NDNP OCR Profile

Version 1.7

CHANGES IN 1.7:

1. Added clarification about column organization.

1. OCR text must be reflect columns of the original newspaper and be ordered column-by-column (that is, in a natural reading order).
2. OCR text will be encoded using the ALTO (Analyzed Layout and Text Object) schema, Version 1-1-041 or greater, with the additional clarifications stated below.
3. The value for MeasurementUnit will be "inch1200," which is 1/1200 of an inch.
4. The use of the SourceImageInformation\fileName element is required. This should include the path if the path contains useful information (e.g., identifying the newspaper title and/or issue).
5. The use of the OCRProcessing element is encouraged.
6. If the OCRProcessing element is used, the use of the ProcessingSoftware element is required. If the software does not have a commercial name, the name of the executable may be used.
7. For all applicable elements, the use of STYLEREFS and language are encouraged.
8. For the PageElement, the use of Height, Width, PRINTED_IMG_NR, QUALITY, POSITION, and PROCESSING are encouraged.
9. For the PageElement, the entire page may be included in the PrintSpace. (Thus, use of TopMargin, LeftMargin, RightMargin, and BottomMargin are not required.)
10. The use of Illustration, GraphicalElement, and ComposedBlock are not required.
11. The use of non-rectangular blocks is not encouraged.
12. The use of SP and HYP are encouraged.
13. For a TextLine, the use of BASELINE is discouraged.
14. For a String, the use of ALTERNATIVE, WC, and CC is encouraged if available.
15. For a String, the use of HEIGHT, WIDTH, HPOS, and VPOS is required.

Additional clarifications:

1. For the ProcessingStepSettings, the settings can be specified as the command-line arguments given to the processing software.
2. For a String, the CONTENT should be a word, not a character.