OAIS: možnosti a limity aplikace

Marek Melichar
Multidata Praha s.r.o

Jan Hutař
Archives New Zealand

Text vznikl v rámci práce v pracovní skupině pro dlouhodobou ochranu v projektu Koncepce rozvoje knihoven ČR na léta 2011-2015

Anotace

V tomto textu se pokusíme ukázat, že existující různé výklady normy OAIS mají základ v odlišnostech prostředí, ve kterých se OAIS aplikuje. Archivy kulturních a vědeckých dat a archivy dokumentů firemních a správních se v praktickém přístupu k řešení požadavků dlouhodobé archivace velmi liší. Ačkoli používají společnou terminologii vycházející z této normy, vykládají odlišně dokonce i některé základní pojmy. Některé implementace OAIS jsou zaměřeny výhradně na ochranu bitů tvořících ukládané dokumenty, jiné se soustředí na ochranu jejich obsahu. Některé implementace OAIS počítají s nutností prokázaní autenticity obsahu pomocí technických prostředků, jako jsou certifikované elektronické podpisy, jinde stačí prokázání autenticity informací v metadatech objektu. Zásadně rozdílné je chápání pojmu "důvěryhodný". Současné technologie umožňují naplnit každý jednotlivý cíl OAIS archivu tak, že mohou být jednotlivé cíle dohromady neslučitelné. Každý archiv s dlouhodobým mandátem se musí rozhodnout, na co bude klást větší důraz.

Klíčová slova: OAIS, Dlouhodobá archivace digitálních informací, digitální repozitář,důvěryhodnost, certifikace dlouhodobých digitálních repozitářů

English

In this article we discuss existing differences of OAIS reference model interpretations as results of different environments constraints and expectations. Culture heritage or research data archives and business or administration archives interpret some requirements of the long-term preservation quite distinctively. They may use common terminology stemming from OAIS, but understand differently even some of its basic concepts. Some OAIS implementations focus solely on bit level preservation, others strive to preserve the accessibility of the content information. Some OAIS implementations expect that the authenticity of the archived content has to be demonstrated using technical tools, for example by the certified electronic signatures, while other archives get by with the information in the metadata of the archived objects. Quite discrepant is the understanding of the concept of "trustworthiness."Current technologies enable the realization of the requirements of OAIS archive perfectly, but could contradic the intended goals of OAIS. In the end, each archive has to decide, where to put the focus.

Keywords: OAIS, Long-term Preservation of Digital Information, Digital Repository, Trustworthiness, Long-term Digital Repository Certification

1. Vyhovění OAIS

Knihovnické veřejnosti vzdělané v posledních letech v oboru knihovnictví a informační věda není pravděpodobně nutné OAIS (ISO 14721:2003) nijak rozsáhle představovat. Norma OAIS vznikala od poloviny devadesátých let na žádost ISO v koordinaci *Consultative Committee for Space Data Systems* (CCSDS), první verze textu je dnes dostupná v Internet archivu[1]. Draft ISO normy byl publikován v roce 1999 a norma pak byla vydána jako ISO standard v roce 2003[2]. V roce 2009 byla aktualizována [3].
Stručně řešeno, OAIS je konceptuální model dlouhodobého archivu, který popisuje základní komponenty dlouhodobého archivu a související funkce a vazby, a také informační model. Informační model popisuje jednotlivé typy informačních balíčků, a to, jaké informace obsahují. OAIS tedy předepisuje jaká metadata mají být spolu s ochraňovanými informacemi ukládána (podrobněji viz např. [4].
OAIS je zatím nezpochybněným základem dlouhodobé ochrany digitálních dat. Základní specifikace cílů OAIS archivu, tedy archivu, který odpovídá vlastní normě, v úvodu textu říká, že OAIS archiv musí:

- *vyjednávat a získávat vhodné informace od producentů;*
- *získat nad poskytnutými informacemi dostatečnou kontrolu tak, aby byl schopen zajistit dlouhodobou ochranu;*
- *rozhodnout buď samostatně, nebo ve spolupráci s dalšími stranami, které komunity by se měly stát uživateli (designated community) a tedy, které komunity by měly být schopny porozumět archivované informaci;*
- *zajistit, že informace, které jsou archivovány, budou nezávisle srozumitelné uživatelům. Jinými slovy, uživatelé by měli být schopni porozumět informaci bez potřeby pomoci expertů, kteří informaci vytvořili;*
- *následovat dokumentované politiky a procedury, které zajistí, že ukládaná informace bude ochráněna proti všem předpokládaným rizikům, a která zajistí, že informace určená k distribuci z archivu bude šířena v autentických kopiích originálu, nebo s jasným vztahem k originálu;*
- *zpřístupňovat ochraňované informace uživatelům.*

Archiv, který odpovídá OAIS, by měl naplnit minimálně výše uvedené cíle a měl by prokázat a vysvětlit, jak konkrétně je naplňuje. Dále ve specifikaci shody nebo kompatibility archivu s normou samotná norma OAIS nejde. Pokud archiv implementuje informační model a doloží, jak plní výše uvedené, je možné jej považovat za archiv v souladu s OAIS. Ovšem ne vždy to znamená, že archiv plně odpovídá OAIS a podporuje všechny procesy tak, jak je OAIS definuje v částech 3.2, 4. a dále. Další části normy OAIS jsou však pouze ilustrativní, popisují podrobněji principy fungování OAIS archivu a slouží jako příklady. Tvůrci normy měli pravděpodobně za to, že není možné předepisovat jednotlivé funkce podrobněji všem možným typům archivů spravujícím velmi rozličné dokumenty v různých kontextech.
V realitě jednotlivé archivy a systémy naplňují požadavky OAIS v různé míře. Některé archivy disponují jen některými funkčními entitami OAIS.

1.1 Jak se tedy pozná digitální archiv, který OAIS odpovídá?

Na celou problematiku existují minimálně tři pohledy [5]. První pohled říká, že kompatibilita je popsána přímo v referenčním rámci OAIS, v částech 1.4, 2.2 a 3.1. Příkladem takového pojetí může být prohlášení kompatibility systému LOCKSS s OAIS, které popisuje, jak LOCKSS odpovídá nárokům z částí 2.2 a 3.1 OAIS referenčního rámce[6]. Druhý pohled je takový, že digitální repozitář odpovídá OAIS tehdy, pokud repozitář má funkcionalitu takovou, jak je specifikována v OAIS nebo alespoň implementuje hlavní charakteristiky modelu OAIS, jako je šest klíčových funkčních komponent/modulů[1] a specifikované odpovědnosti. Třetí pohled je v podstatě odmítavý a říká, že kompatibilita OAIS je velmi vágní termín a referenční rámec není konceptem, ke kterému by měla být jakákoliv kompatibilita vztahována.

Jak je tedy patrné, skutečnost, že konkrétní SW nebo digitální repozitář proklamuje, že odpovídá OAIS, neznamená, že má plnou funkcionalitu z OAIS vyplývající a tím méně to znamená, že se jedná o systém pro dlouhodobou ochranu digitálních dokumentů (LTP - *Long-term Preservation*). Kompatibilita OAIS není rovna funkcionalitám LTP systému. Ve velké většině digitálních repozitářů chybí klíčový modul Plánování ochrany a s ním spojené procesy.

V praxi se tedy může stát, že jeden archiv (nebo systém), o kterém jeho tvůrci prohlašují, že OAIS odpovídá, implementuje přesně všechny požadavky vyplývající z OAIS nebo alespoň ty popsané v kapitole 3.2 a dále v normě OAIS a jiný archiv, o kterém se tvrdí totéž, nemusí nakonec mít s podrobnými požadavky OAIS nic společného. V některých oblastech dlouhodobé archivace digitálních informací se tvůrci systémů snaží svoje archivy budovat s požadavky OAIS jako základní kostrou a explicitně naplňují všechny podrobné funkce vyplývající z OAIS. Jiní tvůrci archivních systémů se na OAIS příliš neohlížejí a pouze svoje archivní řešení popisují pojmy šesti základních komponent OAIS.

Některé archivy implementují velkou část podrobných funkcí OAIS procesně, vně informačního systému, jiná řešení se snaží procesy v archivu maximálně automatizovat a podporovat všechny OAIS funkce.

K orientaci a posouzení toho, zda ten nebo onen archiv splňuje požadavky OAIS a lze jej považovat za dlouhodobý archiv ve smyslu této normy (tj. archiv schopný zajistit logickou ochranu obsahu, nejen ochranu bitů, které obsah tvoří), nám tedy mohou posloužit v podstatě jen vnější indicie. Z dokumentace, kterou archiv publikuje pro svoje *stakeholdery* (především pro ty, kteří jeho provoz financují, kteří mu data svěřují, a kteří data používají – tedy v případě knihoven i běžná veřejnost) by mělo být jasné:

- do jaké míry a jak implementuje archiv datový a informační model OAIS (jednotlivé moduly, metadata, procesy, jak jsou jednotlivé části integrovány mezi sebou a s okolím). Archiv by měl veřejně publikovat informace popisující způsob vyhovění požadavkům normy OAIS;
- zda archiv implementuje současné *best practices* (tj. implementuje-li uznávané a široce používané standardy v oblasti dlouhodobé ochrany, a to jak v oblasti zpracování digitálních formátů dat, tak v oblasti používaných metadat).

[1] Jde o OAIS moduly *Ingest* (Příjem); *Administration* (Administrace); *Archival Storage* (Archivní sklad); *Data Management* (Správa dat); *Preservation Planning* (Plánování dlouhodobé ochrany); *Access* (Zpřístupnění).

Nejlepším způsobem prokázání kvality systému pro dlouhodobou archivaci digitálních informací je ale samozřejmě certifikace. V oblasti dlouhodobé ochrany digitálních dat je v Evropě k dispozici metodika podpořená Evropskou komisí [7], která doporučuje postupovat k certifikaci postupně od deklarace záměru data dlouhodobě uchovávat (*Data Seal of Approval*, http://www.datasealofapproval.org/), přes self-audit (podle DIN 31644 nebo ISO 16363:2012 norem, viz dále) k externímu auditu akreditovaným externím auditorem (také akreditovaným podle normy ISO 16919).
Archivy, které berou dlouhodobý aspekt své činnost skutečně vážně, se snaží si vybudovat a udržet statut důvěryhodného dlouhodobého repozitáře digitálních informací a o oficiální externí certifikaci akreditovaným certifikátorem usilují. Samozřejmě akreditace certifikovanou autoritou je finančně i časově náročná (viz např. [8]), a proces popsaný normou ISO 16363:2012 dosud není v plném běhu. Existuje již sice orgán, který má externí certifikátory podle ISO 16919 akreditovat (PTAB, www.16363.org/), ale norma 16919 nebyla dosud (červen 2012) vydána jako platný standard. Doposud jsou akreditace přidělovány skrz americkou organizaci *Center for Research Libraries* (CRL, http://www.crl.edu/).
K posouzení spolehlivosti a schopnosti archivu uchovávat dlouhodobě digitální informace také přispívá prokázání kvality v dalších oblastech jako je informační bezpečnost, management elektronických záznamů nebo prokázaní kvality činnosti organizace podle ISO 9001. To jsou doplňující vnější indicie, které vypovídají o tom, jak vážně svojí dlouhodobou misi instituce bere.
V případě, že základem systému pro dlouhodobou archivaci digitálních informací je komerční (softwarové) řešení, mají význam také měkká kritéria, jako je schopnost producenta prokázat reference z relevantních projektů řešících logickou dlouhodobou archivaci relevantního typu obsahu nebo aktivita producenta systému v komunitě zabývající se výzkumem v oblasti dlouhodobé ochrany.

V následujícím textu se zamyslíme nad několika body, nad kterými tvůrci dlouhodobých archivů digitálních informací musí učinit zásadní rozhodnutí. Některé požadavky a očekávání spojená s dlouhodobou ochranou digitálních informací nelze jednoduše spojit a implementovat společně v jednom systému. Každý tvůrce musí vyhovět specifickým potřebám oblasti, pro kterou archiv buduje. Tyto požadavky nemusí ale mít v kontextu dlouhodobého archivu smysl a mohou jít proti jeho hlavnímu cíli: zajištění trvale dostupné nezávisle srozumitelné informace v autentické podobě pro budoucí uživatele navzdory změnám technologií. Každý tvůrce dlouhodobého archivu se také musí rozhodnout, jaké z dostupných technologií použije.

2. Lokální repozitář vs. cloudové, distribuované služby

Jak jsme řekli, norma OAIS nepředepisuje, jak přesně mají být realizovány jednotlivé funkce a moduly. Tradičně je OAIS archiv popisován jako samostatně stojící repozitář, který má externí dodavatele dat, data ukládá lokálně a zálohuje je (ve více lokalitách), a data také zpřístupňuje určené komunitě.
Budování OAIS archivu jako lokálního repozitáře ovšem není jedinou cestou jak data nebo dokumenty dlouhodobě ukládat. Gridové nebo cloudové způsoby ukládání dat ale staví instituce při aplikaci OAIS před nové problémy. Norma OAIS vůbec neřeší problémy organizací, které fungují v distribuovaném nebo cloudovém prostředí, nebo chtějí distribuované ukládání používat v modulu Archivní sklad. Už vůbec tato norma není připravena na provoz dlouhodobé ochrany (nebo jen uložení dat) jako služby

poskytované třetí stranou (SaaS - *Storage as a service*). OAIS neřeší situace, kdy se k dodavatelům dat a k archivu připojují např. další partneři jako poskytovatelé cloudového Archivního skladu jako služby , a kdy ke komponentům archivu přibývají další vrstvy nebo systémy v oblasti Archivního skladu. Z hlediska filosofie dlouhodobého uchovávání v OAIS je klíčové, zda OAIS archiv má nad svými daty plnou kontrolu a co všechno s nimi může dělat[2]. [9]

Je třeba rozlišit různé možné přístupy k využívání cloudových nebo gridových technologií v dlouhodobé ochraně digitálních informací:

1) použití soukromého cloudu nebo gridu pro řešení modulu Archivní sklad,

2) použití cloud/grid storage jako služby, nebo převedení i nejen storage ale i aplikace pro správu archivu zcela do cloudu.

Každý z těchto přístupů je asi vhodný pro jiný typ organizace a má dopady na způsob financování dlouhodobé ochrany digitálních dat. A každý nese rizika, se kterými je třeba počítat.

První přístup se konceptuálně příliš neliší od běžného modelu lokálního repozitáře. Jen na úrovni OAIS modulu Archivního skladu funguje složitější technologie. Další archivem spravovaná a plně kontrolovaná vrstva odděluje fyzické uložení a služby OAIS archivu. To v podstatě není nic jiného než přístup k datům přes API, nějaké chytřejší storage technologie nebo HSM (*Hierarchical Storage Management*), která zakrývá několik typů fyzických úložišť, různě rychlé disky a pásky. Ovšem technologie zde používané budou jiné, jako HDFS (*Hadoop Distributed File System*) apod.

Druhý přístup znamená vždy vytvoření závislosti na poskytovateli služeb, s tím že různé činnosti archivu jsou převedeny na úroveň služby.

Nejznámějším systémem pro dlouhodobou archivaci, který využívá distribuované ukládání je LOCKSS (*Lots of Copies Keep Stuff Safe*). Dlouhodobým ukládáním se ale zabývají také propagátoři cloudových technologií. Velmi aktivní je např. organizace SNIA (www.snia.org), která mimo jiné vytváří také standardy a metodiku (XAM, XFDU, SIRF atd.) pro dlouhodobé uchovávání. Zajímavý je jejich koncept nezávislého přenositelného digitálního objektu pro dlouhodobé ukládání SIRF (*Self-contained Information Retention Format*)[10].SIRF je budoucím ISO standardem pro samostatně použitelný, sebepopisující a rozšiřitelný logický kontejner určený pro dlouhodobé ukládání. Tento model ale řeší spíše problémy související s ukládáním v cloudu než s udržením dlouhodobé dostupnosti digitálních informací v nějakých konkrétních formátech. Tyto snahy jistě budou směřovat k postupné standardizaci dlouhodobé ochrany digitálních dat jako služby [11] nebo stanovení jasnějších metodik v této oblasti.

Některé technologie a přístupy v oblasti distribuovaného ukládání řeší z hlediska dlouhodobé ochrany pouze fyzickou rovinu uložení, jiné již dnes směřují k tomu, že budou nabízet stejné nástroje pro správu obsahu archivu jako lokální repozitáře. Debata kolem využití cloudových technologií pro dlouhodobou ochranu digitálních informací upozorňuje na některá rizika. Poskytovatelé storage vcloudu (v případě externích služeb) nemusí být vždy ideálními partnery pro dlouhodobou archivaci z

[2] Také v případě lokálního repozitáře využívají organizace pro zajištění služeb spojených s provozem lokální HW úložiště další dodavatele služeb. Ovšem v případě použití archivního skladu v cloudu jako SaaS je situace mnohem komplexnější, archiv je od svých dat ještě o pár kroků dál a bez dodavatele služby se k nim nedostane. Situace má jak organizační a smluvní aspekty, tak technické aspekty – systém archivu například při uložení v cloudu nemusí mít možnost kontrolovat integritu dat přímo, ale někdy jen prostřednictvím „služby pro fixity check“. Archiv je tak mnohem více závislý na poskytovateli/lích služeb než v případě využití lokálního repozitáře.

několika důvodů. Mají kontrolu nad daty organizace, mohou používat proprietární technologie, zažívají výpadky služeb a ztráty dat, mohou zaniknout, data mohou fyzicky skladovat kdekoli apod. Pro státní instituce může být podstatným problémem také skutečnost, že data jsou mimo konkrétní zemi a podléhají legislativně jiné.

Z hlediska dlouhodobé archivace digitálních informací je ale především rozhodující, jestli je při cloudovém ukládání možné realizovat ekvivalentní služby, jako při lokálním uložení, tj. jestli je možné k datům v rozumném čase přistupovat, zda je možné provádět kontroly *fixity*(integrity či neměnnosti bitů) ukládaných objektů, jestli je možné provést na uložených datech opakovanou analýzu rizik uložených dat, zda je možné re-validovat formáty digitálních dat, zda je možné snadno plánovat a realizovat formátové migrace. Zda je možné data rychle a jednoduše exportovat (*exit policy*) nebo zda lze dostat vždy kompletní informace související s archivními daty i bez systémů archivu.

3. Kulturní data vs. firemní data

Předmětem dlouhodobé ochrany archivu navrženého podle OAIS je informace. V dlouhodobé perspektivě má ochraňovaná informace delší životnost než nějaký konkrétní informační systém označovaný za OAIS odpovídající archiv. Informaci definuje OAIS jako "jakoukoli jednotku znalostí, které je možné si vyměňovat" a informace je při komunikaci reprezentována v datech. OAIS je referenční model, který má platit pro všechny typy archivů. Přirozeně, různé typy archivovaného obsahu vyžadují jiný způsob implementace OAIS.

Data z kulturních institucí jsou z mnoha pohledů odlišná od dat firemních. Především v dlouhodobé perspektivě jsou kulturní data předmětem neustálé interpretace. Kulturní objekty, reprezentované v datech, digitalizované nebo výhradně digitální povahy, jsou součástí kultury a jejich význam se v čase mění. Badatelé a uživatelé kulturní data stále re-interpretují a jejich interpretace se stávají součástí významu kulturních objektů samotných. V praxi OAIS archivu to znamená, že:

1) se mění některé informace, které podle požadavků OAIS musí archiv obsahovat vedle obsahové informace: mění se kontextová informace (*context information*), vyvíjí se informace o původu (*provenance information*).

2) mění se znalostní základna komunity uživatelů (*knowledge base, designated community*) což v praxi může velmi dramaticky omezit budoucí nezávislou srozumitelnost ochraňovaných informací.

3) může se měnit i informace o formě reprezentace (*representation information*), resp. je třeba ji měnit tak, aby archivovaný objekt byl stále samostatně srozumitelný v novém technickém prostředí, ale i v situaci, kdy se změnila znalostní základna uživatelské komunity.

Pokud uvažujeme o kulturních datech jako o objektech skutečně dlouhodobé archivace, musíme s těmito změnami počítat, a musíme mít nástroje, které nám umožní archivované objekty obohacovat a přitom neztratit srozumitelnou informaci o původu dokumentu nebo nepřijít o možnost prokázat jeho autenticitu.

Celá oblast dlouhodobé ochrany digitálních dat vznikla na popud institucí zabývajících se vesmírným výzkumem. Pro příklady dopadů re-interpretace na archivované objekty ale není třeba chodit do oblastí s tak komplexními daty, jako jsou např. data z měření v *takamoku* nebo z vesmírného výzkumu, ale vystačíme i s jednoduchými knihovními daty.

Některé dokumenty jsou v knihovnách poprvé popisovány a elektronicky katalogizovány až v okamžiku digitalizace. V případě masové digitalizace (například ve spolupráci s Googlem), desítky specialistů na bibliografické zpracování historických dokumentů vytvářejí k těmto dokumentům poprvé elektronické záznamy. Když jsou data naskenována a dostanou se k badatelům online, je přirozené, že jejich interpretace si vyžádají další úpravy metadat, možná i strukturální změny - jeden dokument se může stát několika samostatnými dokumenty, změní se typ dokumentu, změní se data o původcích, názvech, edicích atd. přibudou plné texty, objeví se související dokumenty. U historických dokumentů je takový proces přirozený. Správci obsahu a knihovníci mají rádi pořádek a standardy. Ovšem standardy se mění a mění se jejich interpretace, a při zpracování dat dochází k chybám. Tedy i dokumenty, které do archivu přijdou se zdánlivě dokonalým popisem, budou vyžadovat v budoucnu změny. To se již reálně děje, viz např. obohacování záznamů, překlady originálních textů a edice v digitální knihovně NK ČR Manuscriptorium.

V zájmu zachování dostupnosti obsahu dokumentu, bude také potřeba měnit jeho formát nebo měnit některé vlastnosti formátu, který obsah kóduje.[3]OAIS archiv spravující tento typ dat musí takové změny umožnit při zachování informací o původu dokumentu a musí být schopen provedené změny zachytit, uchovat a dodat uživateli jako součást kompletního archivního balíčku.

Firemní data jsou samozřejmě také součástí kultury, a pokud vstoupí do dlouhodobého archivu, musí se na ně uplatit podobné požadavky, zvláště pokud projdou skartačním řízením a stanou se z nich archiválie v digitální podobě. Budoucí historikové budou potřebovat i tato data, ovšem asi není možné čekat, že všechny informace firemní povahy budou ukládány navždy. Firemní archivace má za cíl vyhovět současným zákonným požadavkům a lhůtám a firemní archivy počítají obvykle s uživateli jen velmi výjimečně. Re-interpretace firemních dat probíhá v předcházejících fázích jejich životního cyklu, v dokument management systémech a ECM systémech, a do archivu přicházejí skutečně již jen „mrtvá data“.

Kulturní (a také například vědecká) archivní data mají tedy mnohem delší životní cyklus než data firemní. Rozpracování modelů životního cyklu kulturních a vědeckých dat se věnuje řada zahraničních publikací z oblasti *digital curation* a *data managementu*[13]. Z dlouhodobého hlediska musí být archivy kulturních dat připraveny na trvalou re-interpretaci uloženého obsahu a musí OAIS naplňovat jinak, než archivy dat firemních. Firemní data také obvykle mají kratší perspektivu a instituce, které je ukládají, nemají obvykle na rozdíl od knihoven nebo archivů mandát a ani potřebu ukládat je navždy.[4]

[3]Některé typy formátů bohužel umožňují předstírat validní objekt v souladu se specifikací (typicky například PDF viz video [12]). Dnes bezproblémově zobrazitelný obsah může mít vlastnosti, které mohou budoucí použitelnost značně ztížit, a při tom se formát může zdát v souladu se specifikací (v případě PDF například problém verzí objektů uvnitř souboru, v podstatě libovolného umístění koncových značek, možnost přidávat bity do míst, kde by být neměly, což se při validaci a dnešním používání nepozná atd.). Pokud bychom do archivu přijímali PDF obsah od více dodavatelů dat a ten by procházel jen základní validací, lze očekávat, že budoucí uživatelé budou mít co do činění s mnoha variantami nestandardností. Vstupující data je vhodné normalizovat (v případě PDF překódovat všechny vstupující PDF pomocí jednoho nástroje a jedním nastavením například) právě proto, aby nestandardností bylo co nejméně typů.

[4] Podle toho pak vypadají archivy a systémy, které jsou vydávány za OAIS archivy v komerčním kontextu jako rozšíření dokument managementu. Obvykle není jejich cílem uchování srozumitelnosti a dostupnosti obsahu dokumentu po dlouho dobu navzdory technologickým změnám. Firmy většinou dokumenty po 5ti, 10ti nebo max. 20ti letech zlikvidují nebo předají některé z nich do státního archivu. Problémem je, pokud takovýto přístup k řešení digitálního archivu má i samotný archiv nebo jiná paměťová instituce.

Tabulka 1. Rozdíly mezi kulturními a firemními daty z hlediska dlouhodobé archivace

	Kulturní a vědecké archivy	Podnikové a správní archivy
Data	Kulturní a vědecká data	Firemní nebo správní dokumenty
OAIS	Plná podpora požadavků v informačním systému	Implementace metodikou nebo v procesech
Úroveň ochrany	Udržení nezávisle použitelného obsahu	Dodání bitového streamu v původním stavu (SIP=AIP=DIP)
Ochrana obsahu	Definice„signifikantních vlastností“ obsahu a snaha je udržet v budoucnu	Udržení bitů
Access	Obsah se používá (Preservation for Access)	Obsah se ukládá (Dark archive)
Autenticita, provenience	Hashe,metadata + audit trail PREMIS events, agents aj.	Šifrování obsahu + kvalif. certifikáty, el. podpisy, časová razítka, „objektivní dokazovací postupy“ a „právní nezpochybnitelnost“
Důvěryhodnost	Ve smyslu ISO 16363:2012 “to provide reliable, long-term access to managed digital resources to its designated community, now and into the future”	Schopnost objektivně prokázat, že s obsahem se nemohl seznámit nikdo nepovolaný, případně jej neměl možnost měnit
Archivní entita	Logická entita založená na standardech (například na PREMIS a METS)	Balíky bitů
Změna digitálního formátu	Nová reprezentace uvnitř AIP	Nové verze balíků – nová verzeAIP

4. Ochrana dostupnosti obsahu/informace vs. ochrana autenticity/provenience

Informace ochraňovaná v OAIS archivu má být:

1) nezávisle srozumitelná definovaným uživatelům v průběhu času,

2)autentická, tedy archiv má být schopen prokázat autenticitu šířené informace, resp. její vztah k nějakému “originálu”.

Dlouhodobý archiv na jednu stranu musí počítat s tím, že se bude měnit prostředí, ve kterém bude existovat a na druhou stranu musí prokázat autenticitu a provenienci ukládaného obsahu. Jak již jsme řekli výše, v dlouhodobé perspektivě je třeba očekávat, že se změní technologické prostředí a očekávání (schopnosti) definovaných uživatelů.

Udržet informaci nezávisle srozumitelnou po dlouhou dobu většinou znamená ji změnit (tj. například změnit způsob kódování informace do bit streamu) nebo ji obohatit o novou informaci. Ve skutečně

dlouhodobé perspektivě to může být změna velmi komplexní. Udržet informaci autentickou, prokázat vztah k originálu, je nejjednodušší udržením bitů ve stejné podobě (SIP=AIP=DIP), jakou měly při vložení. Spojení těchto dvou požadavků také řeší metadatový standard PREMIS (http://www.loc.gov/standards/premis/), který umožňuje obojí. PREMIS event je schopen udržet údaje o všech změnách (formátové migrace, např. z důvodu nových technologií pro zpřístupnění, přemístění apod.) včetně času a tzv. agenta, který událost (*event*) provedl. Umožňuje tak uchovávat informace vedoucí až k originální podobě digitálního objektu.

Ovšem prokázání autenticity a provenience pomocí metadat není dostatečná strategie pro všechny typy archivů. Současné technologie umožňují, a některé typy dokumentů vyžadují, zcela jiné a z určitého pohledu mnohem spolehlivější metody zajištění autenticity. Autenticitu a nezměněnost lze prokázat pomocí certifikovaných podpisů a časových razítek. Tento přístup je obvyklejší u firemních archivů, částečně také správních archivů. Existuje model dlouhodobé certifikace dokumentů, ovšem i kdyby existovaly autority, které by se zavázaly trvale poskytovat prostředky pro certifikované podepisování, certifikační politiky se musí měnit a měnit se musí i se změnami technologií. Důvěryhodnost a schopnost prokázat autenticitu je tedy zde převedena na důvěryhodnost poskytovatele certifikátu.

Pro knihovní a kulturní data stačí, když jsou neměnnost a původ prokázány metadaty a pomocí kontrolních součtů jako je md5, SHA-1 aj. Pro komerční nebo správní obsah je důležité, aby neměnnost byla potvrzena externě, nějakou autoritou, která informaci o původním kontrolním součtu ochraňuje a poskytuje jen v kryptované podobě. K samotnému digitálnímu objektu je tak přidávána další vrstva různých certifikovaných podpisů, která může být z hlediska zajištění dlouhodobé dostupnosti digitálních informaci problematická: může omezit možnost budoucích správců archivu nakládat s archivovaným obsahem tak jak je v danou chvíli potřeba, a vytváří závislost archivu na externím poskytovateli nějaké služby.

Podobný problém je s implementací omezení přístupových práv. Na jednu stranu některý archiv může chtít ukládat obsah tak, aby k němu měla prokazatelně přístup jen definovaná skupina uživatelů, čehož může chtít dosáhnout technickými prostředky jako je šifrování obsahu a nikoli pouze metadaty. Pokud ale má dlouhodobý archiv podniknout takové kroky, aby byl obsah v budoucnu nezávisle srozumitelný uživatelům, nemůže přijímat šifrovaný obsah. Tedy může, ovšem v tom okamžiku se musí vzdát ambice na dlouhodobou ochranu obsahové informace ve smyslu OAIS a musí se smířit s tím, že poskytuje pouze ochranu bitové úrovni obsahu.

5. Důvěryhodnost vs. důvěryhodnost

Pojem důvěryhodnost má velmi různé konotace a nepoužívá se v oblasti archivace a repozitářů jednoznačně. Ve vlastním textu OAIS není důvěryhodnost definována přímo, norma ISO 16363:2012 definuje důvěryhodný repozitář jako repozitář, který má za cíl “poskytnout spolehlivý a dlouhodobý přístup k uloženým digitálním zdrojům určené komunitě uživatelů, a to dnes a v budoucnosti”. Dosažení a udržování důvěryhodnosti je proces, který vyžaduje a očekává, že:

- repozitář rozumí nebezpečím a rizikům, která jeho systémům a datům hrozí,
- repozitář neustále monitoruje, plánuje a udržuje svoje data a systémy s cílem dlouhodobě data uchovat,

- repozitář (provozující instituce) zveřejňuje výsledky auditu a udržuje transparentní procesy a dokumentaci, dokládající mimo jiné také například to, jak archiv přistupuje k finančnímu managementu, personálními řízení atd.
- na splnění cíle dlouhodobého uchování obsahu spolupracuje s dalšími repozitáři, těmi, kteří ho financují, uživateli a dalšími *stakeholdery*.

Důvěryhodnost je zde definována především z hlediska zajištění trvalé dostupnosti svěřeného digitálního obsahu a je součástí širšího vývoje repozitáře, který směřuje ke větší zralosti při správě svěřených dat a schopnosti je ochránit. Tento proces často vyžaduje postupnou transformaci celé instituce.[14]
Jiný přístup k důvěryhodnosti je vidět například v prezentaci Ing. Širla [15], který je uznávaným českým odborníkem na archivaci v oblasti e-governementu. Ten říká, že:

„*Důvěryhodné úložiště má poskytnout důvěryhodné prostředí pro práci s elektronickými dokumenty založené na objektivních dokazovacích postupech. Dále má zajistit právní nezpochybnitelnost uložených dokumentů*“.

Jak vidno zde v souvislosti s důvěryhodností není žádný odkaz na dlouhodobost. Širlovo dlouhodobé úložiště je dlouhodobé „po celou dobu životního cyklu vašich dokumentů…, tedy kdykoliv je budete potřebovat.“ Základní koncepty OAIS v prezentaci Ing. Širla zcela chybí, ale pojmy „dlouhodobé“ a „důvěryhodné“ jsou používány hojně. Naopak z výše uvedené definice je vidět důraz na důvěryhodnost jako v podstatě bezpečnost prostředí archivu, a na objektivní prokazovací postupy a právní nezpochybnitelnost. To odpovídá výše uvedenému popisu zajištění autenticity dat ve firemních archivech (certifikované podpisy apod.). V prostředí správních a firemních dat jsou to zcela jistě legitimní požadavky, ovšem jejich smysl pro kulturní data je velmi malý. Objektivně a právně nezpochybnitelně prokazovat původ nebo nezměněnost nebo to, že nikdo nepovolaný nepřistoupil ke skenovaného časopisu z 19. století má jen malý význam. Ovšem například u smluv nebo úředních dokumentů může mít schopnost prokázat nezpochybnitelně autenticitu dokumenty smysl i po uplynutí skutečně dlouhé doby. Mechanismy pro certifikaci a časové razítkování obsahu musí být připraveny pro skutečně dlouhodobé fungování, ovšem nemá smysl je aplikovat všude na všechny typy dat.
Systémy pro důvěryhodné dlouhodobé archivy kulturních dat tak nemusí vypadat jako dostatečně objektivně prokazatelně důvěryhodné pro archivy správní a firemní, systémy pro firemní a správní archivy zase nemusí kulturním připadat dostatečné z pohledu poskytování dlouhodobého přístupu k informacím. Základní rozpor lze stále vidět v tom, že např. knihovníci vidí dlouhodobou ochranu dat ve smyslu schopnosti zpřístupnit *obsah* dokumentu v budoucnu s tím, že je možné dohledat všechny na něm provedené změny, jejich důvody, tedy prokázat autenticitu obsahu. Archiváři a firmy častěji vnímají dlouhodobou ochranu jako ochranu samotného dokumentu (digitálního souboru), který naopak nesmí doznat žádné změny. V oblasti dlouhodobé archivace digitálních informací bychom se měli držet definice důvěryhodnosti založené na ISO 16363.

6. OAIS: kam s ní?

Výše uvedené příklady ukazují, že i když norma OAIS je dnes stále základem pro většinu úvah o dlouhodobé archivaci digitálních informací, neexistuje jednotná interpretace této normy, ani jednotný

pohled na to, jaké technické prostředky použít k realizaci jejích požadavků. Tvrzení, že archiv odpovídá OAIS, je samo o sobě nic neříkající. Moderní technologie a rozšiřující se cloudové služby staví provozovatele dlouhodobých archivů před nová rozhodnutí a OAIS jim v orientaci příliš nepomůže. OAIS je velmi obecně formulovaný referenční model, který navíc povinně předepisuje jen velmi málo obecných požadavků. Implementace OAIS je vždy aplikací v konkrétním prostředí a pro konkrétní typ obsahu. Tvůrce systému pro dlouhodobou ochranu digitálních informací (LTP) se tak musí rozhodnout, jak dlouho chce data (resp. jejich obsah) uchovávat, jakým způsobem a jaké vlastnosti svěřených dat chce uchovat.

Budování LTP systému je balancováním mezi různými možnostmi, které nejsou vždy zcela snadno splnitelné zároveň v jednom informačním systému. Chceme uchovat obsah nezměněný a zcela objektivně prokazatelně autentický, bezpečný nebo vždy v budoucnu použitelný? Má smysl pro naše velmi strukturovaná data používat uložení v cloudu? Počítáme s neustálým používáním obsahu nebo ukládáme skutečně jen pro uložení samo? Jak dlouhý a jak složitý je životní cyklus našich dat, archivem končí nebo začíná?

Když ve Velké Británii zakládali *Digital Curation Centre* (DCC)[16], definovali dlouhodobou ochranu digitálních informací (*digital curation*) jako „interoperabilitu s budoucností“ nebo „komunikaci do budoucnosti“, kde uživatelé toho, co současné repozitáře ukládají, jsou příští generace. OAIS je pak možné chápat jako komunikační model – viz Obrázek 1. Informační zdroj (v OAIS tedy producent) předává zprávu (SIP) do *transmiteru* (OAIS modul Příjem), kde se vytvoří signál (AIP). Na signál působí *noice source* (čas, technologické změny, zastarávání formátů a medií, atd.), pak signál pokračuje do *recieveru* (OAIS modul Zpřístupnění), kde je transformován do podoby zprávy (DIP) a je dodán do určeného cíle (uživatel). Někde kolem je *technical and physical noise* a algoritmus, který umožňuje kódování zprávy do signálu a zpět (*knowledge base*, *representation information*). Doufejme, že navzdory odlišnostem v tom, jak tuto komunikaci s budoucností dnes v různých oblastech archivace chápeme, dorazí přece jen naše zprávy do svých destinací bezpečně a že budou srozumitelné.

Obrázek 1 – Matematická teorie komunikace [17]

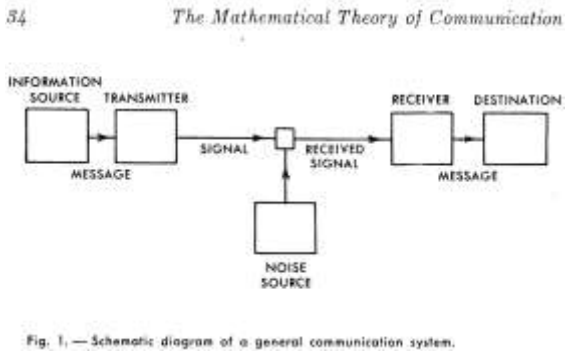

Fig. 1. — Schematic diagram of a general communication system.

7. Literatura a odkazy:

[1]**REICH, Lou a Don SAWYER.1995.***Digital-Archiving Information Services Reference Model* [online]. 14. 9. 1995 [cit. 2012-07-29]. Dostupné z: http://web.archive.org/web/19970122200753/http://bolero.gsfc.nasa.gov/nost/isoas/us01/p004.html

[2]**CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEMS. 2002.***Reference Model for an Open Archival Information System (OAIS): CCSDS 650.0-B-1. Blue Book*[online]. Washington (DC): Consultative Committee for Space Data Systems, January 2002 [cit. 2012-07-29]. 148 s. Dostupné z: http://public.ccsds.org/publications/archive/650x0b1.PDF

CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEMS. 2012.*Reference Model for an Open Archival Information System (OAIS): CCSDS 650.0-M-2. Magenta Book*[online]. Washington (DC): Consultative Committee for Space Data Systems, June 2012 [cit. 2012-07-29]. 135 s. Dostupné z: http://public.ccsds.org/publications/archive/650x0m2.pdf

[3]**CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEMS. 2001.***Reference Model for an Open Archival Information System (OAIS): CCSDS 650.0-P-1.1. Pink Book*[online]. Washington (DC): Consultative Committee for Space Data Systems, August 2001[cit. 2012-07-29]. 131 s. Dostupné z: http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206500P11/Attachments/650x0p11.pdf

[4]**LAVOIE, Brian. 2004.** *The Open Archival Information System Reference Model: Introductory Guide: Technology Watch Report* [online]. Dublin (OH): OCLC & DPC, 2004 [cit. 2012-07-29]. 20 s. DPC Technology Watch Series Report 04-01. Dostupné z: http://www.dpconline.org/docs/lavoie_OAIS.pdf

[5] **OCLC/RLG PREMIS WORKING GROUP. 2004.***Implementing Preservation Repositories for Digital Materials: Current Practice And Emerging Trends In The Cultural Heritage Community: Report by the joint OCLC/RLG Working Group Preservation Metadata: Implementation Strategies (PREMIS)* [online]. Dublin (OH): OCLC, September 2004 [cit. 2012-07-29]. s. 26-27. Dostupné z: http://www.oclc.org/research/activities/past/orprojects/pmwg/surveyreport.pdf

[6] **LOCKSS. 2004.***Formal statement Of Conformance to ISO 14721:2003* [online]. Stanford, CA: LOCKSS, 2004 [cit. 2012-07-29]. Dostupné z: http://www.lockss.org/locksswp/wp-content/uploads/2011/11/OAIS-LOCKSS-Conformance.pdf

[7]**Trusted Digital Repository**. **2010.** In: *TrustedDigitalRepository.eu* [online]. TrustedDigitalRepository.eu, 2010 [cit. 2012-07-29]. Dostupné z: http://www.trusteddigitalrepository.eu/Site/Trusted%20Digital%20Repository.html

[8] **MASSOL, Marion, Olivier, ROUCHON a Lorène BECHARD. 2011**. Certification and Quality: A French Experience. In *iPRES 2011: 8th International Conference on Preservation of Digital Objects, 1.-4.11.2011, Singapore*. Singapore: National Library Board Singapore & Nanyang Technology University, 2011. s. 11-19. ISBN 978-981-07-0441-4. Dostupné také z:http://getfile3.posterous.com/getfile/files.posterous.com/temp-2012-01-02/dHqmzjcCGoexvmiBzJDCyhrlhIgswoffzvsfnpEAxjHFEesarvwahEHrmyvj/iPRES2011.proceedings.pdf

[9]**ASKHOJ, Jan, Mitsuharu NAGAMORI a Shigeo SUGIMOTO. 2010.***Reconsidering the OAIS Reference Model for Record Management and Archiving in a Cloud Computing Environment* [online]. Edinburgh: Digital Curation Centre, 2010 [cit. 2012-07-29]. Dostupné z: http://www.dcc.ac.uk/webfm_send/301

[10]**RABINOVICI-COHEN, Simona. 2010.***Self-contained Information Retention Format (SIRF): Use Cases and Functional Requirements* [online]. Working Draft. Version 0.5a. San Francisco, CA: Storage Networking Industry Association, 2010 [cit. 2012-07-29]. 29 s. Dostupné z:https://www.snia.org/sites/default/files/SIRF_Use_Cases_V05a_DRAFT.pdf

RABINOVICI-COHEN, Simona et al. 2010.*Towards SIRF: Self-contained Information Retention Format* [online]. 2010 [cit. 2012-07-29]. 10 s. Dostupné z: https://www.research.ibm.com/haifa/projects/storage/datastores/papers/systor56-rabinovici-cohen.pdf

[11]**TESSELLA. 2012.** *Preservica: Digital Preservation as a Service* [online]. Abingdon, UK: TESSELLA, 2012 [cit. 2012-07-29]. 2 s. Dostupné z: https://www.digital-preservation.com/wp-content/uploads/Paas-Description-V3-Alternate-Web.pdf

[12]**WOLF, Julia. 2011.***OMG WTF PDF* [online]. 2011 [cit. 2012-07-29]. Video dostupné na YouTube: http://youtu.be/54XYqsf4JEY

[13] **HARVEY, Ross. 2010.***Digital curation: a how-to-do-it manual.* New York: Neal-Schuman, 2010. xxii, 225 s. How-to-do-it manuals for libraries, no. 170. ISBN 978-1-55570-694-4.

[14] **STOKLASOVÁ, Bohdana et al. 2012**. Czech National Digital LibraryEconomic, Strategic and International Aspects of Digital Preservation[online]. In *Aligning National Approaches to Digital Preservation*. Atlanta: Educopia Institute Publication, 2012 [cit. 2012-07-29]. s. 255-269. ISBN 978-0-9826653-1-2. Dostupné z:http://educopia.org/public/resources/ANADP/ANADP_pre_print_07242012.pdf

[15] **ŠIRL, Miroslav. 2009.** Důvěryhodné úložiště elektronických dokumentů [online]. In *15. sympozium EDI, 2009*. Praha: Hospodářská komora ČR, 2009 [cit. 2012-07-29]. 18 slidů PowerPointová prezentace. Dostupné z: http://www.komora.cz/Files/FITPRO/Prezentace/06_Edifact_SIRLv4.pdf

[16] **RUSBRIDGE, Chris et al. 2005.**The Digital Curation Centre: A Vision for Digital Curation [online]. Edinburgh: Digital Curation Centre, 2005 [cit. 2012-07-29]. 11 s. Dostupné z: http://eprints.erpanet.org/82/01/DCC_Vision.pdf

[17] **SHANNON, C. E. 1948.** A mathematical theory of communication. *The Bell System Technical Journal*. 1948, Vol. 27, July October, s. 379–423, 623–656. Dostupné také z: http://cm.bell-labs.com/cm/ms/what/shannonday/shannon1948.pdf

8. Zmíněné normy

ISO 14721, Reference Model for an Open Archival Information System (OAIS)
viz Literatura a odkazy

ISO 16919,Requirements For Bodies Providing Audit And Certification Of Candidate Trustworthy Digital Repositories
http://public.ccsds.org/sites/cwe/rids/Lists/CCSDS%206521R1/Attachments/652x1r1.pdf

ISO 16363, AUDIT AND CERTIFICATION OF TRUSTWORTHY DIGITAL REPOSITORIES
http://public.ccsds.org/publications/archive/652x0m1.pdf