DOI: http://dx.doi.org/10.12775/SE.2016.020

Małgorzata Kowalska

Transkrypcja tekstów w środowisku elektronicznym

Przegląd wybranych narzędzi

Uniwersytet Mikołaja Kopernika w Toruniu, kontakt: koma@umk.pl

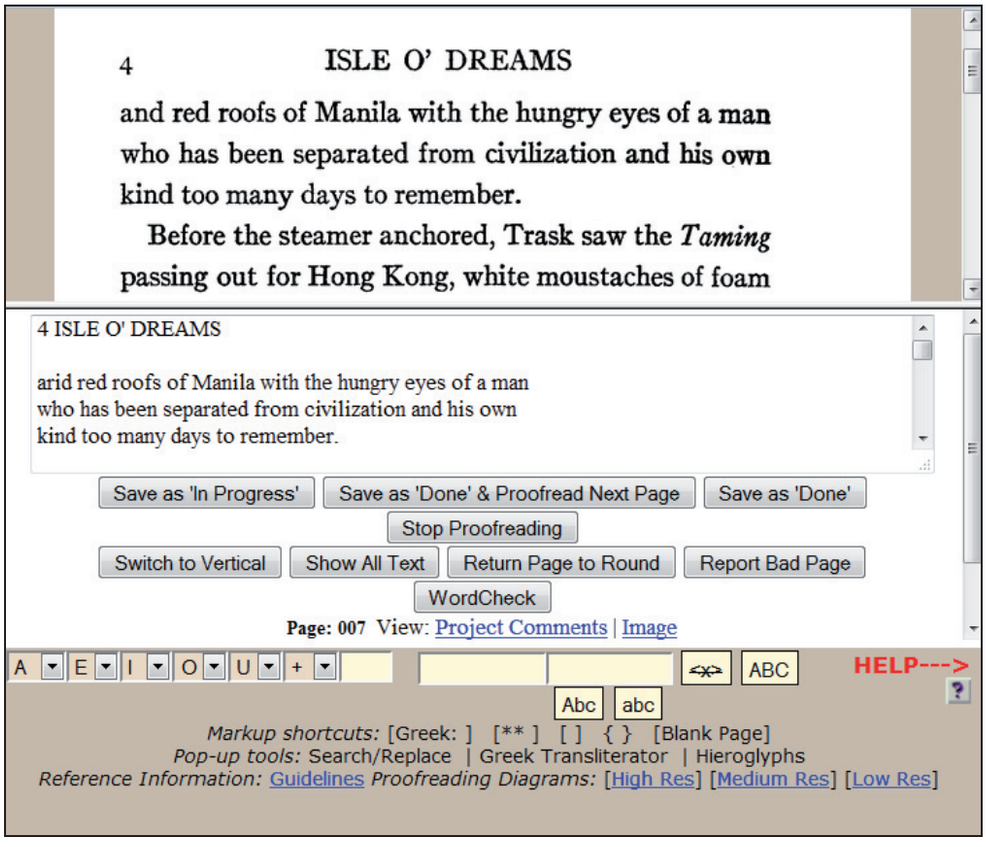
Kiedy w latach siedemdziesiątych i osiemdziesiątych XX wieku zaczęły powstawać pierwsze kolekcje cyfrowe (1971 rok – Project Gutenberg, 1987 rok – Perseus Digital Library), ich twórcom towarzyszyła przede wszystkim troska o zachowanie dziedzictwa dokumentalnego dla przyszłych pokoleń i chęć zapewnienia nielimitowanego dostępu do niego[1]. Nic więc dziwnego, że następstwem tak pojmowanej misji była konwersja cyfrowa pozycji szczególnie cennych z historycznego, kulturowego i naukowego punktu widzenia, trudno osiągalnych na rynku wydawniczym, zaczytanych i cieszących się dużą popularnością wśród czytelników, a przy tym z uregulowanym statusem autorsko-prawnym (dzieła z domeny publicznej)[2]. O ile w pionierskich projektach dygitalizacyjnych udostępnianie klasyki literackiej i dziedzictwa historycznego oraz zapewnianie nieograniczonego dostępu do zasobów stanowiło istotny wyznacznik sukcesu wielu projektów, a jednocześnie znaczący determinant satysfakcji ich użytkowników, o tyle w projektach dojrzałych założenia te okazały się niewystarczające. Stale rosnące wymagania użytkowników formułowane wobec jakości i funkcjonalności zasobów cyfrowych sprawiły bowiem, że głównymi wyznacznikami powodzenia projektów dygitalizacyjnych rozwijanych po przełomie milenijnym stały się przede wszystkim użyteczność i zaspokajanie potrzeb ściśle określonych grup użytkowników[3].

Dziś potencjalny użytkownik materiałów zdygitalizowanych to zazwyczaj internauta przyzwyczajony do intuicyjnego posługiwania się wyszukiwarkami pełnotekstowymi, któremu trudno wyobrazić sobie pracę z dokumentem mającym wyłącznie postać cyfrowego obrazu stron i dla którego poza wysoką jakością kopii

Sztuka Edycji 2/2016
ISSN 2084-7963 (print)
ISSN 2391-7903 (online)
s. 65–74

cyfrowej, jej funkcjonalnością czy multi- i hipermedialnością[4] liczy się także, a może przede wszystkim, możliwość pełnotekstowego przeszukiwania dokumentu (według fraz i wyrażeń, z uwzględnieniem wielkich i małych liter, wykorzystaniem operatorów logicznych) i jego automatycznej analizy. Sprostanie tym wymaganiom wymusza na instytucjach decydujących się na konwersję cyfrową swoistą redefinicję filozofii dygitalizacji, a co za tym idzie odchodzenie od konwersji dokumentów wyłącznie do formatów graficznych na rzecz udostępniania materiałów w formatach wspierających przechowywanie informacji tekstowych (np. PDF czy DjVu). Podążając w tym kierunku, część instytucji dygitalizujących podejmuje proces konwersji cyfrowej, wdrażając do niego od początku techniki optycznego rozpoznawania pisma (Optical Character Recognition, dalej: OCR). Inne (te, których nie stać na zakup komercyjnego oprogramowania, bądź te, które decydują się na redygitalizację swoich zasobów), w celu osiągnięcia podobnego efektu, wykonują transkrypcję tekstów już zeskanowanych, korzystając ze specjalnych programów i narzędzi elektronicznych oraz nierzadko angażując do tego procesu użytkowników sieci (*crowdsourcing*)[5].

Dobrym przykładem udanej inicjatywy transkrypcyjnej jest platforma Distributed Proofreaders, Project Gutenberg: Creation of Ebooks[6], uruchomiona w 2002 roku przez firmę Distributed Proofreaders. Głównym zadaniem tej powstałej w 2000 roku firmy miało być wspieranie dygitalizacji książek z domeny publicznej przeznaczonych dla zasobów Project Gutenberg. Z czasem jej działalność przybrała jednak takie rozmiary, że firma stała się głównym dostawcą treści do tego projektu. Platforma Distributed Proofreaders stwarza wszystkim użytkownikom sieci możliwość konwersji tekstów zdygitalizowanych. Każda zeskanowana książka jest dzielona na pojedyncze strony, co sprawia, że jej korekty może jednocześnie dokonywać wielu użytkowników. Na jednym ekranie komputera prezentowane są wolontariuszom strony zeskanowane oraz zapisane w postaci tekstu powstałego z wykorzystaniem techniki OCR. Dzięki temu istnieje możliwość łatwego porównywania tekstów, ich korekty i „pozostawiania" w sieci w tym samym miejscu (przesyłanie i udostępnienie na tej samej stronie WWW). Kolejny użytkownik, widząc efekty pracy swojego poprzednika, może je korygować, jeśli zachodzi taka potrzeba. Książka przechodzi proces forma-

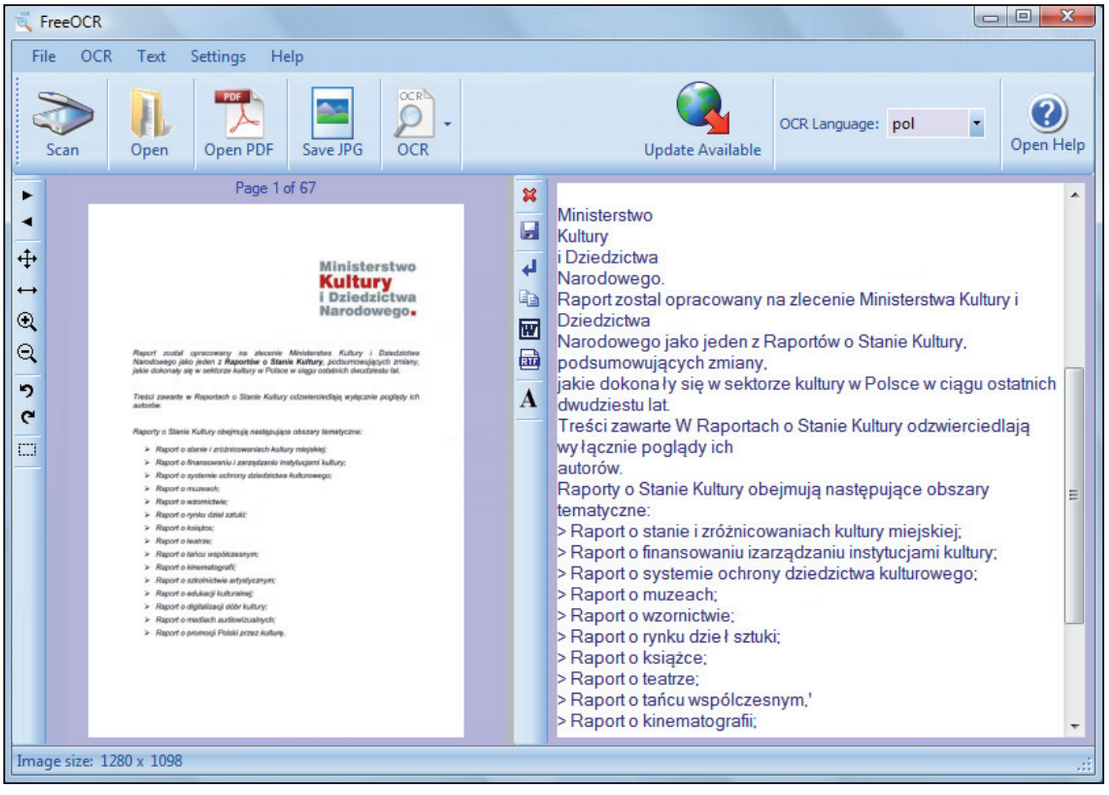


Rys. 1. Distributed Proofreaders – standardowy interfejs do korekty tekstu
Źródło: *Distributed Proofreaders*, online (dostęp: 23.05.2014), http://www.pgdp.net/d/walkthrough/04_Proof.htm.

towania dwukrotnie, lecz z wykorzystaniem tego samego interfejsu. Gdy wszystkie strony przejdą procedurę korekty, za pomocą postprocesora (specjalny typ oprogramowania) są przekształcane w e-booki, które następnie wysyła się do archiwum Project Gutenberg bądź udostępnia w taki sposób, by użytkownicy mogli robić dodatkowe uwagi/notatki, gdy zauważą błędy (*smooth reading*)[7]. Inicjatywa wzbudza ogromne zainteresowanie internautów, o czym świadczą jej efekty: 33 521 – projektów ukończonych, 31 168 – projektów w trakcie korekty, 28 148 – projektów po wstępnym opracowaniu przed zatwierdzeniem, 27 747 – projektów włączonych do Project Gutenberg (stan na 23 maja 2014 roku). Zaledwie w przeciągu jednego tygodnia statystyki odnotowują aktywność około tysiąca wolontariuszy, którzy dokonują korekty ponad czterdziestu książek[8].

Funkcje podobne do tych dostępnych w ramach platformy Distributed Proofreaders oferują również specjalne narzędzia do elektronicznej edycji tekstu. W tym zakresie istnieje szereg rozwiązań zarówno komercyjnych, jak i darmowych. Ze względu na powszechną dostępność, szeroki wachlarz funkcji oraz intuicyjność obsługi warto zwrócić szczególną uwagę na ostatnie z wymienionych narzędzi, a zwłaszcza proste edytory do konwertowania pisma widocznego w plikach graficznych (FreeOCR, GenScriber, Trascript) oraz aplikacje bardziej złożone, umożliwiające całościową konwersję dokumentów (T-Pen, Wirtualne Laboratorium Transkrypcji, DigitLab).

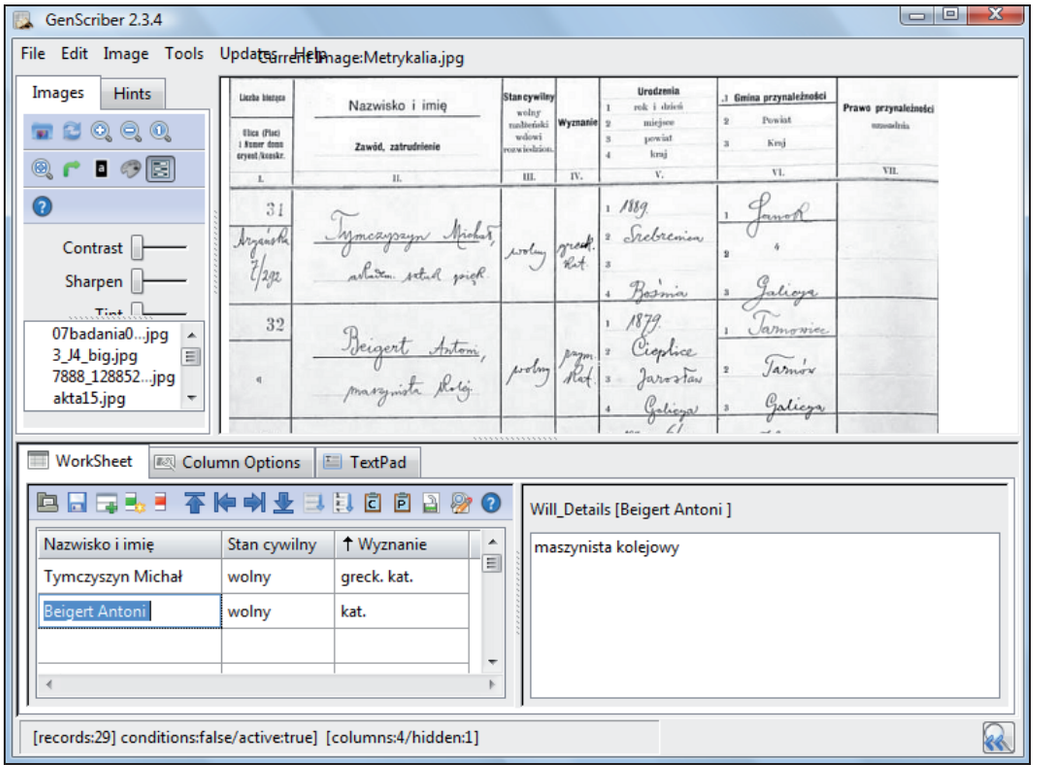
Pierwszą grupę narzędzi otwiera program FreeOCR[9]. Jest to prosty edytor przeznaczony do konwertowania plików graficznych na edytowalne dokumenty. Program zawiera łatwy w obsłudze anglojęzyczny interfejs i współpracuje z wszystkimi wersjami systemów operacyjnych Windows. Działa na podstawie Tesseract 3.01 – nowoczesnego i darmowego open-sourcowego silnika OCR, udostępnionego przez Google, pozwalającego na konwertowanie tekstu w ponad sześćdziesięciu językach[10]. Źródłem odczytu tekstu dla programu może być zarówno kartka papieru, umieszczona w skanerze, jak i plik graficzny zapisany w pamięci komputera. Program odczytuje pliki w formatach GIF, BMP, TIFF i JPG, przy czym do przetworzenia informacji wymagane jest zdjęcie w rozdzielczości minimum 200 dpi. Nowo powstały plik tekstowy można wyeksportować bezpośrednio do formatów DOC, TXT i RTF. Ponieważ program pozwala również na edycję plików PDF, stanowi doskonałą alternatywę dla komercyjnego produktu



Rys. 2. Interfejs programu FreeOCR (wersja 4.2)
Źródło: opracowanie własne.

ABBYY FineReader. Za jego rekomendacją przemawiają także bezproblemowa instalacja, prostota obsługi oraz szybkość ładowania dokumentów skonwertowanych. Niestety, podobnie jak inne edytory OCR, program nie jest pozbawiony wad. Największą z nich jest brak zachowywania struktury konwertowanego tekstu (por. rys. 2), co pociąga za sobą konieczność jego ręcznego formatowania po zakończeniu konwersji. Inną niedogodnością – występującą również w innych programach OCR – są pojawiające się niekiedy problemy z odczytem polskich znaków diakrytycznych, wielkich liter i ligatur (wyświetlają się jako ciągi znaków). Tę wadę eliminują jednak nakładki na program przygotowane przez internautów, dostępne w sieci. Chociaż program często nie radzi sobie z formułami matematycznymi i znakami umieszczonymi w tabelach, manualna poprawa kilku nierozpoznanych elementów z pewnością będzie wymagać od użytkownika znacznie mniej nakładu pracy i czasu niż ręczne przepisywanie całego tekstu.

Kolejnym programem mogącym znaleźć zastosowanie przede wszystkim w instytucjach podejmujących prace dygitalizacyjne jest aplikacja desktopowa GenScriber[11]. Stanowi ona nieocenioną pomoc w opracowywaniu dokumentów archiwalnych. Program został dostosowany do systemów operacyjnych Linux i Windows, a do jego korzystania nie jest wymagana instalacja (należy go jedynie rozpakować i uruchomić). Interfejs, w języku angielskim, ma formę dużego okna podzielonego na kilka mniejszych okien: w górnym – jest widoczny zeskanowany obraz, w dolnym – transkrybowane dane, które mogą przybierać postać arkusza do transkrypcji (analogicznego w układzie do arkusza kalkulacyjnego – WorkSheet) bądź tekstu (TextPad) (por. rys. 3). Program pozwala kopiować dane genealogiczne z sieci, a także pobierać i wyświetlać obrazy z komputera zapisane w formatach JPG, PNG, TIFF, GIF i PDF. Domyślnym formatem zapisu dokumentów edytowalnych jest CSV. Dużą dogodnością dla użytkownika jest możliwość powiększania i pomniejszania obrazu, zmiany jego kontrastu, ostrości i odcienia (aż do skali szarości). Decydując się na wybór programu, trzeba jednak pamiętać, że nie jest to typowy automatyczny konwerter obrazów do plików tekstowych, a raczej program do przepisywania rękopisów i wspomagający indeksację, notowanie i tłumaczenie. Jego instalacja nie powinna sprawić użytkownikom żadnego problemu. Pewnych trudności może przysporzyć praca z arkuszem transkrypcji, zachodzi tu bowiem konieczność samodzielnej redefinicji kolumn. Z pomocą w tym zakresie przychodzą jednak gotowe szablony standardowych dokumentów archiwalnych, w które GenScriber został wyposażony. Program jest godny

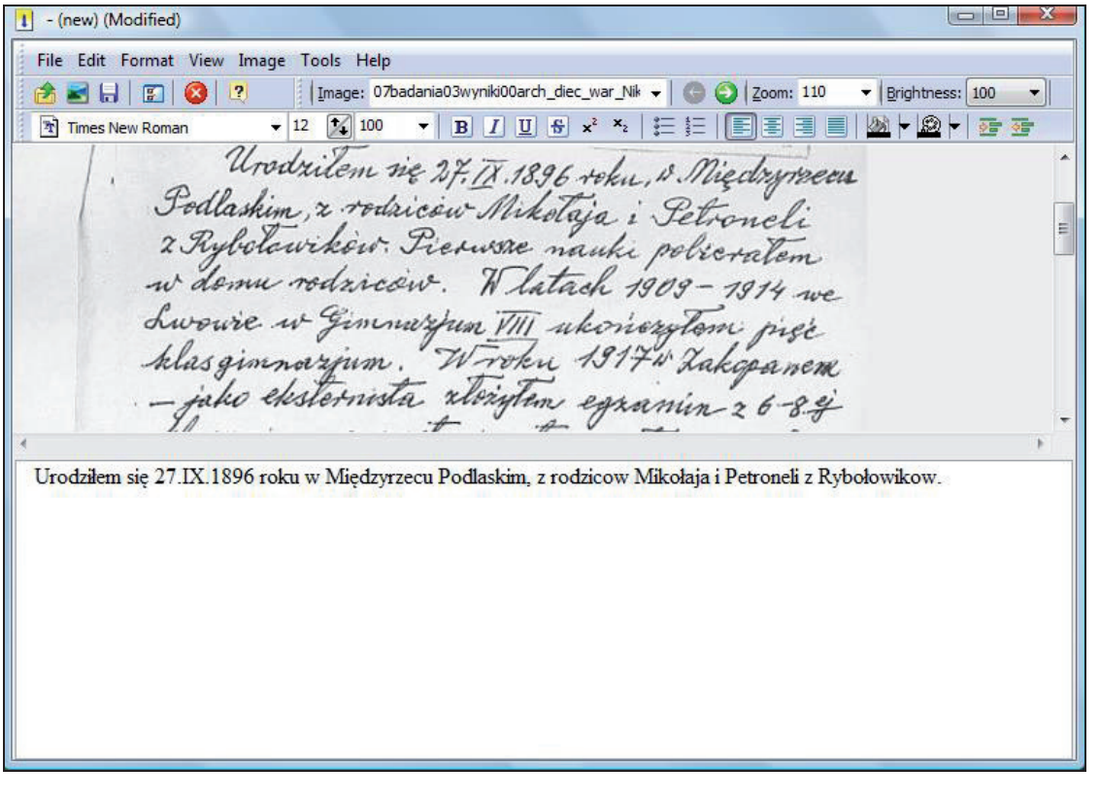


Rys. 3. Interfejs programu GenScriber (wersja 2.3.4, widok arkusza transkrypcji)
Źródło: opracowanie własne.

polecenia przede wszystkim genealogom oraz historykom pracującym z archiwaliami, choć z zastrzeżeniem, że brakuje w nim polskiej wersji językowej interfejsu oraz biblioteki polskich znaków/słownika.

Ostatnim z pierwszej grupy omawianych programów jest edytor Transcript[12]. Podobnie jak jego poprzednik, nie jest to automatyczny konwerter obrazów do postaci tekstowej, lecz narzędzie mające pomóc wszystkim opracowującym rękopisy i zeskanowane dokumenty w ich rozpisywaniu. Program współpracuje z systemem Windows i oferuje interfejs w sześciu wersjach językowych: angielskiej, niemieckiej, francuskiej, holenderskiej, duńskiej i fińskiej. Praca w programie odbywa się z podziałem ekranu na dwie części. W górnej połowie jest wyświetlany obraz cyfrowy, w dolnej – pole edycji tekstu (por. rys. 4). Do okna podglądu (górnego) mogą zostać zaimportowane obrazy w takich formatach, jak JPG, BMP, GIF, PNG i TIFF. Do zapisu edytowanego dokumentu domyślnie stosowany jest format RTF, choć możliwe jest także wyeksportowanie pliku do programu Microsoft Word lub edytora Writer pakietu LibreOffice. W programie można korzystać z większości funkcji znanych z innych edytorów (zmniejszanie, powiększanie obrazu, zmiana nasycenia, ostrości, odcienia). Program jest łatwy w instalacji i niezwykle intuicyjny w obsłudze. Automatycznie zapamiętuje ostatnie miejsce edycji i wraca do tej pozycji po ponownym uruchomieniu programu. Jego wadą jest brak kilku polskich znaków diakrytycznych („ś", „ź", „ć", „ó").

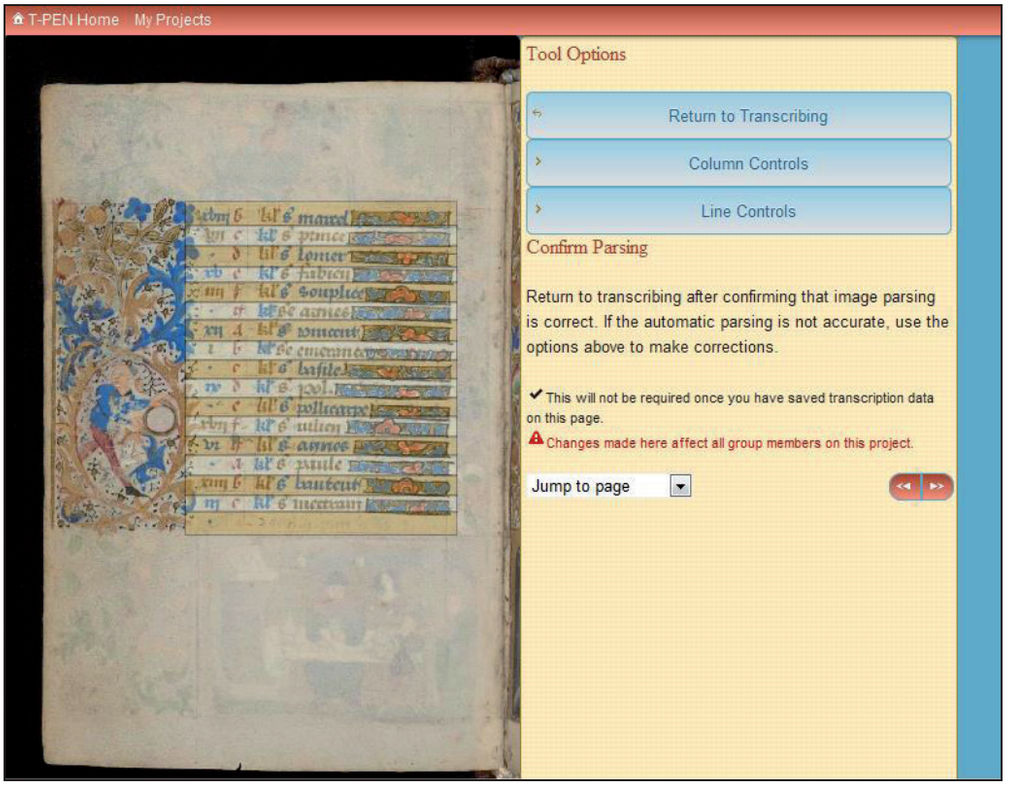
Odrębną kategorię narzędzi umożliwiających transkrypcję stanowią kompleksowe programy, których nadrzędnym celem jest wspomaganie tworzenia pełnotekstowych wersji dokumentów. Pierwszym z tego rodzaju narzędzi jest aplikacja T-PEN (Transcriptio for Paleographical and Editorial Notation), stworzona do odczytu i transkrypcji rękopisów w Center for Digital Theology na Uniwersytecie w Saint-Louis[13]. Aby rozpocząć pracę, należy założyć konto użytkownika na stronie WWW projektu (http://t-pen.org/TPEN/), a następnie, po otrzymaniu linku aktywującego, potwierdzić rejestrację i zalogować się. W systemie można dokonywać transkrypcji własnoręcznie dodanych plików (tu wymagana jest umiejętność tworzenia archiwów ZIP oraz przygotowania serii plików w formacie JPG, które będzie można „załadować" do programu), jak i 4 117 manuskryptów (stan na 23 maja 2014 roku) udostępnionych w projekcie, a pochodzących ze współpracujących repozytoriów. O ile dodawanie i konwersja własnych plików są darmowe, o tyle dostęp do zdygitalizowanych rękopisów należących do poszczególnych instytucji i ich



Rys. 4. Interfejs programu Transcript (wersja 2.4.0.88)
Źródło: opracowanie własne.

transkrypcja wymagają uiszczenia opłaty (zgodnie z umowami nie są własnością T-PEN). Po załączeniu własnego pliku (będzie on widoczny jako plik prywatny i nie znajdzie się w ogólnym katalogu projektów T-PEN) lub wyborze danego manuskryptu z wykazu T-PEN można przystąpić do transkrybowania. Ponieważ aplikacja nie przechowuje obrazów dokumentów w pamięci, każdorazowo dokument jest pobierany i analizowany w czasie rzeczywistym. W pierwszym etapie program określa położenie każdego wersu na stronie, a następnie wyświetla je, oznaczając naprzemiennie kolorami (por. rys. 5). W tym miejscu użytkownik może dokonywać wielu operacji na tekście: usuwać, dodawać i zmieniać szerokość kolumn tekstu oraz wstawiać, łączyć i zmieniać szerokość wersów. Po nadaniu ostatecznego kształtu dokumentowi można rozpocząć jego przepisywanie. W tym celu należy wybrać odpowiedni wers. W efekcie tego zabiegu wyświetli się odrębne okno pozwalające na wpisywanie odczytanego fragmentu (por. rys. 6). T-PEN ma wbudowany zestaw trzynastu narzędzi programistycznych, co pozwala m.in. na dostosowywanie systemu kodowania znaków do potrzeb użytkownika (np. Unicode, UTF-8). Przepisany dokument można wyeksportować do pliku w formatach PDF, RTF i XML. Niewątpliwą zaletą programu jest możliwość samodzielnego wyodrębniania przez użytkownika własnych narzędzi (np. słownik, baza tekstów), które będą widoczne za transkrypcją, podobnie jak słownik abrewiacji. Mimo że aplikacja oferuje dość skromny system znaków specjalnych oraz niewielki wybór słowników skrótów czy językowych, jej twórcy przewidzieli możliwość integracji interfejsu z innymi wykorzystywanymi bądź preferowanymi przez użytkownika narzędziami (wystarczy dodać ich nazwę i URL). Aplikacja T-PEN została pomyślana jako projekt crowdsourcingowy, umożliwiający współpracę wielu osób jednocześnie. Dlatego dzięki tzw. dziennikowi projektu wszystkie zmiany nanoszone przez pojedynczych użytkowników są rejestrowane i odpowiednio oznaczane.

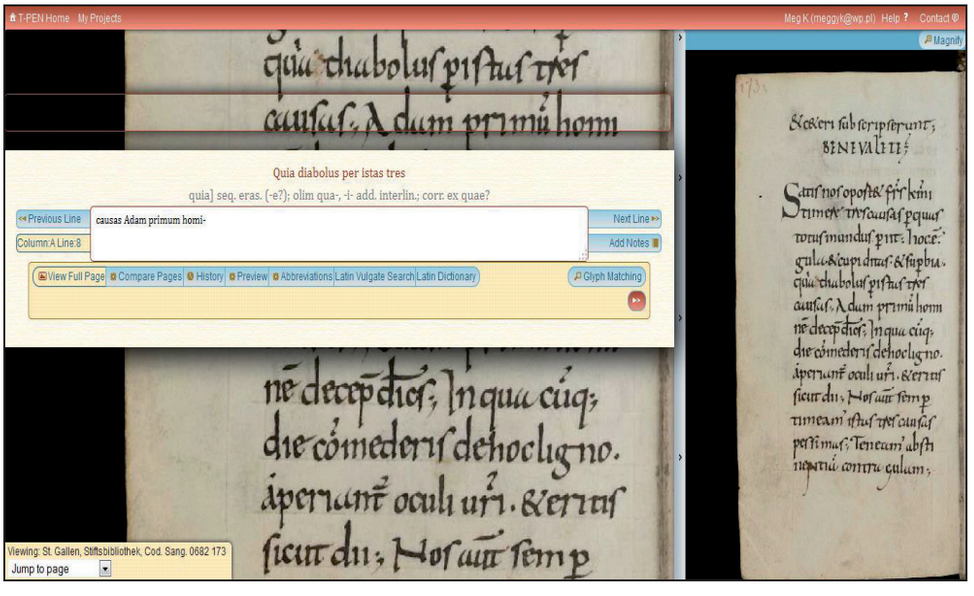
Za inne bardzo obiecujące rozwiązanie należy uznać udostępnione przez Poznańskie Centrum Superkomuterowo-Sieciowe w październiku 2012 roku narzędzie o nazwie Wirtualne Laboratorium Transkrypcji. Aplikacja stanowi część rozbudowanego projektu SYNAT, którego głównym założeniem jest stworzenie uniwersalnej, otwartej, repozytoryjnej platformy hostingowej i komunikacyjnej dla sieciowych zasobów wiedzy dla nauki, edukacji i otwartego społeczeństwa wiedzy. Wirtualne Laboratorium Transkrypcji



Rys. 5. T-PEN (wersja 2.0), widok strony dokumentu z podziałem na wersy

Źródło: *Kalendar – Geneva lat 33* (autor projektu: R. Sanderson), online (dostęp: 23.05.2014), http://t-pen.org/TPEN/transcription.jsp?projectID=3599&p=2583620.

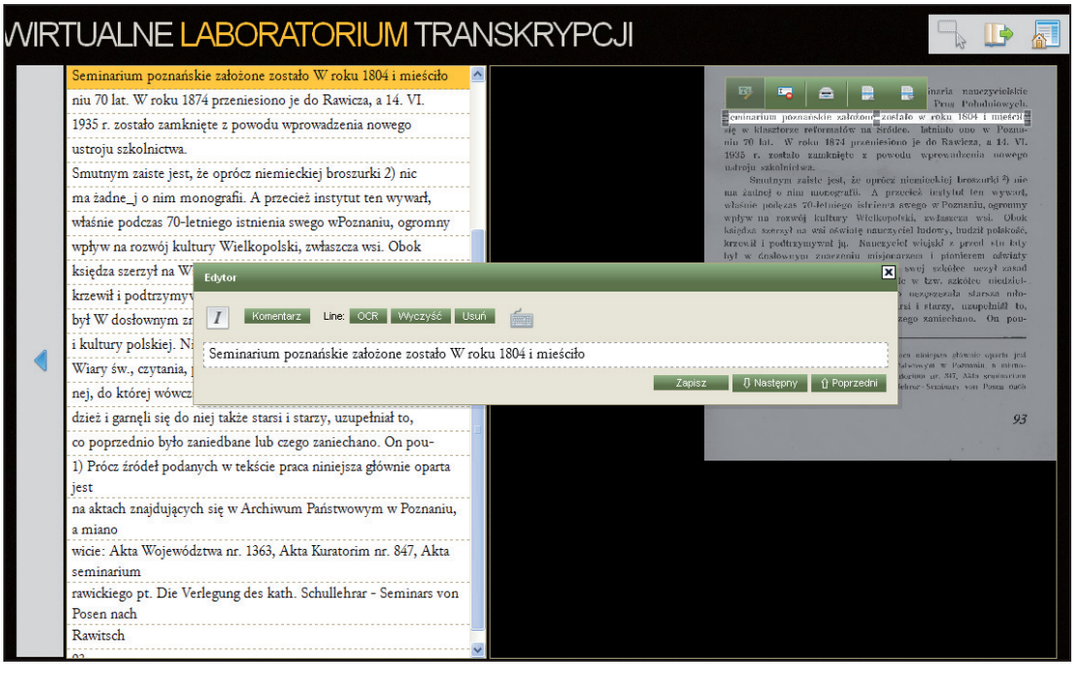
to darmowe narzędzie, udostępnione w wersji testowej, które pomaga tworzyć cyfrowo przeszukiwalne teksty z dokumentów historycznych. Aby rozpocząć pracę[14] z Wirtualnym Laboratorium Transkrypcji na stronie WWW projektu (http://wlt.synat.pcss.pl/wlt-web/index.xhtml), należy założyć konto użytkownika, a następnie – po otrzymaniu linku aktywującego, potwierdzeniu chęci rejestracji i zalogowaniu – we własnym profilu stworzyć nowy projekt, opisując go stosownymi metadanymi (nazwa, autor, tytuł, słowa kluczowe, data publikacji, język/i tekstu, typ tekstu). Po zaakceptowaniu regulaminu i wyborze tytułu projektu można przystąpić do pracy: dodawać pliki, poddawać je rozpoznaniu techniką OCR, transkrybować, załączać istniejące transkrypcje, zarządzać metadanymi. Najważniejszą funkcją Wirtualnego Laboratorium Transkrypcji jest automatyczne rozpoznawanie tekstów w plikach graficznych. Do programu można zaimportować pojedyncze pliki w formatach PNG, GIF, TIFF, JPG i DjVu. Można także załączyć całe archiwum ZIP, zawierające kilka plików w formatach JPG i PNG, lub pobrać dokument w formacie DjVu z pięciu polskich bibliotek cyfrowych znajdujących się w serwisie Federacji Bibliotek Cyfrowych (Wielkopolskiej, Małopolskiej, Dolnośląskiej, Śląskie i Jagiellońskiej Biblioteki Cyfrowej). Po „załadowaniu" pliku na ekranie ukazuje się strona zeskanowanego dokumentu, w obrębie której można dokonać wyboru fragmentu, jaki ma zostać poddany procesowi rozpoznawania znaków. Wybór opcji „Zacznij OCR" inicjuje proces konwersji. Najczęściej trwa ona kilka chwil, a po jej zakończeniu autor projektu pocztą elektroniczną jest informowany o powstaniu pliku wsadowego. Rozpoznany tekst wyświetla się w edytorze transkrypcji w postaci numerowanych wersów. Po kliknięciu w dowolny wers można rozpocząć proces edycji (por. rys. 7). Każdą edytowaną linię tekstu należy zatwierdzić enterem. Dla ułatwienia okno edytora transkrypcji zostało podzielone na dwie części, po lewej stronie są wyświetlane rezultaty procesu rozpoznawania znaków, po prawej – podgląd całej transkrypcji. Dzięki opcjom, takim jak lupa, zoom, przechodzenie do kolejnej strony, wyszukiwanie w tekście, zaznaczanie fragmentu, wygenerowany tekst można poddawać dodatkowym operacjom. Poza możliwością wyświetlania wyników procesu OCR w edytorze transkrypcji nowo powstały dokument można zapisać w postaci ciągłego tekstu w pamięci komputera (format ePUB). Dużą zaletą programu jest możliwość pracy z drukami wielokolumnowymi. Za zaletę należy uznać także rejestrację autorów wprowadzanych w tekście zmian, co pozwala na śledzenie kolejnych etapów transkrypcji. Wśród minusów trzeba z kolei wymienić brak możliwości importowania plików w formacie PDF i eksportowania ich do takiej postaci oraz brak odpowiednich słowników ułatwiających pracę osobom posługującym się dokumentami w języku łacińskim czy złożonymi dawną polszczyzną. Mimo że Wirtualne Laboratorium Transkrypcji, w odróżnieniu od omówionego już programu T-PEN, nie radzi sobie również z odczytem rękopisów, jest to narzędzie bardzo przydatne



Rys. 6. T-PEN (wersja 2.0) – interfejs do transkrypcji
Źródło: *Homiliae (Irish-influenced?)* (autor projektu: T. O'Sullivan), online (dostęp: 23.05.2014), http://t-pen.org/TPEN/transcription.jsp?projectID=64.

podczas pracy z tekstem, pozwala bowiem na dużą oszczędność czasu. Przykładowo, opracowanie, a więc przeprowadzenie procesów skanowania, normalizacji skanów, utworzenia projektu, OCR-owania, korekty rezultatów i eksportu do pliku, osiemnastostronicowej broszury trwa około 1,14 godziny, a przy „załadowaniu" tekstu z biblioteki cyfrowej – nie przekracza godziny zegarowej[15].

Narzędziem, a raczej systemem operacyjnym, rekomendowanym w procesie cyfryzacji zbiorów może być także DigitLab. System, podobnie jak Wirtualne Laboratorium Transkrypcji, powstał w poznańskim centrum Superkomputerowo-Sieciowym i został udostępniony w 2012 roku. Działa na podstawie opensourcowego oprogramowania Linux Ubuntu. Można go pobrać w formie obrazu ISO i wypróbować, nagrywając na pendrive'a lub płytę DVD bez konieczności instalacji na komputerze. Domyślnym językiem systemu jest język angielski. Dodatkowo zainstalowano języki chorwacki, serbski, grecki, albański, turecki oraz polski. System składa się z aż trzydziestu jeden programów narzędziowych, które mogą być przydatne w procesie cyfryzacji zasobów, w tym m.in. narzędzia umożliwiającego obróbkę wyników skanowania, narzędzia ułatwiającego przygotowanie plików w formatach DjVu i PDF, skryptu pozwalającego na tworzenie zoomifów oraz silnika OCR. Do systemu zostały także dołączone trzy przykładowe biblioteki cyfrowe, stworzone na podstawie oprogramowania DSpace, GreenSrone i Libra. Pracę nad dokumentem cyfrowym[16] można rozpocząć albo od procesu skanowania oryginału, albo od przesłania do systemu pliku zawierającego graficzny obraz dokumentu (zeskanowany lub sfotografowany). Następnie skan/plik graficzny należy poddać obróbce technicznej (zmiana orientacji, podział na strony, wyrównanie, zaznaczenie marginesów, eliminacja zanieczyszczeń, por. rys. 8). Podczas wstępnej obróbki trzeba także zaznaczyć te fragmenty czy pola dokumentu, które będą poddane procesowi rozpoznawania pisma. Po zakończeniu technicznego opracowania plik wynikowy należy zapisać (domyślnie zapis następuje w tym samym miejscu, z którego wczytywany jest skan). W miarę potrzeb przygotowany plik wsadowy można poddać optycznemu rozpoznawaniu znaków. By uruchomić ten proces, po wczytaniu pliku z opcji paska narzędzi wybiera się funkcję „OCR", a następnie wskazuje język tekstu na obrazie i zakres stron, które mają zostać poddane konwersji. Po ukończeniu procesu rozpoznawania w zakładce „OCR Output" zostaje wyświetlony jego wynik. Ponieważ każde słowo jest wyświetlane w osobnym polu, by dokonać korekty rezultatów OCR-owania, należy wybrać dowolne z pól. Ostatnim krokiem jest stworzenie pliku końcowego z dokumentem cyfrowym. W tym celu, podobnie jak w innych programach,



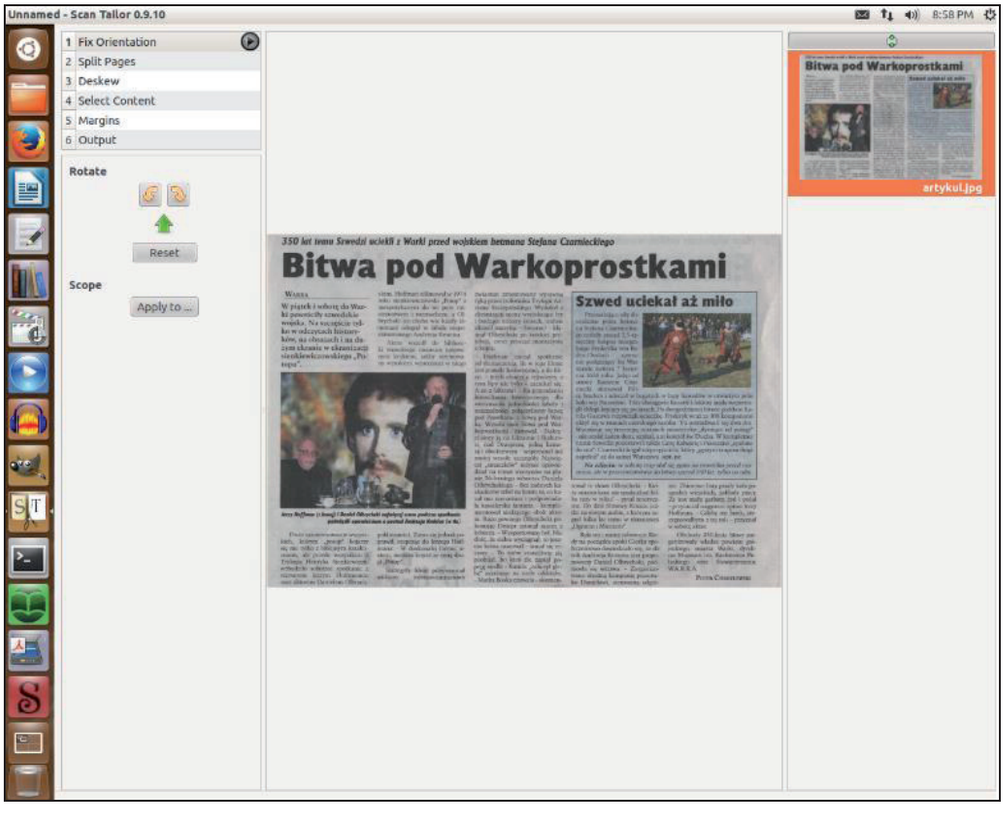
Rys. 7. Wirtualne Laboratorium Transkrypcji – interfejs umożliwiający korektę tekstu

Źródło: *Strona projektu „Kronika Miasta Poznania – 1939 R. 17 Nr 2"* (online), *Wirtualne Laboratorium Transkrypcji* (dostęp: 23.05.2014), http://wlt.synat.pcss.pl/wlt-web/project.xhtml?project=81.

z menu „File" trzeba wybrać opcję „Save" (format DjVu lub PDF). Niewątpliwą zaletą systemu DigitLab jest jego modularność. Dzięki szerokiemu wachlarzowi zintegrowanych programów użytkownik zyskuje zestaw narzędzi pozwalających na kompleksowe przygotowanie dokumentu cyfrowego na wszystkich jego etapach, od skanowania po prezentację w sieci. System pozwala na obróbkę nie tylko materiałów tekstowych, ale także audio i wideo, wymagających niekiedy opracowania fragment po fragmencie. Podobnie jak Wirtualne Laboratorium Transkrypcji, stwarza również szansę na pracę z dokumentami wielkoformatowymi. Wydaje się dobrym rozwiązaniem zarówno do cyfryzacji domowych archiwów, jak i dużych zasobów dokumentów historycznych. Niestety, ze względu na integrację różnych rodzajów narzędzi wymaga znajomości wielu środowisk programistycznych przynajmniej na poziomie średniozaawansowanym.

Zaprezentowane wyżej rozwiązania technologiczne przeznaczone do wspomagania procesów cyfryzacji, a zwłaszcza transkrypcji tekstów, mogą być wykorzystywane przez instytucje GLAM (Galleries, Libraries, Archives, Museums) oraz każdego potencjalnego użytkownika sieci na różne sposoby i do różnych celów. Możliwość ich zastosowania do poprawy jakości dygitalizatów daje niebywałą szansę na wzbogacenie istniejących zasobów internetu dokumentami w pełni przeszukiwalnymi. Wydaje się to szczególnie istotne z uwagi na małą widoczność polskich zasobów cyfrowych w internecie, a także sugerowaną w narodowym programie dygitalizacji konieczności archiwizacji zasobów polskiego internetu[17]. W tym kontekście warte rozważenia jest włączenie do prac konwersyjnych i transkrypcyjnych potencjalnych użytkowników sieci (oczywiście pod warunkiem nadzorowania tych prac przez specjalistów). Jak pokazują doświadczenia australijskich czy amerykańskich instytucji kultury i dziedzictwa, zaangażowanie społeczności wirtualnych w poprawę jakości danych przyczynia się bowiem nie tylko do zwiększania wartości przechowywanych danych i podnoszenia relewantności wyszukiwania, ale także do wzrostu prestiżu tych instytucji i ich znaczenia społecznego, a co ważniejsze – budowy poczucia wspólnoty publicznej i odpowiedzialności za dziedzictwo kulturowe[18].

Key Words: digital documents, digital collections, tools for edition and text analysis, transcription, special programming tools



Rys. 8. *DigitLab* – interfejs programu do obróbki skanów Scan Tailor (wersja 0.9.10)
Źródło: opracowanie własne.

Małgorzata Kowalska

Abstract: Users' requirements formulated in reference to quality and functionality of digital documents are constantly growing. Whereas in the first, pioneering projects of digital collections it was sufficient to have access to digital image of pages of a given document, at present creators of digital documents are expected to fit them with advanced tools of edition and text analysis, as well as ensuring their searchability. A chance to meet those requirements is, on the one hand, scanning documents with the use of optical character recognition and, on the other hand, subjecting the digital texts to the process of transcription using special programming tools. The article presents the possibilities of selected solutions in this area, pointing to possible areas of their use, and outlines the potential advantages and disadvantages of their functionality.

[1] W. M. Kolasa, *Biblioteki cyfrowe na świecie – powstanie i rozwój*, w: *Biblioteki cyfrowe*, pod red. M. Janiak, M. Krakowskiej i M. Próchnickiej, Warszawa 2012, s. 67–70.

[2] M. Kowalska, *Dygitalizacja zbiorów bibliotek polskich*, Warszawa 2007, s. 249–250.

[3] M. Nahotko, *Zasady tworzenia bibliotek cyfrowych*, „Biuletyn EBIB" 2006, nr 4 (74), online (dostęp: 23.05.2014), http://www.ebib.info/2006/74/nahotko.php.

[4] M. Kowalska, *Dygitalizacja zbiorów bibliotek polskich*, s. 42–43.

[5] Na temat transkrypcji tekstów wykonywanej przez społeczności wirtualne zob. eadem, *Wykorzystywanie koncepcji mądrości tłumu w działalności bibliotek*, „Toruńskie Studia Bibliologiczne" 2012, nr 2 (9), s. 99–112.

[6] *Distributed Proofreaders*, online (dostęp: 23.05.2014), http://www.pgdp.net/c/.

[7] Ibidem.

[8] *Distributed Proofreader – Statistics Central*, online (dostęp: 23.05.2014), http://www.pgdp.net/c/stats/stats_central.php.

[9] Program do pobrania pod adresem: http://www.freeocr.net/.

[10] Więcej: *tesseract-ocr*, online (dostęp: 23.05.2014), https://code.google.com/ p/tesseract-ocr/.

[11] Program do pobrania pod adresem: http://genscriber.com/genappsd/.

[12] Program do pobrania pod adresem: http://www.jacobboerema.nl/en/Freeware.htm.

[13] Szczegółowy opis aplikacji: J. Ginther, *A New Tool for Transcription of Digitized Manuscripts*, online (dostęp: 23.05.2014), http://earlymodernonlinebib.wordpress.com/2012/10/22/t-pen-a-new-tool-for-transcription-of-digitized-manuscripts/.

[14] Szczegółowy opis poszczególnych kroków postępowania w: A. Dudczak, B. Wróż, *Wprowadzenie do Wirtualnego Laboratorium Transkrypcji*, online (dostęp: 23.05.2014), https://confluence.man.poznan.pl/community/display/WLT/Wprowadzenie+do+Wirtualnego+Laboratorium+Transkrypcji.

[15] A. Dudczak, *Od skanów do tekstu w kilku prostych krokach i dwóch smakach*, online (dostęp: 23.05.2014), http://lib.psnc.pl/Content/444/adudczak-thatcamp-lublin.pdf.

[16] Szczegółowy opis poszczególnych kroków postępowania w: A. Dudczak, *DigitLab Wiki. Dokumentacja po polsku*, online (dostęp: 23.05.2014), https://confluence.man.poznan.pl/ community/display/DIG/Dokumentacja+po+polsku.

[17] *Raport o digitalizacji dóbr kultury: Program digitalizacji dóbr kultury oraz gromadzenia, przechowywania i udostępniania obiektów cyfrowych w Polsce 2009–2020* (online), Warszawa 2009, s. 67 (dostęp: 23.05.2014), http://www.kongreskultury.pl/library/File/RaportDigitalizacja/Program digitalizacji 2009-2020.pdf.

[18] R. Holley, *Crowdsourcing: How and Why Should Libraries Do It?*, „D-Lib Magazine" 2010, Vol. 16, No. 3/4, online (dostęp: 23.05.2014), http://www.dlib.org/dlib/march10/holley/03holley.print.html.