\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_­­­­­­­­­\_\_\_\_\_\_\_\_\_\_\_

PD.260.01.2024

**Załącznik nr 7 - Wymagania dotyczące wykonania rozpoznania tekstu OCR oraz plików ALTO**

1. Każda strona, każdego obiektu cyfrowego, zawierająca co najmniej jedno słowo drukowane, poddana będzie rozpoznaniu OCR,
2. Wymóg wykonania rozpoznania OCR nie dotyczy stron zawierających wyłącznie : grafiki, fotografie, pismo odręczne lub zapis nutowy.
3. Wymagana skuteczność prawidłowego rozpoznania OCR słów na pojedynczej stronie: przynajmniej 80%,
4. Niższy próg trafności rozpoznania OCR dopuszczony jest jedynie dla stron zawierających wzory   
   i wyrażenia matematyczne, fizyczne, chemiczne i/lub odwrócone o 90° względem poziomu zestawienia tabelaryczne,
5. Rozpoznanie OCR należy prowadzić wyłącznie z uwzględnieniem układu graficzno-logicznego strony (pełne treści kolejnych artykułów wydrukowanych w kolejnych obszarach/kolumnach strony) i w taki też sposób rozpoznany tekst musi być zapisany w plikach prezentacyjnych   
   w formacie PDF,
6. Rozpoznany tekst pojedynczej strony należy zachować jako prosty plik tekstowy zawierający wyłącznie rozpoznane na stronie słowa. W takich plikach tekstowych należy stosować kodowanie liternictwa UTF-8. Pliki tekstowe zapisane będą w folderze „TEXT”.
7. Ponadto, rozpoznany tekst strony należy zapisać jako tekstowy plik XML odpowiadający wymogom standardu ALTO w wersji 4.0 lub nowszej. Plik zapisany będzie w folderze „ALTO”,
8. Sposób nazewnictwa pliku dla każdej strony publikacji opisany jest w Załączniku nr 11.

**Wykonanie pliku ALTO dla rozpoznanego tekstu pojedynczej strony.**

Każdy wykonany plik ALTO musi zawierać :

* + - 1. Współrzędne kolejnych słów wyrażonych w jednostce inch/1200 lub w pikselach,
      2. Określenie z nazwy pliku obrazu, którego dotyczy rozpoznany tekst (element fileName),
      3. Określenie identyfikatora pliku obrazu, a nadanego w pliku METS (element fileIdentifier),

1. Określenie nadanego w pliku METS identyfikatora obiektu cyfrowego, którego dotyczą : w/w plik obrazu i plik ALTO (element documentIdentifier),
2. Określenie użytego oprogramowania (elementy contentGeneration oraz postOperation),
3. Wyznaczenie wartości elementów TopMargin, LeftMargin, RightMargin, BottomMargin,
4. Przypisanie – w oparciu o metadane deskryptywne - atrybutowi LANG w kontenerach „ComposedBlock”, „TextBlock” lub ew. „TextLine” kodu języka, którego dotyczy rozpoznany tekst.
5. Podział każdego słowa na sylaby należy zapisywać w atrybucie HYP, któremu jako wartość, należy przypisać znak myślnika, a nie numer kodu znaku podziału, co było znamiennym dla starszych wersji schematu ALTO,
6. Dla elementów string, których zawartością jest początkowa lub końcowa część słowa podzielonego na sylaby, należy przypisać odpowiednie wartości do atrybutów SUBS\_CONTENT oraz SUBS\_TYPE.