Specification for using ALTO in newspaper digitization projects (Digidaily and SAP) at Kungl. biblioteket.

Version 2.0 (2013-08-29)

ALTO

ALTO is a metadata standard used to store information about both the content and the layout. The full ALTO standard is included in the metadata specification for Digidaily and SAP. However, not all the elements and attributes contained in the standard are to be used. The element and attribute deemed necessary or desirable for a good metadata level has been specified below. The majority of these elements and attributes are requirements and the use of a small part depends on whether the software can automatically deliver them. It is possible that the software used for OCR and segmentation in a good way also can supply elements and attributes that are not explicitly included in the specification below. When this is the case, these elements and attributes may be included in the specification after a valuation. To enable this, working closely with MKC is a necessity. Elements and attributes used will be specified in a separate attachment to this document (table).

There must always be one ALTO file for each image file. This also applies to pages that are missing in the physical original as these are replaced with blank pages in the archive package.

Each ALTO file starts with the element alto.

- Attributes on this alto element level:
  - xmlns:xsi
  - xmlns:xlink
  - xmlns
  - xsi:schemaLocation

- Elements (sections):
  - Description
  - Styles
  - Layout

Description

This section contains metadata of a more general nature.

- Elements:
  - MeasurementUnit
  - SourceImageInformation
  - OCRProcessing

MeasurementUnit (mm10)

As a unit of measure 1/10 mm is used, because other units are both more difficult to visualize and to translate into SI units.

SourceImageInformation

Elements:

- fileName

The filename of the file that was used to generate the information in the ALTO file should always be specified. While it is possible to connect the ALTO file with the original file through information in the METS file, but by including the file name an insurance against data loss is created because the filename Contains the Libris record number for the entry of the digitized publication. This also makes it easier to identify the contents of a file in an independent ALTO file.

OCRProcessing/ocrProcessingStep

All image processing that is done before the OCR is reported in the METS file (under MIX), why this metadata does not need to be repeated inside the ALTO file as a ocrPreProcessingStep attribute. Other settings are specified in an instance of OCRprocessing. In such an instance only one description of the parameters used in character recognition is allowed. Since the ALTO file describes a single image file, this is no problem. If the ALTO file instead had described several pages and the pages characters had been interpreted with different settings, several instances of this element had been needed. If the result from the character recognition has been post processed, this must be stated within this element. Note that several different settings for the post processing can be indicated. It is important that the parameters and settings, both for character recognition, and post processing, are described in detail to allow for reinterpretation and valuation of the result.

- Elements**:**
  - processingDateTime
  - processingAgency
  - processingStepSettings
  - processingSoftware

OCRProcessing/ocrProcessingStep/processingSoftware

- Elements**:**
  - softwareCreator
  - softwareName
  - softwareVersion
  - applicationDescription (Description of the key features for the software. Used for eg Non-commercial software)

OCRProcessing/postProcessingStep

- Elements**:**
  - processingDateTime
  - processingAgency
  - processingStepSettings

- processingSoftware

OCRProcessing/postProcessingStep/processingSoftware

- Elements**:**
  - softwareCreator
  - softwareName
  - softwareVersion
  - applicationDescription (Description of the key features for the software. Used for eg Non-commercial software)

To reduce the ALTO file's scope, a number of font sizes are defined. In the text references to these sizes are then entered. The granularity of the sizes should initially be limited to 0.5 points and only font sizes appearing on newspaper page will be included in the list. Granularity should be evaluated to determine if there is enough with a difference of 1 point.

Styles

- Elements:
  - TextStyle

Styles/TextStyle

- Attributes:
  - ID
  - FONTSIZE
  - FONTFAMILY
  - FONTSTYLE

Layout

In this section, all metadata is saved that derives from the object's structure, such as dimensions, text content, images, titles, etc. In ALTO the hierarchy PAGE> Print Space> ComposedBlock> TextBlock-> Text Line> String are used to describe the content of a page

Some metadata is listed on page level. In the attribute QUALITY a rough measure of the page's quality may be stated (OK, Damaged, Missing).
It is also possible to supplement the measure with a descriptive text if needed, e.g. "the page contains a hole." If pagination is automatically identified, this should be included. Page position should also be indicated (right, left, etc.) if this can be done automatically. The estimated accuracy of the character recognition on page level should be specified if the software can deliver this. The accuracy can be specified in two places in the standard (<Accuracy> and Page Confidence, <PC>) and it is difficult to see why this is so. It is possible that it is so simple that the two attributes are distinguished in that they use different units (<Accuracy> given in percentages and <PC> as a number between 0 and 1). Both measures should be specified if the software supports it. If this is not the case, only the attribute that is supported should be entered.

[Used by Zissor:] The page margins are not stated explicitly as they can be easily calculated from the far side and printSpace (the extent of the text, not including page numbers and similar text).

Page

- Attributes:
  - ID
  - HEIGHT
  - WIDTH
  - PHYSICAL_IMG_NR (page number within the publication, normal serial numbers of the image file)
  - PRINTED_IMG_NR (the printed page number desirable if it can be identified)
  - PROCESSING (Link to OCRProcessing-element)
  - QUALITY (The quality of the page in the physical original)
  - QUALITY_DETAIL (free text)
  - POSITION
  - ACCURACY
  - PC

- Elements:
  - TopMargin
  - BottomMargin
  - LeftMargin
  - RightMargin
  - PrintSpace

Page/TopMargin, BottomMargin, LeftMargin, RightMargin

- Attributes used for each margin element:
  - HEIGHT
  - WIDTH
  - HPOS
  - VPOS

Page/PrintSpace

A physical page should be split in the hierarchy ComposedBlock/TextBlock/TextLine/String (where string can be replaced by SP for space and HYP for hyphens).

- Attributes:
  - ID
  - HEIGHT
  - WIDTH
  - HPOS
  - VPOS

- Elements:
  - ComposedBlock

Page/PrintSpace/ComposedBlock

ComposedBlock i defined in the ALTO standard as “a block that consists of other blocks”. In the Digidaily and Alto projects ComposedBlock is used to mark the area on the page which has been identified as “an article”. This identification is done entirely by machine.

- Attributes:
  - ID
  - HEIGHT
  - WIDTH
  - HPOS
  - VPOS
  - TYPE

- Elements:
  - TextBlock

Page/PrintSpace/ComposedBlock/TextBlock

- Attributes:
  - ID
  - HEIGHT
  - WIDTH
  - HPOS
  - VPOS
  - ROTATION (The text block's slope. The value is given as degrees counterclockwise)
  - language

It is possible to specify the language for a text block. If the software can automatically recognize the language it is recorded as a three-digit code according to ISO639-2b. Example: "eng", "swe". If not possible to identify: "und".

- Elements:
  - TextLine

Page/PrintSpace/ComposedBlock/TextBlock/TextLine

- Attributes:
  - ID
  - HEIGHT
  - WIDTH
  - HPOS
  - VPOS

- Elements (with attributes):
  - String (One word, separated by spaces or hyphen)
    - ID
    - HEIGHT
    - WIDTH

        - HPOS
        - VPOS
        - CONTENT (The content (the word) in the string)
        - STYLEREFS
        - SUBS_TYPE (Contains the whole word in hyphenation or abbreviation (HypPart1, HypPart2 and Abbreviation)
        - WC (Word confidence)
    - ALTERNATIVE (Alternative Spellings that the OCR interpreter can generate)
    - SP (space)
        - ID
        - WIDTH
        - HPOS
        - VPOS
    - HYP (hyphen)
        - WIDTH
        - HPOS
        - VPOS
        - CONTENT (hyphen design)

For each word, alternative spelling can be specified. The number of alternative spellings should be minimized to two because too many incorrect spelling choices can result in an impaired searchability (we expect that the alternative spellings will be searchable).

It is possible to specify the estimated accuracy for character interpretation on both the character and the word level. We have not found any use of accurate character level that justifies its storage. However, we have seen a few cases where word accuracy is estimated at 100%, but where alternative and incorrect spellings yet have been set. In these cases the word accuracy have been used to exclude the alternative spellings.

Hyphens (Hyp) may only be entered as the last character on a line (hyphen included in/as a regular String if they appear anywhere else on the line). The hyphen indicates that the last string continues on the next line. Since the hyphen may have different design (long, short, etc.) then also its appearance should be given in The Content attribute. In at least the U.S. they seem not to use Hyphen to explicitly mark hyphenation which may indicate that not all software supports this. It is important to use Subs_type (for the two String elements) at hyphenating, and complete the hyphenated word so that it is searchable in the text.

If an abbreviation is found in the text, subs_type should, if possible also be used to include the unabridged word. As a result, searches performed on the unabridged word also include hits on the acronym.

Attachements

ALTO elements and attributes (table)
Dictionaries for ALTO