
CCS

CCS
Content Conversion Specialists

METS / ALTO introduction

Why would I use METS and ALTO?

- “I digitize lots of different items and each type is digitized to a different format. Some are Word, some are PDF, some are XML, some are just JPG – I need it all to be the same, but what is the best way?”

- “I’d like to offer full text search to the scientists and researchers in the complete collection, so I need to build a text index of the complete collection – and I might want to change the presentation system in three years, so I need the source data in a non-proprietory format”

- “I need to organize my digitization project accordingly to long term preservation standards – how long will PDF exist? I might need something more robust”

What is METS?

- METS -> Metadata Encoding and Transmission Standard
- Established in 2001
- XML based open standard
- Schema is hosted at Library of Congress (LOC)
- Maintained by METS Editorial Board
- Current version: 1.10
- Usesd for longterm preservation

What does METS do?

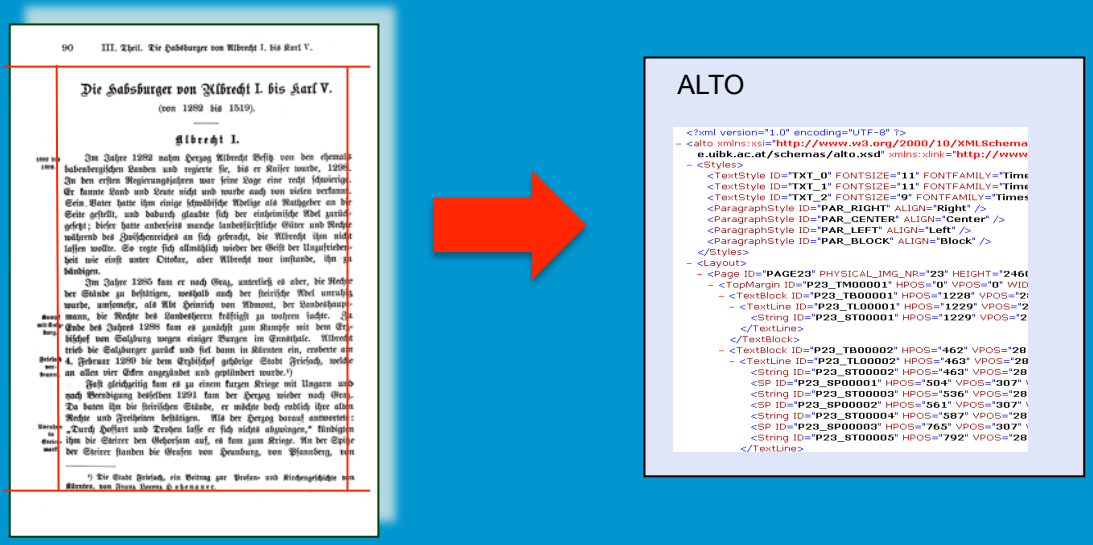
- METS files are used to describe a digital object
  - printed media (book, newspaper, journal), audio or video media
  - any other item
- METS usually embeds different kinds of metadata / metadata standards
  - Descriptive metadata (DC, MODS, MARC21)
  - Administrative metadata (Mix, PREMIS, …)
- May contain structural information
  - Physical structure (page based / file based)
  - Logical structure (chapter based / article based / track based)
- May link to any other digital objects
  - Image files, Audio / video files, Text files
  - External metadata objects

What is ALTO?

- ALTO –> Analyzed Layout and Text Object
- XML based open standard
- Schema is hosted at Library of Congress (LOC)
- Maintained by ALTO Board
- Current version: 2.1 (Draft 3.0)

What does ALTO do?

- Contains the content of a single page
- Describes the layout of a printed page to re-build the original page
- Describes the styles, layout and block type information
- May contain tags which contain more information about content (e.g. **named entities**)



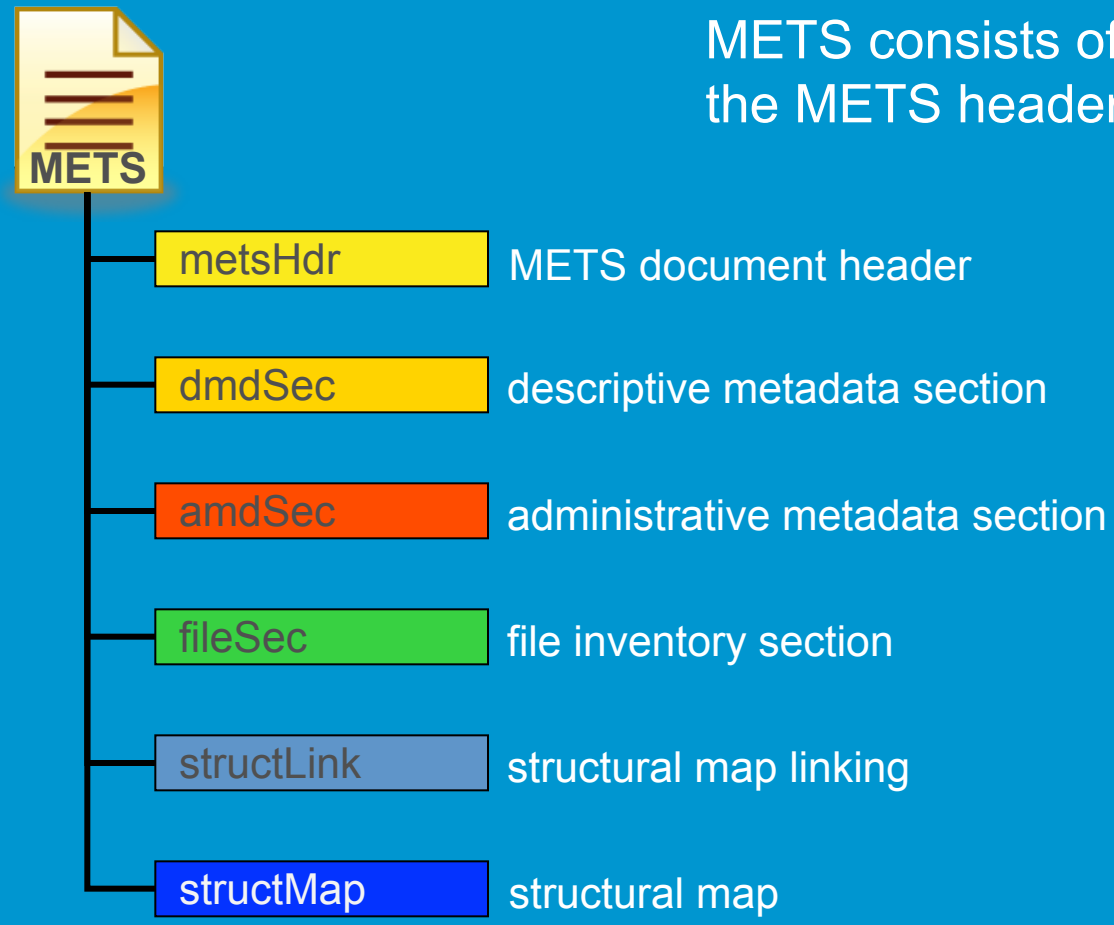
Why would you use XML standards?

- Fully documented XML format
- Can be used by any IT party later on
- Can be transformed to other formats in the future (for longterm preservation)
- Readable by humans

What are your benefits using METS/ALTO?

- It is the current industry standard for digitization used by hundreds of libraries and content providers
- The long-term sustainability of your digital objects is greatly enhanced
- Supports article and chapter segmentation
- You can handle objects in an easy way and exchange them with other parties
- You can create PDF, EPUB, DAISY and other formats from METS/ ALTO
- It is constantly maintained and developed (hosted by the Library of Congres)

How does a METS file looks like?



METS consists of five sections plus the METS header

What information include the METS sections?

- dmdSec – Descriptive Metadata Section, per logical unit
  - title and author, publishing date, language, copyright details, etc
- amdSec – Administrative Metadata Section, per image
  - resolution, bit depth, dimensions, source file name, scanner name, etc
- fileSec – File Inventory Section
  - File inventory and referencing (pointers to each image file, each ALTO file, derivatives)
- StructLink – Structural Link
  - Describes the logical structure of a complex digital object
- StructMap – StructuralMap
  - Decribes page by page the structure of the digital object

Summary | METS and ALTO?

- Established for the description of digitization of printed material
- The idea was to split the descriptive information and the content itself
- In order to handle objects in an easy way
- If all data would be in one XML file (e.g. TEI) the XML file would be huge
- METS and ALTO are maintained and continuously developed by libraries (hosted at Library of Congress)
- METS/ALTO data can be exchanged with other parties
- Many vendors are offering products for handling METS/ALTO data
- You can make PDF, EPUB, DAISY and other formats from METS/ALTO

More Information

- http://www.loc.gov/standards/mets/
- http://www.loc.gov/standards/alto/
- http://en.wikipedia.org/wiki/ALTO_(XML)
- http://www.veridiansoftware.com/knowledge-base/metsalto/
- http://www.nla.gov.au/ndp/project_details/documents/ANDP_Use_of_METSv2.pdf

Contact

CCS Content Conversion Specialists GmbH
Weidestr. 134
22083 Hamburg
Germany

dWsupport@content-conversion.com
www.content-conversion.com

Disclaimer

The information contained in this document has been prepared for the sole purpose of providing information about theme described in the following title. The material herein contained has been prepared in good faith; however, CCS disclaims any obligation or warranty as to its accuracy and/or suitability for any usage or purpose other than that for which it is intended.


CCS