1. Przedmiot zamówienia

1.1 Wykrywanie dokumentów zawierających dane wrażliwe

Dane wrażliwe – wg definicji art. 9 RODO.

Informacje dodatkowe, które mogą być zwracane, ale nie podlegają ocenie konkursowej:

- oznaczenie typu danych wrażliwych,
- zwrócenie/wskazanie zdania zawierającego dane wrażliwe,
- oznaczenie osoby, której dane wrażliwe dotyczą.

1.2 Rozwiązanie musi zostać dostarczone w formie (np. zewnętrznej biblioteki, zewnętrznego serwisu typu REST) umożliwiającej późniejszą integrację z systemem EZD. Wykonawca ma więc umożliwić i uwzględnić taką integrację w przedmiocie konkursu, natomiast stworzenie i wdrożenie komunikacji z EZD leży po stronie Zamawiającego.

1.3 Rozwiązanie musi uwzględniać możliwość późniejszej samodzielnej aktualizacji oraz usprawniania działania przez Zamawiającego.

1.4 Dopuszcza się przekazanie kodu źródłowego na licencji zwrotnej. Kod musi obejmować zarówno model i oprogramowanie dokonujące ewaluacji przy jego pomocy. Przewiduje się także douczanie modelu (trening) w warunkach produkcyjnych i wykorzystanie niezbędnego w tym celu oprogramowania.

1.5 Dokumenty, których przedmiot zamówienia dotyczy, są w języku polskim. Jeden dokument zawiera od 1 do 20 stron tekstu.

2. Wymagania techniczne, ocena skuteczności oraz ograniczenia czasowe

2.1 Do oceny skuteczności modelu stosowana będzie metryka bazująca na f-score.

2.2 Ograniczenia techniczne podczas klasyfikacji dokumentu:

- brak karty graficznej,
- maksymalnie 8 rdzeni, 16 GB RAM,
- konieczność działania przy braku dostępu do Internetu.

2.3 Czas trwania konkursu (maksymalny czas na stworzenie rozwiązania) – nie więcej niż 6 miesięcy od daty ogłoszenia.

2.4 Czas od podpisania umowy do dostarczenia ostatniego elementu wdrożonego produktu – 3 miesiące.

2.5 Maksymalny czas klasyfikacji 1 strony dokumentu – 30s. Maksymalny czas przetwarzania 1 strony dokumentu przy douczaniu – 10 min.

3. Sposób dostarczenia rozwiązania

Rozwiązanie zostanie dostarczone w postaci dockera, uwzgledniające możliwość integracji i komunikacji z API systemu EZD RP. Na potrzeby konkursu należy przygotować rozwiązanie klasyfikujące pliki z zadanego katalogu, zawierającego dwa

kolejne katalogi. Pierwszy zawiera pliki oryginalne w formacie pdf, a drugi – tekstowe, wynik optycznego rozpoznawania znaków (np. OCR Tesseract). Jeśli dostawca zamierza dokonać własnego OCR i przedłożyć jako część rozwiązania, to biblioteka OCR musi być bezpłatna.

4. Zbiory danych

Postuluje się, że zamawiający przygotuje 2 zbiory danych:

4.1 Zbiór przykładowy, zawierający po 10-20 dokumentów na klasę/typ danych wrażliwych (np. 15 dokumentów zawierających dane wrażliwe dotyczące stanu zdrowia, 10 dokumentów dotyczących wyznania itp.).

- zbiór ten zostanie udostępniony uczestnikom konkursu,
- powinien zawierać dokumenty zarówno w formie przetworzonej (po OCR) - tekst, jak i oryginalnej,
- dane udostępnione uczestnikom będą więc w 2 formatach – tekst oraz pdf,
- zbiór powinien zawierać przykłady korespondencji reprezentujące rzeczywiste kierunki komunikacji, np. Urząd - Urząd, Obywatel - Urząd,
- dane powinny być zaszumione.

4.2 Zbiór testowy – służy do oceny jakości modelu i nie zostanie udostępniony uczestnikom.

5. Wersje i opcje do kosztorysu

Uczestnicy wstępnych konsultacji technicznych dokonali wyceny następujących modułów systemu: