Automatyzacja procesu publikowania w bibliotece cyfrowej

Jakub Bajer
Biblioteka Politechniki Poznańskiej

Krzysztof Ober
Poznańska Fundacja Bibliotek Naukowych

Plan prezentacji

1. Cel prezentacji

2. Proces tworzenia publikacji

3. Narzędzia:
   - Document Express Enterprise 5.1
   - ABBYY Recognition Server 2.0
   - narzędzie do dodawania plików do publikacji planowanej (dostępne w systemie dLibra od wersji 4.0.10)

4. Automatyzacja procesu publikowania

Cel prezentacji

Celem prezentacji jest zapoznanie uczestników warsztatów z głównymi funkcjami oprogramowania wykorzystywanego do przygotowania publikacji dla potrzeb biblioteki cyfrowej (m. in. Document Express Enterprise 5.1, Recognition Server 2.0) oraz zaprezentowanie możliwości automatyzacji procesu publikowania na przykładzie rozwiązań wdrożonych w Wielkopolskiej Bibliotece 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 formatu DjVu + OCR,
- umieszczanie publikacji w bibliotece cyfrowej, publikowanie.

Narzędzia

Pewne etapy pracy redaktora można zautomatyzować.

Z pomocą przychodzą narzędzia programistyczne:

- zewnętrzne:
  - Document Express Enterprise 5.1
  - ABBYY Recognition Server 2.0

- wbudowane w system dLibra narzędzie do dodawania treści do opisów publikacji planowanych (dostępne od wersji 4.0.10)

Document Express Enterprise 5.1

DocumentExpress to rodzina aplikacji do tworzenia i manipulowania dokumentami skanowanymi i generowanymi elektronicznie o bardzo dużym stopniu kompresji zapisanych w formacie DjVu.

Enterprise Edition to wersja Document Expressa przeznaczona dla instytucji, które przetwarzają większe ilości dokumentów – możliwość automatycznego przetwarzania wsadowego.

Komponenty DocumentExpress EE

1) Graficzne (tylko Windows):

Configuration Manager – interfejs graficzny do zarządzania profilami (zestawami parametrów przetwarzania) - umożliwia modyfikację istniejących, tworzenie nowych oraz testowanie działania profili;

Workflow Manager – oparty na platformie .NET program pozwalający zorganizować konwersję wsadową. Obsługuje WatchFolders (aktywne foldery), profile konwersji, konwersję PDF-ów (z profilami), OCR, Watermarks (znaki wodne), generowanie plików XML i TXT, seryjną zmianę nazw, obsługę błędów, log operacji;

Komponenty DocumentExpress EE

2) Programy uruchamiane z linii poleceń (wszystkie platformy):

documenttodjvu – konwersja obrazów rastrowych do formatu djvu z obsługą warstw;

photododjvu – konwersja obrazów rastrowych do formatu djvu bez obsługi warstw;

djvutotext – ekstrahowanie warstwy tekstowej do pliku tekstowego;

djvudecode – konwersja plików djvu do obrazów rastrowych;

djvutoxml – ekstrahowanie adnotacji, metadanych oraz warstwy tekstowej do pliku XML;

Komponenty DocumentExpress EE

djvubundle – konwersja pliku DjVu do formatu *bundled* (opcjonalnie tworzenie warstwy OCR oraz osadzanie miniatur);

djvujoin - konwersja pliku DjVu do formatu *indirect* (opcjonalnie tworzenie warstwy OCR oraz osadzanie miniatur);

djvuparsexml – przetwarzanie informacji tekstowych zawartych w pliku XML, import do pliku djvu;

watermarkdjvu – osadzanie znaku wodnego w dokumencie DjVu.

ABBYY Recognition Server 2.0

ABBYY Recognition Server jest zaawansowanym rozwiązaniem serwerowym, które automatyzuje proces rozpoznawania tekstu i konwersji dokumentów PDF. Może on wykonywać wiele zadań równocześnie w obrębie instytucji, natomiast ich monitorowanie odbywa się z jednego centralnego punktu administracji.

Komponenty Recognition Server 2.0

- Server Manager
- Processing Station
- Verification Station
- Remote Administration Console
- COM-based API
- Web Service

Narzędzie do dodawania plików do publikcji planowanej

Narzędzie znajduje się w dystrybucji dLibry począwszy od wersji 4.0.10 i służy do dodawania plików do publikacji planowanej. Pliki jakie mają zostać dodane, użytkownika w imieniu którego pliki będą dodawane oraz publikację do której pliki zostaną dodane wskazywane są w parametrach konfiguracyjnych tego narzędzia. Narzędzie to uruchamiane jest z linii poleceń i jest dedykowane do wykorzystania w mechanizmach automatyzacji pracy redaktorów biblioteki cyfrowej.

Narzędzie do dodawania plików do publikacji planowanej

- ***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

Uruchamianie narzędzia:

run <PREFIX>\<USER_ID>\out\<PUB_ID>\<FILE> false|true

np.

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

<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

<FILE> jest nazwą pliku głównego publikacji

Publikacje planowane

Elementem niezbędnym do prawidłowego działania systemu automatycznego wprowadzania publikacji jest identyfikator publikacji planowanej.

Publikacja planowana – posiada tylko opis, nie posiada treści.

Tworzenie opisów publikacji – ręczne wprowadzanie metadanych lub wykorzystanie mechanizmu importu zaimplementowanego 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.

Umieszczanie plików na serwerze konwersji

- Dla poszczególnych rodzajów publikacji można skonfigurować odpowiednie profile przetwarzania w Document Express'ie.
- Każdemu profilowi przetwarzania Document Expressa zostaje przyporządkowany określony katalog w systemie plików na serwerze. Taka sama struktura katalogów zostaje odwzorowana na dysku lokalnym komputera redaktora.
- Pliki publikacji (TIFF, JPG) muszą zostać umieszczone w katalogach o nazwach odpowiadających identyfikatorom publikacji planowanych w systemie dLibra.
- Redaktor decyduje o parametrach konwersji umieszczając publikację w określonym katalogu na dysku lokalnym.
- Przesyłanie plików na serwer odbywa się za pomocą FTP.
- Na dysku lokalnym komputera redaktora archiwizowane są oryginalne pliki TIFF, na dysku serwera archiwizowane są pliki djvu w trybie bundle oraz – opcjonalnie – pliki TIFF.

Konwersja plików do formatu DjVu + OCR

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),

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.

Logi operacji

Wyniki działania programów realizujących kolejne etapy procesu publikowania są zapisywane do plików log.

Analiza logów pozwala zdiagnozować przyczynę problemu – jeśli taki wystąpi .

Trwają prace nad opracowaniem narzędzia do raportowania błędów – bieżąca analiza wpisów w logach i wysyłanie komunikatu na adres(y) e-mail w przypadku wystąpienia problemu.