Doporučení pro uchovávání dat pořízených v programu VISK 7

Tento text doplňuje základní prohlášení uvedené v textu výzvy k dotačnímu programu VISK 7 ve znění „Příjemce dotace musí zajistit uchování vytvořených dat samostatně, s nejvyšší možnou péčí, a to bez ohledu na předání dat do Národní knihovny ČR.“ Toto doporučení nevyvolává předkladatelům projektu žádné mandatorní povinnosti. Obsahuje souhrn doporučení k základním pravidlům ochrany digitálního obsahu, jež by měla postupně být integrována do interních pravidel všech organizací participujících na digitalizaci svých fondů nejen s ohledem na nově vytvářený obsah, ale i na starší data, jež tvoří důležitou součást služeb, které nabízejí svým uživatelům. Tímto doporučením se nijak nemění základní parametry programu VISK 7, včetně práv a povinností Národní knihovny ČR. Digitalizující organizace by tedy nadále měla v první řadě poskytovat maximální možnou součinnost pro zajištění podmínek dlouhodobé ochrany dat v Národní knihovně ČR – zejména důsledným dodržováním standardizace NDK a poskytnutím součinnosti při uzavírání smluv formálně definujících práva a povinnosti Národní knihovny ČR a digitalizující instituce vůči datům odevzdaným do Národní knihovny ČR (v souladu s normou ČSN ISO 14721).

Předmětem doporučení je tak zejména sdílení osvědčených postupů (best practices) pro zajištění bitové ochrany, a zajištění vyšší úrovně ochrany dat v souladu s normou ČSN ISO 14721.

1. BITOVÁ OCHRANA

Bitová ochrana dat je nedílnou součástí správy digitálního obsahu. Rozumí se jí uchovávání všech souborů tvořících digitalizovaný obsah (dat, metadat, souborů obsahujících kontrolní součty ad.) v původní podobě tak, aby byla zajištěna jejich ochrana před nedovoleným smazáním, zničením, zastaráním nebo degradací datových nosičů, na kterých jsou uloženy a zajištěna jejich dostatečná organizace tak, aby soubory byly vždy dohledatelné a kompletní. Nezbytnou součástí bitové ochrany je dostatečné finanční zajištění instituce a opatření proti bezpečnostním rizikům (záložní napájecí zdroje, ochrana proti počítačovým útokům, monitorování infrastruktury, omezení přístupu do prostor s úložnými datovými nosiči apod.). Příjemce by měl vzniklá data uchovávat minimálně na této úrovni. Bitová ochrana by měla probíhat v souladu s vnitřní politikou instituce týkající se bitové ochrany. Do politiky bitové ochrany patří také střednědobé odhady nákladů na uchování dat. Příjemce by se měl vykonávání bitové ochrany přinejmenším řídit mezinárodními normami pro řízení informační bezpečnosti (normy z řady ISO/IEC 27001) a řízení kvality (normy z řady ISO 9000, resp. ISO 9001).

Bitová ochrana může být realizována vlastními prostředky instituce příjemce nebo využitím externí služby veřejně dostupné infrastruktury (například datová úložiště CESNETu nebo Krajská úložiště či prostředky z jiných veřejných projektů). Realizace tohoto stupně ochrany je nezávislá na dále zmiňovaných vyšších stupních ochrany.

Doporučené postupy bitové ochrany

Bez ohledu na objemy dat nebo použité technologie je obecně doporučovanou metodou bitové ochrany udržování několika vzájemně nezávislých kopií dat a pravidelné kontroly jejich použitelnosti a konzistence. Předpokladem trvalého ukládání dat je také evidence jednotlivých kopií.

Základní vstupní parametry

1) Více kopií

Doporučujeme udržování dvou, ideálně tří kopií uložených balíků. Jako třetí kopii je možno považovat kopii odevzdanou do Národní knihovny ČR.

2) Nezávislost kopií (jiná lokalita, jiné technologie)

Doporučujeme udržování jednotlivých kopií dat na různých typech datových nosičů a technologií pro uchovávání dat (disková pole, páskové knihovny, off-line zálohy, cloudová úložiště atd.), neboť použití podobných úložných technologií zvyšuje riziko ztráty dat v důsledku stejné chyby. Jinými slovy: ideální je uložení dat ve dvou či více odlišných datových centrech (tedy v geograficky vzájemně vzdálených lokalitách), za použití úložných technologií dvou či více různých výrobců.

3) kontroly použitelnosti/neměnnosti kopií

Periodické kontroly uložených kopií vyžadují značnou lidskou práci a výpočetní výkon. Instituce by měly mít plán rozsahu pravidelných kontrol integrity (neporušenosti) uložených dat a dokumentované postupy pro provádění takových kontrol. Kontroly integrity lze realizovat na několika úrovních – jak na úrovni celého informačního balíčku, tak na úrovni jednotlivých souborů.

Doporučené hashovací algoritmy jsou md5, sha-1, crc32 případně obdobně spolehlivé mechanismy. Kontrola integrity celého informačního balíčku může například probíhat každé dva roky, kontrola použitelnosti jednotlivých souborů pak v rozsahu alespoň 10% obsahu úložiště ročně.

Existuje celá řada nástrojů pro generování a kontrolu integrity, například DROID, MD5 summer, python cavpp/md5tool, Linux md5sum nebo Java MessageDigest.

4) organizace a evidence kopií

Evidence kopií je předpokladem schopnosti instituce úspěšně využít zálohovaná data. Minimálním požadavkem by měla být evidence balíků a médií v Excelu. V případě kopií uložených na jiných než offline nosičích je nezbytné využívat zavedené softwarové systémy pro správu obsahu.

2. LOGICKÁ OCHRANA DAT

Bitová ochrana dat je nutným předpokladem pro logickou ochranu dat, která představuje vyšší stupeň ochrany, kterou mohou provádět pouze kvalifikovaní odborníci. Další nezbytné kroky nezbytné pro možnost vykonávání logické ochrany dat, jako jsou identifikace, validace a charakterizace souborových formátů určených k uchovávání jsou požadovány již při samotném vytváření dodavatelských balíčků SIP (PSP). Rovněž dodržení doporučených postupů při vytváření datových souborů (zejména archivních kopií ve formátu JP2) vede k vysoké míře homogenity a validity vytvářených dat již při produkci a je prvním a nezbytným krokem k úspěšné dlouhodobě vedené logické ochraně dat.

Pro zajištění plnohodnotné logické ochrany ukládaných dat je již potřeba provozovat tzv. LTP (long-term preservation) systém splňující standardy norem ISO 16363 a ISO 14721. Vybudování a provozování vlastního LTP systému je odborně, organizačně i finančně náročné a lze ho doporučit jen těm institucím, které disponují v dlouhodobém časovém horizontu dostatečnými finančními, personálními a odbornými zdroji. Zajištění logické ochrany odevzdaných dat proto zajišťuje v první řadě Národní knihovna ČR provozující svůj vlastní systém pro dlouhodobou archivaci digitálního obsahu. Mimo tohoto řešení jsou k dispozici další komerční (například Rosetta, Tessela) i open-sourcová řešení (například Archivematica), které však pro své použití vyvolávají potřebu výše uvedených zdrojů. Jakákoli ochranná opatření v oblasti logické ochrany (např. formátová migrace) vykonávaná institucí, která svá data uchovává samostatně, je tak s ohledem na efektivní vynakládání veřejných prostředků vždy lepší předem konzultovat s odbornými pracovišti Národní knihovny ČR, zejména s Odborem novodobých digitálních sbírek.

3. DOPORUČENÝ OBSAH PLÁNU BITOVÉ OCHRANY

Stručný plán bitové ochrany, který by měl být obsažen v interní politice žadatele a být dostatečně podrobně zdokumentován**,** by měl obsahovat:

1) popis způsobu uložení - kolik kopií, na jakých technologiích, v jakých lokalitách, způsob vytvoření kopií, způsob využití externí služby, případně další relevantní informace

2) popis kontrol neměnnosti a použitelnosti dat – frekvence kontrol integrity informačních balíčků a kontrol integrity jednotlivých souborů, metodika kontroly úložných datových nosičů a technologií.

3) popis způsobu evidence kopií - použité identifikátory, způsob zavedení do evidence, odpovědnost za udržování evidence v aktuálním stavu, kontroly samotné evidence

Pokud žadatel podobný plán ve svých interních pravidlech pro bitovou ochranu již má zaveden a zdokumentován, bylo by vhodné, aby jej zmínil ve své žádosti. Odborní pracovníci Národní knihovny ČR po seznámení s těmito pravidly mohou doporučit další úpravy či doplnění, která dále přispějí ke zvýšení kvality ochrany digitálního obsahu.