

Závěrečná zpráva

ŘEŠENÍ PROCESU DLOUHODOBÉ ARCHIVACE DAT V NÁRODNÍM MUZEU

Projekt Fondu rozvoje CESNET

Instituce:	Národní muzeum
Sídlo:	Václavské náměstí 68, 110 00, Praha 1
Řešitel:	Ing. Martin Souček, Ph.D.
Spoluřešitelé:	Mgr. Adam Cironis, Filip Šír, DiS., Lukáš Bělka

Oblast:	I
Tematický okruh:	B

Obsah

Terminologie	3
Postup při řešení, způsob řešení.....	3
Výchozí stav	3
Architektura pracovní linky	4
Popis použitého procesu	4
Struktura archivačního balíčku	5
Struktura BagIt balíčku.....	6
Struktura katalogizačního záznamu	6
Diagram uložení do systému LTP CESNET za pomoci nástroje ProArc	8
Rozšíření systému ProArc o nové funkce.....	8
Vytvoření BagIt balíčku	8
Export BagIt balíčku do LTP CESNET	8
Možnost zápisu o uložení v LTP do záznamu v systému Verbis	9
Dosažené cíle	9
Zdůvodnění případných změn v projektu	9
Výstupy (konkrétní výstupy, další využitelnost)	9
Přínosy projektu, vlastní zhodnocení.....	10
Tisková zpráva	10

Terminologie

ID nosiče – jedná se o unikátní identifikátor, který je použit pro jeden zvukový nosič. V našem případě se jedná o kombinaci písmen a čísel např. NMCMH00001

SIP – Submission Information Package – jedná se o balíček digitálních objektů pojmenovaných dle unikátního ID, obsahuje obrazové a zvukové digitalizáty, technická metadat, OCR soubory. Národní muzeum ho používá jako vstupní balíček dat importovatelných do ProArc

AIP/SIP-NDK – Archival Information Package – Obecné pojmenování pro archivační balíček. V rámci systému ProArc dle aktuálního standardu se jedná od tzv. SIP-NDK balíček. Tento balíček je pak zabalen do BagIt a poslán do LTP CESNET systému.

UUID – Universal Unique Identifier – slouží k jednoznačné identifikaci digitálních objektů v systému ProArc a následných balíčků.

ProArc – Je produkční a archivační systém, který umožňuje řídit dokumentaci a dodržovat nejlepší postupy pro kontrolu dokumentů. Systém poskytuje nástroje pro řízení změn, plánování a sledování pokroku projektů a spolupráci.

Kramerius – Je digitální knihovna využívaná ke zpřístupňování zdigitalizovaného obsahu v Národním muzeu. Obsah je zpřístupněn koncovému uživateli dle autorského zákona.

Postup při řešení, způsob řešení

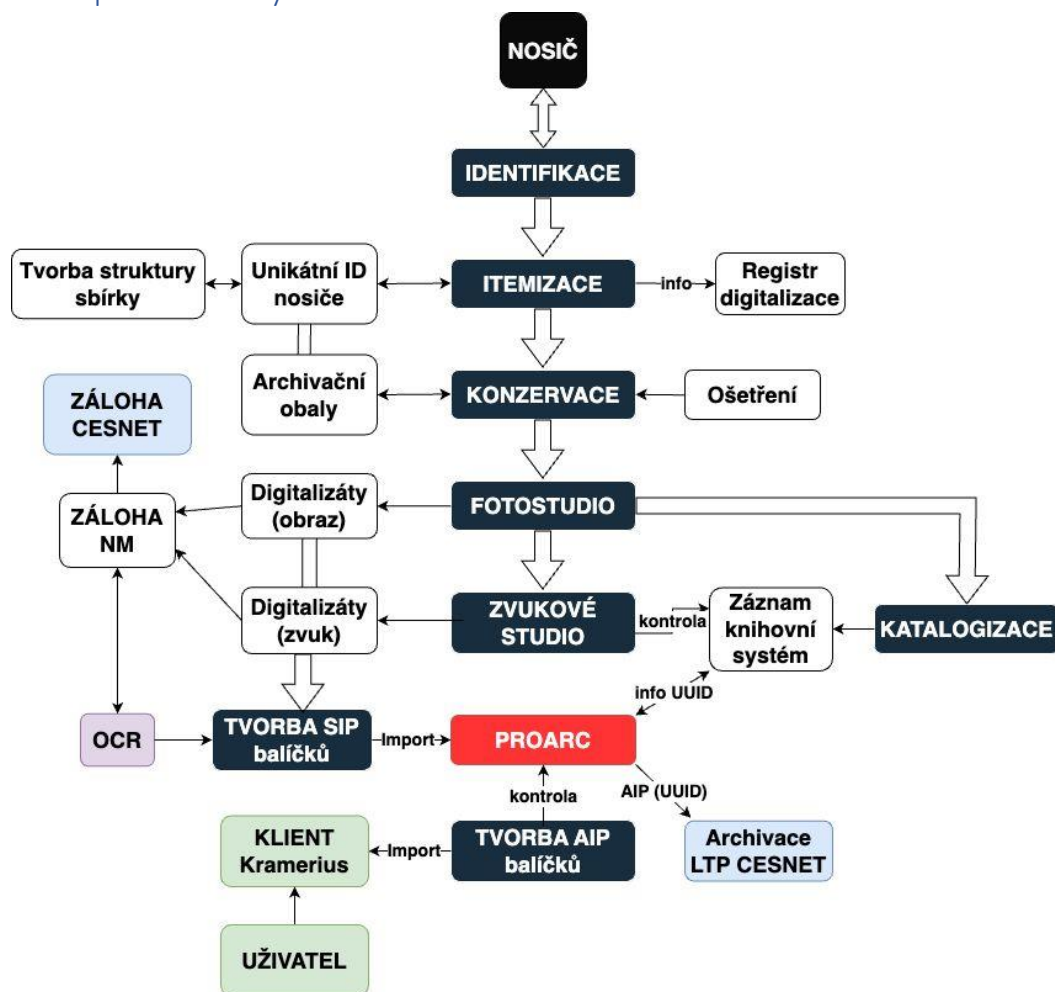
Výchozí stav

Národní muzeum ve svých sbírkách uchovává velké množství zvukových nosičů a je tedy na místě, že se koncepčně zabývá jejich záchranou. Záchrana probíhá nejen směrem k obsahu zvukových nosičů, ale i směrem ke stránce fyzické. Vzhledem k tomu, že se jedná v mnoha případech o unikátní nahrávky na fonografických válečkách nebo o komerčně vydané gramofonové desky s bohemikálním obsahem, bylo potřeba vybudovat takové pracoviště – Zvuková laboratoř Národního muzea – které dokáže obsáhnout celý proces záchrany obsahu a formy. Identifikovat, ochránit, zdigitalizovat a zpřístupnit daný obsah.

Zvuková laboratoř provádí nejen samotné ochranné reformátování analogových nosičů, ale celý řetězec se podařilo vybudovat tak, aby ve všech krocích bylo možné dodržet nejvyšší standardy a doporučení uznávané na mezinárodní úrovni. Současně s tím se pracovišti podařilo vytvořit metodologické postupy, které umožní za pomoci unikátního identifikátoru držet informaci o digitalizátu a fyzickém nosiči na jednom místě v tzv. trackovacím systému. Linka tedy umožňuje digitalizovat zvukové a obrazové dokumenty, vytvářet a zpracovávat všechny typy metadata a v neposlední řadě kompletace v nástroji pro tvorbu tzv. archivačních balíčků (AIP/SIP-NDK). K uzavření celého procesu záchrany je potřeba daný obsah zpřístupnit a dlouhodobě uchovat. Ke zpřístupnění obsahu nám slouží digitální knihovna Kramerius, kde se daný titul zpřístupní dle autorského zákona.

Dlouhodobá archivace za pomoci systému LTP byla poslední částí z celého řetězce, která před samotným projektem nebyla vyřešena. Proto bylo přistoupeno k tomuto projektu a snaze o nalezení nejlepšího řešení pro dlouhodobou archivaci.

Architektura pracovní linky



Obrázek 1: Architektura pracovní linky

Popis použitého procesu

Celý proces začíná kompletací digitalizátů (audio a obraz) s potřebnými soubory z OCR a následně jsou soubory vloženy do SIP balíčku. Po importu těchto vstupních dat do systému ProArc se v rámci na tvorby struktury archivačního balíčku stáhne z knihovního systému Verbis potřebný katalogizační záznam – popisná metadata, a současně se načtou technická metadata z obrazu a audia. Po vytvoření samotné struktury balíčku je dokument/titul připraven ke zpřístupnění v digitální knihovně Kramerius a k následné archivaci. V rámci importu samotného balíčku (SIP-NDK) do digitální knihovny se využívá pouze určitá část dat k prezentaci. Digitální knihovna Kramerius, kterou Národní muzeu využívá ke zpřístupňování sbírkových předmětů, především povahy tištěné, byla v posledních letech rozšířena i možností poslechu zvukových nahrávek, a proto se v rámci strategie NM přistoupilo k využívání této digitální knihovny i ke zpřístupňování tohoto typu sbírkových předmětů.

Pro dlouhodobou archivaci se využívá služba LTP provozovaná sdružením CESNET, která slouží jako archiv, do kterého se ukládají speciálně připravené archivační balíčky. Tato služba pracuje s archivačními balíčky odpovídajícími formátu BagIt. Je proto potřeba nakonfigurovat systém ProArc pro práci s balíčky ve formátu BagIt a s jeho pomocí vygenerovat speciální archivační balíček odpovídající tomuto formátu, aby bylo možné nahrát ho do LTP úložiště od CESNET. Tento balíček obsahuje několik pomocných souborů, kontrolní součet a adresář "data" (struktura BagIt balíčku popsána níže). Po nahrání archivačního balíčku provede automatizovaný systém LTP CESNET validaci: prověří vnitřní kontrolní součty z metadat balíčku a jeho celkový formát. Z ProArc je nejen možné vytvořit BagIt balíček, ale následně zaslat ho přímo do LTP úložiště CESNET viz Technická dokumentace.

Po dokončení exportu BagIt balíčku do LTP CESNET prostřednictvím ProArc je možné ve webovém rozhraní LTP CESNET najít balíčky podle jejich UUID, které bylo přiděleno v systému ProArc. Je zcela zásadní po exportu balíčku vždy zkontrolovat v LTP CESNET, že je balíček systémem korektně přijat. U každého balíčku je možné rozkliknout jeho detaily kliknutím na hlavičku záznamu. Jsou zde obsaženy informace o nahrání, velikosti, kontrolním součtu apod. Zde lze samotný balíček stáhnout.

Následně je možnost ze systému ProArc poslat informaci o uložení v LTP CESNET do knihovního systému Verbis. Tedy, podle ID záznamu se do samotného záznamu propíše do pole 856 informace o identifikátoru použitým v LTP CESNET – UUID, případně informace o digitalizaci.

Proces exportu balíčků do LTP CESNET poté probíhá primárně na zařízeních pořízených v rámci spoluúčasti Národního muzea při financování tohoto projektu.

Struktura archivačního balíčku

Archivační balíček (AIP/SIP-NDK) využívá standardu METS a obsahuje v sobě zdrojové, archivní i prezentační kopie mediálních souborů a několik druhů metadat. Formáty mediálních obsahů odpovídají v současném provozu standardu Národní knihovny pro gramofonové desky a jsou popsány v rejstříku formátů. Pro zvuk jde o WAV, obrázky se ukládají ve formátu jpeg2000 a je nad nimi prováděno OCR s výstupy ve formě textu a popisného XML. Metadata se ukládají v XML souborech. Struktura je vidět na obrázku.

alto	Složka souborů	
amdsec	Složka souborů	
mastercopy	Složka souborů	
mastercopy_audio	Složka souborů	
source_audio	Složka souborů	
txt	Složka souborů	
usercopy	Složka souborů	
usercopy_audio	Složka souborů	
info_abe370-000192	Dokument ve formátu XML	2 kB
md5_abe370-000192.md5	Soubor MD5	1 kB
mets_abe370-000192	Dokument ve formátu XML	17 kB

Obrázek 2: Struktura archivačního balíčku AIP/SIP-NDK

AIP/SIP-NDK balíček obsahuje následující části:

Název složky/souboru	Popis
alto	XML popis rozložení OCR
amdsec	Administrativní metadata k obsahu balíčku
mastercopy	Archivní kopie obrázků
mastercopy_audio	Archivní kopie audia
source_audio	Zdrojová kopie audia
txt	Text OCR
usercopy	Uživatelská kopie obrázků
usercopy_audio	Uživatelská kopie audia
info_ID	Stručné údaje o vzniku balíčku
md5_ID.md5	Seznam kontrolních hashů souborů v balíčku
mets_ID	Hlavní METS záznam

Struktura BagIt balíčku

Balíček BagIt má jednoduchou strukturu, která je popsána na následujícím obrázku. Obsahuje povinné části, díky kterým lze identifikovat použitou verzi a ověřit datovou integritu obsahu. V rámci řetězce slouží jako vrstva nad AIP/SIP-NDK balíčkem.

 data	Složka souborů	
 bag-info	Textový dokument	1 kB
 bagit	Textový dokument	1 kB
 manifest-md5	Textový dokument	1 kB
 tagmanifest-md5	Textový dokument	1 kB

Obrázek 3: Struktura BagIt balíčku

BagIt balíček obsahuje následující části:

Název složky/souboru	Popis
data	Složka s daty AIP balíčku
bag-info	Základní metadata BagIt balíčku
BagIt	Označení verze standardu BagIt
manifest-md5	Seznam kontrolních hashů souborů ve složce data
tagmanifest-md5	Seznam kontrolních hashů metadataových souborů

Struktura katalogizačního záznamu

Katalogizační záznam ve standardu MARC 21 je používán pro popis zvukových nosičů v katalogizačním systému Verbis. Celá struktura [ukázkového záznamu](#) je vidět v tabulce.

Pole	Ind	Obsah pole
leader		-----njm-a22-----i-4500

1 kpw0887016

5 20231011171146.4

7 se imuul||ipuna

8 200624q19001915xx |||||cze||

40 ## \$a ABE370 \$b cze \$e rda

41 0# \$d cze

72 #7 \$a 782 \$x Divadelní hudba. Opera, opereta, muzikál \$2 Konspekt \$9 9

100 1# \$a Dvořák, Antonín, \$d 1841-1904 \$7 jn19981000826 \$4 cmp

245 10 \$a Rusalka ? \$n I., \$b arie Víím, že jsi kouzlo, které mine / \$c [Antonín Dvořák]

264 #1 \$a [Místo vydání není známo] : \$b [nakladatel není známý], \$c [mezi 1900 a 1915?]

300 ## \$a 1 fonováleček : \$b analog, 160 ot/m, mono ; \$c průměr 54 mm, délka 109 mm

336 ## \$a hraná hudba \$b prn \$2 rdacontent

337 ## \$a audio \$b s \$2 rdamedia

338 ## \$a audioválec \$b se \$2 rdacarrier

500 ## \$a Název z krabičky fonoválečku

500 ## \$a Název arie z poslechu fonoválečku

511 0# \$a Pták, zpěv

562 ## \$b MH 715 \$3 Název na krabičce psán rukou

685 \$a fonoválečky

700 1# \$a Pták, Bohumil, \$d 1869-1933 \$7 ola200202882 \$4 sng

856 4# \$d zdigitalizováno \$x proarcld \$y uuid:0d8c130c-a170-4ca7-a20b-c4b8870cea5d

910 ## \$a ABE370 \$b MH 715

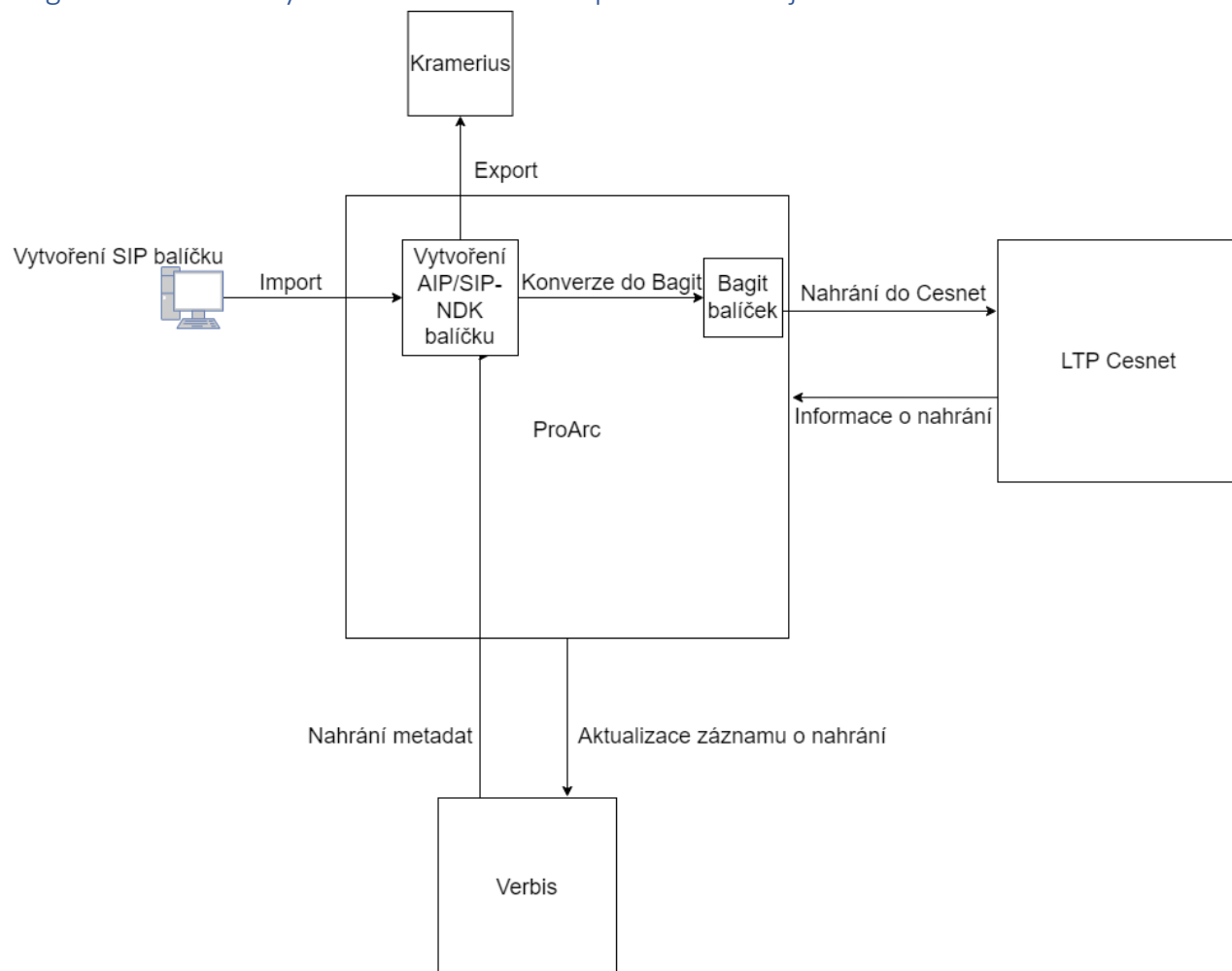
V rámci dlouhodobé archivace je potřeba propsat do metadat informaci o uložení do úložiště LTP CESNET. To je zaznamenáno v **poli 856 - Elektronická lokace a přístup**, která se jinak využívá primárně na určení lokace elektronického zdroje. Struktura položky je popsána níže.

856	4#	\$d zdigitalizováno \$x proarcld \$y uuid:0d8c130c-a170-4ca7-a20b-c4b8870cea5d
------------	-----------	--

Indikátor	Hodnota	Význam
Metoda přístupu	4	HTTP
Vztah	#	nespecifikován
Subfield	Kód	Hodnota
Cesta k souboru	d	Text “zdigitalizováno”

Neveřejná poznámka	x	Text “proarcId”
Text odkazu	y	uuid:<unikátní uuid předmětu>

Diagram uložení do systému LTP CESNET za pomoci nástroje ProArc



Obrázek 4: Diagram procesu uložení do LTP CESNET

Rozšíření systému ProArc o nové funkce

Vytvoření BagIt balíčku

Po vytvoření AIP/SIP-NDK balíčku je nově možnost v ProArc vytvořit BagIt balíček, který je požadovaným formátem služby LTP provozované sdružením CESNET.

Export BagIt balíčku do LTP CESNET

Po vytvoření BagIt balíčku nám ProArc umožňuje přímý export do uložště LTP CESNET.

Možnost zápisu o uložení v LTP do záznamu v systému Verbis

Potom, co je BagIt balíček úspěšně nahraný v LTP CESNET a systém vyhodnotil balíček jako validní, je možnost nahrát informaci o úspěšném uložení do knihovnického systému Verbis do pole číslo 856.

Dosažené cíle

Cílem projektu bylo vytvořit komplexní archivační řetězec pro archivaci, včetně dlouhodobého uložení zvukových nosičů. Řešení mělo navázat na již probíhající digitalizační aktivity v Národním muzeu a vznikalo ve spolupráci se sdružením CESNET, které takovýto LTP systém poskytuje.

Základní snahou bylo na příkladu zvukových nosičů digitalizovaných ze sbírky Národního muzea – Českého muzea hudby ukázat metody, jakými rozšířit stávající postupy digitalizace o dlouhodobou archivaci a připravit tak modelový příklad, který lze snadno rozšířit na další snahy v oblasti dlouhodobé archivace sbírkových předmětů v Národním muzeu, ale i v dalších paměťových institucích. Tento postup byl ověřen v rámci pilotního provozu a zdokumentován. Snahou Národního muzea je nejen tento provoz zpřístupnit online, ale také pokračovat v přípravách na jeho certifikaci.

Zdůvodnění případných změn v projektu

Zpočátku bylo v plánu využít software Archivematica, který nachází využití v obdobných procesech v mnoha institucích po celém světě. Důvodem byla i nativní možnost exportu ve formátu BagIt, který byl potřeba jako vstupní do prostředí LTP úložiště CESNETu. Proběhlo několik konzultací ohledně vhodného navázání Archivematicy na do té doby používaná řešení, primárně v systému ProArc. Následně jsme ale v rámci domluv přešli k nápadu modifikovat systém ProArc pro naše použití. Ten je využíván ve více místních institucích a je již vhodně provázán z dalšími systémy, jako je Kramerius a Verbis. Zároveň je celý postup více centralizovaný a není nutné přecházet mezi aplikacemi. Potřebné funkce vytvoření archivačního balíčku i odeslání do úložiště tak byly přidány firmou Inovatika do open-source verze ProArcu, která je aktivně vyvíjena na a zveřejňována na Githubu (<https://github.com/proarc/proarc>).

Vzhledem k předem neočekávaným okolnostem bylo v průběhu projektu nutno přistoupit ke změně hlavního řešitele, a to z důvodu nástupu na mateřskou dovolenou hlavní řešitelky Mgr. art. Eleonóry Stankové za řešitele Ing. Martina Součka, Ph.D.

Výstupy (konkrétní výstupy, další využitelnost)

Výstupem projektu je dokumentace celého archivačního řetězce, od tvorby archivačního balíčku až po jeho uložení k dlouhodobé archivaci v systému LTP CESNET, a to na modelovém příkladu zvukových dokumentů. Tyto postupy byly ověřeny na archivaci desítkách archivních balíčků, které byly vytvořeny v rámci pracovní linky digitalizace zvukových nosičů a úspěšně prošly celým procesem archivace až po jeho dlouhodobé uložení v LTP CESNET.

Zvolená SW řešení jsou open-source a byla publikována v rámci softwaru ProArc verze 4.1.0. Spolu se zveřejněnou dokumentací jsou tak vhodným základem pro implementaci dlouhodobé archivace pro jiné projekty a instituce u nás.

Zároveň proběhl workshop pro pracovníky muzea, ve kterém byli seznámeni s celým průběhem zpracování předmětu a počítá se s tím, že další digitalizační snahy v Národním muzeu již budou reflektovat poznatky zjištěné v rámci tohoto projektu.

Z projektu také vznikly podněty do nové strategické dokumentace, která byla vytvářena v souladu s konceptem dlouhodobého ukládání podle modelu OAIS. Dokumenty jsou ke stažení na webu projektu a jejich užití bude obsahem nově vznikající Směrnice k dlouhodobému uchovávání digitalizovaných sbírkových předmětů v Národním muzeu.

Z důvodu, že systém ještě není nasazen a aktivně využíván v plné míře, tak je self-audit stále aktivně připravován. Při celé přípravě systému bylo přihlíženo k tomu, že Národní muzeum má o certifikaci do budoucna zájem. Specificky se při návrhu bral zřetel na pečeti od organizace Nestor, která je postavena na normě DIN 31644.

Přínosy projektu, vlastní zhodnocení

V rámci hlavního cíle projektu vzniklo koncepční řešení dlouhodobé archivace tzv. AIP balíčků, které jsou vytvářeny v rámci systému ProArc. Toto řešení bylo odzkoušeno na archivaci zvukových dokumentů, resp. jejich archivačních balíčcích, a v rámci zkušebního provozu se potvrdila jeho funkčnost.

Tato úprava nám umožnila dokončit, a především zdokumentovat ucelené pracovní workflow dlouhodobé archivace zvukových dokumentů, jakožto první komplexní provoz pro zvukové nosiče v ČR.

Tyto postupy je možné využít dalšími paměťovými institucemi. Postupy a funkce vzniklé v rámci projektu jsou zpřístupněny online na adrese www.nm.cz/ltp a jsou plně k dispozici okolním institucím zabývajících se touto problematikou.

Tento projekt, a především jeho cíle dopomohly k vylepšení strategických dokumentů Národního muzea a umožnily tak možnost zpracování a následné archivace i ostatních typů sbírkových předmětů.

Tisková zpráva

V rámci projektu Fondu rozvoje CESNET „ŘEŠENÍ PROCESU DLOUHODOBÉ ARCHIVACE DAT V NÁRODNÍM MUZEU“ byl vytvořen a popsán řetězec pro digitální archiv od akvizice předmětu pro dlouhodobé uložení. Systém je popsán na příkladu archivaci zvukových nosičů a využívá k tomu systému ProArc, který nám umožňuje jak tvorbu archivačních balíčků, tak jejich dlouhodobé uložení v LTP systému poskytovaném sdružením CESNET.

Výstupem procesu je uložení archivního balíčku v systému LTP, zpřístupnění v digitální knihovně Kramerius a zaznamenání informace o digitalizaci a archivaci v knihovním systému Verbis. Informace o projektu jsou dostupné na webu www.nm.cz/ltp.