dr inż. Tomasz Schweitzer
Prezes Polskiego Komitetu Normalizacyjnego
ul. Świętokrzyska 14, 00-050 Warszawa, tel. 22 556 76 00, fax: 22 556 77 80, e-mail: prezsekr@pkn.pl, www.pkn.pl

Warszawa, 2016-04-21

WIF.261.3.2016

Wszyscy Wykonawcy

Dotyczy: postępowania o udzielenie zamówienia publicznego prowadzonego w trybie przetargu nieograniczonego na zaprojektowanie, wykonanie i wdrożenie systemu do wirtualizacji Polskich Norm (PN) w wersji angielskiej (E), (WIF.261.3.2016).

Odpowiedź Zamawiającego
w ramach zgłoszonych wniosków o wyjaśnienie SIWZ cz. 2

Zgodnie z art. 38 ust. 2 ustawy z dnia 29 stycznia 2004 r. Prawo zamówień publicznych (t.j. Dz.U. z 2015 r. poz. 2164) udzielamy wyjaśnień, w związku z pytaniami Wykonawców, dotyczących Specyfikacji Istotnych Warunków Zamówienia w postępowaniu jw.

Pytanie 1.

Czy pliki PDF będące wsadem do systemu są plikami tekstowymi czy PDF'ami ze skanami rastrowymi dokumentów?

Odpowiedź:

Może się zdarzyć, że wśród dokumentów w wersji angielskiej włączonych do PDF z treścią Polskiej Normy, będą PDF ze skanami.

Pytanie 2.

Proszę o przesłanie przykładowych dokumentów zawierających grafikę, tabelę, przypisy i temu podobne.

Odpowiedź:

Zamawiający na stronie internetowej zamieścił próbkę dokumentów pdf.

Pytanie 3.

Zamawiający wskazał, iż posiada następujące licencje:

dr inż. Tomasz Schweitzer
Prezes Polskiego Komitetu Normalizacyjnego
ul. Świętokrzyska 14, 00-050 Warszawa, tel. 22 556 76 00, fax: 22 556 77 80, e-mail: prezsekr@pkn.pl, www.pkn.pl

„licencje RTL na serwer Findreader 8.1 o produktywności 75k PPM, 12 licencji na kontrolkę Pegasus ImageXpress oraz licencję MS Windows Server 2012 Standard i MsSQL Server 2008R2 Enterprise 64 bit"

Prosimy o potwierdzenie, że nie zaszła pomyłka literowa I nie chodzi o produkt Abby Fine Reader 8.1. w wersji Runtime?

Odpowiedź:

Zamawiający dokonał modyfikacji SIWZ, poniżej polegającej na poprawieniu oczywistej omyłki pisarskiej.

Pytanie 4.

Czy mogą Państwo przekazać dokładny opis elementów znajdujących się w załączniku nr 3 do SIWZ (o ile taki opis istnieje)?

Taki opis zawierałby np. informacje dotyczące formatu oraz zakresu danych umieszczanych w elemencie "metadane_g" w pliku XML.

Odpowiedź:

Zamawiający nie dysponuje dokładniejszym opisem elementów znajdujących się w załączniku nr 3 do SIWZ.

Pytanie 5.

Prosimy o dokładniejsze informacje dotyczące następującego fragmentu SIWZ:

- "2. Dialekt XML musi być zgodny z dialektem stosowanym obecnie w PKN (przedstawionym w załączniku nr 3 do SIWZ)".

Chcielibyśmy zwrócić uwagę na to, że nakład pracy wymagany do utworzenia systemu w dużym stopniu zależy od tego, które elementy widoczne w załączniku nr 3 do SIWZ muszą znajdować się w eksportowanym pliku XML. Teoretycznie sama "zgodność z dialektem stosowanym obecnie w PKN" mogłaby być osiągnięta poprzez utworzenie pliku XML pasującego do określonego schematu, ale nieposiadającego wszystkich elementów widocznych w załączniku nr 3 do SIWZ.

A. Prosimy o informacje dotyczące tego, które z następujących elementów muszą być obsługiwane przez system (np. pobierane z pliku PDF, eksportowane do pliku XML):

- "grafika".
- "tekstografika".

dr inż. Tomasz Schweitzer
Prezes Polskiego Komitetu Normalizacyjnego
ul. Świętokrzyska 14, 00-050 Warszawa, tel. 22 556 76 00, fax: 22 556 77 80, e-mail: prezsekr@pkn.pl, www.pkn.pl

- "tabelografika".

Czy wystarczające byłoby użycie tylko elementu "grafika", zamiast trzech wyżej wymienionych elementów?

B. Czy niezbędne jest uzupełnianie elementów oznaczonych napisem "Dane z OCR grafiki"?

Czy konieczne jest wykorzystywanie technologii OCR (rozpoznawanie tekstu znajdującego się w grafikach w pliku PDF, podczas konwertowania do formatu XML)?

C. Jakie informacje dotyczące kolejności odpowiednich elementów muszą znaleźć się w pliku XML?

Rozumiemy, że dane poszczególnych stron zawierają w sobie m. in. dane o tekście występującym na stronie, dane o grafikach i tabelach. Czy niezbędne jest umieszczanie w pliku XML dodatkowych informacji o kolejności odpowiednich elementów? Chodzi np. o dane mówiące o tym, że w pewnym pliku PDF mamy na danej stronie kolejno: tekst, grafikę, drugi tekst, tabelę, a następnie kolejny tekst

Odpowiedź:

Wszystkie elementy widoczne w załączniku nr 3 do SIWZ muszą znajdować się w eksportowanym pliku XML, o ile występują w pliku PDF. Kolejność elementów w pliku XML musi odwzorowywać kolejność elementów w pliku PDF.

Ad. A. Niewystarczające będzie użycie tylko elementu "grafika".

Ad. B. Niezbędne jest uzupełnianie elementów oznaczonych napisem "Dane z OCR grafiki", ze względu na spójność z dialektem XML wykorzystywanym przez Zamawiającego. Zamawiający nie wskazuje technologii wykorzystywanej do rozpoznawania tekstu w grafice.

Ad. C. Niedopuszczalne jest wprowadzenie dodatkowych informacji o kolejności odpowiednich elementów. Kolejność elementów w pliku XML musi odwzorowywać kolejność elementów w pliku PDF.

Pytanie 6.

W jaki sposób będą rozwiązywane problemy, które mogą się pojawić, gdy pliki PDF nie będą posiadały danych zakodowanych w sposób pozwalający na ich automatyczne przetworzenie?

W naszej ocenie jest to poważny problem.

dr inż. Tomasz Schweitzer
Prezes Polskiego Komitetu Normalizacyjnego
ul. Świętokrzyska 14, 00-050 Warszawa, tel. 22 556 76 00, fax: 22 556 77 80, e-mail: prezsekr@pkn.pl, www.pkn.pl

A. Przykładowo - w dokumencie "Dokument_5_20 stron z 60.pdf" mamy sekcję "Normative references". Rozumiemy, że system powinien automatycznie odczytać treść tej sekcji. Nie wiadomo jednak, w jaki sposób wykryć koniec takiej sekcji podczas automatycznego przetwarzania pliku PDF. Czy wystarczające będzie dodanie mechanizmu wykrywającego koniec sekcji na podstawie tekstu podanego przez użytkownika (np. za koniec sekcji możemy uznać podany przez użytkownika napis "Terms and definitions")?
Dla użytkownika oczywiste jest, gdzie znajduje się koniec sekcji - jednak w naszej ocenie, tego rodzaju informacje mogą nie być obecne w pliku PDF.
B. Inny przykładem tego problemu jest sposób przetwarzania tabel.
W pliku PDF, informacje o tym, że dana część pliku (wyglądająca jak tabela), może po prostu nie być opisana jako tabela - może być np. zapisana jako zwykły tekst lub jako grafika (lub połączenie np. tekst na tle grafiki). Z tego powodu wykrycie informacji np. o liczbie tabel występujących w pliku PDF może okazać się niemożliwe.

Czy dopuszczalnym rozwiązaniem w takich sytuacjach jest dodanie do pliku XML danych przepisanych wprost z pliku PDF (czyli np. komórki tabeli zostałyby zapisane w pliku XML jako zwykły tekst). Proszę zwrócić uwagę na fakt, że samo wyciągnięcie tekstu z pliku PDF (bez sprawdzania, które elementy są częścią tabeli) powinno wystarczyć do zapewnienia prawidłowego działania mechanizmów do wyszukiwania pełnotekstowego.

Odpowiedź:

Oferowany system musi spełniać wszystkie wymagania określone w SIWZ. Zamawiający nie narzuca sposobu zaprojektowania mechanizmów systemu wirtualizacji. Nie jest dopuszczalne rozwiązanie polegające na dodaniu do pliku XML danych przepisanych wprost z pliku PDF, ponieważ byłoby to niezgodne z wymaganiami SIWZ.

Pytanie 7.

Prosimy o informację dotyczącą reprezentatywności próbki przykładowych dokumentów.
Rozumiemy, że pliki PDF w wersji angielskiej są tworzone przez różne organizacje normalizacyjne.
A. Z ilu różnych organizacji będą pochodzić przetwarzane pliki PDF?

dr inż. Tomasz Schweitzer
Prezes Polskiego Komitetu Normalizacyjnego
ul. Świętokrzyska 14, 00-050 Warszawa, tel. 22 556 76 00, fax: 22 556 77 80, e-mail: prezsekr@pkn.pl, www.pkn.pl

B. Z ilu różnych organizacji pochodzą pliki PDF wchodzące w skład próbki przykładowych dokumentów?

Odpowiedź:

W skład przekazanej próbki dokumentów wchodziły fragmenty dokumentów PDF z czterech różnych organizacji. Zostały wybrane ze względu na zawartość odzwierciedlającą wszystkie elementy, które mogą wystąpić w pliku XML. Dokumenty do przetworzenia nie są jednorodne, nawet jeżeli pochodzą z jednej organizacji. Zamawiający nie jest w stanie określić liczby organizacji, z których będą pochodzić pliki PDF.

Modyfikacja SIWZ

Zgodnie z art. 38 ust. 4 ustawy z dnia 29 stycznia 2004r. Prawo zamówień publicznych (t.j. Dz.U. z 2015 r. poz. 2164) informuję, że dokonano następującej modyfikacji Specyfikacji Istotnych Warunków Zamówienia w postępowaniu na zaprojektowanie, wykonanie i wdrożenie systemu do wirtualizacji Polskich Norm (PN) w wersji angielskiej (E), (WIF.261.3.2016).

1. **Rozdział 2, pkt 2 „Licencje", otrzymuje brzmienie:**

PKN posiada licencje RTL na serwer FineReader 8.1 o produktywności 75k PPM, 12 licencji na kontrolkę Pegasus ImageXpress oraz licencję MS Windows Server 2012 Standard i MsSQL Server 2008R2 Enterprise 64 bit.

Prezes Polskiego Komitetu Normalizacyjnego

/-/ Tomasz Schweitzer [1]

[1] Podpis elektroniczny weryfikowany certyfikatem kwalifikowanym.