

Technická dokumentace

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

1. Terminologie.....	2
2. Popis použitého procesu	2
3. Rozšíření systému ProArc o nové funkce a jejich konfigurace	7
1. Tvorba BagIt balíčku	7
1. Návod a technické specifikace.....	7
2. Export BagIt balíčku do LTP systému	7
2. Návod a technické specifikace.....	7
3. Zápis informací o uložení v LTP do katalogizačního záznamu	8
3. Návod a technické specifikace.....	9

Terminologie

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.

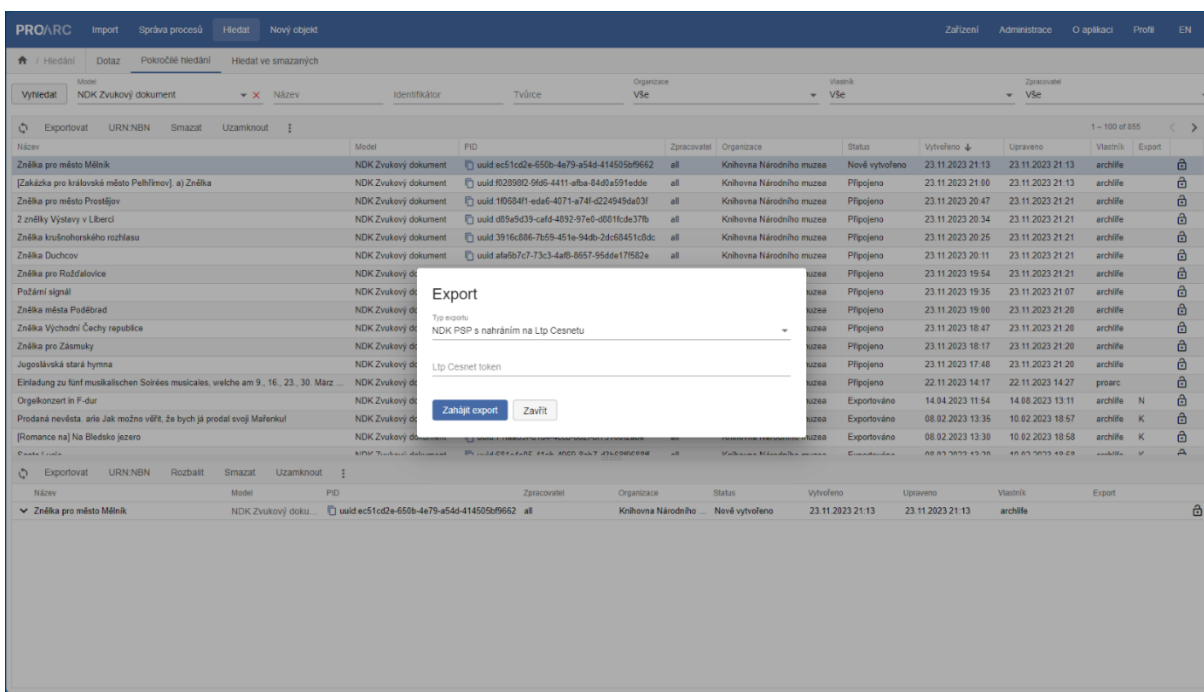
Popis použitého procesu

Celý proces začíná kompletací digitalizátů (audio a obraz) s potřebnými soubory z OCR do SIP balíčku. Po importu SIP balíčku (vstupních dat) do systému ProArc se v rámci na tvorby archivačního balíčku stáhne z knihovního systému Verbis potřebný katalogizační záznam – popisná metadata, a současně se načtou technická metadata z obrazu a audia.

Po vytvoření archivačního balíčku (struktura balíčku je popsána níže), se provádí export samotného balíčku do digitální knihovny Kramerius, kde probíhá import vybrané části kde se nachází uživatelská data použitelná pro prezentaci.

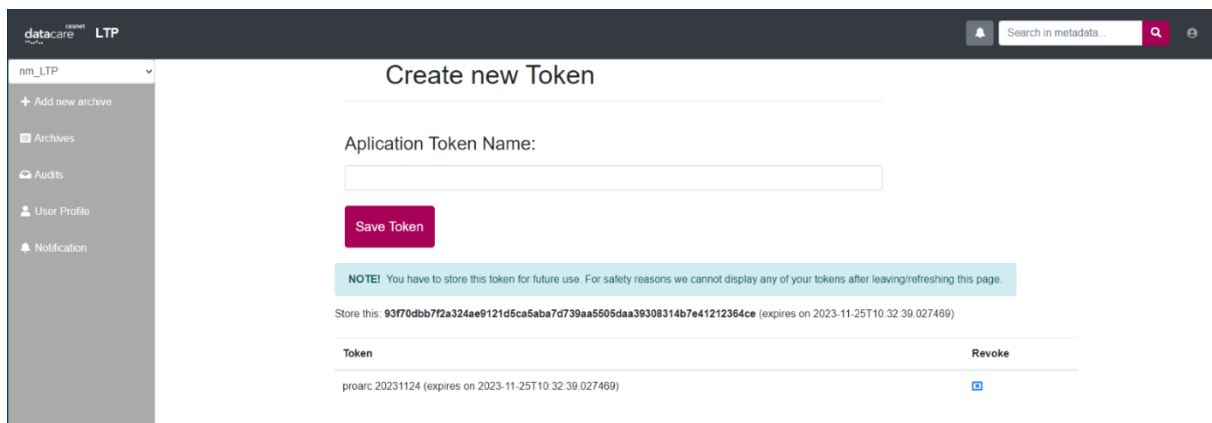
Pro dlouhodobou archivaci se využívá služba LTP provozovaná sdružením CESNET, která slouží jako archiv, do kterého se ukládají speciálně připravené archivační balíčky. Tato služba pracuje s archivačními balíčky odpovídajícími formátu BagIt. Je proto potřeba nakonfigurovat systém ProArc pro práci s balíčky 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. Po nahrání archivačního balíčku provede automatizovaný systém LTP CESNET validaci: prověří vnitřní kontrolní součty z metadat balíčku a jeho celkový formát. Pro další kontroly jeho binární konzistence si systém vytvoří vlastní kontrolní součty. Průběžné kontroly jsou zaznamenávány do auditních zpráv na webu LTP CESNET.

Z ProArc je nejen možné vytvořit BagIt balíček, ale následně zaslat ho přímo do LTP úložiště CESNET.



Obrázek 1: Export balíčku v ProArc

Pro nahrání do LTP úložiště je potřeba vložit token, který slouží jako API klíč. Tento token lze vygenerovat ve webovém rozhraní LTP úložiště CESNET v záložce *User Profile -> Manage tokens*.



Obrázek 2: Vytvoření nového tokenu v LTP CESNET

Je nutné zkopírovat vygenerovaný token na bezpečné místo, protože jej nebude možné z bezpečnostních důvodů znovu zobrazit. Po vložení tokenu v systému ProArc se ze všech parametrů složí příkaz, který se uloží do skriptu a ten se spustí (Nelze spustit příkaz pro nahrání přímo z Javy, protože je potřeba předat parametry v uvozovkách, ale Java to takto nedovolí spustit, proto je to řešeno pomocí využití skriptu. Tento skript je nutné vytvořit ručně a nastavit mu oprávnění na to, aby se dal spouštět).

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. Balíčky mají 4 různé stavy, které se liší barevným rozlišením hlavičky záznamu.

- Šedá barva značí, že záznam byl v pořádku vytvořen a soubor se nahrává.
- Zelená barva značí, že po nahrání proběhly kontroly archivačního balíčku v pořádku.
- Červená barva značí, že kontroly neproběhly správně, nebo se balíček nenahrál z důvodu nesprávného formátu BagIt balíčku. Balíček bude po nějaké době automaticky odstraněn.
- Černá barva značí, že balíček už byl, kvůli neúspěšnému nahrání, smazán.

ID	UUID	Status Integrity	Version	Create date
ID: 651	uuid:930f67c0-c547-4732-ba6c-504cc4b29a4f	Archive is deleted	1	2023-11-28T13:05:25.426062
ID: 643	uuid:0d8c130c-a170-4ca7-a20b-c4b8870cea5d	Unknown	1	2023-11-22T14:55:01.031819
ID: 578	uuid:d1dd68d1-5ecf-46a1-82c9-7c2a38aad04f	Archive is valid	1	2023-08-14T13:12:21.617425
ID: 577	uuid:d1dd68d1-5ecf-46a1-82c9-7c2a38aad04f	Archive is valid	1	2023-08-14T13:08:47.505725
ID: 576	uuid:ed2ae08a-2333-4acb-80e2-c2722171ac9e	Archive is valid	1	2023-08-14T13:05:53.436798
ID: 574	uuid:eebea6d-8bcf-49f8-a2c5-4451f18d3aae	Archive is valid	1	2023-06-08T14:12:14.056765
ID: 573	uuid:ed2ae08a-2333-4acb-80e2-c2722171ac9e	Archive is valid	1	2023-06-07T12:23:37.620304
ID: 572	uuid:0d8c130c-a170-4ca7-a20b-c4b8870cea5d	Archive is valid	1	2023-06-07T12:06:31.200252
ID: 571	uuid:0d8c130c-a170-4ca7-a20b-c4b8870cea5d	Archive is valid	1	2023-06-01T11:36:51.223783
ID: 570	uuid:0d8c130c-a170-4ca7-a20b-c4b8870cea5d	Archive is valid	1	2023-06-01T10:42:36.855795
ID: 569	uuid:0d8c130c-a170-4ca7-a20b-c4b8870cea5d	Archive is valid	1	2023-06-01T10:41:55.612457
ID: 568	uuid:0d8c130c-a170-4ca7-a20b-c4b8870cea5d	Archive is invalid	1	2023-06-01T10:29:34.894228

Obrázek 3: Přehled archivačních balíčků v LTP CESNET

Je tedy zcela zásadní po vkládání balíčku vždy zkontrolovat, že je balíček systémem korektně přijat. Je také možné nastavit si upozornění, v záložce *Notifications*, například pro případ porušení integrity archivního balíčku. 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 pomocí tlačítka *Prepare download*. Archivní balíček musí nejdříve systém připravit pro stažení. Tato příprava zahrnuje jeho vytažení z archivu a kontrolu kontrolního součtu a může trvat několik minut. Jakmile se kontroly dokončí objeví se tlačítko *Download here*, kliknutím dojde ke stažení balíčku.

Rozšíření systému ProArc o nové funkce a jejich konfigurace

Po spuštění aplikačního serveru se v adresáři `$PROARC_HOME` vytvoří soubor `proarc.cfg.template` s přednastavenými parametry systému. Pro vlastní nastavení vytvořte soubor `proarc.cfg` s parametry, které chcete změnit. Případné změny souboru `proarc.cfg.template` budou přepsány.

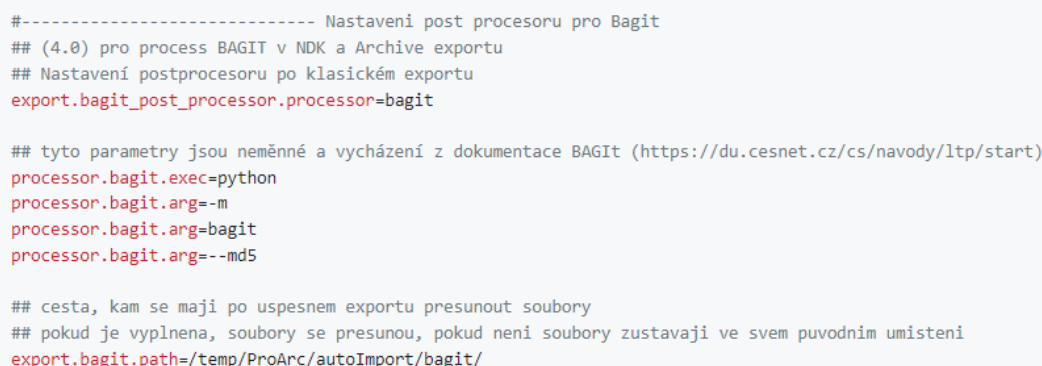
Číslo v závorkách znamená verzi, od které lze nastavení použít.

Tvorba BagIt balíčku

Po vytvoření AIP/SIP-NDK balíčku je nově možnost v ProArc vytvořit BagIt balíček, který je požadovaným formátem služby LTP CESNET. Pro možnost vytvoření těchto balíčků je potřebná následující konfigurace popsaná níže.

Návod a technické specifikace

Pro vytvoření BagIt balíčků je požadována instalace pythonu s rozšířením pip a balíčkem BagIt. Poté je potřeba do konfigurace doplnit konkrétní parametry, které vychází ze samotné dokumentace tohoto formátu (<https://du.CESNET.cz/cs/navody/ltp/start>).

A screenshot of a code editor showing configuration parameters for BagIt. The text is in a light blue box with a copy icon in the top right corner. The code is as follows:

```
#----- Nastavení post procesoru pro BagIt
## (4.0) pro process BAGIT v NDK a Archive exportu
## Nastavení postprocesoru po klasickém exportu
export.bagit_post_processor.processor=bagit

## tyto parametry jsou neměnné a vycházejí z dokumentace BAGIt (https://du.cesnet.cz/cs/navody/ltp/start)
processor.bagit.exec=python
processor.bagit.arg=-m
processor.bagit.arg=bagit
processor.bagit.arg=--md5

## cesta, kam se mají po úspěšném exportu přesunout soubory
## pokud je vyplněna, soubory se přesunou, pokud není soubory zůstávají ve svém původním umístění
export.bagit.path=/temp/ProArc/autoImport/bagit/
```

Obrázek 6: Potřebné parametry v konfiguraci

Export BagIt balíčku do LTP systému

Po vytvoření BagIt balíčku nám ProArc umožňuje přímý export do LTP CESNET a pro správnou funkčnost je nutná konfigurace python scriptu.

Návod a technické specifikace

Je nutné si stáhnout a nainstalovat python script (<https://github.com/CESNET/ltp-api-client-lib>). Poté je potřeba definovat `exec` pro spuštění pythonu a uvést cestu k nainstalované python knihovně. Mezi nutnými parametry je `group token` – stejný bez doby expirace. Kromě něj je při použití klienta nutné při každém exportu doplnit token pro autentizaci (token vydrží jen 24 hodin, a proto není nutné ho ukládat do konfigurace).

```
## (4.0.3) pro upload Bagit balíčku do Ltp Cesnetu
## Nastavení postprocesoru zabalení do Bagit balíčku
export.ltp_cesnet_post_processor.processor=ltpCesnet

## tyto parametry jsou nemenne a vychazeji z dokumentace Ltp Cesnet (https://github.com/CESNET/ltp-api-client-lib)
processor.ltpCesnet.exec=python3
processor.ltpCesnet.arg=/opt/ltpCesnet/ltp-api-client-lib/ltp_api_client.py

export.ltpCesnet.groupToken=[token skupiny]
export.ltpCesnet.scriptPath=/opt/ltpCesnet/upload/cesnetLtpUpload.sh
```

Obrázek 7: Potřebné parametry v ProArc pro Export BagIt do LTP CESNET

Zápis informací o uložení v LTP do katalogizačního záznamu

Nově je v systému ProArc možnost zaslání všech potřebných informací o uložení balíčku v systému LTP CESNET do knihovního systému Verbis, resp. do katalogizačního záznamu daného titulu/zvukového dokumentu.

V tuto chvíli tato funkce funguje pouze s knihovním systémem Verbis, který byl pro tyto účely upraven.

Tedy, podle kontrolního čísla záznamu, které se nachází v poli 001, se do samotného katalogizačního záznamu propíše do pole 856 tyto informace: informace o digitalizaci, identifikátor proarc a UUID daného balíčku.

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

Návod a technické specifikace



```
# (2.1) povinné URL OAI serveru
catalog.oai.url=http://arXiv.org/oai2
# (2.1) povinné označení katalogu
catalog.oai.type=OAICatalog
# (2.1) název katalogu zobrazený uživateli
catalog.oai.name=OAI Katalog
# (2.1) povinný OAI metadata prefix pro data ve formátu http://www.loc.gov/standards/marcxml/schema/MARC21slim.xsd
catalog.oai.metadataPrefix=marc21
# (2.1) nepovinný prefix identifikátoru, který se přidá před hodnotu vyplněnou uživatelem
catalog.oai.identifierPrefix=oai:arXiv.org:
# (2.1) povolená pole: id
catalog.oai.fields=id
# (2.1) název pole zobrazený uživateli
catalog.oai.field.id.title=Identifikátor
#(4.1.0) Potřebné pro autentizaci při editaci
catalog.oai.authorizationUrl=http://arXiv.org/oai2/token
#(4.1.0) Potřebné pro autentizaci při editaci záznamu v katalogu, jinak nechat prázdné
catalog.oai.username=OaiUser
#(4.1.0) Potřebné pro autentizaci při editaci záznamu v katalogu, jinak nechat prázdné
catalog.oai.password=OaiUserPassword
#(4.1.0) Potřebné pro editaci záznamu v katalogu, jinak nechat prázdné
catalog.oai.updateRecordUrl=http://arXiv.org/oai2/updateBulk
# (4.1.0) Potřebné pouze pro editaci záznamu v katalogu, pole kam se píše uuid a info o zdigitalizování
catalog.oai.recordField=856
#(4.1.0) Potřebné pouze pro editaci záznamu v katalogu, podpole kam se píše identifikátor aplikace
catalog.oai.recordSubfieldAppIdentifier=x
#(4.1.0) Potřebné pouze pro editaci záznamu v katalogu, podpole kam se píše identifikátor objektu
catalog.oai.recordSubfieldObjectIdentifier=y
#(4.1.0) Potřebné pouze pro editaci záznamu v katalogu, podpole kam se píše info o zdigitalizování
catalog.oai.recordSubfieldDigitalized=d
```

Obrázek 8: Potřebné parametry v ProArc pro nahrání informace o nahrání do Verbis