

32-BU.610.9.2018

Toruń, dnia 23.01.2019 r.

Opis przedmiotu zamówienia

Przedmiotem zamówienia jest **przeprowadzenie językowej korekty warstwy tekstowej plików po maszynowym OCR.**

Zamówienie dotyczy **obiektów ucyfrowionych** w ramach projektu pn. „Ucyfrowienie zasobów regionu kujawsko-pomorskiego dla potrzeb nauki i dydaktyki całego kraju”, współfinansowanego ze środków Unii Europejskiej, z Europejskiego Funduszu Rozwoju Regionalnego w ramach Poddziałania 2.3.1 „Cyfrowe udostępnienie informacji sektora publicznego ze źródeł administracji i zasobów nauki” (typ projektu: cyfrowe udostępnienie zasobów nauki) Programu Operacyjnego Polska Cyfrowa na lata 2014–2020 (nr umowy: POPC.02.03.01-00-0039/18-00 z dnia 01.08.2018 r.).

Typy obiektów ucyfrowionych: książki oraz czasopisma powstałe w XIX i XX w.

Język wyselekcjonowanych i zdigitalizowanych obiektów: głównie j. polski, ale również niemiecki, francuski, angielski, rosyjski, łacina oraz pozycje drukowane szwabachą.

Liczba skanów: min. 500 tys. – max. 630 tys., tj. 300-380 tys. skanów w języku polskim i 200-250 tys. skanów w innych językach.

Oczekiwana trafność/skuteczność: 100%. Oznacza to, że OCR w 100% jest zgodny z tekstem widocznym na cyfrowej postaci obiektu.

Zamawiający przekazuje Wykonawcy ustaloną uprzednio liczbę plików wyjściowych w formacie TIFF o rozdzielczości z zakresu 300-600 ppi na dyskach zewnętrznych lub przez sieć w terminie co 14 dni kalendarzowych, przy czym pierwsze przekazanie plików nastąpi w ciągu 5 dni roboczych od zawarcia umowy. Wykonawca przygotowuje pliki wynikowe w terminie nie dłuższym niż dwa tygodnie (14 dni kalendarzowe) od momentu dostarczenia plików wyjściowych. Otrzymanie plików wyjściowych od Zamawiającego oraz otrzymanie plików wynikowych od Wykonawcy będzie każdorazowo potwierdzane mailowo z podaną liczbą otrzymanych plików, a w przypadku plików wynikowych także protokołem odbioru. Za termin skutecznego przekazania gotowych plików wynikowych Strony uznają dzień podpisania protokołu odbioru bez zastrzeżeń.

Forma wynikowa jaka należy dostarczyć zamawiającemu:

- Wiązki z FineReadera lub innego programu, w którym wykonano OCR
- Pliki w formacie pdf z warstwa tekstową
- Pliki w formacie ALTO i txt
- Pliki-logi z programu, w którym dokonano OCR

Specyfikacja:

- Nazwy plików wynikowych mają mieć taką samą nazwę jak katalog, z którego pochodzą
- Dostępna informacja dotycząca oprogramowania, przy pomocy którego dokonano OCR-u

Specyfikacja dostarczonych plików pdf:

- Optymalizacja grafiki: Rozdzielczość obrazu nie może być mniejsza niż 300ppi
- Kompresja tła i ilustracji równa 90%, tekst bezstratnie
- W metadanych informacje o wersji PDF, strukturze, formacie grafiki rastrowej, rozdzielczości oraz o rozmiarze dokumentu oryginalnego
- Warstwa tekstowa musi odzwierciedlać podział tekstu w pliku, w szczególności podział tekstu na kolumny

Do opisu przedmiotu zamówienia dołączono próbkę zawierającą pliki w formacie TIFF oraz odpowiadające im pliki w formacie pdf z warstwą tekstową. Można je pobrać, używając następujących łączy:

<https://www.bu.umk.pl/popc/informator1>

https://drive.google.com/file/d/1N_CqeJOymemnZ_F9b7DcT8SFawjKpSyM/view?usp=drive_web

Jeśli warstwa tekstowa plików poddanych procesowi OCR będzie niezgodna z tekstem zawartym w ucyfrowionych obiektach, Wykonawca zobowiązuje się do ich poprawy bez dodatkowych kosztów.

Termin wykonania usługi: 750 dni od momentu zawarcia umowy, jednak zgodnie z harmonogramem projektu PO PC, wymienionego powyżej, nie później niż do 3.05.2021 r.