Technologie bibliotek cyfrowych

Od planowania do publikowania – co można zautomatyzować?

Krzysztof Ober
Poznańska Fundacja Bibliotek Naukowych

Cel prezentacji

Celem niniejszej prezentacji jest wspólne zastanowienie się nad możliwościami automatyzacji pracy redaktora biblioteki cyfrowej.

Punktem wyjścia do rozważań będzie system automatycznego wprowadzania publikacji opracowywany dla potrzeb bibliotek poznańskich tworzących zasób Wielkopolskiej Biblioteki Cyfrowej.

Proces tworzenia publikacji

Tworzenie publikacji dla potrzeb biblioteki cyfrowej jest procesem wieloetapowym:

- opracowywanie planów wprowadzania publikacji,
- tworzenie opisów publikacji planowanych,
- przygotowywanie cyfrowych wersji publikacji,
- konwersja plików do formatów używanych w bibliotece, OCR,
- umieszczanie publikacji w bibliotece cyfrowej, publikowanie.

Narzędzia

Pewne etapy pracy redaktora można spróbować zautomatyzować. Z pomocą przychodzą narzędzia programistyczne:

- wbudowane w system dLibra: dedykowane narzędzie do dodawania plików do publikacji planowanej (dostępne od wersji 4.0.10),

- zewnętrzne – np. Document Express Enterprise i wbudowany w niego mechanizm wsadowego przetwarzania plików (*Workflow Manager*).

Etap I.
Opracowywanie planów wprowadzania publikacji

Różne aspekty procesu planowania:

- profil biblioteki cyfrowej,
- badanie zapotrzebowania użytkowników (np. za pomocą ankiet),
- analiza statystyk wykorzystania już opublikowanych materiałów,
- indywidualne prośby użytkowników o umieszczenie określonych pozycji w bibliotece cyfrowej,
- prawa autorskie.

Stosowanie różnego rodzaju narzędzi programistycznych ułatwia pracę na tym etapie, ale nie zastąpi w całości pracy, którą musi wykonać człowiek.

ETAP II.
Tworzenie opisów publikacji planowanych

Opracowywanie opisów publikacji za pomocą metadanych w formacie Dublin Core jest czynnością czasochłonną – wymagającą posiadania odpowiedniej wiedzy i doświadczenia.

Mechanizmy importu opisów zaimplementowane w systemie dLibra:

- import metadanych z formatu MARC,
- import metadanych z formatu XML,
- import metadanych z formatu BibTeX,
- pobieranie metadanych poprzez rozszerzenie Z39.50,
- wymiana metadanych za pomocą formatu RDF.

Identyfikator publikacji planowanej jest elementem niezbędnym do prawidłowego działania systemu automatycznego wprowadzania publikacji.

ETAP III.
Przygotowywanie cyfrowych wersji publikacji

Digitalizacja (najczęściej skanowanie) materiałów przeznaczonych do umieszczenia w bibliotece cyfrowej jest etapem, który można częściowo zautomatyzować.

Pozwalają na to nowoczesne rozwiązania sprzętowe i programowe stosowane w skanerach:

- profile skanowania,
- automatyzacja zapisu stron,
- przyciski szybkiego dostępu.

Zeskanowane materiały powinny zostać umieszczone w katalogach o nazwach odpowiadających indentyfikatorom publikacji planowanych w systemie dLibra.

ETAP IV.
Konwersja plików, OCR

Redaktor nie musi tracić czasu oczekując na zakończenie konwersji i OCR. Zeskanowane pliki (TIFF) można jednym kliknięciem myszki umieścić na serwerze realizującym zadania systemu automatycznego wprowadzania publikacji.

System automatycznego wprowadzania publikacji wykona – w zależności od katalogu, w którym zostaną umieszczone pliki - następujące zadania:

- skonwertuje pliki do formatu djvu stosując odpowiednie parametry konwersji,
- wykona OCR,
- wygeneruje pliki djvu w trybie indirect (dla potrzeb www),

ETAP V.
Umieszczanie publikacji w bibliotece cyfrowej, publikowanie

- umieści pliki publikacji na serwerze Wielkopolskiej Biblioteki Cyfrowej wykorzystując identyfikator publikacji planowanej,
- jeśli redaktor sobie tego życzy: opublikuje nową publikację.

Warunki:

- pliki publikacji muszą zostać umieszczone w katalogach o nazwach odpowiadających identyfikatorom publikacji planowanych,
- katalogi z plikami publikacji muszą zostać umieszczone w odpowiednich katalogach odpowiadających profilom konwersji zdefiniowanym na serwerze systemu automatycznego wprowadzania publikacji.

Profile konwersji i zadania przetwarzania

- Dla poszczególnych rodzajów publikacji można skonfigurować odpowiednie profile w Document Express'ie.
- Każdy profil Document Expressa odpowiada określonemu katalogowi w systemie plików na serwerze.
- Odwzorowanie odpowiedniej podstruktury katalogów na serwerze znajduje się na dysku lokalnym komputera redaktora.
- Umieszczając pliki publikacji w określonym katalogu redaktor decyduje o parametrach konwersji.
- Przesyłanie plików na serwer odbywa się za pomocą FTP (ncftp).
- Na dysku lokalnym komputera redaktora archiwizowane są oryginalne pliki TIFF, na dysku serwera archiwizowane są pliki djvu w trybie bundle.

Narzędzie do dodawania plików do publikacji planowanej

Narzędzie znajduje się w dystrybucji dLibry począwszy od wersji 4.0.10 i zawiera następujące elementy:

- ***lib*** - katalog zawierający potrzebne biblioteki do uruchomienia narzędzia
- ***config.xml*** - plik zawierający informacje o serwerze do którego narzędzie dodawania plików ma się podłączyć
- ***users.xml*** - informacje o użytkownikach w imieniu których narzędzie będzie dodawało pliki do publikacji planowanej.
- ***run.bat*** - skrypt uruchamiający narzędzie w środowisku systemów z rodziny Windows
- ***run.sh*** - skrypt uruchamiający narzędzie w środowisku systemów z rodziny Linux

Narzędzie do dodawania plików do publikacji planowanej

config.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE properties SYSTEM "http://java.sun.com/dtd/properties.dtd">
<properties>
<comment>
Configuration file for files uploader.
Properties in this file specify: server host and server port.
1. Server host
In order to specify server host place an entry which has 'server' as a key. Value specified for this key
    determines the server host used by the files uploader.
2. Server port
In order to speciy server port place an entry which has 'port' as a key. Value specified for this key determines
    the server port user by the files uploader.
</comment>
<entry key="server">localhost</entry>
<entry key="port">10051</entry>
</properties>
```

Narzędzie do dodawania plików do publikacji planowanej

users.xml

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE properties SYSTEM "http://java.sun.com/dtd/properties.dtd">
<properties>
<comment>
Configuration file for files uploader.
Properties in this file specify information about users on behalf of which this files uploader works
Each entry in this file is composed of key and value. Key indicates user login and value indicates password for
    this user.
</comment>
<entry key="jkowalski">mojehaslo</entry>
</properties>
```

Narzędzie do dodawania plików do publikacji planowanej

Uruchamianie narzędzia:

run <PATH_TO_MAIN_FILE_NAME> false|true

gdzie PATH_TO_MAIN_FILE_NAME:
<PREFIX>/<USER_ID>/out/<PUB_ID>/<MAIN_FILE_NAME>

np.

run C:\pliki\jkowalski\out\22345\directory.djvu true

Narzędzie do dodawania plików do publikacji planowanej

<PREFIX>/<USER_ID>/out/<PUB_ID>/<MAIN_FILE_NAME>

<PREFIX> to pierwsza część ścieżki nieistotna z punktu widzenia narzędzia

<USER_ID> jest katalogiem którego nazwa jest loginem użytkownika w imieniu którego narzędzie ma dodać pliki publikacji

out jest katalogiem zawierającym publikacje danego użytkownika

<PUB_ID> jest katalogiem którego nazwa jest identyfikatorem publikacji planowanej do której mają zostać dodane pliki publikacji; zawiera wszystkie pliki publikacji

<MAIN_FILE_NAME> jest nazwą pliku głównego publikacji

Plany na przyszłość

Automatyczne pobieranie plików publikacji, modyfikacja i podmiana plików publikacji, ponowne umieszczanie zmodyfikowanej publikacji na serwerze WBC (tworzenie nowego wydania) - chcemy ten mechanizm wykorzystać m. in. do wykonania OCR-u w wielu publikacjach, które zostały umieszczone w WBC bez warstwy tekstowej.

Dodawanie warstwy tekstowej do plików djvu w trybie photo.

Konwersja różnych typów dokumentów do formatu djvu z wykorzystaniem drukarki wirtualnej.

Integracja metadanych w publikacjach djvu.

Podsumowanie

Redaktor nie musi zajmować się konwersją i umieszczaniem plików publikacji w bibliotece cyfrowej.

Można zaoszczędzić sporo czasu redaktora automatyzując w całości lub częściowo pewne etapy jego pracy.

Jest to możliwe dzięki programom uruchamianym z linii komend, które można „oskryptować” i uruchamiać wsadowo.

Może inne funkcjonalności dLibry dałoby się zaimplementować w podobny sposób?