Úvod Uživatelské účty Vstupní parametry Kód registrátora Mód registrace Základní OAI URL OAI metadatový prefix OAI set Transformační šablony Úpravy a testování šablon Na co si dát pozor Report.txt a logy Spuštění přes webové rozhraní Spouštění z příkazové řádky Configuration.properties O tomto dokumentu

Úvod

OAI Adapter je nástroj, který slouží k usnadnění hromadné a pravidelné registrace digitálních dokumentů (přiřazení URN:NBN, dále jen *registrace DD*) a importu digitálních instancí (přiřazení URL k URN:NBN, *dále jen import DI*) v systému CZIDLO.

Nástroj pracuje tak, že sklízí metadatové záznamy pomocí protokolu OAI-PMH 2.0 z OAI Repozitářů, typicky digitálních knihoven. Záznamy transformuje a následně používá jako vstupy operací registrace DD a import DI přes CZIDLO API.

OAI Adapter je možné spouštět dvojím způsobem. Buď z terminálu/příkazového řádku, nebo jak proces na serveru spravovaný přes webové rozhraní.

Více informací naleznete zde:

https://github.com/NLCR/CZIDLO/wiki/OAI-Adapter

Uživatelské účty

Pro spuštění OAI Adapteru je nutné použít platný uživatelský účet. Uživatel také musí mít přiřazena práva ke konkrétnímu registrátorovi. Za tvorbu účtů a přiřazování práv je odpovědný administrátor nebo kurátor. Kontakt na něj je k dispozici na https://resolver.nkp.cz/web/#tab0.

Vstupní parametry

Vstupní parametry pro běh OAI Adapteru jsou definovány v souboru .properties. V případě spuštění z webového rozhraní je potřeba vybrané parametry vyplnit do formuláře dialogu. Parametry jsou následující:

Kód registrátora

Kód registrátora, jemuž digitální dokumenty patří. Kód je součástí samotného URN:NBN. Při práci s více registrátory je potřeba spouštět OAI Adapter pro každého z registrátorů zvlášť.

OAI základní URL

Základní URL OAI Poskytovatele¹, např. https://kramerius.mzk.cz/oai/. Pro informace o Poskytovateli stačí za základní URL připojit řetězec "?verb=Identify", tedy https://kramerius.mzk.cz/oai/?verb=Identify, a otevřít ve webovém prohlížeči. Výsledkem bude být popis repozitáře v xml.

OAI metadatový prefix

Identifikátor formátu metadatové reprezentace sklízených záznamů. Např. *oai_dc* (Dublin Core²), *mods* (MODS), *oai_marc* (MARC21), *drkramerius4* (interní formát Krameria), *czidlo* (interní formát CZIDLO), ese (formát Europeany). Seznam podporovaných formátů lze získat následovně: https://kramerius.mzk.cz/oai/?verb=ListMetadataFormats
Transformační šablony jsou vyrobeny na míru metadatového formátu, resp. jeho konkrétnímu dialektu používanému v repozitáři.

OAI set

Set (množina), tedy omezený soubor záznamů. Např. Kramerius používá množiny *monograph*, *periodical, periodicalvolume*, apod. Množina je nepovinný parametr, prakticky ale většinou vyplněn bývá. Množiny jsou totiž častým nástrojem, jak oddělit jednotlivé typy dokumentů (např. monografie a periodika v Krameriu). Ze samotných metadat už pak nemusí být zřejmé, o jaký typ dokumentu jde. Proto dává smysl množiny vyplňovat a používat specifické transformační šablony pro každou množinu. A OAI Adapter pak spouštět pro každou množinu zvlášť. Seznam množin se dá zjistit takto: https://kramerius.mzk.cz/oai/?verb=ListSets.

Transformační šablony

Transformační šablony XSLT definují konverzi metadatového záznamu z OAI na dvě xml, která jsou vstupem operací *registrace DD* a *import DI*. Obě šablony na vstupu očekávají záznam v konkrétním metadatovém formátu a na výstupu vrací záznam ve formátu CZIDLO. Některá data mohou být přímo součástí šablony³. Typicky kořenový element (monograph/periodical/periodicalIssue, ...) pro vstup *registrace DD*, nebo id digitální knihovny pro *import DI*.

Příklady šablon:

monografie ve formátu Kramerius:
 https://github.com/NLCR/CZIDLO/tree/master/oaiAdapter/src/main/resources/xsltSample
 s/kramerius/drkramerius4-mzk

¹ OAI Poskytovatel (Provider) - API pro sklízení metadat z OAI Repozitáře.

² Dublin core je povinný, každý OAI Poskytovatel musí vracet záznamy v tomto formátu.

³ Více v kapitole *Na co si dát pozor*.

- mapy ve formátu MARC21: https://github.com/NLCR/CZIDLO/tree/master/oaiAdapter/src/main/resources/xsltSample s/oldMaps
- články časopisu Duha ve formátu Dublin Core: https://github.com/NLCR/CZIDLO/tree/master/oaiAdapter/src/main/resources/xsltSample-s/duha
- všechny příklady: https://github.com/NLCR/CZIDLO/tree/master/oaiAdapter/src/main/resources/xsltSample
 s

Vždy jsou zde soubory **NAZEV_OAI-FORMAT.xmI** (původní záznam z OAI),

NAZEV_OAI-FORMAT_to_czidloDD.xsI a NAZEV_OAI-FORMAT_to_czidloDI.xsI (samotné šablony) a NAZEV_czidloDd.xmI a NAZEV_czidloDi.xmI (pomocí šablon transformované záznamy, tedy vstupy operací registrace DD a import DI).

Úpravy a testování šablon

Pro tvorbu a úpravy transformačních XSLT šablon budete potřebovat editor XML, který podporuje XLST 1.0 a ideálně i validaci s pomocí XSD schémat. Porovnání možností některých editorů: https://en.wikipedia.org/wiki/Comparison_of_XML_editors.

XSD je potřeba z toho důvodu, že vstupy operací se na serveru validují podle XSD dostupných zde: https://github.com/NLCR/CZIDLO/tree/master/api/src/main/resources/v5. Proto je vhodné i během úprav šablon průběžně validovat výstupy.

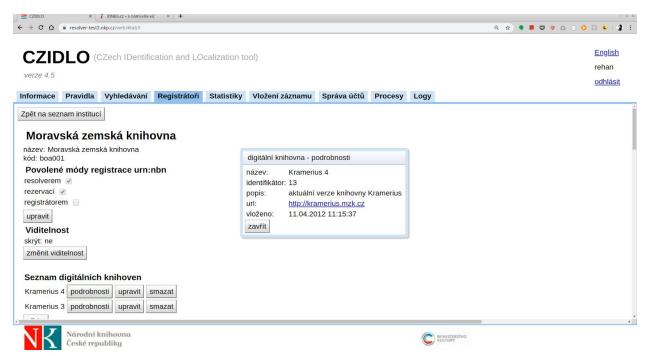
Pro základní transformace a validace mohou stačit online nástroje, třeba:

http://www.freeformatter.com/xsl-transformer.html http://www.freeformatter.com/xml-validator-xsd.html

Na co si dát pozor

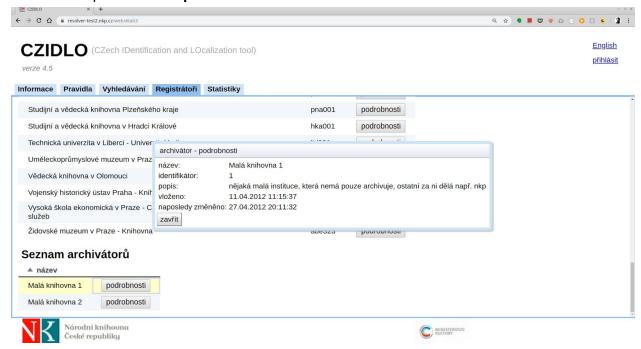
Při tvorbě úpravách šablon je potřeba si dávat pozor zejména na následující elementy ve výstupních xml:

• digitalLibraryld (import DI) - Číselný identifikátor digitální knihovny, v níž se digitální instance nacházejí. CZIDLO eviduje digitální knihovny každého registrátora. Těch může být více (Kramerius 3, Kramerius 4, vlastní webová prezentace, ...). Typicky budou všechny záznamy zpracované v jednom běhu OAI Adapteru ve stejné digitální knihovně. Proto bývá digitalLibraryld v šabloně na pevno. Identifikátor digitální knihovny zjistíme přes webové rozhraní v záložce Registrátoři po rozkliknutí Podrobnosti. V sekci Seznam digitálních knihoven můžeme digitální knihovnu vytvořit tlačítkem přidat. Kliknutím na tlačítko podrobnosti u digitální knihovny se otevře dialog, který mimojiné obsahuje identifikátor digitální knihovny.



Identifikátor digitální knihovny lze také zjisit ze záznamu regsitrátora přes API: https://resolver-test2.nkp.cz/api/v5/registrars/tst02.

 archiverld (registrace DD) - Nepovinný identifikátor archivátora, pokud sám registrátor není zároveň archivátorem. Lze zjistit v záložce Registrátoři dole, v sekci Seznam archivátorů po rozkliknutí podrobnosti u archivátora.



url (import DI) - URL konkrétního digitálního dokumentu v digitální knihovně (tj.
instance). To často bývá odvoditelné z identifikátoru, který je obsažen v metadatovém
záznamu. Např. pro Kramerius se tvoří takto:

- http://kramerius4.nkp.cz/search/handle/ + pid (perzistentí identifikátor), např.: http://kramerius4.nkp.cz/search/handle/uuid:22602f10-ee0c-11dd-9dd9-000d606f5dc6
- titleInfo (registrace DD) Tento element obsahuje různé názvové údaje, které se liší podle typu dokumentu. Třeba monografie má povinný element title a nepovinný subTitle. Ale číslo periodika (periodicalIssue) musí mít hned tři povinné názvové údaje periodicalTitle (název celého periodika), volumeTitle (název ročníku), issueTitle (název čísla). Tento fakt musí OAI Repozitář zohlednit a poskytovat dostatečně popisná metadata.

Registrace DD

Ve verzi 4.4 došlo k zjednodušení podprocesu, který je zodpovědný za registraci/potvrzení URN:NBN. Matoucí vstupní parametr *registrationMode* byl nahrazen parametry *registerDigitalDocumentsWithUrnNbn* a *registerDigitalDocumentsWithoutUrnNbn*. Pokus o registraci DD tedy proběhne tehdy, pokud je splněn jeden z bodů:

- 1. Vstupní dokument pro registraci DD (získaný konverzí OAI záznamu) **obsahuje** URN:NBN a zároveň *registerDigitalDocumentsWithUrnNbn*=true
- 2. Vstupní dokument pro registraci DD (získaný konverzí OAI záznamu) **neobsahuje** URN:NBN a zároveň **registerDigitalDocumentsWithoutUrnNbn=true**

Módy registrace a stavy URN:NBN stále hrají roli, ale uživatel na vstupu definuje jen to, zda se má OAI Adapter pokoušet o registraci, pokud nebo není URN:NBN nalezeno. Vyhodnocení stavu URN:NBN, povoleného módu registrace, porovnání kódu registrátora se děje více skrytě, uživatel je o případném problému informován v reportu.

Například pokud je URN:NBN v záznamu nalezeno a registerDigitalDocumentsWithUrnNbn=true, ale toto URN:NBN je již registrováno, registrace neproběhne.

Nebo pokud je URN:NBN rezervováno, ale registrátor nemá povolen mód registrace BY_RESERVATION, registrace nebude úspěšná.

Může se také stát, že kód registrátora na vstupu (property *oaiAdapter.czidloApi.registrarCode*, resp. formulářové pole *registrátor*) nesouhlasí s kódem registrátora uvnitř URN:NBN. Tehdy registrace DD neproběhne. Pokud ale už URN:NBN registrováno je, ač jiným registrátorem, pokračuje se dál importem DI. Taková situace typicky nastává, když jedna instituce dokument registruje pod svým kódem a poté s již přiděleným URN:NBN distribuuje nejen do své digitální knihovny, ale i do digitální knihovny další instituce. A tato další instituce je také registrátorem v CZIDLO a používá OAI Adapter k doplnění/aktualizování svých DI.

Pokud je splněn druhý bod (záznam neobsahuje URN:NBN a registerDigitalDocumentsWithoutUrnNbn=true) a registrace DD proběhne úspěšně, bude dokumentu přiděleno volné URN:NBN. Potom je na registrátorovi, aby vlastními silami doplnil přidělená URN:NBN do záznamů své digitální knihovny. Může k tomu použít API CZIDLO a registrar-scope identifikátor typu OAI Adapter.

Nahrazování a slučování digitálních instancí

Parametr *oaiAdapter.dilmport.mergeDigitalInstances* umožňuje slučování digitálních instancí. Pokud nastane situace, že se v databázi CZIDLO nachází DI se stejnou kombinací registrátora, digitální knihovny a digitálního dokumentu, jako právě připravená nová DI, a existující DI se od nové liší, je potřeba DI aktualizovat. Pokud doposud DI neexistuje, vloží se nová DI.

Aktualizace DI do verze 4.3 probíhala tak, že byla existující DI deaktivována a nová DI vložena. Od verze 4.4 ale lze záznam staré a nové DI sloučit a vložit tento sloučený záznam. Sloučení znamená doplnění chybějících nepovinných polí (formát, dostupnost, omezení dostupnosti) z existující DI do nové DI.

Existující a nová DI se od sebe liší, pokud se liší URL, nebo některé z nepovinných polí, tedy i pokud je některé nepovinné pole v jedné z DI prázdné v druhé ne. Parametry oaiAdapter.diImport.ignoreDifferenceInAccessibility a

oaiAdapter.dilmport.ignoreDifferenceInFormat lze nastavit, že DI nebudou považovány za odlišné, pokud se liší hodnoty některého z těchto parametrů. Tedy např. pokud se bude lišit formát a oaiAdapter.dilmport.ignoreDifferenceInFormat=true, nedojde k aktualizaci DI.

Access restriction/Omezení přístupu

Ve verzi 4.4 přibylo nové povinné pole pro DI - access restriction. To může nabývat jedné ze tří hodnot: UNKNOWN, UNLIMITED_ACCESS a LIMITED_ACCESS, přičemž pokud není řečeno jinak (a pro všechny stávající záznamy DI zpětně), je hodnota nastavena na UNKNOWN. Hodnota access restriction se používá pro rezolvování. Pokud je během rezolvování nalezeno více vhodných DI, je upřednostněna ta DI, jejíž hodnota access restriction je UNLIMITED_ACCES, případně UKNOWN. Je to proto, aby byl čtenář raději přesměrován do té digitální knihovny, kde je pravděpodobnější, že bude digitální dokument plně přístupný. Přičemž zdroj přístupu je stále silnější, než accessibility - tedy např. z katalogu NKP bude přesměrováno do digitální knihovny NKP i tehdy, když bude existovat DI v digitální knihovně MZK s volnějším omezením přístupu.

V záznamu DI stále existuje podobné pole accessibility (dostupnost), které ale obsahuje jen textovou informaci o dostupnosti DI a nijak se nevyužívá pro rezolvování.

Report.txt a logy

Výstupem OAI Adapteru je soubor report.txt. Ten obsahuje popis průběhu pro každý jednotlivý záznam sklizený přes OAI, dále transformovaný a použitý pro operace CZIDLO API. Na konci reportu je souhrn zpracovaných záznamů a počtů záznamů v jednotlivých výsledných stavech. Příklad reportu:

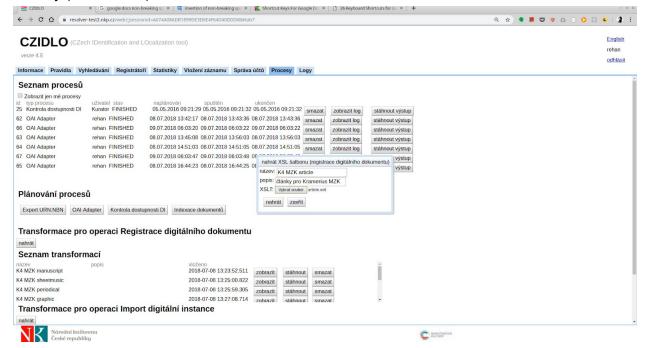
https://github.com/NLCR/CZIDLO/blob/master/oaiAdapter/src/main/resources/otherExamples/report.txt.

Ostatní události OAI Adapter ukládá do logu. Jde zejména o varování a chyby, které způsobí předčasné ukončení celého průběhu OAI Adapteru. Při spuštění procesu přes webové rozhraní je log k dispozici ke stažení, stejně jako report. Při spuštění procesu z terminálu/příkazové řádky je log vypisován na standardní výstup.

Spuštění přes webové rozhraní

Nejjednodušší způsob použití OAI Adapteru je jeho spuštění jako procesu na serveru přes webové rozhraní. To je dostupné na adrese https://resolver.nkp.cz/web/#tab7. Pro práci s procesy je nutné se přihlásit (v pravém horním rohu).

Před samotným spuštěním procesu je nutné nahrát šablony pro transformaci výstupu OAI do vstupu pro operace registrace DD a import DI. To provedete kliknutím na *nahrát novou transformaci* v sekcích Transformace pro operaci Registrace digitálního dokumentu a Transformace pro operaci Import digitální instance v záložce Procesy. V následujícím dialogu je potřeba vyplnit název a popis transformace a nahrát samotný soubor transformační šablony (soubor .xslt).



Nahrané transformační šablony jsou uloženy na serveru a spojeny s přihlášeným uživatelem. Pro další spuštění procesů je již nebude nutné znovu nahrávat.

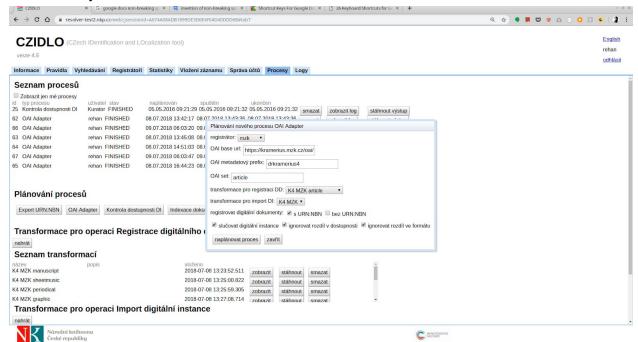
Samotné spuštění procesu se provádí tlačítkem **OAI Adapter** v sekci **Plánování procesů** v záložce **Procesy**.

V dialogu pro konfiguraci procesu je třeba doplnit následující⁴:

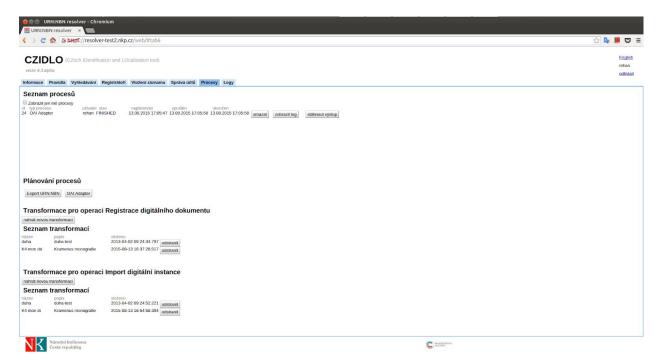
- Vybrat registrátora (zvolit kód registrátora). Pokud v seznamu není žádný registrátor, nebo chybí registrátor, s kterým jste měli v úmyslu pracovat, znamená to, že přihlášenému uživateli není přiřazeno právo spravovat registrátora. S požadavkem na přidělení práva se obraťte na kurátora/administrátora.
- Vyplnit OAI základní URL.
- Vyplnit OAI metadatový prefix.
- Vyplnit OAI set. Toto pole není povinné.
- Zvolit transformační šablonu pro vstup operace registrace DD.

⁴ Více v kapitole Vstupní parametry.

- Zvolit transformační šablonu pro vstup operace import DI.
- Zvolit, zda se má OAI Adapter pokoušet o registraci DD, pokud je/není URN:NBN přítomno v záznamu pro registraci DD získaném konverzí OAI záznamu.
- Povolit/zakázat slučování digitálních instancí
- Povolit/zakázat ignorování rozdílu v dostupnosti pro vyhodnocení, zda se nová DI liší od stávající
- Povolit/zakázat ignorování rozdílu ve formátu pro vyhodnocení, zda se nová DI liší od stávající



Tlačítkem **Naplánovat proces** bude proces uložen a naplánován (stav SCHEDULED). Po nějaké době, která záleží na počtu a prioritě dalších naplánovaných procesů, bude proces spuštěn (RUNNING). Proces může být uživatelem předčasně ukončen (CANCELED), ukončen systémem (KILLED), skončit chybou (FAILED), nebo být úspěšně dokončen (FINISHED). Stavy procesů můžete průběžně kontrolovat v sekci **Seznam procesů** v záložce **Procesy**. Zobrazené procesy lze filtrovat jen na ty, které vlastní aktuálně přihlášený uživatel, a to zaškrtnutím **Zobrazit jen mé procesy**.



Tlačítkem **zobrazit log** se otevře nové okno prohlížeče obsahující logy o průběhu procesu, toto tlačítko je přítomno u probíhajících a dokončených procesů. Probíhající procesy je možné zrušit tlačítkem **zrušit**.

U úspěšně dokončených procesů (FINISHED) je dále tlačítko **stáhnout výstup**, které umožní stáhnout soubor, jenž je výstupem procesu, do vašeho počítače. U procesu OAI Adapter je výstupem soubor *report.txt*.

Záznamy všech dokončených procesů (FINISHED, FAILED, CANCELED, KILLED) lze smazat tlačítkem **smazat**.

Spuštění z příkazové řádky

Instalační balík CZIDLO⁵ obsahuje soubor *oaiAdapter-VERZE.jar*. Ten je určen pro lokální běh OAI Adapteru. Pro jeho spuštění je nutné mít nainstalováno prostředí Java Runtime Environment (JRE), alespoň ve verzi 6 (1.6). To obsahuje program java, který spouští kód v balíku jar. Aktuální verze JRE 8 je k dispozici zde:

http://www.oracle.com/technetwork/java/javase/downloads/jre8-downloads-2133155.html

Spuštění OAI Adapteru vypadá následovně (Linux):

java -jar oaiAdapter-4.5.jar /cesta/k/souboru/configuration.properties

pro Windows:

java -jar oaiAdapter-4.5.jar C:\cesta\k\souboru\configuration.properties

Na výstup jsou vypisovány některé chybové a informační hlášky, proto může být vhodné výstup přesměrovat do souboru:

⁵ Instalační balíky jsou dostupné zde: https://github.com/NLCR/CZIDLO/releases.

java -jar oaiAdapter-4.5.jar /cesta/k/souboru/configuration.properties
>/path/to/logs.txt

Configuration.properties Soubor configuration.properties obsahuje samotné parametry běhu⁶. Jeho obsah vypadá nějak takto: # OAI PROVIDER # oaiAdapter.oai.baseUrl=http://duha.mzk.cz/oai oaiAdapter.oai.metadataPrefix=oai_dc #@optional oaiAdapter.oai.setSpec=monographs # CZIDLO API # oaiAdapter.czidloApi.baseUrl=resolver.nkp.cz/api oaiAdapter.czidloApi.login=czidlo user 123 oaiAdapter.czidloApi.password=***** oaiAdapter.czidloApi.registrarCode=aba001 #@optional(default=false) oaiAdapter.czidloApi.ignoreInvalidCertificate=false ************ # XSL TRANSFORMATIONS # oaiAdapter.digDocRegistrationXsl=/path/to/oaiDc_to_czidloDd.xsl oaiAdapter.digInstImportXsl=/path/to/monograph_oaiDc_to_czidloDi.xsl # XSD FOR VALIDATING CZIDLO API INPUTS # oaiAdapter.digDocRegistrationXsdUrl=http://resolver.nkp.cz/api/v5/digDocRegistration.x sd oaiAdapter.digInstImportXsdUrl=http://resolver.nkp.cz/api/v5/digInstImport.xsd # DD REGISTRATION #

⁶ Více v kapitole Vstupní parametry.

#whether OAI Adapter should try and register DD if URN:NBN is found in input data and its state is discovered to be FREE or RESERVED

#and registrar-code in URN:NBN matches expected registrar-code (typically from oaiAdapter.czidloApi.registrarCode)

#@mandatory

oaiAdapter.ddRegistration.registerDigitalDocumentsWithUrnNbn=true

#whether OAI Adapter should try and register DD if URN:NBN is not found in input data AND digital document has not been registered

#Situation when URN:NBN is not found but digital document has already been registered still can happen due to digital library not updating its OAI records properly after the registration. #Though OAI Adapter can detected this through registrar-scope-id mapping (with type OAI_Adapter).

#@mandatory

oaiAdapter.ddRegistration.registerDigitalDocumentsWithoutUrnNbn=false

DI IMPORT

#whether new DI (OAI record transformed with XSLT) should be merged with current DI before inserting

#merging here means that selected empty fields in DI (accessibility, format) will be filled from old DI if available

#@optional(default=true)

oaiAdapter.dilmport.mergeDigitalInstances=false

#whether difference in accessessibility should be ignored when deciding if digital instance has changed

#@optional(default=false)

oaiAdapter.dilmport.ignoreDifferenceInAccessibility=false

#whether difference in format should be ignored when deciding if digital instance has changed #@optional(default=false)

oaiAdapter.dilmport.ignoreDifferenceInFormat=false

REPORT FILE

oaiAdapter.reportFile=/path/to/report.txt

Kromě základních parametrů popsaných v sekci *Vstupní parametry*, jsou zde ještě navíc parametry *oaiAdapter.czidloApi.**, které slouží k napojení OAI Adapteru na CZIDLO API:

- oaiAdapter.czidloApi.baseUrl Základní URL CZIDLO API bez uvedení protokolu. Protokolem je vždy https, nicméně jej neuvádějte. Na konci by nemělo být lomítko.
- **oaiAdapter.czidloApi.login** Uživatelské jméno (login) uživatele, který má přiřazena práva pro daného registrátora.
- oaiAdapter.czidloApi.password Heslo uživatele.
- oaiAdapter.czidloApi.registrarCode Kód registrátora.
- oaiAdapter.czidloApi.ignoreInvalidCertificate Zda se má pokračovat i tehdy, pokud je neplatný certifikát pro CZIDLO API. Určeno jen pro testovací účely.

O tomto dokumentu

Verze: 2.1

Vytvořeno: 12. 7. 2015 (Martin Řehánek)

Naposledy upraveno: 10. 9. 2018 (Martin Řehánek)