National Library of Sweden

Datum 2014-06-18

Metadata for digitized newspapers in Project SAP
Specification documents

Version 2.4
with updated instructions for missing page, etc. (in English)

Contains:

Packaging METS: Metadata for digitized newspapers in project ”Digitalisering och tillgängliggörande av den svensk-amerikanska pressen (SAP)”
Version 1.1 (corr. 2014-06-18)

Specification for using ALTO in newspaper digitization projects (Digidaily and SAP) at Kungl. biblioteket
Version 2.0 (2013-08-29)

National Library
of Sweden

Packaging METS: Metadata for digitized newspapers in project

”Digitalisering och tillgängliggörande av den svensk-amerikanska pressen (SAP)”

Manual

KB – Riksarkivet/MKC

Version 1.1

(corr. 2014-06-18)

(Added updated attachment: Instruction for missing page, missing issue, missing supplement, missing section, missing newsbill)

Innehåll:

Introduction

This document provides a comprehensive and general description of how the digital files in the project SAP are packaged and described using METS with several standards.

Submission Information Package (SIP)

The result of digitization is packaged in several "delivery packages" . Submission Information Package (SIP). Each SIP will correspond to a newspaper issue. Each SIP will contain all the files that make up a digital image of a newspaper issue.
A newspaper issue is defined as an edition of a newspaper as it was released a specific day. Example: Svenska socialisten November 1, 1905

Each SIP is created, based on the specifications for newspaper digitization project Digidaily that has been developed in cooperation between the Media Conversion Center (MKC), the National Library ((KB) (Kungl. biblioteket in Swedish)) and the National Archives ((RA) (Riksarkivet in Swedish)). Some modifications had been done to suit the SAP project. The specifications are based on KB's basic profile for METS, which in turn is based on SWEIPB - a basic profile developed by the National Archives and others archives and KB together.
The metadata standards used in the packaging are: METS, PREMIS, MODS, MIX. Text files with tagging in ALTO are also included in the packages. More standards may be appropriate**.**

Datamodel

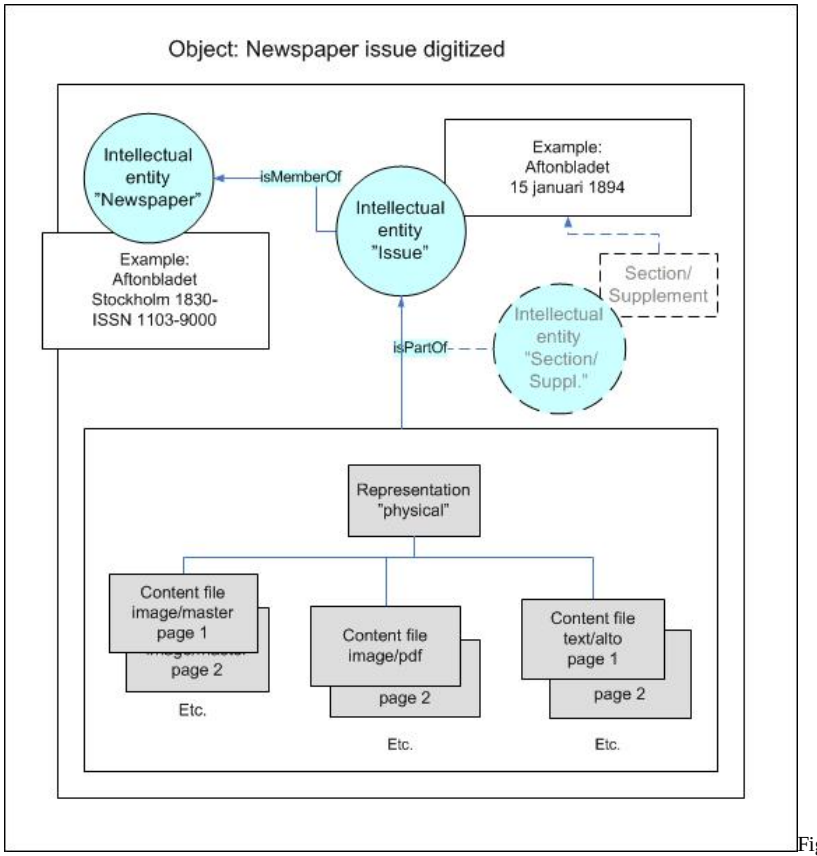


Figure 1

Figure 1 show an issue of a digitized printed newspaper, illustrated based on the PREMIS data model which is composed of five entities: "Intellectual", "Object", "Event", "Agent" and "Rights".
The illustration exemplifies the relationship between the "Intellectual entities" (intellectual entity) and "Objects" (object).

An intellectual entity is defined in PREMIS as a collection of content that can be treated and described as an intellectual entity. In this case, we have several entities that can be described as their own intellectual entities. First and foremost, the newspaper issue Aftonbladet 1894-01-15 (the magazine that it was released on that date), which is in turn an issue of many in Aftonbladet (1830 -). A newspaper issue can be divided into several subordinate units - Section, Supplement, Article, Image, etc. all constitute intellectual entities. (The latter two categories, article and image, but will not be addressed in this project).

Metadata about these intellectual entities, called bibliographic (descriptive) metadata are recorded in the METS document into one or more <dmdSec>.

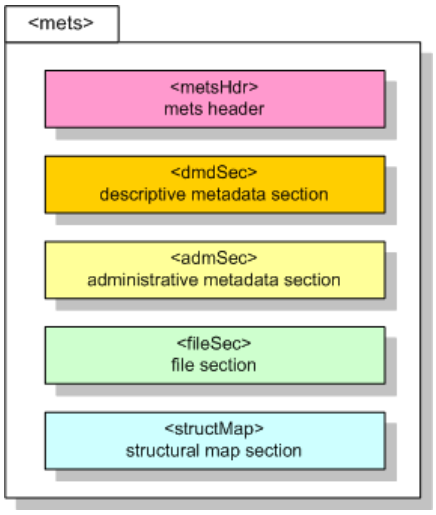
The newspaper issue is represented by a set of image files (a type of object). Each page of the newspaper issue may be mapped separately and saved in different formats for different purposes, such as master files for long term storage, ALTO tagged files for further text processing, pdf-files, etc.. Administrative (structural) metadata in METS <amdSec> and <structMap> keeps track of which the files are, their properties and relations.

A "representation" in PREMIS is another type of object. A representation is the collection of files and administrative metadata required for a complete and understandable representation of an intellectual entity**.**

METS

METS stands for Metadata Encoding and Transmission Standard and is used to encode and pack up the metadata (and data) to digital objects. METS is a flexible standard; the encoding is done in XML and may also include other metadata standards.

For each resource that is packaged in a package at least one METS document (an xml file with the METS-codes) is written. A METS document can be divided into up to seven sections, but in this project, only the following five sections will be used:



Element and attribute in METS

<mets>

A METS document is created for each newspaper issue. A newspaper issue is defined by the specific date when a newspaper title was published. A newspaper issue can be divided into sections, which can have attachments and newsbills, etc.

- For information about ”xmlns” and ”xsi:schemaLocation” Se also Attachements. Metadataelement och attribute i METS…
- Should contain the attributes OBJID, PROFILE, TYPE and LABEL.
- OBJID is the same as the packet identifier and must be unique in its context, ie. in MKC and in contact between MKC and KB.
- In PROFILE is the profile named that the METS-file follows: http://www.kb.se/namespace/mets/kbse_mets_profile_001.xml
- LABEL should contain the newspapers title (title) and date / number (date / number) from dmdSec / @ LABEL = Primary / ... / mods: title info / mods: title. Example: Aftonbladet 1851-12-04

<metsHdr>

The metsHdr element contains information on the METS document itself.
Should include: the CREATEDATE attribute and the elements agent, altRecordID and metsDocumentID.

- CREATEDATE should include date and time when the METS-document was created.
- RECORDSTATUS - write here, where appropriate, any of the values REPLACEMENT or SUPPLEMENT if the package will either replace or supplement a previously delivered packet.
- The element <agent> repeated for the institutions who create and receive a METS document:
  <agent ROLE=”CREATOR” TYPE=”ORGANIZATION”>
  <name>Riksarkivet/MKC</name>
  <note>http://id.kb.se/organisations/SE2021001074-MKC</note>
  </agent>
  <agent ROLE=”ARCHIVIST” TYPE=”ORGANIZATION”>
  <name>Kung. biblioteket</name>
  <note>http://id.kb.se/organisations/SE2021001710</note>
  </agent>
  Repeat this agent for each archiving organization.
- The <altRecordID> element is repeated 3 times with different values in the TYPE attribute:
  - <altRecordID TYPE=”DELIVERYTYPE”> always with the value “AGREEMENT”
  - <altRecordID TYPE=”DELIVERYSPECIFICATION”> always with the value http://www.kb.se/namespace/digark/deliveryspecification/agreement/sap/”
  - <altRecordID TYPE=”SUBMISSIONAGREEMENT”> always with the value “http://www.kb.se/namespace/digark/submissionagreement/DNR_122-KB_270-2013/”

  <metsDocumentID> contains the name of the METS-file. See also Attachments: File naming convention.

Descriptive Metadata Section <dmdSec>

The element is a container for descriptive metadata about the intellectual entity associated with the items that are included in the package. In this case, the metadata about the newspaper and the issue depicted. Descriptive metadata will be drawn from several sources: LIBRIS database, Signe, and the digitization process. Details on which metadata to be in here, see the section on MODS.

- There should always be one dmdSec with metadata in MODS for the newspaper issue. This dmdSec should contain the element mdWrap with attribute LABEL = "Primary" according to the dictionary vcDmdSec_LABEL_kbse.
- There should always be one dmdSec with metadata in MODS for supplier and publisher. This dmdSec should contain the element mdWrap with attribute LABEL = "Local", according to the dictionary vcDmdSec_LABEL_kbse.

- If the accompanying appendices or sections of the newspaper issue are to be described in MODS, these should have their own dmdSec.
- Each dmdSec is numbered sequentially from dmdSec001, dmdSec002, etc. These numbers are recorded as values in the attribute ID. These IDs are used as references in the METS file, eg from <structMap >.
- Metadata in XML is encapsulated in element <xmlData> under <mdWrap >.
- The attribute MDTYPE i <mdWrap> with the value ”MODS” is mandatory for all dmdSec.

Administrative Metadata Section <amdSec>

The element make up a container for administrative metadata about objects (File or representation) and the events, agents and rights associated with the objects. With administrative metadata means technical metadata from image capture, structural metadata that describes relationships between data files, what happened to the files in the work process, restrictions when using, etc.

- Each METS-document include one, and only one, <amdSec>.
- The attribute ID is mandatory with the value ”amdSec001”.
- Contains several elements <techMD>.

Technical Metadata <techMD>

The element is a container for the technical metadata about the files included in the package. Metadata is retrieved from processes with image capture. The metadata Standard is PREMIS with the addition of MIX. Details on which metadata can be found under subsections PREMIS and MIX.

- <techMD> repeated for each objectv (datafile).
- One <techMD> is also created for the representation.
- In The attribute ID the elements are numbeered according to the pattern techMD001, techMD002, etc. These ID are used as references from <fileSec> and <structMap>.
- Metadata in xml is encapsulated in the element <xmlData> under <mdWrap>.
- The attribute MDTYPE i <mdWrap> with the value ”PREMIS:OBJECT” is mandatory in all techMD.

File Section <fileSec>

In the file-section (fileSec) all the files in the package are listed, with file names and details of the size, checksum, etc.
A degree of duplication of information is thus here because some information is also present as PREMIS metadata in <techMD>.
Here also referred to metadata in <techMD>.

- METS-documentet should contain one, and only one, <fileSec>-element.
- The attribute ID with the value fileSec001 is mandatory.
- Should contain at least one <fileGrp>-element, one for each filetype that is included in the package. These are numbered according to pattern fileGrp001, fileGrp002, etc.

Each <fileGrp> should contain the attribute USE who is naming the group according to glossary”vcUSE_kbse”. Se Attachements Metadataelement och attribute i METS… (tabell).
Example: Value ”image/master” indicates that the element contains references to image files for archiving.

- The element <file>, which is subordinated to <fileGrp>, is repeated for each file included in this package. A number of attributes are here mandatory:
  - ID - Contains a numeric value, running from 1, preceded by the prefix "file". Example ” file1”, ”file2”, ”file3”, etc.
  - USE - repeat the value specified in the related <fileGrp>.
  - MIMETYPE - enter value according to http://www.iana.org/assignments/media-types/..
  - SIZE – enter same value as in premis <size> in corresponding mets <techMD>.
  - CREATED - specify when the file was created. Repeat timing for example from mix <dateTimeCreated>. written as YYYY-MM-DD'T'HH:mm:ss+01:00.
  - ADMID - enter ID-value for the <techMD> who Contains metadata about the file, eg. ”techMD002”
  - CHECKSUM – enter same value as in premis <messageDigest> in corresponding mets <techMD>. If several checksums appear in premis, enter the last.
  - CHECKSUMTYPE - enter value ”MD5”, ie. same value as in premis <messageDigestAlgorithm> in corresponding mets <techMD>.
- Under <file> the element <Flocat> is included with a reference to the name of the file. Mandatory attributes are:
  - LOCTYPE – always with the value ”URL”
  - xlink:type - always with the value ”simple”
  - xlink:href - specify the file name preceded by the prefix ”file:”, eg ”file:bib4112678_18760203_1_24_8_m.jp2”. Se also Attachements: File name convention.

Structure Map <structMap>

A "structural map" describes how the data files are related to each other so that a system can reconstruct and provide the digital objects.

- The METS-document should contain a <structMap> which describes the "physical" structure.
- The attribute ID with the value ”structMap001” is mandatory.
- The attribute TYPE should be ”physical” (value from vcStructMap_TYPE_kbse).
- <structMap> are divided into a number of which in turn can be divided into one or more . This process creates a hierarchical structure that shows how the data files are related to each other.
- Mandatory attributes in are:
  - ID – ongoing values according to the pattern div001, div002, etc.
  - TYPE - Each div should be named in TYPE with a value from the dictionary vcDiv_TYPE_kbse. See examples and suggestions on the structure below:
  - LABEL - Can be used for example when a page, or an entire newspaper issue, is missing and has been replaced by a "placeholder" with explanatory text. Values from vcDiv_LABEL_kbse.
  - ORDER - the order in which the file will be displayed in. Values: 1, 2, 3, etc.
  - ORDERLABEL - Optional. Text fields for the page numbering as it appears in the resource.

- DMDID - enter the ID value for the <dmdSec> that contain bibliographic metadata about the things the div element refers to. Example: enter the ID of the dmdSec with LABEL = "Primary". For div TYPE = "supplement" enter the ID of the dmdSec that describes the Supplement, etc.
- A <fptr> under is repeated for each data file. In The attribute FILEID the same value as in the corresponding file/@ID. Example "file1", "file2", "file3", etc.

The structure folder is divided like this with the help of and The attribute TYPE. Example 1:

1. Digitized pages (one jpg and one pdf of each) + one file with ”performance data”:

```
<structMap>
   <div TYPE=”files”> Acts as a "wrapper" around the others
         <div TYPE=”issue”> Highest level for the newspaper issue
               <div TYPE=”page” ORDER=”1”> Here are the references listed for the files on page 1
                     <fptr FILEID=””>Reference to jpg-file for page 1
                     <fptr FILEID=””>Reference to alto-file for page 1
                     <fptr FILEID=””>Reference to pdf-file for page 1
               </div>
               <div TYPE=”page” ORDER=”2”> Here are the references listed for the files on page 2
                     <fptr FILEID=””> Reference to jpg-fil for page 2
                     <fptr FILEID=””> Reference to alto-fil for page 2
                     <fptr FILEID=””>Reference to pdf-file for page 2
               </div>          etc.

         </div>
         <div TYPE=”performance”>On the same level as ”issue”
               <fptr FILEID=””>
         </div>
   </div>
</structMap>
```

Example 2 (see next page):

2. Digitized pages divided in sections + attachment + news bills etc.

```
<structMap>
   <div TYPE=”files”>
         <div TYPE=”issue”> Highest level for the newspaper issue
               <div TYPE=”section”>divided into sections, repeat if more.
                     <div TYPE=”page” ORDER=”1”> Here are the references listed to the files for
                     page 1
                           <fptr FILEID=””> Reference to jpg-fil for page 1
                           <fptr FILEID=””> Reference to alto-fil for page 1
                           <fptr FILEID=””>Reference to pdf-file for page 1
                     </div>
                     <div TYPE=”page” ORDER=”2”> Here are the references listed to the files for
page 2
                           <fptr FILEID=””> Reference to jpg-fil for page 2
                           <fptr FILEID=””> Reference to alto-fil for page 2
                           <fptr FILEID=””>Reference to pdf-file for page 2
                     </div>          etc.
               </div>
               <div TYPE=”supplement”>
                     <div TYPE=”page” ORDER=”1”> Here are the references listed to the files for
page 1
                           <fptr FILEID=””> Reference to jpg-fil for page 1
                           <fptr FILEID=””> Reference to alto-fil for page 1
                           <fptr FILEID=””>Reference to pdf-file for page 1
                     </div>
                     <div TYPE=”page” ORDER=”2”> Here are the references listed to the files for
page 2
                           <fptr FILEID=””> Reference to jpg-fil for page 2
                           <fptr FILEID=””> Reference to alto-fil for page 2
                           <fptr FILEID=””>Reference to pdf-file for page 2
                     </div>          etc.
               </div>
               <div TYPE=”newsbill”>
                     <fptr FILEID=”Reference to jpg-file for newsbill”>
                     <fptr FILEID=”Reference to alto-file for newsbill”>
                     <fptr FILEID=”Reference to pdf-file for newsbill”>
               </div>

         </div>
         <div TYPE=”performance”> On the same level as ”issue”
               <fptr FILEID=””>
         </div>
   </div>
</structMap>
```

MODS

MODS (Metadata Object Description Schema) are an XML scheme for bibliographic metadata that can be used to describe many different types of objects. It is primarily in libraries you encounter MODS. It is often used as an "exchange format" to and from the directory entries in the traditional library format MARC 21, but also to create new descriptions of objects held in libraries. Responsible for the content and development of the standard is the Network Development and MARC Standards Office at the Library of Congress.

Newspaper issue and “related units”

Are recorded in dmdSec/@ID=”dmdSec001”/mdWrap/@LABEL=”Primary”

the MODS-section always begins with metadata about the physical issue but with the addition that the package contains a digital version, either of a hard copy or a microfilmed copy.
The description also includes the registration of an issue of "related entities". This is done in the container <relatedItem> repeated for each type of relationship. Some <relatedItem> is mandatory, other may apply, if necessary.

Element and attributes in MODS

<mods>

<mods> is registred under mets/dmdSec/mdWrap/@MDTYPE=”MODS”/xmlData.

Metadata for the newspaper issue:

typeOfResource

- with the value ”text”.

genre

- with the value ”issue”.
- with attribute authority=”marcgt” (stands for marc genre term)

titleInfo

Includes the elements:

- title
  - with a constructed title for the newspaper issue - same as in <mets@LABEL>, se <mets>

originInfo

- Contains information on when the newspaper issue was issued and whether it belongs to an "edition". The information herein is always the analog original.

Including elements:

- dateIssued – Here a date are composed according to the pattern yyyy-mm-dd.
  - with the attribute encoding=”w3cdtf”.

    - If dating is missing or incorrect in the original add the attribute qualifier, with the value "inferred". As the value of <dateIssued> an assumed date are written down.
- edition - value according to controlled list. Se Attachements File naming convention for more information

physicalDescription

Includes the elements:

- digitalOrigin
  Some of the following values:
    - ”reformatted digital” if the digitization is made from printed original.
    - ”digitized microfilm” if the digitization is made from microfilm.
- note
    - with The attribute type=”reproduction”.
      Value: Digital reproduktion: Stockholm : Riksarkivet/MKC i samarbete med Kungl. biblioteket, yyyy (the year in which the digitization was made).
    - with the attribute type=”script”.
      Chose value according to list (vcScriptType_kbse).

<relatedItem> = Host publication

- with the attribute TYPE=”host”
  Mandatory information for all packages. By "host publication" means the newspaper that the newspaper issue is a part of. Metadata about the host publication will be taken from the LIBRIS database. This is done by KB before digitization begins.

  Here you can find among others, the newspapers title, LIBRIS id, ISSN, if any, and the number and / or date of designation that along with the title identifies the number.
- Recorded within the same <dmdSec> as the newspaper issue otherwise.

Includes the elements:

titleInfo

Includes the elements:

- title
    - with the newpapers title in plain language

genre

- with value ”newspaper”
- with attribute authority=”marcgt”

originInfo

includes the elements:

- dateIssued
    - Value: a date in the form yyyy-mm-dd
    - With the attributes encoding=”w3cdtf”, and point=”start” for the newspapers start date.

- dateIssued repeated if enddate exist, attribute point=”end”.

language

Can be repeated if the newspaper contains multiple languages which are equally "dominant" for the content. Includes element:

- languageTerm.
  - Value: letter code according to iso639-2b, t.ex. ”swe”
  - With attributes type=”code” and encoding=”iso639-2b”

identifier

Specifies the URI that identifies the newspaper in LIBRIS database. If the newspaper has an ISSN number, this is indicated in a separate <Identifier>.

- with attribute type=”uri” för LIBRIS-id.
- <identifier> optionally repeated with ISSN (attribute type=”issn”).

Part

- Specifies the numbering and / or date of the newspaper issue.
  In most cases, reference to the date is doubled and look exactly the same as in mods / @ TYPE = issue / original info / dateIssued.
  Including elements:
  - detail
    - with the attribute type=”issue”
      Include the element:
      - number
        Specifies the numbering of the newspaper if one exists, ie printed numbering or other denomination (eg Sample number) on the newspaper issue. (Note. The number of the microfilm in relatedItem / @ TYPE = 'original").
  - date
    Date for the issue.
    - Write date according to pattern yyyy-mm-dd, example: 1885-12-10.
    - Always the attribute encoding=”w3cdtf”.
    - If the date is missing in the original, also add The attribute qualifier, with the value ”inferred”. As value in <dateIssued> an assumed date is written. As the value of <dateIssued> an assumed date is written.

note

- with the attribute type=”date”
  Here any remark about missing or incorrect date is written.
- with The attribute type=”numbering”
  Here any remark about missing or incorrect numbering.

<relatedItem> = Project

- with the attribute type=”host”
  Here any remark about missing or incorrect date is written.

- With the attribute type = "numbering" Mandatory information for all packages. Metadata about the project will be taken from the LIBRIS database which is done by KB before digitization begins.

Includes the elements:

genre

- with the value ”project”

identifier

Specifies the URI that identifies the project in the LIBRIS database.

- with attribute type=”uri” for LIBRIS-id.

titleInfo

Includes the elements:

- title, with the projects title in plain text

The code looks like this in all cases:

```
<mods:relatedItem type="host">
        <mods:genre>project</mods:genre>
        <mods:titleInfo>
          <mods:title> Digitalisering och tillgängliggörande av den svensk-amerikanska pressen</mods:title>
        </mods:titleInfo>
        <mods:identifier type="uri"> http://libris.kb.se/resource/bib/13983307</mods:identifier>
      </mods:relatedItem>
```

<relatedItem> = microfilm as original

- with attribute TYPE=”original”
  Mandatory if applicable. Here fits information on the original (microfilm) to the digital reproduction.

Includes the elements:

identifier

- Value: the microfilms number.
- attribute type=”reel number”.

physicalDescription

Includes the elements:

- form
  - with attribute authority=”marcform” and value: microfilm.
-

location

Includes:

- holdingSimple
  Includes the element:

- copyInformation, as Includes:
    - note, without attribute.
      Here is space for notes on the original. Is under investigation by the Newspaper Unit
      Details about damage to individual pages saved in ALTO during The attribute QUALITY_DETAIL.
    - note, with attribute type=”condition”.
      Here is information on the originals physical condition, according to a list of coded values. Investigation is under way by the Newspaper Unit.

<relatedItem> = printed newspaper as the original

- with attribute TYPE=”original”
  Mandatory if applicable. Here is information about the original (printed newspaper) to the digital reproduction.

Includes the elements:

identifier

- Value: Code that identifies which sample used as the original. Value in position 1-2 in preparation code from Signe. Example: "S-A "
- attribute type=”local”.

physicalDescription

Includes:

- form
  - with attribute authority=”marcform” and value: print.

location

Includes:

- holdingSimple
  Includes elementet:
  - copyInformation, as Includes:
    - note, without attribute.
      Here is space for notes about the original, for example, that it has been disposed after digitization. Can be repeated if necessary.
      Always a note with a code that identifies the originals physical condition, and if it is a national sample (taken from Signe, such as SA-1).
      Details about damage to individual pages saved in ALTO during The attribute QUALITY_DETAIL.
    - note, with attribute type=”condition”.
      Note on the originals physical condition. Here is the value placed which forms position 4 in preparation code from Signe. Example: "1"

<relatedItem> section, appendix, or news bill

- with attribute type=”constituent”>

<relatedItem TYPE=”constituent”> is registred under its own dmdSec: dmdSec/mdWrap/@MDTYPE=”MODS”/xmlData.

Here information is registered about the parts that are included in the newspaper issue which needs to be described individually: sections, appendices, news bills.

Includes the elements:

genre

Value in the element is chosen according to genre term_kbse: supplement, section, newsbill

titleInfo

Includes the elements:

- partName
  - Name and any part names. KB will provide MKC with more information for each newspaper.

subject

Includes the element:

- Subject
  - Topic.
    Value: select from the controlled vocabulary: "bilagetyp_kbse”.

originInfo

Includes:

- dateIssued
  - Here write the date if it differs from the newspaper issue. Write according to pattern yyyy-mm-dd.
  - with the attribute encoding=”w3cdtf”.
  - If date is missing in the original, ad the attribute qualifier, with the value ”inferred”. As value i <dateIssued> write an adopted date. As value in <dateIssued> write an adopted date.

note

- with the attribute type=”date”
  Here any remark about missing or incorrect date.
- Without attribute
  Eg for information about added title.

Local data on the publisher and supplier

These include name and identifier in Mimer’s *provider registry* for provider and publisher of the delivered resource. This information is for internal use at Kungl. Biblioteket.

Is registred in dmdSec/@ID=”dmdSec002”/mdWrap/@LABEL=”Local”

Elements and attributes in MODS

<mods>

<mods> is registered under mets/dmdSec/mdWrap/@MDTYPE=”mods”/xmlData.

Includes the elements <name> and <role> is repeated for supplier and publisher::

name

- with attributes TYPE=”corporate”, AUTHORITY=”local” and valueURI=”MHS id for the organisation]

Includes:

- namePart

Values for namePart:

- ”Kungl. biblioteket”
- ”Riksarkivet/MKC”

Values for valueURI:

- ”http://id.kb.se/organisations/ SE2021001710 ”
- ”http://id.kb.se/organisations/SE2021001074-MKC”

role

Includes:

- roleTerm with attributes type=”text” and authority=marcrelator or local
    - .. Values:
      ”supplier” (authority=local)
      ”publisher” (authority=marcrelator)
    - ..

PREMIS

PREMIS stand for Preservation Metadata: Implementation strategies. It is a standard developed to effectively manage, find and recover digital information. PREMIS include information about the digital object's technical characteristics, provenance, activities around objects, any restrictions in the access and use, etc.
Library of Congress provides schemes in XML for PREMIS and there is also a Swedish version of which the National Archives is responsible.
The PREMIS model entities Object, Event, Agent and Rights, are represented in the scheme of metadata containers of the same name.
In this project, MKC will work with <object> (see PREMIS: OBJECT) where each digital object's basic data is captured.

PREMIS:OBJECT

Two different types of ”object” is registred in the packages:
1. One <object> relating to the representation. The attribute xsi:file=”representation”
2. One <object> is repeated for each data file which is included in the package. The attribute xsi:file=”file”

<object> is registred under
mets/amdSec/techMD/mdWrap/@MDTYPE=”PREMIS:OBJECT”/xmlData.
Is repeated for every file.

1. object xsi:file=”representation”

Here are collected metadata relating to the representation, which you do not want to repeat for each data file.

- Mandatory attribute in <object> is xsi:file with the value ”representation”.
- Mandatory under <object> is:
  - 1.1. objectIdentifier

1.1. objectIdentifier

Element under <object> who identifies the representation. Here are two under elements:

- objectIdentifierType – Enter here the identifier type, value: “local”.
- objectIdentifierValue – Enter here the same identifier as in mets/@OBJID.

1. object xsi:file=”file”

Here are the metadata that relates to a file collected.

- Mandatory attribute is xsi:file with the value ”file”.
- Mandatory element/container under <object> is
  - 1.1. objectIdentifier
  - 1.5. objectCharacteristics

1.1. objectIdentifier

Identifies the object (file). Contains:

- objectIdentifierType - Enter value “filepath”.
- objectIdentifierValue – Enter here the name of the file. See also Attachment: Naming convention.

1.5. objectCharacteristics

Contains:

- 1.5.1. compositionLevel
- 1.5.2. fixity
- 1.5.3. size
- 1.5.4. format
- 1.5.7. objectCharacteristicsExtension

1.5.1. compositionLevel

Value always “0” (zero) for uncompressed file.

1.5.2. fixity

Information about checksum calculation. Information about current checksum and algoritm is repeated in mets/fileSec/fileGrp/file/@CHECKSUMTYPE and /@CHECKSUM. Contains:

- messageDigestAlgorithm MD5
- messageDigest – the calculated checksum
- messageDigestOriginator - eg. ”Riksarkivet/MKC”.

1.5.3. size

Here is information about the size of the file expressed in byte. Example: ”3120”. The information is repeated in mets/fileSec/fileGrp/file@SIZE.

1.5.4. format

Information about the format. Contains:

- 1.5.4.1. formatDesignation
- 1.5.4.2. formatRegistry

1.5.4.1. formatDesignation

Information about file format and version.

- formatName – Name of the file format. Value from PRONOM.
- formatVersion – Version number of the file format according to PRONOM.

1.5.4.2. formatRegistry

Information from PRONOM about the format. Contains:

- formatRegistryName - always value ”PRONOM”.
- formatRegistryKey – enter identificator/key to format and format version in PRONOM.
- formatRegistryRole - always value ”specification”.

1.5.7. objectCharacteristicsExtension

Container to register object characteristic data using other metadata standards. See also under MIX.

MIX

MIX is a metadata standard for describing technical metadata for image files. It is a translation into XML by another standard, ANSI / NISO Z39.87-2006.

Metadata that describes MIX are stored in a PREMIS object under the tag <objectCharacteristicsExtension>. Each image file should have its metadata stored in its own PREMIS object.

There is a lot of overlap between PREMIS and MIX. Some of the tags used in the MIX are retrieved directly from PREMIS. In cases where something can be described both in PREMIS and MIX we have chosen to use only PREMIS. If we instead chose to use the MIX, we usually still had to use elements from the PREMIS, but with the difference that the elements are embedded in the MIX section. Examples of data that can be stored in both MIX and PREMIS is basic file information (file format, etc.) and checksums.

BasicDigitalObjectInformation

Here is information of general nature stored which is applicable on all types of digital files, not just image files. Much of the information is redundant and is stored in other places in the scheme.

compression
If the image file is compressed, it is stated in the MIX because PREMIS does not support this.

Contains:

- compressionScheme. Values from ANSI/NISO Z39.87-2006
- compressionSchemeLocalList. Enter only if a custom dictionary is created
- compressionSchemeLocalValue. Enter only if a custom dictionary is created
- compressionRatio. Should according to the standard be specified as a positive integer but we also allow floating point when the compression rate in JPEG2000 is not limited to integer.

BasicImageInformation

This section contains basic metadata about the image file.

BasicImageCharacteristics

Metadata that is not related to a specific file format.

Imagewidth/imageHeigth
Width and height of the image in pixels

All the metadata that is related to color space is saved in MIX. It is possible to also store the version number of a color space in the name element (eg AdobeRGB 1998) but this makes it harder to find the version number. If the color space cannot be considered well established (eg eciRGBv2) a URI should be stored to simplify information retrieval. MIX also supports local

color profiles but we believe that such has a negative impact on long-term resistance and therefore should not be used.

PhotometricInterpretation
Contains:

- colorSpace

PhotometricInterpretation/colorProfile/iccProfile
Contains:

- iccProfileName
- iccProfileVersion
- iccProfileURI (Note that the tag end on URI, not URL as ANSI/NISO Z39.87-2006 states).

SpecialFormatCharacteristics

Metatdata which are related to a specific file format.

MIX supports JPEG2000. The fields for the codestreamProfile and complianceClass are not used when we do not see any use for this metadata. Only a small number of parameters that are of importance for JPEG2000 can be specified in dedicated element. Unfortunately, a free text field is lacking where it is possible to store all the parameters used in the conversion. These should therefore be stored in the metadata field that is inside the image file.

JPEG2000/CodecCompliance

Contains:

- codec
- codecVersion

JPEG2000/EncodingOptions

Contains:

- tiles
- qualityLayers
- resolutionLevels

Image Capture Metadata

This section Contains technical metadata about the image capture.

SourceInformation

Metadata which relate from the captured (analog) object.

Object orientation can be specified in MIX, which makes it easier to view the image in the correct orientation

Contains:

- orientation

It is possible that in MIX specify the type of original that has been digitized. This information is usually stored in MODS. However, we can see the extremely rare cases where the majority of the image files have a certain physical original and a single image file has a different (eg a missing journal page which subsequently digitized from microfilm). In cases where the image files have different originals, sourceType should be used to record the originals type for different image files. The field sourceID is used to specify which physical original that has been digitized. sourceID is formatted in the same way as the corresponding material categories in MODS. For media types that are not covered the categories available in the ANSI / NISO Z39.87-2006 can be used. If this becomes necessary, we will need to create a glossary for these fields.

SourceID
Contains:

- sourceType
- sourceID

The size of the physical original can be specified in MIX. This information is redundant if the sampling frequency is stored as the size of the object can be calculated from the size and sampling frequency. If the equipment can automatically measure the physical size while the sampling frequency is unknown the size should be specified (in millimeters). In all other cases, this element can be left blank.

SourceSize/sourceX(Y)Dimension
Contains:

- sourceX(Y)DimensionValue
- sourceX(Y)DimensionUnit

GeneralCaptureInformation

Metadata for image capture, independent of image capture devices.

Time for image capture and the use of image capture equipment will be stored. <imageProducer> is normally left empty when this information stored in MODS. In the same way that a single image can have a different physical original than the majority of the images it can also have another digital origin. In these cases, the image producer should be stated.

Contains:

- dateTimeCreated
- captureDevice
- imageProducer

ScannerCapture

Metadata for image capture and technical equipment (scanner).

We make no use of the fields maximumOpticalResolution or scannerSensor, therefore these are not included. We recommend that the scanner's serial number is included to simplify the quality assurance of multiple scanners if the same model is used. To facilitate searches, scanner software and software version are specified in separate fields.

Contains:

- scannerManufacturer

ScannerModel
Contains:
- scannerModelName
- scannerModelNumber
- scannerModelSerialNo

ScanningSystemSoftware
Contains:
- scanningSystemSoftwareName
- scanningSystemSoftwareVersionNo

DigitalCameraCapture

Metadata for image capture and technical equipment (scanner).

We make no use of the fields maximumOpticalResolution or scannerSensor, which therefore are not included. We recommend that the scanner's serial number is included to simplify the quality assurance of multiple scanners if the same model is used. To facilitate searches, scanner software and software version are specified in separate fields.

Contains:
- digitalCameraManufacturer

ScannerModel
Contains:
- digitalCameraModelName
- digitalCameraModelNumber
- digitalCameraModelSerialNo

A selection of the metadata derived from the actual image capture with the camera has been made. Metadata that may be useful or in demand have been included, while all other fields (eg OECF and spectralSensitivity) are not included in the specification. Some fields in the standard are also duplicated because they describe the same thing but using different devices and formats (eg ½ s and 0.5 s for the shutter speed).

CameraCaptureSettings/ImageData
Contains:
- fNumber
- exposureTime
- isoSpeedRatings
- exifVersion
- exposureBiasValue
- lightsource
- focalLength
- autoFocus

ImageAssessmentMetadata

Metadata dealing with the quality of the image.

SpatialMetrics

If the equipment can measure the practical sampling rate, this should be stated and prevail over the originals physical size (see above). It is important to note that the sampling frequency changes with the distance to the object.

samplingFrequencyUnit
Contains:

- xSamplingFrequency
- ySamplingFrequency

ImageColorEncoding

The images color depth should be stated both in the total number of bits and the number of bits for each color channel. It is possible to specify more advanced color information, but we see no benefit to this.

Contains:

- samplesPerPixel

bitsPerSample
Contains:

- bitsPerSampleValue
  This element is repeated for each channel.
- bitsPerSampleUnit

TargetData

The image quality is measured with help from targets and to enable a subsequent verification, information from the quality measurement is included in the archive package. This is done by one output file that is saved in the archive package. It is possible to point to the result file from MIX but to create uniform package, this is done instead from fileSec/structMap . Information about the used target should always be saved. An external target is captured alone, while an internal target appears in the same picture as the object. Identifiers for individual objectives should be included in the metadata. For more information about quality measurement, see separate documentation
Contains:

- targetType (”external”/”internal”, not 0/1 as stated in ANSI/NISO Z39.87-2006)

TargetID
Contains:

- targetManufacturer
- targetName
- targetNo

Change History

If the image file is processed this should be specified in ChangeHistory . It is important that both describe the software used, and all the parameters and purpose with post processing. The aim, however, should only be indicated if it deviates from normal processing. Only the last processing is saved in <ImageProcessing>. Earlier processing is moved to a block under the Previous Image Metadata. If multiple processes are carried out, they are stored chronologically (most recent at the top of Previous Image Metadata and the first last). Information saved in Previous Image Metadata should never be erased.

ImageProcessing

Contains:

- dateTimeProcessed
- processingAgency
- processingRationale (explanatory text in image processing that is not performed routinely)
- processingActions

processingSoftware

Contains:

- processingSoftwareName
- processingSoftwareVersion

Links/References

PREMIS, http://www.loc.gov/standards/premis/

MODS, http://www.loc.gov/standards/mods/

METS, http://www.loc.gov/standards/mets/

MIX, http://www.loc.gov/standards/mix/

ALTO, http://www.loc.gov/standards/alto/

The National Librarys profiles and schemes for packaging in METS, http://www.kb.se/namespace/mets/

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

Attachments

Projekt Digitalisering och tillgängliggörande av den svensk-amerikanska pressen (SAP): Metadataelement och attribut i METS, MODS, PREMIS och MIX (table)

File naming convention

Instruction for missing page, missing issue, missing supplement, missing section, missing newsbill

Projekt Digitalisering och tillgängliggörande av den svensk-amerikanska pressen (SAP): Metadataelement och attribut i METS, MODS, PREMIS och MIX				**Updated 2013-10-08**				
Scheme	**Element name**	**Attribute**	**Value in attribute**	**Value in element**	**M/O**	**Repeated?**	**Notes/comments**	**Comments**
					M: mandatory MA: mandatory if applicable O=Optional			
							Blue text represents the metadata that comes from KB and/or KB's system	
							Red text = updates	
							Purple text = needs to be checked	
METS	mets	**Namespaces and schemas for validating:**			M		Contains: metsHdr, dmdSec, amdSec, fileSec, StructMap	
		xmlns:mets	http://www.loc.gov/METS/					
		xmlns:mods	http://www.loc.gov/mods/v3					
		xmlns:premis	info:lc/xmlns/premis-v2					
		xmlns:mix	http://www.loc.gov/mix/v20					
		xmlns:xlink	http://www.w3.org/1999/xlink					
		xmlns:xsi	http://www.w3.org/2001/XMLSchema-instance					
		xsi:schemaLocation	http://www.kb.se/namespace/mets/kbse_mets_001.xsd					This schema is updated (version 0.3) with vocabularies for altRecordID/@TYPE and @RECORDSTATUS. This reference, however, will keep the same name/uri.
			http://www.kb.se/namespace/mods/kbse_mods_001.xsd					Empty spacing between the address pairs
			http://www.kb.se/namespace/premis/kbse_premis_001.xsd					
			http://www.kb.se/namespace/mix/kbse_mix20_001.xsd					
		PROFILE	http://www.kb.se/namespace/mets/kbse_mets_profile_001.xml					This profile is updated (version 0.3) with vocabularies for altRecordID/@TYPE and @RECORDSTATUS. This reference, however, will keep the same name/uri.
		ID			M		The name of the mets file. Se separate instructions for file naming. Repeated in metsHdr/@metsDocumentID	
		OBJID			M		Unique identifier for the packet which is based on the filename.	

		LABEL			M		Get the paper's title from dmdSec/mdWrap/ @LABEL=Primary/.../mods:title. Example: Aftonbladet 1851-12-04	
		TYPE	SIP		M	NR	Value should always be "SIP"	
	metsHdr				M	NR	Under: mets. Contains: agent, altRecordID, metsDocumentID	
		CREATEDATE			M	NR	Date and time when the mets file was created. Written YYYY-MM-DD'T'HH: mm: ss +01:00	
		RECORDSTATUS			MA	NR	Used to indicate if a package is a replacement or an addition to a previously delivered package. RECORD STATUS is written only if any of the values SUPPLEMENT or REPLACEMENT should be assigned to the package	RECORD STATUS is written only if any of the values SUPPLEMENT or REPLACEMENT should be assigned to the package
			REPLACEMENT				Replacement - The entire package, with all data exchanged	The entire package, with all data is replaced
			SUPPLEMENT				supplementation - parts of the package are replaced / and or supplemented with more data	Parts of the package is replaced/and or supplemented with more data. See special instructions
	agent				M	R	Under: metsHdr. Contains name and note. Agent registered 2 times for the MKC National Archives (role = creator) and the National Library (role = archivist).	
		ROLE			M			
			CREATOR					
			ARCHIVIST					
	name				M		Under: agent. Obligatory attribute type with the value: organization	
		TYPE	ORGANIZATION		M			
				Riksarkivet/MKC			With ROLE="CREATOR"	
				Kungl. biblioteket			With ROLE="ARCHIVIST"	
	note				M		Under:agent. Mandatory element containing the organization number for MKC and KB respectively	
				http://id.kb.se/organisations/SE2021001074-MKC			The organisation number for Riksarkivet/MKC	
				http://id.kb.se/organisations/SE2021001710			The organisation number for Kungl. biblioteket	

	altRecordID				M	R	altRecordID is repeated 3 times with 3 different TYPE attributes: DELIVERYTYPE, DELIVERYSPECIFICATION and SUBMISSIONAGREEMENT	
		TYPE						
			DELIVERYTYPE				AGREEMENT	
			DELIVERYSPECIFICATION				http://www.kb.se/namespace/digark/deliveryspecification/agreement/sap/	
			SUBMISSIONAGREEMENT				http://www.kb.se/namespace/digark/submissionagreement/DNR_122-KB_270-2013/	
	metsDocumentID				M		Under: metsHdr. Same value as in mets/@ID	
	The newspaper issue							
	dmdSec				M	R	Under: mets. Contains: mdWrap	
		ID	dmdSec001, dmdSec002, etc.		M	NR	Consecutive numbering of dmdSec according to pattern "dmdSec001", "dmdSec002", "dmdSec003" etc.	
	mdWrap				M	NR	Under: demdSec.Contains xmlData	
		LABEL	Primary					
		MDTYPE	MODS		M	NR		
	xmlData				M	NR	Under: mdWrap. Contains: mods	
MODS	mods				M	NR	Contains: identifier, typeOfResource, genre, originInfo, physicalDescription, relatedItem	
	identifier	type	local	[mkc:s resource id]	M	NR	Under: mods.	This id is in most cases the same as mets@objid.
	typeOfResource			text	M	NR	Under: mods.	
	genre	authority	marcgt	issue	M	NR	Under: mods.	
	titleInfo				M	NR	Under: mods. Contains: title	
	title				M	NR	Under: titleInfo . The newspaper's title and date/number (date/number). Same value as in mets:mets@LABEL. Example: Aftonbladet 1851-12-04	

	originInfo				M	NR	Under: mods. Contains: dateIssued, note, edition. The data on originalInfo regards the analog (printed) original.	
	dateIssued				M	NR	Under: originInfo. Example: 1834-08-11.	
		encoding	w3cdtf		M	NR		
		qualifier	inferred		MA	NR	If dating is missing, or incorrect, in the original, state an expected date and QUALIFIER attribute with the value: "inferred".	
	note	*type*	*date*		*O*	*R*	*Under: originalInfo . Are discussed: In a "note" to enter if the date in the original is missing / incorrect. Instead, use the note in relatedItem type=host*	*Repetition. Instead use note in relatedItem type=host*
	edition			string	MA	NR	Under: originalInfo . Value is the same as bibliographic ID in Signe	
	physicalDescription				M	NR	Under: mods. Contains: digitalOrigin, note	
	digitalOrigin				M	NR	Under: physicalDescription	
				reformatted digital			digitization from printed original	
				digitized microfilm			digitization from micro film	
	note	type	reproduction	Digital reproduktion: Stockholm : Riksarkivet/MKC i samarbete med Kungl. biblioteket, [yyyy]	M	NR	Under: physicalDescription.	
	note	type	script	According to vcScriptType_kbse	M	NR	Under: physicalDescription.	
	Micro film as original							
	relatedItem	type	original		MA	R	Under: mods. Contains: identifier, location, physicalDescription	
	identifier	type	reel number	string	M	NR	Under: relatedItem. Mikrofilmens nummer	
	location				O	NR	under: relatedItem. Contains holdingSimple	

	holdingSimple				O	NR	Under: location. Contains copyInformation	
	copyInformation				O	NR	Under: holdingSimple. Contains note	
	note				O	R	Under: copyInformation. Notes about the originals. Can be repeated	Can be used when necessary for comments in free text
	note	type	condition		O	NR	Under: copyInformation. Notes about the originals physical condition.	Can be used when necessary for comments in free text
	physicalDescription				M	NR	Under: relatedItem. Contains: form	
	form	authority	marcform	microfilm			Under: physicalDescription	
	Printed copies as original							
	relatedItem	type	original		MA	R	Under: mods. Contains: identifier, location, physicalDescription. If more than one copy of the newspaper delivered , relatedItem is repeated	If more than one copy of the newspaper has been delivered , relatedItem is to be repeated
	identifier	type	local	string	M	NR	Under: relatedItem. Code that identifies the copy used for original. The values in positions 1-2 in the preparation code from Signe. Example: "S-A"	See updated guidance for relatedItem@type="original"
	location				O	NR	Under: relatedItem. Contains holdingSimple.	
	holdingSimple				O	NR	Under: location. Contains copyInformation	
	copyInformation						Under: holdingSimple.Contains note	
	note				O	R	Under: copyInformation. Notes about the originals. Can be repeated	Can be used when necessary for comments in free text
	note	type	condition		O	NR	Under: copyInformation. Note on originals physical condition. Here are the value at the position 4 of the preparation code from Signe. For example, "1"	See updated guidance for relatedItem@type="original"
	physicalDescription				M	NR	Under: relatedItem. Contains: form	
	form	authority	marcform	print	M	NR	Under: physicalDescription	
	Host publication (newspaper) including numbering and date of newspaper issue:							
	relatedItem	type	host		M	NR	Under: mods. Contains: genre, titleInfo, originInfo, language, identifier, part, note	Note. <relatedItem> are the parent (container-element) to genre, titleInfo , originalInfo, language identifier, part, note

	genre	authority	marcgt	newspaper	M	NR	Under: relatedItem	
	titleInfo				M	NR	Under: relatedItem. Contains: title	
	title				M	NR	Under: titleInfo. The title of the host publication. Metadata from KB/LIBRIS	
	originInfo				M	R	Under: relatedItem. Contains: dateIssued	
	dateIssued	encoding	w3cdtf		M	NR	Under originInfo. Date according to pattern YYYY-MM-DD	
		point	start		M	NR	Date for the first ed number (start date)	
		point	end		M	NR	Date of the last ed number (end date)	
	language				M	R	Under: relatedItem. Contains: languageTerm	
	languageTerm	type	code		M	NR	Under: language	
		encoding	iso639-2b		M	NR	A three-digit code according to ISO639-2b	
	identifier	type			M	R	Under: relatedItem. Identifiers to the newspaper in its entirety (host publication)	The Newspaper Unit needs to provide MKC with proper id:s:
			uri	http://libris.kb.se/resource/bib/[librisid]	M	NR	URI to the libris record of the digital newspaper (host publication)	Svenska socialisten has: http://libris.kb.se/bib/13983363
			issn		MA	NR	ISSN to the newspaper online . The metadata from KB/LIBRIS. Printed with hyphen. The element i excluded if there is no ISSN	Most of the historical Swedish American newspapers lack ISSN
	part				MA	NR	Under: relatedItem. Contains: detail, date	
	detail	type	issue		MA	NR	Under: part. contains: number	
	number				MA	NR	Under: detail. The newspaper's numbering here, if there is one. Ie printed numbering or other denotation (eg Sample number) on the newspaper number.	
	date	encoding	w3cdtf		MA	NR	Number of the microfilm is in relatedItem / TYPE = "" original "" / identifier / TYPE = "" reel number "" "	

		qualifier	inferred		MA	NR	If dating is missing, or incorrect, in the original, state an expected date and the attribute QUALIFIER with the value: "inferred".	
	note	type	date		O	R	Under: relatedItem. Under discussion: In a "note" to enter if the date in the original is missing/is incorrect. Standard Phrase: "Dating is missing or incorrect"	
	note	type	numbering		O	R	Under: relatedItem. Under discussion: In a "note" enter if numbering in the original is missing/is incorrect. Standard Phrase: "Numbering missing or incorrect"	
	The Project:							
	relatedItem	type	host		M	NR	Under: mods. Reference to information in LIBRIS about the project. Contains: genre, title info, identifiable	Note. <relatedItem> are the parent (container-element) to genre, titleInfo, identifier
	genre			project	M	NR	Under: relatedItem. Value from list genreterm_kbse	
	titleInfo				M	NR	Under: relatedItem. Contains: title	
	title			Digitalisering och tillgängliggörande av den svensk-amerikanska pressen	M	NR	Under: titleInfo. The title of the project. Always the same value	
	identifier	type	uri	http://libris.kb.se/bib/13983307	M	NR	Under: relatedItem. URI till LIBRIS-post about the project. Will always have the same id	
Appendix, section, newsbill:							newsbill could be excluded here and instead only be recorded in structMap	
METS	dmdSec				MA	R	Under: mets. Contains: mdWrap	
		ID	dmdSec001, dmdSec002, etc.		M	NR	Consecutive numbering of all dmdSec: dmdSec001, dmdSec002 etc.	
	mdWrap	MDTYPE	MODS		M	NR	Under: dmdSec. Contains:xmlData	
	xmlData				M	NR	Under: mdWrap. Contains: mods	
MODS	mods				M	NR	Under: xmlData. Contains: relatedItem	Note. <xmlData> must here be followed by <mods>

	relatedItem	type	constituent		M	NR	Under: mods. Contains: titleInfo, subject, genre, originInfo, note	Note. <relatedItem> are the parent (container-element) to to other elements within this the mods section
	titleInfo						Under: relatedItem. Contains: partName	
	partName				MA	NR	Under: titleInfo . Name of the appendix or section. Here are also any letter or number designation	newsbills have no names.
	PartNumber				*MA*	*NR*	*Not used*	
	subject				MA	R	Under: relatedItem. Keywords that describe the content of the appendix. Contains topic	
	topic	authority	bilagetyp_kbse	string	MA	NR	Under: subject. KB provides the terms	
	genre			supplement	MA	NR	Under: relatedItem. Value according to genreterm_kbse	
				section	MA	NR		
				newsbill	MA	NR		
	originInfo				O	NR	Under: relatedItem. Contains: dateIssued	
	dateIssued	encoding	w3cdtf		O	NR	Under: dateIssued. The data on originalInfo regards the analog (printed) original. Listed if date differs from issue.	
		qualifier	inferred		O	NR	If dating is missing, or incorrect. In the original an expected date and the attribute QUALIFIER with the value: "inferred" is used.	
	note	type	date		O	R	Under: relatedItem. In a "note" enter if the date in the original is Missing/incorrect	
	note				O	R	Under: relatedItem. Standard Phrase: "Title added by hand, probably temporary appendix"	
Local information about delivery:								Local information for the digital archive of KB.
METS	dmdSec				M	R	Under: mets. A locall dmdSec, to handle name + id for supplier. Contains: mdWrap	

		ID	dmdSec001, dmdSec002, etc		M	NR		
		LABEL	Local		M	NR		
	mdWrap	MDTYPE	MODS		M	NR	Under: dmdSec. Contains: xmlData	
	xmlData				M	NR	Under: mdWrap. Contains: mods	
MODS	mods				M	NR	Under: xmlData. Contains: name	Note. <xmlData> must here be followed by <mods>
	name	type	corporate		M	R	Under: mods. Contains: namePart, role	
		authority	local		M	NR		
		valueURI	http://id.kb.se/organisations/SE2021001710		M	NR	The form for organization id: SE[organization number]-[suffix]	The form changed in consultation with RA in order to harmonize with the HSA code used by local governments.
	namePart			Kungl. biblioteket	M	NR	Under: name. The name of the publisher, here always Kungl. biblioteket	
	role				M	NR	Under: name. Contains: roleTerm	
	roleTerm	type	text		M	NR	Value always publisher	
		authority	marcrelator	publisher				
	name	type	corporate		M	R	Under: mods.Contains: namePart, role	
		authority	marcrelator		M	NR		
		valueURI	http://id.kb.se/organisations/SE2021001074-MKC		M	NR	The form for organization id: SE[organization number]-[suffix]	The form changed in consultation with RA in order to harmonize with the HSA code used by local governments.
	namePart			Riksarkivet/MKC	M	NR		
	role				M	NR	Under: name. Contains: roleTerm	
	roleTerm	type	text		M	NR	Under: role. Value always supplier	
		authority	marcrelator	supplier				
METS PREMIS:OBJECT								
METS	amdSec				M	NR	Under: mets. Contains: techMD	
		ID	amdSec001		M	NR		
	techMD				M	R	TechMD repeated for each object included in the package. Contains: mdWrap	
		ID	techMD001, techMD002 etc.		M	NR		
	mdWrap	MDTYPE	PREMIS:OBJECT		M	NR	Contains: xmlData	
	xmlData				M	NR	Contains: premis	
PREMIS	premis				M	NR	Contains: object	

	object				M	NR	Contains: objectIdentifier, objectCharacteristics, relation	
		xsi:type			M	NR		
			file		MA	NR	Repeat techMD/..object=file for each data file inclluded in the package.	
			representation		MA	NR	Always a techMD/..object=representation for each package.	
	objectIdentifier				M	NR	Contains: objectIdentifierType, objectIdentifierValue	
	objectIdentifierType				M	NR		
				filepath			object=file	
				uri				
				local			object=representation	
	objectIdentifierValue				M	NR	When object=file the value becomes the same as the file name (filepath). When object=representation the value becomes the same as in mets/@OBJID	
	objectCharacteristics				MA	NR	Contains: compositionLevel, fixity, size, format, objectCharacteristicsExtension	
	compositionLevel			0	M	NR	0(noll) = Uncompressed file	
	fixity				M	R	Contains: messageDigestAlgorithm, messageDigest, messageDigestOriginator	
	messageDigestAlgorithm			MD5	M	NR	KB:s metsprofile support MD5 och SHA-1	
	messageDigest				M	NR	checksum	
	messageDigestOriginator			Riksarkivet/MKC	M	NR		
	size				M	NR	a number in the unit byte	
	format				M		Contains: formatDesignation, formatRegistry	
	formatDesignation				M	NR	Contains: formatName, formatVersion	
	formatName				M	NR	enter the name of the format (Name:) according to PRONOM, http://http://www.nationalarchives.gov.uk/PRONOM/	
				JPEG2000				
				Extensible Markup Language				

	formatVersion				M	NR	enter the version of the format according to PRONOM, or leave blank if information is missing	
				1.0			Format version for xml	
				[No information]			Format version leave blank for jpg2000	
	formatRegistry				M	NR	Contains: formatRegistryName, formatRegistryKey, formatRegistryRole	
	formatRegistryName			PRONOM	M	NR	Value, always from PRONOM	
	formatRegistryKey				M	NR	Formatets ID i PRONOM	
				x-fmt/392			Formatkey för jpg2000	
				fmt/101			Formatkey för xml	
	formatRegistryRole			specification	M	NR	Always "specification"	
	objectCharacteristicsExtension				M		Container for metadata according to other standards. Contains: mix	
objectCharacteristicsExtension: MIX							Technical metadata for image files	
MIX	mix							
	BasicDigitalObjectInformation							
	compression				M	R	(for each filel)	
	compressionScheme			string/enumerated in local list	M	NR	Initially, the values of ANSI / NISO are used. If these are not sufficient, create a custom list	
	compressionSchemeLocalList			reference/URL	MA	NR	location of the file containing the local enumerated list	
	compressionSchemeLocalValue				MA	NR	Values from the local glossary	
	compressionRatio			float (actually int)	MA	NR	6, 8, 10 etc. 6 corresponds to 6:1. JPEG2000 may have e.g. 1:13.3 so I suggest we use the float. O o standard, MA with us	
	BasicImageInformation							
	BasicImageCharacteristics				M	NR		
	imageWidth			int	M	NR	Width in pixels. (Specifies the width of the image in horizontal or X dimension Should be registrered in pixels)	
	imageHeigth			int	M	NR	Height i pixels (Specifies the length of the image in vertical or or Y dimension, should be registrered in pixels)	
	BasicImageInformation							
	BasicImageCharacteristics							

	PhotometricInterpretation							
	colorSpace			string	M	NR	RGB, CMYK, Lab, etc. Should be from list even if it is not in the standard	
	BasicImageInformation							
	BasicImageCharacteristics							
	PhotometricInterpretation							
	ColorProfile				MA	NR		
	iccProfile				MA	NR		
	iccProfileName			string	MA	NR	sRGB, adobeRBG etc	
	iccProfileVersion			string	MA	NR	T.ex 1998	
	iccProfileURI			string/URL	MA	NR	If the profile is not well known, a URI must be specified.	
	SpecialFormatCharacteristics				MA	NR	Only for JPEG2000	
	JPEG2000							
	CodecCompliance				MA (O)	NR	(Some O in the standard, MA with us)	
	codecVersion			string	MA (O)	NR	Kakadu	
	codecVersion			string	MA (O)	NR	3.1	
	SpecialFormatCharacteristics							
	JPEG2000							
	EncodingOptions				MA	NR		
	tiles			string	MA	NR	1024x1024	
	qualityLayers			int	MA	NR	14	
	resolutionLevels			int	MA	NR	6	
							Note: The parameters used in the encoding should be saved inside the file in an XML field.	
	ImageCaptureMetadata							
	SourceInformation							
	SourceID							
	sourceType			string	MA	NR	Used only if the current image has different origin than the majority of the files in the package (eg previously missing page scanned from micro film). Normal values are digitized microfilm born digital and reformatted digital.	
	sourceID			string	MA	NR	Identifier to the , physical object, such as which microfilm roll is used. Formatted in the same way as the corresponding materials category in MODS	

	ImageCaptureMetadata							
	SourceInformation							
	SourceSize							
	sourceXDimension							
	sourceXDimensionValue			real	O	NR	Shall be entered if the equipment can automatically deliver the value and if not sampling rate can be set. Can be used for scaling.	
	sourceXDimensionUnit			enumerated	MA	NR	millimeter	
	ImageCaptureMetadata							
	SourceInformation							
	SourceSize							
	sourceYDimension							
	sourceYDimensionValue			real	O	NR	Shall be entered if the equipment can automatically deliver the value and if not sampling rate can be set. Can be used for scaling.	
	sourceYDimensionUnit			enumerated	MA	NR	millimeter	
	ImageCaptureMetadata							
	orientation			enumerated	M	NR	Enter a value according to http://www.loc.gov/standards/mix/mix20/mix20.xsd, typeOfOrientationType. See Glossary, row 137ff	See glossary row 137ff
	GeneralCaptureInformation				M	NR		
	dateTimeCreated			w3cdtf	M	NR	Date and tiime for the image capture. Is entereed according to w3cdtf som YYYY-MM-DD'T'HH:mm:ss+01:00.	
	captureDevice			enumerated	M	NR	List of allowed values to be developed in cooperation with MKC, eg Scanner, digital camera, etc	
	imageProducer			string	MA	NR	Used only if the image has a different origin than the the majority of the images, ie. analogous to the source Type/sourceID	
	ScannerCapture				MA	NR	All are R in the standard	
	scannerManufacturer			string	MA	NR	Manufacturer	
	ScannerCapture							
	ScannerModel							

	scannerModelName			string	MA	NR	Name of model.	
	scannerModelNumber			string	MA	NR	Number of model	
	scannerModelSerialNo			string	R	NR	Serial number. Can be used for QA if multiple devices of the same scanner type are used.	
	ScannerCapture							
	ScanningSystemSoftware							
	scanningSoftwareName			string	MA	NR	Name of scanner software.	
	scanningSoftwareVersionNo			string	MA	NR	Version name of scanner software.	
	DigitalCameraCapture				MA	NR	All are R in the standard	
	digitalCameraManufacturer			string	MA	NR		
	DigitalCameraCapture							
	DigitalCameraModel							
	digitalCameraModelName			string	MA	NR	Model name	
	digitalCameraModelNumber			string	MA	NR	Model number	
	digitalCameraModelSerialNo			string	R	NR	Serial number. Can be used for QA if multiple devices of the same scanner type are used.	
	DigitalCameraCapture							
	CameraCaptureSettings							
	ImageData				MA		O in the standard	
	fNumber			real	MA	NR	8.0	
	exposureTime			real	MA	NR	0.5	
	isoSpeedRatings			int	MA	NR	Eg. 100. Values from ISO12232	
	exifVersion			enumerated	MA	NR	Eg. 0220 (2.2). The values retrieved from EXIF 2.2 and formatted a bit differently	
	exposureBiasValue			rational	MA	NR	1.0	
	lightsource			enumerated	MA	NR	Eg. flourescent. Values from EXIF 2.2	
	focalLength			real	MA	NR	Eg. 0.08. In meters	
	autoFocus			enumerated	MA	NR	Eg. Manual. Values from DIG35, B3.2.5	
	ImageAssessmentMetadata							
	SpatialMetrics							
	samplingFrequencyUnit	enumerated	1,2,3		MA	NR	1 = no unit, 2 = inch, 3 = centimeter Is this an attribute?)	

	xSamplingFrequency			rational	MA	NR	Enter only if the value can be calculated automatically. Overrides sourceDimension	
	ySamplingFrequency			rational	MA	NR	Enter only if the value can be calculated automatically. Overrides sourceDimension	
	ImageColorEncoding							
	bitsPerSample				M	NR		
	bitsPerSampleValue			string	M	R	Eg. 8. Specifies the number of bits per color channel. Repeat for each channel	This change is already implemented
	bitsPerSampleUnit			int	M	NR	Specifies the unit for bitsPerSample	
	ImageColorEncoding							
	samplesPerPixel			int	M	NR	Specifies the number of color channels	
	TargetData						(R in the standard)	
	targetType			enumerated	MA	NR	MA. 0 = external, 1 = internal. An external target takes up its own image, an internal target appear in the same picture as the object	
	TargetData							
	TargetID							
	targetManufacturer			string	MA	NR	GreytagMacbeth	
	targetName			string	MA	NR	ColorChecker	
	targetNo			string	MA	NR	Which copy that is used	
	ChangeHistory				MA	NR	Somewhat complicated structure. See the documentation of the MIX/NISO for details	
	ImageProcessing				MA	R	Stored in chronological order if the order is significant	
	dateTimeProcessed			w3cdtf	MA	NR	Last time of modification is saved. Written according to YYYY-MM-DD'T'HH: mm: ss +01:00	
	processingAgency			Riksarkivet/MKC	MA	R		
	processingRationale			string	R	NR	Possibility to enter explanatory text at the image processing, which is not performed routinely	
	ImageProcessing							
	processingSoftware				R	R		
	processingSoftwareName			string	R	NR		
	processingSoftwareVersion			string	R	NR		
	ImageProcessing							

	processingActions			string	R	R	Eg. deskew 1.6%. Should if possible be the command sent to the program	
METS: fileSec								
METS								
	fileSec	ID	fileSec001		M	NR	Under: mets. One fileSec per METS. Contains: fileGrp	
	fileGrp				M	R	Under: fileSec. fileGrp repeated, one per USE. Contains: file	Create one fileGrp for each format type: jpg (USE="image/master"), alto (USE="text/alto") , pdf (USE="text/pdf")
		ID	fileGrp001, fileGrp002 etc.		M	NR		
		USE	string		M	NR	Under: fileGrp. Pick value from glossary ID = "vcUSE_kbse"	
	file				M	R	Group files with the same USE within a fileGrp. Contains: Flocat	
		ID	file1, file2, file3, etc.		M	NR	Number from 1, 2, 3 etc. Prefix "file"	
		USE			M	NR	according to vocabulary ID="vcUSE" (same as fileGrp)	
		MIMETYPE	text/xml		M	NR	Value according tot MIME	
			image/jp2					
			application/pdf					mimetype for pdf
		SIZE			M	NR	same as in premis/…/size	
		CREATED			M	NR	YYYY-MM-DD'T'HH:mm:ss+01:00	
		ADMID			M	NR	ID for the techMD that contains metadata about the file	
		CHECKSUM			M	NR	Same checksum as in premis/fixity/messageDigest	
		CHECKSUMTYPE	MD5		M	NR		
	Flocat	LOCTYPE	URL		M	NR	Under: file.	
		xlink:type	simple		M	NR		
		xlink:href	file:[filnamn]		M	NR		
METS: structMap								
	structMap				M	R	Under: mets. Is repeatable, but normally only one structMap describing how objects relate to each other "physically. Divided into a hierarchical structure using the elements div. See examples in the manual. Contains: div	
		ID	structMap001		M	NR		

		TYPE	physical		M	NR	value "physical" from glossary vcStructMap_TYPE_kbse	
	div				M	R	Under: structMap. Each div may contain none, one or more div. Hierarchies are built with values from dictionary vcDiv_TYPE_kbse. Contains: fptr	
		ID	div001, div002 etc.		M	NR		
		TYPE	string		M	NR	Each div named in TYPE with a value from vcDiv_TYPE_kbse	
		LABEL	string		MA	NR	If div has the type-value page or issue and if page number for example is missing, value can be selected from vcLABEL_kbse	
		ORDER	1,2,3, etc.		M	NR	The order in which the file will appear in	
		ORDERLABEL			M	NR	Text fields. Here you can add page numbering (probobly not an issue for the the newspapers)	
		DMDID			M	NR	ID for the DMDID containing metadata at the level that the div element represents	
		ADMID	techMDXXX		M	NR	ID for techMD containing metadata representation.	
	fptr	FILEID	file1, file2, file3, etc.		M	R	Under: div. Enter the same value as in the corresponding file@ID	

Projekt SAP: Dictionaries				**Updated 2013-10-08**		
Scheme	**Element name**	**Attribute**	**Value in attribute**	**Agreed value in element**	**Reference**	**Comments**
METS						
	metsHdr	RECORDSTATUS	REPLACEMENT		vcRECORDSTATUS_kbse	
			VERSION			
	agent	ROLE	CREATOR		SWEIPB	
			ARCHIVIST		SWEIPB	
				Riksarkivet/MKC		
	name	ROLE			SWEIPB	
		TYPE	ORGANIZATION		SWEIPB	
	name			Kungl. biblioteket		
				Minnesota Historical Society		
				American Swedish institute		
				Swenson Swedish Immigration Research Center		
	altRecordID	TYPE	DELIVERYTYPE		FGS	
			DELIVERYSPECIFICATION		FGS	
			SUBMISSIONAGREEMENT		FGS	
	dmdSec	LABEL	Primary		vcDmdSec_LABEL_kbse	
			Local		vcDmdSec_LABEL_kbse	
	fileGrp	USE				
	file		image/master		vcUSE_kbse	image file only intended for long term preservation
			image/reference		vcUSE_kbse	image file intended for display (used when master file is also used for display)
			image/dynamic		vcUSE_kbse	image file with dynamic characteristics for display (used when master file is also used for display)
			text/alto		vcUSE_kbse	ALTO
			text/performance		vcUSE_kbse	File that contains the results of quality measurement
			text/pdf		vcUSE_kbse	text file in PDF format

			text/metadata		*vcUSE_kbse*	*added 20110301 :) file that contains metadata such as original metadata from the supplier. Used internally for KB*
		MIMETYPE			MIME type, http://www.iana.org/assignments/media-types/	
			image/jp2			
			text/xml			
			application/pdf			
	structMap	TYPE				
			physical		vcSTRUCTMAP_TYPE_kbse	
			logical		vcSTRUCTMAP_TYPE_kbse	
	div	TYPE				
			files	http://id.kb.se/organisations/SE2021001074-MKC	vcDiv_TYPE_kbse	
			issue	http://id.kb.se/organisations/SE2021001710	vcDiv_TYPE_kbse	
			pdf		vcDiv_TYPE_kbse	
			section		vcDiv_TYPE_kbse	
			page		vcDiv_TYPE_kbse	
			supplement		vcDiv_TYPE_kbse	
			newsbill		vcDiv_TYPE_kbse	
			performance		vcDiv_TYPE_kbse	
			origmetadata		vcDiv_TYPE_kbse	*added 20110301. Internaly for KB*
	div	LABEL	missingpage		vcDiv_LABEL_kbse	"tom sida" replace missing page.
			missingissue		vcDiv_LABEL_kbse	"tom sida" replace a missing newspaper number

			damagedpage		vcDiv_LABEL_kbse	Page with "broken" content
			misplaced		vcDiv_LABEL_kbse	page that is misplaced / unknown - sorted in the end
MODS						
	edition			[edition designation]	Unik lista för varje värdpublikation	
	dateIssued	qualifier	approximate		MODS	a date that may not be exact, but is approximated, such as "ca. 1972".
			inferred		MODS	a date that has not been transcribed directly from a resource, such as "[not before 1852]".
			questionable		MODS	a questionable date for a resource, such as "1972?".
	digitalOrigin			reformatted digital	MODS	Digitization of printed original
				digitized microfilm	MODS	Digitization from microfilm
				born digital	MODS	
	note	type	reproduction	[text]	Unik lista byggs upp efter behov	
			script	gothic	vcScriptType_kbse	The text in the original is in fracture-style.
				roman	vcScriptType_kbse	The text in the original is in antikva-style.
				mixed	vcScriptType_kbse	The text in the original contain both fracture and antikva
			condition	[Position 4 in preparation code from Signe]		Note about the originals physical condition
	identifier	type	uri	[one uri]	MODS	URL, URN, etc.
			local	[unspecified]	MODS	
			issn		MODS	
			reel number	[number of the mircofilm]	Signe	
	physicalLocation			[Library sigel]		Exampel: S
	topic			[Category Words that describe the content of he attachement]	bilagetyp_kbse	Locla topic list
	genre			supplement	genreterm_kbse	
				section	genreterm_kbse	
				newsbill	genreterm_kbse	
				project	genreterm_kbse	

				issue	marcgt	
				newspaper	marcgt	
	roleTerm			supplier	roleterm_kbse	
				publisher	marcrelator	
PREMIS						
	formatName			JPEG2000	PRONOM, http://http://www.nationalarchives.gov.uk/PRONOM/	
				Extensible Markup Language		
	formatRegistryName			PRONOM	PRONOM, http://http://www.nationalarchives.gov.uk/PRONOM/	
	formatRegistryKey			x-fmt/392	PRONOM, http://http://www.nationalarchives.gov.uk/PRONOM/	
				fmt/101		
MIX						
	compressionScheme			Uncompressed	NISO Z39.87-2006	(Selection). If additional values must be used, a separate list is needed
				LZW		
				JPEG Baseline sequential		
				JPEG 2000 lossy		
				JPEG 2000 lossless		
	colorSpace			RGB	NISO Z39.87-2006	(Selection). If additional values must be used, a separate list is needed
				sRGB		
				CIELab		
				Lab		
				CMYK		
				YCbCr		
				Other		(replace with appropriate value)
	iccProfileName				[Unique list along with the version number]	Depends on the workflow of MKC and restrictions in JPEG 2000
	iccProfileVersion					

	codecVersion				[Unique list to be expanded as needed]	
	codecVersion					
	sourceType			[Unique list, the same values as the digitalOrigin]		reformatted digital, digitized microfilm, born digital
	orientation				http://www.loc.gov/standards/mix/mix20/mix20.xsd	
				normal		
				normal, image flipped		
				normal, rotated 180		
				normal, image flipped, rotated 180		
				normal, image flipped, rotated cw 90		
				normal, rotated ccw 90		
				normal, image flipped, rotated ccw 90		
				normal, rotated cw 90		
				unknown		
						cw=clockwise; ccw=counterclockwise;
	imageProducer			Riksarkivet/MKC		
	scannerManufacturer				[Unique list to be expanded as needed]	Value list?
	scannerModelName				[Unique list to be expanded as needed]	Value list?
	scannerModelNumber				[Unique list to be expanded as needed]	Value list?
	scanningSystemSoftwareName				[Unique list to be expanded as needed]	Value list?
	scanningSystemSoftwareVersionNo				[Unique list to be expanded as needed]	Value list?

	digitalCameraManufacturer				[Unique list to be expanded as needed]	Value list?
	digitalCameraModelName				[Unique list to be expanded as needed]	Value list?
	digitalCameraModelNumber				[Unique list to be expanded as needed]	Value list?
	isoSpeedRatings				ISO12232	
	exifVersion				EXIF 2.2	
	lightsource				EXIF 2.2	
	autofocus				EXIF 2.2	
	samplingFrequencyUnit			The numbers 1, 2, 3	NISO Z39.87-2006	1 = no unit, 2 = inch, 3 = centimeter
	targetType			The numbers 0 och 1	NISO Z39.87-2006	0=external, 1=internal
	targetManufacturer				[Unique list to be expanded as needed]	Value list?
	targetName				[Unique list to be expanded as needed]	Value list?
	processingAgency			Riksarkivet/MKC		
	processingRationale				[Unique list to be expanded as needed]	Value list?
	processingSoftwareName				[Unique list to be expanded as needed]	Value list?
	processingSoftwareVersion				[Unique list to be expanded as needed]	Value list?
	processingActions				[Unique list to be expanded as needed]	Value list?
JPEG2000	XML-box					Should contain the parameters that were used when the file was created.
Noteringar						
	ISO8601			yyyymmdd		Several different formatting options are used in the standard, the current here is what is called "basic"
	w3cdtf			yyyy-mm-dd		Profile of ISO8601 which adds the "-" between the numbers, http://www.w3.org/TR/NOTE-datetime

National Library of Sweden

File naming convention for Swedish American Newspapers v.1.2

Change history

Date	Changes	Changes made by
2013-12-11		Jonas Ahlberg
2014-01-30	Naming convention PDF-files	Jonas Ahlberg

All filenames start with the prefix bib (file names beginning with a digit is not allowed).

libris = Libris numbers for host publication (not including the entry for the digital code or the printed original), preceded by the prefix "bib".

yyyymmdd = date of the printed newspaper number, as specified in field date in the METS file

edition = edition designation according to a unique list for the host publication. The list is constructed by the Newspaper Unit at KB. The original edition designation cannot be stored in the filename as the name often contains characters that are not allowed (eg "*"). For each host publication therefore, a translation between the edition number and a serial number is provided. An edition designation must always be mentioned in all file names. The project Swedish American Newspapers normally digitizes the main edition. If the edition name is missing, the edition is always indicated by the number 0 if no other instructions are given.

number = newspaper's serial number, as specified in field part/number of the METS file. If the number name is missing, enter the letter “s”

METS-file

Mimer (KB's ingest and storage platform) want a consistent naming convention for files that contain metadata so that the system can easily recognize them. The names shall be designed so that the metadata file should not be confused with other files. It has been decided to replace the file extension ". xml", with the ". [Name of the metadata standard]. Metadata”.

libris_yyyymmdd_edition_number.mets.metadata

Example
bib4112678_18760203_1_24.mets.metadata (standard case)
bib4112678_18760203_0_24.mets.metadata (missing edition designation on the main edition)
bib4112678_18760203_1_s.mets.metadata (missing number name)

National Library of Sweden

Image file

libris_yyyymmdd_edition_number_serialnumber.jp2

serial number = sequence number for the image in the Journal number
The serial number must be four figured

Example
bib4112678_18760203_1_24_0008.jp2

Pdf file

Issue level:
libris_yyyymmdd_edition_number.pdf
page level:
libris_yyyymmdd_edition_number_serialnumber.pdf

Example
bib4112678_18760203_1_24.pdf
bib4112678_18760203_1_24_0008.pdf

ALTO file

libris_yyyymmdd_edition_number_serialnumber_alto.xml

Example
bib4112678_18760203_1_24_0008_alto.xml

ObjektID

Each package must have a unique objectID which is registered in the attribute OBJID in the field mets in the Mets file. This id has the same structure as the file name.

libris_yyyymmdd_edition_number

Example
bib4112678_18760203_1_24

National Library of Sweden

Instruction for missing page, missing issue, missing supplement, missing section, missing newsbill

Datum	Ändring	Ändrad av
2014-06-11	Added " Missing supplement, section, newsbill"	Jonas Ahlberg

Missing page

If a page is missing, an image should be inserted containing the title of the newspaper, publication date, and the text "Missing Page". If possible, also include page numbers in the image. The image file should have the same dimensions as the other images derived from the journal issue.

The inserted image is treated as an ordinary image file, both in terms of metadata and file name. Some technical metadata will obviously be missing.

In structMap the page is assigned a LABEL with the value "missing page".

An ALTO file is created for the missing page. The file need to include only the data stored in the metadata tag <sourceImageInformation>.

Missing issue

If a whole issue is missing, still a SIP should be constructed according to the specification. The package will contain a single image file that shows the title of the newspaper, publication date, and the text "Missing issue". The image file should have similar dimensions as the image files derived from the equipment in which the newspaper would likely to have been captured. All bibliographic metadata in code level is handled as usual. Since the only image file is not captured in the usual way, however, some metadata fields are left blank (e.g. technical metadata).

The inserted image is treated as a regular image file. Data about the edition in the file name is specified with the main edition unless otherwise announced.

In structMap the image file is assigned a LABEL with the value "missingissue".

National Library of Sweden

Missing supplement, section, newsbill

If an entire supplement, section or newsbill of the newspaper is missing, insert an image that contains the title of the newspaper, publication date, and the text "Missing supplement", "Missing section" or "Missing newsbill."

The inserted image is treated as a standard image file, both in terms of metadata and the file name. The file has the same name as the first "real" page that the supplement/section would have received. For newsbills there is no numbering. Since the only image file is not an image captured in the usual way, some non-mandatory metadata fields have to be omitted (eg, MIX metadata in the premis element objectCharacteristicsExtension).

In structMap supplement/section/newsbill is assigned a "div" with one of the values "supplement", "section" or "newsbill." The inserted image file that serves as a "placeholder" is given the attribute LABEL with the value "missing page".

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

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	