

Znak sprawy: SP-271-4/18

Załącznik nr 3 do OPZ

Wykonywanie i zachowanie tekstu rozpoznanego w kolejnych plikach macierzystych

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, zapis nutowy lub tekst wydrukowany czcionkami cyrylicznymi, gotyckimi (szwabachą), arabskimi lub greką,
3. Wymagana skuteczność prawidłowego rozpoznania OCR słów na pojedynczej stronie: przynajmniej 84%,
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 oraz strony zawierające tekst zarówno czcionkami alfabetu łacińskiego jak i czcionkami cyrylicznymi, arabskimi lub greką,
5. Rozpoznanie OCR należy prowadzić wyłącznie z uwzględnieniem układu graficzno-logicznego strony (pełne treści kolejnych artykułów wydrukowanych w kolejnych szpaltach/obszarach strony) i w taki też sposób rozpoznany tekst musi być zapisany w plikach prezentacyjnych w formatach DjVu oraz PDF,
6. Rozpoznawanie tekstu bez uwzględnienia układu szpalt, kolumn lub innych wydzielonych na stronach obszarów, jest niedopuszczalne,
7. Rozpoznany tekst pojedynczej strony należy zachować jako prosty plik tekstowy zawierający wyłącznie rozpoznane na stronie słowa. W takich plikach tekstowych, należy stosować kodowanie literactwa UTF-8. W archiwizowanych pakietach informacji, pliki tekstowe zapisane będą w folderze „TEXT”.
8. Ponadto, rozpoznany tekst strony należy zapisać jako tekstowy plik xml odpowiadający wymogom standardu ALTO w najnowszej jego wersji. Plik zapisany będzie w folderze „ALTO”,
9. Sposób nazewnictwa pliku dla każdej strony publikacji opisany jest w załączniku nr 1 do OPZ.

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*),

4. Określenie nadanego w pliku METS identyfikatora archiwizowanego pakietu informacji, którego dotyczą : w/w plik obrazu i plik ALTO (element *documentIdentifier*),
5. Określenie użytego oprogramowania (elementy *ocrProcessingStep* oraz *postProcessingStep*),
6. Wyznaczenie wartości elementów *TopMargin*, *LeftMargin*, *RightMargin*, *BottomMargin*,
7. Przypisanie – w oparciu o metadane deskryptywne - atrybutowi LANG w kontenerach „*ComposedBlock*”, „*TextBlock*” lub ew. „*TextLine*” kodu języka, którego dotyczy rozpoznany tekst.
8. 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 znamienym dla starszych wersji schematu ALTO,
9. 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.