POLSKIE BIBIOTEKI CYFROWE 2008 • str. 81-87

Od planowania do publikowania – co można zautomatyzować?

KRZYSZTOF OBER

Poznańska Fundacja Bibliotek Naukowych
krzychu@pfsl.poznan.pl

Streszczenie

Tworzenie publikacji dla potrzeb biblioteki cyfrowej jest procesem wieloetapowym. Podczas gdy realizacja pewnych etapów tego procesu wymaga dużego nakładu pracy redaktora, to inne etapy można spróbować zautomatyzować. Z pomocą przychodzą narzędzia programistyczne: a) wbudowane w system *dLibra*: dedykowane narzędzie do dodawania plików do publikacji planowanej (dostępne od wersji 4.0.10), b) zewnętrzne – np. *Document Express Enterprise* i wbudowany w niego mechanizm wsadowego przetwarzania plików (*Workflow Manager*). Celem niniejszego opracowania jest zaprezentowanie możliwości automatyzacji pracy redaktora biblioteki cyfrowej za pomocą wyżej wymienionego oprogramowania. Punktem wyjścia do rozważań będzie system automatycznego wprowadzania publikacji opracowywany dla potrzeb bibliotek poznańskich tworzących zasób Wielkopolskiej Biblioteki Cyfrowej.

Słowa kluczowe: *dLibra, DocumentExpress, DjVu*, automatyzacja, redaktor biblioteki cyfrowej

Proces tworzenia publikacji

Etapy tworzenia publikacji dla potrzeb biblioteki cyfrowej w bardzo ogólnym ujęciu można wypunktować w następujący sposób:

- 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.

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*).

Przyjrzyjmy się, jak wyżej wymienione etapy przygotowywania materiałów cyfrowych dla biblioteki cyfrowej poddają się procesom automatyzacji.

Etap I: Opracowywanie planów wprowadzania publikacji

Wybór materiałów do umieszczenia w bibliotece cyfrowej uzależniony jest od profilu biblioteki i od przyjętych przez bibliotekę założeń co do gromadzenia określonej zawartości cyfrowej. Proces planowania przebiega w bardziej lub mniej sformalizowany sposób, ale zawsze powinien być poprze-

dzony badaniem zapotrzebowania użytkowników (np. za pomocą ankiet) i analizą statystyk wykorzystania już opublikowanych materiałów. Takie badania i analizy mają duże znaczenie przy podejmowaniu decyzji o wyborze materiałów do digitalizacji. Niejednokrotnie na proces decyzyjny wpływ mają również indywidualne prośby użytkowników o umieszczenie określonych pozycji w bibliotece cyfrowej. Ważnym elementem, który należy uwzględnić na etapie opracowywania planów wprowadzania publikacji, są prawa autorskie. Zgoda autora lub wydawnictwa wymaga podpisania stosownych umów, a jej brak uniemożliwia umieszczenie publikacji w bibliotece cyfrowej.

Etap planowania zawartości biblioteki jest moim zdaniem mało podatny na jakiekolwiek próby automatyzacji. Stosowanie różnego rodzaju narzędzi programistycznych może oczywiście ułatwić pracę, 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. Tutaj raczej nie ma mowy o możliwości jakiejkolwiek automatyzacji. Wyjątkiem jest sytuacja, w której umieszczamy w bibliotece cyfrowej publikację, której opis bibliograficzny istnieje już w jakimś innym katalogu elektronicznym. Wtedy przychodzą nam z pomocą mechanizmy importu opisów bibliograficznych 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.

Należy podkreślić, że mowa tutaj o przygotowywaniu opisów bibliograficznych publikacji planowanych. Identyfikatory tych publikacji będą wykorzystywane na dalszym etapie wprowadzania publikacji – są 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, m.in.:

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

Profile skanowania

Dobór odpowiednich parametrów skanowania zależy od jakości skanowanego materiału oraz jego przeznaczenia. Inaczej będą skanowane materiały przeznaczone do prezentacji na WWW, inaczej materiały przeznaczone do wydruku, a jeszcze inaczej materiały, na których w dalszej kolejności będzie przeprowadzany proces rozpoznawania tekstu. Istnieje możliwość zapisania parametrów skanowania w tzw. profilu skanowania i następnie wielokrotnego wykorzystywania takiego profilu podczas skanowania innych materiałów o zbliżonej charakterystyce.

Automatyzacja zapisu stron

Przed rozpoczęciem skanowania definiuje się m.in. format plików, w którym będą zapisywane skany oraz nazwę katalogu wyjściowego. Kolejne skany są automatycznie numerowane i zapisywane w wybranym katalogu, a nazwy plików, oprócz numeru, mogą zawierać również wcześniej zdefiniowany prefiks.

Przyciski szybkiego dostępu

Istnieje możliwość uruchamiania często używanych funkcji za pomocą przycisków na obudowie skanera – bez konieczności szukania odpowiedniej opcji w menu oprogramowania, np.: szybki podgląd, skanowanie w kolorze, skanowanie w odcieniach szarości, skanowanie tekstu, OCR, otwieranie skanu w programie do obróbki grafiki, itp.

Etap IV: Konwersja plików do formatów używanych w bibliotece cyfrowej, OCR

Konwersję plików do formatu *DjVu*, który jest formatem najczęściej stosowanym dla publikacji umieszczanych w Wielkoplskiej Bibliotece Cyfrowej, można przeprowadzić całkowicie automatycznie – bez najmniejszej ingerencji redaktora. Rozważany system automatycznego wprowadzania publikacji zakłada konwersję do formatu *DjVu* plików graficznych (TIFF, JPG, GIF) oraz plików PDF. Redaktor jednym kliknięciem myszy umieszcza skany publikacji (najczęściej są to pliki TIFF) 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 zadane parametry konwersji,
- wykona OCR,
- wygeneruje pliki *DjVu* w trybie *indirect* (dla potrzeb prezentacji na stronie www).

Etap V: Umieszczanie publikacji w bibliotece cyfrowej, publikowanie

Po zakończeniu konwersji i opcjonalnym rozpoznaniu tekstu, pliki publikacji muszą zostać umieszczone w bibliotece cyfrowej. Ten etap również nie wymaga udziału redaktora. System automatycznego umieszczania publikacji wysyła gotowe pliki na serwer Wielkopolskiej Biblioteki Cyfrowej wykorzystując identyfikator publikacji planowanej. Dodatkowo, jeśli zostanie ustawiony odpowiedni parametr, nowo umieszczona publikacja może zostać automatycznie opublikowana.

Aby etapy: IV i V mogły zostać zrealizowane w sposób automatyczny, redaktor musi zadbać o następujące rzeczy:

- 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.

Szczegóły dotyczące nazewnictwa plików publikacji oraz definiowania parametrów konwersji poprzez umieszczanie plików w odpowiednich katalogach omówione zostaną w kolejnych rozdziałach.

Profile konwersji i zadania przetwarzania

Automatyzacja IV etapu procesu tworzenia publikacji cyfrowej (konwersja do formatu *DjVu* oraz OCR) została zrealizowana w oparciu o oprogramowanie *DocumentExpress Enterprise* oraz własne skrypty napisane w języku *Perl*.

Każda operacja konwersji do formatu *DjVu* za pomocą oprogramowania *DocumentExpress* opiera się na dość pokaźnym zestawie parametrów, od których zależą wyniki konwersji. Parametry dobiera się w zależności od jakości materiałów źródłowych oraz przeznaczenia materiałów wynikowych. Często doboru tych parametrów dokonuje się metodą prób i błędów. *DocumentExpress* umożliwia zapisywanie parametrów konwersji w tzw. profilach konwersji. Raz utworzony profil konwersji można wielokrotnie wykorzystywać. Dla poszczególnych rodzajów publikacji umieszczanych w bibliotece cyfrowej (np. pocztówki, skrypty, gazety) można skonfigurować i zapisać odpowiednie profile konwersji. Te profile są następnie wykorzystywane w definiowaniu tzw. zadań przetwarzania. Każdy profil może być wykorzystany w wielu zadaniach przetwarzania. Dla każdego redaktora, który posiada konto w systemie automatycznego wprowadzania publikacji, tworzy się zestaw zadań przetwarzania (w zależności od potrzeb). Dla różnych rodzajów publikacji tworzy się oddzielne zadania przetwarzania. I tak np. oddzielne zadania przetwarzania zostaną utworzone dla starodruków, oddzielne dla książek i skryptów oraz oddzielne dla gazet i czasopism. Z każdym zadaniem przetwarzania związany jest katalog w systemie plików, z którego pobierane są pliki do przetwarzania i do którego zapisywane są pliki wynikowe konwersji. Tak więc redaktor może wybrać odpowiednie parametry konwersji poprzez umieszczenie plików w wybranym katalogu na serwerze systemu. Np. redaktor, który w systemie automatycznego wprowadzania publikacji posiada konto o nazwie *jkowalski,* umieści skany pocztówek w katalogu c:\pliki\jkowalski\in\photo_300 (zdjęcia w rozdzielczości 300 dpi), a skan dobrej jakości skryptu w katalogu c:\pliki\jkowalski\in\book_ocr (skan tekstu, na którym zostanie wykonany OCR). Dodatkowo skany publikacji muszą zostać umieszczone w katalogach o nazwach odpowiadających identyfikatorom publikacji planowanych, np. c:\pliki\jkowalski\in\photo_300\22472 (publikacja o identyfikatorze 22472), czy c:\pliki\jkowalski\in\book_ocr\12355 (publikacja o identyfikatorze 12355). Umieszczanie plików skanów na serwerze odbywa się w ten sposób, że redaktor gromadzi pliki przeznaczone do konwersji na dysku lokalnym swojego komputera, a w wybranym przez siebie momencie przesyła wszystkie pliki za pomocą programu *ncftp* na serwer. Na dysku lokalnym komputera redaktora odwzorowana jest struktura katalogów, którą każdy użytkownik systemu posiada na swoim koncie na serwerze. Klikając w ikonę na pulpicie swego komputera, redaktor uruchamia program *ncftp*, który przenosi przygotowane pliki (skany publikacji) do odpowiednich katalogów na serwerze. Parametry połączenia FTP (adres serwera, nazwa użytkownika oraz hasło) zapisane są w pliku konfiguracyjnym programu *ncftp*, tak więc redaktor nie musi wprowadzać tych danych przy każdym przesyłaniu plików na serwer. Po umieszczeniu skanów publikacji na serwerze, są one poddawane obróbce przy użyciu oprogramowania *DocumentExpress Enterprise* (z wykorzystaniem przetwarzania wsadowego i mechanizmu tzw. aktywnych folderów) oraz własnych skryptów w języku *Perl*, w wyniku której w katalogu wyjściowym pojawiają się gotowe do umieszczenia w bibliotece cyfrowej pliki *DjVu*. Katalog, który w omawianym IV etapie procesu tworzenia publikacji cyfrowej jest katalogiem wyjściowym, w etapie V (przesyłanie plików do biblioteki cyfrowej) staje się katalogiem wejściowym. Dla przykładowego redaktora *jkowalski*, pliki wynikowe IV etapu zostaną umieszczone w katalogu c:\pliki\

jkowalski\out\. Poszczególne publikacje będą oczywiście umieszczone w katalogach o nazwach odpowiadających identyfikatorom publikacji planowanych, np.: c:\pliki\jkowalski\out\22472, czy c:\pliki\jkowalski\out\12355.

Narzędzie do dodawania plików do publikacji planowanej

Automatyzacja V etapu procesu tworzenia publikacji cyfrowej (umieszczanie publikacji w bibliotece cyfrowej i publikowanie) została zrealizowana w oparciu o narzędzie do dodawania plików do publikacji planowanej (wbudowane w system *dLibra*: http://dlibra.psnc.pl/community/pages/viewpage.action?pageId=4259865) oraz własne skrypty.

Narzędzie znajduje się w dystrybucji systemu *dLibra* począwszy od wersji 4.0.10 i zawiera następujące elementy:

lib – katalog zawierający biblioteki potrzebne do uruchomienia narzędzia,
config.xml – plik konfiguracyjny zawierający informacje o serwerze,
users.xml – plik konfiguracyjny zawierający informacje o użytkownikach,
run.bat – skrypt uruchamiający narzędzie w środowisku systemów z rodziny Windows,
run.sh – skrypt uruchamiający narzędzie w środowisku systemów z rodziny Linux.

Pliki konfiguracyjne narzędzia:

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 specify 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>
```

Jest to plik XML zawierający informacje o serwerze *dLibra*, do którego narzędzie będzie się podłączać. W pliku definiuje się adres serwera *dLibra* (klucz server) oraz port, na którym serwer *dLibra* oczekuje na połączenia (klucz port). W powyższym przykładzie serwer *dLibra* uruchomiony jest na lokalnym komputerze (*localhost*) i oczekuje połączeń na porcie *10051*.

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>
```

Jest to plik XML zawierający informacje o użytkownikach, w imieniu których narzędzie będzie logować się do serwera *dLibra* w celu dodania plików do publikacji planowanej. W pliku definiuje się nazwy użytkowników oraz hasła dostępu. Nazwa klucza odpowiada nazwie użytkownika, natomiast hasło jest wartością klucza. W powyższym przykładzie zdefiniowano użytkownika *jkowalski*, któremu przyporządkowano hasło *mojehaslo*.

Uruchamianie narzędzia:

`run <PATH_TO_MAIN_FILE_NAME> false|true`

Pierwszy parametr `(PATH_TO_MAIN_FILE_NAME)` oznacza ścieżkę do pliku głównego publikacji, natomiast drugi parametr określa, czy po dodaniu plików wydanie ma zostać opublikowane (*true*) czy nie *(false).*

Ścieżka do pliku głównego publikacji również zawiera w sobie kilka parametrów. Jest ona zbudowana w następujący sposób:

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

gdzie:

<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.

I tak np. wydanie polecenia:

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

spowoduje umieszczenie na serwerze biblioteki cyfrowej plików publikacji o identyfikatorze *22345* i opublikowanie nowego wydania publikacji. Publikacja zostanie umieszczona w imieniu użytkownika *jkowalski*.

Narzędzie do dodawania plików do publikacji planowanej jest uruchamiane za pośrednictwem skryptów napisanych w języku *Perl*. Skrypty kontrolują zawartość katalogów *out* (katalogów wyjściowych) wszystkich użytkowników systemu automatycznego wprowadzania publikacji i uruchamiają narzędzie, jeśli w katalogu *out* pojawi się nowy katalog z plikami publikacji.

Plany na przyszłość

Opisany tutaj system automatycznego wprowadzania publikacji jest opracowywany dla potrzeb poznańskich bibliotek tworzących zasób Wielkopolskiej Biblioteki Cyfrowej. W chwili obecnej dobrze wspomaga pracę redaktorów WBC. Myślimy jednak nad zwiększeniem zakresu jego

możliwości. Oto kilka pomysłów, które w najbliższym czasie chcielibyśmy wdrożyć, częściowo przy współpracy programistów tworzących platformę *dLibra*:

- 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 redaktorowi zaoszczędzić sporo wysiłku, automatyzując w całości lub częściowo pewne etapy jego pracy. Jest to możliwe dzięki programom uruchamianym z linii komend, które można oskryptować i uruchamiać wsadowo. System *dLibra* wspiera takie rozwiązania i dostarcza odpowiednich narzędzi do ich realizacji. Być może jeszcze jakieś inne funkcjonalności systemu *dLibra* dałoby się zaimplementować w podobny sposób?