Newspaper digitization NL Sweden: ALTO elements and attributes			**Updated 2013-08-29**			
Element	**Attribute**	**Value in attribute**	**Value in element**	**M/O**	**Repeated?**	**Notes/comments (all examples are from the DD and SAP projects)**
alto	xmlns:xsi	"http://www.w3.org/2001/XMLSchema-instance"				Schemas and namespaces for validating
	xmlns:xlink	"http://www.w3.org/1999/xlink"				
	xmlns	"http://www.loc.gov/standards/alto/ns-v2#"				
	xsi:schemaLocation	"http://www.loc.gov/standards/alto/ns-v2# http://www.kb.se/na				Copy of http://www.loc.gov/standards/alto/alto-v2.0.xsd
Description						Contains information of general character.
Description/MeasurementUnit			string	M	NR	Should be "mm10". Unit for the measurements (1/10 mm)
Description/SouceImageInformation						
Description/SouceImageInformation/FileName			string	M	NR	File name for file that has been OCR recognitioned
Description/OCRProcessing				M	R	Can occur in several instances but should not do so because each file should only contain the result from one page.
	ID	OCR1		M	NR	Attribute is included and contains the same value, "OCR1", in all files.
Description/OCRProcessing/ocrPreProcessingStep						*Covered by MIX and Image Processing. Not included.*
Description/OCRProcessing/ocrProcessingStep				M	NR	The actual OCR interpretation may appear only in one instance (within each OCRPROcessing instance).
Description/OCRProcessing/ocrProcessingStep/processingDateTime			dateTimeType	M	NR	Format w3cdtf: YYYY-MM-DD
Description/OCRProcessing/ocrProcessingStep/processingAgency			string	M	NR	Constant value: "Riksarkivet/MKC"
Description/OCRProcessing/ocrProcessingStep/processingStepSettings			string	M	NR	This is a textual description of the OCR settings used. Constant value: "Internal layout; Normal mode; Text type Normal"
Description/OCRProcessing/ocrProcessingStep/processingSoftware				M	NR	
Description/OCRProcessing/ocrProcessingStep/processingSoftware/softwareCreator			string	M	NR	Constant value: "Zissor AS"
Description/OCRProcessing/ocrProcessingStep/processingSoftware/softwareName			string	M	NR	Example: "ContentOCR-10". Probably constant value, although the 10 refers to the version of the Abbyy engine, so if we decide to upgrade to version 11 of the engine, this might change to "ContentOCR-11"
Description/OCRProcessing/ocrProcessingStep/processingSoftware/softwareVersion			string	M	NR	Example: "1.0.4"
Description/OCRProcessing/ocrProcessingStep/processingSoftware/applicationDescription			string	MA		Description of the key features of the software. Used for, for example Non-commercial software. Constant value: "OCR Component for Zissor Content Conversion System"
Description/OCRProcessing/postProcessingStep				MA	R	Mainly applicable if OCR interpretation is followed by a manual step for eg segmentation. Can be repeated, several steps may be performed
Description/OCRProcessing/postProcessingStep/processingDateTime			dateTimeType	M	NR	Format w3cdtf: YYYY-MM-DD
Description/OCRProcessing/postProcessingStep/processingAgency			string	M	NR	Constant value: "Riksarkivet/MKC"
Description/OCRProcessing/postProcessingStep/processingStepSettings			string	M	NR	Values from Riksarkivet/MKC will be presented later

Description/OCRProcessing/postProcessingStep/processingSoftware			string	M	NR	Reference. Used by PROCESSING
Description/OCRProcessing/postProcessingStep/processingSoftware/softwareCreator			string	MA	NR	Constant value: "Zissor AS"
Description/OCRProcessing/postProcessingStep/processingSoftware/softwareName			string	MA	NR	Constant value: "MediaArticleAnalyzerSeg"
Description/OCRProcessing/postProcessingStep/processingSoftware/softwareVersion			string	MA	NR	Example: "1.0.7"
Description/OCRProcessing/postProcessingStep/processingSoftware/applicationDescription			string	MA	NR	Description of the key features of the software. Used for, for example Non-commercial software. Constant value: "Segmentation Component for Zissor Content Conversion System"
Styles						
Styles/TextStyle					R	Used to define the text size. Repeated for each text size that can be found in the text. Should only be integers, although floating is allowed
	ID	style1, style2, etc.			NR	Serial number preceded by the prefix "style"
	FONTSIZE				NR	Narrow to whole and half numbers. Examples: "22", "7", "float"
	FONTFAMILY	Times New Roman; Arial; Courier New			NR	
	FONTSTYLE	bold italics underline; bold italics; bold underline; italics underline; italics; underline; bold			NR	Values according to alto schema restrictions. Example: "bold", "bold italics"
Layout						
Layout/Page				M	R	An ALTO file can contain multiple pages. In the SAP and Digidaily projects one ALTO file will typically contain only one page.
	ID	PAGE1,PAGE2, etc.		M		Serial number preceded by the prefix "PAGE". Normally only one page = "PAGE1"
	HEIGHT	int		M		Measured in Measurement Unit (1/10 mm)
	WIDTH	int		M		
	PHYSICAL_IMG_NR	int		M		The page number in the publication. Should be the same as the page sequence number
	PRINTED_IMG_NR	string		O		The printed page number. Desirable but can not be extracted automatically, so not included in DD and SAP.
	PROCESSING	IDREF		MA		Link to the Description/OCRProcessing entry for this page. In DD and SAP projects there is only one entry so the value will always be set to the ID of that entry, "OCR1".
	QUALITY	list of standard		O		Not used in DD and SAP
	QUALITY_DETAIL	string		O		Free text for extended description of the physical originals quality. Should be as controlled as possible, procedures must be developed
	POSITION	list of standard		O		left, right etc. Not used in DD and SAP

	ACCURACY	float		O		OCR, in %. Not used in DD and SAP
	PC	float		MA		The value is limited to the range 0-1. Example: "0,97"
						(Both ACCURACY and PC must be specified if the software supports it. If this is not the case, only the supported attribute should be specified)
Layout/Page/TopMargin						
	HEIGHT	int		M		
	WIDTH	int		M		
	VPOS	int		M		
	HPOS	int		M		
Layout/Page/BottomMargin						
	HEIGHT	int		M		
	WIDTH	int		M		
	VPOS	int		M		
	HPOS	int		M		
Layout/Page/LeftMargin						
	HEIGHT	int		M		
	WIDTH	int		M		
	VPOS	int		M		
	HPOS	int		M		
Layout/Page/RightMargin						
	HEIGHT	int		M		
	WIDTH	int		M		
	VPOS	int		M		
	HPOS	int		M		
Layout/Page/PrintSpace						Normally only one print per page and file
	ID	"PRINTSPACE1"		M		Serial number preceded by the prefix "PRINTSPACE"
	HEIGHT	int		M		
	WIDTH	int		M		
	HPOS	int		M		Horizontal position
	VPOS	int		M		Vertical position
Layout/Page/PrintSpace/ComposedBlock				MA	R	Identified article.
	ID	"ARTICLE1", "ARTICLE2", etc.		M	NR	Serial number preceded by the prefix "ARTICLE"
	HEIGHT	int		M	NR	
	WIDTH	int		M	NR	
	HPOS	int		M	NR	
	VPOS	int		M	NR	
	TYPE			O	NR	Will not be used. Changed from Mandatory to Optional
Layout/Page/PrintSpace/ComposedBlock/TextBlock				MA	R	Contiguous blocks of text, at least one per page (if the page has text).
	language	[language code]		M	NR	A three-digit code according to ISO639-2b. Example: "eng", "swe". If not possible to identify: "und"

	ID	"ZONE1", "ZONE2", etc.		M	NR	Serial number preceded by the prefix "ZONE".
	HEIGHT	int		M	NR	
	WIDTH	int		M	NR	
	HPOS	int		M	NR	
	VPOS	int		M	NR	
	ROTATION	float		MA	NR	Stated if the text block is leaning and not straightened up. The value is set as degrees counterclockwise
Layout/Page/PrintSpace/ComposedBlock/TextBlock/TextLine				MA	R	one line of text
	ID	"Line1", "Line2", etc.		M	NR	Serial number preceded by the prefix "Line"
	HEIGHT	int		M	NR	
	WIDTH	int		M	NR	
	HPOS	int		M	NR	
	VPOS	int		M	NR	
Layout/Page/PrintSpace/ComposedBlock/TextBlock/TextLine/String				MA	R	One word, separated by spaces or hyphen
	ID	"STR1", "STR2", etc.		M	NR	Serial number preceded by the prefix "STR"
	HEIGHT	int		M	NR	
	WIDTH	int		M	NR	
	HPOS	int		M	NR	
	VPOS	int		M	NR	
	CONTENT	string		M	NR	The content (the word) in the string
	STYLEREFS	IDREF		MA		Refer to the Text Style
	SUBS_TYPE	enumeration		MA		Contains the whole word in hyphenation or abbreviation (HypPart1, HypPart2 and Abbreviation)
	WC	float		MA		Word confidence
Layout/Page/PrintSpace/ComposedBlock/TextBlock/TextLine/String/ALTERNATIVE			string	O	R	Includes alternative spellings that the OCR interpreter can generate. Is not used in DD and SAP
Layout/Page/PrintSpace/ComposedBlock/TextBlock/TextLine/SP				MA	R	Blanc spaces
	ID	"SP1", "SP2", etc.		M	NR	Serial number preceded by the prefix "SP"
	WIDTH	int		M	NR	
	HPOS	int		M	NR	
	VPOS	int		M	NR	
Layout/Page/PrintSpace/ComposedBlock/TextBlock/TextLine/HYP				MA	NR	Hyphens, can only occur at the end of a line
	WIDTH	int		O	NR	Not used in DD and SAP. Comment from Zissor: "By default, the actual hyphen character itself is not saved from the OCR. This is because it is usually not needed in the output, just the information that there is a hyphen. However, there is an option in the OCR program to save the hyphen character with the OCR text, in which case these fields will be populated, but the hyphen character will then also be included as a character in the text ouput".
	HPOS	int		O	NR	
	VPOS	int		O	NR	
	CONTENT	string		M	NR	Hyphens may look different.

Projekt DD and SAP: Dictionaries for ALTO				Updated 2013-08-29		
Container	**Element name**	**Attribute**	**Value in attribute**	**Agreed value in element**	**Reference**	**Comments**
ocrProcessingStep						
	processingDateTime			yyyy-mm-dd	w3cdtf	
	processingAgency			Riksarkivet/MKC		
	processingStepSettings			Internal layout; Normal mode; Text type Normal		
	softwareCreator			Zissor AS		
	softwareName			ContentOCR-10	[Unique list to be expanded as needed]	
	softwareVersion			1.0.4	[Unique list to be expanded as needed]	
	applicationDescription			OCR Component for Zissor Content Conversion System		
postProcessingStep						
	processingDateTime			yyyy-mm-dd	w3cdtf	
	processingAgency			Riksarkvet/MKC		
	processingStepSettings			Standard newspaper	[Unique list to be expanded as needed]	Value list will be presented later by Riksarkivet/MKC
	softwareCreator			Zissor AS		
	softwareName			MediaArticleAnalyzerSeg		
	softwareVersion			1.0.7	[Unique list to be expanded as needed]	
	applicationDescription			Segmentation Component for Zissor Content Conversion System		
Styles						
	TextStyle	FONTFAMILY	Times New Roman			
			Arial			
			Courier New			
		FONTSTYLE	bold italics underline			
			bold italics			
			bold underline			
			italics underline			
			italics			
			underline			
			bold			

Layout						
	TextBlock	language	[language code]		ISO639-2b	