Automatizované zpracování autoritních záznamů

Stránka se upravuje

Princip řešení

Cílem řešení je automatizovat práci s autoritními záznamy z NK ČR a zajistit jejich dávkové vkládání a aktualizaci. Vzhledem k plánované četnosti provádění těchto úkonů jednou měsíčně (popř. méně často) se počítá s využíváním exportu všech autoritních záznamů z NK ČR. Data z NK ČR jsou následně čištěna (export z NK ČR obsahuje jak platné úplné záznamy, tak některé neúplné záznamy; kromě standardních polí označených číslicemi obsahuje i pole označená písmeny). Takto předzpracovaná data jsou prostřednictvím identifikátorů záznamů srovnávána s exportem autoritních záznamů z Evergreenu (tedy záznamů, které byly do Evergreenu nahrány již dříve). Z toho vznikne aktualizační balík pro autority v Evergreenu. Ze souboru všech autoritních záznamů z NK ČR jsou při této operaci odstraněny právě tyto záznamy využité k aktualizaci. Zatím nepoužité záznamy jsou potom srovnávány s exportem bibliografických dat z Evergreenu – hledají se záznamy pro automatické nahrání do systému podle identifikátorů autoritních záznamů v bibliografických záznamech stažených prostřednictvím protokolu Z39.50 (především z bází NKC a SKC). Zbytek autoritních záznamů z NK ČR (ten již neobsahuje autoritní záznamy použité v předešlém kroku) potom prochází fulltextovým prohledáváním (prohledáváním podle řetězců), které vytvoří další balík návrhů pro import do Evergreenu. Všechny autoritní záznamy, které jsou získány v jednotlivých krocích, jsou následně vloženy do Evergreenu, a to buď jako aktualizace dříve vložených záznamů, nebo jako nové záznamy pomocí standardních prostředků Evergreenu pro import autoritních záznamů. Autoritní záznamy v Evergreenu potom mohou být ručně nebo automaticky navázány na bibliografické záznamy.

Technická implementace

Řešení je postaveno na skriptech v jazycích Bash a Python. Očekává se, že skripty budou spouštěny na serveru, na němž je provozován knihovní software Evergreen. Vlastní implementace byla testována na Evergreenu řady 2.12, ale není znám problém, který by bránil jejímu použití například ve verzi 2.10 nebo 3.0.

Prerekvizity

Postup předpokládá, že jsou nainstalovány operační systém Debian 7 a Evergreen 2.12. Pro běh skriptů se očekává spouštění pod uživatelem opensrf. Při vykonávání skriptů musí být Evergreen spuštěn. Kromě balíků instalovaných při instalaci Evergreenu je navíc potřeba pouze balík **xml-twig-tools**. Ten je třeba nainstalovat příkazem *„apt-get update && apt-get install xml-twig-tools“.*

Základní nastavení skriptů počítá s umístěním programu na cestě *„/home/opensrf/Autority/bin“* a s umístěním dat na cestě *„/home/opensrf/Autority/work“*. Tyto cesty musejí být vytvořeny ručně a balík

s programovým vybavením se musí rozbalit do adresáře *„/home/opensrf/Autority/bin“*. Cesty mohou být změněny v konfiguračním souboru **0_env.sh**.

Pro běh programu je nutná existence souboru s přihlašovacími údaji k databázi evergreen *„~/.pgpass“* a v základním nastavení počítá s databázovým uživatelem evergreen a databází evergreen puštěnou na databázi Postgres na lokálním serveru. Změny v umístění a jménu databáze či uživatele databáze je možné nastavit v souboru **0_env.sh**.

Pro získání exportu autoritních dat z NK ČR je třeba zažádat si o přístupové údaje na FTP server NK ČR.

Postup a kontaktní osoby jsou k dispozici na stránce http://autority.nkp.cz/kooperace/soubory-k-aktualizaci-lokalnich-bazi. Tyto údaje se potom musí vložit do konfiguračního souboru **0_env.sh**.

Pro vkládání do Evergreenu je třeba uživatelské jméno a heslo administrátora. Toto se rovněž nastavuje v souboru **0_env.sh**. Nastavení polí, která jsou srovnávána při hledání podle identifikátorů, se provádí v souboru **pole.txt**. Nastavení polí, která jsou srovnávána při fulltextovém vyhledávání, se provádí v souboru **bibaut.txt**.

Postup aktualizace autoritních záznamů

Aktualizace autoritních záznamů je rozdělena do devíti kroků, přičemž krok 0 je nastavení a export proměnných a krok 8 je volitelné automatické navázání autoritních záznamů na bibliografické záznamy.

0_env.sh

V souboru se nastavují proměnné, uživatelské údaje a cesty k souborům.

Spouští se pomocí příkazu *„source /home/opensrf/Autority/bin/0_env.sh“*.

1_stahni.sh

Skript stáhne komprimovaný balík autoritních záznamů z FTP NK ČR a autoritní záznamy rozbalí.

Spouští se pomocí příkazu *„/home/opensrf/Autority/bin/1_stahni.sh“*.

2_cisti.sh

Skript nejprve z autoritních záznamů z NK ČR odstraní pole označená písmeny. Poté odstraní záznamy bez pole 040. (Záznamy s poli označenými písmeny a záznamy bez pole 040 nejsou pokládány za validní a nelze je do Evergreenu automaticky naimportovat.)

Spouští se pomocí příkazu *„/home/opensrf/Autority/bin/2_cisti.sh“*.

3_soucasne_bib_a_aut.sh

Skript z Evergreenu vyexportuje aktuální nesmazané autoritní a bibliografické záznamy. Ty jsou v dalších krocích použity pro vyhledávání aktualizací a nových autorit.

Spouští se pomocí příkazu *„/home/opensrf/Autority/bin/3_soucasne_bib_a_aut.sh“*.

4_hledani_autorit.sh

Skript volá tři skripty v jazyce Python, a to **a.py**, **b.py** a **c.py**. Skript **a.py** hledá aktualizace do Evergreenu již vložených autoritních záznamů (skript neřeší aktuálnost záznamu podle data; záznam je vždy přepsán obsahem z exportu autoritních záznamů z NK ČR).

Skript **b.py** podle mapování v souboru pole.txt hledá autoritní záznamy odpovídající identifikátorům autoritních záznamů nalezených v exportu bibliografických dat z Evergreenu.

Skript **c.py** autority hledá fulltextově na základě parametrů nastavených v programu (ignorované znaky a délka souhlasného hashe). Pro nastavení vztahů autoritních a biliografických záznamů se používá soubor **bibaut.txt**.

Výstupem skriptu jsou:

- soubor s aktualizací autoritních záznamů již v Evergreenu vložených (soubor je ve formátu TSV),
- soubor s autoritními záznamy nalezenými podle identifikátorů v bibliografických záznamech (soubor je ve formátu MARCXML),
- soubor s autoritními záznamy nalezenými pomocí fulltextového vyhledávání (soubor je ve formátu MARCXML).

Skript se pouští pomocí příkazu *„/home/opensrf/Autority/bin/4_hledani_autorit.sh“*.

5_aktualizace_autorit.sh

Skript pomocí SQL příkazů nahraje aktualizační data do databáze a poté provede aktualizaci autoritních záznamů.

Skript se pouští pomocí příkazu *„/home/opensrf/Autority/bin/5_aktualizace_autorit.sh“*.

6_vlozeni_mapovanych_autorit.sh

Skript z autoritních záznamů ve formátu MARCXML (tyto záznamy byly nalezeny podle svého identifikátoru) vytvoří sekvenci SQL příkazů pro vložení nových autoritních dat a ty vloží do databáze.

Skript se pouští pomocí příkazu *„/home/opensrf/Autority/bin/6_vlozeni_mapovanych_autorit.sh“*.

7_vlozeni_fulltextove_vyhledanych_autorit.sh

Skript z fulltextově vybraných autoritních záznamů ve formátu MARCXML vytvoří sekvenci SQL příkazů pro vložení nových autoritních záznamů a ty vloží do databáze.

Skript se pouští pomocí příkazu *„/home/opensrf/Autority/bin/7_vlozeni_fulltextove_vyhledanych_autorit.sh“*.

8_spusteni_aktualizace_vazeb.sh

Skript spustí standardní nástroje Evergreenu **authority_authority_linker.pl** a **authority_control_fields.pl**.

Nástroj **authority_authority_linker.pl** propojí autoritní záznamy mezi sebou.

Nástroj **authority_control_fields.pl** automaticky přidá vazby mezi bibliografickými a autoritními záznamy.

Nastavení propojování (vazeb) se provádí úpravou kódu těchto standardních nástrojů Evergreenu. V českém prostředí je třeba přidat podpole 7, které obsahuje identifikátor autoritního záznamu.

Skripty ke stažení

autority.zip

Poznámky

Výběr autoritních záznamů a jejich propojování jsou operace, které jsou poměrně náročné na výpočetní výkon. Je proto vhodné je provádět mimo běžnou provozní dobu knihovny.

*Autorem skriptů **a.py** a **b.py** je Václav Maixner, autorem skriptu **c.py** je Ing. Miloslav Nič, Ph.D. Ostatní skripty zpracoval Ing. Václav Jansa. Dokumentaci zpracoval Ing. Václav Jansa. Skripty jsou dostupné pod licencí GNU General Public License verze*

Skripty pro automatizované zpracování autoritních záznamů z Národní knihovny ČR (NK ČR) a jejich aktualizace a vkládání do knihovního softwaru Evergreen vznikly jako součást řešení projektu Zkvalitnění služeb Knihovny Jána Langoše poskytovaných prostřednictvím online katalogu Evergreen. Projekt realizovala Knihovna Jána Langoše, která je součástí Ústavu pro studium totalitních režimů, v roce 2017, a to s využitím dotace z programu VISK 3.