Przetwarzanie i OCR czasopism drukowanych gotykiem - krok po kroku

Tomasz Kalota, Rafał Raczyński, Paweł Rękar

www.BibliotekaCyfrowa.pl

Plan prezentacji

Proces digitalizacji materiałów bibliotecznych można podzielić na pięć etapów:

- digitalizacja,
- przygotowanie plików źródłowych,
- rozpoznanie tekstu – OCR,
- przygotowanie plików prezentacyjnych,
- publikacja w bibliotece cyfrowej.

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. Dodatkowych 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.

Digitalizacja

Efektywna digitalizacja mikrofilmów możliwa jest do zrealizowania przy pomocy 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.

Przy pomocy tego typu sprzętu można skanować od kilku do kilkunastu standardowych rolek mikrofilmowych dziennie.

Digitalizacja

Zadania przy realizacji digitalizacji mikrofilmów:

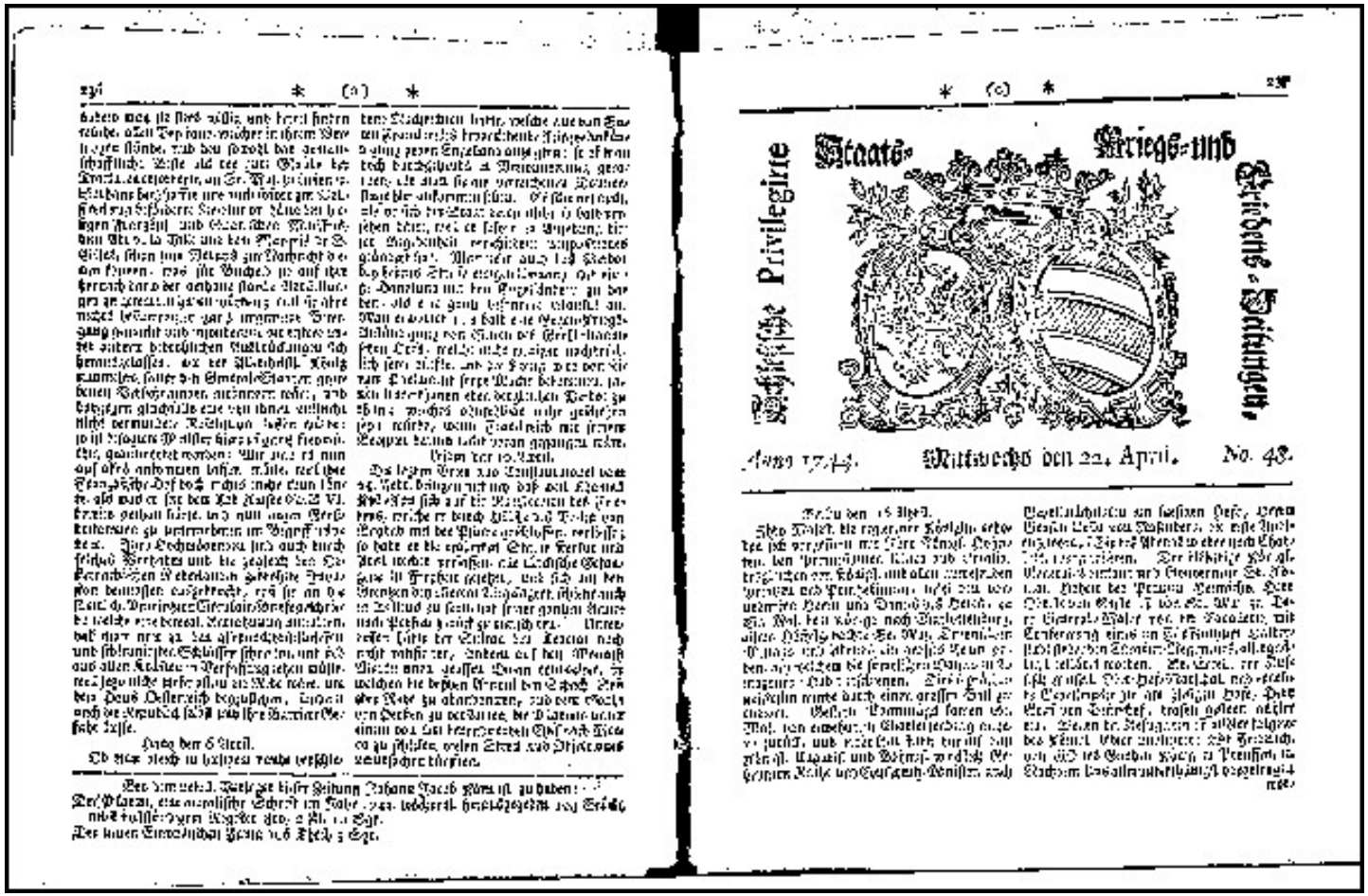
- określenie parametrów wynikowych plików źródłowych (tif, 600 dpi, grayscale)
- ocena i przygotowanie materiału źródłowego - mikrofilmu,
- dobranie parametrów digitalizacji, które zapewnią dobrą jakość zapisu cyfrowego
- kontrola parametrów i jakości plików źródłowych
- przygotowanie odpowiedniej ilości miejsca na przechowywanie plików roboczych.

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.

Przygotowanie plików źródłowych

Plik przed obróbką



Przygotowanie plików źródłowych

Pliki po obróbce

236 * (o) *

dahero man sie stets willig und bereit finden würde, allen Beystand, welcher in ihrem Vermögen stünde, und den sowohl das gemeinschaffliche Beste als der gute Glaube der Tractaten erforderte, an Se. Maj. zu leisten ꝛc. Sothane hertzhaffte und auch sofort zur Vollstreckung beförderte Resolution hätte den hiesigen Französ. und Spanischen Ministris, dem Abt de la Ville und dem Marquis de S. Gilles, schon zum Voraus zur Nachricht dienen können, was für Bescheid sie auf ihre hernach darwider gethane starcke Vorstellungen zu gewarten haben würden; weil sie aber nichts destoweniger gantz ungemeine Bewegung gemacht und insonderheit der erstere unter andern bedrohlichen Ausdrückungen sich herausgelassen: wie der Allerchrist. König nunmehro seiner den General-Staaten gegebenen Versicherungen entbunden wäre, und darzegen gleichfalls eine von ihnen vielleicht nicht vermuthete Resolution fassen würde; so ist besagtem Minister hierauf gantz freymüthig geantwortet worden: Wie man es nun auf alles ankommen lassen müste, weil der Französische Hof doch nichts mehr thun könte, als was er seit dem Tod Kaiser Carls VI. bereits gethan hätte, und nun gegen Großbritannien zu unternehmen im Begriff stünde ꝛc. Ihro Hochmögenden sind auch durch solches Vorhaben und die zugleich den Oesterreichischen Niederlanden gedrohete Invasion dermassen aufgebracht, daß sie an die sämtliche Provintzien Circulair-Briefe geschrieben, welche eine bewegl. Ermahnung enthalten, daß man nun zu den allernachdrücklichsten und schleunigsten Schlüssen schreiten, und sich aus allen Kräften in Verfassung setzen müste, weil jetzo nicht mehr allein die Rede wäre, um dem Haus Oesterreich beyzustehen, sondern auch die Republick selbst und ihre Barriere Gefahr liesse.

Haag den 6 April.

Ob man gleich in hiesigem Lande verschiedene Nachrichten hatte, welche eine von Seiten Franckreichs bevorstehende Kriegs-Ankündigung gegen Engelland anzeigten: so ist man doch durchgehends in Verwunderung gerathen, als man sie am verwichenen Donnerstage hier ankommen sehen. Es scheinet auch, als ob sich der Staat deren nicht so bald versehen hätte, weil er sofort in Ansehung dieser Begebenheit verschiedene Dispositiones geändert hat. Man sieht auch das Verbot bey Lebens-Strafe einigen Umgang oder einige Handlung mit den Engelländern zu haben, als eine gantz besondere Clausul an. Man erwartet nun bald eine Gegen-Kriegs-Ankündigung von Seiten des Großbritannischen Hofs, welche nicht weniger nachdrücklich seyn dürfte, und der König wird von seinem Parlament freye Macht bekommen, seinen Unterthanen eben dergleichen Verbot zu thun; welches ohnfehlbar nicht geschehen seyn würde, wenn Franckreich mit seinem Beyspiel darinn nicht voran gegangen wäre.

Leiden den 10. April.

Die letzten Brief aus Constantinopel vom 25. Febr. bringen mit sich, daß weil Thamas Kuli-Kan sich auf die Ratification des Friedens, welche er durch Hülffe des Bacha von Bagdad mit der Pforte geschlossen, verliesse; so habe er die eroberten Städte Kerkut und Arril wieder verlassen, alle türckische Gefangene in Freyheit gesetzet, und sich auf den Grentzen bey Mercat Ali gelagert, schiene auch in Willens zu seyn, mit seiner gantzen Armee nach Persien zurück zu marschiren. Unterdessen hätte der Sultan den Tractat noch nicht ratificiret, sondern auf den Monath Martii einen grossen Divan convociret, in welchen die beyden Articul den Schach Sefi oder Rade zu abandoniren, und dem Sophi von Persien zu verstatten, die Pilgrims unter einem von ihm dependirenden Chef nach Mecca zu schicken, vielen Streit und Objectiones verursachen dürfften.

Bey dem privil. Verleger dieser Zeitung Johann Jacob Korn ist zu haben:
Der Pilgrim, eine moralische Schrift im Jahr 1743. wöchentl. herausgegeben 104 Stücke nebst vollständigem Register 8vo, 2 Fl. 10 Sgr.
Der neuen Europäischen Fama 106 Theil, 3 Sgr.

* (o) * 23

Schlesische Privilegirte Staats- Kriegs- und Friedens-Zeitungen,

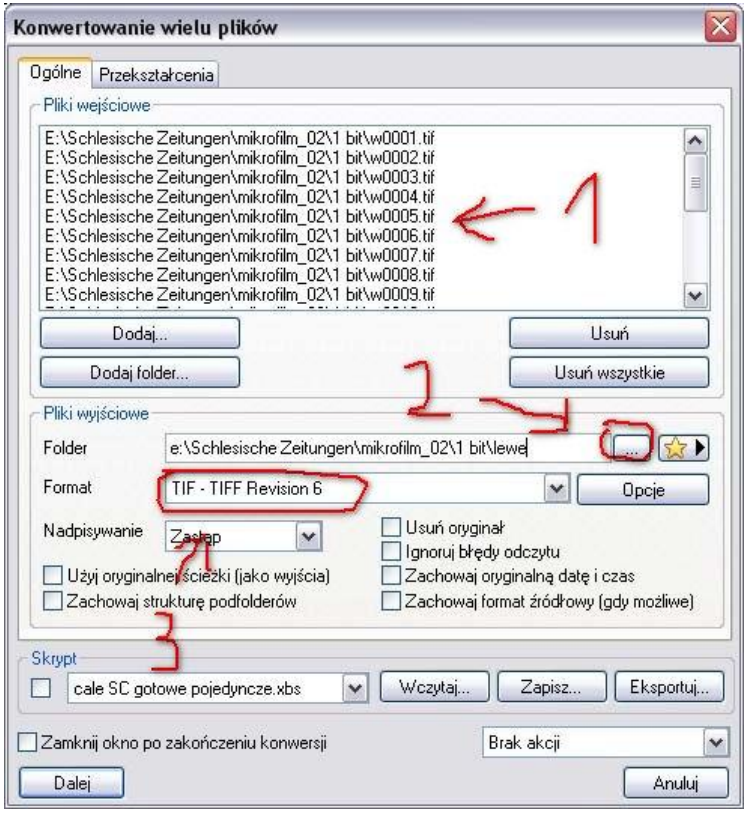
Anno 1744. Mittwochs den 22. April. No. 48.

Berlin den 16 April.

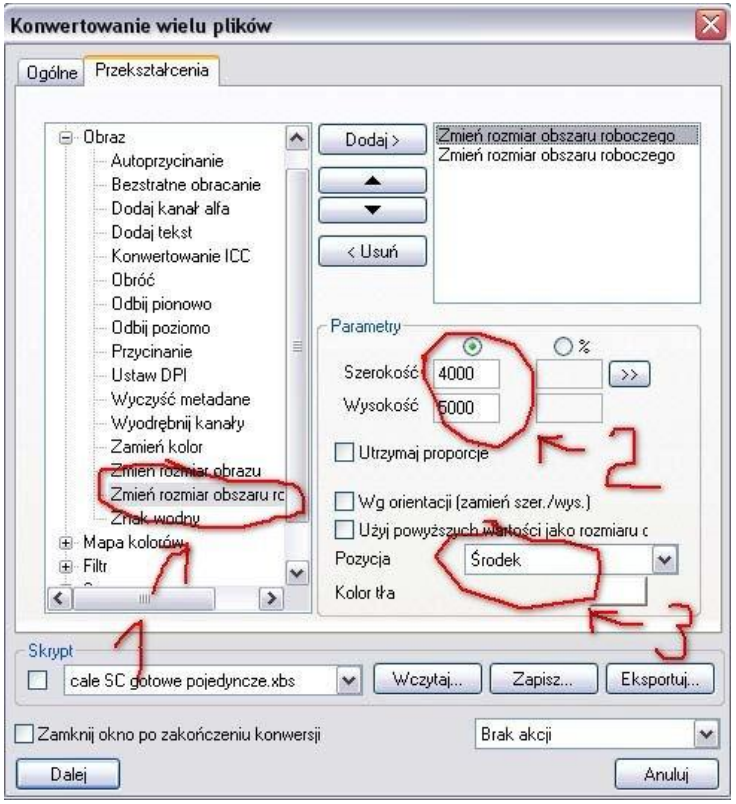
Ihro Majest. die regierende Königin, erhoben sich vorgestern mit Ihro Königl. Hoheiten, den Printzessinnen Ulrica und Amalia, desgleichen den Königl. und allen anwesenden Printzen und Printzessinnen, nebst den vornehmsten Herrn und Dames des Hofes, zu Sr. Maj. dem Könige nach Charlottenburg, allwo Höchstgedachte Se. Maj. Denenselben Mittags und Abends ein grosses Festin gaben, bey welchem die sämtlichen Dames in Amazonen-Habit erschienen. Dieses prächtige Festin wurde durch einen grossen Ball geendiget. Gestern Vormittags kamen Se. Maj. von erwehntem Charlottenburg anhero zurück, und ertheilten kurtz darauf dem Königl. Ungarisch. und Böhmisch. wircklich Geheimten Rathe und Conferentz-Minister, auch Gevollmächtigten am hiesigen Hofe, Herrn Grafen Ursin von Rosenberg, die erste Audientz, worauf Sie des Abends wieder nach Charlottenburg reiseten. Der bisherige Königl. General-Adjutant und Gouverneur Sr. Königl. Hoheit des Printzen Heinrichs, Herr Obriste von Stille, ist von Sr. Maj zu Dero General-Major von der Cavallerie, mit Conferirung eines im Fürstenthum Halberstadt stehenden Cürassier-Regiments, allergnädigst erklärct worden. Se. Excell. der Russisch Kayserl. Ober-Hof-Marschall und ernennte Gevollmächtigte am hiesigen Hofe, Herr Graf von Bestuchef, trafen gestern allhier ein. Wegen der Refugirten ist allhier folgendes Königl Edict publiciret: Wir Friderich, von GOttes Gnaden König in Preussen, ꝛc. Nachdem Uns allerunterthänigst vorgetragen wor-

Przygotowanie plików źródłowych

Do uzyskania takiego efektu wykorzystamy konwerter plików XnView



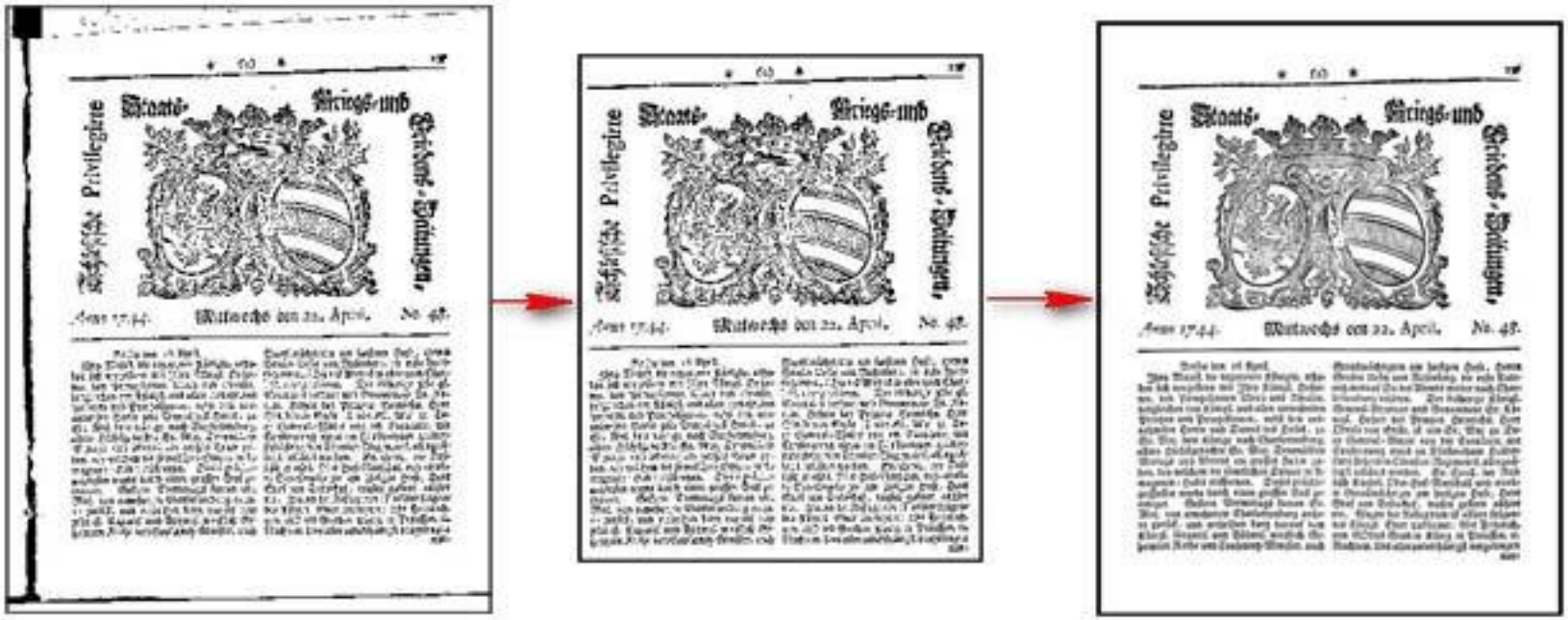
Wczytujemy pliki (1), ustawiamy lokalizację w której mają być zapisywane pliki wynikowe (2), podajemy format zapisu (3).



Zmieniamy rozmiar obszaru roboczego (1), ustalamy szerokość i wysokość (2), ustalamy krawędź przycięcia (3).

Przygotowanie plików źródłowych

Po tych operacjach przystępujemy do kadrowania pojedynczych stron, czyli wracamy do 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.



Rozpoznanie tekstu

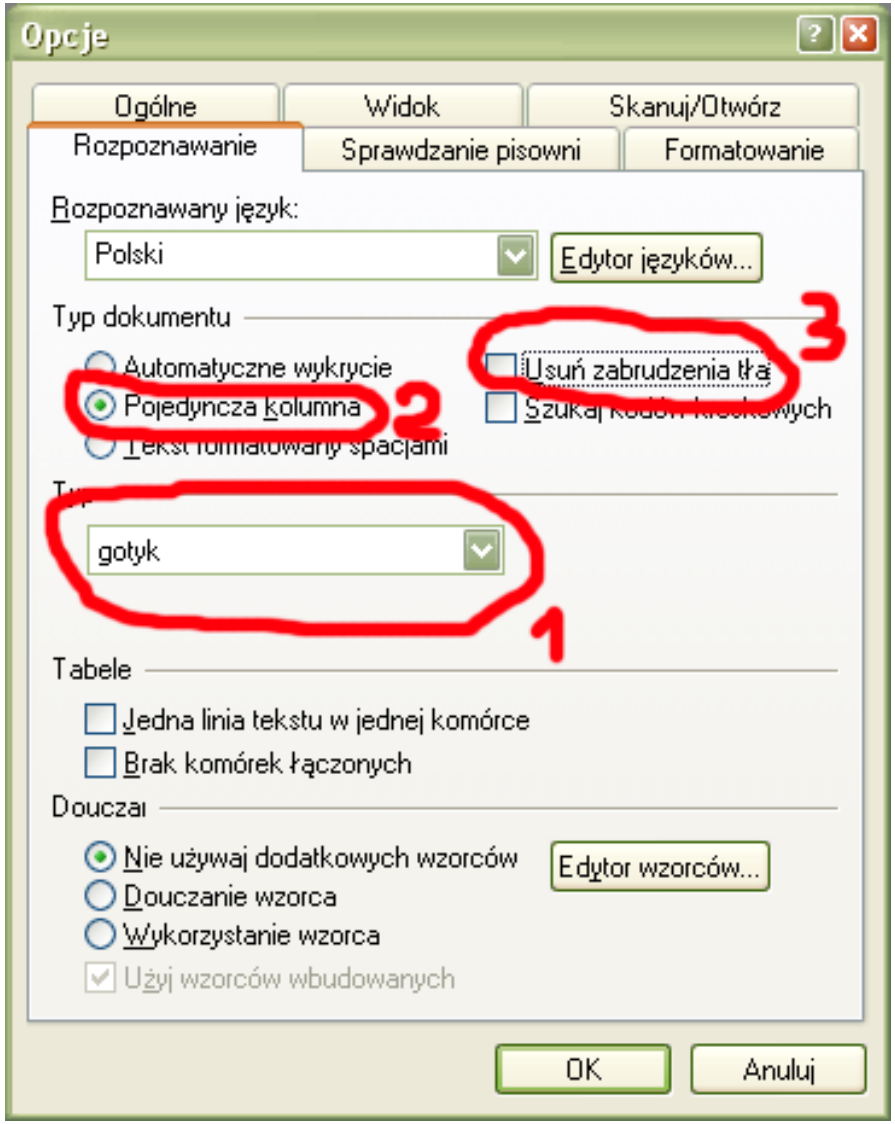
Rozpoznanie tekstu drukowanego czcionką gotycką jest procesem dosyć kosztownym ze względu na sposób licencjonowania oprogramowania wykorzystywanego do obróbki OCR - FineReader XIX. Producent określa ile stron można przetworzyć w ramach jednej licencji i w związku z tym należy zadbać o to, aby rozpoznawania 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.

Rozpoznanie tekstu

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:

1. Typ druku – gotyk
2. Typ dokumentu - pojedyncza kolumna
3. Typ dokumentu - usuń zabrudzenia tła (wyłączone)

Rozpoznanie tekstu



Rozpoznanie tekstu

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 go w czasie najmniejszego obciążenia sprzętu np. w godzinach nocnych.

W polskich bibliotekach cyfrowych najpopularniejszym formatem prezentowania publikacji cyfrowych jest format DjVu (rzadziej PDF).

Jaki format wybrać do prezentacji czasopism?

Pliki prezentacyjne w formacie DjVu

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

1. **Document Express Enterprise** - http://www.djvu.com.pl/de_family.php

2. **Serwis any2djvu** - http://any2djvu.djvuzone.org

2. **Djvudigital** - http://djvu.sourceforge.net/doc/man/djvudigital.html

3. **Pdf2djvu** - http://code.google.com/p/pdf2djvu/

Zgodnie z dostępnym w sieci porównaniem http://code.google.com/p/pdf2djvu/wiki/DjVuDigital na chwilę obecną, pdf2djvu wydaje się być najkorzystniejszym rozwiązaniem do zrealizowania celów postawionych przy digitalizacji czasopism drukowanych gotykiem.

Pliki prezentacyjne w formacie DjVu

Najważniejsze zalety tego rozwiązania to:

1. do tworzonego dokumentu dołączany jest niewidoczny tekst oraz metadane,

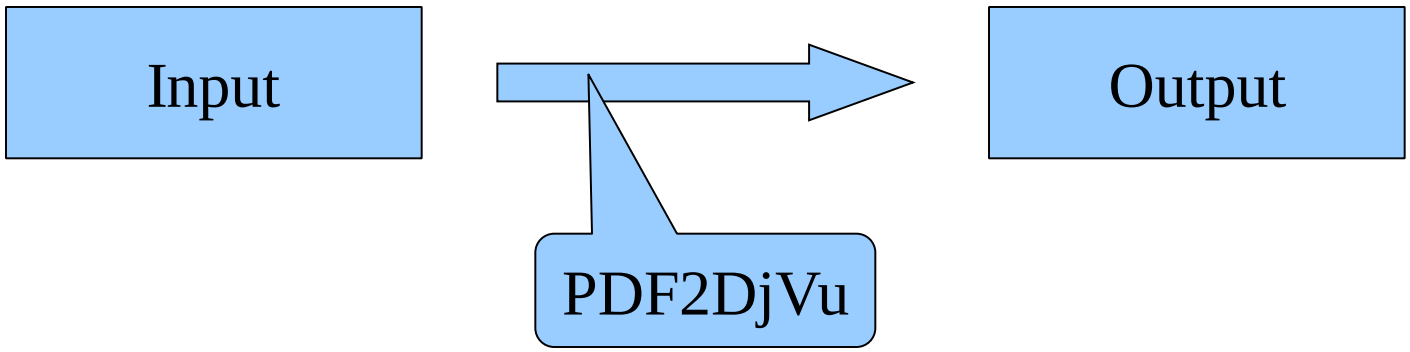
2. duże możliwości wyboru kompresji grafiki,

3. do działania nie wymaga komercyjnego oprogramowania,

4. Dostęp do obszernej dokumentacji autorstwa Jakuba Wilka
http://students.mimuw.edu.pl/~jw209508/papers/thesis/thesis.pdf

Pliki prezentacyjne w formacie DjVu

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 600dpi, pliki scalone, wyłączony antyaliasing). Wyniki jego pracy zapisują się w folderze Output.



Pliki prezentacyjne w formacie DjVu

Do zautomatyzowania pracy przy tworzeniu publikacji DjVu wykorzystywany jest skrypt jazdaDjVu.bat, którego zadaniem jest:

1. ustawianie koloru nagłówka i stopki w plikach wygenerowanych przez program PDF2DjVu,

2. stworzenie miniaturek,

3. rozdzielenie scalonych plików i przekopiowanie nowo powstałych do osobnych katalogów,

4. dołączenie do katalogów z rozdzielonymi plikami, plików opisujących publikację (publication.properties, directory.rdf).

Pliki prezentacyjne w formacie PDF

Do zautomatyzowania pracy przy tworzeniu publikacji PDF wykorzystywany jest skrypt jazdaPDF.bat, którego zadaniem jest:

1. przeniesienie otrzymanych z FineReadera plików PDF do katalogów o nazwach plików,

2. zmiana nazw plików w katalogach na directory.pdf,

3. dołączenie do katalogów z plikami directory.pdf, plików opisujących publikację (publication.properties, directory.rdf).

Prezentacja czasopism – PDF czy DjVu

W przypadku czasopism, które prezentowane są w postaci pojedynczych numerów składających się z kilku do kilkunastu stron warto rozważyć ich prezentacje w formacie PDF.

Argumenty przemawiające na korzyść formatu PDF:

1. wielkość pliku PDF w przypadku pojedynczych numerów czasopisma oscyluje wokół 1 MB, co nie jest obecnie problemem przy prezentowaniu treści w internecie,

2. PDF jest bardziej popularny od DjVu,

3. PDF lepiej się indeksuje w wyszukiwarkach internetowych,

4. krótszy czas przygotowania publikacji w formacie PDF.

Publikacja w bibliotece cyfrowej

Publikowanie w bibliotece cyfrowej dużej liczby 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**. Gotowa struktura publikacji jest wynikiem działania omówionych wcześniej skryptów jazdaDjVu.bat lub jazdaPDF.bat.

directory.rdf

```
<?xml version="1.0"?>
<rdf:RDF
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:dlibra_avs="http://www.dlibra.psnc.pl/">
  <rdf:Description>
    <dlibra_avs:Title xml:lang="pl">Schlesische Privilegirte Staats- Kriegs-
und Friedens-Zeitung 1744-12-02 [Jg.3] Nr 143</dlibra_avs:Title>
    <dlibra_avs:Date xml:lang="pl">1744-12-02</dlibra_avs:Date>
  </rdf:Description>
</rdf:RDF>
```

publication.properties

publication.published=true
publication.collections=
publication.destination.parentPublicationId=29103
publication.name=Schlesische Privilegirte Staats- Kriegs- und Friedens-Zeitung 1744-12-02 [Jg.3] Nr 143
publication.destination.directoryId=22
publication.notes=
publication.mainFile=directory.pdf
publication.secured=false
publication.actorsRights.public=
publication.metadataFile=directory.rdf

Podsumowanie

Zaprezentowany 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 na organizowanie 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.

Dziękuję za uwagę i zapraszam do dyskusji

Biblioteka 2.0 - http://forum.biblioteka20.pl/

Forum dLibra - http://dlibra.psnc.pl/forum/

Digitalizacja.pl - http://www.digitalizacja.pl/

Tomasz Kalota

www.Tomasz.Kalota.pl