Korekta OCR – problemy i rozwiązania

Edyta Kotyńska
eTEKA.com.pl

* (o) * 343

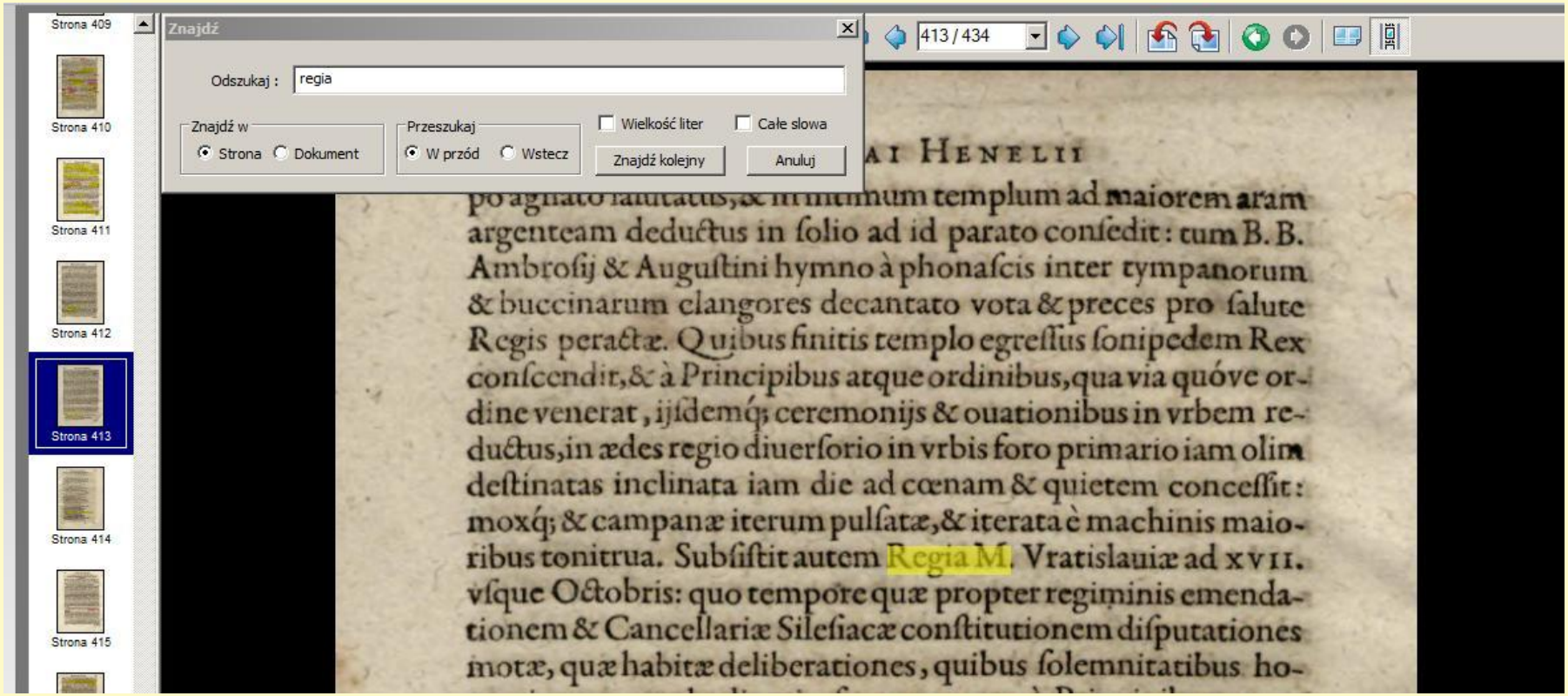
Schlesische Privilegirte Staats- Kriegs- und Friedens-Zeitungen.

Anno 1745. Sonnabends den 3. Jul. No. 77.

Breßlau den 2. Jul.

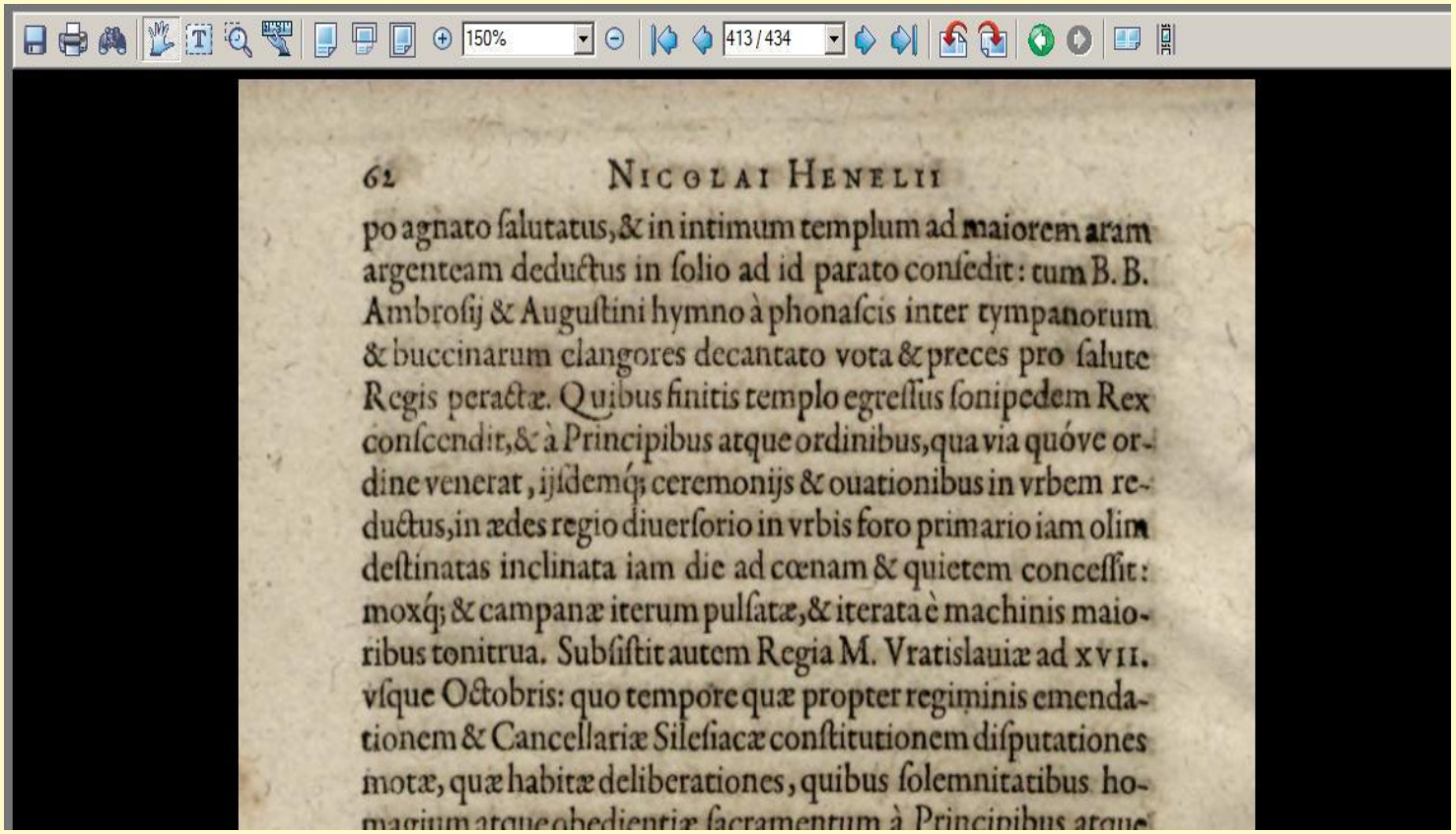
Aus Böhmen enthalten die Nachrichten vom 25. Jun. folgendes: Als wir am 15ten das Lager bey Nowiengeslow oder Böhmisch Neustädtel bezogen, setzten sich die Feinde 15000. Mann starck, theils regulirte, theils Panduren, Husaren und Insurgenten auf den Bergen, Kirchhöfen und in den Gehölzen recht um unser Lager herum, und liessen uns die Zelter in guter Ruhe aufschlagen. Die Husaren formirten die beyden Flügels, die Insurgenten das Centrum, u. hinter ihnen stund die regulaire Mannschafft. Es wurde also unsers Avant-Garde, welche weiter vorwerts rücken solte, etwas verstärckt. Der Herr General von Jeetz commandirte selbige, und der Herr General von Lehwald hatte den lincken und der Herr General-Major von Winterfeld den rechten Flügel; beyde richteten ihren Marsch so ein, daß sie die feindl. Husaren von den Ungarn trenneten und abschnitten. Als nun die regulaire Mannschafft dieses sahe, nahm selbige sogleich die Flucht; während dieser Zeit rückte der Hr. General von Jeetz gegen die Insurgenten, u. delogirte dieselben; sie retirirten sich aber in der Meinung, daß ihre Flanquen noch gedeckt wären, in ein nah gelegen Gehölze, allein auch da wurden sie gar bald vertrieben; hierauf suchten sie ein Dorf zu erreichen, die Unsrigen kamen ihnen aber zuvor, so daß sie sich auf freyem Felde umzingelt sahen. Da sich nun diese Ungarn solchergestalt in die Enge getrieben befanden, gab ihnen die Desperation ein, einen Crayß zu formiren, und in solcher Stellung marschirten sie, ohne einen Schuß zu thun, auf unsere Dragoner und Husaren loß, sie machten dabey ein solch Gebrüll und Geschrey, daß auch die Pferde davon scheu wurden,

Tekst, który nie jest cyfrowy - nie istnieje w sieci

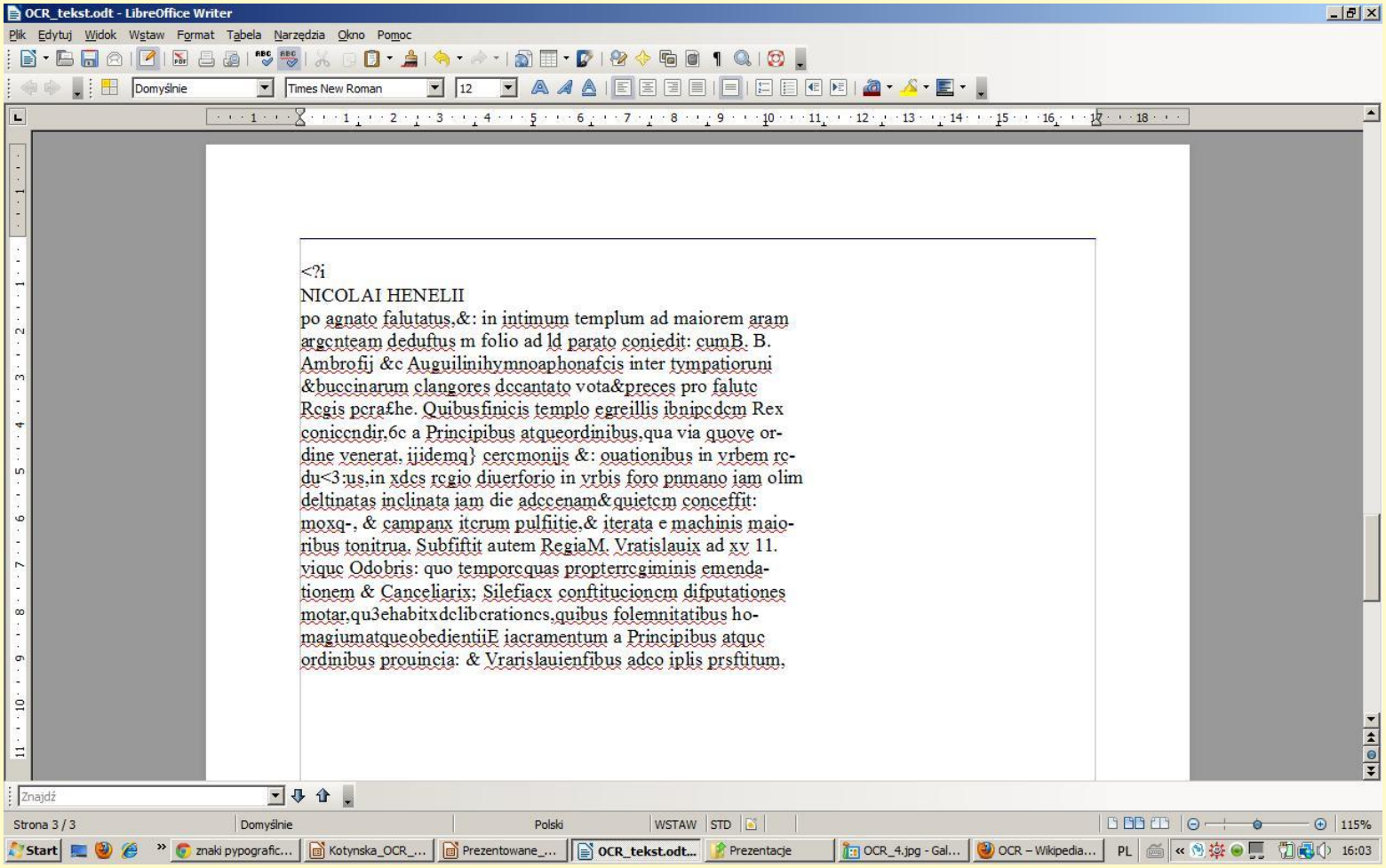


OCR umożliwia korzystanie z materiałów historycznych w formie przeznaczonej do odczytu komputerowego.

Materiały historyczne: obraz



Materiały historyczne: tekst



W efekcie OCR-u możemy uzyskać:

- pełnotekstowe indeksowanie i przeszukiwanie,
- możliwość kopiowania tekstu,
- powszechne udostępnienie materiałów „trudnych” do czytania,
- możliwość korzystania z translatorów,
- możliwość modyfikacji, uzupełnień,
- hiperlinkowanie w tekście,
- udostępnienie materiałów użytkownikom z dysfunkcją wzroku,
- umożliwienie wyeksportowania treści do formatów mobilnych.

Źródło: T. Kalota

OCR - optyczne rozpoznawanie znaków (ang. Optical Character Recognition) – zestaw technik lub oprogramowanie służące do rozpoznawania znaków oraz całych tekstów z obrazów cyfrowych i przetworzenie ich na teksty cyfrowe.
Zadaniem OCR-u jest rozpoznanie tekstu w zeskanowanym dokumencie:
- drukowanym,
- rękopiśmiennym,
- tabelarycznym.

ICR – zaawansowane rozpoznawanie kroju pisma, stopnia pisma, interlinii.

Terminy powiązane:

digitalizacja

zachowanie dziedzictwa kulturowego

dokumenty cyfrowe reprodukcja **obrazy cyfrowe**

teksty cyfrowe

badania naukowe **transkrypcja**

interpretacja ***biblioteki cyfrowe***

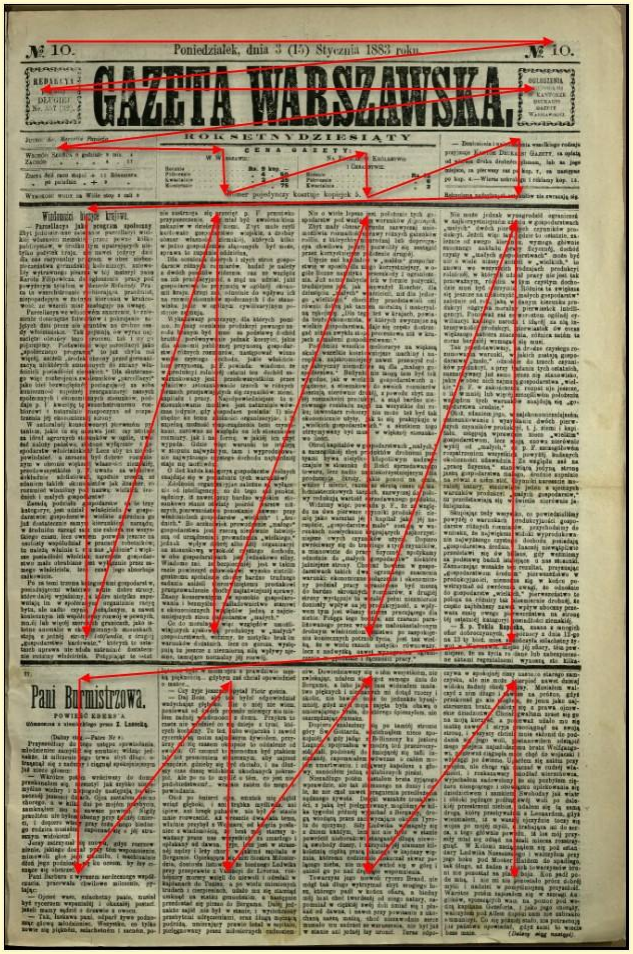
korekta **konwersja** słowniki

technologia

...

Dobry efekt rozpoznawania znaków zależy od:

- jakości tekstu i ogólnego stanu materiałów (zagięcia, pofałdowania),
- języka tekstu, alfabetu, kroju pisma,
- układu tekstu (kolumny),
- formatowania tekstu (marginesy, interlinie),
- rozdzielczości skanowania i zastosowania głębi kolorów.


№ 10. Poniedziałek, dnia 3 (15) Stycznia 1883 roku. № 10.
GAZETA WARSZAWSKA.
ROK SETNY DZIESIĄTY
Pani Burmistrzowa.

Źródło: G. Bednarek

Problemy techniczne dot. obrazu:

- wybór tekst (pomijanie grafiki),
- przebijający tekst,
- plamy, zalania, wandalizm,
- deformacja obiektu,
- kolorystyka i nasycenie druku,
- kolorystyka i nasycenie tła.

Konieczna obróbka techniczna.


WYCHODZI CO CZWARTEK.

„POKÓJ LUDZIOM DOBREJ WOLI".

Niech będzie pochw

Ewangelja

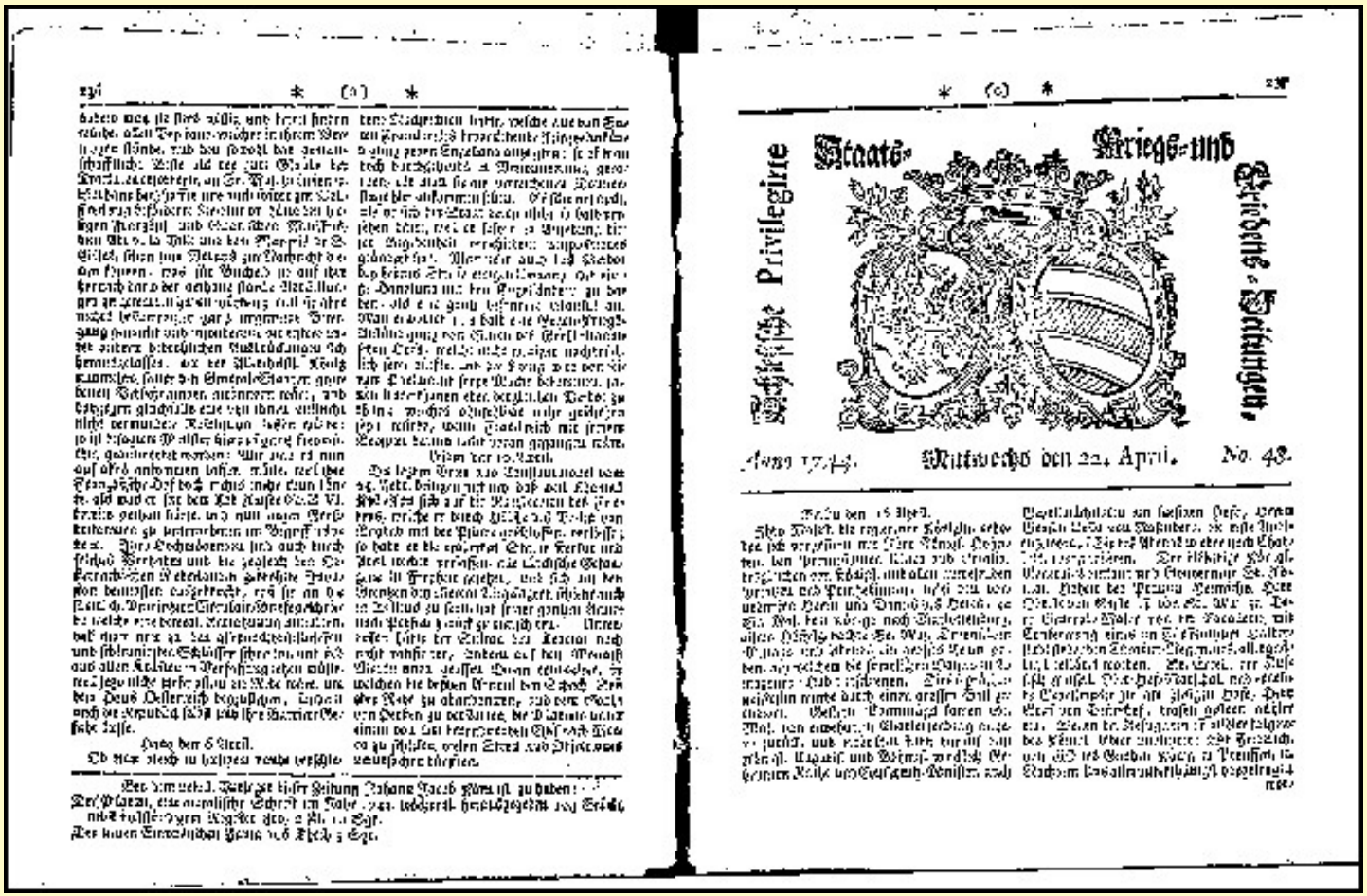
Na Niedzielę jedenastą po Św

zapisana u św. Marka w rozdz. VII. 32—

Onego czasu: Wyszedłszy Jezus z granic Tyru, przyszedł skiego, środkiem granic Dekapolskich. I przywiedli mu głuchego na niego rękę włożył. A wziąwszy go na stronę od rzeszy, wp a splunąwszy, dotknął języka jego: a wejrzawszy w niebo, westc otwórz się. I wnet otworzyły się uszy jego, i rozwiązała się zw

Źródło: G. Bednarek

Plik przed obróbką



Źródło: T. Kalota

Plik 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 leisen 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 Allerchristl. 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 liefe.

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 Briefe 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 Alt 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.

* (o) *

Hochfürstliche 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 erkläret worden. Se. Excell. der Russisch Kaiserl. Ober-Hof-Marschall und erneñte 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-

Źródło: T. Kalota

Problemy techniczne dot. języka tekstu i alfabetu:

- łączenie słów,
- zamiana znaków,
- złe odczytywanie znaków,
- nieużywane współcześnie znaki i słowa.

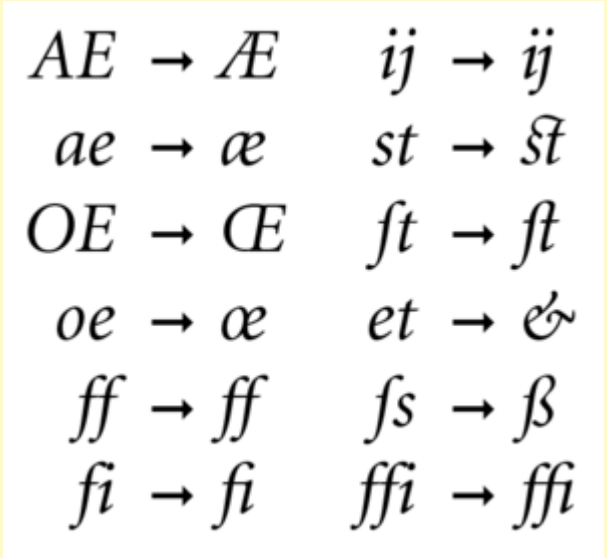
ß ſ Æ &

ss s AE et

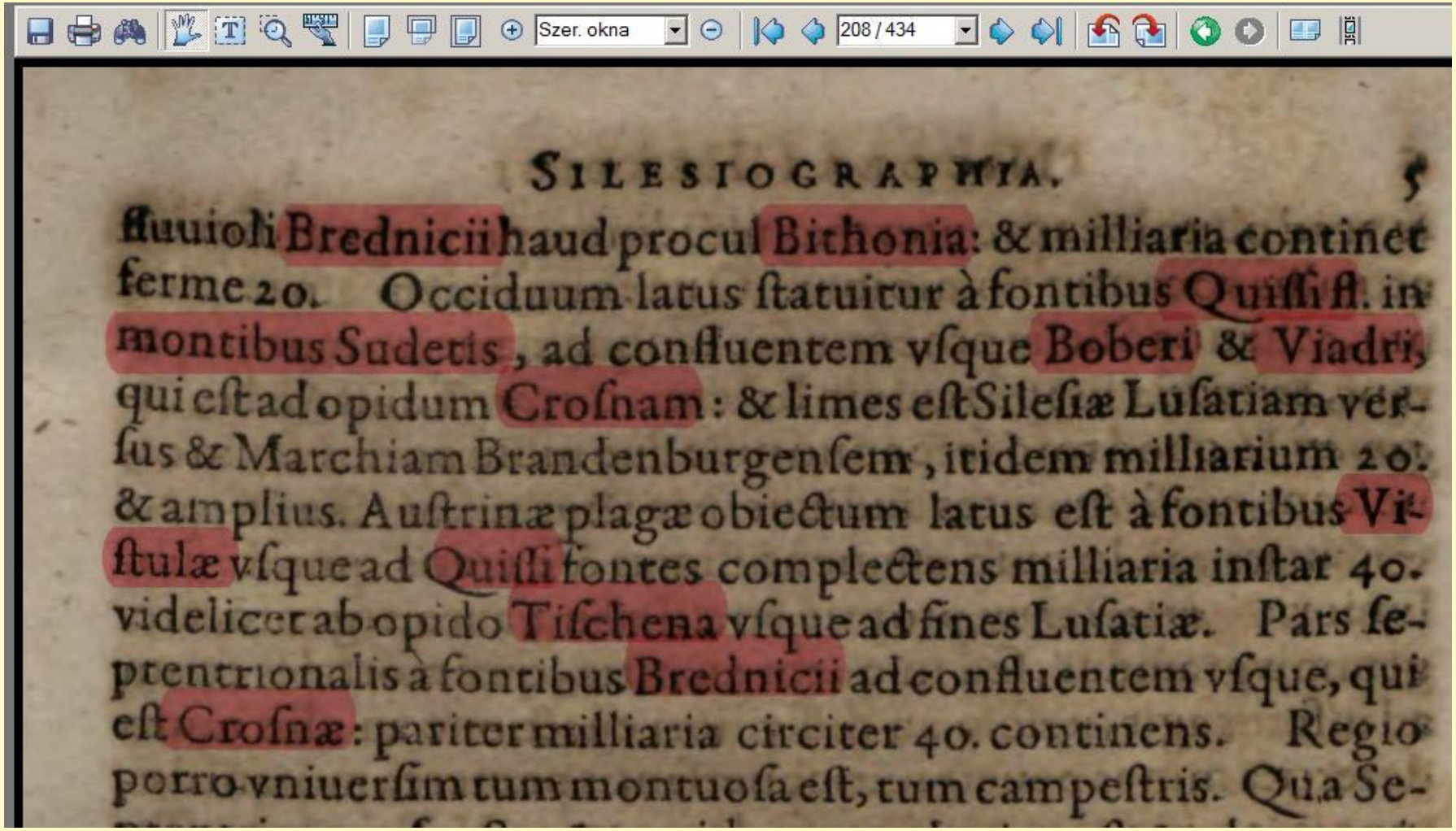
Konieczna korekta

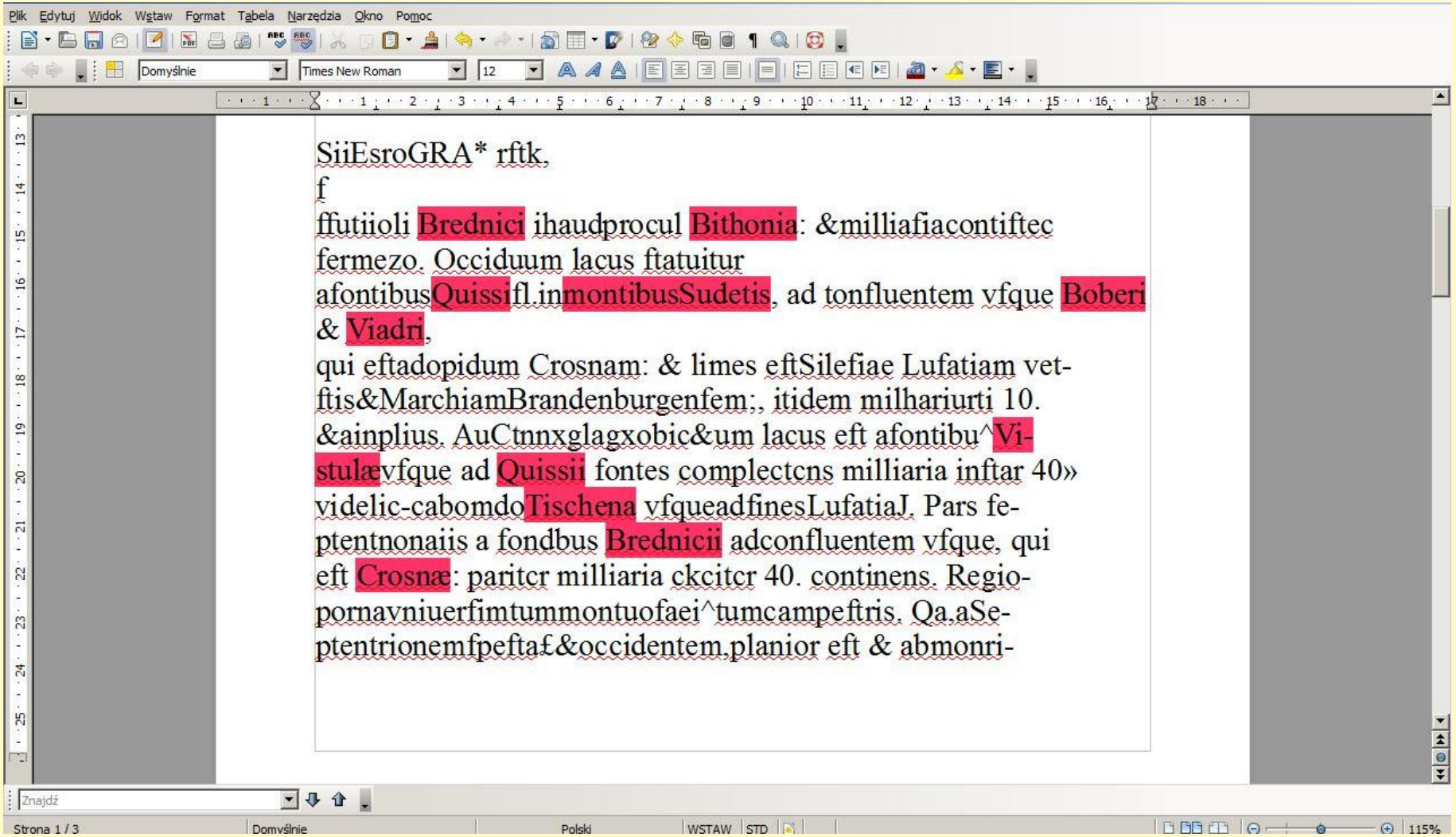
Problemy merytoryczne: obraz czy tekst?

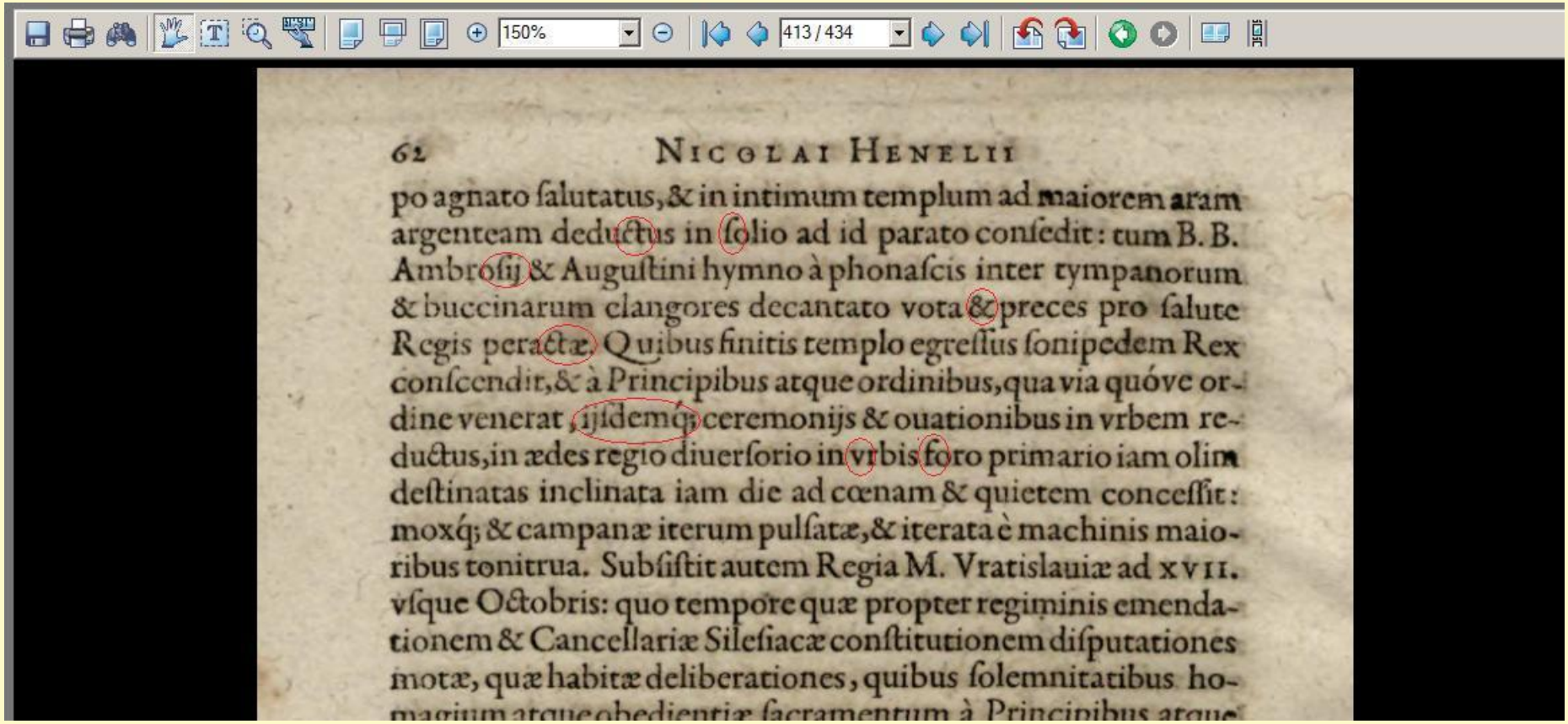
- znaki specjalne,
- litery z akcentami,
- skróty (abrewiatury, ligatury),
- dawna pisownia nazw geograficznych i osobowych,
- dawna gramatyka i fleksja,
- zasób typograficzny drukarni,
- nieużywane współcześnie słowa.



Typowe ligatury alfabetu łacińskiego


Szer. okna
208 / 434
SILESIOGRAPHIA. 5
fluuioli Brednicii haud procul Bithonia: & milliaria continet ferme 20. Occiduum latus ſtatuitur à fontibus Quiſſi fl. in montibus Sudetis, ad confluentem vſque Boberi & Viadri, qui eſt ad opidum Croſnam: & limes eſt Sileſiæ Luſatiam verſus & Marchiam Brandenburgenſem, itidem milliarium 20. & amplius. Auſtrinæ plagæ obiectum latus eſt à fontibus Viſtulæ vſque ad Quiſſi fontes complectens milliaria inſtar 40. videlicet ab opido Tiſchena vſque ad fines Luſatiæ. Pars ſeptentrionalis à fontibus Brednicii ad confluentem vſque, qui eſt Croſnæ: pariter milliaria circiter 40. continens. Regio porro vniuerſim tum montuoſa eſt, tum campeſtris. Qua Se-


Plik Edytuj Widok Wstaw Format Tabela Narzędzia Okno Pomoc
Domyślnie
Times New Roman
12
SiiEsroGRA* rftk,
f
ffutiioli Brednici ihaudprocul Bithonia: &milliafiacontiftec
fermezo. Occiduum lacus ftatuitur
afontibusQuissifl.inmontibusSudetis, ad tonfluentem vfque Boberi
& Viadri,
qui eftadopidum Crosnam: & limes eftSilefiae Lufatiam vet-
ftis&MarchiamBrandenburgenfem;, itidem milhariurti 10.
&ainplius. AuCtnnxglagxobic&um lacus eft afontibu^Vi-
stulævfque ad Quissii fontes complectcns milliaria inftar 40»
videlic-cabomdoTischena vfqueadfinesLufatiaJ. Pars fe-
ptentnonaiis a fondbus Brednicii adconfluentem vfque, qui
eft Crosnæ: paritcr milliaria ckcitcr 40. continens. Regio-
pornavniuerfimtummontuofaei^tumcampeftris. Qa.aSe-
ptentrionemfpefta£&occidentem.planior eft & abmonri-
Znajdź
Strona 1 / 3
Domyślnie
Polski
WSTAW
STD
115%

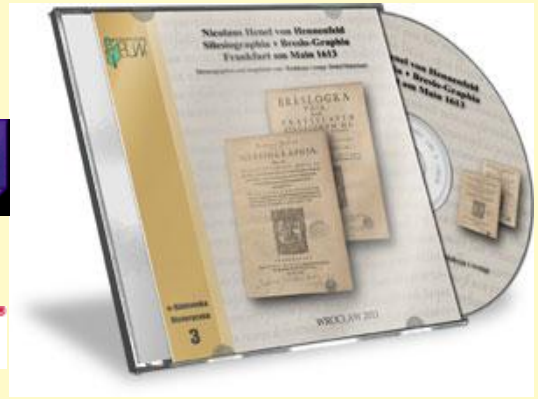


Heduuige czy **Hedwige**
Cvnradvs czy **Cunradus**
Iablvncka czy **Jabluncka** czy **Jablunka**
VVarta czy **Warta** czy **Wartha**

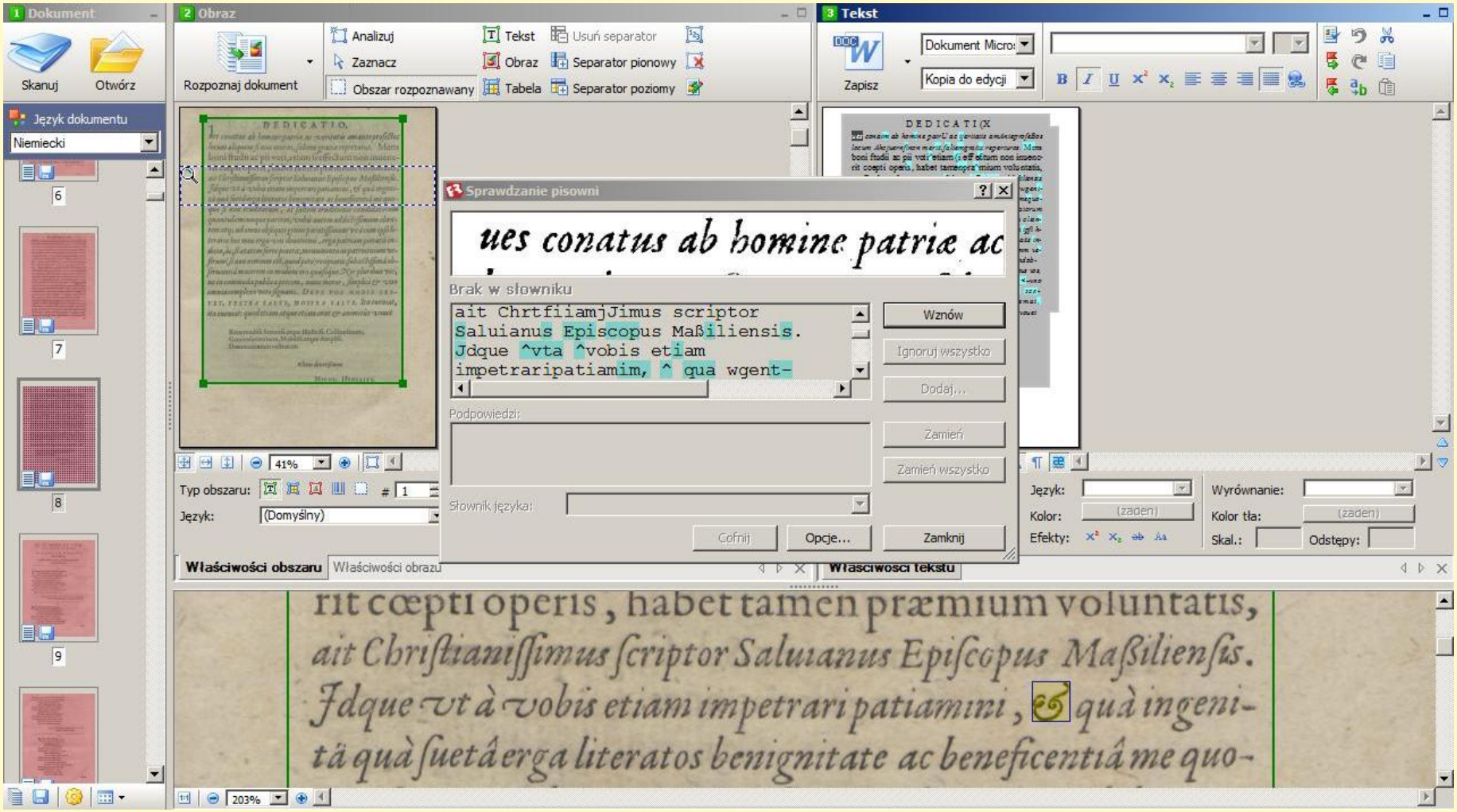
Korekta = Ekonomia

(sprzęt, oprogramowanie, licencje, praca, czas, efekty):

- korekta dla bibliotek cyfrowych,
- korekta dla projektów badawczych i wydawniczych,
- korekta społeczna,
- reCAPTCHA.



Korekta OCR Abbyy FineReader

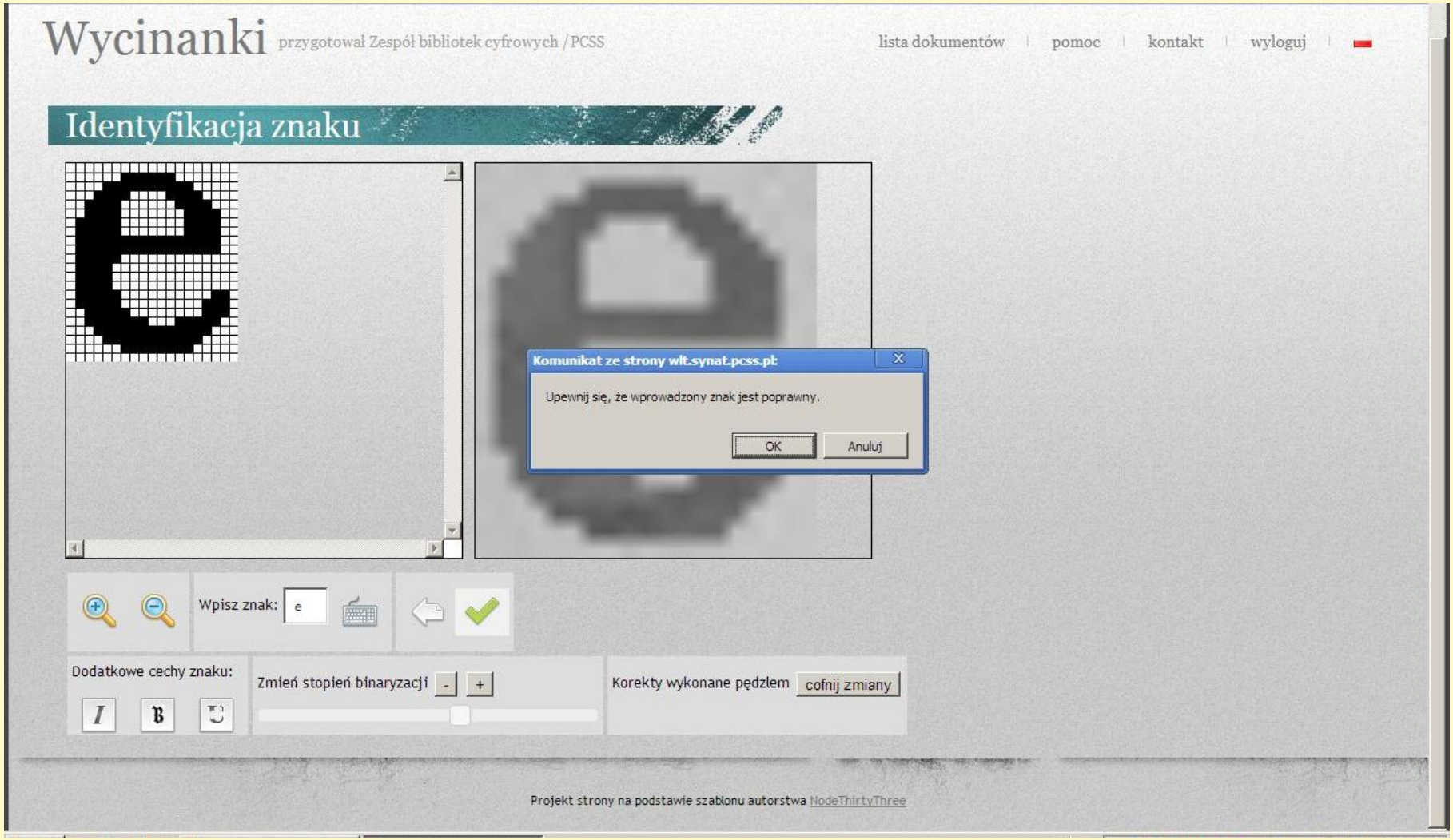


Korekta OCR ReCAPTCHA

Autoryzacja komentarza pod artykułem http://wiadomosci.ngo.pl/wiadomosci/833233.html

Korekta OCR w WLT http://wlt.synat.pcss.pl

Korekta znaków w WLT



Podsumowanie:

Cel – dążenie do pełnej użyteczności historycznych materiałów, które są już dostępne w sieci.

OCR jest nadal w fazie bardzo intensywnego rozwoju z powodu dobrych narzędzi, ale ciągle jeszcze niedoskonałych efektów ich pracy. Korekta OCR-u wykonywana przez człowieka jest efektywna, ale bardzo kosztowna. Masowa digitalizacji, aby była opłacalna wymaga automatyzacji procesów, a tym samym ciągle doskonalszych narzędzi wspomagających i „uczących" OCR. Dodatkowo generalnych ustaleń wymagają problemy merytoryczne.

Wybrane źródła:

Repozytorium Instytucjonalne PCSS: http://lib.psnc.pl/dlibra

Biblioteka 2.0: http://forum.biblioteka20.pl/index.php

Digitalizacja.pl: http://digitalizacja.pl/

Format Djvu: http://www.djvu.com.pl/

Federacja Bibliotek Cyfrowych: http://fbc.pionier.net.pl/owoc

Wybrani autorzy publikacji dot. OCR-u:

Grzegorz Bednarek, Adam Dudczak, Tomasz Kalota, Tomasz Parkoła, Paweł Rękar, Marcin Szala

* (o) * 343

Schlesische Privilegirte Staats- Kriegs- und Friedens-Zeitungen.

Anno 1745. Sonnabends den 3. Jul. No. 77.

Breßlau den 2. Jul.

Aus Böhmen enthalten die Nachrichten vom 25. Jun. folgendes: Als wir am 15ten das Lager bey Nowiengestoa oder Böhmisch Neustädtel bezogen, setzten sich die Feinde 15000. Mann starck, theils regulirte, theils Panduren, Husaren und Insurgenten auf den Bergen, Kirchhöfen und in den Gehölzen recht um unser Lager herum, und liessen uns die Zelter in guter Ruhe aufschlagen. Die Husaren formirten die beyden Flügels, die Insurgenten das Centrum, u. hinter ihnen stund die regulaire Mannschafft. Es wurde also unsers Avant-Garde, welche weiter vorwerts rücken solte, etwas verstärckt. Der Herr General von Jeetz commandirte selbige, und der Herr General von Lehwald hatte den lincken und der Herr General-Major von Winterfeld den rechten Flügel; beyde richteten ihren Marsch so ein, daß sie die feindl. Husaren von den Ungarn trenneten und abschnitten. Als nun die regulaire Mannschafft dieses sahe, nahm selbige sogleich die Flucht; während dieser Zeit rückte der Hr. General von Jeetz gegen die Insurgenten, u. delogirte dieselben; sie retirirten sich aber in der Meinung, daß ihre Flancken noch gedeckt wären, in ein nah gelegen Gehölze, allein auch da wurden sie gar bald vertrieben; hierauf suchten sie ein Dorf zu erreichen, die Unsrigen kamen ihnen aber zuvor, so daß sie sich auf freyem Felde umzingelt sahen. Da sich nun diese Ungarn solchergestalt in die Enge getrieben befanden, gab ihnen die Desperation ein, einen Crayß zu formiren, und in solcher Stellung marschirten sie, ohne einen Schuß zu thun, auf unsere Dragoner und Husaren loß, sie machten dabey ein solch Gebrüll und Geschrey, daß auch die Pferde davon scheu wurden,

Dziękuję za uwagę. Edyta Kotyńska

eTEKA.com.pl
edyta.kotynska@eteka.com.pl
Digitalizacja.pl