat nezávisle na absolutním čase archivace. ERS jako první v LTA zavádí systém zřetězených (provázaných) hashů (tzv. linkage (linking) hashes). ERS zavádí vhodná doporučení jak vytvářet stromy (resp. grafy) z hashů jednotlivých objektů, jejich ověřování a zejména jak provádět jejich následnou redukci (pokud je

vyžadována). Provázané hashe Dostálek [26] je popisuje jako: „vytváření závislé a ověřitelné datové struktury“.

ERS je svojí aplikací provázaných hashů a stromových schémat do oblasti LTA jedním z hlavních podkladů pro tuto práci. Původní ERS bylo v roce 2007 nahrazeno (doplněno) o XML verzi (RFC 6283), tj. verzi, která se opírá o silnou základnu značkovacího jazyka XML a jeho možností. Poslední verzí draftu je době psaní této práce verze 12 z července 2011 [9]. Na vytváření tohoto standardu se podílela společnost SETCCE[72] při výzkumném Jože Štefan Institute v Lublani, s níž autor během svého zahraničního působení spolupracoval (viz kapitola 2.1 „Zdůvodnění výběru předmětného tématu a motivace“). U (XLM) ERS je třeba si uvědomit to, že se nejedná o ochranný mechanismus samotných objektů. To musí být zajištěno v archivu jinými prostředky. Nicméně ERS je nástroj, který zajistí, aby bylo možné ověření integrity objektu (a následně potvrzena či vyvrácena jeho autenticita).

4.3.7 Výběr vhodného hashovacího algoritmu

Všechny existujícího schémata provázaných hashů předpokládají jako první krok výběr vhodného (v době výběru nejvíce bezpečného, tzv. collision-free či collision-resistant) hashovacího algoritmu. Hash lze v obecné rovině definovat jako:

- Nechť existuje polynom *P* a pro každé číslo *n* existuje v čase *T(n,|x|,|y|)* funkce $h_n(x,y)$ pro všechny $x \in Xn\ a\ y \in Yn$. V daném řase *T* (představujícím dobu bezpečnosti použité hashovací funkce) neexistuje žádný další polynom P’ pro který by platilo pro dané číslo n pro pár $(x,y) \in X_n\ x\ Y_{n,}a\ y' \in Y_n$ najít takové $x' \in X_n$ aby platilo $h_n(x,y) = h_n(x',y')$. [18]

U algoritmu je zásadní nejen délka hashe (tzv. message digest size), ale především maximální bitová délka zprávy, ze které je možné hash vytvořit. Ta je dána následujícím vztahem:

[72] http://www.setcce.si/

$$p = 2^{r-1} \leq p < 2^r$$ 3-6: Maximální bitová délka zprávy

Kde:

p = bitová délka příslušné zprávy (dokumentu).

r = počet bitů.

Funkce, resp. použitý algoritmus vedoucí k vytvoření výstupu, bude z hlediska informatické zákonitosti a dostupného výpočetního výkonu bezpečný maximálně několik let. Pro legislativy jednotlivých států v EU (včetně ČR) je doporučujícím dokumentem v této oblasti ETSI TS 102-176-1 v aktuální verzi 2.1.1 [28]. Ten mj. v sekci „*Time period resistance of hash functions and keys*“ stanovuje doporučené období, po které jsou aktuálně používané hashovací funkce (v době psaní této práce) ještě bezpečné (viz Tabulka 5). S ohledem na LTA je nutné podotknout, že se jedná o spekulativní rovinu vycházející pouze z dostupných údajů nárůstu výpočetní kapacity a kryptografické bezpečnosti. Po uplynutí doporučené doby předpokládané bezpečnosti se obvykle podaří nalézt duplicitu a algoritmus se tak stává nedůvěryhodným. Tento fakt jde zcela proti logice LTA, kde třeba důvěryhodnost zachovat po celou dobu archivace ED. Jinými slovy v LTA je téměř jisté, že hashovací algoritmus bude třeba v době existence dlouhodobého archivu měnit.

Výše uvedený odstavec vyjadřuje jeden ze stavebních kamenů této práce, který představuje včasné odhalení duplicity a zajištění z toho vyplývající bezpečnosti dlouhodobě archivovaného objektu, zapříčiněné nevyhnutelným a postupným stárnutím kryptografických algoritmů.

Tabulka 5: Bezpečnosti hashovacích funkcí dle ETSI

Název hashovací funkce	1 rok	3 roky	6 let	10 let
sha1 (nedoporučuje se)	Nepoužitelné	Nepoužitelné	Nepoužitelné	Nepoužitelné
*sha224**	Použitelné	Použitelné	Použitelné	Nelze určit
sha256	Použitelné	Použitelné	Použitelné	Nelze určit
*sha384***	Použitelné	Použitelné	Použitelné	Použitelné
sha512	Použitelné	Použitelné	Použitelné	Použitelné
Whirlpool	Použitelné	Použitelné	Použitelné	Použitelné

Zdroj: [28]

Obecně jsou všechny hashovací funkce rodiny SHA-2[73] v době psaní této práce považovány za bezpečné (nebyla nalezena kolize). Jak je ovšem vidět z Tabulka 5, funkce SHA-224 a SHA-256 jsou označeny v dlouhodobém pohledu jako „unknown“, čili od 1. 1. 2018 je již nebude možné považovat za bezpečné. SHA-224, resp. SHA-386 mají stejnou specifikaci jako SHA-256, resp. SHA-512. Liší se pouze v počátečních hodnotách a bitové délce výsledného hashe (SHA-224, resp. SHA-384) kdy výsledek zkrátí na příslušný počet bitů. Proto se doporučuje (pokud je to možné používat) vždy funkce s výsledkem mocniny 2^x. Parametry rodiny funkcí SHA-2 viz následující tabulka.

Tabulka 6: SHA-2 parametry

	SHA - 224/256	**SHA – 384/512**
Hash size	224/256b	384/512b
– string	56/64B	96/128B
Max input	2^{64}–1b	2^{128}–1b

Zdroj: [28]

Hashovací funkce, která má vysoký bezpečnostní potenciál a vychází ze symetrické blokové šifry AES je známá pod označením WHIRPOOL [23]. Její předurčení k využití v LTA je dáno v prvé řadě možnou dvojnásobnou bitovou délkou oproti zprávě (dokumentu) určené k hashovaní, než nejpokročilejší funkce rodiny SHA-2 a to SHA-512. Jak bude ukázáno v experimentální části práce, počet bitů nemá na rychlost vytvoření hashe u rodiny SHA-2 velký vliv.

Hashovací funkcí o níž se doporučení ETSI nezmiňuje je funkce Tiger[74]. Jedná se o starší hashovací funkci, na níž, na rozdíl od MD-5, nebyl do této doby proveden žádný známý a úspěšný útok. Funkce vyniká nejen svojí rychlostí (viz kapitola vědecký experiment), ale též je známá jejím využitím pro výpočet stromů (více

[73] http://networking.answers.com/security/the-evolution-of-the-nist-secure-hash-algorithm-from-sha-1-to-sha-3
[74] http://link.springer.com/chapter/10.1007%2F3-540-60865-6_46

v kapitole pokročilým schématům řetězení) ve vybraných P2P aplikacích jako jsou DC++[75] či Gnutella[76].

4.3.8 SHA-3

Prvotním impulsem pro zahájení aktivit v oblasti tvorby nové hashovací funkce byly kolem rok 2004 velmi intenzivní útoky na existující hashovací funkce (SHA-1, SHA-2). Bylo deklarováno, že nově vzniklá funkce se musí od již existujících a používaných zásadně lišit a to především z bezpečnostních důvodů. SHA-3 je první hashovací funkce, jejíž výsledná podoba vzejde z výsledku třetího kola veřejné soutěže vypsané NIST[77]. První kolo bylo vyhlášeno v roce 2. 10. 2007, druhé v roce 9. 12. 2010. Z něho vzešlo 5 finalistů, kteří se utkali ve třetím kole o konečnou podobu SHA-3 [39]. NIST 2. 10 2012 vybrala vítěze, jímž byla funkce s názvem „KECCAK“ [15] vycházející z hashovacích funkce RadioGatún[78]. Funkce je při HW implementaci velmi rychlá (dokumentace [8] deklaruje 12,5 cyklu na bajt), což byla také jedna z výhod, která ji pomohla k vítězství[79] (např. konkurenční BMW[80] dává max 7,33 cyklu/bajt na shodném HW). V říjnu 2013 byla stanovena definitivní podoba, avšak pouze ve verzi draftu. Finální specifikace (standard) je očekávána v roce 2014. Hashovací funkce SHA-3 nemá být chápána jako náhrada (zranitelnost nebyla prokázána). Díky téměř pětiletému čekání se na svět dostala funkce SHA-4 a přeskočila tak vlastně SHA-3. Naneštěstí se jedná pouze o upravenou funkci SHA-2 [41].

4.3.9 Obecná doporučení pro zajištění vyšší bezpečnosti standardní funkcí

1. Kombinací funkcí z rozdílných rodů.
   Například kombinací funkcí SHA-2 a TIGER.
2. Oříznutím výsledného hashe na nižší bitovou délku než předepisuje specifikace.

[75] http://cs.wikipedia.org/wiki/DC++
[76] http://cs.wikipedia.org/wiki/Gnutella
[77] National Institute of Standard and Technology
[78] http://csrc.nist.gov/groups/ST/hash/sha-3/documents/Keccak-slides-at-NIST.pdf
[79] http://keccak.noekeon.org/
[80] Blue Midnight Wish

Například: MD5 ← SHA1 ← SHA2 ← Whirlpool (32 ← 40 ← 56/64/96/128 ← 128).

3. Kombinací výše uvedených metod.

4.4 Systémový audit

Archiv musí být připraven procházet pravidelným i nepravidelným obecným systémovým auditem, v rámci kterého se prověří shody stavu systému, postupů a správa archivu s relevantními předpisy a normami. Základním typem auditu je audit systémový, který je schopen v jakémkoliv časovém okamžiku prověřit výstupní hodnoty hashovacích funkcí. Jinými slovy prověří integritu jakéhokoliv archivovaného objektu za v jakémkoliv časovém okamžiku od doby archivace (v čase *t+1)*. Jedním z přístupů pro komplexní audit, je Audit Control Environment („*ACE*“) [62] nebo AbsolutProof [29]. Ten podporuje nezávislé auditorské nástroje (programy) třetích stran respektující platné standardy. Jinými slovy nástroje jako ACE, AbsoluteProof a další založené na standardu ERS, který je jedním ze stavebních kamenů této práce, mohou přispět k úspěšnosti komplexního auditu digitálního archivu. Na druhou stranu na schéma pro vytvoření důvěryhodného digitálního objektu (TDO) byla v únoru 2013 podána patentní přihláška [31], jež se velmi podobá způsobu zajištění integrity právě u již zmíněného ACE. Lze se tedy domnívat, že princip doporučovaný při auditu dlouhodobých digitálních archivů není původní.

„Best Practices“ jakým způsobem auditovat jakýkoliv archiv lze nalézt například v [3]. Z dokumentu je patrné, že se jedná o kontrolu integrity objektů, resp. v terminologii OAIS jednotlivých balíčků. Za nejvíce vypovídající se považují doporučení a metodiky DRAMBORA či TRAC pro provedení interního auditu dlouhodobých archivů.

5 Zabezpečení elektronického dokumentu během životního cyklu

Bezpečnost dokumentů, jejich metadat a dalších prvků zajišťujících nepopiratelnost jsou elementární požadavky na archiv kladené. Průkaznost existence ED (a k němu přidružených informací) v čase je nezbytnou vlastností, stejně jako zajištění jeho integrity po celou dobu úschovy. Na zajištění, že objekt nebyl od data a času přijetí do archivu modifikován jsou použity známé postupy [53] založené na bezpečném kryptografickém mechanismu důvěryhodném právě v relativním čase archivace dokumentu. Tyto předpoklady byly již vyjádřeny v předchozích kapitolách. Z dlouhodobého pohledu je hlavní schopnost prokázat, předtím než se použité kryptografické algoritmy stanou slabými (tudíž nedůvěryhodnými), že na nich založené ověřovací mechanismy existovaly a byly platné. U hashů se bude především jednat o výstupy hashovací funkce do doby, než bude nalezena jejich kolize (duplicita). To lze zajistit například implementací Data Structure for the Security Suitability of Cryptographic Algorithms („*DSSC*“) [47]. Je třeba na tomto místě zdůraznit, že sám Buldas tvrdí [21], že není nezbytně nutné, aby samotné schéma provázaných hashů bylo nutně postaveno na jednocestné hashovací funkci. Tj. že jednocestnost nemá vliv na bezpečnost tohoto řešení. Dokonce v [22] lze najít, že bezkoliznost hashovací funkce (nalezeni duplicity) nemá na bezpečnost systému provázaných hashů vliv.

Cílem této kapitoly, je analyzovat, následně omezit a definovat ty způsoby zabezpečení ED, které mají pro tuto práci zásadní význam.

5.1 Metody prokazování existence objektu v čase

Proces ověřování existence daného dokumentu (objektu) v čase se dělí na metody:

- Absolutní.

  Ověření přesného časového okamžiku, kdy byl dokument orazítkován, resp. archivován. Časové razítko („*TS*“) dle ANSI ASC X9.95 či [1].
- Relativní.

Je reprezentováno provázanou datovou strukturou dle ERS.

Ověření existence objektu Doc_n v čase *t* po dokumentu existujícím v čase *t-1* a před dokumentem existujícím v čase *t+1*. Jinými slovy pro ∀ Doc_n archivované v čase *t-1* ∃ ($Doc_{n(t-1)} < Doc_{n(t)} < Doc_{n(t+1)}$).

- Hybridní.

  Ověření kombinace obou předchozích metod, resp. doplnění relativního času v pravidelných intervalech, které stanoví politika archivu, o absolutní časový údaj – časové razítko, které zafixuje v čase řetězec relativních časových razítek.

5.2 Zajištění integrity archivovaných objektů

Pro zajištění exaktní bitové struktury všech objektů, je nutné aplikovat takový nástroj, který je na jedné straně co nejméně citlivý na vývoj v oblasti výpočetních a kryptografických technik a na straně druhé co nejvíce univerzální. Jinými slovy aplikovat takové řešení, které bude průběhu času stále důvěryhodné. V dnešní době se zmiňují v zásadě dva hlavní přístupy:

- Přerazítkovávání („Continuous Time Stamping“).

  O způsobech přerazítkování, jeho výhodách, nevýhodách a dalších aspektech, viz kapitola 4.3.3.1.
- *Systém provázaných hashů („Linking (Linkage) Hashes“).*

Oba výše uvedené způsoby mají společný cíl, kterým je zajištění autenticity a integrity všech dlouhodobě archivovaných dokumentů a příslušných metadat. To v čem se ale zásadně odlišují je způsob a metody vedoucí k nepopiratelnosti svěřených dokumentů. Obě metody je samozřejmě možné kombinovat. Obecně však již při aplikování metody provázaných hashů není přerazítkovávání nutné. Provázané hashe je možné aplikovat I na elektronicky podepsané a časově orazítkované archivované objekty. V následujících pasážích se tato práce zaměřuje pouze na druhý jmenovaný přístup.

Většina existujících schémat časových razítek (příp. hashů) jsou tzv. provázaná. Tj. v čase *t* jsou jednocestně závislá na časovém razítku vydaném v čase *t-1*. Z toho plyne zavedení systému relativního časové autentizace (RTA). Ta je zásadní pro celý systém provázaných hashů. Relativní časová autentizace pracuje jak pro jednotlivé objekty, tak pro kola („*rounds*“). Ty představují agregovanou frontu několika žádostí (hashů) TSA o časové razítko. To je pak vydáváno souhrnně pro všechny dokumenty (resp. jejich žádosti) doručených TSA během kola *r*. Systém front lze vyjádřit:

- Nechť existují takové žádosti, u kterých platí, že pro každou jejich dvojici *v* a *w*, kde $v < w$, existuje takové spojení těchto žádostí v_h a w_h, které obsahuje cestu mezi příslušnými vydanými časovými razítky S_v a S_w.

Každý prvek *v* (vstupní dokument) orientovaného grafu *G* odpovídá řetězci H_v. Stejně tak certifikát $C_{(v)}$ se skládá z takového (minimálního) množství dat na vstupu, aby bylo možné provést ověření, zdali se tento prvek nachází na cestě mezi *i* (řetězec odpovídající orazítkování posledního kola) a *j* (řetězec odpovídající razítku aktuálního kola).

Objekt, který byl archivován v relativním čase *n* a byl k němu vytvořen příslušný hash *H* za pomoci „*collision-resistant*“ („*collision-free*“ nebo „*collision-freedom*“) hashovací funkce *h* (platí tedy $H_n = h(Doc_n)$) a tento hash nebude v posloupnosti označen jako první (H_0) pak musí existovat hash H_{n-1}, z dokumentu Doc_{n-1} a musí existovat též hash H_{n+1}. Tento předpoklad musí platit pro všechny následné *n+m* kroky.

5.3 Rozbor existujících schémat provázaných hashů

Provázané hashe představují alternativní a z dlouhodobého hlediska praktičtější (a také levnější) metodu k již zmíněnému přerazítkovávání dle PKI [71]. Jednocestnost, absence tajného klíče a rychlost dělají z provázaných hashů daleko robustnější nástroj.

Zajišťují také tři základní stavební kameny, na kterých staví mj. elektronický podpis. Integrita je zajištěna sestavením shodné hodnoty kořenového (root) hashe. Čas je

zajištěn posloupností archivovaných objektů, resp. jejich hashů (jedná se tedy o relativní čas archivace, nikoliv absolutní[81]). Identita je zajištěna politikou archivu a lze ji vyvodit z pořadí postupného řetězení na jednotlivých úrovních hierarchické (stromové se složitostí vyhledávání log_n) nebo lineární (s asymptotickou složitostí $O_{(n)}$) struktury [5].

Tato disertační práce se soustředí primárně na pokročilá schémata provázaného hashování, která jsou vhodná pro zajištění integrity v digitální archivaci. Při předpokladu lineárního trendu nárůstu dokumentů v archivu jsou jako akceptovatelné (vzhledem k výsledné velikosti grafu) považovány dvě hlavní skupiny, které jsou rozebrány podrobněji v následujících podkapitolách.

5.3.1 Částečně uspořádaná (stromová) schémata

Jejich zásadní znaky oproti plně uspořádaným schématům jsou:

- Obecně menší složitost.
- Jsou vzájemně porovnatelná časová razítka z různých kol[82].

Typickými představiteli jsou schémata stromová („Merkle Tree“ nebo „Binary Tree“) a jejich společným znakem je nelineárnost hashovacího procesu [18]. Dále sdílí vlastnost vyžadující minimální počet žádostí („*requests*“) během kola rovnající se dvěma. Princip každého stromu je, že jako vstupní hodnoty (listy) jsou použity hashe, které jsou dále řetězeny a opakovaným použitím hashovacích funkce je vygenerován jediný kořenový („*root*“) hash. Obecně se výše uvedené dá vyjádřit následujícím vztahem:

$$h_{12} = h(x_1|x_2)$$

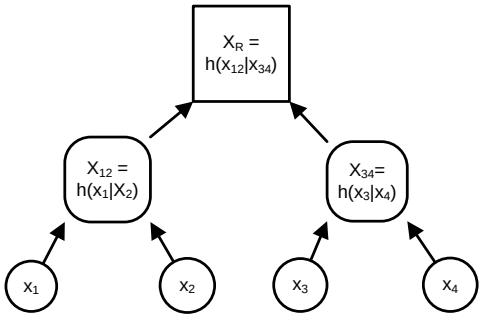
Je zásadní si uvědomit, že x_1 a x_2 tvoří na vstupu uspořádanou dvojici, tj. $h(x_1|x_2) \neq h(x_2|x_1)$.

[81] O sub-absolutní by se jednalo, pokud by jednotlivá kola byla odvislá od časového elementu a nikoliv od počtu archivovaných objektů.
[82] Pokud je v rámci jednoho kola orazítkováno více než *n* dokumentů kde n>1.

U stromového (stromového binárního) schématu (někdy též nazývaného jako „*height balanced*“) provázané hashování neprobíhá v časové posloupnosti, jak jsou postupně dokumenty archivovány, ale po jednotlivých kolech („rounds“) definovaných diskrétním časovým intervalem (např. hodina) nebo událostí (např. dosažení určitého počtu archivovaných dokumentů). Používá se také v případě, kdy je požadavek na archivaci n-tého dokumentu archivem obdržen ve stejnou chvíli jak *n+1* dokumentu. To se dá vyjádřit následující definicí:

- Nechť je kolo *r* pro příslušné časové razítko *Tr* kumulovaný hash všech dokumentů archivovaných v předchozím kole *r-1* a vytvořeným kumulativním hashem platným pro předchozí kolo T_{r-1}. Existuje strom S_r, jehož listy jsou tvořeny hashy H_{rn} vytvořenými z archivovaných dokumentů během kola *r*. Hash je vytvořen (v té době „*colission-free*“) hashovací funkcí *h* z *n-tého* dokumentu (*Doc*) archivovaného během kola *r*. Mezilehlé hashe MH_n pak nejsou nic jiného než použití stejné hashovací funkce *h* na výsledný hash levého *l* a pravého *p* listu. Tj: $MH_n = h(H_{rl}, H_{rp})$.

Stromová schémata jsou definována ve dvou modifikacích a to pod spojenými názvy jejich tvůrců Haber-Stornetta (oba pracují pro Surety Technologies a na základě tohoto schématu je postaven hlavní produkt, více v kapitole 3.2) a Benaloh de-Mare [6]. Obecný princip u stromového schématu se liší se pouze tím, že v případě Haber-Stornetta modelu, je výsledný hash z předchozího kola (T_{r-1}) podkladem pro vytvoření hashe z *r-tého* kola spolu se všemi dokumenty archivovanými v tomto kole. Z toho je jasné, že pro potřeby LTA, není tento model optimální, protože pro rekurzi je nutné projít znovu všechny listy stromu *r-ntých* kol. Více v kapitole věnující se binárnímu schématu (kapitola 5.3.2.3).



Obrázek 11: Stromové schéma Zdroj: [29], [autor]

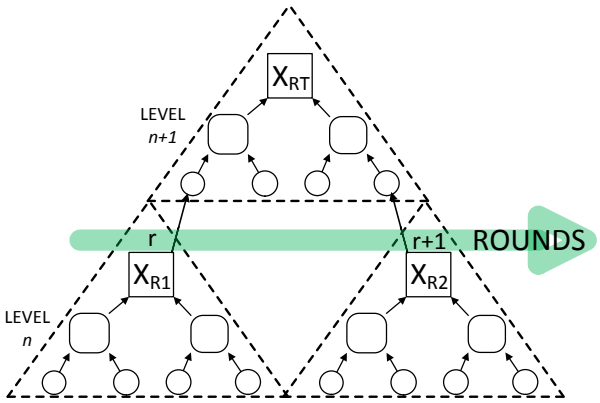
Se schématem výše souvisí také pojem kostra stromu. Jedná se o takovou část grafu, ze které je možné zrekonstruovat kořen z jakéhokoliv vstupního listu. To lze obecně vyjádřit jako $log_2(n)$ kde n je počet listů. Tj. pokud má například vlastník hashe x_2 (tj. dokumentu Doc_2) k dispozici hash x_1 a hash x_{34}, je schopen získat kořenovou hodnotu X_R. Tím lze dokázat, že hash x_2 (ale nikoliv dokument, resp. objekt Doc_2) byl použit, resp. podílel se, na původním výpočtu.

Z předchozího odstavce plyne, že pro to, aby si klient mohl (bereme ve vztahu k TSA) ověřit kořenovou hodnotu, musí mít k dispozici také mezilehlé hashe, které se podílely na jejím vzniku spolu s pořadím, v jakém docházelo k jejich řetězení. Ty jsou součástí potvrzení od TSA (časového razítka).

Pro potřeby auditu (kapitola 4.4) navrhuje [20] uschovávat mimo schémata množinu všech binárních řetězců (objektů), které byly během cyklu označeny časovým razítkem. Zásadní v tomto přístupu je fakt, že jakmile je uzavřen cyklus r a začne $r+1$, není již možné žádný bitový řetězec (posloupnost) do provázané struktury přidat.

5.3.1.1 Distribuovaný přístup

Každý graf uvedený v předchozí podkapitole (Obrázek 16), lze začlenit do distribuované hierarchické struktury, kde každý kořen X_{Rr} se stane listem dalšího samostatného „podgrafu“. Vznikne tím další ucelený systém, kde vrchol grafu na nižší úrovni (n) tvoří list základy grafu na úrovni $n+1$. Výsledkem je pak tzv. vrchní kořenový hash („*Top Root Hash*“) X_{RT}.



Obrázek 12: Schéma hierarchického distribuovaného stromového systému Zdroj: [autor], [29]

Stromové grafy jsou libovolně stohovatelné a tím pádem rozšiřitelné. Lze tedy do předchozího znázornění přidat další podgrafy (trojúhelníky vyznačené čárkovaně). Ty lze pak dle potřeby provázat tak, aby byla zajištěna redundance (například tím, že jedno X_R bude listem pro více podgrafů na úrovni *n+1*, resp. *n+m*).

5.3.2 Plně uspořádaná schémata

Oproti částečně uspořádaným mají vyšší složitost a dovolují porovnání jakýchkoliv absolutních časových razítek (i ze stejného kola). Platí, že čím delší je kolo (tj. vyšší počet žádostí), tím je pak menší počet dat k ověřování. Základní schémata provázaných hashů popsal již v roce 1999 (a později doplnil) Buldas[83] [17] a Pugh [56].

5.3.2.1 Lineární provázané (Linear (Simple) Linking) schéma.

Lineární schéma, představuje základní princip systému provázaného hashování. Autor se zmíní o vlastnostech tohoto schématu v praktické části. Předpoklad je, že se bude jednat o nejpomalejší a tudíž nejméně efektivní způsob. V prvním kroku se hash vypočítává pouze z původního dokumentu. Za hodnotu hashe jednoho objektu v n-tém kroku označíme $\mathrm{h_n} = h(Doc_n)$.

$$H_n = h(h(Doc_n), h(Doc_{n-1})) \qquad \text{4-7: Hodnota hashe v n-tém kroku}$$

[83] Profesor kryptografie na technické univerzitě v Talinu a na univerzitě v Tartu (Estonsko).

Službu zajišťující označení (orazítkování) objektu razítkem s_n v relativním čase označme *T* za použití funkce *ts*. Jinými slovy s_n je produktem *T*. Z předchozích vztahů, kdy jsme dokument na vstupu (resp. objekt, resp. posloupnost bitů) označili Doc_n, který má v LTA systému svůj jednoznačný identifikátor ID_n. Ovšem na vstupu není označen časovým razítkem pouze samotný dokument, ale také celá řada dalších údajů. Označení pak probíhá způsobem, který reprezentuje následující vztah:

$$s_n = ts(n, t_n, ID_n, Doc_n; L_{n-1})$$

4-8: Lineární schéma hashování

Kde platí:

$$L_{n-1} = (t_{n-1}, ID_{n-1}, Doc_{n-1}; L_{n-2})$$

4-9: Vstupní hodnota z předchozího kroku

Kde:

n = pořadové číslo dokumentu.

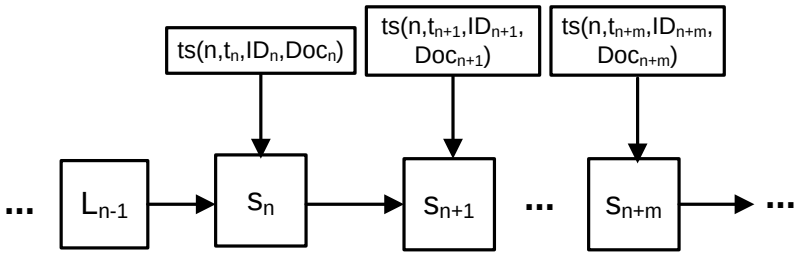
t_n = absolutní čas archivace n-tého dokumentu (UTC).

ID_n = jednoznačný identifikátor příslušného objektu.

Doc_n = příslušná vstup (bitová posloupnost).

s_{n-1} = hash z předchozího kroku. Zároveň představuje relativní čas archivace (s_{n-1} časové razítko vzniklo před s_n).

Z výše uvedeného je vidět, že v každém kroku je potřeba vypočítat hash z nově archivovaného dokumentu, plus z kontrolního součtu vytvořeného v předchozím kroku.



Obrázek 13: Lineární schéma

Zdroj: [17],[26],[autor]

5.3.2.2 Přeskakovací (Skip-List Linking) schéma a jeho modifikace.

Přeskakovací schéma je modifikací lineárního schématu provázaného hashování. Nevytváří se kontrolní součet z každého dokumentu na vstupu a hashe z kroku *n-1*, ale nový hash se vytvoří z již existujícího v posloupnosti o *m* dokumentů zpět. Je evidentní, že tento postup je výhodný z hlediska rychlejšího ověřování pravosti jednotlivých objektů (je-li požadováno) a to především díky přítomnosti pointerů[84]. Ty se využívají v případě hledání určitého objektu pro přeskakování dlouhých bloků. Princip přeskakovacího schématu je vyjádřen následujícím vztahem:

$$H_n = h(h(Doc_{n-m}), h(Doc_n)) \qquad \text{4-10: Přeskakovací schéma}$$

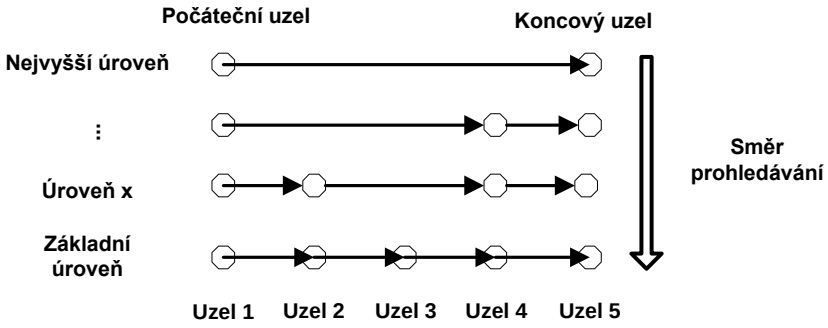
Pro tvorbu hashů je toto schéma výrazným rychlostním vylepšením. Bohužel rychlost prohledávání nedosahuje efektivity pokročilých (např. binární či závitové, viz následující podkapitoly) schémat. Ideálem je, aby se svojí rychlostí prohledávání přiblížilo ideálu $log_2 n$.

Princip přeskakovacího schématu má několik specifických variant. První z nich je tzv. dokonalé přeskakovací schéma („*Perfect Skip List*“), jehož konstrukce říká, že element *e* spadá do úrovně *l* pokud jeho index je násobkem 2^l. Dalším speciálním případem je tzv. přeskakovací schéma s indexováním. To ukládá počet přeskočených dokumentů (resp. délku jednotlivých úrovní) při vytváření hashů. Představuje tak vlastně přechod mezi lineárním a stromovým schématem. Procházení je vždy od nejvyšší úrovně až po základní podrobnosti [56].

U tohoto schématu je třeba si uvědomit, že každý uzel má následující vlastnosti:

- Hodnotu (hash, resp. žádost pro TSA).
- Úroveň (rozsah od 0 po nejvyšší úroveň přeskakovací listu, resp. schématu vertikálně).
- Index (pozice uzlu v přeskakovacím listu horizontálně).
- Značku (hash jednosměrně závislý na značkách předchozích uzlů).

[84] Jednoznačný ukazatel na umístění objektu v datové struktuře.



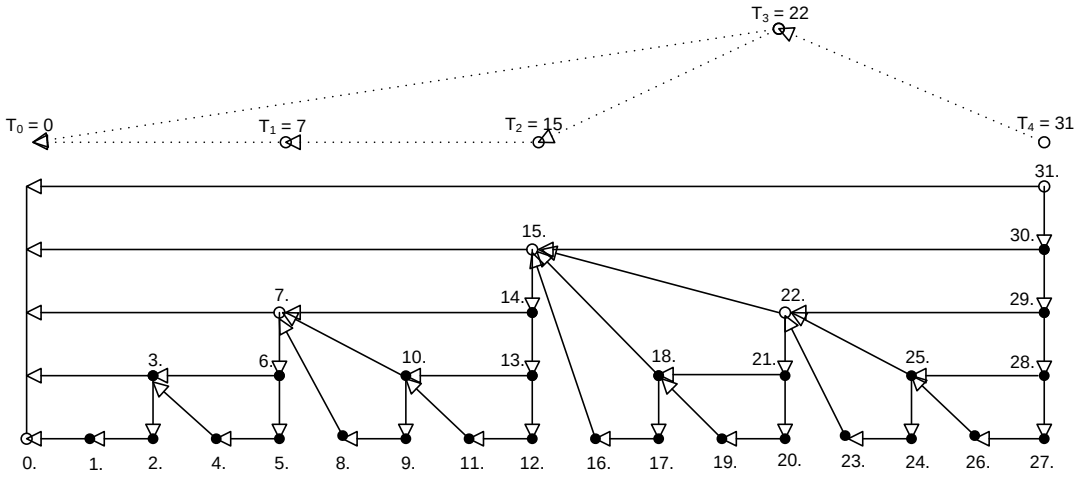
Obrázek 14: Přeskakovací schéma s indexováním Zdroj: [56], autor

Toho využívá například schéma definované v [12]. Po bližším prostudování lze konstatovat, že autory deklarované „*totally new*“ zase tak „*totally*“ není.

5.3.2.3 Binární (Binary Linking) schéma.

U schématu binárního se jedná o proprietální kombinaci lineárního a stromového. Zavedl ho a poprvé popsal Buldas [18]. Jeho základ tvoří „stanovení horní meze délky nejkratší cesty mezi kterýmikoliv dvěma časovými razítky“. Principielně tedy představuje další způsob provázaných hashů, jehož základ tvoří graf, nazvaný v tomto případě *Tk*. Ten představuje vždy souhrnný hash tvořený předem stanoveným počtem dokumentů pro dané kolo. Hlavní přínosem je kompletní provázanost, ovšem za cenu zdvojení hashů. Rozdíl mezi kontrolními součty je v tom, že jsou vždy provázány nejen navzájem, ale také s výsledným hashem z kroku *k+1*, respektive *k-1*.

Velkou nevýhodou binárního schématu je menší efektivita projevující se především vyšší časovou náročností jak při samotném razítkování, tak při ověřování ve srovnání se schématem přeskakovacím s a to především díky dodatečnému řetězení (viz Obrázek 15).

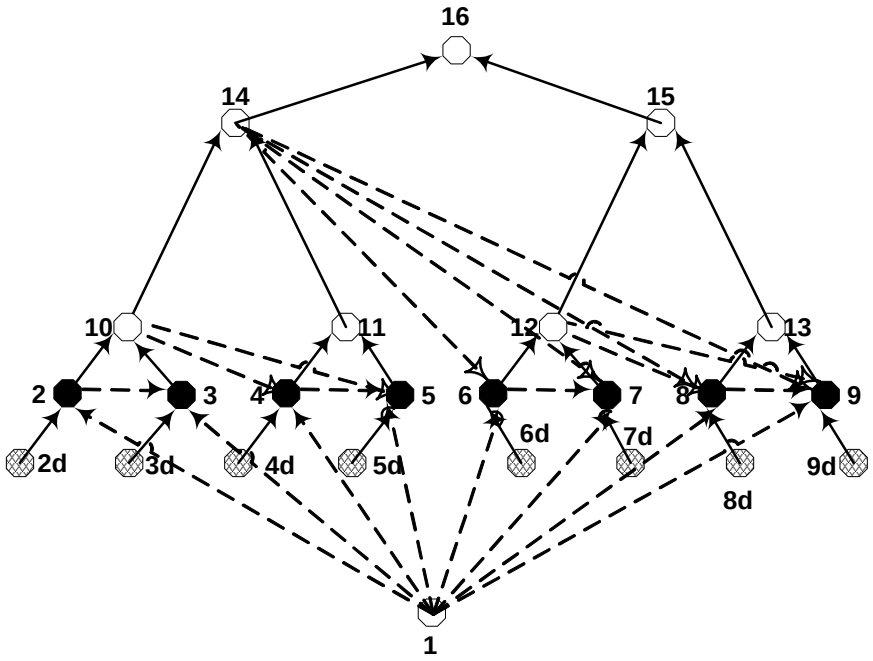


Obrázek 15: Binární schéma provázaných hashů **Zdroj: [18]**

Jednou z nejznámějších aplikací je tzv. „Tiger Tree Hash“. Používá standardní systém rodiče a dvou potomků s tím, že hashe jsou vypočítány z jednotlivých datových bloků (listů) o velikosti 1024 bajtů, s využitím kryptografické funkce Tiger (viz kapitola „Výběr vhodného hashovacího algoritmu“).

5.3.2.4 Závitové (Threaded Authentication) schéma.

Závitové schéma je speciálním případem schématu stromového. Jeho implementace je jednodušší než u binárního či stromového, ale stále má vysokou míru komplexnosti kvůli dodatečným řetězením. Stále tedy platí (stejně jako v případě „*Merkle Tree*“), že pro strom T_d o d úrovních velikost relativního časového razítka pro vrchol v je $k(d+2)$, kde k je velikost výstupu hashovací funkce h. V dalším kroku by velikost byla $k(d+3)$. Závitové schéma se vrací k původnímu grafu bez navyšovaní velikosti výstupu, ale implementací vstupu na každém indexu. V [19] lze nalézt detailní matematický popis a Obrázek 16 představuje grafické znázornění.



Obrázek 16: Závitové schéma Zdroj: [19]

5.4 Zajištění bezpečnosti hashovací funkce

K zajištění bezpečnosti hashovací funkce v modelu navrženém v kapitole 6 bude využito standardu DSSC[85] definované v RFC 5698 [47]. Jedná se o datovou strukturu, která umožňuje automatickou analýzu bezpečnosti vybraných kryptografických algoritmů v daném časovém okamžiku nebo k určitému datu. K určení vhodnosti nasazení dané kryptografické funkce slouží jako podklad zvolené bezpečností politiky. Ty musí splňovat:

- Automatickou interpretaci (vyhodnocení bezpečnosti dané kryptografické funkce).
- Flexibilitu (musí umožnit podporu nových algoritmů).
- Ověření zdroje politiky (daná politika musí důvěryhodná, veřejně akceptovaná bezpečností politika, buď na národní úrovni, či úrovni EU (kapitola 4.3.7).
- Autenticita a integrita dané politiky musí být ověřitelná, včetně data vydání.

[85] Data Structure for the Security Suitability of Cryptographic Algorithms.

Samotné ověřování bezpečnosti zvoleného algoritmu pak probíhá ve třech krocích:

1. Ověření politiky.
2. Stanovení platnosti kryptografického algoritmu.
3. Vyhodnocení důvěryhodnosti daného algoritmu.

Jako vstupní parametry pro ověření bezpečnosti pak pro DSSC slouží následující informace:

- Politika.
- Současný čas.
- Identifikace algoritmu a jeho parametrů (např. délka klíče).
- Požadovaný čas zjištění stavu bezpečnosti daného kryptografického algoritmu.

Výstupy DSSC jsou buď ve standardu XML schématu (RFC 3275) či ve standardu ASN.1 (RFC 5912). Exaktní aplikace DSSC v navrženém modelu je v kapitole 6.1.2.

5.5 Omezující podmínky upřesňující předmětnou oblast vědeckého bádání

Pro řešení vlastního vědeckého úkolu bylo žádoucí jasně vymezit technologické (legislativní a další již byly stanoveny dříve, viz kapitola 3.1) podmínky, za kterých lze ke splnění vytyčených cílů práce dospět. Omezujícími technologickými podmínkami přímo souvisejícím s integritou archivu jsou:

- Použitý algoritmus v čase *t* bude považován za absolutně bezpečný až do času *t+1*.
- Ve stanovené časové periodě (úseku) bude vždy archivován alespoň jeden objekt. Tato podmínka je závislá na volbě parametru.
    - dle času

      nebo
    - počtu archivovaných dokumentů.
- Hash pro jakkoliv archivovaný objekt bude moci být kdykoliv zpětně verifikován a vypočten (implikuje premisu neexistence fyzického vymazání).

Další ne přímo omezující podmínkou je předpoklad, že počet archivovaných objektů není v absolutním vyjádření omezen (tj. může růst nad všechny meze a je omezen pouze úložnou kapacitou).

5.6 Závěr z analýzy současného stavu vědeckého poznání

Z provedené analýzy a rešeršního šetření jednoznačné vyplývá, že archivace ED z pohledu dlouhodobosti není dořešeným tématem. Přístupy, kde se s ad-hoc změnou hlavního bezpečnostního elementu počítá, sice existují, ale neřeší ji napříč celým systémem uchovávání. Jinými slovy, to co bylo archivováno v čase *t-1*, má nezměněnou integritu také v čase t a bude ji mít i v čase *t+1*.

O možnosti napadení posloupnosti zřetězených hashů podvrženou duplicitou v čase *t* již nedůvěryhodné funkce z doby *t-1* se prostudovaná literatura ani v náznacích nezmiňuje. Některé z metod zmíněných v rešeršní části (například ACE [62], který si vede informaci (ID) použitého hashovacího algoritmu) sice řeší integritu LTA objektů formou částečně nezávislou na absolutním čase archivace, ale neřeší univerzálnost použitého způsobu zabezpečení.

6 Vlastní výzkum a návrhy řešení stanovených vědeckých úkolů

Jak již bylo uvedeno v kapitole 2, vlastní výzkum v oblasti LTA a jejích paradigmat, začal již během zahraničního výměnného pobytu v institutu Jože Štefana (Slovinsko) a předkládaná práce na tyto aktivity de facto navazuje. Prostudované literární prameny a návrhy níže uvedených cílů byly průběžně konzultovány s komunitou zainteresovanou v této oblasti vědeckého bádání (např. odborníky z firmy HP, NBÚ, či z již zmíněného institutu). Dílčí návrhy byly předneseny i na konferencích věnujících se archivaci výsledků z eParticipation projektu EU (viz publikační činnost autora).

Návrhy jak hlavního, tak dílčího cíle a požadavky na něj kladené jsou formulovány v podkapitole 2.5.1 a 2.5.2. Metoda chránící objekty v průběhu přechodu po změně původně použitého hashovacího algoritmu v důkazní struktuře archivu musí zajistit:

- Kontinuitu nepopiratelnosti.
- Zabránění záměny dokumentu či objektu, jeho neoprávněné modifikace nebo smazání.
- Nezpochybnitelnost původních důkazů pravosti (grafů provázaných hashů).
  Toto již dokázal Buldas v [21] a [22], kdy jasně vymezil, že použitý hashovací algoritmus nemá vliv na strukturu, resp. bezpečnost provázaného hashování. Co ovšem neřeší je vliv této skutečnosti na samotný dokument. Díky podstatě každého hashe[86] není pravděpodobná pouhá modifikace původního digitálního objektu útočníkem, ale podvržení za zcela jiný, jehož hash bude duplicitní. Je třeba tento systém dále rozvinout a obohatit o takový mechanismus, který strukturu relativní časové autentizace rozšíří o nepopiratelnou relaci dokument <-> hash.

Je zřejmé, že uvedené požadavky splňuje pouze takové řešení, které je založeno na minimálně jenom z následujících paradigmat:

[86] Malá změna na vstupu způsobí velkou změnu na výstupu.

1) Umožní paralelně spravovat *n* grafů.

   Každý systém provázaných hashů bude využívat jiného hashovacího algoritmu identifikujícího se unikátním ID. Tyto algoritmy budou dle [28] voleny tak, aby jejich předpokládaný „konec životnosti“ byl diametrálně odlišný. Dále díky využití metody (DSSC [46]) pro zjištění blížícího se zpochybnění bezkoliznosti použité hashovací funkce lze paralelně v souběhu implementovat nahrazující hashovací funkci nové generace a nahradit tak původní graf. Toto paradigma tak reprezentuje vlastnost systému: škálovatelnost.

2) Bude obsahovat jednoznačnou identifikaci archivovaného objektu. Tato identifikace musí být založena na relaci „*objekt <–> důkaz*“. To implikuje nezávislost na použitém hashovacím algoritmu. Musí tedy v jakýkoliv časový okamžik zajistit, že původní objekt je stále originální.

6.1 Návrh řešení hlavního cíle

Tato podkapitola řeší v dlouhodobém časovém horizontu zachování nepopiratelnosti archivovaného objektu po změně hashovací funkce, jejíž bezpečnost od určitého časového okamžiku již není zaručena. K tomu využívá aparát dvojí systémové ochrany spravovaných objektů.

Zjednodušeně cílem této návrhové podkapitoly je zabránění jakékoliv (i oprávněné) změny původního objektu v průběhu přechodu na jiný (bezpečnější) algoritmus. Dodržujíc první paradigma z předchozí podkapitoly, bude vždy existovat bezpečná hashovací funkce bez ohledu na použitý typ grafu. Bude tedy platit následujícím definice:

- Nechť ∃! taková bezkolizní hashovací funkce $h_x()$ v čase *t* o které můžeme říci, že je bezpečná pokud v čase *t* ∃! taková $h_y()$ o které platí že její výstup je shodný s původní funkcí $h_x()$ (viz kapitola 4.3.5). Předpoklady a omezující podmínky návrhu:

  - Nutnou vstupní podmínkou tohoto návrhu je paralelní existence alespoň dvou bezkolizních hashovacích funkcí, jejichž časový odstup pravděpodobnosti nálezu duplicity, je vzdálen minimálně o t[87].
- Každý objekt musí mít v systémových metadatech odpovídající relativní index, jenž definuje pořadí archivovaného objektu vzhledem k ostatním objektům (záleží na typu grafu).
- Nový graf bude zřetězen, až bude plně dokončeno generování nových hashů všech archivovaných objektů.
- Prohození primárního grafu za sekundární (tj. „již“ potenciálně nebezpečného za „ještě“ bezpečný) se bud řídit minimalizací času potřebného ke znovuvytvoření grafu založeného na funkci nové. V systémové teorii se tak bude tento stav považovat za vychýlení systému z normálu a snahou bude tedy minimalizovat čas, po který sekundární graf bude plnit funkci primárního.

    V době této systémové změny stavu se předpokládá, že služby poskytované dlouhodobým archivem budou pozastaveny (profylaxe). Primárním důvodem je zafixování konečného počtu objektů v čase po dobu nezbytně nutnou ke znovuvytvoření předmětného grafu.
- Primární a sekundární hashovací funkce, jež tvoří základní stavební kámen paralelně existujících grafů v systému, budou vždy z hlediska předpokládané doby bezkoliznosti od sebe vzdáleny o p_{max}, kde nejmenší uvažovanou jednotkou je jeden den (24h).

    Definice: Nechť v čase t ∃ grafy $g_1(h_1)$ a $g_2(h_2)$ a jejich bezpečnost $s(p_1)$, resp. $s(p_2)$. Předpokládejme, že ∃ taková p_{min} a *platí že* $p_1 << p_2$[88], pak:

$$p_{min} = \lfloor log_{365} (p_2 - p_1) \rfloor > 1$$

5-11: Minimální doba bezpečnosti použité hashovací funkce

[87] Je určeno politikou archivu. Většinou se bude jednat hodnotu času nezbytně nutnou k provedení „přehashování“ objektů, resp. pro zničení staré struktury grafů a vytvoření nové.
[88] Je tak zřejmé, že hlavním grafem se vždy stane ten, jenž je postaven na hashovací funkce, která má předpokládanou dobu bezkoliznosti kratší než graf záložní.

Nechť ∃ skupina h_1, h_2, …, h_x k nimž ∃ příslušné p_1, p_2, …, p_x pak pro p_{max} platí:

$$p_{max} = p_x - p_{x-n} > p_x - p_{x-m} > p_{min}$$

5-12: Maximální časová vzdálenost bezpečnosti použitých hashovací funkcí

kde ve výše uvedeném:

g = graf provázaných hashů (graf s indexem 1 je vždy primární).

h = příslušná hashovací funkce.

p = doba bezkoliznosti (bezpečnosti) příslušné hashovací funkce.

- Bezpečnost systému LTA monitoruje vždy primární graf. Bezkoliznost sekundárního grafu je ověřována separátně, nezávisle na grafu hlavním. Pokud je třeba vyměnit hashovací funkci sekundárního (záložního) grafu, lze to provést pouze za předpokladu, že v daném časovém okamžiku nenahrazuje funkci grafu hlavního (tj. nemá funkci hlavního bezpečnostního prvku zajišťujícího integritu archivu).
- V čase t=0, tj. na počátku spuštění služeb poskytovaných dlouhodobým archivem bude primární graf vždy vytvořen před grafem sekundárním.
- Volitelnou součástí přepisovatelných metadat bude hodnota udávající bezpečnostní klasifikaci[89] dokumentu.

  Na základě hodnoty tohoto parametru může být prováděno hashování a tvorba nového grafu. S ohledem na provedený experiment (kapitola 7) kdy je snaha o minimalizaci funkce času nutného k vytvoření nového grafu lze díky tomuto parametru mít vytvořeno *n* provázaných struktur, pro každý stupeň bezpečnosti dokumentu zvlášť. Blíže viz kapitola diskuse.

Pokud ovšem bude v každém časovém okamžiku zaručena bezkoliznost minimálně jedné hashovací funkce, znamená to, že není možné provést jakoukoliv neoprávněnou záměnu ani modifikaci archivovaného objektu. Systém provázaných hashů lze však stále po jistou dobu považovat za bezpečný ([22]) i přesto, že

[89] Tajné, přísně tajné atd.

duplicita k použité hashovací funkci již byla pravděpodobně nalezena. A to i bez existence jeho jednoznačné identifikace (není nezbytná – neexistuje duplicitní objekt).

Je také nezbytné zmínit, že jako důkazní struktura bude sloužit pouze jeden graf. Z toho plyne, že například při rekonstrukci důkazu integrity, bude použita příslušná kostra pouze z jednoho grafu. Který bude ve výsledku použit (tj. ten, u něhož se očekává nalezení duplicity použité hashovací funkce dříve či později) určuje opět archivní politika. Pokud by byla například stanovena vždy ta nejbezpečnější funkce, znamenalo by to také dřívější zničení dalších grafů a jejich přepočet (přehashování archivovaných objektů).

Souběhem aplikace obou těchto funkcí na nově archivované dokumenty v rámci systému má přímou nevýhodu ve dvojnásobné výpočetní náročnosti. Součástí návrhu je proto tvorba grafů kde $t_1 \neq t_2$, kde rozdíl mezi jednotlivými časy bude dán:

1) Nejmenší mírou „tresholdu“ zatížení stroje (v případě jeden stroj = jeden archiv).
2) V případě rozdělení zátěže mezi více strojů je definován hlavní (primární) a vedlejší (sekundární) stroj. Každý z nich je zodpovědný za vlastní strukturu provázaných hashů. Primárním strojem je vždy ten, jehož aktuálně používaná hashovacích funkce je dle DSSC bezpečnější (tj. v průběhu času se o svoji funkci střídají).

6.1.1 Systémový přístup

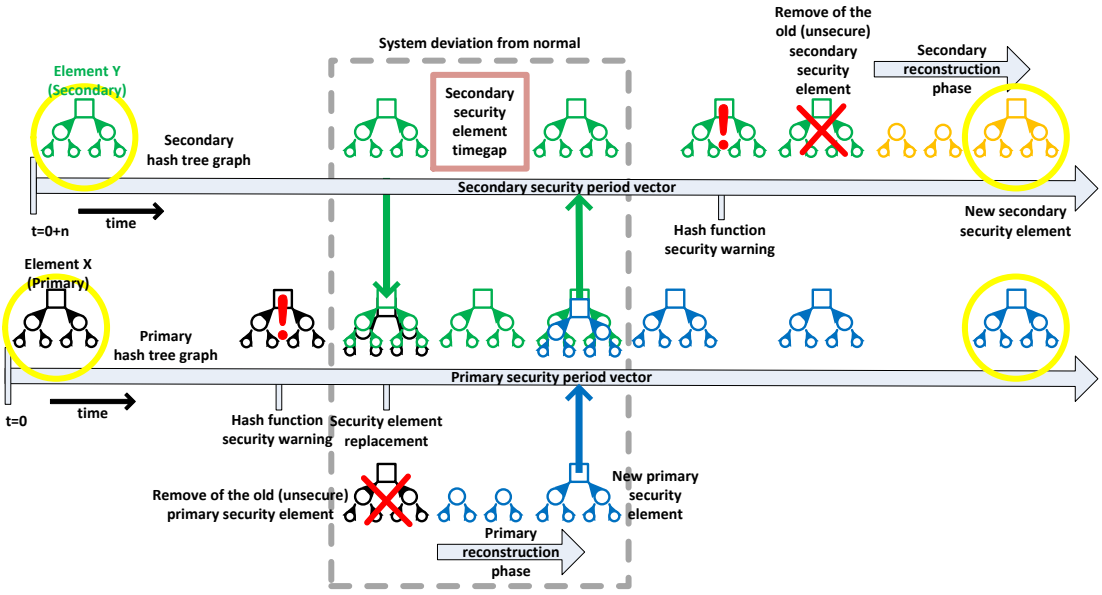
Z předchozí podkapitoly plyne, že v teorii systémů se jedná o tzv. řízený spojitý dynamický deterministický systém. Ten je zároveň časově neinvariantní s více vstupy a výstupy (MIMO). Jak bylo již napsáno v omezujících podmínkách tohoto návrhu, bude záměna grafů (prvků) systému považována za ovlivnění jeho stability a cílem tedy je minimalizovat tuto destabilizaci vrácením do výchozího stavu (primární graf bude opět primárním). S použitím matematického aparátu lze stabilitu tohoto systému, ve kterém koexistují 2 prvky, jež jsou ovlivněny okolím vyjádřit jako takové

řešení, kde můžeme o systému X(t) (stav systému X v čase *t*) říci že je stabilní, pokud každý stav systému Y v čase t+1 je při respektování min(*t)* blízké k X(t). Jinými slovy vychýlení z výchozího stavu netrvá déle než je nezbytně nutné. Stavovou proměnnou je v tomto případě čas, jako prvek vektoru.

Grafické znázornění systémové záměny prvků (Obrázek 17) nám udává, že změna hashovací funkce v čase nám ovlivní archivní systém nucenou záměnou jeho prvků. Primární a sekundární prvky systému si tak budou dynamicky měnit svoji roli. Nad výše uvedenými prvky systému bude „bdít" struktura DSSC, monitorující jejich dobu jeho „životnosti". Zjednodušeně se dá tedy životní cyklus (záměna) zapsat prvků systému zapsat:

$$X_0 \rightarrow Y_0 \rightarrow X_1 \rightarrow Y_1 \ldots$$

kde indexy nám udávají pořadová čísla jednotlivých hashovacích funkcí a chronologické využití v systému LTA.



Obrázek 17: Schéma záměny bezpečnostních prvků v systému archivu Zdroj: [autor]

Prezentovaný model výše vyjadřuje tvorbu primární bezpečnostní struktury (založené na hashovací algoritmu *X, např. SHA-512*) začínající v čase $t_1=0$ právě

tehdy, kdy je archivován první dokument (objekt). V čase $t_2=0+n$ začíná konstrukce záložní bezpečnostní struktury (hashovací algoritmus Y, např. SHA3-256). U obou aplikovaných hashovacích funkcí je již implementována DSSC, jež v čase t_s zajistí včasnou detekci nalezení kolize a započne fáze rekonstrukce primární bezpečnostní struktury založené na nové, bezpečnější hashovací funkci. Bezpečnost celého systému a z něho vyplývající integrita archivu je během této doby stále zajištěna záložním grafem Y.

Experimentální ověření návrhu řešení (a následné statistické zpracování výsledků) lze nalézt v kapitole 7.

6.1.2 Využití funkce DSSC v navrženém modelu

V kapitole 5.4 je uvedeno jaké parametry na výstupu vrací funkce pro včasnou detekci bezpečnosti hashovacích algoritmů (DSSC). K identifikaci použitého hashovacího (kryptografického) algoritmu je v navrženém modelu využito OID identifikátoru pro následující hashovací funkce:

- SHA-1 (OID: 1.3.14.3.2.26)
- SHA-384 (OID: 2.16.840.1.101.3.4.2.2)
- SHA-512 (OID: 2.16.840.1.101.3.4.2.3)

Pro hashovací algoritmy funkce rodiny SHA-3 prozatím (v době psaní této práce) neexistuje v IOD databázi záznam. Hashovací funkce TIGER bude využita v části experimentální ověření (kapitola 7) pouze jako informativní pro demonstraci rychlosti výpočtu a její nasazení v modelu navrženém v předchozí kapitole není uvažováno. DSSC také neumožňuje (ve standardu přímo zakazuje) násobný výskyt stejné hashovací funkce při ověřování, což je plně ve shodě s navrženým modelem. Výsledný zápis v XML standardu výše uvedených hashovacích funkcí využitých v modelu a v následném experimentálním ověření:

```
<SecuritySuitabilityPolicy xmlns="urn:ietf:params:xml:ns:dssc"
     xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
    <PolicyName>
     <Name>ETSI TS 102-176-1</Name>
    </PolicyName>
```

```
    <Publisher>
     <Name>European Telecommunications Standards Institute</Name>
    </Publisher>
    <PolicyIssueDate>2011-07-01T00:00:00</PolicyIssueDate>
    <Usage>Hash functions security suitability</Usage>
    <Algorithm>
     <AlgorithmIdentifier>
      <Name>SHA-1</Name>
      <ObjectIdentifier>1.3.14.3.2.26</ObjectIdentifier>
     </AlgorithmIdentifier>
     <Evaluation>
      <Validity>
       <End>2008-06-30</End>
      </Validity>
     </Evaluation>
    </Algorithm>
    <Algorithm>
     <AlgorithmIdentifier>
      <Name>SHA-384</Name>
      <ObjectIdentifier>2.16.840.1.101.3.4.2.2</ObjectIdentifier>
     </AlgorithmIdentifier>
     <Evaluation>
      <Validity>
       <End>2017-12-31</End>
      </Validity>
     </Evaluation>
    </Algorithm>
    <Algorithm>
     <AlgorithmIdentifier>
      <Name>SHA-512</Name>
      <ObjectIdentifier>2.16.840.1.101.3.4.2.3</ObjectIdentifier>
     </AlgorithmIdentifier>
     <Evaluation>
      <Validity>
       <End>2017-12-31</End>
      </Validity>
     </Evaluation>
    </Algorithm>
</SecuritySuitabilityPolicy>
```

Jako *Security Policy* (element `<PolicyName></PolicyName>`) je využito *ETSI TS 102-176-1* (kapitola 4.3.7), kde původcem (element `<Publisher></Publisher>`) je ETSI[90]. Dále s ohledem na kapitolu 6.1.2, zjednodušeně napsáno, červený vykřičník v navrženém modelu (Obrázek 17) je důsledkem dosažení stanoveného časového tresholdu parametru `<End></End>` u sub-elementu `<Validity></Validity>` náležícího pod element `<Algorithm></Algorithm>` datové struktury DSSC.

[90] European Telecommunications Standards Institute.

6.2 Návrh řešení dílčího cíle

Podstatou tohoto návrhu je zabránění záměrného i neúmyslného útoku na podvržení objektu, resp. zabránění jeho neoprávněné záměny. Název této podkapitoly vychází z druhého paradigmatu z kapitoly 6. V tomto případě bez ohledu na použitý způsob zajištění integrity objektu, resp. systému provázaných hashů, musí existovat minimálně lokálně[91], unikátní identifikace dané entity[92]. Zjednodušeně řečeno, podstata návrhu tedy spočívá v zabezpečení zdroje (archivovaného objektu), tak aby jej nebylo možné podvrhnout.

Omezující podmínky pro tento návrh jsou:

1. Úplná nezávislost na hashovací funkci a entropii bezkoliznosti.
2. Veličiny, tvořící důkaz existence objektu, budou k dispozici minimálně po dobu existence archivu[93].
3. Dokument bude vždy pevně svázán s metadaty. Pokud se metadata budou dělit na permanentní a editovatelná dokument bude vždy svázán s nepřepisovatelnými.
4. Velikost dokumentu na vstupu bud mít minimálně 1024 bajtů. Pokud bude soubor menší, bude archivem odmítnut[94].

Prezentovaný návrh si klade za cíl identifikátor objektu definovat univerzálně, bez ohledu na OAIS, či jiný model dlouhodobé archivace.

6.2.1 Existující přístupy

V oblasti otevřených informatických věd již existují různé přístupy k jednoznačné identifikaci objektu. Jedná se například o DOI, UUID nebo GUID. Ty však mají jedno společné – používají k identifikaci obsahu hashovací funkci, což porušuje první omezující podmínku. Stejně tak v oblasti patentovaných a komerčních řešení

[91] Ideálně globálně, tj. napříč archivy.
[92] Entita v tomto případě je představována buď celým objektem, jeho částmi, nebo jejich kombinací.
[93] Tj. dlouhodobě, ovšem s ohledem na deterministický charakter veličin a ukládání jednotlivých datových typů nelze s nimi pracovat bez stanovení horní meze.
[94] Nutné pro vytvoření identifikátoru, jenž je primárně koncipován pro větší objekty.

(například iTernity[95], resp. HP a jejich patentované CSC ID nebo již zmiňované ACE a jejich MDO). V patentu [31] je jednoznačný identifikátor dokumentu vyjádřen jako UUID („Identifier block“). Naproti tomu [62] neidentifikuje dokument jinak než pomocí hashů.

Dalším potenciálním problémem jsou generátory čísel (například u GUID) kde náhodnost je sub-optimální a generovaná čísla jsou tedy pouze pseudonáhodná[96]. Úplnou náhodnost nelze nikdy úplně zajistit - ani na základě přírodních jevů[97] a GUID svými čtyřmi pilíři na základě kterých unikátní identifikátor generuje, jde proti doporučení RFC 1750[98].

6.2.2 Parametry identifikátoru

Zajištění integrity jednotlivých objektů archivovaných v relativní časové posloupnosti bude muset být zabezpečeno pomocí nového nástroje. Ten bude muset splňovat minimálně následující body:

- Nezávislost na hashovací funkci.
- Jednoznačná identifikace binární struktury objektu.
- Jednoznačná identifikace absolutního času.
- Jednoznačný identifikátor relativního pořadí objektu v archivu.
- Vazbu na předchozí objekt.

Učinit z archivovaného dokumentu unikátní objekt lze jeho permanentním svázáním s nepřepisovatelnými systémovými metadaty. Tento jednoznačný identifikátor na straně klienta se bude skládat z následujících elementů:

1. „*Identifikace osoby*“ (původce). Pole „*s1*“.

[95] http://www.iternity.com/technology.html
[96] Generovaná na základě deterministické matematické funkce.
[97] Nothing in Nature is random. A thing appears random only through the incompleteness of our knowledge (Baruch Spinoza).
[98] http://www.ietf.org/rfc/rfc1750.txt

Je pro ni vyhrazeno pole o velikosti 32 znaků. Typicky unikátní „username“. Pokud je toto pole větší než 32 znaků provede se operace „*shrink*“, pokud menší provede se doplnění nulami.

2. „*Identifikace stroje*“ (klienta). Pole „*s2*“.

   Představuje jedinou volitelnou položku.

   Je založena na kombinaci hardwarová adresa (lokální síťová identifikace) stroje + hardwarová adresa síťové brány + veřejná IPv4 (IPV6 adresa) z API „http://myexternalip.com/raw“ nebo z „http://api.externalip.net/ip“. Skládá se pouze z číslic a písmen (tj. bez teček, dvojteček a dalších znaků). Neaplikují se metody pro zkrácený zápis IPv6 adresy. Velikost pole odpovídá maximálnímu počtu 56 znaků. Pokud bude použita IPv4 adresa pak bude zbytek bajtů do velikosti IPv6 doplněn nulami.

3. „*Čas*“. Poskytne TSA. Pole „*s3*“.

   Velikost 16 znaků ve formátu „*yyyymmddhhmmssxx*“. Dva znaky jsou rezervovány na případné zpřesnění časového údaje.

4. „*První dokument*“. Pole „*s4*“.

   Příznak zdali se jedná o první archivovaný dokument v nejmenší časové jednotce či nikoliv. Pokud ano hodnota pole je *1*.

   Velikost pole 1 znak.

5. „*Pořadové číslo*“ archivovaného dokumentu v dané nejmenší časové jednotce. Pole „*s5*“.

   Pokud se jedná o první dokument, pak toto pole obsahuje samé nuly.

   Velikost pole 7 znaků.

6. Kopie každého $2^b - 2^{b-2}$ bajtu (HEX) dokumentu, kde:

   $(b_n)_{n=2}^{n=x-1}$

   $x \doteq \log_2 \gamma \leq 60$[99],

   γ = velikost dokumentu v bajtech. Pole „*s6*“.

[99] Uvažovaná maximální velikost objektu, je 1024 EB.

Tato kopie se stane nedílnou součástí nepřepisovatelných metadat, která budou umístěna na začátku dokumentu (dle logiky pramenící z tradičních archivů). U souborů menší než maximální uvažovaná velikost bude zbytek pole metadat doplněn nulami. Pro toto bude vyhrazeno 60 znaků.

7. „*Důkaz předchozího dokumentu*“. Pole „*s7*“.
   Nepřepisovatelná metadata budou také obsahovat identifikátor *n-1* archivovaného objektu. To bude prvních 128 bitů (tj. první blok) z celkového identifikátoru. Velikost pole bude odpovídat 128 znakům. V případě že neexistuje předchozí dokument (jednoznačně dán indexem v metadatech) bude pole vyplněno nulami.
8. Zbývajících 20 bitů je rezervováno a vyplněno nulami. Pole „s8“

Tyto prvky budou sloučeny výše uvedeným uspořádaným způsobem „*concat(char *s1, *s2, …, *s8)“ do celku o velikosti odpovídající 256 bitům. Na tento řetězec bude aplikována symetrická bloková šifra AES[100]. Vstupní řetězec bude tedy rozdělen na 2 části a ty budou zašifrovány stejným klíčem. Jeho navrhovaná délka je 256 bitů. Sdílené tajemství bude hardwarově svázáno s archivem a každý archiv (pokud jich bude existovat více, viz diskuse) bude mít svůj unikátní klíč. Jeho sestavení v případě ztráty bude možné pouze z jeho fragmentů. Jejich umístění a vlastnictví definuje politika archivu.

S výstupním identifikátorem, jehož se dosáhne zřetězením výše uvedených polí, je možné dále pracovat a aplikovat na něj další techniky z oblasti bezpečnosti či kryptografie. Vstupní pole musí být také součástí metadat aby byla rekonstrukce možná, a nemohou být modifikována.

[100] AES provádí kryptografické operace vždy nad 128 bitovými bloky.

6.2.3 Algoritmus návrhu

Návrh využívá následující funkce:

Tabulka 7: Funkce využívané pro identifikace objektu a jejich popis

Název funkce	Popis funkce
GetSize()	Vrátí velikost vstupní proměnné.
GetCheck()	Vrátí zda-li proměnná je prázdná, či nikoliv.
Shrink$_x$()	Vrátí oříznutou vstupní proměnnou na zadaný (x) počet znaků.
GetTime$_x$()	Vrátí aktuální čas s předem definovanou přesností.
GetValue()	Vrátí hodnotu zadané proměnné.
GetBin()	Zásadní funkce – vrací hodnotu bajtu vstupu na dané pozici. Jejím hlavním cyklem je „FOR n=2 TO m WHERE m=$\log_2 Y \leq 60$ DO“
ZeroFill$_x$()	Vrací zadaný počet (x) nul – využívá se při doplnění na danou velikost.
Conc()	Sloučí stringy ze zadaných proměnných v definovaném pořadí.

Zdroj: [autor]

Se znalostí předchozí tabulky je vlastní algoritmus následující:

FOR α DO
IF α < 1025 THEN
A=GetSize(α);
B=ZeroFill $_{(1024 - A)}$;
α=Conc(*α, *B);
END;
A=GetSize(s1);
IF A>32 THEN Shrink(s1);
ELSE IF A<32 THEN s1 = Conc (*s1; *ZeroFill$_{(32 - A)}$);
GetCheck(s2)≠0;
THEN A = GetSize(s2);
IF A < 29 ;
THEN B = ZeroFill$_{(28 - A)}$;
s2 = Conc(*s2; *B);
ELSE A = GetTime$_{14}$();
s3 = Conc(*A; *ZeroFill$_{(16 - \text{GetSize(A)})}$;
IF s3≠s3$_{(n-1)}$ THEN s4=1 AND s5=ZeroFill$_{(7)}$;
ELSE s4=0;
IF s4=0 THEN s5=(GetValue(s5$_{(n-1)}$)+1;
s6=GetBin(α) WHERE Y=GetSize(α);
s6=Conc(*s6; *ZeroFill$_{(60-m)}$);
IF EXIST s7$_{(n-1)}$ THEN s7=Shrink$_{128}$(β$_{(n-1)}$);

```
            ELSE s7=0;
            s8=ZeroFill16();
        END;
        β = Conc(*s1; *s2; *s3; *s4; *s5; *s6; *s7; *s8);
```

kde:
α = vstupní dokument

β = výstupní řetězec

Pokud není uvedeno jinak, jsou hodnoty vždy uvedeny v bajtech.

Příklady generování navrhovaného jednoznačného identifikátoru na testovacím dokumentu lze nalézt v příloze A.

7 Vědecký experiment

Úkolem této kapitoly je experimentálně ověřit, návrh z kapitoly 6.1. V rámci experimentu bude prověřena čistá výpočetní (časová) náročnost na rychlost získání výstupu dané hashovací funkce. Experiment byl proveden za jinak stejných podmínek (tj. byly zachovány konzistentní výpočetní zdroje, vstupní množina testovaných objektů atd.).

7.1 Prostředí pro provedení experimentu

Empirické ověření matematické náročnosti návrhu bylo provedeno na hlavním počítačovém uzlu („clusteru“) Jihočeské Univerzity v Českých Budějovicích (dále jen „Hermes“). Ten je součástí distribuované výpočetní infrastruktury Metacentra sdružení CESNET. Instituce jako vlastník tohoto stroje má přístup do tzv. privilegované fronty, která mu umožňuje využívat jeho plný potenciál počítáním v dlouhých časových intervalech a má přednost před ostatními uživateli. Hermes tvoří celkem 11 výpočetních uzlů a jeden řídící. K dispozici je v tomto clusteru celkem 96 procesorových jader, čehož je tedy 88 výpočetních. Byl testován výpočet na jednom stroji (tj. při dostupnosti 8 jader) Paralelní zpracování výpočtu v tomto případě není možné (lineární graf se vytváří postupně z archivovaných objektů z jednoho úložiště).

7.2 Popis experimentu

V rámci provedeného vědeckého experimentu bylo pracováno s kvantifikovatelnými daty. Empirická data pro následné statistické zpracování s využitím převážně metod vícerozměrné domény. Primární kolekce dat, byla sestavena především z vlastního výzkumu a to konkrétně z následujících hodnot:

- Časových veličin reprezentujících dobu potřebnou k vytvoření výstupu jednotlivých bezpečnostních prvků systémové struktury.
- Bezpečnosti jednotlivých datových struktur založených na aktuálně bezpečných hashovacích funkcích.

Datové soubory, vzniklé z provedené výzkumné činnosti (kvantitativní data) byly zpracovány za použití následujících statistických nástrojů:

- Pro práci s velkými datovými soubory pokud bude třeba, lze využít metody pro redukci velkých datových souborů a to za použití předzpracování a normalizaci datového souboru.
- Pro zjištění závislosti mezi proměnnými byla na normalizovaný výstupní datový soubor aplikována vícefaktorová analýza rozptylu[101] s faktory typ objektu a hashovací funkce. Jako koeficientu byla použita doba výpočtu výstupu příslušné hashovací funkce (resp. dobu tvorby grafu jako součet tvorby všech zřetězených hashů) a koeficientu doby bezpečnosti.
- V případě zamítnutí nulové hypotézy je vhodné doplnit analýzu tzv. Tukey HSD[102] test pro zjištění statistické významnosti rozdílů středních hodnot jednotlivých proměnných. Tento test provedl mnohonásobné vzájemné porovnání všech dvojic vyskytujících se skupin.

Výše uvedené metody pro statistické zpracování dat byly zvoleny na základě prostudování relevantní odborné literatury ([33], [60]). Byly však zvoleny tak, aby vyhovovaly stanovenému vědeckému problému a veličinám získaných z experimentálního ověření.

Nosnou část experimentu reprezentuje samotná aplikace, jež byla naprogramována v jazyce Java. Ke spuštění bylo nutné mít na cílovém stroji nainstalováno Java JRE 1.7. Aplikace dovoluje stanovit také jako parametr stanovit počet opakování jako podklad získání průměrného času výpočtu u jednotlivých objektů a směrodatné odchylky pro stanovení statistické disperze. Primární kolekce dat je pak jako výstup uložena v (parametricky) daném CSV souboru. Každý řádek (záznam) ve výstupním souboru obsahuje pro každý použitý hashovací algoritmus 4 pole a to:

a. Název použitého hashovacího algoritmu.

[101] Analysis of Variance („ANOVA“).
[102] „Honestly“ Significant Difference Test

b. Výstupní hash s využitím daného algoritmu.
c. Průměrná délka výpočtu (v mikrosekundách) – x0 = sum(x)/n .
d. Směrodatná odchylka délky výpočtů s = sqrt((x-x0)^2/(n-1)).

V době provedení experimentu jsou podporovány následující hashovací funkce:

- MD5
- **sha-1 (SHA1)**
- **Tiger**[103]
- sha-224 (SHA2 family)
- sha-256 (SHA2 family)
- **sha-384 (SHA2 family)**
- **sha-512 (SHA2 family)**
- sha3-224 (SHA3 family)
- **sha3-256 (SHA3 family)**
- sha3-284 (SHA3 family)
- **sha3-512 (SHA3 family)**

Tučně jsou vyznačeny ty funkce, jež byly v v průběhu experimentu skutečně použity.

7.2.1 Omezující podmínky

Zásadní omezující podmínku pro průběh experimentu jsou I/O diskové operace, které mají přímý vliv na rychlost výpočtu hodnot hashů jednotlivých funkcí. Při testování se potvrdilo, že výpočetně nejjednodušší (tj. procesor nejméně zatěžující) algoritmy byly při čtení z datového úložiště na I/O operace nejnáročnější. Graf pro každou použitou hashovací funkci byl extenzivně vytvářen na dedikovaném stroji ze zdrojových objektů a to až ve chvíli kdy doběhl proces na stroji jiném (aby nedocházelo k ovlivnění paralelních I/O operací ze stejného zdroje). Tj. graf g_x a jemu příslušný krok tvorby hashe $g_x h(Doc_n)$ musí předcházet tvorbě hashe v grafu g_y o časovou jednotku $g_y h(Doc_n)$.

7.3 Průběh a statistické zpracování

Experiment proběhl na testovací množině digitálních objektů, reprezentovaných ED (bitstreamy) různých verzí, formátů, velikostí a důležitosti. Objekty měly taktéž

[103] Hashovací funkce Tiger byla vybrána záměrně, z důvodu demonstrace využití alternativní hashovací funkce oficiálně nepodporované ETSI.

různou informativní hodnotu a rozdílnou bitovou reprezentaci. Celkem bylo archivováno 5686 objektů, jež byly podkladem pro vytvoření zřetězených hashů zformovaných do celkem 6 grafů dle použité hashovací funkce. Pro dosažení statistické významnosti byl experiment opakován stokrát. Odlišnosti v rychlosti výpočtů při použití jednotlivých hashovacích funkcí jsou vyjádřeny směrodatnou odchylkou.

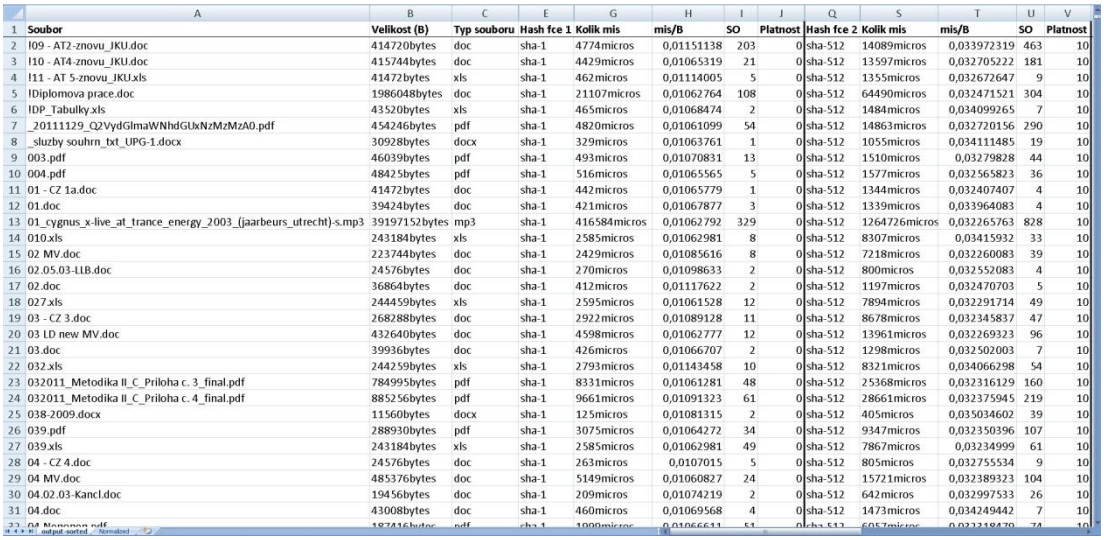
7.3.1 Statistické (před)zpracování

V průběhu samotného experimentu byla získána primární kolekce dat reprezentující výstupy jednotlivých hashovacích funkcí aplikovaných na testovací množinu. Aby nad tuto kolekcí dat mohly být aplikovány vybrané metody statického aparátu, většinou z vícerozměrné domény, bylo nutné provést její předzpracování. To představovalo provést následující kroky právě v tomto pořadí:

1. Normalizace dat: doba trvání výpočtu v µs / bajt
2. Stanovení koeficientu bezpečnosti použitých hashovacích funkcí. Ten je odvozen od predikované doby bezpečnosti předmětné funkce dle doporučení ETSI (viz. kapitola 4.3.7) v letech.
3. Rozdělení zpracovaných (hashovaných) objektů dle archivovaného typu (dokument, video, audio, text).
4. Extrakce závislých a nezávislých proměnných.

Výsledný datový soubor ve formátu CSV byl importován to MS Excel 2007 a zde předzpracován pro následný import do Statistiky. Obrázek 18 představuje částečně pro potřeby zobrazení upravená[104] původní zdrojová. Vystupují zde proměnné, kde každý řádek byl reprezentován jedním souborem, jehož výsledný hash byl vstupem do řádku (kroku) následujícího. Lze si povšimnout proměnných: velikost souboru v bajtech, celková výpočetní dobu v µs, typ použité hashovací funkce, její bezpečnost a směrodatnou odchylku pro 100 provedených opakování.

[104] Samotné výstupy hashovacích funkcí v jednotlivých krocích byly skryty pro jejich prostorovou náročnost.



	A	B	C	E	G	H	I	J	Q	S	T	U	V
1	Soubor	Velikost (B)	Typ souboru	Hash fce 1	Kolik mis	mis/B	SO	Platnost	Hash fce 2	Kolik mis	mis/B	SO	Platnost
2	!09 - AT2-znovu_JKU.doc	414720bytes	doc	sha-1	4774micros	0,01151138	203	0	sha-512	14089micros	0,033972319	463	10
3	!10 - AT4-znovu_JKU.doc	415744bytes	doc	sha-1	4429micros	0,01065319	21	0	sha-512	13597micros	0,032705222	181	10
4	!11 - AT 5-znovu_JKU.xls	41472bytes	xls	sha-1	462micros	0,01114005	5	0	sha-512	1355micros	0,032672647	9	10
5	!Diplomova prace.doc	1986048bytes	doc	sha-1	21107micros	0,01062764	108	0	sha-512	64490micros	0,032471521	304	10
6	!DP_Tabulky.xls	43520bytes	xls	sha-1	465micros	0,01068474	2	0	sha-512	1484micros	0,034099265	7	10
7	_20111129_Q2VydGlmaWNhdGUxNzMzMzA0.pdf	454246bytes	pdf	sha-1	4820micros	0,01061099	54	0	sha-512	14863micros	0,032720156	290	10
8	_sluzby souhrn_txt_UPG-1.docx	30928bytes	docx	sha-1	329micros	0,01063761	1	0	sha-512	1055micros	0,034111485	19	10
9	003.pdf	46039bytes	pdf	sha-1	493micros	0,01070831	13	0	sha-512	1510micros	0,03279828	44	10
10	004.pdf	48425bytes	pdf	sha-1	516micros	0,01065565	5	0	sha-512	1577micros	0,032565823	36	10
11	01 - CZ 1a.doc	41472bytes	doc	sha-1	442micros	0,01065779	1	0	sha-512	1344micros	0,032407407	4	10
12	01.doc	39424bytes	doc	sha-1	421micros	0,01067877	3	0	sha-512	1339micros	0,033964083	4	10
13	01_cygnus_x-live_at_trance_energy_2003_(jaarbeurs_utrecht)-s.mp3	39197152bytes	mp3	sha-1	416584micros	0,01062792	329	0	sha-512	1264726micros	0,032265763	828	10
14	010.xls	243184bytes	xls	sha-1	2585micros	0,01062981	8	0	sha-512	8307micros	0,03415932	33	10
15	02 MV.doc	223744bytes	doc	sha-1	2429micros	0,01085616	8	0	sha-512	7218micros	0,032260083	39	10
16	02.05.03-LLB.doc	24576bytes	doc	sha-1	270micros	0,01098633	2	0	sha-512	800micros	0,032552083	4	10
17	02.doc	36864bytes	doc	sha-1	412micros	0,01117622	2	0	sha-512	1197micros	0,032470703	5	10
18	027.xls	244459bytes	xls	sha-1	2595micros	0,01061528	12	0	sha-512	7894micros	0,032291714	49	10
19	03 - CZ 3.doc	268288bytes	doc	sha-1	2922micros	0,01089128	11	0	sha-512	8678micros	0,032345837	47	10
20	03 LD new MV.doc	432640bytes	doc	sha-1	4598micros	0,01062777	12	0	sha-512	13961micros	0,032269323	96	10
21	03.doc	39936bytes	doc	sha-1	426micros	0,01066707	2	0	sha-512	1298micros	0,032502003	7	10
22	032.xls	244259bytes	xls	sha-1	2793micros	0,01143458	10	0	sha-512	8321micros	0,034066298	54	10
23	032011_Metodika II_C_Priloha c. 3_final.pdf	784995bytes	pdf	sha-1	8331micros	0,01061281	48	0	sha-512	25368micros	0,032316129	160	10
24	032011_Metodika II_C_Priloha c. 4_final.pdf	885256bytes	pdf	sha-1	9661micros	0,01091323	61	0	sha-512	28661micros	0,032375945	219	10
25	038-2009.docx	11560bytes	docx	sha-1	125micros	0,01081315	2	0	sha-512	405micros	0,035034602	39	10
26	039.pdf	288930bytes	pdf	sha-1	3075micros	0,01064272	34	0	sha-512	9347micros	0,032350396	107	10
27	039.xls	243184bytes	xls	sha-1	2585micros	0,01062981	49	0	sha-512	7867micros	0,03234999	61	10
28	04 - CZ 4.doc	24576bytes	doc	sha-1	263micros	0,0107015	5	0	sha-512	805micros	0,032755534	9	10
29	04 MV.doc	485376bytes	doc	sha-1	5149micros	0,01060827	24	0	sha-512	15721micros	0,032389323	104	10
30	04.02.03-Kancl.doc	19456bytes	doc	sha-1	209micros	0,01074219	2	0	sha-512	642micros	0,032997533	26	10
31	04.doc	43008bytes	doc	sha-1	460micros	0,01069568	4	0	sha-512	1473micros	0,034249442	7	10
32	04 Nanonon.pdf	187416bytes	pdf	sha-1	1999micros	0,01066611	51	0	sha-512	6957micros	0,033218479	74	10

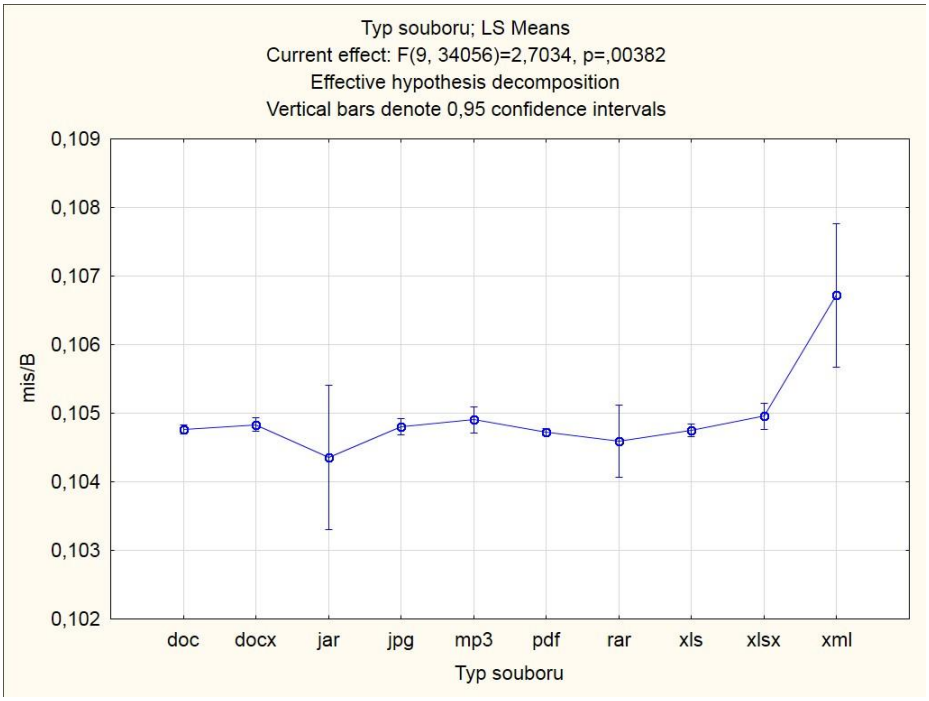
Obrázek 18: Výstup z experimentální aplikace **Zdroj: [autor]**

7.4 Vyhodnocení experimentu

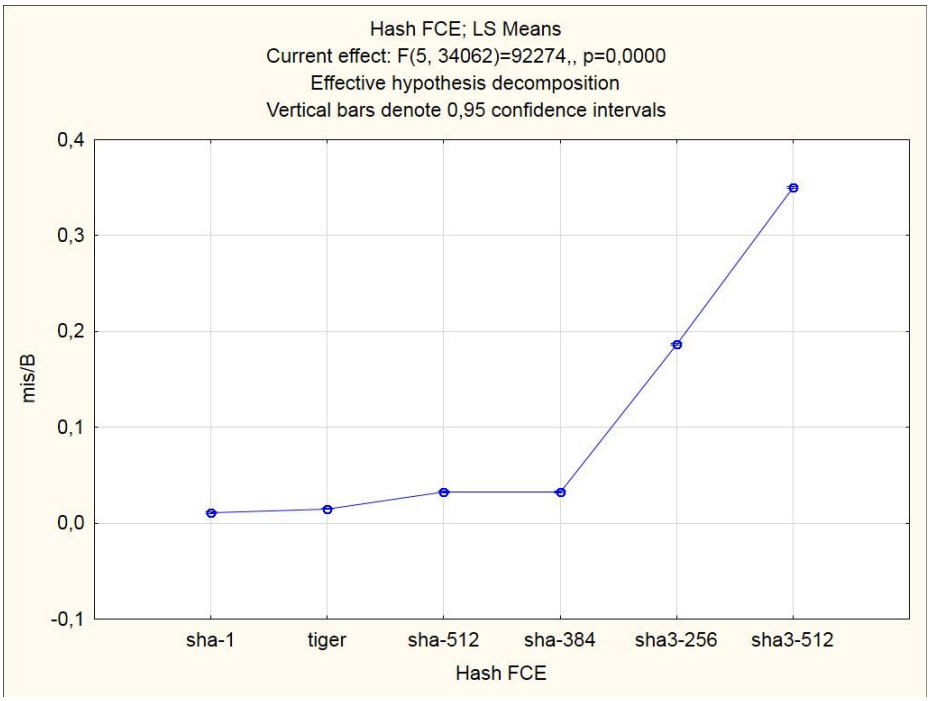
Předzpracovaný datový soubor byl dále vyhodnocen v programu Statistica V12[105]. Výsledky byly rozděleny do dvou skupin a to podle typu objektu a bezpečnosti použité hashovací funkce, kde závislé proměnná bude vždy normalizovaná doba potřebná k výpočtu příslušného hashe a faktory typ souboru a typ hashovací funkce. Vyhodnocení samotného experimentu je dále demonstrováno v níže uvedených tabulkách a grafech spolu s jejich popisem.

Na grafu (Obrázek 19) je patrný nárůst průměru normalizované doby výpočtu u textového souboru formátu XML. Také je patrný rozptyl na intervalu spolehlivosti $\alpha = 0,95$ zejména u souborů typu JAR, RAR a XML. Velmi zajímavý, je také závěr, že rychlost výpočtu hashů z dokumentů typu OOXML[106] (*.*x), jehož formát představuje ZIP kontejner, byla vždy horší, než u dokumentů proprietálního MS Office předchozí generace[107].

[105] http://www.statsoft.cz/
[106] Office Open XML, standard ISO/IEC 29500
[107] Microsoft Office 2003 a nižší.



Obrázek 19: Závislost normalizované rychlosti výpočtu hashe na typu souboru. **Zdroj: [autor]**



Obrázek 20: Druh hashovací funkce a normalizovaná rychlost výpočtu **Zdroj: [autor]**

Na grafu výše je jasně patrný nárůst výpočetního času nutný pro nové hashovací funkce SHA3 oproti předchozí generaci SHA2. Vzhledem k faktu, že se jedná o novou hashovací funkci, která ještě v době psaní této práce nebyla plně standardizována, tudíž se lze domnívat, že její implementace do programovacího jazyka Java verze 1.7. není optimální, proto výsledky plynoucí z Obrázek 20 týkající se SHA3 je nutné brát s rezervou. Velmi zajímavým je ovšem fakt výpočetní náročnosti hashovacích funkcí rodiny SHA2, kde velikost výstupu v bitech má na normalizovanou rychlost výpočtu v µs/B zanedbatelný, až téměř nulový vliv.

Effect	Univariate Tests of Significance for mis/B (output-normalized-new-predzpracovano.sta) Sigma-restricted parameterization Effective hypothesis decomposition SS	Degr. of Freedom	MS	F	p		
Intercept	16,82847	1	16,82847	1633024	0,000000		
Typ souboru	0,00025	9	0,00003	3	0,003817		
Hash FCE	24,05918	5	4,81184	466938	0,000000		
Typ souboru*Hash FCE	0,00055	45	0,00001	1	0,190348		
Error	0,35095	34056	0,00001				

Obrázek 21: ANOVA pro všechny faktory **Zdroj: [autor]**

Obrázek 21 zobrazuje jednorozměrnou přehledovou tabulku ANOVA se všemi faktory a jednou závislou proměnou (čas výpočtu hashe) a příslušným stupněm volnosti (Degr. of Freedom).

- Pro typ souboru: F(9, 34056), p = (0,0038) z čehož vyplývá, že je rozdíl mezi typy souboru. Tímto se **zamítá** nulová hypotéza H_{01} (není rozdíl mezi typy (formáty) souborů).
- Pro typ hashovací funkce: F(5, 34056), p = (10^{-35}) je zřejmé, že je staticky významný rozdíl mezi typy hashovacích funkci. Tímto se **zamítá** druhá nulová hypotéza H_{o2} (tj. že není rozdíl mezi bezpečností hashovacích funkcí).

Z testu analýzy rozptylu je zřejmé, že jsou statisticky významné rozdíly jak mezi typy jednotlivých souborů, tak mezi hashovacími funkcemi. Jak již bylo vymezeno v kapitole 7.3, v případě zjištění rozdílu mezi faktory (zamítnutí statistických hypotéz o neexistenci rozdílů) pro chyby 1. druhu α bude využitu Tukey HSD testu. Tento

test umožní odpovědět na otázku, mezi kterými typy souborů a hashovacími funkcemi je staticky významný rozdíl (na hladině významnosti 0,05).

Tukey HSD test; variable mis/B (output-normalized-new-predzpracovano.sta)
Approximate Probabilities for Post Hoc Tests
Error: Between MS = ,00001, df = 34056,

Cell No.	Typ souboru	{1} ,10476	{2} ,10483	{3} ,10436	{4} ,10480	{5} ,10491	{6} ,10472	{7} ,10459	{8} ,10475	{9} ,10496	{10} ,10672
1	doc		0,975182	0,999099	0,999885	0,934501	0,998417	0,999781	1,000000	0,696105	0,009771
2	docx	0,975182		0,996760	0,999997	0,999786	0,695930	0,996542	0,965923	0,985708	0,016404
3	jar	0,999099	0,996760		0,998112	0,991744	0,999580	0,999997	0,999331	0,984468	0,056184
4	jpg	0,999885	0,999997	0,998112		0,997161	0,979686	0,998886	0,999450	0,951352	0,013719
5	mp3	0,934501	0,999786	0,991744	0,997161		0,762443	0,984219	0,914938	0,999998	0,029186
6	pdf	0,998417	0,695930	0,999580	0,979686	0,762443		0,999971	0,999995	0,424334	0,007524
7	rar	0,999781	0,996542	0,999997	0,998886	0,984219	0,999971		0,999899	0,957265	0,013603
8	xls	1,000000	0,965923	0,999331	0,999450	0,914938	0,999995	0,999899		0,669068	0,009114
9	xlsx	0,696105	0,985708	0,984468	0,951352	0,999998	0,424334	0,957265	0,669068		0,039365
10	xml	0,009771	0,016404	0,056184	0,013719	0,029186	0,007524	0,013603	0,009114	0,039365	

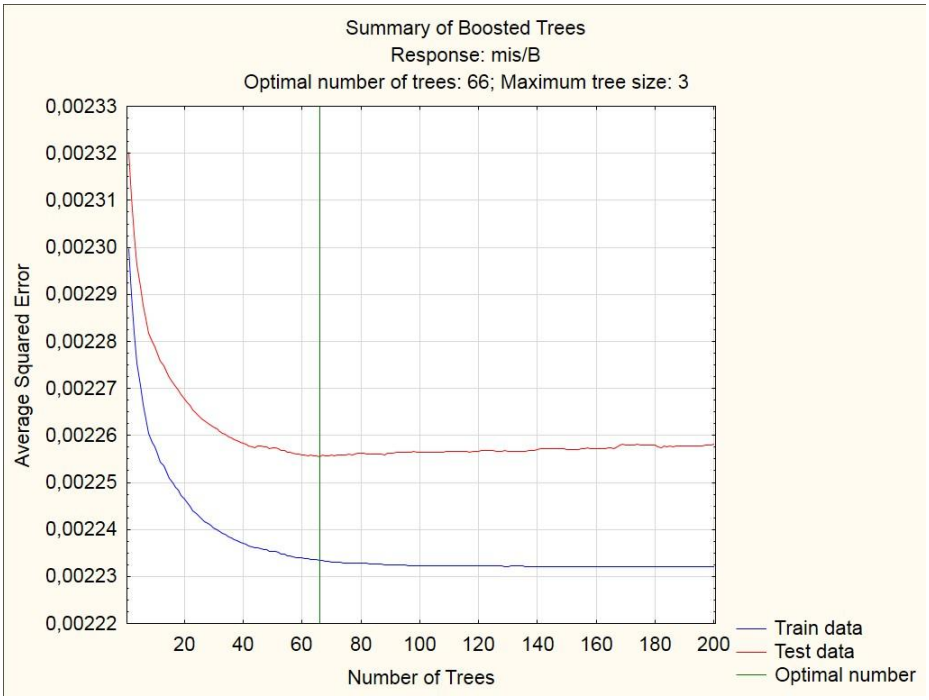
Obrázek 22. Tukey HSD test pro typ souboru **Zdroj: [autor]**

Obrázek 22 reprezentuje Tukey HSD test pro jednotlivé tyty souborů. Je patrné, že staticky významné rozdíly jsou v případě souboru XML se všemi ostatními objekty, kromě souboru typu JAR s tím, že jen těsně přesáhne stanovenou hranici statistické významnosti 0,05 (přesně (5,6184%). Obdobně bylo postupováno pro faktor „typ hashovací funkce“, kde bylo zjištěn statisticky významný rozdíl (opět na hladině významnosti p = 0,05) mezi všemi hashovacími funkcemi, kromě hashovacích funkcí SHA-384 a SHA-512. Z pohledu výpočetních nároků jsou obě hashovací funkce srovnatelné.

Na predikci chování navrženého modelu při archivaci dalších objektů a s tím spojený následný nárůst souborů v archivu bylo využito pokročilých metod regresní analýzy, běžně využívaných v dataminingu. Vstupní data představují trénovací (historickou) množinu a na základě ní predikovat chování modelu v budoucnosti. Použitá metoda „Boosting trees“[108], vytváří posloupnost různých typů modelů a jejich kombinací, jimž jsou přiřazeny váhy na základě počtu chybně klasifikovaných případů. Jinými slovy každý následný strom „trénuje“ na neúspěšných vzorech stromů předchozích (výsledný graf je složen ze stromu každé předchozí iterace) a složením pak vznikne jeden ucelený model. Na grafu níže lze vidě tzv. cross-validation model, kde optimální počet stromů je 66, kde testovací data dosahují minimální prahové

[108] http://www.statsoft.com/Textbook/Boosting-Trees-Regression-Classification

hodnoty a od tohoto bodu se již začínají vzdalovat od dat trénovacích. Jako závislá proměnná byla použita normalizovaná doba výpočtu hashe, nezávislé proměnná pak kategoriální typ objektu a prediktivní proměnná bezpečnost jednotlivých hashovacích funkcí dle ETSI.



Obrázek 23: Regresní analýza kombinovaný Boosted Trees **Zdroj: [autor]**

Z výše uvedených statisticky zpracovaných dat a dosažených výsledků lze konstatovat následující závěry:

1. Typ souboru má zanedbatelný vliv na rychlost výpočtu hashe.
2. Pro rychlost výpočtu hashe je nevýznamně podstatné jak je graf již velký, resp. kolik má úrovní, resp. kolik archiv spravuje objektů.

V této fázi je již možné odpovědět na vědecké hypotézy stanovené v úvodu této práce. Hypotéza I (kapitola 2.5.1) stanovila, že počet archivovaných objektů v systému nebude mít vliv na rychlost výpočtu hashovacích funkcí a tuto hypotézu lze na základě dosažených výsledků **potvrdit**. Odůvodnění: Vzhledem k uvažované statistické chybě predikce chování modelu je patrné, že testovací data od cca 140 stromu jsou již lineární. Hypotéza II (kapitola 2.5.2) stanovila, že nový hashovací

algoritmus bude vždy výpočetně náročnější než starý. Tuto hypotézu lze opět na základě výsledků **potvrdit**. Odůvodnění: na základě dostupných výsledků lze hypotézu potvrdit, ovšem s otazníkem implementace hashovací funkce SHA3 do jazyka Java i vzhledem o tomu, že knihovny Crypto++[109] v. 5.6.0 pro programovací jazyk C++ nemají pro SHA3 ještě vůbec podporu. Co se týká výsledků rychlosti výpočtu funkcí SHA-384 a SHA-512, tyto funkce jsou výpočetně stejně náročné a funkce SHA-384 je pouze oříznutou (zkrácenou) funkcí SHA-512.

[109] http://www.cryptopp.com/benchmarks.html

8 Naplnění cílů práce a její přínosy

Kapitola zhodnocuje dosažené výsledky v předchozích kapitolách s důrazem na oblasti jejich uplatnění. Dále taxativně rozebírá její přínosy jak v doméně teoretické tak praktické.

8.1 Přínosy doktorské disertační práce

V této části jsou v bodech shrnuty věcné přínosy předkládané práce do oblasti počítačové vědy a praktického využití závěrů, ke kterým autor v průběhu vypracování dospěl.

8.1.1 Teoretický (vědecký) přínos

- Obohacení počítačové vědy o systém zaručující bezpečnost libovolného datového objektu bez jeho přímého ovlivnění slábnutím kryptografických algoritmů.
- Přispění do badatelské domény počítačové bezpečnosti mechanismem jednoznačné identifikace dokumentu s vyloučením závislosti na hashovacích funkcích.

8.1.2 Praktický (faktický) přínos

- Univerzálnost řešení hlavního cíle pro budoucí návrhy dlouhodobých archivů.
- Využití jednoznačné identifikace dokumentu také mimo oblast archivace (například pro DMS).
- Díky postavení systému LTA na otevřených standardech byl prezentován prostředek pro nasazení certifikovaných, nezávislých nástrojů třetích stran (např. auditorů).

8.1.3 Původní autorův přínos

Tato podkapitola taxativně v bodech formuluje původní přínosy autora předkládané disertační práce.

- Prezentování obecného, univerzálně aplikovatelného modelu, řešícího změnu hlavního bezpečnostního prvku (hashovací funkce) ve struktuře, jenž zajišťuje integritu archivu.
- Využití funkce DSSC, v oblasti dlouhodobé archivace elektronických dokumentů, pro včasnou detekci možného výskytu kolize výstupu hashovací funkce a tím pádem nedůvěryhodnosti použitého hashovacího algoritmu.
- Navržení obecného identifikátoru datové struktury bez využití hashovací funkce.

8.2 Aplikace navržených řešení

- Implementace navrženého způsobu zajištění integrity archivovaných objektů do systémových návrhů a koncepce dlouhodobých archivů.
- Využití identifikátoru spravovaných dokumentů v prostředí jejich dlouhodobé úschovy a všude tam, kde prezentované řešení není možné.

8.3 Vazba na ostatní práce a projekty.

- Tendr na vybudování národního digitálního archivu ČR (NDA ČR).
- Policie ČR – poptávka po dlouhodobém digitálním archivu pro ukládání logových záznamů ze Schengenského informačního systému druhé generace.
- Vědecká a odborná činnost, z níž vznikly produkty „*eKeeper*“ (Slovinsko, Ljubljana, Jože Štefan Institut, SETCEE) a „GuardTime“ (Estonsko, Tallinn, Cybernetica AS).
- Národní bezpečnostní úřad („*NBÚ*“) – dlouhodobý elektronický archiv dokumentů s různou úrovní bezpečnostní klasifikace.

9 Diskuse a otevřené otázky

- Oblast auditu integrity archivu.

Již bylo v práci zmíněno, že pro kontrolu důvěryhodných dlouhodobých archivů jako celku je nutný certifikovaný nástroj, který bude zodpovědný za ověření jejich integrity. Takovýto exaktní nástroj v době psaní této disertační práce není veřejně dostupný a auditoři digitálních archivů jsou tak odkázání na neexaktní metody kontroly integrity archivu. Pro potřeby a verifikaci výsledků této práce a jejich případného uplatnění je třeba vytvořit takový nástroj, založený na platných standardech, který musí být univerzální a aplikovatelný na libovolnou implementaci archivu. Předkládaná práce svými výstupy umožňuje aplikaci těchto nástrojů a představuje inspiraci pro návaznou badatelskou a vědeckou činnost v této oblasti. Tvorba takovéhoto nástroje, stejně jako hlubší analýza předmětné oblasti je mimo rámec této disertační práce, ale může být vodítkem pro případně téma práce navazující.

- Oblast citlivosti archivovaných informací

Tato problematika souvisí s rozdělováním dokumentů a vlivu „zubu“ času na jejich bezpečnostní klasifikaci (případně podrobnější členění, záleží na politice archivu). Z pohledu systémové bezpečnosti, je toto věc nepodstatná, ale má význam na „hodnotový obsah“ archivu. Konkrétně zdali dokument, podléhající v době archivace *t* absolutnímu utajení kvůli nejvyšší důležitosti (možno kvantifikovat) má v čase například *t+100*, stejnou důležitost jako na počátku, nebo požadavek na utajení časem pominul a blíží se nule (tj. archivujeme něco, co už není tajné). Bohužel se autorovi není do doby psaní této práce známo, zdali existuje analýza přímé souvislosti stárnutí obsahu na jeho důležitost.

- Oblast separátní důkazní struktury (souvisí s předchozím bodem)

Pokud by předchozí paradigma neplatilo, vyvstává otázka, zdali pro každou bezpečnostní kategorii dokumentů nezaložit a udržovat v systému LTA speciální provázané důkazní struktury, které by se při změně hashovacího algoritmu rekonstruovaly (znovu vytvářely) dle stupně bezpečnosti (např. přísně tajné jako první). Došlo by tak k výrazné úspoře času při jejich rekonstrukci v případě využití nové (bezpečné) hashovací funkce. Jistě by ale pro jednotlivé bezpečnostní implementace dlouhodobých archivů platily zcela odlišné nároky na přístup, fyzické umístění atd.

V obecné rovině bylo pro autora předložené disertační práce jednou z mnoha otázek, jak konkrétní má být navržené modelové řešení, aby práce nebyla označena či považována za spíše inženýrskou, nebo řešící ryze inženýrský, či technický problém. Autor se snažil v nejvyšší možné míře zohlednit připomínky z malé obhajoby doktorské disertační práce, konané dne 16. 12. 2013, kde mj. zazněla připomínka o konkretizace implementace funkce DSSC v navrženém modelu. Reakcí na tuto poznámku je podkapitola 6.1.2, stejně tak jako zařazení kapitoly 8.1.3, jež v bodech taxativně vyjmenovává původní autorův přínos v této disertační práci.

10 Závěr

Předložená disertační práce se soustředila na fenomén bezpečnosti dlouhodobé archivace elektronických dokumentů, především pak na integritu jak jednotlivých archivovaných objektů, tak digitálního archivu jako celku. Východiskem pro tuto práci byl předpoklad existence a využití metod relativní časové autentizace.

Práce se ve své úvodní části usiluje o konkretizaci předmětné oblasti vymezením základního odborného pojmového aparátu tak, abych nedocházelo k záměnám pojmů, s nimiž je v této práci operováno s jejich v běžné řeči používaném významu. Je zde také formulován výzkumný problém, jímž je problematika integrity digitálního archivu dále vymezena. Definování výzkumného problému vede k výzkumným otázkám. Jejich zodpovězením se práce dostává k nosné části úvodní pasáže, ve které jsou posléze stanoveny hlavní a dílčí cíle práce a z nich vyplývající úkoly, které je nutné splnit, aby mohlo dojít k dosažení těchto cílů. Formulací nosných prvků disertační práce bylo možné stanovit metody jejich řešení a vyřknout výzkumné hypotézy.

Tato práce jako celek je tvořena třemi stavebními kameny, které jsou reprezentovány částí analytickou, návrhovou a experimentální. V těchto částech práce byly provedeny úkony, vedoucí k dosažení stanovených cílů, vědeckých úkolů a k zodpovězení vymezených výzkumných otázek. V předposlední kapitole jsou v diskusi zmíněny oblasti, které autor považuje za vhodné k dořešení a navázaní na to, čeho bylo v této práci dosaženo. Dále tato kapitola v krátkosti shrnuje reakci na malou obhajobu disertační práce.

Teoreticky orientovaný analytický blok práce je koncipována jako přehledovo-syntetická studie současného stavu problematiky dlouhodobé archivace jak ve světě, tak v ČR. Nedílnou součástí práce je souhrnný pohled na legislativní rámec, jenž především v prostředí EU mapuje implementaci směrnic Evropského společenství tykajících se předmětného tématu do národních legislativ. Další úsek

této části práce se soustřeďuje na projekty, jež svým řešením přispívají k obohacení předmětného tématu. Zde bylo nutné je rozdělit na projekty čistě komerční, většinou založené na proprietálních technologiích a projekty veřejné, dodržujících platné standardy a doporučení. Krátký přehled standardů a norem vztahujících se k dlouhodobé elektronické archivaci uzavírá tuto kapitolu. Tato část práce **splnila první a druhý vědecký úkol** práce, stanovený v kapitole 2.4.3.

Druhá část analytického bloku se pak soustředila především na referenční model OAIS a jeho jednotlivé komponenty. Model OAIS je považován za průkopníka moderního pojetí elektronické archivace, proto je mu v této části práce věnována náležitá pozornost. Definováním workflow ED v dlouhodobém archivu se analyticky orientovaná část dostává k příčinám ztrát průkazních hodnot archivovaných dokumentů a způsobům ochrany během jejich životního cyklu. Zahrnují metody jak na straně původce (elektronický podpis) tak na straně archivu (časové razítko). Současně rozdělují tyto metody na časově relativní (ERS) a časově absolutní (časové razítko). Z tohoto přehledu jednoznačně vyplynulo, že všechny tyto metody využívají rozdílným způsobem hashovacích funkcí jako prostředek, jenž zajišťuje autenticitu daného objektu. Následovala proto kapitola 4.3.7., která mapuje bezkoliznost existujících hashovacích funkcí a jejich vhodnost pro potřeby dlouhodobé archivace elektronických dokumentů.

Závěrečná část analytického celku se věnovala současným přístupům k řetězení výstupů hashovacích funkcí, jenž svojí komplexností a úrovní složitosti přímo ovlivňují jak rychlost výpočtu použité hashovací funkce, tak její zpětné verifikace. Jako nejméně komplikované a zároveň robustní řešení, bylo pro svoji názornost do vědeckého experimentu zvoleno schéma lineární. Analytický blok uzavírá popis datové struktury DSSC, která zajišťuje dohled nad bezkolizností dané hashovací funkce. Předchozí dvě části analytického bloku vedly ke **splnění třetího stanoveného úkolu** práce.

V analytické části práce bylo zjištěno, že v oblasti bezpečnosti dlouhodobé archivace elektronických dokumentů (archivovaných objektů), resp. její podmnožině zabývající se integritou archivu jako systémového celku, neřeší všechny autorovi známé přístupy fakt nutné změny hashovacího algoritmu v čase. Závěry z analytické části práce pomohly stanovit obecný výchozí rámec pro tvorbu systému zabezpečení dlouhodobě archivovaného digitálního objektu a vedly tak je **splnění prvního dílčího cíle** práce.

Část návrhová představuje jádro této práce. Skládá se ze dvou celků, návrhu modelu záměny bezpečnostních struktur v archivu a jednoznačného identifikátoru objektu, který není založen na hashovacích funkci. Model záměny zabezpečovacího prvku implementuje výstupy z DSCC pro včasnou detekci výskytu kolize použité hashovací funkce (s tím souvisejícím zpochybnění její bezpečnosti). Podrobný popis přechodu spolu s detailní implementací DSSC je popsán v kapitole 6.1.1 a 6.1.2 V rámci této části práce došlo ke splnění hlavního cíle, jímž bylo navržení systému zabezpečení, garantujícího nepopiratelnost dlouhodobě uchovaného archivního objektu a zajišťujícího integritu archivu jako celku. Navržením modelu, jenž paralelně spravuje *n* datových struktur (grafů či stromů) a zohledňuje bezpečný přechod od hashovacího algoritmu k němuž by mohla být v daném čase nalezena kolize k novému, dlouhodobě bezpečnému při zachování trvalé integrity archivu jako celku byl **dosažen hlavní cíl** této disertační práce a zároveň **splněn druhý dílčí cíl**. Stanovením jednoznačného identifikátoru archivovaného objektu pak došlo k **naplnění třetího dílčího cíle** této práce.

Experimentální část předložené práce je reprezentována provedením experimentálního ověření části návrhové. Experiment sestává ze dvou složek – samotné fáze provedení experimentu a fáze následného statistického zpracování výstupního souboru tvořícího primární kolekci dat. Pro účel první fáze byla naprogramována aplikace v jazyce Java, jenž představuje zjednodušený model tvorby lineární zřetězené struktury. Archivováno bylo celkem 5686 objektů

tvořených deseti různými typy souborů. Následné statistické zpracování výsledků mělo odpovědět na otázku, zdali typ použité hashovací funkce a typ souboru budou mít vliv na normalizovanou dobu výpočtu. Vyhodnocení testování modelu z návrhu mělo také odpovědět na to, zdali se potvrdí či vyvrátí **hypotézy** (stanovené v kapitole 2.5) stanovené v kapitole 2.5. Z vyhodnocení vyplynulo, že na stanovené hladině významnosti $\alpha = 0{,}05$ se zamítá hypotéza, že použitá hashovací funkce neovlivní rychlost výpočtu a zároveň se potvrdilo, že normalizovaná rychlost výpočtu hashe není závislá na typu archivovaného objektu. Průběh experimentálního ověření včetně provedeného statistického zpracování výsledného datového souboru je podrobně popsán v kapitole 7. Realizováním vědeckého experimentu byl **dosažen čtvrtý úkol** a statistickým zpracováním výsledného datového souboru byl **splněn čtvrtý dílčí cíl práce.**

Vzhledem ke zjištěným skutečnostem lze konstatovat, že aplikace návrhového modelu z kapitoly 6, při dodržení stanovených omezujících podmínek, není ovlivněna typem archivovaných objektů, ale pouze úrovní bezpečnosti, jenž je reprezentována danou hashovací funkcí. Z vyhodnocení experimentu je záhodno akcentovat, že v době psaní této práce nelze nasazení hashovací funkce SHA-3 doporučit, pro její vysokou výpočetní náročnost, neexistenci kryptografickou komunitou uznávaného standardu a pravděpodobnou nízkou optimalizací implementace v programovacím jazyce Java V 1.7. Tento odstavec vede ke **splnění posledního**, **pátého dílčího cíle**.

V návaznosti na výše uvedené interpretace a závěry z provedených analýz a experimentálních úkonů lze formulovat odpovědi na výzkumné otázky položené v úvodu práce (kapitola 2.3.2). Z analytické části práce vyplývá, že elektronický dokument, u něhož lze určit původce (bez ohledu na použitou technologii) má stejné právní účinky jako písemnost podepsaná. Jako příklad lze uvést rozhodnutí Nejvyššího správního soudu České republiky ze dne 17. 2. 2012 vedeným pod č. j. 8 As 89/2011-31 [54]. Tento odstavec **odpovídá na první výzkumnou otázku**.

Odpověď na **druhou výzkumnou otázku** dává kapitola 3.2, jenž analyzuje v současné době existující řešení dlouhodobé archivace využívající ERS. S ohledem na zjištění v kapitole 5, lze formulovat odpovědi na **výzkumné otázky 3, 4 a 5** vztahující se zejména k bezpečnosti digitálních objektů. Lze konstatovat, že digitální objekt, jenž je dlouhodobě archivován, může být z aplikačního pohledu ochráněn proti neoprávněnému porušení integrity, či smazání. Tento stav je ovšem plně závislý též na použité hardwarové instalaci archivu, jenž ovšem není předmětem zkoumání. Odpověď na otázku závislosti bezpečnosti archivovaného objektu na stárnutí použitých kryptografických algoritmů lze nalézt například v kapitole 4.3 a v kapitole 7, kde je tento vztah také předmětem zkoumání. Zjednodušeně lze napsat, že vztah mezi těmito veličinami je, ale lze zajistit, aby nedošlo ke zpochybnitelnosti (prahu) pravosti spravovaného objektu. Na **výzkumné otázky 6 a 7** pak odpovídá návrhová část práce, která jasně prezentuje, že je možné předejít okamžiku zpochybnění nepopiratelnosti a lze zajistit integritu spravovaných objektů.

Z celkového pohledu jsou nedílnou součástí každé disertační práce také oblasti **uplatnění v praxi** s co nejvyšší mírou konkretizace. V případě předmětné práce se jedná o kapitolu 8.2 a 8.3, kde je deklarována možná aplikace autorem navrženého modelu v celé řadě projektů a řešení problematiky v oblasti dlouhodobé archivace elektronických dokumentů.

Hlavním přínosem disertační práce je vytvoření nového autorem navrženého modelu, pomocí nějž lze zajistit permanentní integritu digitálního archivu v čase. Dalšími, vybranými **dílčími přínosy** jsou návrh jednoznačného identifikátoru digitálního objektu a experimentální ověření a vyhodnocení nejnovějšího typu hashovací funkce SHA-3. Přehled všech přínosů práce je v kapitole 8, resp. 8.1.

Disertační práce se vzhledem k absenci ucelené odborné literatury zabývající se dlouhodobou archivací elektronických dokumentů z pohledu bezpečnosti a systémové integrity a také díky komplexnosti předmětné oblasti nemohla věnovat všem jejím aspektům. Je tak ve své podstatě prvním krokem a k dalšímu

podrobnějším a návaznému rozpracování této problematiky, především pak z pohledu relativní časové autentizace. Jako velmi zajímavou oblastí pro budoucí výzkum se jeví například otázka standardizované metodiky auditu dlouhodobých digitálních archivů bez ohledu na skutečně použité technické řešení. Jednou z hlavních operací, jež by takovýto audit zjišťoval, by samozřejmě byla celistvost a neporušenost archivu, resp. archivovaných a archivem spravovaných objektů. Potřebným se též jeví soustředit se na další rozvoj ERS jako důkazního prvku v relativní časové struktuře, která by nebyla závislá především na hashovací funkci, ale na alternativních důkazech unikátního výskytu objektu.

11 Použitá literatura a další informační zdroje

[1] ADAMS Carlisle, Patric CAIN, Denis PINKAS a Robert ZUCCHERATO. *Internet X.509 Public Key Infrastructure Time-Stamp Protocol (TSP)* [online]. Internet Engineering Task Force, The Internet Society, 2001, August 2001 [cit. 2012-08-13]. Dostupné z: http://www.ietf.org/rfc/rfc3161.txt.

[2] ALA (AMERICAN LIBRARY ASSOCIATION). *Definitions of digital preservation* [online]. Chicago: American Library Association, 2007. [cit. 2012-12-10] Dostupné z: http://www.ala.org/ala/mgrps/divs/alcts/resources/preserv/defdigpres0408.pdf

[3] AUDIT AND CERTIFICATION OF TRUSTWORTHY DIGITAL REPOSITORIES. RECOMMENDED PRACTICE. CCSDS 652.0-M-1 [online]. Washington DC: Magenta Book, 2011, September 2011[cit. 2012-08-23]. Dostupné z: http://public.ccsds.org/publications/archive/652x0m1.pdf

[4] BAYER, Dave, Stuart HABER, and W. Scott STORNETTA. *Improving the efficiency and reliability of digital time-stamping* In: R.M. Capocelli, A. DeSantis, and U. Vaccaro, editors, Sequences II: Methods in Communication, Security, and Computer Science. Pages 329-334, New York: Springer-Verlag 1993.

[5] BENALOH, Josh a Michael de MARE. *One-way Accumulators: A Decentralized Alternative to Digital Signatures* in: EUROCRYPT'93, LNCS 765, pages 274-285. Springer-Verlag Berlin Heidelberg. Berlin: 1994. ISBN 3-540-56413-6.

[6] BENALOH, Josh a Michael de MARE. *Efficient broadcast time-stamping*. Technical Report 1. TR-MCS-91-1. Postdam: Clarkson Univ. Dep. of Mathematics and Computer Science, 1991, srpen 1991.

[7] BERNAS, Jiří: *Národní digitální archiv* [online]. Národních archiv ČR, Praha, 2013. Naposledy změněno: 30.9.2013. [cit. 2013-11-12]. Dostupné z: http://www.cnz.cz/ke-stazeni/2013/konference-2013/CNZ-2013Bernas.pdf

[8] BERTONI Guido, Joan DAEMEN, Michaël BEETERS, Gilles VAN ASSCHE a Ronny VAN KEER. *KECCAK implementation overview* [online]. Version 3,2, poslední aktualizace 29. 5. 2012. [cit. 2014-02-12]. Dostupné z: http://keccak.noekeon.org/Keccak-implementation-3.2.pdf

[9] BLAŽIČ Jerman A., Svetlana SALJIC a Terence GODROM. *Extensible Markup Language Evidence Record Syntax (XMLERS)* [online]. RFC 6283. Internet Engineering Task Force (IETF), 2011 [cit. 2012-08-07]. ISSN 2070-1721. Dostupné z: http://tools.ietf.org/html/rfc6283

[10] BLAŽIČ Jerman Aleksej. *Long Term Archiving Implementation – Slovenian Experience with Long Term Archiving*. Conference proceedings, 20th conference,

Czech Open System Users' Group EurOpen.CZ. Plzeň: 2007 [cit. 2012-07-18]. Pages 121-129. ISBN 978-80-86583-12-9

[11] BLAŽIČ JERMAN A. *Long-term Archive Protocol (LTAP)* [online]. Internet Engineering Task Force, 2009, last updated July 13. 2009 [cit. 2012-07-15], Dostupné z: http://tools.ietf.org/html/draft-ietf-ltans-ltap-08

[12] BLIBECH Kaouthar a Alban GABILLON. *A New Timestamping Scheme Based on Skip Lists* [online]. Université de Pau – IUT de Mont de Marsan, France. 2006. [cit. 2013-06-05]. Dostupné z: http://pages.upf.pf/Alban.Gabillon/articles/blibech_final.pdf

[13] BORGHOFF, Uwe M., Jan SCHEFFCZYK, Lothar SCHMITZ a Petr RÖDIG. Long-Term Preservation of Digital Documents. New York: Springer-Verlag, Berlin Heidelberg, 2005 [cit. 2012-07-29]. Vydáno 26. března 2010. ISBN 978-3-540-33639-6

[14] BORKAR Shekhar a Andrew A. CHIEN. *The Future of Microprocessors* [online] In: Communications of the ACM, Vol. 54 No. 5, Pages 67-77. New York: ACM, 2011 [cit. 2012-08-19]. Dostupné z: http://cacm.acm.org/magazines/2011/5/107702-the-future-of-microprocessors/fulltext

[15] BOUTIN Chad. *NIST Selects Winner of Secure Hash Algorithm (SHA-3) Competition* [online]. The National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, 2012. Naposledy aktualizováno: 10. 10 2012. [cit. 2014-01-12]. Dostupné z: http://www.nist.gov/itl/csd/sha-100212.cfm

[16] BSI TR-03125 Preservation of Evidence of Cryptographically Signed Document. Federal Office for Information Security [online]. [cit. 28.11.2012]. Dostupné z: https://www.bsi.bund.de/EN/Publications/TechnicalGuidelines/TR03125/BSITR03125.html

[17] BULDAS, Ahto a Märt SAAREPERA. *On Provably Secure Time-Stamping Schemes* [online] In: Asisacrypt 2004. Tartu: University of Tartu 2004, 9. 12. 2004 [cit. 2012-07-31]. Dostupné z: http://www.iacr.org/conferences/asiacrypt2004/data/Asiacrypt2004/12 Protocols/01_Ahto Buldas.pdf

[18] BULDAS, Atho, Peeter LAUD, Helger LIPMAA a Jan VILLEMSON. *Time-stamping with Binary Linking Schemes. Advances in Cryptology* In: CRYPTO'98, LNCS 1462. Santa Barbara: Springer Lecture Notes in Computer Science, 1998, srpen 1998 [cit. 2012-07-30]. Pages 486-501, vol. 1462

[19] BULDAS, Atho, Helger LIPMAA a Berry SCHOENMAKERS. *Optimally Efficient Accountable Time-Stamping*. LNCS 1751: 293–305. Third International Workshop on Practice and Theory in Public Key Cryptosystems, PKC 2000, Melbourne, Victoria, Australia, January 18-20, 2000. Proceedings. [cit 2012-10-12]. ISBN: 978-3-540-46588-1

[20] BULDAS, Ahto, Peeter LAUD, Märt SAAREPERA a Jan WILLEMSON. *Universally Composable Time-Stamping Schemes with Audit* [online]. Londýn: Lecture Notes in Computer Science, 2005 [cit 2012-08-01]. Volume 3650/2005. Pages 359-373, DOI: 10.1007/11556992_26

[21] BULDAS, Ahto a Sven LAUR. *Do broken hash functions affect the security of time-stamping schemes?* 4th International Conference on Applied Cryptography and Network Security – ACNS'06, LNCS 3989, pp. 50–65. Springer Berlin Heidelberg New York, 2006. [cit. 2013-04-20]. ISBN: 978-3-540-34703-3

[22] BULDAS, Ahto a Aivo JÜRGENSON. *Does Secure Time-Stamping Imply Collision-Free Hash Functions?* First International Conference, ProvSec 2007, Wollongong, Australia, November 1-2, 2007. Pages 138-150. [cit. 2013-06-12]. ISBN: 978-3-540-75669-9

[23] ČSN ISO/IEC 10118-3. *Informační technologie - Bezpečnostní techniky - Hašovací funkce - Část 3: Dedikované hašovací funkce* [online]. Praha: Český normalizační institut, 2004. Dostupné z: http://nahledy.normy.biz/nahled.php?i=61281

[24] ČSN ISO/IEC 17799. *Informační technologie - Bezpečnostní techniky - Soubor postupů pro management bezpečnosti informací*. ICS 35.040 [online]. Praha: Český normalizační institut, 2006, srpen 2006 [cit. 2012-08-15]. Dostupné z: http://csnonlinefirmy.unmz.cz/html_nahledy/36/75901/75901_nahled.htm

[25] CUBR, Ladislav. *Dlouhodobá ochrana digitálních dokumentů*. 1. vydání, 154 stran. Národní knihovna České republiky, Praha, 2010 [cit. 2013-04-23]. ISBN: 978-80-7050-588-5

[26] DOSTÁLEK, Libor a kol. *Velký průvodce infrastrukturou PKI a technologií elektronického podpisu.* Praha: Computer press, 2006, říjen 2006 [cit. 2012-08-02]. 3. doplněné vydání. 536 stran. ISBN: 80-251-0828-7

[27] DOSTÁLEK, Libor a Marta VOHNOUTOVÁ. *Long-term Archive Architecture* [online]. Paříž, Internet Engineering Task force: 2003, listopad 2003 [cit. 2012-07-21]. Dostupné z: http://ltans.edelweb.fr/draft-ietf-ltans-arch-00.pdf

[28] Electronic Signatures and Infrastructures (ESI); Algorithms and Parameters for Secure Electronic Signatures - Part 1: Hash functions and asymmetric algorithms [online]. ETSI TS 102 176-1 V2.1.1 Nice: European Telecommunications Standards Institute, 2011, July 2011 [cit. 2012-08-02]. Dostupné z: http://www.etsi.org/deliver/etsi_ts/102100_102199/10217601/02.01.01_60/ts_10217601v020101p.pdf

[29] Ensuring Record Integrity with AbsoluteProof [online]. Surety, 2010, 8. 4. 2010 [cit. 2012-08-04]. Dostupné po registraci z: http://www.surety.com.

[30] GHAURI, Pervez a Kjell GRONHAUG. *Research Methods in Business Studies A Practical Guide*: *Third Edition*. Londýn: 2005. ISBN: 978-0-273-68156-4.

[31] GLADNEY Henry. *Method and System for Preparing Digital Information For Long-Term Preservation* [online]. Patent Application Publication. United States, 2013. Pub. Date 28. Feb. 2013. Pub. No. US 2013/0054607 A1. Dostupné z: http://www.freepatentsonline.com/y2013/0054607.html

[32] GONDROM Tobias, Ralf BRANDNER a Ulrich PORDESCH. *Evidence Record Syntax (ERS)* [online]. RFC 4998. Internet Engineering Task Force (IETF) Trust, 2007, August 2007 [cit. 2012-08-18]. Dostupné z: http://tools.ietf.org/html/rfc4998

[33] HEBÁK Petr a kolektiv, 2005: *Vícerozměrné statistické metody* [1-3]. Praha: Informatorium, 2004-2005. ISBN: 80-7333-025-3 [1], 80-7333-036-9 [2], 80-7333-039-3 [3]

[34] HESLOP, Helen, Simon DAVIS a Andrew WILSON. *An Approach to the Preservation of Digital Records* [online]. Sydney: National Archives of Australia, 2002, 13. 12. 2002 [cit. 2012-08-05]. Dostupné z: http://www.naa.gov.au/Images/An-approach-Green-Paper_tcm16-47161.pdf

[35] HÖNIG, Petr. *Trustworthy electronic archive for long-term archivation and requirements for their solution* [online]. Liberec: 2007, May 2007 [cit. 2012-07-14]. Dostupné z: http://www.akvs.cz/akp-2007/14-honig.pdf

[36] How a digital timestamp works. DigiStamp, Inc. *DigiStamp* [online]. Dallas, 2012 [cit. 2012-08-10]. Dostupné z: http://www.digistamp.com/technical/how-a-digital-time-stamp-works/

[37] HUTAŘ, Jan. *Digitalizace, popis pomocí metadat a jejich formáty*. Praha, 2012. Univerzita Karlova, Filozofická fakulta, Ústav informačních studií a knihovnictví. [cit. 2014-01-24] Dostupné z: https://is.cuni.cz/webapps/zzp/detail/25756/.

[38] JEŽEK David. *Zasáhne proti agresivní patentové válce Applu sám prezident Obama?* [online]. Praha: CD-R server s.r.o., 2011, 26. 8. 2011 [cit. 2012-08-16]. ISSN 1213-2225. Dostupné z: http://diit.cz/clanek/zasahne-proti-agresivni-patentove-valce-applu-sam-prezident-obama

[39] KELSEY John: *The New SHA3 Hash Functions* [online]. NIST, 2013. [cit. 2014-1-25]. Dostupné z: http://csrc.nist.gov/groups/SMA/ispab/documents/minutes/2013-12/new_sha3_functions.pdf

[40] KING Ros, Rainer SCHMIDT, Christoph BECKER a Mark GUTTENBRUNNER. *Research on Digital Preservation within projects co-funded by the European Union in the ICT programme* [online]. Brusell: Europen Union, 2011, May 2011 [cit. 2012-08-21]. Dostupné z: http://cordis.europa.eu/fp7/ict/telearn-digicult/report-research-digital-preservation_en.pdf

[41] KLÍMA Vlastimil. *Digitální otisk SHA-4 přeskočil SHA-3* v Sdělovací technika, kryptologie pro praxi 05/2012. Praha: Sdělovací technika 2012 [cit. 2012-08-11]. ISSN 1801-2140

[42] KNIJFF, Johan van der. *EPUB for archival preservation* [online]. National Library of the Netherlands, 2012. Naposledy upraveno 20. července 2012. [cit. 2013-02-01]. Dostupné z: http://www.openplanetsfoundation.org/system/files/epubForArchivalPreservation20072012ExternalDistribution.pdf

[43] KLOBUČAR, Toma a Aleksej Jerman BLAŽIČ. *Long-term Trusted Preservation Services Using Service Interaction Protocol and Evidence Record* In: Computer Standards & Interfaces. Volume 29 Issue 3. Ljubljana: 2006, March 2007 [cit. 2012-07-15]. Pages 398-412. doi:10.1016/j.csi.2006.06.004

[44] KUČERA, Roman. *Elektronická archivace elektronických dokumentů* [online]. Praha: První certifikační autorita, 2011, 12. 10. 2011 [cit. 2012-07-22]. Dostupné z: http://www.ica.cz/Userfiles/files/zpravy/bezpecna_archivace.pdf

[45] KUNDER, Maurice De. *The Size of the World Wide Web (The Internet)* [online]. Tilburg: Tilburg Univerity, 2012, 5. 7. 2012 [cit. 2012-07-05]. Dostupné z: http://www.worldwidewebsize.com/

[46] KUNSTOVÁ, Renáta. *Efektivní správa dokumentů. Co nabízí Enterprise Content Management*. Praha: Grada, 2009. ISBN 978-80-247-3257-2

[47] KUNZ Thomas, Susanne OKUNICK a Ulrich PORDESCH. *Data Structure for the Security Suitability of Cryptographic Algorithms (DSSC)* [online]. RFC 5698. Internet Engineering Task Force (IETF), 2009, November 2009 [cit. 2012-08-09]. Dostupné z: http://tools.ietf.org/html/rfc5698

[48] LAPPIN, James. *Why a link between MoReq2010 and the OAIS model would benefit both records managers and archivists* [online]. Londýn: 2012, 13. 7. 2012 [cit. 2012-07-26]. Dostupné z: http://thinkingrecords.co.uk/2012/07/13/why-a-link-between-moreq2010-and-the-oais-model-would-benefit-both-records-managers-and-archivists/

[49] LEE, Christopher A. *Open Archival Information System (OAIS) Reference Model* [online]. Chapel Hill: University of North Carolina, 2010 [cit. 2012-07-22]. Dostupné z: http://ils.unc.edu/callee/p4020-lee.pdf

[50] Long-Term and Independent Proof of Record Integrity [online]. Prove Ownership of Your Digital IP, Electronic Notary, Digital Notary. Surety, LLC, 2012 [cit. 2012-08-04]. Dostupné z: http://www.surety.com/digital-copyright-protection/prove-ownership.aspx

[51] NEWBOLD Paul, William L. CARLSON a Betty THORNE. *Statistics for Business and Economics: Seventh Edition*. Londýn: Pearson, 2010. ISBN: 978-0-13-507248-6

[52] PARR, Ben. Google: *There Are 129,864,880 Books in the Entire World* [online]. New York: Mashable Inc. 2012, 6. 8. 2012 [cit. 2012-07-04]. Dostupné z: http://mashable.com/2010/08/06/number-of-books-in-the-world/

[53] PETERKA, Jiří. *Nástroje prokazování platnosti dokumentů: dlouhodobě ověřitelné elektronické podpisy a časová razítka* [online]. 2011, 24. 5. 2011 [cit. 2012-07-21] Dostupné z: http://earchiv.cz/papers/p55/slide.php3?l=1&me=1

[54] PROŠEK David, PIRÁTI.*cz*: *Nejvyšší správní soud potvrdil stížnost Pirátů proti Kalouskovi* [online]. eGov.cz nezávislý informační portál. Naposledy upraveno: pátek, 9. března 2012 07:51. [cit. 2014-05-12]. Dostupné z: http://www.egov.cz/index.php?option=com_content&view=article&id=378:nejvyi-spravni-soud-potvrdil-stinost-pirat-proti-kalouskovi&catid=9:datove-schranky&Itemid=3

[55] *Průlom v elektronické archivaci dokumentů* [online]. Průlom v elektronické archivaci dokumentů | CFOwolrd.cz. Praha: IDG, 2012, 11. 7 2012 [cit. 2012-08-25]. Dostupné z: http://cfoworld.cz/ostatni/prulom-v-elektronicke-archivaci-dokumentu-1785

[56] PUGH, William. *A Skip List Cookbook* [online]. University of Maryland, College Park, červenec 1989, revisited 1990, issue date 15. říjen 1998. CS-TR-2286.1 UMIACS; UMIACS-TR-89-72.1. [cit. 2013-05-26] Dostupné z: http://cg.scs.carleton.ca/~morin/teaching/5408/refs/p90b.pdf

[57] RABINOVICI-COHEN, Simona, John Marberg a Kenneth Nagin. *Preservation DataStores in the Cloud (PDS in the Cloud): Longterm Digital Preservation in the Cloud* [online]. IBM Research Division. Haifa: 20. leden 2013. H 0318 (HAI 301-006). Dostupné z: http://domino.research.ibm.com/library/cyberdig.nsf/papers/F1A0F33F7354D53B85257B01004E9AEF/$File/h-0318.pdf

[58] REFERENCE MODEL FOR AN OPEN ARCHIVAL INFORMATION SYSTEM (OAIS). RECOMMENDED PRACTICE. CCSDS 650.0-M-2 [online]. Washington DC: Magenta Book, 2012, July 2012 [cit. 2012-08-22]. Dostupné z: http://public.ccsds.org/publications/archive/650x0m2.pdf

[59] REQUIREMENTS FOR BODIES PROVIDING AUDIT AND CERTIFICATION OF CANDIDATE TRUSTWORTHY DIGITAL REPOSITORIES. RECOMMENDED PRACTICE. CCSDS 652.1-M-1 [online]. Washington DC:Magenta Book, 2011, November 2011. [cit. 2012-08-24]. Dostupné z: http://public.ccsds.org/publications/archive/652x1m1.pdf

[60] SAUNDERS Mark, Philip LEWIS a Adrian THORNHILL. *Research Methods for Business Students: Fifth Edition*. Londýn: Person, 2009. ISBN: 978-0-273-71686-0.

[61] SIMONITE Tom. *Math Advances Raise the Prospect of an Internet Security Crisis* [online]. MIT Technology Review, Cambridge, MA, 2013. Poslední aktualizace 2. srpna 2013, [cit. 2013-08-20]. Dostupné z: http://www.technologyreview.com/news/517781/math-advances-raise-the-prospect-of-an-internet-security-crisis/

[62] SMORUL Michael, Sangchul SONG a Joseph JAJA. *An Implementation of the Audit Control Environment (ACE) to Support the Long Term Integrity of Digital Archives* [online]. Chapel Hill: Proceedings of DigCCurr2009 Digital Curation: Practice, Promise and Prospects, 2009 [cit. 2012-08-03]. University of North Carolina. Dostupné z: http://www.umiacs.umd.edu/publications/implementation-audit-control-environment-ace-support-long-term-integrity-digital

[63] SØRENSEN, Jan Dalsten. *A national standard for digital records -- a Danish perspective* [online]. Publikováno 18. 4. 2012. Dostupné z: http://www.youtube.com/watch?v=A2Wx6WDTCbI

[64] SPEIRS William Robert. *Dynamic Cryptographic Hash Functions*. West Lafayette: Purdue University, Proquest, 2007 [cit. 2012-08-20]. 131 stran. ISBN: 978-054-916-6740.

[65] STRODL, Stephan, Christoph BECKER, Robert NEUMAYER a Andreas RAUBER. *How to Choose a Digital Preservation Strategy: Evaluating a Preservation Planning Procedure*. [online] Vienna: Vienna University of Technology, 2010 [cit. 2012-07-21]. Dostupné z: http://www.ifs.tuwien.ac.at/~strodl/paper/FP060-strodl.pdf

[66] STYBLIŃSKA, Maria. *Long-Term Preservation of Digital Assets – Some Specific Aspects* [online]. 2006. [cit. 2012-07-20]. Dostupné z: http://www.proceedings2006.imcsit.org/pliks/149.pdf

[67] Trusted Digital Repositories: Attributes and Responsibilities, An RLG-OCLC Report [online]. Dublin: Research Libraries Group, 2002, May 2002 [cit. 2012-08-14]. Dostupné z: http://www.oclc.org/resources/research/activities/trustedrep/repositories.pdf

[68] VERHEUL, Ingeborg. *Networking for Digital Preservation: Current Practice in 15 National Libraries*. Munchen: De Gruyter Saur, Germany, 2006. 269 stran. [cit. 2013-01-22]. ISBN: 978-3-598-21847-7

[69] VOHNOUT, Rudolf. *Thin Client Usage in Long-term Archivation Environment*. Ekonomie a Management, Liberec: 2008, roč. XI., č. 3, s. 127–131. ISSN: 1212-3609

[70] VOHNOUT, Rudolf. *Zajištění integrity systému dlouhodobé archivace.* Praha: Vysoká škola ekonomicka v Praze, Oeconomica, 2012 [cit. 2012-08-25]. Sborník

prací účastníků vědeckého semináře doktorského studia FIS, VŠE. Strany 31–32. ISBN 978-80-245-1862-6.

[71] VOHNOUTOVÁ, Marta a Rudolf VOHNOUT. *Zajištění integrity archivu, část I. Provázané haše a jejich využití v digitálních archivech*. DSM - Data Security Management. © TATE International, s.r.o., 2012, číslo 4/2012, s. 40-45, ISSN 1211-8737.

[72] VOHNOUTOVÁ, Marta a Rudolf VOHNOUT. *Zajištění integrity archivu, část II*. DSM - Data Security Management. © TATE International, s.r.o., 2013, číslo 1/2013, ISSN 1211-8737.

[73] VRÁBEL, Tomáš. *Implementace dlouhodobého elektronického podpisu*. [online]. Praha, 2008 [cit. 2012-07-26]. Dostupné z: http://digitool.is.cuni.cz/R/-?func=dbin-jump-full&object_id=829689&silo_library=GEN01. Diplomová práce. Univerzita Karlova v Praze. Vedoucí práce Libor Dostálek.

[74] WALLACE, Carl. Using the Server-Based Certificate Validation Protocol (SCVP) to Convey Long-Term Evidence Records [online]. RFC 5276. Internet Engineering Task Force (IETF), 2008 [cit. 2012-08-08]. Dostupné z: http://tools.ietf.org/html/rfc5276

[75] What is Digital Preservation? Digital Preservation Europe, 2006. Copyright © 2006 DPE. Naposledy změněno: 28. dubna 2006. [cit. 2012-12-11]. Dostupné z: http://www.digitalpreservationeurope.eu/what-is-digital-preservation/

[76] WINN, Joss. *Orbital and the OAIS reference model In: Orbital and the OAIS reference model | Orbital* [online]. Lincoln: 2012, 10. 2. 2012 [cit. 2012-07-17]. Dostupné z: http://orbital.blogs.lincoln.ac.uk/2012/02/10/orbital-and-the-oais-reference-model/

[77] Xena – Digital Preservation Software [online]. Sydney: National Archives of Australia, 2012 [cit. 2012-08-03]. Dostupné z: http://xena.sourceforge.net/index.php

[78] Zásady dlouhodobé archivace elektronických dokumentů [online]. Naposledy změněno 12. 6. 2012. [cit. 2013-02-15-05]. Dostupné z: http://www.602.cz/newsletter-dlouhodoba-archivace-lta

12 Seznam použitých zkratek a termínů

ACE = Audit Control Environment

AIP = Archival Information Package

ANOVA = Analysis of Variance

ANSI = American Nation Standards Institute

CAdES = CMS Advanced Electronic Signatures

CMS = Content Management System

CRL = Certification Revocation List

DC = Dublin Core

DIP = Dissemination Information Package

DLM = Document Management Lifecycle

DMS = Document Management System

DRM = Document Rights Management

DRAMBORA = Digital Repository Audit Method Based on Risk Assessment

ES = Electronic Signature

DSSC = Data Structure for the Security Suitability of Cryptographic Algorithms

ECM = Enterprise Content Management

ED = Electronic Document

EU = European Union

EDRMS = Electronic Document and Records Management System

ERMS = Electronic Record management System.

ERS = Evidence Record Syntax.

ETSI = European Telecommunications Standards Institute

FIS = Fakulta informatiky a statistiky

IDPF = International Digital Publishing Forum

ISAD(G) = General International Standard Archival Description

KSI = Keyless Signature Infrastructure

LMER = Long-term preservation Metadata for Electronic Resources

LTA = Long-Term Archivation = Dlouhodobá archivace

LTAP = Long-term Archive Protocol

METS = Metadata Encoding & Transmission Standard

METS = Metadata Encoding and Transmission Standard

MoReq = Model requirements for Electronic Records Management

NBÚ = Národní bezpečnostní úřad

NEDLIB = Networked European Deposit Library

NIST = National Institute of Standard and Technology

OAI-ORE = Open Archives Initiative Object Reuse and Exchange

OAI-PMH = Open Archives Initiative Protocol for Metadata Harvesting

OAIS = Open Archival Information System

PAdES = PDF Advanced Electronic Signatures

PDF = Portable Document Format

PDI = Preservation Description Information

PKI = Public Key Infrastructure

PREMIS = Preservation Metadata Implementation Strategies

PTAB = The Primary Trustworthy Digital Repository Authorisation Body

RIPEMD = RACE Integrity Primitives Evaluation Message Digest

RTA = Relative Temporal Authentication.

RXP = Repository Exchange Package

SCVP = Server-based Certificate Validation Protocol

SHA = Secure Hash Algorithm

SIP = Submission Information Package

TAA = Trustworthy Archival Authority

TEI = Text Encoding Initiative

TDO = Trustworthy Digital Object

TRAC/TDR = Trustworthy Repositories Audit & Certification/ Trustworthy Digital Repository Checklist

TS = Time Stamp; Technical Specification

(T)TSA = (Trusted) Time Stamp(ing) Authority

UCLA = University of California v Los Angeles

UETA = Uniform Electronic Transaction Act

UML = Unified Modelling Language

UTC = Coordinated Universal Time

VEO = VERS Encapsulated Object

VERS = Victorian Electronic Records Strategy

XAdES = XML Advanced Electronic Signatures

XENA = XML Electronic Normalising for Archives

XML = eXtensible Markup Language

XMLERS = Extensible Markup Language Evidence Record Syntax

13 Seznam obrázků

14 Seznam tabulek

15 Seznam vzorců, rovnic a nerovnic

16 Přílohy

Příloha A

Pro demonstraci navrženého identifikátoru objektu byl vybrán následující vstup:

Objekt[1]: Main_text-small_defense-v15.pdf

Typ objektu: PDF

Velikost objektu (B): 2 667 256

Identifikace osoby (s1): rvohnout000000000000000000000000

Identifikace stroje (s2):
00217 0EF223654E6FCA8C4A016021712913600000000000000000000

Čas (s3): 2014052022164700

První dokument (s4): 1

Pořadové číslo (s5): 0000000

Pole (s6): 462EB50D32663130202FAB7331A75041F1F905FB00000000000000000000

Důkaz předchozího dokumentu (s7):
0000000000000000000000000000000000000000000000000000000000000000

Pole (s8): 00000000000000000000

Výsledný identifikátor:

rvohnout000000000000000000000000002170EF223654E6FCA8C4A0160217129136000000
00000000000000201405202216470010000000462EB50D32663130202FAB7331A75041F1F
905FB00000000000000000000000000000000000000000000000000000000000000000000
000000000000000000000000000000000000

[1] V tomto případě dokument s embedded metadaty