POLSKIE BIBLIOTEKI CYFROWE 2010 • **str. 55-62**

Przetwarzanie i OCR czasopism drukowanych gotykiem krok po kroku

TOMASZ KALOTA, RAFAŁ RACZYŃSKI, PAWEŁ RĘKAR

Biblioteka Uniwersytecka we Wrocławiu

Streszczenie

Referat prezentuje krok po kroku proces przygotowania publikacji cyfrowych będących odwzorowaniem czasopism drukowanych czcionką gotycką. Tego typu materiały stanowią dosyć pokaźny zasób polskich bibliotek i w związku z tym celowe jest opracowanie metody umożliwiającej sprawne przygotowanie funkcjonalnej publikacji cyfrowej spełniającej najlepsze standardy jakościowe. Głównym celem referatu jest skonfrontowanie przyjętych schematów organizacyjnych i logistycznych oraz zastosowanych rozwiązań technicznych z metodami digitalizacji przyjętymi w innych instytucjach zajmujących się digitalizacją i tworzeniem bibliotek cyfrowych. Autorzy referatu wyrażają przekonanie, że jednym z kluczowych czynników wpływającym na efektywność i obniżenie kosztów procesów digitalizacji jest ich automatyzacja. Niniejszy referat ma więc być zachętą do analizy poszczególnych kroków cyfryzacji obiektów bibliotecznych oraz podjęcia próby ich udoskonalania i usprawniania.

Słowa kluczowe: digitalizacja, OCR, logistyka, przetwarzanie danych, biblioteki cyfrowe

1. Wstęp

Cały proces digitalizacji materiałów bibliotecznych można podzielić na pięć etapów:

- digitalizacja,
- przygotowanie plików źródłowych,
- rozpoznanie tekstu – OCR,
- przygotowanieLINK\l"id.e581szvkkizv" plików prezentacyjnych,
- publikacja w bibliotece cyfrowej.

Każdy z wymienionych etapów wymaga zaprojektowania i skonfigurowania warsztatu pracy (wybór sprzętu i oprogramowania) oraz skoordynowania poszczególnych działań oraz zapewnienia płynności prac w wymiarze całego procesu. Autorzy referatu zaprezentują własne doświadczenia zdobyte podczas konfigurowania oraz obsługi linii technologicznej dedykowanej dla digitalizacji czasopism drukowanych gotykiem w Bibliotece Uniwersyteckiej we Wrocławiu.

2. Digitalizacja

Digitalizacja dziewiętnastowiecznych czasopism jest trudnym zadaniem ze względu na ich jakość i stan zachowania. Podstawowym utrudnieniem, a zarazem powodem konieczności szybkiego zabezpieczania tych czasopism jest kruchy i rozsypujący się kwaśny papier, na którym były drukowane. Dodatkowe trudności przysparzają często opasłe oprawy introligatorskie, którymi trudno manipulować podczas skanowania. W związku z tym, planując digitalizację tego typu materiałów, warto rozważyć możliwość wykorzystania form pośrednich, jakimi są mikrofilmy. Znaczna część tego typu zbiorów została już zabezpieczona za pomocą technologii mikrofilmowania. Wiele instytucji posiada jeszcze kamery mikrofilmowe, które można wykorzystać do szybkiego zabezpieczenia treści narażonej na zniszczenie. Poza tym, kamery mikrofilmowe, dzięki swojej konstrukcji lepiej radzą sobie z ułożeniem wspomnianych obszernych opraw introligatorskich. W archiwum Forum EBIB dostępna jest dyskusja na ten temat: http://ebib.oss.wroc.pl/phpBB/viewtopic.php?t=3969.

Czasopismo Schlesische Privilegirte Staats-, Kriegs- und Friedens-Zeitung, którym posłużyliśmy się jako przykładem do opisania całego procesu digitalizacji, zostało już wcześniej zmikrofilmowane przez Bibliotekę Uniwersytecką we Wrocławiu, co w oczywisty sposób zadecydowało o wyborze mikrofilmu jako źródła pozyskania zapisu cyfrowego.

Efektywna digitalizacja mikrofilmów możliwa jest do zrealizowania za pomocą specjalnych skanerów, które w sposób automatyczny skanują całe zwoje mikrofilmów. Przykładami takich skanerów są:

- SunRise – http://www.sunriseimaging.com/
- Zeutschel OM 1600 – http://www.zeutschel.com/products/microfilm_scanner_om1600.html

Za pomocą tego typu sprzętu można skanować od kilku do kilkunastu standardowych rolek mikrofilmowych dziennie. Są to jednak drogie urządzenia. Warto więc rozważyć inwestycję we własne urządzenie, względem zlecania takich usług na zewnątrz.

Jednym z ważniejszych etapów podczas planowania masowej digitalizacji mikrofilmów jest ocena i przygotowanie materiału źródłowego, a następnie dobranie parametrów, które zapewnią dobrą jakość zapisu cyfrowego.

Pierwsze zadanie, które pochłania czas, ale jest konieczne do wykonania, to przegląd mikrofilmów w celu sprawdzenia odpowiedniej długości „rozbiegówki" oraz defektów w postaci naderwań, rozklejeń itp. Każda tego typu usterka wykryta podczas skanowania oraz konieczność jej usunięcia powoduje znaczne straty czasowe, więc przygotowanie materiału okazuje się konieczne do zapewnienia płynności całego procesu.

Kolejnym zadaniem jest ustawienie parametrów skanera takich jak format, rozdzielczość, jasność skanowania, kontrast, przycięcie itp. Te parametry zmieniają się w zależności od jakości mikrofilmu i wymagają ciągłej kontroli. Przydatne w tym kontekście jest zapisywanie profili z różnymi ustawieniami i następnie wykorzystywanie ich stosownie do potrzeb. Tutaj bardzo cenne jest doświadczenie operatora skanera, który powinien umieć ocenić jakość mikrofilmu i dobrać odpowiedni zestaw parametrów digitalizacji.

Przed przystąpieniem do skanowania należy przygotować odpowiednią ilość miejsca na przechowywanie plików, które produkowane są bardzo szybko i zajmują sporo powierzchni dyskowej. Odpowiednia ilość dysków, stanowiących bufor do tymczasowego przechowywania półproduktów, jest niezbędna do zapewnienia płynności całego procesu, tak samo jak gotowość do dalszego przetwarzania plików oraz ich archiwizowania.

Po wykonaniu prac przygotowawczych można przystąpić do skanowania, którego celem, w przypadku wspomnianego wcześniej czasopisma, jest wyprodukowanie plików tif w rozdzielczości 600 dpi oraz w trybie *grayscale* (odcienie szarości).

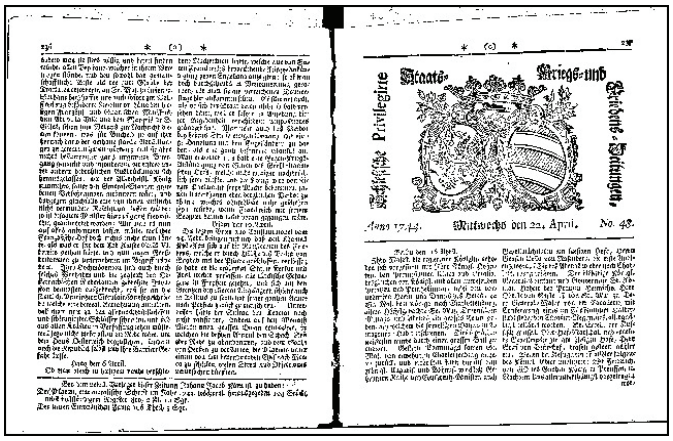
3. Przygotowanie plików źródłowych

Przygotowanie plików źródłowych to zadanie, którego celem jest stworzenie jak najlepszego materiału, który następnie zostanie poddany obróbce OCR (ang. *Optical Character Recognition*). Jakość rozpoznanego tekstu w znacznym stopniu zależy od jakości materiału wejściowego. Należy więc zadbać o to, aby pliki źródłowe zostały przygotowane z należytą starannością oraz z uwzględnieniem wszystkich szczegółów, mających wpływ na jakość wynikowej publikacji cyfrowej. Ustalenie odpowiednich parametrów skanowania i przetwarzania jest czynnością żmudną i czasochłonną, ponieważ wymaga przeprowadzenia odpowiedniej ilości prób. Ponadto zmienność parametrów wejściowych materiałów bibliotecznych powoduje konieczność ciągłej kontroli i korygowania wypracowanych wcześniej parametrów digitalizacji.

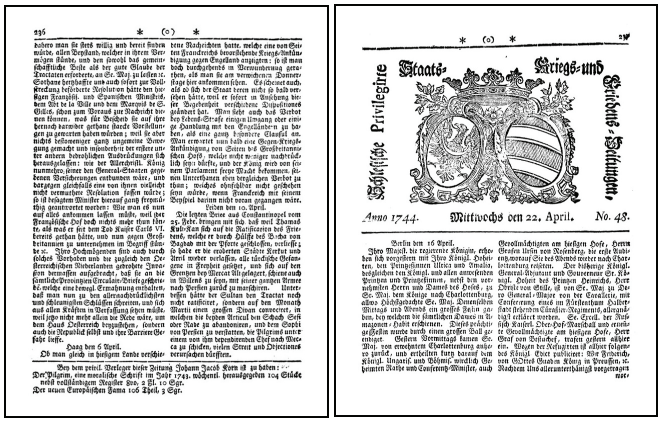
Pierwszym etapem obróbki jest wstępne wyprostowanie tekstu na zeskanowanych stronach oraz konwersja plików do wersji jednobitowej. Do tej operacji został wybrany program FineReader 10, który podczas otwierania plików wyrównuje wiersze z tekstem i zapisuje do dowolnego formatu. Pliki wstępnie przygotowane przez FineReader trafiają do progamu XnView (httpNK"http://www.google.com/url?q=http%3A%2F%2Fwww.xnview.com%2Fen%2Fdownloadwin32.html&sa=D&sntz=1&usg=AFQjCNFUJjnm3TUDnJUuePvv8trpXUjfdw":://www.xnview.com/en/downloadwin32.html), jest to w zasadzie przeglądarka plików, ale wyposażona w bardzo bogate funkcje do wsadowego przetwarzania plików.

4. Opis obróbki plików źródłowych w programie XnViev i FineReader 10

Celem obróbki jest uzyskanie wyprostowanych oraz jednakowo wykadrowanych pojedynczych stron czasopisma, które następnie zostaną przekazane do rozpoznania tekstu w programie FineReader XIX.

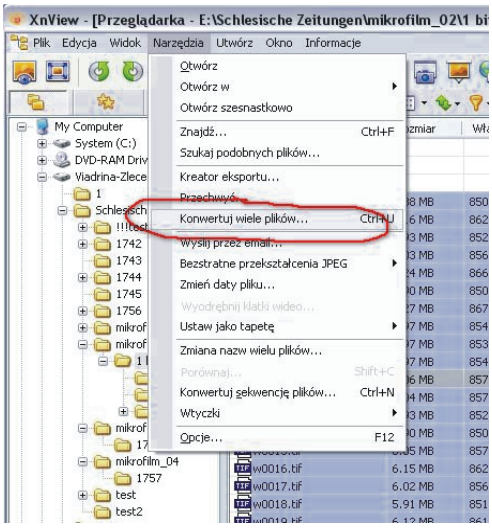


Ryc. 1. Plik przed obróbką



Ryc. 2. Pliki po obróbce

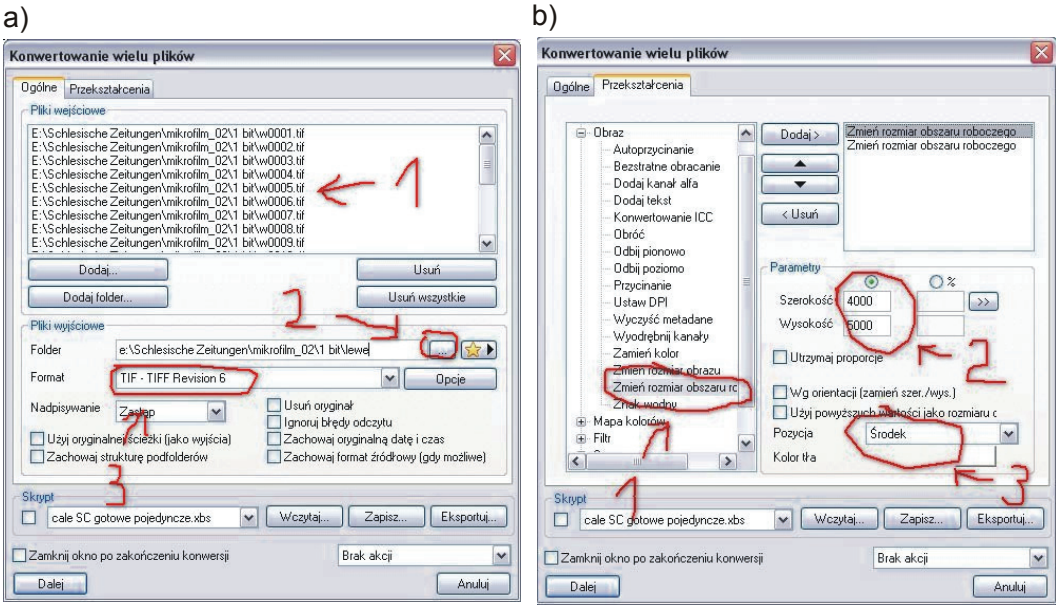
Do uzyskania takiego efektu wykorzystamy konwenter plików XnView (ryc. 3.).



Ryc. 3

Pliki wczytujemy do naszego konwertera (ryc. 4a – 1), ustawiamy lokalizację w której mają być zapisywane pliki wynikowe (ryc. 4a – 2) oraz podajemy format zapisu (ryc. 4a – 3).

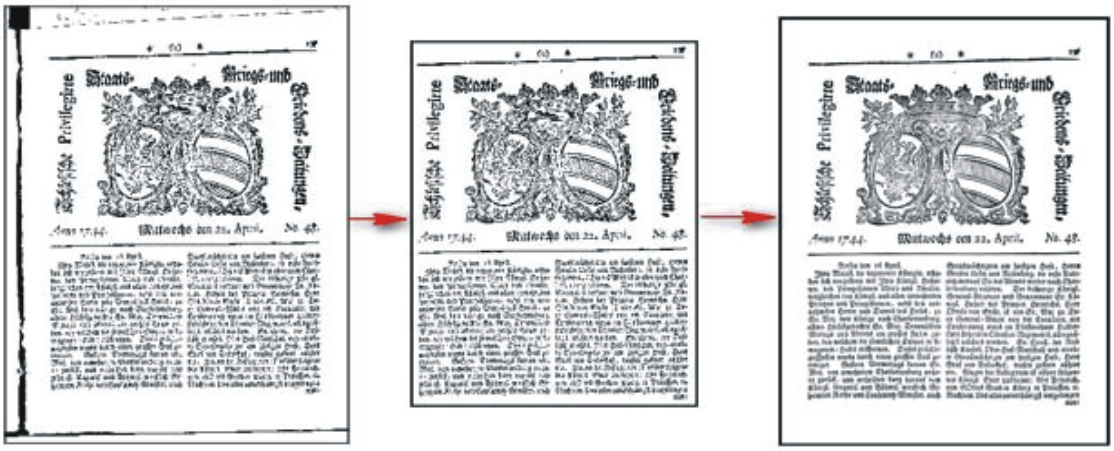
Następnie przechodzimy do zakładki „przekształcenia", gdzie ustalamy rozmiar strony, wybieramy „Zmień rozmiar obszaru roboczego" (ryc. 4b – 1), wpisujemy szerokość i wysokość (ryc. 4b – 2) i od której krawędzi ma ścinać plik, lewa prawa lub środek (ryc. 4b – 3).



Ryc. 4

Funkcje „Zmień rozmiar obszaru roboczego" możemy stosować wielokrotnie w tej samej akcji, co nam ułatwi dokładniejsze przycięcie pliku. Najpierw przycinamy plik na połowę. Gdy otrzymamy lewe i prawe pliki, wyrównujemy je ponownie w programie FineReader 10.

Po tych operacjach możemy już na gotowo przyciąć plik, czyli wracamy do naszego konwertera i ustalamy wymiar na pojedynczy plik, funkcja "Zmień rozmiar obszaru roboczego", przycinamy do tekstu, uwzględniając możliwość przesuwania się tekstu na stronie, po czym dodajemy białe tło.



Ryc. 5

5. Rozpoznanie tekstu

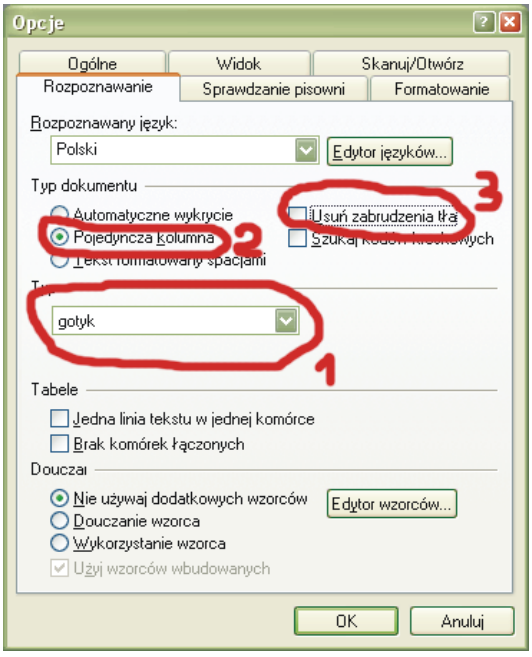
Rozpoznanie tekstu drukowanego czcionką gotycką jest procesem dosyć kosztownym ze względu na sposób licencjonowania oprogramowania wykorzystywanego do obróbki OCR -FineReader XIX. Producent określa, ile stron można przetworzyć w ramach jednej licencji i w związku z tym należy zadbać o to, aby rozpoznawania tekstu nie trzeba było powtarzać ze względu na niezadowalające efekty spowodowane niską jakością materiału wejściowego. Ponadto w niektórych przypadkach warto rozważyć wykorzystanie różnych wersji oprogramowania, aby nie eksploatować droższych licencji do wykonywania czynności, które tych licencji nie wymagają.

Po wczytaniu plików źródłowych do wiązki w programie FineReader XIX należy ustawić odpowiednie opcje rozpoznawania. Podczas prac nad przygotowaniem cyfrowych wersji czasopisma Schlesische Privilegirte Staats-, Kriegs- und Friedens-Zeitung zauważono, że istotnymi opcjami mającymi wpływ na jakość rozpoznania tekstu są:

- typ druku – gotyk,
- typ dokumentu – pojedyncza kolumna,
- typ druku – usuń zabrudzenia tła.

Wybór parametru gotyk jako typ druku jest oczywisty, ale trzeba o tym pamiętać, ponieważ domyślnie nie jest on wybrany i rozpoczęcie rozpoznawania tekstu przy ustawieniach domyślnych powoduje wykorzystanie limitu przydzielonego w ramach licencji.

Ustawienie funkcji „Pojedyncza kolumna” w „Typ dokumentu” jest uzasadnione tym, że w przypadku starszych czasopism oprogramowanie ma kłopot z jednoznacznym wykryciem obszaru z tekstem do rozpoznania. Zdarzały się przypadki, w których pewne fragmenty tekstu zostały zakwalifikowane jako grafika, co powodowało wykluczenie ich z procesu rozpoznania tekstu. Ponadto na znacznej części stron tekst był wykrywany jako „pojedyncza kolumna”, mimo że faktycznie było tych kolumn więcej. Te obserwacje zadecydowały o wyborze ustawienia pojedynczej kolumny. Usuwanie zabrudzeń tła jest bardzo przydatna funkcją, ale w przypadku druków współczesnych. Pozostawienie tej opcji włączonej powodowało usuwanie drobnych punktów, które w rzeczywistości były fragmentami druku, co zmniejszało skuteczność rozpoznania tekstu. Ta opcja domyślnie jest włączona, więc należy zwrócić uwagę na to, aby przed rozpoczęciem rozpoznawania tekstu zmienić jej ustawienie.



Ryc. 6

Po zakończeniu rozpoznawania tekstu wynik pracy zostaje wyeksportowany do plików PDF zawierających warstwę graficzną oraz tekstową.

6. Przygotowanie plików prezentacyjnych

Przygotowanie plików prezentacyjnych polega na wyprodukowaniu gotowych publikacji cyfrowych przeznaczonych do udostępnienia w bibliotece cyfrowej. Proces ten można w znacznym stopniu zautomatyzować wykorzystując przetwarzanie wsadowe oraz realizując je w czasie najmniejszego obciążenia sprzętu, np. w godzinach nocnych.

Poniżej opisano proces przygotowania plików prezentacyjnych w dwóch wariantach DjVu i PDF.

6.1. Pliki prezentacyjne w formacie DjVu

W celu konwersji plików z formatu PDF na DJVu można posłużyć się następującymi programami:

- Document Express Enterprise – http://www.djvu.com.pl/de_family.php
- Serwis any2djvu – http://any2djvu.djvuzone.org
- Djvudigital – http://djvu.sourceforge.net/doc/man/djvudigital.html
- Pdf2djvu – http://code.google.com/p/pdf2djvu/

Zgodnie z dostępnym w sieci porównaniem (http://code.google.com/p/pdf2djvu/wiki/DjVuDigital) w tej chwili, pdf2djvu wydaje się najkorzystniejszym rozwiązaniem do zrealizowania celów postawionych przy digitalizacji czasopism drukowanych gotykiem Najważniejsze zalety tego rozwiązania to:

- do tworzonego dokumentu dołączono niewidoczny tekst oraz metadane (jeśli jest), co umożliwia używanie go do dalszej obróbki plików wynikowych programu ABBY FineReader,
- większe możliwości wyboru kompresji grafiki (djvu używa tylko trybu bezstratnego dla obrazów monochromatycznych),
- do działania nie wymaga komercyjnego oprogramowania (bądź też oprogramowania na licencji niekompatybilnej z GPL),
- dostęp do obszernej dokumentacji autorstwa Jakuba Wilka – http://students.mimuw.edu.pl/~jw209508/papers/thesis/thesis.pdf

Dalszy ciąg obróbki plików wygląda następująco. Na serwerze konwersji, udostępnione są katalogi: wejściowy (Input) oraz wyjściowy (Output). Przygotowane pliki pdf kopiowane są do folderu Input. Wykonujący się cyklicznie (co 10 minut) skrypt sprawdza, czy w katalogu Input są jakieś pliki pdf, a jeśli tak, to uruchamia konwerter pdf2djvu z ustalonymi wcześniej parametrami (jakość 600 dpi, pliki scalone, wyłączony antyaliasing). Wyniki jego pracy zapisują się w folderze Output. Jeśli skrypt nie zdąży skonwertować danej partii materiału przed swoim kolejnym wywołaniem, następna instancja konwertera nie jest wywoływana, co zapobiega duplikacji pracy nad tymi samymi plikami. Po zakończeniu obróbki katalog wejściowy jest opróżniany i gotowy na przyjęcie następnej paczki plików. Wszystkie operacje podczas pracy są zapisywane do osobnego pliku logu, dzięki czemu możliwe jest lepsze monitorowanie pracy.

Do dalszego zautomatyzowania pracy wykorzystywany jest kolejny skrypt, którego zadaniem jest:

- ustawianie koloru nagłówka i stopki,
- tworzenie miniaturek,
- rozdzielenie tak przygotowanych plików i przekopiowanie nowo powstałych do osobnych katalogów,
- dołączenie do katalogów publikacji plików opisujących (publication.properties, directory.rdf).

Poniżej zaprezentowano kod skryptu jazdaDjVu.bat, który automatycznie przygotowuje wsad do dLibry.

6.2. Pliki prezentacyjne w formacie PDF

W tym wariancie otrzymane z FineReadera pliki PDF są poprzez skrypt przenoszone do katalogów o nazwach plików, a same pliki przemianowuje się na directory.pdf. Kolejnym krokiem jest dołączenie do katalogów plików opisujących publikację (jak w wariancie z DJVu).

```
@echo off

if not exist djvused.exe goto BrakPlikow
if not exist djvmcvt.exe goto BrakPlikow
if not exist publica
tion.properties goto BrakPlikow
if not exist directory.rdf goto BrakPlikow
if not exist color_header_ant.txt goto BrakPlikow

echo.
echo Zadanie 1 - Ustawianie koloru, nglowka i stopki

for /R %%x in (*.djvu) do djvused %%x -f color_header_ant.txt -s

echo Zrobione

echo Zadanie 2 - tworzenie miniaturek

for %%x in (*.djvu) do djvused -e 'set-thumbnails' -s %%x

echo Zrobione

echo Zadanie 3 - Rozdzielanie plikow DjVu

for %%x in (*.djvu) do mkdir %%~nx
for %%x in (*.djvu) do djvmcvt -i %%x %%~nx directory.djvu

echo Zrobione

echo.
echo Zadanie 4 - Kopiowanie plikow opisujacych publikacje
rem R.R.

for /R %1 %%i in (.) do copy publication.properties "%%i"
for /R %1 %%i in (.) do copy directory.rdf "%%i"

echo Zrobione

echo.
echo Wszystkie zadania wykonane poprawnie
echo.
pause

goto koniec

:BrakPlikow
echo.
echo Brak ktoregos programu lub pliku w katalogu !!!
echo.
pause

:koniec
```

Poniżej zaprezentowano kod skryptu jazdaPDF.bat, który automatycznie przygotowuje wsad do dLibry. Ostatnim etapem przygotowania plików prezentacyjnych PDF jest ich optymalizacja do przeglądania w Internecie wykonywana przy pomocy programu Adobe Acrobat.

7. Publikacja w bibliotece cyfrowej

Publikowanie w bibliotece cyfrowej dużej ilości numerów czasopism możliwe jest do zrealizowania w sposób automatyczny dzięki funkcji masowego ładowania publikacji. Konieczne jest wcześniejsze przygotowanie wsadu do biblioteki składającego się ze struktury publikacji oraz plików publication.propertis i directory.rdf.

```
@echo off
if not exist publication.properties goto BrakPlikow
if not exist directory.rdf goto BrakPlikow
for /R %1 %%i in (*.pdf) do (
	for /F %%a in ("%%~Ni") do (
		mkdir %%a
		copy publication.properties %%a
		copy directory.rdf %%a
		copy "%%a.pdf" "%%a\directory.pdf"
		rmdir /q /s directory
	)
)
echo.
echo Wszystkie zadania wykonane poprawnie
echo.
pause
goto koniec
:BrakPlikow
echo.
echo Brak ktoregos programu lub pliku w katalogu !!!
echo.
pause
:koniec
```

Pliki te zawierają informacje dla dLibry potrzebne do właściwego umiejscowienia publikacji w bibliotece cyfrowej oraz do wygenerowania opisu publikacji. Gotowa struktura publikacji jest wynikiem działania programu jazdaDjVu.bat lub jazdaPDF.bat, które zostały opisane wcześniej i które przygotowują strukturę katalogów z umieszczonymi w nich odpowiednimi numerami czasopisma oraz kopiami plików opisujących publikację. Aby przygotować gotowy wsad do dLibry, należy przeedytować każdy plik opisujący: publication.propertis i directory.rdf i zamieścić w nich odpowiednie wartości opisujące konkretny numer czasopisma. Cała operacja sprowadza się do edycji pliku, np. w notatniku i wpisanie odpowiedniej daty i numeru czasopisma. Ważne jest, aby notatnik obsługiwał kodowanie znaków UTF-8, w przeciwnym razie dLibra będzie informowała o błędach i przerwie publikowanie plików. Nie może więc to być standardowy notatnik dostępny w systemie Widows, ale np. programy: Notaped2 lub Notaped++. Po przygotowaniu struktury publikacji oraz plików opisujących można przystąpić do ich wysyłania do biblioteki cyfrowej. W aplikacji redaktora należy wybrać opcję masowego ładowania publikacji, po czym wskazać lokalizację przygotowanej struktury plików.

8. Podsumowanie

Opisany powyżej proces przygotowania publikacji cyfrowych został zaprojektowany dla konkretnego typu zbioru, ale każdy z jego etapów może być realizowany niezależnie i być wykorzystany w projektowaniu alternatywnych linii technologicznych, dedykowanych dla innych typów zbiorów archiwalnych i bibliotecznych. Autorzy referatu liczą na dyskusję dotyczącą udoskonalania procesów digitalizacji oraz alternatywnych pomysłów dotyczących organizowania linii technologicznych umożliwiających automatyzację digitalizacji. W tym celu przygotowywany jest serwis internetowy www.digitalizacja.pl, który w zamierzeniu twórców ma się stać miejscem prezentacji i analizowania pomysłów na digitalizację różnego rodzaju materiałów.