Changes to 2013-07-15 PAGE XML Format

Christian Clausner

July 2016

1 "Page" Element

- New attributes (same type/content as TextRegion counterparts; lower-level definitions override the page-level definitions)
  - "primaryLanguage"
  - "secondaryLanguage"
  - "primaryScript"
  - "secondaryScript"
  - "readingDirection"
  - "textLineOrder" (see below)
  - "externalRef" (for "Metadata" sub-element)

2 "TextRegion" Element

- New attributes:
  - "dataType" (for "TextEquiv" sub-element)
    - Type of text content
    - String, optional
    - Constraints: xsd:decimal, xsd:float, xsd:integer, xsd:boolean, xsd:date, xsd:time, xsd:dateTime,xsd:string,other
  - "dataTypeDetails" (for "TextEquiv" sub-element)
    - Refinement for "dataType", e.g. a regular expression
    - String, optional
  - "xHeight" (for "TextStyle" sub-element)
    - "the x-height or corpus size refers to the distance between the baseline and the mean line of lower-case letters in a typeface";
    - Integer, optional, Unit: pixel
  - "textLineOrder"
    - Inner-block order of text lines (in addition to "readingDirection" which is the inner-text line order of words and characters)
    - String, optional
    - Constraints: top-to-bottom, bottom-to-top, left-to-right, right-to-left
- Modified attributes:
  - "primaryScript", "secondaryScript"
    - New constraints using ISO 15924 http://unicode.org/iso15924/iso15924-codes.html

        - Requires migration from Chinese-Traditional and Chinese-Simplified to Han (Traditional variant) and Han (Simplified variant)

3 “TextEquiv” Element (Text Content)

- “TextEquiv” (sub-element of text region, text line, word and glyph)
    - Can now occur multiple times to reflect OCR / text variants
    - Additional “index” attribute can be used to order the variants (lowest index is used as main text content)
    - Additional “comments” attribute (explanations, e.g. if different OCRs are used or an annotated version of the text content is entered)
    - “conf” (OCR confidence) is attached to “TextEquiv” as before

4 “TextLine”, “Word” and “Glyph” Elements

- New attributes:
    - “script” for Glyph
        - Same type / content as primaryScript / secondaryScript
    - “primaryScript”, “secondaryScript” for TextLine and Word