ALTO changes 2.1

Version 21-01-2014

ALTO Board

Table of Contents

1. Introduction

Active use of the ALTO standard over the past years has highlighted areas requiring improvement and new additions. This document lists all eight proposals, and shows the difference between the current standard and the one to be implemented.

Please review the changes in the next two weeks and report any positive or negative aspects that come to mind until February 3rd 2014, considering your own particular usage of the ALTO standard in your projects.

Your feedback will be carefully reviewed and integrated into the final and official version, ALTO 2.1.

2. Tag element reference

There was no reference to a new element group "Tags/Tag" until now. Introducing one will allow additional information to be added to the elements referring to these tag elements.

Backwards compatible: Yes

2.1. The changes

Current Schema ALTO 2.0:	Proposed change:
n/a - not existing yet	<xsd:element name="Tags" type="TagsType"> <xsd:annotation> <xsd:documentation>Tag define properties of additional characteristic. The tags are referenced from related content element on Block or String element by attribute TAGREF via the tag ID. </xsd:documentation> </xsd:annotation> </xsd:element>
n/a - not existing yet	<!-- various element definitions for <Tags> --> <xsd:complexType name="TagsType"> <xsd:sequence> <xsd:choice minOccurs="0" maxOccurs="unbounded"> <xsd:element name="LayoutTag" type="TagType"/> <xsd:element name="StructureTag" type="TagType"/> <xsd:element name="RoleTag" type="TagType"/> <xsd:element name="NamedEntityTag" type="TagType"/> <xsd:element name="OtherTag" type="TagType"/> </xsd:choice> </xsd:sequence> <xsd:annotation> <xsd:documentation>There are following variation of tag types available: LayoutTag – criteria about arrangement or graphical appearance StructureTag – criteria about grouping or formation RoleTag – criteria about function or mission NamedEntityTag – criteria about assignment of terms to their relationship / meaning (NER) OtherTag – criteria about any other characteristic not listed above, the TYPE attribute is intended to be used for classification within those. </xsd:documentation> </xsd:annotation> </xsd:complexType>

	<!-- base type for all tag elements--> <xsd:complexType name="TagType"> <xsd:sequence> <xsd:element name="XmlData" minOccurs="0"> <xsd:annotation> <xsd:documentation xml:lang="en"> The xml data wrapper element XmlData is used to contain XML encoded metadata. The content of an XmlData element can be in any namespace or in no namespace. As permitted by the XML Schema Standard, the processContents attribute value for the metadata in an XmlData is set to “lax”. Therefore, if the source schema and its location are identified by means of an XML schemaLocation attribute, then an XML processor will validate the elements for which it can find declarations. If a source schema is not identified, or cannot be found at the specified schemaLocation, then an XML validator will check for well-formedness, but otherwise skip over the elements appearing in the XmlData element. </xsd:documentation> </xsd:annotation> <xsd:complexType> <xsd:sequence> <xsd:any namespace="##any" processContents="lax" maxOccurs="unbounded"/> </xsd:sequence> </xsd:complexType> </xsd:element> </xsd:sequence> <xsd:attribute name="ID" type="xsd:ID" use="required"/> <xsd:attribute name="TYPE" type="xsd:string" use="optional"> <xsd:annotation> <xsd:documentation>Type can be used to classify and group the information within each tag element type. </xsd:documentation> </xsd:annotation> </xsd:attribute> <xsd:attribute name="LABEL" type="xsd:string" use="required"> <xsd:annotation> <xsd:documentation>Content / information value of the tag.</xsd:documentation> </xsd:annotation> </xsd:attribute> <xsd:attribute name="DESCRIPTION" type="xsd:string" use="optional"> <xsd:annotation> <xsd:documentation>Description text for tag information for clarification.</xsd:documentation> </xsd:annotation> </xsd:attribute> <xsd:attribute name="URI" type="xsd:anyURI" use="optional"> <xsd:annotation> <xsd:documentation>Any URI for authority or description relevant information. </xsd:documentation> </xsd:annotation> </xsd:attribute> </xsd:complexType>

Current Schema ALTO 2.0: BlockType definition	Proposed change:
<xsd:extension base="BlockType"> <xsd:sequence minOccurs="0"> <xsd:element name="TextLine" maxOccurs="unbounded"> <xsd:annotation> <xsd:documentation>A single line of text.</xsd:documentation> </xsd:annotation> <xsd:complexType> <xsd:sequence> <xsd:sequence maxOccurs="unbounded"> <xsd:element name="String" type="StringType"/>	<xsd:extension base="BlockType"> <xsd:sequence minOccurs="0"> <xsd:element name="TextLine" maxOccurs="unbounded"> <xsd:annotation> <xsd:documentation>A single line of text.</xsd:documentation> </xsd:annotation> <xsd:complexType> <xsd:sequence> <xsd:sequence maxOccurs="unbounded"> <xsd:element name="String" type="StringType"/>

<xsd:element name="SP" minOccurs="0"> <xsd:annotation> <xsd:documentation>A white space.</xsd:documentation> </xsd:annotation> <xsd:complexType> <xsd:attribute name="ID" type="xsd:ID" use="optional"/> <xsd:attribute name="WIDTH" type="xsd:float" use="optional"/> <xsd:attribute name="HPOS" type="xsd:float" use="optional"/> <xsd:attribute name="VPOS" type="xsd:float" use="optional"/> </xsd:complexType> </xsd:element> </xsd:sequence> <xsd:element name="HYP" minOccurs="0"> <xsd:annotation> <xsd:documentation>A hyphenation char. Can appear only at the end of a line.</xsd:documentation> </xsd:annotation> <xsd:complexType> <xsd:attribute name="WIDTH" type="xsd:float" use="optional"/> <xsd:attribute name="HPOS" type="xsd:float" use="optional"/> <xsd:attribute name="VPOS" type="xsd:float" use="optional"/> <xsd:attribute name="CONTENT" type="xsd:string" use="required"/> </xsd:complexType> </xsd:element> </xsd:sequence> <xsd:attribute name="ID" type="xsd:ID"/> <xsd:attribute name="STYLEREFS" type="xsd:IDREFS" use="optional"/> <xsd:attribute name="HEIGHT" type="xsd:float" use="required"/> <xsd:attribute name="WIDTH" type="xsd:float" use="required"/> <xsd:attribute name="HPOS" type="xsd:float" use="required"/> <xsd:attribute name="VPOS" type="xsd:float" use="required"/> <xsd:attribute name="BASELINE" type="xsd:float" use="optional"/> ...	<xsd:element name="SP" minOccurs="0"> <xsd:annotation> <xsd:documentation>A white space.</xsd:documentation> </xsd:annotation> <xsd:complexType> <xsd:attribute name="ID" type="xsd:ID" use="optional"/> <xsd:attribute name="WIDTH" type="xsd:float" use="optional"/> <xsd:attribute name="HPOS" type="xsd:float" use="optional"/> <xsd:attribute name="VPOS" type="xsd:float" use="optional"/> </xsd:complexType> </xsd:element> </xsd:sequence> <xsd:element name="HYP" minOccurs="0"> <xsd:annotation> <xsd:documentation>A hyphenation char. Can appear only at the end of a line.</xsd:documentation> </xsd:annotation> <xsd:complexType> <xsd:attribute name="WIDTH" type="xsd:float" use="optional"/> <xsd:attribute name="HPOS" type="xsd:float" use="optional"/> <xsd:attribute name="VPOS" type="xsd:float" use="optional"/> <xsd:attribute name="CONTENT" type="xsd:string" use="required"/> </xsd:complexType> </xsd:element> </xsd:sequence> <xsd:attribute name="ID" type="xsd:ID"/> <xsd:attribute name="STYLEREFS" type="xsd:IDREFS" use="optional"/> <xsd:attribute name="TAGREFS" type="xsd:IDREFS" use="optional"/> <xsd:attribute name="HEIGHT" type="xsd:float" use="required"/> <xsd:attribute name="WIDTH" type="xsd:float" use="required"/> <xsd:attribute name="HPOS" type="xsd:float" use="required"/> <xsd:attribute name="VPOS" type="xsd:float" use="required"/> <xsd:attribute name="BASELINE" type="xsd:float" use="optional"/>

Current Schema ALTO 2.0: StringType definition	Proposed change:
<xsd:complexType name="StringType" mixed="false"> <xsd:annotation> <xsd:documentation>A sequence of chars. Strings are separated by white spaces or hyphenation chars.</xsd:documentation> </xsd:annotation> <xsd:sequence minOccurs="0"> <xsd:element name="ALTERNATIVE" maxOccurs="unbounded"> <xsd:annotation> <xsd:documentation>Any alternative for the word.</xsd:documentation> </xsd:annotation> <xsd:complexType> <xsd:simpleContent> <xsd:extension base="xsd:string"> <xsd:attribute name="PURPOSE" type="xsd:string" use="optional"> <xsd:annotation> <xsd:documentation>Identifies the purpose of the alternative.</xsd:documentation> </xsd:annotation> </xsd:attribute> </xsd:extension> </xsd:simpleContent> </xsd:complexType> </xsd:element>	<xsd:complexType name="StringType" mixed="false"> <xsd:annotation> <xsd:documentation>A sequence of chars. Strings are separated by white spaces or hyphenation chars.</xsd:documentation> </xsd:annotation> <xsd:sequence minOccurs="0"> <xsd:element name="ALTERNATIVE" maxOccurs="unbounded"> <xsd:annotation> <xsd:documentation>Any alternative for the word.</xsd:documentation> </xsd:annotation> <xsd:complexType> <xsd:simpleContent> <xsd:extension base="xsd:string"> <xsd:attribute name="PURPOSE" type="xsd:string" use="optional"> <xsd:annotation> <xsd:documentation>Identifies the purpose of the alternative.</xsd:documentation> </xsd:annotation> </xsd:attribute> </xsd:extension> </xsd:simpleContent> </xsd:complexType> </xsd:element>

</xsd:sequence> <xsd:attribute name="ID" type="xsd:ID" use="optional"/> <xsd:attribute name="STYLEREFS" type="xsd:IDREFS" use="optional"/> <xsd:attribute name="HEIGHT" type="xsd:float" use="optional"/> <xsd:attribute name="WIDTH" type="xsd:float" use="optional"/> <xsd:attribute name="HPOS" type="xsd:float" use="optional"/> <xsd:attribute name="VPOS" type="xsd:float" use="optional"/> ... 	</xsd:sequence> <xsd:attribute name="ID" type="xsd:ID" use="optional"/> <xsd:attribute name="STYLEREFS" type="xsd:IDREFS" use="optional"/> <xsd:attribute name="TAGREFS" type="xsd:IDREFS" use="optional"/> <xsd:attribute name="HEIGHT" type="xsd:float" use="optional"/> <xsd:attribute name="WIDTH" type="xsd:float" use="optional"/> <xsd:attribute name="HPOS" type="xsd:float" use="optional"/> <xsd:attribute name="VPOS" type="xsd:float" use="optional"/> ...

The TextBlockType is based on BlockType and is inheriting the TAGREF attribute by this as well.

2.2. Example

Note: Short extract of ALTO xml to illustrate result of actual schema change.

```
<Tags>
<!-- Structure tags-->
<StructureTag ID="Tag01" TYPE="Logical" LABEL="Title" DESCRIPTION="Title"/>
<StructureTag ID="Tag02" TYPE="Logical" LABEL="RunningTitle" DESCRIPTION="Repeating text on each site"/>
<StructureTag ID="Tag03" TYPE="Logical" LABEL="TOC" DESCRIPTION="Table of content"/>
<StructureTag ID="Tag04" TYPE="Logical" LABEL="PageNumberReference" DESCRIPTION="Reference to page number of other page"/>
<StructureTag ID="Tag05" TYPE="Logical" LABEL="Footnote"/>
<StructureTag ID="Tag06" TYPE="Logical" LABEL="FootnoteReference" DESCRIPTION="Reference to footnote"/>
<StructureTag ID="Tag07" TYPE="Logical" LABEL="Headline" DESCRIPTION="Headline of article" />
<StructureTag ID="Tag08" TYPE="Logical" LABEL="Subheadline" DESCRIPTION="Subheadline of article" />
    <!-- Layout tags-->
<LayoutTag ID="Tag09" LABEL="MusicalScore" DESCRIPTION="Musical notation"/>
<LayoutTag ID="Tag10" LABEL="MathFormula" DESCRIPTION="Mathematical formula"/>
<LayoutTag ID="Tag11" TYPE="Graphic" LABEL="ChemFormula" DESCRIPTION="Chemical formula"/>

    <!-- Named Entity tags -->
       <NamedEntityTag ID="Tag5" TYPE="Person" LABEL="Dr Reynolds"/>
       <NamedEntityTag ID="Tag14" TYPE="Organization" LABEL="Central Kentucky Tobacco Warehouse Company"/>
       <NamedEntityTag ID="Tag17" TYPE="Location" LABEL="Montana"/>
<NamedEntityTag ID="Tag12" TYPE="person" LABEL="John Johnson"/>
<NamedEntityTag ID="Tag13" TYPE="location" LABEL="London"/>
<NamedEntityTag ID="Tag14" TYPE="person" LABEL="Zachariah Jackson">
 <XmlData>
  <myXML>Any XML can be added here.</myXML>
 </XmlData>
</NamedEntityTag>
<NamedEntityTag ID="Tag15" TYPE="person" LABEL="Johann Christian Bach">
   <MADS >
      <authority>
 <name type="personal" authority="naf">
  <namePart>Bach, Johann Christian</namePart>
  <namePart type="date">1735-1782</namepart>
 </name>
 <titleInfo authority="naf">
  <title>Bach, Johann Christian, Symphonies, E ♭ major</title>
  <partNumber>W. C 19b, E ♭ major</partNumber>
 </titleInfo>
   </MADS>
</NamedEntityTag>
<NamedEntityTag ID="Tag16" TYPE="Person" LABEL="James M Bigstaff ">
 <XmlData>
  <MADS xmlns="http://www.loc.gov/mads/">
   <authority>
```

```
      <name type="personal" authority="naf">
        <namePart>Bigstaff, James M</namePart>
        <namePart type="date">1835-1882</namePart>
      </name>
    </authority>
  </MADS>
  <NEC>0.7</NEC>
 </XmlData>
</NamedEntityTag>
<NamedEntityTag ID="Tag17" TYPE="Location" LABEL="Kentucky" URI="http://www.geonames.org/6254925">
 <XmlData>
 <NEC>1.0</NEC>
 <Variants>Kenekuke Kentaki Kentákii Hahoodzo Kèntòki Kentórkị Kentucky Kentucky
  Kentucky suyu Kentucky suyu Kentuki Kentukia Kentukio Kentukis Kentukki Kentukki
  Kéntukki Shitati Khén-thap-kî كنتاكي كنتاكى كنتاكى يۇنتاكى كنتاكى يۇنتاكىكى يۇنڭكىكى كەنتكي
  קענטאקי Κεντάκι Кентаки Кентаки Кентакки Кентуки Кентукки Кентукки Кентукки
  Кентукки Кентукки Кентуккі Кентъки Кинтааки Штат Кентукі Կենտուկի კენტუკი
  केन्टकी കെന്റക്കി കെന്റുക്കി കെന്റാകി ケンタッキー州 肯塔基州</Variants>
 </XmlData>
</NamedEntityTag>
</Tags>
```

Note: Short extract of ALTO xml to illustrate result of actual schema change

```
<TextBlock ID="P7_TB00003" HPOS="76" VPOS="358" WIDTH="870" HEIGHT="32" STYLEREFS="TXT_0 PAR_CENTER" TAGREFS="Tag01" >
   <TextLine ID="P7_TL00003" HPOS="78" VPOS="358" WIDTH="868" HEIGHT="32">
   <String ID="P7_ST00005" HPOS="402" VPOS="358" WIDTH="100" HEIGHT="32" CONTENT="Bach" WC="0.70" CC="5063" TAGREFS="Tag15"/>
   <SP ID="P7_SP00003" HPOS="553" VPOS="358" WIDTH="22"/>
   <String ID="P7_ST00006" HPOS="458" VPOS="358" WIDTH="95" HEIGHT="30" CONTENT="J." WC="0.74" CC="50" TAGREFS="Tag15"/>
   <SP ID="P7_SP00004" HPOS="553" VPOS="358" WIDTH="22"/>
   <String ID="P7_ST00007" HPOS="458" VPOS="358" WIDTH="95" HEIGHT="30" CONTENT="S." WC="0.74" CC="50" TAGREFS="Tag15"/>
   <SP ID="P7_SP00005" HPOS="434" VPOS="358" WIDTH="24"/>
   <String ID="P7_ST00008" HPOS="575" VPOS="358" WIDTH="371" HEIGHT="31" CONTENT="ILLUSTRATIONS" WC="0.88"
CC="0000400500104"/>
   </TextLine>
</TextBlock>
```

3. Consistent data type definition: positions and dimensions (int vs floating values)

The ALTO schema allows different measurement units.

Floating values are not required for pixels, as a pixel is the smallest describable unit.

But while also inch1200 (1/1200 of inch) and 10mm (10th of mm) are possible, integer values are not sufficient for describing the exact position of a point.

For the following two cases, the integer is defined:

```
<xsd:element name="Page" maxOccurs="unbounded">"
 <xsd:attribute name="HEIGHT" type="xsd:int" use="optional"/>
 <xsd:attribute name="WIDTH" type="xsd:int" use="optional"/>
```

and

```
<xsd:complexType name="BlockType">
 <xsd:attribute name="HEIGHT" type="xsd:int" use="required"/>
 <xsd:attribute name="WIDTH" type="xsd:int" use="required"/>
 <xsd:attribute name="HPOS" type="xsd:int" use="required"/>
 <xsd:attribute name="VPOS" type="xsd:int" use="required"/>
```

However, "xsd:float" is pre-defined for all smaller units.

To ensure that each point can be described as precisely as the existing information allows (e.g. digital born) as well as to simplify the standard and make easy to get it, "xsd:float" is used for all HPOS, VPOS, HEIGHT and WIDTH attributes.

Backwards compatible: Yes

Changing xsd:int to xsd:float is backwards compatible. Values even without ".0" are valid!

3.1. The schema changes

Current schema ALTO 2.0:	Proposed schema change:
<xsd:element **name**="Page" **maxOccurs**="unbounded"> <xsd:annotation> <xsd:documentation>One page of a book or journal. </xsd:documentation> </xsd:annotation> [...] <xsd:attribute name="STYLEREFS" type="xsd:IDREFS" use="optional"/> <xsd:attribute name="HEIGHT" type="xsd:int" use="optional"/> <xsd:attribute name="WIDTH" type="xsd:int" use="optional"/> <xsd:attribute name="PHYSICAL_IMG_NR" type="xsd:int" use="required"> <xsd:annotation> <xsd:documentation>The number of the page within the document.</xsd:documentation> </xsd:annotation> </xsd:attribute>	<xsd:element name="Page" maxOccurs="unbounded"> <xsd:annotation> <xsd:documentation>One page of a book or journal. </xsd:documentation> </xsd:annotation> [...] <xsd:attribute name="STYLEREFS" type="xsd:IDREFS" use="optional"/> <xsd:attribute name="HEIGHT" type="xsd:float" use="optional"/> <xsd:attribute name="WIDTH" type="xsd:float" use="optional"/> <xsd:attribute name="PHYSICAL_IMG_NR" type="xsd:int" use="required"> <xsd:annotation> <xsd:documentation>The number of the page within the document.</xsd:documentation> </xsd:annotation> </xsd:attribute>
<xsd:complexType name="BlockType"> <xsd:annotation> <xsd:documentation>Base type for any kind of block on the page.</xsd:documentation> </xsd:annotation> <xsd:sequence minOccurs="0"> <xsd:element name="Shape" type="ShapeType"/> </xsd:sequence> <xsd:attribute name="ID" type="xsd:ID" use="required"/> <xsd:attribute name="STYLEREFS" type="xsd:IDREFS"/> <xsd:attribute name="HEIGHT" type="xsd:int" use="required"/> <xsd:attribute name="WIDTH" type="xsd:int" use="required"/> <xsd:attribute name="HPOS" type="xsd:int" use="required"/> <xsd:attribute name="VPOS" type="xsd:int" use="required"/> <xsd:attribute name="ROTATION" type="xsd:float" use="optional"> <xsd:annotation> <xsd:documentation>Tells the rotation of the block e.g. text or illustration. The value is in degree counterclockwise. </xsd:documentation> </xsd:annotation> </xsd:attribute> <xsd:attribute name="IDNEXT" type="xsd:IDREF" use="optional"> <xsd:annotation> <xsd:documentation>The next block in reading sequence on the page. </xsd:documentation> </xsd:annotation> </xsd:attribute> <xsd:attributeGroup ref="xlink:simpleLink"/> </xsd:complexType>	<xsd:complexType name="BlockType"> <xsd:annotation> <xsd:documentation>Base type for any kind of block on the page.</xsd:documentation> </xsd:annotation> <xsd:sequence minOccurs="0"> <xsd:element name="Shape" type="ShapeType"/> </xsd:sequence> <xsd:attribute name="ID" type="xsd:ID" use="required"/> <xsd:attribute name="STYLEREFS" type="xsd:IDREFS"/> <xsd:attribute name="HEIGHT" type="xsd:float" use="required"/> <xsd:attribute name="WIDTH" type="xsd:float" use="required"/> <xsd:attribute name="HPOS" type="xsd:float" use="required"/> <xsd:attribute name="VPOS" type="xsd:float" use="required"/> <xsd:attribute name="ROTATION" type="xsd:float" use="optional"> <xsd:annotation> <xsd:documentation>Tells the rotation of the block e.g. text or illustration. The value is in degree counterclockwise. </xsd:documentation> </xsd:annotation> </xsd:attribute> <xsd:attribute name="IDNEXT" type="xsd:IDREF" use="optional"> <xsd:annotation> <xsd:documentation>The next block in reading sequence on the page. </xsd:documentation> </xsd:annotation> </xsd:attribute> <xsd:attributeGroup ref="xlink:simpleLink"/> </xsd:complexType>

4. CircleType - annotation change and type definition

A clear definition of the point described by HPOS / VPOS is not included in the current ALTO version. This will be changed, with ALTO 2.1 having an annotation and documentation to define this point.
In addition, a type definition will also be included.

Backwards compatible: Yes (as long as attributes have been provided)

4.1. The changes

Current Schema ALTO 2.0:	Proposed change:
<xsd:complexType name="CircleType"> <xsd:annotation> <xsd:documentation>A circle shape.</xsd:documentation> </xsd:annotation> <xsd:attribute name="HPOS"/> <xsd:attribute name="VPOS"/> <xsd:attribute name="RADIUS"/> </xsd:complexType>	<xsd:complexType name="CircleType"> <xsd:annotation> <xsd:documentation>A circle shape. HPOS and VPOS describe there the center of the circle.</xsd:documentation> </xsd:annotation> <xsd:attribute name="HPOS" type="xsd:float" use="required"/> <xsd:attribute name="VPOS" type="xsd:float" use="required"/> <xsd:attribute name="RADIUS" type="xsd:float" use="required"/> </xsd:complexType>

5. EllipseType - annotation change and type definition

As for the CircleType , there is currently no clear definition of the point described by HPOS / VPOS. ALTO 2.1 will include annotation and documentation for this point, as well as a type definition.

Backwards compatible: Yes (as long as values have been provided)

5.1. The changes

Current schema ALTO 2.0:	Proposed changed schema:
<xsd:complexType name="EllipseType"> <xsd:annotation> <xsd:documentation>An ellipse shape.</xsd:documentation> </xsd:annotation> <xsd:attribute name="HPOS"/> <xsd:attribute name="VPOS"/> <xsd:attribute name="HLENGTH"/> <xsd:attribute name="VLENGTH"/> </xsd:complexType>	<xsd:complexType name="EllipseType"> <xsd:annotation> <xsd:documentation>An ellipse shape. The point described is the center of the shape. HLENGTH and VLENGTH are the width and height of the described ellipse. </xsd:documentation> </xsd:annotation> <xsd:attribute name="HPOS" type="xsd:float" use="required"/> <xsd:attribute name="VPOS" type="xsd:float" use="required"/> <xsd:attribute name="HLENGTH" type="xsd:float" use="required"/> <xsd:attribute name="VLENGTH" type="xsd:float" use="required"/> </xsd:complexType>

6. Type definition for attribute "CONTENT" of HYP element

As for 4. And 5., there is currently no definition of the point described by HPOS / VPOS.

The attribute value "CONTENT for the complexType "StringType" and "HYP" element is undefined.

ALTO 2.1 will include a definition, in order to prevent wrong or erroneous usage.

Backwards compatible: Yes, as no other usage is allowed by XML syntax.

6.1. The changes

Current schema ALTO 2.0:	Proposed schema change:
<xsd:element name="HYP" minOccurs="0"> <xsd:annotation> <xsd:documentation>A hyphenation char. Can appear only at the end of a line. </xsd:documentation> </xsd:annotation> <xsd:complexType> <xsd:attribute name="WIDTH" type="xsd:float" use="optional"/> <xsd:attribute name="HPOS" type="xsd:float" use="optional"/> <xsd:attribute name="VPOS" type="xsd:float" use="optional"/> <xsd:attribute name="CONTENT" use="required"/> </xsd:complexType> </xsd:element>	<xsd:element name="HYP" minOccurs="0"> <xsd:annotation> <xsd:documentation>One or more hyphenation character. Can appear only at the end of a line. </xsd:documentation> </xsd:annotation> <xsd:complexType> <xsd:attribute name="WIDTH" type="xsd:float" use="optional"/> <xsd:attribute name="HPOS" type="xsd:float" use="optional"/> <xsd:attribute name="VPOS" type="xsd:float" use="optional"/> <xsd:attribute name="CONTENT" type="xsd:string" use="required"> </xsd:attribute> </xsd:complexType> </xsd:element>

7. CS attribute to be defined on String

Use case

A CS (Correction Status, boolean) attribute already exists on TextLine element (ALTO LoC V2.0). It specifies whether manual correction has been done or not. This attribute should be available on String too. The Line level is not precise enough as only some words might have manually corrected.

Backwards compatible: Yes

Remarks:

This attribute is part of the "production family" attributes: CS, ILLS, DBTS

To be consistent, it should be defined on TextBlock, TextLine, String with a recommendation:

Always use the higher level available to set the attribute (ie: do not set an attribute on all the sub-elements of a specific level)

7.1. The changes

Current Schema ALTO 2.0:	Proposed change:
<xsd:complexType name="StringType" mixed="false"> ... <xsd:attribute name="CC" type="xsd:string" use="optional"> <xsd:annotation> <xsd:documentation> Confidence level of each character in that string. A list of numbers, one number between 0 (sure) and 9 (unsure) for each character. </xsd:documentation> </xsd:annotation> </xsd:attribute> <xsd:attribute name="LANG" type="xsd:language" use="optional"> <xsd:annotation> <xsd:documentation> Attribute to record language of the string. The language should be recorded at the highest level possible. </xsd:documentation> </xsd:annotation> </xsd:attribute> </xsd:complexType>	<xsd:complexType name="StringType" mixed="false"> ... <xsd:attribute name="CC" type="xsd:string" use="optional"> <xsd:annotation> <xsd:documentation>Confidence level of each character in that string. A list of numbers, one number between 0 (sure) and 9 (unsure) for each character.</xsd:documentation> </xsd:annotation> </xsd:attribute> <xsd:attribute name="LANG" type="xsd:language" use="optional"> <xsd:annotation> <xsd:documentation>Attribute to record language of the string. The language should be recorded at the highest level possible.</xsd:documentation> </xsd:annotation> </xsd:attribute> <xsd:attribute name="CS" type="xsd:boolean" use="optional"> <xsd:attribute> <xsd:annotation> <xsd:documentation>Correction Status. Indicates whether manual correction has been done or not. The correction status should be recorded at the highest level possible (Block, TextLine, String). </xsd:annotation> </xsd:attribute> </xsd:documentation> </xsd:complexType>

<xsd:complexType name="BlockType"> <xsd:annotation> <xsd:documentation>Base type for any kind of block on the page.</xsd:documentation> </xsd:annotation> <xsd:sequence minOccurs="0"> <xsd:element name="Shape" type="ShapeType"/> </xsd:sequence> <xsd:attribute name="ID" type="xsd:ID" use="required"/> <xsd:attribute name="STYLEREFS" type="xsd:IDREFS"/> <xsd:attribute name="TAGREFS" type="xsd:IDREFS" use="optional"/> <xsd:attribute name="HEIGHT" type="xsd:float" use="required"/> <xsd:attribute name="WIDTH" type="xsd:float" use="required"/> <xsd:attribute name="HPOS" type="xsd:float" use="required"/> <xsd:attribute name="VPOS" type="xsd:float" use="required"/> <xsd:attribute name="ROTATION" type="xsd:float" use="optional"> <xsd:annotation> <xsd:documentation>Tells the rotation of the block e.g. text or illustration. The value is in degree counterclockwise.</xsd:documentation> </xsd:annotation> </xsd:attribute> <xsd:attribute name="IDNEXT" type="xsd:IDREF" use="optional"> <xsd:annotation> <xsd:documentation>The next block in reading sequence on the page.</xsd:documentation> </xsd:annotation> </xsd:attribute> <xsd:attributeGroup ref="xlink:simpleLink"/> </xsd:complexType>	<xsd:complexType name="BlockType"> <xsd:annotation> <xsd:documentation>Base type for any kind of block on the page.</xsd:documentation> </xsd:annotation> <xsd:sequence minOccurs="0"> <xsd:element name="Shape" type="ShapeType"/> </xsd:sequence> <xsd:attribute name="ID" type="xsd:ID" use="required"/> <xsd:attribute name="STYLEREFS" type="xsd:IDREFS"/> <xsd:attribute name="TAGREFS" type="xsd:IDREFS" use="optional"/> <xsd:attribute name="HEIGHT" type="xsd:float" use="required"/> <xsd:attribute name="WIDTH" type="xsd:float" use="required"/> <xsd:attribute name="HPOS" type="xsd:float" use="required"/> <xsd:attribute name="VPOS" type="xsd:float" use="required"/> <xsd:attribute name="ROTATION" type="xsd:float" use="optional"> <xsd:annotation> <xsd:documentation>Tells the rotation of the block e.g. text or illustration. The value is in degree counterclockwise.</xsd:documentation> </xsd:annotation> </xsd:attribute> <xsd:attribute name="IDNEXT" type="xsd:IDREF" use="optional"> <xsd:annotation> <xsd:documentation>The next block in reading sequence on the page.</xsd:documentation> </xsd:annotation> </xsd:attribute> <xsd:attribute name="CS" type="xsd:boolean" use="optional"> <xsd:annotation> <xsd:documentation>Correction Status. Indicates whether manual correction has been done or not. The correction status should be recorded at the highest level possible (Block, TextLine, String). </xsd:annotation> </xsd:attribute> </xsd:documentation> <xsd:attributeGroup ref="xlink:simpleLink"/> </xsd:complexType>

8. LANGUAGE attribute on TextLine and String

Use case

The LANGUAGE attribute is defined on the BlockType element. This attribute should be available on TextLine and String too.

Backwards compatible: Yes

8.1. The changes

Current Schema ALTO 2.0:	Proposed change:
<xsd:complexType name="StringType" mixed="false"> <xsd:annotation> <xsd:documentation>A sequence of chars. Strings are separated by white spaces or hyphenation chars.</xsd:documentation> </xsd:annotation> <xsd:sequence minOccurs="0"> <xsd:element name="ALTERNATIVE" maxOccurs="unbounded"> <xsd:annotation> <xsd:documentation>Any alternative for the word.</xsd:documentation> </xsd:annotation> <xsd:complexType> <xsd:simpleContent> <xsd:extension base="xsd:string"> <xsd:attribute name="PURPOSE" type="xsd:string" use="optional"> <xsd:annotation> <xsd:documentation>Identifies the purpose of the alternative.</xsd:documentation> </xsd:annotation> </xsd:attribute> </xsd:extension> </xsd:simpleContent> </xsd:complexType> </xsd:element> </xsd:sequence> <xsd:attribute name="ID" type="xsd:ID" use="optional"/> <xsd:attribute name="STYLEREFS" type="xsd:IDREFS" use="optional"/> <xsd:attribute name="HEIGHT" type="xsd:float" use="optional"/> <xsd:attribute name="WIDTH" type="xsd:float" use="optional"/> <xsd:attribute name="HPOS" type="xsd:float" use="optional"/> <xsd:attribute name="VPOS" type="xsd:float" use="optional"/> <xsd:attribute name="CONTENT" use="required">	<xsd:complexType name="StringType" mixed="false"> <xsd:annotation> <xsd:documentation>A sequence of chars. Strings are separated by white spaces or hyphenation chars.</xsd:documentation> </xsd:annotation> <xsd:sequence minOccurs="0"> <xsd:element name="ALTERNATIVE" maxOccurs="unbounded"> <xsd:annotation> <xsd:documentation>Any alternative for the word.</xsd:documentation> </xsd:annotation> <xsd:complexType> <xsd:simpleContent> <xsd:extension base="xsd:string"> <xsd:attribute name="PURPOSE" type="xsd:string" use="optional"> <xsd:annotation> <xsd:documentation>Identifies the purpose of the alternative.</xsd:documentation> </xsd:annotation> </xsd:attribute> </xsd:extension> </xsd:simpleContent> </xsd:complexType> </xsd:element> </xsd:sequence> <xsd:attribute name="ID" type="xsd:ID" use="optional"/> <xsd:attribute name="STYLEREFS" type="xsd:IDREFS" use="optional"/> <xsd:attribute name="HEIGHT" type="xsd:float" use="optional"/> <xsd:attribute name="WIDTH" type="xsd:float" use="optional"/> <xsd:attribute name="HPOS" type="xsd:float" use="optional"/> <xsd:attribute name="VPOS" type="xsd:float" use="optional"/> <xsd:attribute name="CONTENT" use="required">

<xsd:simpleType> <xsd:restriction base="xsd:string"> <xsd:whiteSpace value="preserve"/> </xsd:restriction> </xsd:simpleType> </xsd:attribute> <xsd:attribute name="STYLE" type="fontStylesType" use="optional"/> <xsd:attribute name="SUBS_TYPE" use="optional"> <xsd:annotation> <xsd:documentation>Type of the substitution (if any).</xsd:documentation> </xsd:annotation> <xsd:simpleType> <xsd:restriction base="xsd:string"> <xsd:enumeration value="HypPart1"/> <xsd:enumeration value="HypPart2"/> <xsd:enumeration value="Abbreviation"/> </xsd:restriction> </xsd:simpleType> </xsd:attribute> <xsd:attribute name="SUBS_CONTENT" type="xsd:string" use="optional"> <xsd:annotation> <xsd:documentation>Content of the substiution.</xsd:documentation> </xsd:annotation> </xsd:attribute> <xsd:attribute name="WC" use="optional"> <xsd:annotation> <xsd:documentation>Word Confidence: Confidence level of the ocr for this string. A value between 0 (unsure) and 1 (sure). </xsd:documentation> </xsd:annotation> <xsd:simpleType> <xsd:restriction base="xsd:float"> <xsd:minInclusive value="0"/> <xsd:maxInclusive value="1"/> </xsd:restriction> </xsd:simpleType> </xsd:attribute> <xsd:attribute name="CC" type="xsd:string" use="optional"> <xsd:annotation> <xsd:documentation>Confidence level of each character in that string. A list of numbers, one number between 0 (sure) and 9 (unsure) for each character.</xsd:documentation> </xsd:annotation> </xsd:attribute> </xsd:complexType>	<xsd:attribute name="LANG" type="xsd:language" use="optional"> <xsd:simpleType> <xsd:restriction base="xsd:string"> <xsd:whiteSpace value="preserve"/> </xsd:restriction> </xsd:simpleType> </xsd:attribute> <xsd:attribute name="STYLE" type="fontStylesType" use="optional"/> <xsd:attribute name="SUBS_TYPE" use="optional"> <xsd:annotation> <xsd:documentation>Type of the substitution (if any).</xsd:documentation> </xsd:annotation> <xsd:simpleType> <xsd:restriction base="xsd:string"> <xsd:enumeration value="HypPart1"/> <xsd:enumeration value="HypPart2"/> <xsd:enumeration value="Abbreviation"/> </xsd:restriction> </xsd:simpleType> </xsd:attribute> <xsd:attribute name="SUBS_CONTENT" type="xsd:string" use="optional"> <xsd:annotation> <xsd:documentation>Content of the substiution.</xsd:documentation> </xsd:annotation> </xsd:attribute> <xsd:attribute name="WC" use="optional"> <xsd:annotation> <xsd:documentation>Word Confidence: Confidence level of the ocr for this string. A value between 0 (unsure) and 1 (sure). </xsd:attribute> <xsd:attribute name="CC" type="xsd:string" use="optional"> <xsd:annotation> <xsd:simpleType> <xsd:restriction base="xsd:float"> <xsd:minInclusive value="0"/> <xsd:maxInclusive value="1"/> </xsd:restriction> </xsd:simpleType> </xsd:attribute> <xsd:attribute name="CC" type="xsd:string" use="optional"> <xsd:annotation> <xsd:documentation>Confidence level of each character in that string. A list of numbers, one number between 0 (sure) and 9 (unsure) for each character.</xsd:documentation> </xsd:annotation> </xsd:attribute> </xsd:complexType>

Current Schema ALTO 2.0:	Proposed change:
<xsd:element name="TextLine" maxOccurs="unbounded"> <xsd:annotation> <xsd:documentation>A single line of text.</xsd:documentation> </xsd:annotation> <xsd:complexType> <xsd:sequence> <xsd:sequence maxOccurs="unbounded"> <xsd:element name="String" type="StringType"/> <xsd:element name="SP" minOccurs="0"> <xsd:annotation> <xsd:documentation>A white space.</xsd:documentation> </xsd:annotation> <xsd:complexType> <xsd:attribute name="ID" type="xsd:ID" use="optional"/> <xsd:attribute name="WIDTH" type="xsd:float" use="optional"/> <xsd:attribute name="HPOS" type="xsd:float" use="optional"/> <xsd:attribute name="VPOS" type="xsd:float" use="optional"/> </xsd:complexType> </xsd:element> </xsd:sequence> <xsd:element name="HYP" minOccurs="0"> <xsd:annotation> <xsd:documentation>A hyphenation char. Can appear only at the end of a line.</xsd:documentation> </xsd:annotation> <xsd:complexType> <xsd:attribute name="WIDTH" type="xsd:float" use="optional"/> <xsd:attribute name="HPOS" type="xsd:float" use="optional"/> <xsd:attribute name="VPOS" type="xsd:float" use="optional"/> <xsd:attribute name="CONTENT" use="required"/> </xsd:complexType> </xsd:element> </xsd:sequence> <xsd:attribute name="ID" type="xsd:ID"/> <xsd:attribute name="STYLEREFS" type="xsd:IDREFS" use="optional"/> <xsd:attribute name="HEIGHT" type="xsd:float" use="required"/> <xsd:attribute name="WIDTH" type="xsd:float" use="required"/> <xsd:attribute name="HPOS" type="xsd:float" use="required"/> <xsd:attribute name="VPOS" type="xsd:float" use="required"/> <xsd:attribute name="BASELINE" type="xsd:float" use="optional"/> <xsd:attribute name="CS" type="xsd:boolean" use="optional"> <xsd:annotation> <xsd:documentation>Correction Status. Indicates whether manual correction has been done or not.</xsd:documentation> </xsd:annotation> </xsd:attribute> </xsd:complexType> </xsd:element>	<xsd:element name="TextLine" maxOccurs="unbounded"> <xsd:annotation> <xsd:documentation>A single line of text.</xsd:documentation> </xsd:annotation> <xsd:complexType> <xsd:sequence> <xsd:sequence maxOccurs="unbounded"> <xsd:element name="String" type="StringType"/> <xsd:element name="SP" minOccurs="0"> <xsd:annotation> <xsd:documentation>A white space.</xsd:documentation> </xsd:annotation> <xsd:complexType> <xsd:attribute name="ID" type="xsd:ID" use="optional"/> <xsd:attribute name="WIDTH" type="xsd:float" use="optional"/> <xsd:attribute name="HPOS" type="xsd:float" use="optional"/> <xsd:attribute name="VPOS" type="xsd:float" use="optional"/> </xsd:complexType> </xsd:element> </xsd:sequence> <xsd:element name="HYP" minOccurs="0"> <xsd:annotation> <xsd:documentation>A hyphenation char. Can appear only at the end of a line.</xsd:documentation> </xsd:annotation> <xsd:complexType> <xsd:attribute name="WIDTH" type="xsd:float" use="optional"/> <xsd:attribute name="HPOS" type="xsd:float" use="optional"/> <xsd:attribute name="VPOS" type="xsd:float" use="optional"/> <xsd:attribute name="CONTENT" use="required"/> </xsd:complexType> </xsd:element> </xsd:sequence> <xsd:attribute name="ID" type="xsd:ID"/> <xsd:attribute name="STYLEREFS" type="xsd:IDREFS" use="optional"/> <xsd:attribute name="HEIGHT" type="xsd:float" use="required"/> <xsd:attribute name="WIDTH" type="xsd:float" use="required"/> <xsd:attribute name="HPOS" type="xsd:float" use="required"/> <xsd:attribute name="VPOS" type="xsd:float" use="required"/> <xsd:attribute name="BASELINE" type="xsd:float" use="optional"/> <xsd:attribute name="CS" type="xsd:boolean" use="optional"> <xsd:attribute name="LANG" type="xsd:language" use="optional"> <xsd:annotation> <xsd:documentation>Correction Status. Indicates whether manual correction has been done or not.</xsd:documentation> </xsd:annotation> </xsd:attribute> </xsd:complexType> </xsd:element>
<xsd:complexType name="TextBlockType"> <xsd:annotation> <xsd:documentation>A block of text.</xsd:documentation> </xsd:annotation> <xsd:complexContent> <xsd:extension base="BlockType"> <xsd:sequence minOccurs="0"> <xsd:element name="TextLine" maxOccurs="unbounded"> ... </xsd:sequence>	<xsd:complexType name="TextBlockType"> <xsd:annotation> <xsd:documentation>A block of text.</xsd:documentation> </xsd:annotation> <xsd:complexContent> <xsd:extension base="BlockType"> <xsd:sequence minOccurs="0"> <xsd:element name="TextLine" maxOccurs="unbounded"> ...

<xsd:attribute name="language" type="xsd:language" use="optional"/> </xsd:extension> </xsd:complexContent> </xsd:complexType>	</xsd:sequence> <xsd:attribute name="language" type="xsd:language" use="optional"> <xsd:annotation> <xsd:documentation>Attribute deprecated. LANG should be used instead.</xsd:documentation> </xsd:annotation> </xsd:attribute> <xsd:attribute name="LANG" type="xsd:language" use="optional"> <xsd:annotation> <xsd:documentation>Attribute to record language of the textblock.</xsd:documentation> </xsd:annotation> </xsd:attribute> </xsd:extension> </xsd:complexContent> </xsd:complexType>

8.2. Example

Note: Short extract of ALTO xml to illustrate result of concrete schema change

```
<TextBlock ID="P7_TB00003" HPOS="76" VPOS="358" WIDTH="870" HEIGHT="32" STYLEREFS="TXT_0 PAR_CENTER" language="EN" LANG="en">
   <TextLine ID="P7_TL00003" HPOS="78" VPOS="358" WIDTH="868" HEIGHT="32"  LANG="en">
   <String ID="P7_ST00005" HPOS="78" VPOS="358" WIDTH="356" HEIGHT="32" CONTENT="RESTORATIONS" WC="0.70" CC="506043800302" LANG="en"/>
   <SP ID="P7_SP00003" HPOS="434" VPOS="390" WIDTH="24"/>
   <String ID="P7_ST00006" HPOS="458" VPOS="360" WIDTH="95" HEIGHT="30" CONTENT="AND" WC="0.74" CC="501"/>
   <SP ID="P7_SP00004" HPOS="553" VPOS="390" WIDTH="22"/>
   <String ID="P7_ST00007" HPOS="575" VPOS="359" WIDTH="371" HEIGHT="31" CONTENT="ILLUSTRATIONS" WC="0.88" CC="0000400500104"/>
   </TextLine>
</TextBlock>
```

9. MeasurementUnit - annotation change and define as mandatory

The definition of the values in the ALTO came under discussion in the NPO project.

There is some concern about the annotation of "default" definition in these cases.

The conclusion is to make it a required value however this will cause a problem on backwards compatibility.

Backwards compatible: No

9.1. The changes

Current schema ALTO 2.0:	Proposed changed schema:
<xsd:element name="MeasurementUnit" minOccurs="0"> <xsd:annotation> <xsd:documentation>All measurement values inside the alto file except fontsize are related to this unit. The default is 1/10 of mm </xsd:documentation> </xsd:annotation> <xsd:simpleType> <xsd:restriction base="xsd:string"> <xsd:enumeration value="pixel"/> <xsd:enumeration value="mm10"/> <xsd:enumeration value="inch1200"/> </xsd:restriction> </xsd:simpleType> </xsd:element>	<xsd:element name="MeasurementUnit" minOccurs="1"> <xsd:annotation> <xsd:documentation> All measurement values inside the alto file are related to this unit, except the font size. Coordinates as being used in HPOS and VPOS are absolute coordinates referring to the upper-left corner of a page. The upper left corner of the page is defined as coordinate (0/0). values meaning: mm10: 1/10th of millimeter inch1200: 1/1200th of inch pixel: 1 pixel The values for pixel will be related to the resolution of the image based on which the layout is described. Incase the original image is not known the scaling factor can be calculated based on total width and height of the image and the according information of the PAGE element. </xsd:documentation> </xsd:annotation> <xsd:simpleType> <xsd:restriction base="xsd:string"> <xsd:enumeration value="pixel"/> <xsd:enumeration value="mm10"/> <xsd:enumeration value="inch1200"/> </xsd:restriction> </xsd:simpleType> </xsd:element>

10. Conclusion

The changes for various annotation, data type definition, as well as making attributes consistent available on TextLine and String level are covering missing information and inconsistency of former schema definition.

The new tag element references implementation will be an improvement to the ALTO standard, making it more widely applicable. It covers a variety of change requests addressed to the ALTO board.

Defining MeasurementUnit as mandatory cause the new version not being backwards-compatible. But as the transformation of ALTO files can easily be performed the importance for correct provided measurement information was voted as more important. The comment with default unit type of inch was declared as not sufficient from ALTO board.