Proč jsou české digitální repozitáře nespolehlivé?

Jan Hutař / Národní knihovna ČR – oddělení archivace webu / e-mail: jan.hutar@nkp.cz

Úvod: Digitální repozitář a jeho důvěryhodnost

V rámci repozitáře, zvláště pokud se jedná o certifikovaný důvěryhodný repozitář, se počítá s tím, že informace v něm budou uloženy velmi dlouho, přičemž by měly být stále v použitelné, tedy v čitelné podobě, kterou je možno zpřístupnit uživatelům. Vyhovět tomuto cíli není v době překotného vývoje technologií vůbec jednoduché. Technologie se vyvíjejí obrovským tempem, což přináší nejen zastarávání hardwaru, ale hlavně zastarávání softwaru a formátů. Je téměř jisté, že formáty jakýchkoli digitálních dokumentů (text, obraz, video, audio), které jsou dnes rozšířeny a obecně využívány (.doc, .pdf, .mp3, .avi, .jpg apod.), budou v blízké budoucnosti nahrazeny formáty jinými, takže dříve či později může být zpřístupnění současných digitálních dokumentů vážným nebo dokonce neřešitelným problémem.

Nemluvíme zde o něčem, co se teprve stane, ale o něčem, co se již dlouhou dobu děje. Pokud bychom šli do extrému v oblasti hardwaru, můžeme za takovou změnu považovat náhradu děrných štítků magnetickými páskami, později pevnými paměťmi, optickými disky apod. Co si kdo dnes počne s děrnými štítky, pokud by z nich chtěl interpretovat nějakou informaci? Nemusíme ani chodit tak daleko do minulosti, problémem dnes mohou být i obyčejné diskety, které jsme před několika lety používali denně a které jsou dnes vytlačeny flashovými paměťmi. Na současných počítačích již disketová mechanika zcela chybí.

Totéž platí i pro softwarové aplikace. Ačkoliv se to nezdá, k takové markantní proměně softwaru již došlo, například v oblasti textových editorů. Kdo si dnes vzpomene na textový editor AmiPro, běžící na počítačích Amiga a využívaný ještě před nástupem MS Office a jeho Wordu? To samé se týká obrazových digitálních dokumentů procházejících ještě bouřlivějším vývojem, který žene kupředu snaha o co největší komprimaci při co nejlepším zachování kvality.

Zatímco u klasických dokumentů je relativně snadné zajistit jejich autenticitu a jejich ohrožení lze zjistit pouhým okem při prohlídce skladiště, u digitálních dokumentů je obojí podstatně složitější, stejně jako jejich zajištění proti neoprávněnému užití. Degradace papírových nosičů informací je pomalá, ztráty v digitálním světě jsou naopak rychlé, nevratné a ne vždy snadno a včas zachytitelné. Proto je v poslední době v mezinárodním kontextu věnována stále větší pozornost důvěryhodným digitálním repozitářům[1].

Správa vlastních digitálních dokumentů i souvisejících metadat je složitý, permanentní a nákladný proces. Novou a prozatím málo prozkoumanou a zdokumentovanou oblastí je plánování ochrany digitálních dokumentů s ohledem na rychlý proces stárnutí a změn těchto dokumentů, resp. jejich formátů (nutnost včasné migrace). Stárne ovšem i HW a SW potřebný pro jejich zpřístupnění (následná nutnost emulace v případě potřeby zachování nejen obsahu, ale i nosiče a celkového prostředí, tzv. look and feel).

Co vlastně má označovat slovo „*důvěryhodný*“? Označuje skutečnost, že podle konkrétních kritérií byla uznána (prokázána) schopnost repozitáře zachovat digitální dokumenty v dlouhodobém horizontu přístupné a použitelné. Tedy repozitář musí být navenek důvěryhodný, což se netýká pouze technického řešení, právě naopak. Jde o to, jak je zajištěno např. financování, v jaké instituci a v jakých podmínkách je repozitář provozován, jak schopný má instituce personál, jaké má zabezpečení apod. (více viz níže).

[1] STOKLASOVÁ, Bohdana; HUTAŘ, Jan. Nové směry v dlouhodobém uchovávání dokumentů ... s. 87.

Obecné problémy s dlouhodobou archivací digitálních objektů[2]

1. Archivovaná informace musí být použitelná uživateli, kteří jsou vzdáleni v prostoru i čase a nemají podporu producenta informace.

- Může se stát, že producent informace již neexistuje, v tom případě nemůže odpovědět na žádné otázky. Jediné co máme, je digitální objekt a jeho metadata.
- SW, na kterém byla informace (digitální objekt) vytvořena, již nemusí být podporován. Může se tedy stát, že informace zaznamenané tímto SW jsou zcela nedostupné (nikdo např. nezná jejich kódování a dokumentace se během let ztratila). Nemusí být také vůbec jasné, zda jde o text, obraz nebo zvuk.

2. Uživatelská komunita se bude během doby měnit.

- Nová komunita nebude znát pozadí vzniku konkrétních informací (vznik, SW, HW, účel …).
- Může chtít kombinovat informace z mnoha zdrojů.
- Nová komunita může používat naprosto odlišné pracovní prostředí.

3. Archiv se bude měnit během doby.

- Může dojít k přechodu na nové technologie, což vždy souvisí s reorganizací informací v archivu a ta zase s možnými změnami ve vztazích mezi informacemi.
- Může také dojít k jeho přenesení do jiných institucí, tzn. možné změny v managementu, datové struktuře i formátu souborů.

Základní funkce důvěryhodného digitálního repozitáře

Máme-li hledat odpověď na otázku, proč jsou vybudování a provoz důvěryhodného digitálního repozitáře finančně i personálně tak náročné, že často překračují možnosti jednotlivých institucí, je třeba pochopit jeho základní funkce. V úplnosti je mapuje (a do detailů rozvíjí) referenční model OAIS[3] .

Digitální repozitář je definován jako organizace, která uchovává informace s cílem jejich zpřístupnění a využití. Nejde tedy o prostý sklad digitálních informací, ale o komplex funkcí podobný funkcím klasické knihovny, která získává (přijímá) dokumenty od producentů, zajišťuje jejich popis, uložení a správu ve skladišti a jejich zpřístupnění uživatelům, kteří si mohou informace o dokumentech vyhledat prostřednictvím katalogů nebo portálů a následně si je v knihovně vypůjčit.

Digitální repozitář rovněž přijímá dokumenty (digitální) od producentů, zpravidla včetně dohodnutých metadat (soubor informací pro dodávání), tzv. SIP balíček. Výsledkem následné kontroly digitálních dokumentů a příslušných metadat a obohacení metadat (zde nehovoříme jen o metadatech popisných, ale i administrativních, technických a ochranných) je soubor informací pro archivaci – AIP balíček, při požadavku na zpřístupnění informací z repozitáře je generován soubor informací pro šíření – DIP balíček.

Kritéria hodnocení důvěryhodnosti digitálního repozitáře

Kritéria hodnocení důvěryhodnosti byla původně obsažena ve dvou základních materiálech:
- Trusted Digital Repositories : Attributes and Responsibilities (RLG, OCLC, 2002)
- Trustworthy Repositories Audit & Certification : Criteria and Checklist (OCLC, CRL, 2007).

Starší z obou dokumentů vydaný v roce 2002 definuje obecné vlastnosti důvěryhodného repozitáře následovně. Důvěryhodný repozitář musí:

[2] GIARETTA, David. ISO/CCSDS Open Archival Information System (OAIS) …
[3] STOKLASOVÁ, Bohdana; HUTAŘ, Jan. Nové směry v dlouhodobém uchovávání dokumentů ... s. 87.

- Přijmout odpovědnost za dlouhodobou péči o svěřené digitální dokumenty a za jejich zpřístupnění současným i budoucím uživatelům.
- Organizačně zajistit dlouhodobou životnost nejen pro vlastní repozitář, ale i pro svěřené digitální informace.
- Prokázat finanční zajištění v současnosti i trvale udržitelný rozvoj.
- Navrhnout systém pro správu digitálního repozitáře v souladu s obecně platnými konvencemi a standardy v zájmu zaručení trvalé správy, zpřístupnění a zabezpečení uložených digitálních dokumentů.
- Stanovit metodiku hodnocení důvěryhodnosti systému.
- Jasně a srozumitelně prezentovat svoji odpovědnost za dlouhodobou ochranu a zpřístupnění dokumentů uživatelům i subjektům, které své dokumenty v repozitáři ukládají.
- Disponovat strategií, pracovními postupy a službami, které umožňují snadné hodnocení a měření.

Navazující dokument Trustworthy Repositories Audit & Certification : Criteria and Checklist již představuje velmi detailní a propracovaný systém kritérií pro hodnocení repozitáře, který se rozpadá do tří hlavních sekcí:
1. Organizace (řízení, struktura, udržitelnost, finance)
2. Správa digitálních objektů
3. Technologie, technická infrastruktura, bezpečnost

V jednotlivých sekcích jsou sledovány okruhy, v jejichž rámci je třeba zodpovědět poměrně konkrétní otázky. Z uvedeného výčtu je zřejmé, že pro vybudování a provoz důvěryhodného repozitáře zdaleka nestačí zakoupení drahého technického a programového vybavení, právě naopak. Jedná se o složitý a dlouhodobý proces, který musí být řádně zakotven ve strategických prioritách i organizační struktuře instituce, která aspiruje na vybudování a provoz důvěryhodného digitálního repozitáře. Musí být adekvátně finančně a personálně zajištěn nejen po dobu vzniku repozitáře a po krátkou dobu po něm, ale dlouhodobě[4].

Metody certifikace a auditu důvěryhodných digitálních repozitářů:

1) **DRAMBORA (Digital Repository Audit Method Based on Risk Assessment)**
- viz dále

2) **Trustworthy Repositories Audit & Certification (TRAC) : Criteria and Checklist**

- zdůrazňuje kritéria doporučeného postupu (best practice) pro důvěryhodné repozitáře ve třech různých oblastech,
- bere OAIS za své východisko a vztažný bod pro úspěšné porovnávání,
- aspiruje na to stát se standardem,
- klade důraz na proces certifikace,
- certifikace probíhá na objednávku, třetí nezávislou stranou (externím auditorem),
- proces je placený a velmi nákladný – zatím takto proběhlo několik auditů především v USA, dále v Nizozemsku a na Novém Zélandě,
- ideální je navázat na interní audit (DRAMBORA).

3) **NESTOR Criteria Catalogue**
- obsahuje 14 kritérií doplněných detailním vysvětlením a konkrétními příklady,
- dělí celou problematiku do skupin,
--- organizační rámec,
--- správa digitálních objektů,
--- infrastruktura a zabezpečení,
- odráží německý kontext (právní, finanční apod.).

[4] STOKLASOVÁ, Bohdana; HUTAŘ, Jan. Nové směry v dlouhodobém uchovávání dokumentů ... s. 87.

4) **International Audit and Certification Birds of a Feather Group**
- mezinárodní snaha o vytvoření jednotného ISO standardu, podle kterého by se v budoucnu prováděl kompletní audit a certifikace repozitářů,
- pravděpodobně zároveň s certifikací OAIS,
- snaží se integrovat již existující nástroje (viz výše) do jednoho dokumentu nebo souboru propojených dokumentů (vytvářejí různá srovnání, crosswalks apod.).

Na základě těchto nástrojů lze definovat deset základních principů důvěryhodnosti, které jsou jakýmsi jejich průnikem[5]. Mohou být prvotním východiskem, než se opravdu přistoupí k serióznímu auditu nebo procesu certifikace. Těchto deset bodů můžeme považovat za desatero jakéhokoliv repozitáře, který chce být považován za důvěryhodný. Takový archiv:

1. se musí zavázat k neustálému opatrování/správě digitálních objektů pro určitou cílovou komunitu,
2. musí prokázat svou životaschopnost/způsobilost (včetně financování, personálních otázek, struktury, procesů), aby dostál stanovenému závazku,
3. musí si osvojit a dodržovat potřebná smluvní a zákonná práva a dostát všem z nich plynoucím závazkům,
4. musí mít efektivní a dostačující rámcovou strategii,
5. získává a ukládá digitální objekty na základě stanovených kritérií, které odpovídají cílům a schopnostem instituce,
6. neustále udržuje/zajišťuje integritu, autenticitu a využitelnost digitálních objektů, které trvale uchovává,
7. vytváří a uchovává potřebná metadata o událostech souvisejících s uloženými digitálními objekty v průběhu jejich uchování, jakož i metadata o samotném vytvoření digitálních objektů, podmínkách zpřístupnění a kontextu využití digitálních objektů,
8. musí splnit nezbytné požadavky na zpřístupnění objektů z repozitáře určité komunitě,
9. musí mít strategii pro plánování ochrany a souvisejících procesů včetně záchranných prací,
10. musí mít technickou infrastrukturu adekvátní účelu neustálé údržby a zajištění digitálních objektů.

Důvěryhodnost je klíčovou vlastností, kterou musí certifikovaný repozitář demonstrovat a doložit. Ochrana digitálních dat je v podstatě uvědomování si organizačních, procedurálních, technologických a jiných nejistot a jejich přeměna na zvladatelná (řešitelná) rizika. *K získání takové důvěryhodnosti má napomoci právě nezávislý audit*.

Audit by měl pomoci:

- rozpoznat a stanovit prioritní (největší) rizika, která ohrožují jeho aktivity,
- vypořádat se s riziky tak, aby se snížila možnost jejich výskytu,
- stanovit možné nepředvídatelné události, aby se snížil efekt případného rizika; pak je repozitář velmi pravděpodobně připraven obdržet status důvěryhodnosti..

[5] CRL Core Requirements for Digital Archives ...

Interní audit a nástroj DRAMBORA[6]

První verze nástroje DRAMBORA (*Digital Repository Audit Method Based on Risk Assessment*) vznikla počátkem roku 2007 ve spolupráci s Digital Curation Centre (DCC[7]) a projektu DigitalPreservationEurope (DPE[8]), kterého se účastní i Národní knihovna ČR. V dalším textu jsou popsány jednotlivé kroky procesu interní certifikace pomocí nástroje DRAMBORA verze 1. Nástroj má záměrně dosti styčných bodů s certifikací TRAC i NESTOR (viz výše). Tvoří jej velmi podrobný návod a šablony k použití.

DRAMBORA nevznikla jako další certifikační nástroj, ale jako nástroj, který by měl pomoci instituci plánující certifikaci svého repozitáře. Pomoci tím, že instituce provede důkladný interní audit a sama odhalí slabiny i silné stránky svého repozitáře. Nejlepší postup je, že audit udělá někdo (např. pracovník konkrétní instituce), kdo má odpovídající znalosti, odborné školení, nebo někdo, kdo již takový audit dělal, třeba i s využitím nástroje DRAMBORA. Tento „samoodhad" může do velké míry snížit náklady na následnou externí certifikaci, která je placená (např. pro TRAC se hovoří o částkách v desítkách tisíc dolarů), a také ji může významně urychlit.

Interní audit je důležitý zvláště v situaci, kdy se o repozitářích velmi mluví, řada institucí si je zřizuje a začíná provozovat. Většina dnešních repozitářů se však nachází (měřeno parametry lidského věku) v raném dětském období a prodělává první vážné dětské nemoci, v lepším případě pak ve stadiu před pubertou. DRAMBORA je strukturována jako odpověď pro další vývoj repozitářů a hlavně pro jejich funkčnost v nedaleké i vzdálenější budoucnosti.

Struktura celého procesu auditu

DRAMBORA má **6 fází**, které obsahují celkem **10 úkolů**:

1. identifikace východisek (kontext organizace/instituce)
2. identifikace strategie a regulačního rámce
3. identifikace aktivit, prostředků a jejich „vlastníků" [10]
4. identifikace rizik souvisejících s aktivitami a prostředky
5. vyhodnocení rizik
6. zvládnutí rizik

Jak je vidět ze základní struktury, audit je v podstatě „risk management". Výhoda nástroje DRAMBORA je i v tom, že se audit může po určité době opakovat. DRAMBORA vyjadřuje rizika možná pomocí číselných hodnot, což je ideální pro případ, že si instituce udělá první audit svého repozitáře, odhalí jeho slabiny, vyhodnotí rizika, jejich možné následky a přijme opatření typu „zlepšit to a ono do dvou měsíců" a následně za tyto dva měsíce může nechat repozitář projít znovu celým procesem a výsledky obou auditů porovnat. Viz obrázek na další straně[11].

[6] http://www.repositoryaudit.eu/

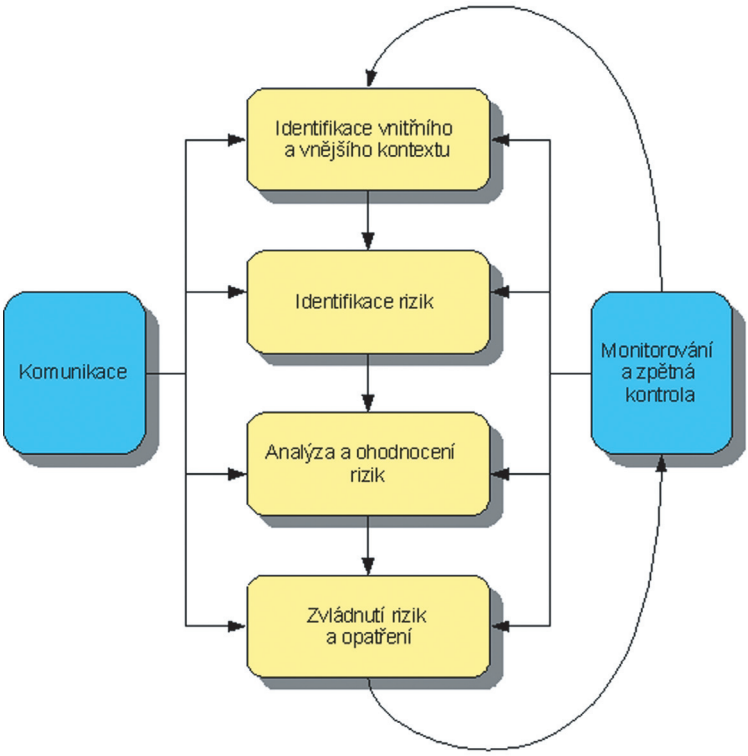
[7] http://www.dcc.ac.uk/

[8] http://www.digitalpreservationeurope.eu/

[9] Základ tohoto textu vznikl v prosinci 2007, DRAMBORA od té doby dospěla do verze 2 (duben 2008), ovšem celý proces auditu ve verzi 2 se liší od první verze minimálně. Hlavním rozdílem je to, že verze 2 je k dispozici jako interaktivní online nástroj a ne pouze jako návod (verze 1). Následující text tak může být ideálním pomocníkem i pro druhou verzi nástroje DRAMBORA.

[10] Jde o vlastníky v tom smyslu, že mají s aktivitou/prostředky něco do činění, jsou za ni např. odpovědní.

[11] ROSS, S. et al. Building Trust in Digital Repositories Using DRAMBORA …



Zmíněných 10 úkolů v rámci šesti fází není rozmístěno rovnoměrně. Nejvíce úkolů se objevuje v první (2) a v druhé fázi (4), ostatní fáze obsahují vždy pouze jeden úkol.

Každý z úkolů (tasks) je vypracováván se zřetelem na **8 funkčních tříd** z hlediska funkčnosti repozitáře. Tyto funkční třídy se dělí na **procesní** (4) a **podpůrné** (4). Ve výsledku tedy dostáváme rizika, která jsou rozdělena podle těchto tříd. Jsou tak daleko přehlednější a dá se na ně lépe reagovat konkrétními nápravnými opatřeními.

Procesní funkční třídy:
- akvizice a ingest (příjem dokumentů do repozitáře)
- ochrana a uložení
- management metadat
- přístup a šíření

Podpůrné funkční třídy:
- organizace a management
- personální otázky
- finanční management
- technická infrastruktura a bezpečnost

Fáze 1

úkol 1. Specifikujte mandát/poslání vašeho repozitáře nebo organizace, kde je provozován.
- např. proč instituce vůbec existuje, co je její primární úkol apod.

úkol 2. Vyjmenujte cíle a účel repozitáře/organizace.

Fáze 2

úkol 3. Vyjmenujte dokumenty, které se týkají strategického plánování vašeho repozitáře.
- různé strategie, manuály, koncepce apod.

úkol 4. Uveďte všechny právní, smluvní a regulační rámce nebo dohody, které se byť jen okrajově dotýkají vašeho repozitáře.
- smlouvy s třetí stranou, zákony – např. zákoník práce, knihovní zákon, povinná nařízení, autorský zákon, zákon o povinném výtisku, normy apod.

úkol 5. Vyjmenujte dobrovolná nařízení, standardy a manuály, ke kterým se repozitář hlásí a řídí se podle nich.
- využívané postupy – nekodifikované např. normou, vnitřní nařízení atd.

úkol 6. Vyjmenujte ostatní dokumenty a principy, se kterými je repozitář „ve shodě“. Tj. jakékoliv jiné dokumenty a skutečnosti, které ovlivňují, jak repozitář dělá to, co by dělat měl.

Fáze 3

úkol 7. Vyjmenujte všechny aktivity a prostředky potřebné k jejich realizaci a jejich „majitele“ (osoby odpovědné). Tj. jaké jsou aktivity, které v repozitáři probíhají, aby plnil svůj cíl a účel v rámci vnějšího kontextu, a jaké prostředky využívá pro zabezpečení těchto aktivit, včetně personálních otázek, zkušeností, znalostí a technologií.
- vznikne koncepční model toho, co repozitář zajišťuje,
- z obecného popisu se dostáváme ke konkrétní analýze aktivit a pracovních procesů,
- hierarchická nebo procesní analýza.

Fáze 4

úkol 8. Identifikujte rizika spojená s aktivitami a prostředky vašeho repozitáře.
- jde v podstatě o aktivity a procesy z předešlé fáze s vyjádřením jejich zranitelnosti popsané ve formě rizika,
- jedno riziko může ovlivňovat či vytvářet druhé,
- identifikují se vnitřní i vnější rizika,
- škálovatelnost rizik podle jejich šířivého efektu na své okolí (výbušná, nakažlivá, sdružená, protikladná a minimální).

Fáze 5

úkol 9. vyhodnoťte/odhadněte rizika (pomocí číselných hodnot z tabulky).
- pravděpodobnost rizika (viz přehled 1),
- jeho dopad (viz přehled 2),
- seskupování rizik.

Bodové ohodnocení se musí udělat pro každé riziko ve vzniklém seznamu – je to vlastně součin pravděpodobnosti, že hypotetické riziko se stane realitou, a jeho očekávaného dopadu (viz níže).

Přehled 1: Hodnocení pravděpodobnosti vzniku rizika se vyjadřuje číselnou stupnicí (1–6).

1 *minimální pravděpodobnost*, riziko se objeví jednou za 100 let
2 *velmi nízká pravděpodobnost*, jednou za 10 let
3 *pravděpodobnost nízká*, jednou za 5 let
4 *pravděpodobnost střední*, jednou za 1 rok
5 *pravděpodobnost vysoká*, výskyt jednou za měsíc
6 *pravděpodobnost velmi vysoká*, vyskytuje se vícekrát měsíčně

Přehled 2: Tabulka hodnocení dopadů – možného vlivu konkrétních rizik.

1 *nulový dopad*, vede k nulovým ztrátám ohledně autenticity a srozumitelnosti digitálních objektů
2 *zanedbatelný dopad*, vede k izolovaným ale zcela obnovitelným ztrátám autenticity a srozumitelnosti digitálních objektů
3 *malý dopad*, vede k širokým ale zcela obnovitelným ztrátám autenticity a srozumitelnosti digitálních objektů
4 *střední dopad*, vede k úplným ale zcela obnovitelným ztrátám autenticity a srozumitelnosti digitálních objektů
5 *velký dopad*, vede k izolovaným ztrátám autenticity a srozumitelnosti digitálních objektů – včetně nějakých neobnovitelných
6 *značný dopad*, vede k značným ztrátám autenticity a srozumitelnosti digitálních objektů, včetně ztrát neobnovitelných nebo obnovitelných pouze třetí stranou
7 *katastrofický dopad*, vede k úplné a neobnovitelné ztrátě autenticity a srozumitelnosti digitálních objektů

Fáze 6

úkol 10. Čelit objeveným rizikům; seznam opatření.
- vyhnout se okolnostem, za kterých se rizika objevují
- snížit pravděpodobnost výskytu rizika
- snížit potenciální dopad rizik

Zjednodušeně řečeno: v prvních 3 fázích (identifikace východisek; identifikace strategie a regulačního rámce; identifikace aktivit, prostředků a jejich „vlastníků“) jde o nashromáždění dostatku údajů k tomu, abychom mohli přistoupit k identifikaci jednotlivých rizik ve fázi 4, která poté vyústí v jejich bodové ohodnocení ve fázi 5, jež by mělo naplno ukázat, které z nich jsou potenciálně největším rizikem. V poslední fázi 6 by mělo dojít k návrhu určitého řešení pro nalezená rizika. Všech 6 fází je zpracováváno vzhledem k již výše zmíněným 8 funkčním třídám.

Výstup z celého auditu DRAMBORA a jeho výhody

Provozovatel repozitáře získá ucelený pohled na všechny procesy, které mají s repozitářem souvislost, a to ve formě seznamu rizik pro jednotlivé již zmíněné funkční třídy, kde je každé riziko uvedeno s bodovým skóre. Ukáží se tak silné a naopak nejslabší oblasti, které mohou znamenat riziko. Někdy výsledky mohou být překvapivé, velmi často se za nejrizikovější oblast považují technologie, ovšem daleko větší riziko může být v nedostatečně zvládnutém a nejistém financování repozitáře nebo instituce, která jej provozuje, nebo také např. v personálních otázkách (nedostatek odborníků z důvodu nízkých platů apod.). To jsou věci, které mohou potenciálně ochromit chod celého systému stejně jako technická závada. Dalšími výstupy jsou mj.:

- Zvýšené vědomí/znalost o chodu repozitáře, struktuře procesů v něm.
- Seznam rizik, která repozitáři hrozí.
- Porozumění silným stránkám i nedostatkům v infrastruktuře.
- Připravenost na kompletní externí audit (certifikaci).
- Ověření, zda všechny procesy a celková strategie jsou nastaveny dobře a zda systém pracuje optimálně.

Výsledky auditu v NK ČR

Ještě před vydáním nástroje DRAMBORA proběhly dva zkušební audity v Evropě a dva v USA. Během léta 2007 pak proběhly pilotní audity v institucích partnerů projektu DPE, kteří jsou spoluautory celého nástroje. Jedním z takových partnerů je i Národní knihovna

České republiky, která provedla audit v červenci a srpnu 2007. Následující tabulky jsou výstupem tohoto auditu.

Následuje přehled všech rizik, které byly výsledkem auditu v NK ČR. Dělení odpovídá jednotlivým funkčním třídám (viz výše).

Výsledky ve 4 procesních funkčních třídách

- akvizice a ingest (příjem dokumentů do repozitáře)
- ochrana a uložení
- management metadat
- přístup a šíření

Poznámka – číselné ohodnocení jednotlivých rizik je součinem pravděpodobnosti výskytu rizika a jeho možného dopadu. Není to tedy ve všech případech stav, který by byl na repozitáři právě aktuální. Může se jednat i o stav, který se bude moci vyskytnout i v budoucnu během rutinního provozu.

Akvizice a ingest		
R08	Systém neprovádí transformaci digitálních objektů do „archivních balíčků“	20
R03	Chybějící/nedostatečná metadata v nově ukládaných digitálních objektech (balíčcích)	15
R06	Externí změny ukládaných dat během ingestu	12
R07	Kontrola integrity ukládaných objektů není dostatečná	10
R04	Nevhodná forma digitálních objektů určených k ingestu (uložení)	8
R05	Nekompletní SIP	8
R09	Přebíraná/ukládaná data nejsou kontrolována antivirovým programem	6
R01	Dodavatel digitálních objektů zastaví spolupráci nebo změní její podmínky	3
R02	Omezení digitalizace nebo sklízení webu	3

Z tabulky problémů, které vzešly z auditu pro oblast akvizice a ingestu, se po ohodnocení jeví jako nejpalčivější ty, které souvisí s tím, že repozitář NK ČR nemá zatím implementován žádný DOMS (Digital object management system). Repozitář v podstatě zatím funguje jako „skladový“ prostor pro digitální dokumenty s file systémem. DOMS, podporující OAIS, dodá repozitáři potřebné vlastnosti, které se očekávají a mají zaručit dlouhodobou využitelnost uložených dokumentů a jejich formátů.

Proto stávající systém nepodporuje funkce, které by systém odpovídající OAIS podporovat měl, např. není schopen transformovat data, která do repozitáře během ingestu přicházejí, na archivní balíčky AIP apod. Problémem se může stát i nedostatečné vybavení digitálních objektů metadaty, zvláště v těch případech, kdy budou mít právo do repozitáře ukládat i jiné instituce než NK.

Ochrana a uložení		
R10	Integrita a autenticita digitálních objektů v repozitáři není dostatečně kontrolována	20
R18	Není jasné, co obsahuje balíček AIP	20
R19	Identifikátory digitálních objektů nejsou persistentní	20
R11	Plán ochrany není možné implementovat	16
R14	Ztráta integrity informací	15
R15	Ztráta autenticity informací	15
R12	Ztráta použitelnosti/schopnosti zobrazit informace	10
R16	Ztráta informace o původu informace	10
R13	Ztráta dostupnosti informací a služeb	6
R20	Strategie ochrany má za následek ztrátu informací	6
R17	Ztráta nebo nevhodná záloha	5

V oblasti Ochrany a uložení je hlavním problémem opět neexistence DOMS. Nemáme archivní balíček AIP, sporná je také otázka autenticity a integrity digitálních objektů v repozitáři (neoprávněná chtěná i nechtěná manipulace s objekty apod.). Další spornou otázkou je problematika globálních persistentních identifikátorů, spolu s otázkou identifikátorů interních, která je řešena pouze částečně. Dosti velká je i pravděpodobnost, že v budoucnu nebude možné zcela zavést požadovaný plán ochrany dokumentů, a to z různých důvodů (viz níže).

Management metadat		
R21	Ochranná metadata pro archivované objekty nejsou vytvářena	20
R26	Historie změn je nekompletní nebo nesprávná	20
R22	Definice balíčků SIP, AIP, DIP nejsou odpovídající	18
R23	Není jasné, co digitální objekty obsahují (formát)	10
R24	Správci repozitáře nesledují poslední vývoj a trendy v oblasti metadat	9
R25	Potenciální dodavatelé dat nemají úplné informace o metadatech, která jsou potřeba (mají dodat společně s digitálními objekty)	6
R27	Identifikace dokumentů není dostatečná	3

Problematika metadat náležejících k digitálním objektům skrývá také mnoho rizik. Předně, digitální objekty nyní nemají ochranná metadata, s výjimkou Krameria. U Krameria jsou na vstupu do repozitáře základní ochranná metadata, ovšem ta by měla být automaticky v repozitáři doplňována v průběhu „životního cyklu“ objektu. Toto zajistí opět pouze DOMS, který nemáme. Zatím není dokončena definice jednotlivých balíčků SIP, AIP a DIP. Určitým nebezpečím je i to, že digitální objekty nebudou v budoucnu použitelné i přes všechna opatření (migrace, emulace apod.).

Přístup a šíření informací		
R29	Transformace archivovaných objektů (AIP) do podoby dokumentů pro uživatele (DIP) nepracuje bezchybně	10
R31	Integrita digitálních objektů nabízených uživateli není známa	10
R33	Nefunkčnost ověřovacího a autorizačního podsystému	8
R32	Problémy s migrací do formátů určených pro poskytování uživatelům	6
R28	Mechanismy pro hledání, výběr a přístup k obsahu repozitáře nefungují odpovídající formou nebo vůbec neexistují	3
R30	Obsah repozitáře je poskytován uživatelům bez dodržování legislativy (autorský zákon aj.)	3

V poslední procesní funkční třídě není celkové bodové ohodnocení rizik tak vysoké. Nejvíce bodů dostala nemožnost proměny AIP do balíčku určeného uživatelům. Je to logické, když repozitář nedisponuje žádným DOMS. Ovšem i pokud disponovat bude, tato transformace může být problémem. Další rizika víceméně souvisejí s neexistencí DOMS.

Výsledky ve čtyřech podpůrných okruzích:

- organizace a management
- personální otázky
- finanční management
- technická infrastruktura a bezpečnost

Organizace a management		
R34	Celková koncepce Národní digitální knihovny není schválena vládou	21
R43	Nedostatečné vyčlenění prostředků na zajištění chodu repozitáře (nejen finančních)	20
R44	Pozice „digitální ochrany“ v organizační struktuře Národní knihovny ČR	15
R35	Smlouvy s producenty, dodavateli nejsou naplněny	12
R38	Dokumentace procesů, metod a procedur upravujících různé aspekty všech aktivit	12
R39	Mandát repozitáře	12
R40	Certifikace digitálního repozitáře jakožto důvěryhodného nebude úspěšná	12
R41	Selhání managementu	12
R37	Plán pro následné aktivity digitální ochrany po skončení životnosti repozitáře	10
R36	Odpovědnosti a povinnosti vyplývající z legislativy nebo jiných nařízení nejsou naplněny	9
R42	Ztráta důvěryhodnosti	9
R45	Nereálná očekávání rozsahu využití/úspěchu repozitáře	9

Z hlediska organizačního byla v době provedení auditu (léto 2007) **zásadním problémem skutečnost, že celková koncepce Národní digitální knihovny, pro niž je centrální digitální repozitář základním stavebním kamenem, nebyla schválena a finančně zajištěna navzdory tomu**, že je hotova (v souladu s objednávkou Ministerstva kultury ČR) již od roku 2005. **V současné době představuje určitou naději vládní schválení záměru Národní digitální knihovny v rámci Strukturálních fondů (14. 5. 2008).**

Podobně na tom byla i problematika „digital preservation“, která se řešila v době auditu v rámci několika oddělení **(od 1. 6. 2008 vzniká nový Odbor digitální ochrany, byť zatím s minimálním personálním obsazením)**. Zajímavým, i když očekávaným zjištěním bylo, že citelně chybí psaná a schválená dokumentace k procesům, které se okolo repozitáře odehrávají. Toto se může obrátit proti instituci tehdy, pokud je určitý proces zajišťován pouze jediným člověkem, který ovšem může instituci opustit, pak je těžké „obnovovat“ jeho postup a učit jej někoho jiného. Psané postupy mohou být ovšem nápomocné v zajištění odpovídající kvality určitého procesu. Rizikem může být i při rutinním běhu repozitáře selhání managementu nebo nekompetentnost vedoucích pracovníků, které může vést až ke ztrátě dat. Velmi důležitá je také otázka horizontu existence repozitáře. Je nutné mít výhledový plán, který říká, co bude následovat, až repozitář dosloužíapod.

Personální zajištění		
R47	Nedostatek kvalitních zaměstnanců	20
R50	Ztráta klíčového zaměstnance/ů	15
R46	Vypovězení nebo změna smluv ze strany externích dodavatelů	9
R48	Nedostatečný počet zaměstnanců v odboru digitální ochrany	9
R49	Zkušenosti a znalosti zaměstnanců zastarají	9
R51	Dobré pracovní prostředí a odměny zaměstnanců	8

Aby repozitář měl odpovídající úroveň, důvěru a spolehlivost, je samozřejmě nutné, aby se o něj staral kvalitní a dostatečně proškolený odborný personál. Dosti nepříjemností i vážných problémů může způsobit nedostatek takového personálu. Toto riziko je přímo spojeno s otázkou dostatečného financování celého repozitáře a věcí okolo něj, ale také se schopností a ochotou managementu udržet v instituci schopné lidi.

S již zmiňovanou dokumentací souvisí i hrozba ztráty klíčového zaměstnance, kterého nebude lehké nahradit, protože pouze on zná přístupová hesla, pouze on programoval celý systém, pouze on má odpovídající znalosti určité věci apod.). Zaměstnanci si v tomto oboru stále musejí doplňovat znalosti, protože pokrok jde velmi rychle kupředu. Musíme si také uvědomit, co vlastně zaměstnance drží u konkrétní firmy (v našem případě instituce), nejsou to jen peníze, ale i celková atmosféra a spousta dalších okolností, které práci zpříjemňují.

Finanční management		
R52	Nejistota získání financí na zakoupení DOM systému pro repozitář	20
R56	Nedostatek financí pro splnění závazků a cílů repozitáře	15
R54	Omezení rozpočtu a/nebo finanční deficit nebo omezení příjmů/dotací	9
R53	Transparentnost finančních operací a zásad organizace	5
R55	Náchylnost k nedodržování zákonů upravujících finanční toky a jiných nařízení	5

Je jasné, že finanční otázky působí problémy ve všech institucích a to nejen v ČR. Někde ovšem působí problémy menší, někde větší. Z hlediska získání digitálního repozitáře, který by se mohl ucházet o certifikaci (tj. externí potvrzení spolehlivosti a aktivního přístupu k digital preservation), je primární získání a implementace nějakého DOMS. Ten by zaručil, že všechny nárokované funkce bude repozitář schopen na různé úrovni plnit. V době auditu (léto 2007) byla otázka jeho zakoupení kvůli neschválené koncepci NDK značně nejistá. Po schválení záměru v květnu 2008 by bylo riziko ohodnoceno méně body, než jak je vidět ve výše uvedené tabulce. Z tabulky také vyplývá, že i pokud budeme zmíněný systém mít, bude stále repozitář velmi závislý na odpovídajícím a hlavně pravidelném řádně ukotveném systémovém financování. Není to otázka pouze zakoupení HW a SW a zprovoznění celé věci, ale dalšího financování. Jakékoliv snížení nebo omezení prostředků by mohlo mít negativní vliv na chod celého repozitáře a jeho cíl, tj. ochránit digitální objekty z dlouhodobého hlediska.

Technická infrastruktura a bezpečnost		
R59	Havárie nebo zastarání HW	15
R57	Záložní data storage a zálohy nejsou dostatečné	12
R58	Nedokonalý krizový plán, plán obnovy a malá připravenost na havárii	12
R62	Zneužití bezpečnostní zranitelnosti	12
R63	Interní nebo externí napadení SW	12
R60	Havárie nebo zastaralost SW	10
R61	Problémy s podporou operačního systému	10
R64	Destruktivní problém místního prostředí (přírodní pohromy)	8
R65	Nedostupnost základních služeb (elektřina, síťové připojení apod.)	6

Samozřejmou podmínkou toho, aby repozitář fungoval a nabízel své služby, musí být dostatečná a kontrolovaná technická infrastruktura. HW celého repozitáře je samozřejmě neustále monitorován (externí firma). Tato věc ovšem souvisí i se schopností interních pracovníků, kteří by měli být schopni při sebemenším problému jednat a hlavně jej rozpoznat, k čemuž musí mít odpovídající znalosti. Dalším viditelným rizikem je možný nedostatek kapacit v repozitáři (např. na zálohování). S bezpečností souvisí reálné riziko poškození důsledkem nějaké nehody či neštěstí. Pro tyto případy musí mít pracoviště vypracován tzv. bezpečnostní plán postupu, který by měl eliminovat možnost vzniku velkých a nevratných poškození. Tento plán musí být průběžně aktualizován. To, že celý repozitář by měl být „za zavřenými dveřmi“ s omezenými možnostmi vstupu, je samo

zřejmé. Po určité době se může problémovou ukázat otázka podpory jednotlivých využívaných SW, ovšem tato věc by měla být kontrolována a řešena v průběhu životnosti repozitáře. Ne příliš častou situací je výpadek dodávek elektřiny nebo připojení k síti. Pracoviště by mělo mít záložní zdroj energie, aby tato eventualita neovlivnila přístup uživatelů k dokumentům a hlavně, aby neohrozila samotný repozitář.

Závěr

Z výsledků jednotlivých částí auditu je zřejmé, že zásadním problémem není zakoupit a rozběhnout technologie k provozování repozitáře. Právě tuto podmínku většina lidí přeceňuje. Zásadním problémem není ani udržovat vše v odpovídajícím chodu. Je jasné, že je nutné zajistit omezení přístupu k repozitáři, ochranu před jeho napadením zvenčí (ale i zevnitř) a zamezit jeho poškození, např. živelní událostí.

Vážnějším problémem je organizační stránka instituce, která je oním důležitým jazýčkem na vahách, který může ovlivnit funkčnost a existenci celého repozitáře. Lze říci, že právě organizační, finanční a personální zázemí jsou nejdůležitější pro to, aby repozitář byl schopen plnit svou funkci – tj. uchovávat digitální objekty v dlouhodobém horizontu a zajišťovat k nim odpovídající přístup.

Z hlediska důvěryhodnosti a následné případné certifikace repozitáře zůstalo největším problémem neschválení koncepce Národní digitální knihovny vládou včetně zajištění financí. Dalším, úzce souvisejícím problémem je neexistence DOMS, který by pracoval „nad“ repozitářem. Repozitář tedy neobsahuje balíčky odpovídající OAIS a nemá ani archivní modul. Proto není možné kontrolovat autenticitu do takové míry, jaká by byla žádoucí.

V otázce metadat je chronickým a obecným nedostatkem malá míra zastoupení tzv. ochranných metadat, do repozitáře se spolu s digitálními objekty dostávají jen některá.

Třetím vážným problémem je relativní nedostatek odborného IT personálu. IT pracovník s potřebnými znalostmi se většinou nespokojí s tabulkovým platem státní instituce, přesto je ale nutné mít pracovníky s odpovídající kvalifikací přímo v instituci i v případě vysoké míry outsourcingu řady činností. I když údržba a monitoring budou zajištěny externí firmou, potřebuje tato firma schopnou styčnou osobu přímo na místě. Je dobré připomenout ještě jeden fakt. Vedení knihovny musí chápat důležitost a náročnost celé problematiky a podporovat ji všemi prostředky, nejen finančně. Tj. musí se ztotožnit s tím, že repozitář (v případě NK ČR) je nedílnou a velmi důležitou součástí knihovny, musí být ukotven v jejím statutu a její organizační struktuře tak, aby byla zajištěna jeho kontinuita.

Pro upřesnění přikládám tabulku rizik, která získala minimálně 15 bodů. Veškerá tato rizika byla již popsána v předchozím textu.

	Název rizika	
R34	Celková koncepce Národní digitální knihovny není schválena vládou	21
R08	Systém neprovádí transformaci digitálních objektů do „archivních balíčků“	20
R10	Integrita a autenticita digitálních objektů v repozitáři není dostatečně kontrolována	20
R18	Není jasné, co obsahuje balíček AIP	20
R19	Identifikátory digitálních objektů nejsou persistentní	20
R21	Ochranná metadata pro archivované objekty nejsou vytvářena	20
R26	Historie změn je nekompletní nebo nesprávná	20
R43	Nedostatečné vyčlenění prostředků na zajištění chodu repozitáře (nejen finančních)	20
R52	Nejistota získání financí na zakoupení DOM systému pro repozitář	20
R47	Nedostatek kvalitních zaměstnanců	20
R22	Definice balíčků SIP, AIP, DIP nejsou odpovídající	18
R11	Plán ochrany není možné implementovat	16
R03	Chybějící/nedostatečná metadata v nově ukládaných digitálních objektech (balíčcích)	15
R09	Přebíraná/ukládaná data nejsou kontrolována antivirovým programem	15
R14	Ztráta integrity informací	15
R15	Ztráta autenticity informací	15
R44	Pozice „digitální ochrany“ v organizační struktuře Národní knihovny ČR	15
R50	Ztráta klíčového zaměstnance/ů	15
R56	Nedostatek financí pro splnění závazků a cílů repozitáře	15
R59	Havárie nebo zastarání HW	15

Budoucnost nástroje DRAMBORA

Na základě auditů provedených v roce 2007 v rámci testování první verze nástroje byla DRAMBORA doplněna, upravena a počátkem dubna 2008 byla veřejnosti zpřístupněna verze 2. DRAMBORA již není jen „pouhý“ návod v papírové podobě se šablonami, ale jedná se o mocný interaktivní online nástroj, který zpracování auditu velmi ulehčuje a umožňuje jeho neustálé doplňování. Každý auditor má v systému svůj účet a má neustálý přístup k auditu. Na jaře roku 2009 by měla následovat finální verze 3, která bude reflektovat připomínky uživatelů k verzi 2.

Koncem roku 2008 by se mělo oficiálně začít s certifikací auditorů. Školení a certifikace auditorů nebude samozřejmě nutná podmínka k provedení samotného auditu, ovšem absolventi školení by měli být schopni provést audit rychle a odpovídající formou, případně předat zkušenosti dalším potencionálním auditorům.

Podrobnější informace o nástroji DRAMBORA verze 2 online naleznete ve sborníku konference Inforum 2008.

Použitá literatura:

Catalogue of Criteria for Trusted Digital Repositories [online]. Version 1. Goettingen : Nestor Working Group, 2006. 48 s.[cit. 2007-12-20]. Přístup z WWW: <http://edoc.hu-berlin.de/series/nestor-materialien/8en/PDF/8en.pdf >.

CRL Core Requirements for Digital Archives [online]. CRL, 2002-2007 [cit. 2007-12-20]. Přístup z WWW: <http://www.crl.edu/content.asp?l1=13&l2=58&l3=162&l4=92>.

Digital Curation Centre and DigitalPreservationEurope. Digital Repository Audit Method Based on Risk Assessment [online]. Version 1.0. 221 s. Glasgow : DPE, 2007 [cit. 2007-12-20].
Přístup z WWW: <http://www.repositoryaudit.eu>.

Digital Repository Audit and Certification [online]. 2007 [cit. 2007-12-19]. Přístup z WWW: <http://wiki.digitalrepositoryauditandcertification.org/>.

GIARETTA, David. *ISO/CCSDS Open Archival Information System (OAIS) Reference Model* [online]. 1997. Powerpointová prezentace [cit. 2007-12-20]. Přístup z WWW: <http://www.sstd.rl.ac.uk/ccsdsp2/isoas/bnsc97/oais1/OAIS1.PPT>.

ROSS, Seamus et al. *Building Trust in Digital Repositories Using DRAMBORA*. Powerpointová prezentace z DPE Workshopu v Haagu, 4. 5. 2007. Interní materiál.

STOKLASOVÁ, Bohdana; HUTAŘ, Jan. Nové směry v dlouhodobém uchovávání dokumentů v mezinárodním kontextu. In *Automatizace knihovnických procesů 11*. Liberec 16.-17. 5. 2007. Praha : ČVUT, 2007. s. 83-93. Přístup také z WWW: <http://www.akvs.cz/akp-2007/11-stoklasova-hutar.pdf>. ISBN 978-80-01-03691-4.

Trusted Digital Repositories : Attributes and Responsibilities. An RLG-OCLC Report [online]. RLG : Mountain View (CA), 2002. 70 s. [cit. 2007-12-19].
Přístup z WWW: <http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf>.

Trustworthy Repositories Audit & Certification (TRAC) : Criteria and Checklist [online]. Dublin (OH); Chicago (IL) : OCLC, CRL, 2007. 94 s. [cit. 2007-12-19]. Přístup z WWW: <http://www.crl.edu/PDF/trac.pdf>.