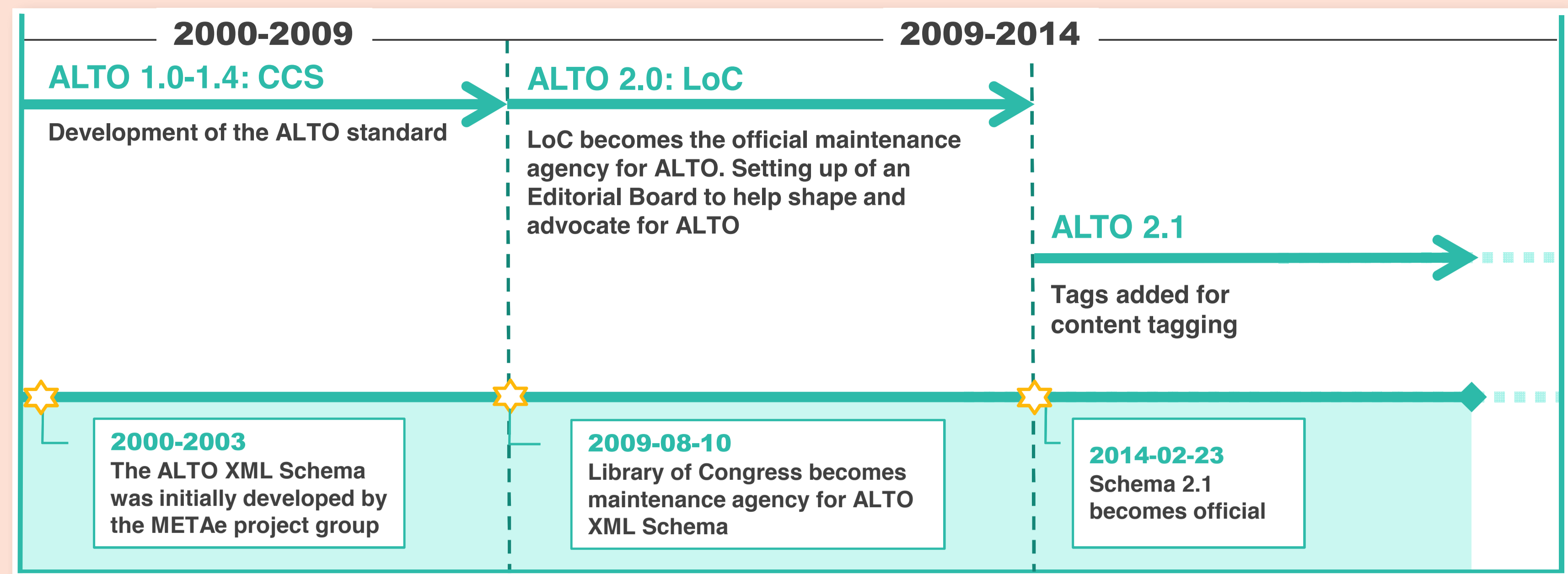


New features in ALTO XML: Tags for layout, structure, roles, named entities...

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 newspapers. 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.



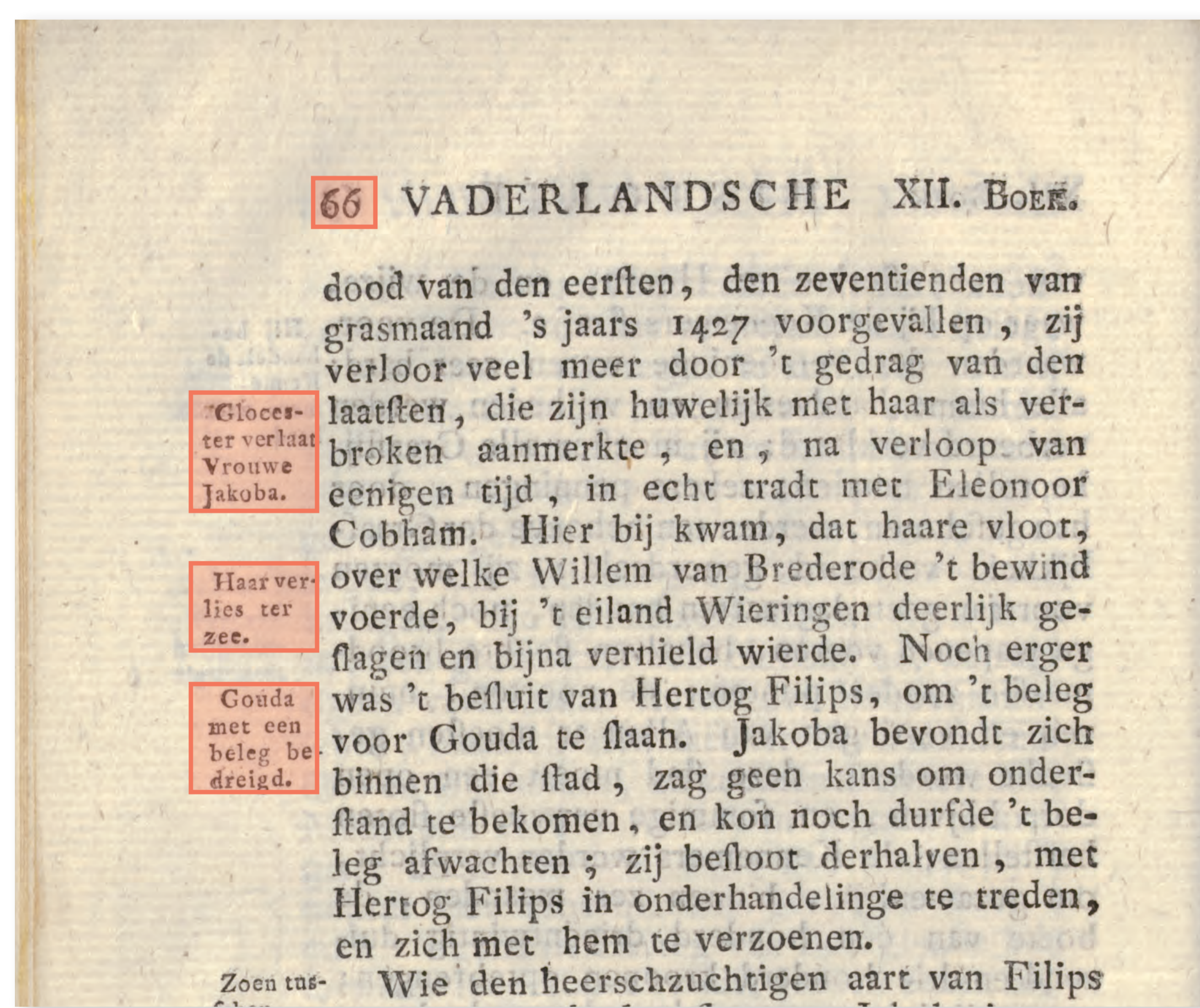
3 Use Cases for Tags in the Digital Library Community (Institutions, Service Providers, Research Projects)

● 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.



<http://kb.nl> / National Library of the Netherlands

● Layout/Content Tagging

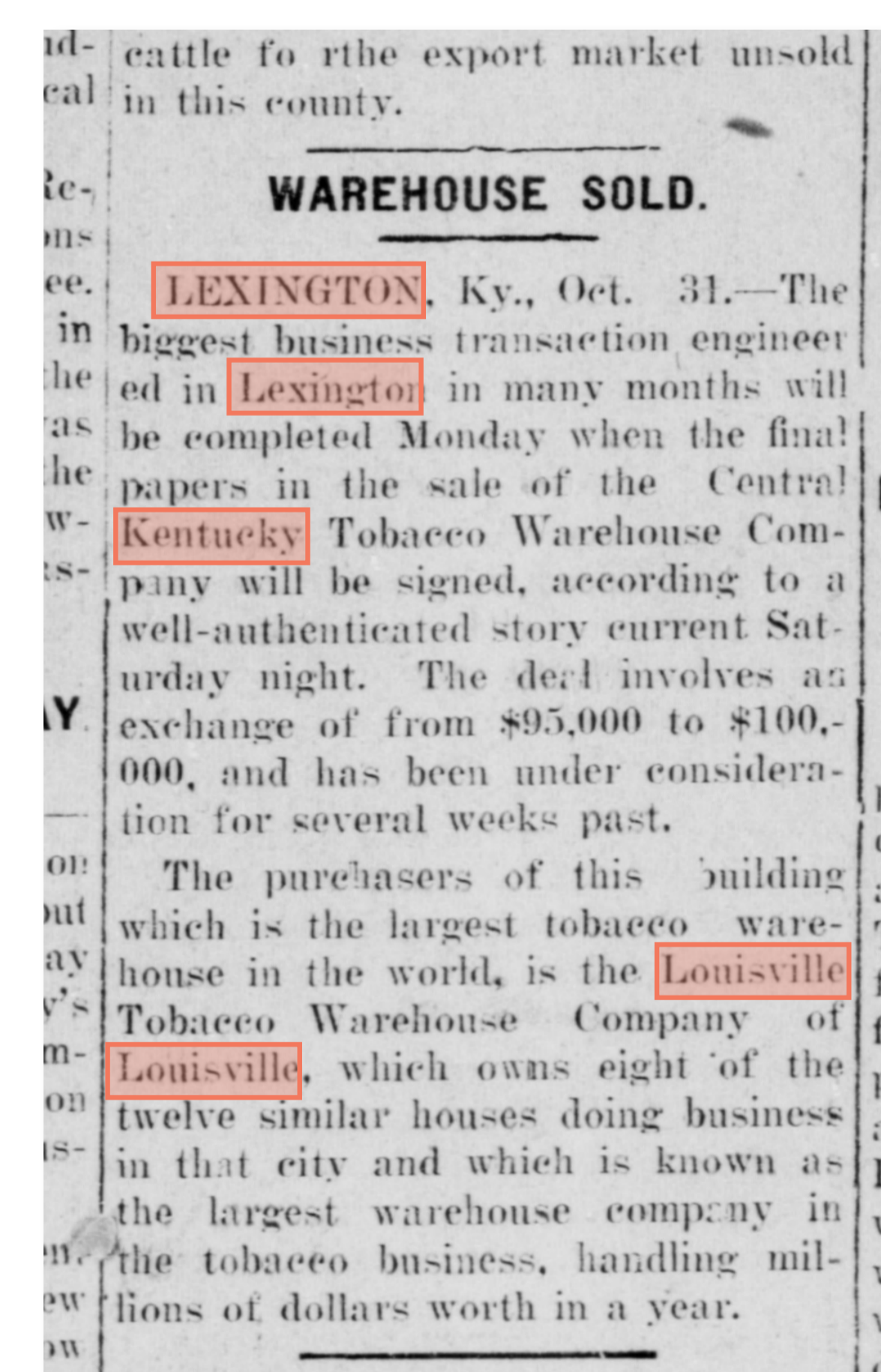
Before the actual character recognition starts, most OCR software analyses the page structure to identify regions. Modern OCR software supports numerous region types that ALTO can record: tables, graphics, music scores, maps, etc. These regions can be labelled either by the software or manually by operators.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



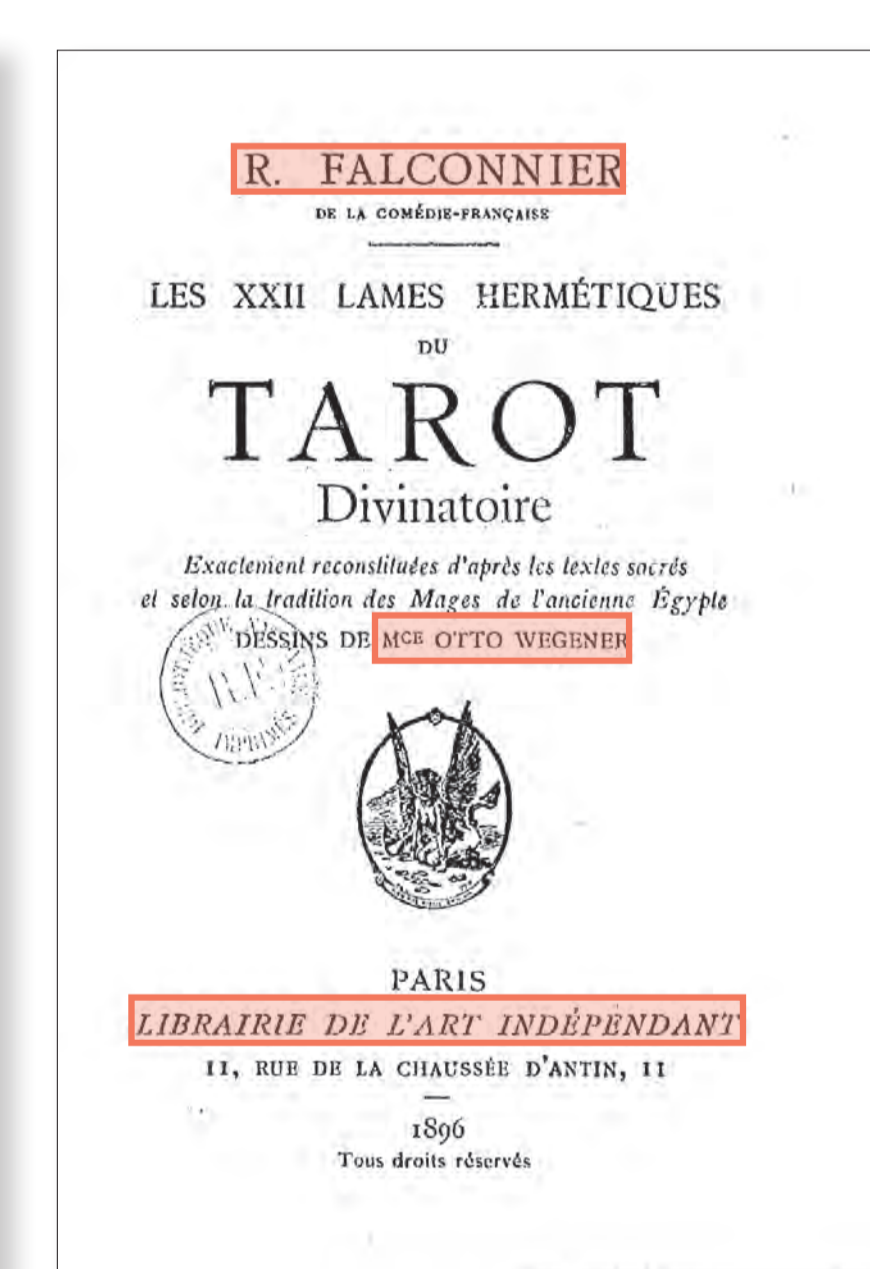
Gallica.bnf.fr / Bibliothèque nationale de France

● 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.



Chronicle America: Historic American Newspapers. Lib. of Congress



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 Bibliothèque nationale de France large-scale digitization program makes use of NER tagging to identify meaningful words in OCR results and to concentrate the post-correction task on these specific words.

