NDNP OCR Profile

Version 1.17

CHANGES in 1.17:

1. Added clarification for languages that have both "b" (bibliographic) and "t" (terminology) ISO 639-2 language codes.

CHANGES in 1.16:

1. Expanded language code specification information to include any language with a valid ISO 639-2 alpha-3 language code.

CHANGES in 1.15:

1. Added language code specification information for Danish, Hungarian, Norwegian, Portuguese, and Swedish text.

CHANGES in 1.14:

1. Added language code specification information for German text.
2. Added clarification about non-English text language codes.

CHANGES in 1.13:

1. Changed version of ALTO to 2.0.
2. Added language code specification for non-English text and acceptable language codes.

CHANGES in 1.12:

1. Required use of HEIGHT and WIDTH for Page element.

CHANGES IN 1.11:

1. Changed version of ALTO to 1-4.

CHANGES IN 1.10:

1. Added clarification of hyphenation.

CHANGES IN 1.9:

1. Added prohibition on multiple Strings in same location.
2. Further clarified natural reading order.

CHANGES IN 1.8:

1. ALTO version updated to 1-2.

CHANGES IN 1.7:

1. Added clarification about column organization.
2. OCR text will be encoded using the ALTO (Analyzed Layout and Text Object) schema, Version 2.0, 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 Page element, the use of PRINTED_IMG_NR, QUALITY, POSITION, and PROCESSING are encouraged.
9. For the Page element, the use of HEIGHT and WIDTH are required.
10. For the Page element, the entire page may be included in the PrintSpace. (Thus, use of TopMargin, LeftMargin, RightMargin, and BottomMargin are not required.)
11. The use of Illustration, GraphicalElement, and ComposedBlock are not required.
12. The use of non-rectangular blocks is not encouraged.
13. The use of SP and HYP are encouraged.
14. For a TextLine, the use of BASELINE is discouraged.
15. For a String, the use of ALTERNATIVE, WC, and CC is encouraged if available.
16. For a String, the use of HEIGHT, WIDTH, HPOS, and VPOS is required.
17. The coordinates of Strings should not overlap. In other words, for each location on a page, there should only be one String. (If alternatives are desired, ALTERNATIVE should be utilized, not multiple Strings.)
18. OCR text must be in natural reading order. Thus, OCR text should reflect columns of the original newspaper and be ordered column-by-column. In addition, the ordering of all elements should reflect the original newspaper. (That is, reading order should be indicated by the ordering of elements, for example, Strings should be in reading order.)
19. Non-English text must be encoded at the TEXTBLOCK, using ISO 639-2 alpha-3 language codes (http://www.loc.gov/standards/iso639-2/). For languages that have both "b" (bibliographic) and "t" (terminology) codes, use the "b" (bibliographic) code. Language search support will vary according to the tools/technologies in use by the Library of Congress' web sites. Fraktur/black letter fonts must incorporate OCR technical processing that includes Fraktur/black letter specific tools.

Note: a single ALTO document may have multiple languages encoded within individual TEXTBLOCKs (e.g. bilingual newspaper pages), but a single TEXTBLOCK may only have a single language.

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.
3. If a hyphen splits a word at the end of a line, the OCR file should represent both fragments of the word, the hyphen, and the complete word. See the following example, where the word "experts" was split at the end of a line.

```
<String ID="P5_ST00015" HPOS="5508" VPOS="24344" WIDTH="170"
```

```
HEIGHT="61" CONTENT="ex" SUBS_TYPE="HypPart1"
SUBS_CONTENT="experts" WC="0.96" CC="111"/>
<HYP CONTENT="-"/>
</TextLine>
<TextLine ID="P5_TL00003" HPOS="3146" VPOS="24425" WIDTH="2532"
HEIGHT="108">
<String ID="P5_ST00016" HPOS="3146" VPOS="24439" WIDTH="288"
HEIGHT="94" CONTENT="perts" SUBS_TYPE="HypPart2"
SUBS_CONTENT="experts" WC="0.99" CC="00001"/>
```

4. If the hyphenated word occurs in the middle of a line, the hyphen should be left in place. See the following example where the word is "re-examination" occurred in the middle of a line.

```
<String ID="P2_ST03691" HPOS="11428" VPOS="15727" WIDTH="897"
HEIGHT="89" CONTENT="re-examination" WC="0.97"
CC="01001011010110" STYLEREFS="TXT_5"/>
```

5. Because a single reel or batch may contain variations in language and OCR quality, automated language recognition (if used) should be applied to individual titles, rather than across entire reels or batches.