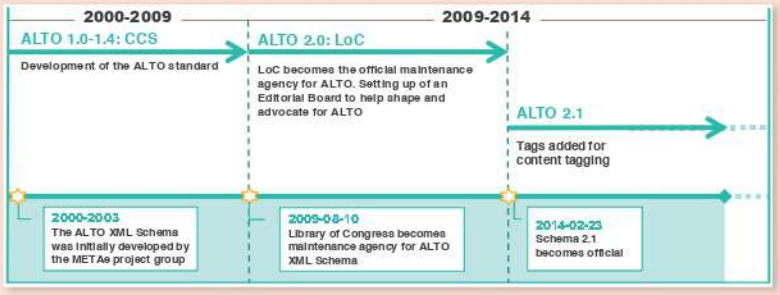
New features in ALTO XML

Jean-Philippe Moreux, Frederick Zarndt

New features in ALTO XML:
Tags for layout, structure, roles, named entities…

ALTO (Analyzed Layout and Text Object) is a widely used XML based open standard for describing the layout and content of text images. ALTO is most often used with METS XML and TIFF or JPEG2000 images to create digital surrogates for books, journals, magazines and newpapers. ALTO XML is implemented in text digitisation projects around the world.



ALTO is an open standard, first administered and maintained by Content Conversion Specialists (CCS), one of the MetaE project members. In 2009, CCS transferred its administration to the ALTO Editorial Board whose members are from libraries and private organizations around the world. Library of Congress is the maintenance agency for ALTO Schema.

3 Use Cases for Tags in the Digital Library Community (Institutions, Service Providers, Research Projects)

● Logical Labelling of Structural Elements

The ALTO format captures the layout and the full text of a page. Although this OCRing provides full text retrieval, that retrieval can benefit from marking additional structural elements such as:

- running titles,
- page numbers
- captions
- marginalia, footnote

Unlike entities which record the physical and logical structure of an entire document (in a format such as METS), these elements are specific to the page.


DE LIBRA 10

Producatur FG vsq; ad mundi centrum, quod sit S. & à puncto S circulum AFBG contingens ducatur, neq; enim linea à puncto S circulum contingere potest in A; sam ducta AS triangulum ACS duos haberet angulos rectos, nempè SAC ACS, quod est impossibile. neq; supra punctum A in circumferentia AF continget; circulum enim secaret. tanget igitur infra, sitq; SO. connectatur deinde SD SL. quæ circumferentiam AOG in punctis KH secent. & CK CH contingantur. Et quoniam pondus, quanto propius est ipsi F, magis quoque initiatur centro; vt pondus in D magis versionis puncto C innitur tanquam centro; hoc est in D magis supra lineam CD grauitat, quàm si esset in A supra lineam CA; & adhuc magis in L supra lineam CL; Nam cùm tres anguli cuiuscunq; trianguli duobus rectis sint æquales, & trianguli DCk æquicruris angulus DCk

18 Tertii.

Gallica.bnf.fr / Bibliothèque nationale de France

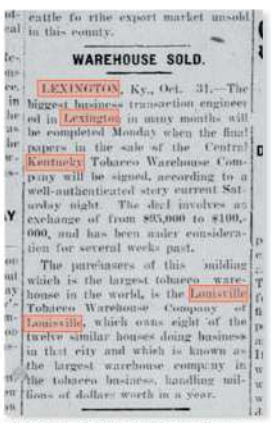
● Layout/Content Tagging

Before the actual character recognition starts, most OCR software analyses the page structure to identify regions. Modern OCR software supports numerous region types that ALTO can record: tables, graphics, music scores, maps, etc. These regions can be labelled either by the software or manually by operators. Like labelling elements, labelling regions helps improve retrieval.

Gallica.bnf.fr / Bibliothèque nationale de France

● Named Entities & Roles

Named-entity recognition (NER) is a subtask of information extraction that seeks to locate and classify discrete textual elements into predefined categories (names of persons, organizations, locations, expressions of times, quantities, etc.). Role tagging describes how entities were involved in the content creation: author, editor, publisher, distributor, etc.

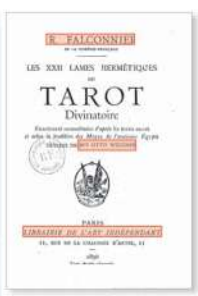

cattle to the export market unsold in this county.

WAREHOUSE SOLD.

LEXINGTON, Ky., Oct. 31.—The biggest business transaction engineered in Lexington in many months will be completed Monday when the final papers in the sale of the Central Kentucky Tobacco Warehouse Company will be signed, according to a well-authenticated story current Saturday night. The deal involves an exchange of from $05,000 to $100,000, and has been under consideration for several weeks past.

The purchasers of this building which is the largest tobacco warehouse in the world, is the Louisville Tobacco Warehouse Company of Louisville, which owns eight of the twelve similar houses doing business in that city and which is known as the largest warehouse company in the tobacco business, handling millions of dollars worth in a year.

Chronicling America: Historic American Newspapers. Lib. of Congress


S. FALCONNIER

LES XXII LAMES HERMÉTIQUES DU

TAROT

Divinatoire

PARIS

Gallica.bnf.fr / Bibliothèque nationale de France

Implementers of ALTO 2.1

European Newspapers project

- Named Entities Recognition: EN project has implemented an ALTO 2.1 output for its NER task. Persons, locations and organizations are recognized from hundred thousands of European newspapers.

BnF Mass Digitization program

- Markup of important words: The next BnF large-scale digitization program makes use of NER tagging to identify meaningfull words in OCR results and to concentrate the post-correction task on these specific words.

The ALTO Board involves members from various origins

LIBRARY OF CONGRESS · BnF · 16 40 THE NATIONAL LIBRARY OF FINLAND · CCS · BRITISH LIBRARY · KB · NLB · CDNC California Digital Newspaper Collection