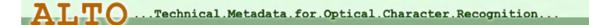
ALTO tags Use Cases

Version 16-01-2014



ALTO Board

Copyright © 2014 ALTO Board. All rights reserved.

No part of this publication may be reproduced, stored in databases, or transferred in any form (electronically, photo-mechanically, chemically, manually, or otherwise) without the express written permission of the ALTO Board. The software described in this manual is licensed software that may be used only in compliance with the licensing terms and conditions. The ALTO Board reserves the right to make changes to the content of this manual without notice. The ALTO Board makes no guarantee regarding the accuracy of the information provided in this manual. Microsoft, MS-DOS, and Windows are registered trademarks of the Microsoft Corporation.

Parts of the software uses the Duden-Proof-Factory of the Brockhaus Duden Neue Medien GmbH for syllable separation.

Product or company names that are mentioned may be trademarks or registered trademarks of the respective company. The ALTO Board uses these names and trademarks in the following manual merely for explanatory purposes and for the benefit of the respective user, and such use does not imply trademark infringement.

Under this software license, you are only permitted to reproduce materials that are not protected by copyright laws. This excludes only materials where you hold the copyright and/or legal permission to reproduce copyrighted materials. If you are uncertain about the copyright status of certain materials then please seek legal counsel. The ALTO Board holds no liability over copyright violations resulting from the use of this software.

Last updated: 15.01.2014

The ALTO Board



Table of Contents

| 1. | Introduction | 4 |
|------|----------------------------|----|
| 1.1. | XML schema | 4 |
| 1.2. | Good pratices, limitations | 4 |
| 1.3. | Uses cases | 4 |
| 2. | Layout tagging | 5 |
| 2.1. | XML schema | 5 |
| 2.1. | Good pratices, limitations | 5 |
| 3. | Structural tagging | 21 |
| 3.1. | XML schema | 21 |
| 3.2. | Good pratices, limitations | 21 |
| 4. | Named Entities tagging | 45 |
| 4.1. | XML schema | 45 |
| 4.2. | Good pratices, limitations | 45 |
| 5. | Reference documents | 49 |



1. Introduction

The ALTO version 2.1 introduces a new mechanism: tag (or annotation). The aim of this mechanism is to cover several change requests concerning the ALTO format:

- Layout labelling ALTO format should record various region types (tables, graphics, maths formulas, music scores, etc.)
- Logical Labelling Of Structural Elements Structural tagging of documents offers a large number of benefits. Full text search can be done in a much more focused way, reprinting of digitised books could be made easier, etc.
- Named Entity markup Post-OCR processing on page images often includes named-entity recognition attempts.

XML schema

This mechanism has been introduced in ALTO version 2.1 schema. In XML terms, a reference to a new element group "Tags/Tag" allows to add additional information to the contents which are referring to these tag elements:

Good pratices, limitations

- The tag reference must be set on the upper level available (eg on TextBlock, not on all TextLines or Strings of the block).
- A content crossing two or more pages can be described: the blocks composing this content must reference the same tag ID.
- If the IDs of the ALTO content elements change, the tag references must be updated.



For readibility purpose, IDs have been omitted in the following examples.

ALTO tags Version: 23.01.2014 4 / 53



2. Layout tagging

Before the actual character recognition starts, any OCR software analyzisis the page structure to identify regions. Modern OCR software supports various region types than ALTO could record: tables, graphics, music scores, etc. Some of these regions can be categorized for the OCR software to use a specific algorithm or process when recognizing their characteristics; they can also be identify with a manual labelling.



ALTO tag element: <LayoutTag>

Tag attributes:

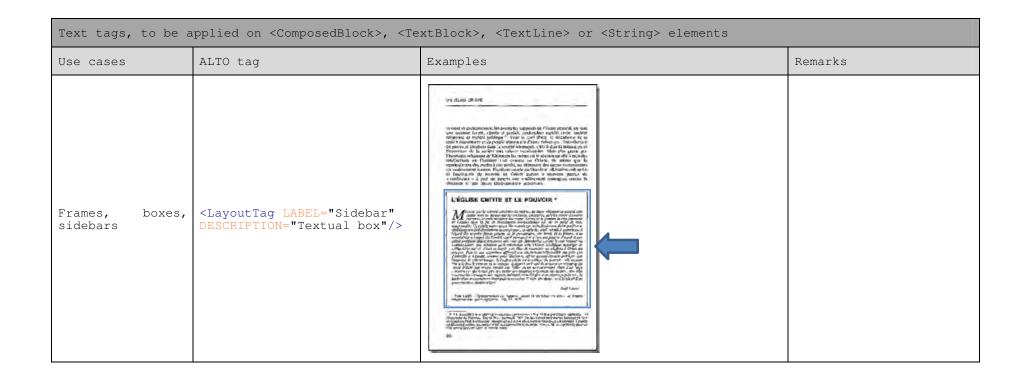
- ID: Tag id
- LABEL: Region type
- TYPE (optional): Suppertype or category of the region type
- DESCRIPTION (optional): Description of the region, description of the region content, or the content itself
- XmlData (optional): Any XML encoded metadata helping to describe the tag, or the vocabulary used to describe the tag
- URI (optional): Any authority URI relevant for the tag

Good pratices, limitations

The intent of layout tagging is to identify types of content. Sometimes, specific types of content can also have a
functional value (sidebars, separators, etc.) and the question of using a layout tag or a structural tag may be
raised.

ALTO tags Version: 23.01.2014 5 / 53





ALTO tags Version: 23.01.2014 6 / 53



| Tables | <layouttag description="Table of numbers " label="Table"></layouttag> | Situation de la Banque d'Angleterre, aux dates suivantes : | |
|-----------------------|---|---|---|
| Mathematical formulas | <layouttag description="4 maths formulas" label="MathFormula" type="Formula"></layouttag> | Cela posé, le tétraedre devant être en équilibre, sous l'action des forces élastiques qui s'exercent sur ses quatre faces, et des forces qui sollicitent sa masse, les sommes des composantes de ces forces, estimées suivant chaque axe, devront être nulles. A la somme des composantes suivant l'axe des x, la face inclinée fournira le terme Xw; la face a, le terme — N ₁ .mw, la face b, —T ₃ .nw, la face c, —T ₂ .pw; les forces qui agissent sur la masse donneront un terme égal à p X ₀ multiplié par le volume du tétraèdre, qui est un infiniment petit du troisième ordre; ce cinquient terme disparatira donc à la sinte des quatre autres, qui sont des infiniment petit du sesoni ordre. Realant la soume fouvée à site, et disparatir par x, on obtient la première des équations. | On <textline> elements On <string> elements</string></textline> |

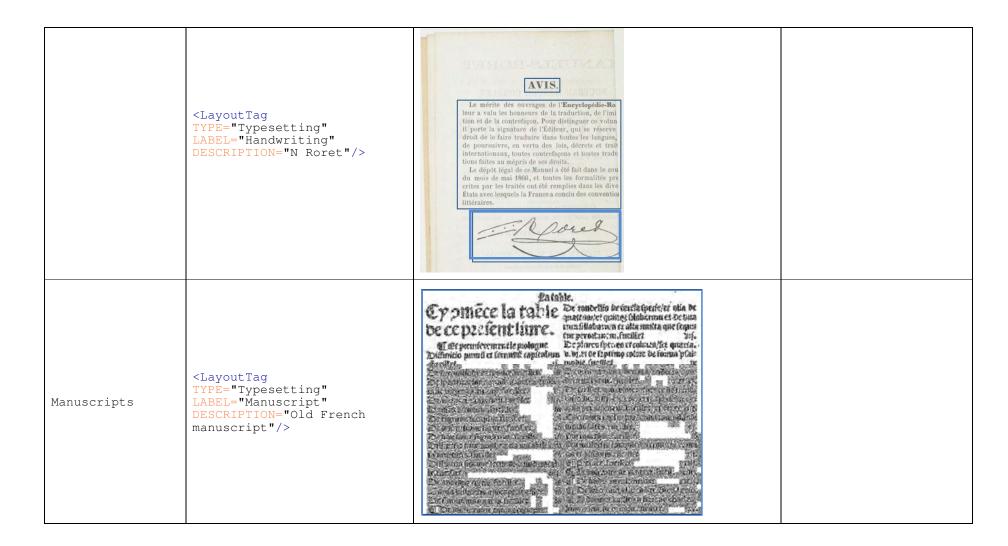
ALTO tags Version: 23.01.2014 7 / 53



| Chemical formulas (text) | <pre><layouttag description="Ka formula" label="ChemFormula" type="Formula"></layouttag></pre> | $K_{A} = \frac{\left[A^{-}\right]_{q} \cdot \left[H_{3}O^{+}\right]_{q}}{\left[AH^{-}\right]_{q}}$ | |
|---|--|---|--|
| Physics formulas | <layouttag description="e and v formulas" label="PhyFormula" type="Formula"></layouttag> | 18. Exercices: Les formules qui donnent l'espace et la vitesse dans le mouvement de chute libre, le corps tombant sans vitesse initiale, sont $(9) \qquad e = \frac{1}{2} \ gt^2, \qquad (10) \ v = gt. \qquad ,$ En éliminant t entre ces deux équations, on a $(11) \qquad e = \frac{v^2}{2g} \qquad \text{ou} \qquad v = \sqrt{2ge},$ expression utile de la vitesse après une chute d'une hauteur e . S'il y a une vitesse initiale v_0 , les formules deviennent | |
| Text content under a stamp | <layouttag description="PAR" label="TextStamped"></layouttag> | L'ORGANICI | See also "Stamps" in the GraphicalElement section |
| Handwriting contents, annotations, initials, etc. | <layouttag description="43g8" label="Handwriting" type="Typesetting"></layouttag> | ICHE 4398 | Handwriting could also be described with a <graphicalelement></graphicalelement> |

ALTO tags Version: 23.01.2014 8 / 53





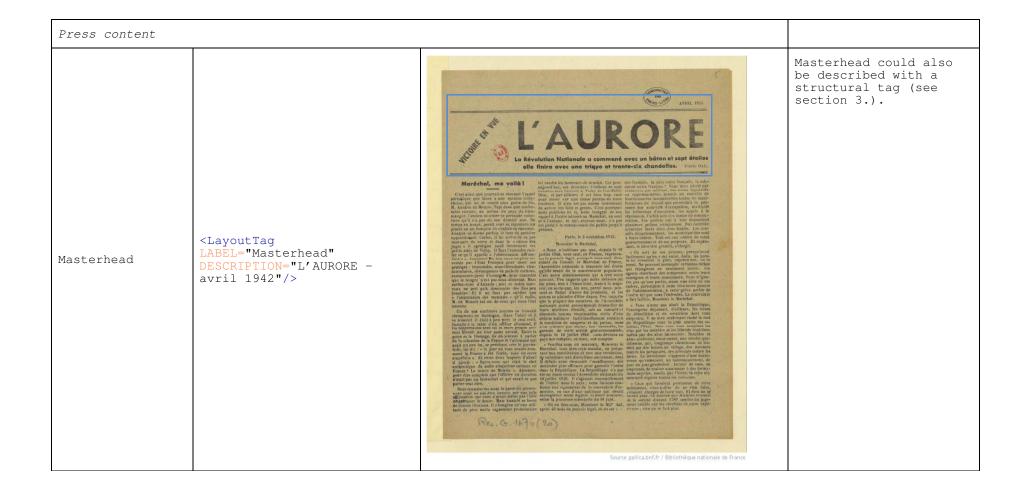
ALTO tags Version: 23.01.2014 9 / 53



| Script fonts | <layouttag description="Script font content" label="ScriptFonts" type="Typesetting"></layouttag> | Platinum Gifts for Her Precious Gems Legacy Collection Exoile Diamond& Silver | |
|-----------------|--|---|-------------------------------|
| Non-latin fonts | <layouttag description="Chinese ideograms" label="NonLatinFont" type="Typesetting"></layouttag> "" | INDEX 1NDEX 29 | Pages 43 32 10 34. 9,10 42 17 |

ALTO tags Version: 23.01.2014 10 / 53





ALTO tags Version: 23.01.2014 11 / 53



Advertisements, announcements, public relation pages

<LayoutTag
LABEL="Advertisements"
DESCRIPTION="AVIS"/>



ALTO tags Version: 23.01.2014 12 / 53



| Small ads, obituaries, weather forecast, train tables, etc. | <pre><layouttag description="Petites annonces" label="SmallAds"></layouttag></pre> | The Comment of the control of the co | |
|---|--|--|--|
| | | | |

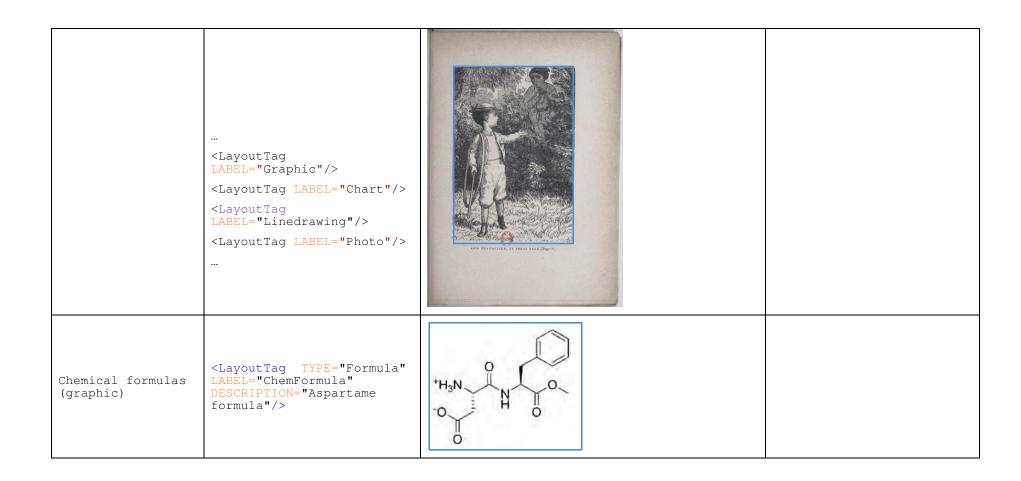
ALTO tags Version: 23.01.2014 13 / 53



| Graphical tags, to be applied on <composedblock> or <illustration> elements</illustration></composedblock> | | | |
|--|--|---|---------|
| Use cases | ALTO tag | Examples | Remarks |
| Graphics, charts or graphs, linedrawings, maps, photographs, engravings, etc. | <pre><layouttag label="Map"></layouttag> <layouttag label="Engraving"></layouttag></pre> | CONSIGNED HARD STATES AND | |

ALTO tags Version: 23.01.2014 14 / 53





ALTO tags Version: 23.01.2014 15 / 53



| Musical notations | <layouttag description="Maurice Ravel" label="MusicalScore"></layouttag> | | |
|------------------------|--|---|---------|
| | | | |
| Other graphical tag | s to be applied on <composedb< td=""><td>lock>, <graphicalelement> elements</graphicalelement></td><td></td></composedb<> | lock>, <graphicalelement> elements</graphicalelement> | |
| Use cases | ALTO tag | Examples | Remarks |
| Ornaments, tail pieces | no need | | |

ALTO tags Version: 23.01.2014 16 / 53



| Horizontal separator lines between paragraphs | <pre><layouttag label="TransitionSep"></layouttag></pre> | atteignatial circuiron 1 2.2 a Banque, consignation 2 2.3 a Banque, a Banque | |
|---|--|--|--|
| Horizontal separator lines between text block and footnotes | <layouttag label="FootnoteSep"></layouttag> | ELETTRES DE JEAN CHAPELAIN. Byer', de Giry', ne vous y sees pas imolié. Ces sutheurs seed purs, et l'en ne pout errete nie se suyvant, pourveu que, comme friscit à Père Le Moine, on reu porte point l'imitation un étà de borsons qu'il s' y sont prescrition. Métà de de sade pour une parsonne ususi occupée que vous. In anis, Mon Réserted Père, etc. Bu-Beste, es ma pina-sépal. 1 Sur l'andémicien Dierre De Rye, vois dans aobet tons l'. In onts a de la page 35. 2 Sur l'andémicien Dierre De Rye, vois dans aobet tons l'. In onts a de la page 35. 3 Sur l'andémicien Dierre De Rye, vois dans aobet tons c'il not est de la page 35. 4 Sur l'andémicien Louis Giry, vois (skir) la cité de Cheffe à M. de la Papositir es el. avi sod sur de la page 35. 5 Le 15 juin, Chapelin (18 2gy v) 'paré sind de Cheffe à M. de la Papositir es el. avi sod sur de me étite en conseque d'épuès seront apparement in la criscoure certain d'une si home main, Vous me me étite seron excertain d'une si home main, Vous me me étite seron excertain d'une si home main, Vous me me étite seron excertain d'une si home main, Vous me me étite seron excertain d'une si home main yeu de la couver en gredit est pas de la chape que l'annément de ma de la chape que l'annément de ma de l'annément de la chape que l'annément de ma de l'annément de la chape que l'annément de me faire nichter avec le maint de me de l'annément de la chape que l'annément de ma de l'annément de la chape que l'annément de l'annément de la chape que l'annément de l'annément de la chape que l'annément de la maint de cante de l'annément de l'annément de l'annément de la chape que l'annément de la chape que l'annément de l'annément d | |

ALTO tags Version: 23.01.2014 17 / 53



| Stamps | <layouttag description="Dépôt légal Vosges 1891" label="Stamp"></layouttag> | PELLERIN & C**, imp6dit. A-NE |
|------------------|---|---|
| Dropped initials | <layouttag description="A" label="DropCap"></layouttag> | JAN MARTIN SECRETAIRE DE MONSEI- gneur le Reverendissime Cardinal de Lenoncourt. AUX LECTEURS. I vous desirez, Messeigneurs, entendre à peu de parolles ce qui est contenu en cest œuvre, sachez que Poliphile dict avoir veu en songe des choses admirables, entre lesquelles il en descrit plusieurs antiques dignes de memoire, comme Pyramides, Obelisques, grandes ruines d'edifices, la difference des colonnes, leurs mesures, piedestalz, bazes, et chapiteaux dont elles sont ornées. Puis les architraves, frizes, cornices, et frontispices avec leurs ouvrages. Un grand cheval, un Elephant de merveilleuse grandeur, un Colosse, et une porte magnifique, avec son plant, ordonnance, moulures, et besongne de taille. Apres comment cinq belles Nymphes le menerent |

ALTO tags Version: 23.01.2014 18 / 53



| Misc, to be applied | on any elements | | |
|---------------------|---|--|---|
| Use cases | ALTO tag | Examples | Remarks |
| Illegible contents | <layouttag description="scan quality problem" label="Illegible"></layouttag> | Applianter tole one ied a mo potentishin , e trè altre oncre di oglio commune, e vi fi mette dentro va podrecolloquinitala, poi pollo titto al fuoco in pignatti-lidecia, che relitato i decido, e ficola, fer soi patto e di nacine il corpo, il desellare callefo inicito, como di quel indicali espa, e fi mette na fico e con di quel indicali inche il corpo, il desellare callefo inicito e con di quel indicali ombelico e fi rong po con per mare hora e caracteria, e glazante all inche. Si e provino, che apera, e con con di quel indicali ombelico e fi rong po con per mare persone. Il l'accuse na consentatione e fi aper montre corpo. | On <textblock> or <textline> elements.</textline></textblock> |
| | <layouttag description="illegible word, may be 'Scribe'" label="Illegible"></layouttag> | faites et à faire, et non de trans cela résulte de tous les documents d la correspondance et de l'acte de illite de Flament, déclarée le 1er oct mbre 1859, où Scribe demandé à n re, comme étant nanti d'un gage, d du Code de commerce; | On <string> element.</string> |

ALTO tags Version: 23.01.2014 19 / 53



| Noise | <layouttag description="scan noise" label="Noise"></layouttag> | itable nuée de reporters, non mais de tous les pays du obligé de l'interner sur-le-cl II UN ARTICLE DE PIERRE HUMBOURG, DANS LE « MATIN », — LA SIGNATURE DU VAMPIRE, — LE SOUVENIR DE JACK « L'ÉVENTREUR. — ENCORE UNE FAUSSE PISTE. — LA PIOCHE ET LA PELLE. — UN TYPE ÉTRANGE. — L'HOMME-FEMME. — ATMOS-PHÈRE DE LÉGENDE. — LE CROQUIS RÉVÉLATEUR. — LES QUARANTE COMMISSAIRES. — LA FEMME EN SOIE ROSE. — LA GUINGUETTE DU MOULIN. — « NITCHS », — LA FEMME MYS-TÉRIEUSE. — LE VAMPIRE ÉCRIT ÉCRIT — UN FOU OU UN FARCEUR ? OUTIEUR de ses informations, de ses lecteurs en envoyant à meilleurs rédacteurs, Pierre « Los neuf victimes porte tonu à la partie la plus sens | data, only artifacts on the document or scanner noise. |
|-----------------|--|--|--|
| Unknown content | <pre><layouttag description="text?" label="Unknown"></layouttag></pre> | | To be used if the region type cannot be ascertained. |

ALTO tags Version: 23.01.2014 20 / 53



3. Structural tagging

The ALTO format captures the layout and the full text of a page. One intention of OCRing is full text retrieval. But full text retrieval may benefit from marking additional labelling elements such as page headers (running titles), page numbers, signature marks, etc. These are different type of structural elements than intellectual entities which record the intellectual structure of a document (The intellectual structure is not recorded in ALTO, but in a container format such as METS.)

If these elements are labelled, a full text search could either be restricted only to these elements, or the elements could be excluded from a search. A running title that appears on every page could e.g. manipulate the ranking of a resource or delivery invalid hits. Other labelled information such as the page number could be used for quality assurance purpose, document navigation, etc.

This structural tagging can be implemented in a mass digitisation workflow to extract structural or functional features (TOC entries, headings, headers, footnotes, article headings, etc.) or as an editing and correction facility for the improvement of already digitised books. Apart from METS/ALTO, enhanced PDFs and ebooks can also be generated if some structural information is available.



ALTO tag element: <StructureTag>

Tag attributes:

- ID: Tag id
- LABEL: Structural tag
- TYPE (optional): Suppertype or category of the tag
- DESCRIPTION (optional): Description of the tag, description of the content, or the content itself
- XmlData (optional): Any XML encoded metadata helping to describe the tag or the vocabulary used to describe the tag.
- URI (optional): Any authority URI relevant for the tag element

Good pratices, limitations

- Structural tagging in ALTO can be considered as awkward (compared to the classic and recommanded METS-ALTO solution), as ALTO does not foresee the storing of any semantic information. But it could be useful:
- If a METS container is not used in the digitization project.
- If a simple and limited capturing of structural information is needed.
- Besides, tagging in ALTO has some benefits: engineering simplicity (compared to a METS-ALTO implementation), autonomy (content and structure are stored in a unique file).

ALTO tags Version: 23.01.2014 21 / 53



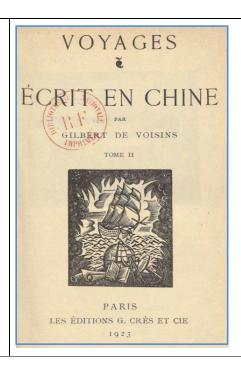
| Text tags, to be applied on <composedblock>, <textblock>, <textline> or <string> elements</string></textline></textblock></composedblock> | | | | |
|---|--|--|---|--|
| Use Cases | ALTO tag | Examples | Remarks | |
| Covers | <pre> <structuretag label="Cover" type="Functional"></structuretag> <structuretag label="cover" type="Functional"> <xmldata></xmldata></structuretag></pre> | FARIS PARIS PA | http://gallica.bnf.fr/ark :/12148/bpt6k6207567r/f1. planchecontact Description of the vocabulary used. | |

ALTO tags Version: 23.01.2014 22 / 53



Title pages TYPE

<StructureTag
TYPE="Functional"
LABEL="TitlePage"/>



ALTO tags Version: 23.01.2014 23 / 53



<StructureTag TYPE="Functional" LABEL="Catalog"/> DU MÊME AUTEUR La petite Angoisse, roman. Pour l'Amour du Laurier, roman. Frontmatter Le Démon secret, roman. (foreword, Sentiments, critique.
Les Moments perdus de John Shag. prologue, notice to the reader, Le Bar de la Fourche, roman. L'Enfant qui prit peur, roman. warning, <StructureTag Le Mirage, roman. advertisements, TYPE="Functional" L'Esprit impur, roman. Fantasques, petits poèmes. La Conscience dans le Mal, roman. copyright page, LABEL="CopyrightPage"/> publisher <StructureTag catalogs, etc.) TYPE="Functional" Prochainement. LABEL="Foreword"/> Le Jour naissant. <StructureTag TYPE="Functional" LABEL="Notice"/>

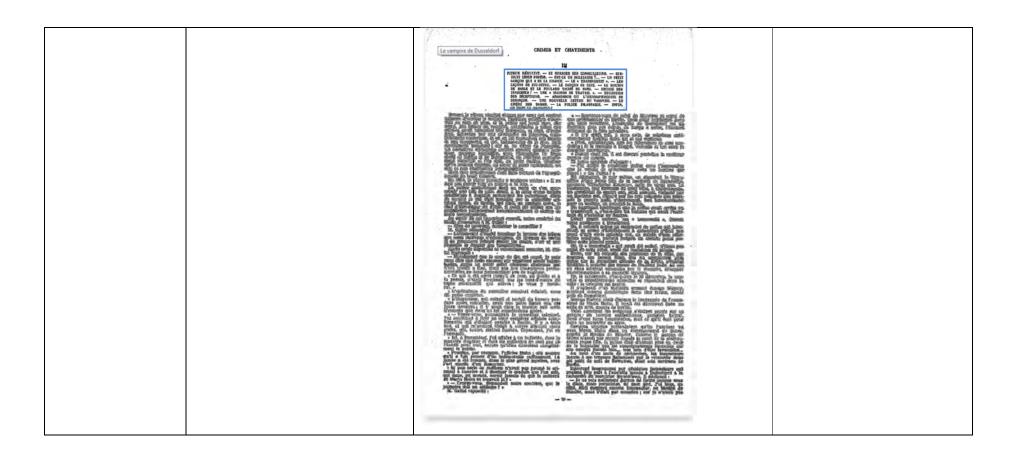
ALTO tags Version: 23.01.2014 24 / 53



Global tables of content TABLE DES MATIÈRES TOME I TOME II <StructureTag Tables of content TYPE="Functional" IV. De Lan-tcheou à Tchong-king LABEL="TOC"/> V. Le Belour Le Souvenir de Segalen. . . Internal tables content (chapter beginnings, etc.)

ALTO tags Version: 23.01.2014 25 / 53





ALTO tags Version: 23.01.2014 26 / 53



| Body matter | <pre><structuretag label="BodyMatter" type="Structural"></structuretag></pre> | M'y voilà donc, dans ce canton lointain que je m'étais fixé comme but! Voulant rêver d'aventures, j'avais ouvert mon atlas sur le plus grand des continents et, par jeu, tout au hasard, j'avais piqué la page, parmi les fleuves, les villes, les montagnes Je regardai le petit point noir. « Heureux, pensai-je, qui peut se trouver là, avant vécu dans le bruit que font les hommes de son pays! ». Sans cesse, j'y songeais, ne quittant plus la carte des yeux et revenant toujours à ce petit | Beginning of the main content |
|-------------|---|--|-------------------------------|
|-------------|---|--|-------------------------------|

ALTO tags Version: 23.01.2014 27 / 53



| <pre> <structuretag label="Glossary" type="Functional"></structuretag> <structuretag cstructuretag="" indicates="" of="" pro<="" property="" th="" the="" type="Functional"></structuretag></pre> |
|--|
|--|

ALTO tags Version: 23.01.2014 28 / 53



| Index | <structuretag label="Index" type="Functional"></structuretag> | Pia de Tolomei (la): Amour, I, 132, 134; Promenades, II, 41. Pia (porta): Promenades, I, 24, 127, 256; II, 54, 66; III, 260, 324, 329; Idées, 293; Correspondance, IX, 329; X, 149. Piacrna: Journal, V, 293. Pianoin: Pages, 56. Pianoin: Pages, 56. Piano d'Erra: Rome, I, 114. Pian d'Erra: Rome, I, 114. Pian d'Erra: Rome, I, 114. Pianoin (géogri): Journal, IV, 326, 327. Pianono (géogri): Journal, IV, 326, 327. Piansano: Correspondance, X, 206. Pianoid Armonia (II): Rossini, I, 64; II, 237, 245, 329. Piant-Desvials (Mme): Brulard, I, 127. Piant: Correspondance, X, 206. Pianye (la): Napoléon, II, 313; Ecoles, II, 125. Piaze (la): Napoléon, II, 313; Ecoles, II, 170, 171. |
|----------------|---|--|
| Running titles | <pre><structuretag label="RunningTitle" type="Functional"></structuretag></pre> | The National To District Control of the District Contr |

ALTO tags Version: 23.01.2014 29 / 53



PHYSIOLOGIE CATHOLIQUE ET PHILOSOPHIQUE <StructureTag TYPE="Structural" LABEL="Part" PREMIÈRE PARTIE DESCRIPTION="I"/> DE LA VIE DE RELATION Fonctions sensoriales, ou sensations. - Entendement humain. Fonctions intelreductions sensoraises, ou sensations, — Emendement humain, reductions inter-lectuelles et morales. Passions, Psychologie ou idéologie physiologique. Système phrénologique, Méthode phrénométrique. — Fonctions vocales. La voix, La parole. — Fonctions motrices et locomotrices, etc. <StructureTag TYPE="Structural" LABEL="Chapter" DESCRIPTION="I"/> CHAPITRE I Volumes, parts, chapters DES FONCTIONS SENSORIALES OU DES SENS C'EST-A-DIRE DES SENSATIONS. Réflexions préliminaires. Dominateur du globe et des éléments, l'homme doit entretenir avec toute la nature les relations nécessaires à son existence et à son bien-ètre; il est donc important qu'il apprécie ces rapports, afin qu'il recherche et attire les objets qu'il a intérêt de connaître, et qu'il fuie et repousse ceux qui lui nuisent. Les sens sont des sentinelles avancées placées aux limites de l'existence, lesquelles, à l'aide de la correspondance active et presque incessante des nerfs, transmettent à la souveraine assise sur son trône tout ce qui se passe à l'étranger, c'est-à-dire dans le monde extérieur. Or cette souveraine, c'est l'âme; son trône ou son palais, c'est le cerveau, qui est en même temps le centre de son gouvernement (1). (1) Que l'on ne pense pas que nous prétendions assigner à l'âme un siège matériel et anatomique, Voici à ce sujet ce que nous avons dit dans un autre ouvrage

ALTO tags Version: 23.01.2014 30 / 53



<StructureTag TYPE="Structural" LABEL="FullTitle"/> Les XXII Lames Hermétiques DU TAROT DIVINATOIRE <StructureTag TYPE="Structural" LABEL="Title1"/> INTRODUCTION Titles, subtitles L'Astrologie est un traité de la Providence. ST JEROME. Le Tarot, du sanscrit TAR-O, étoile fixe (probablement la polaire qui dans l'astrologie antique symbolisait la tradition immusble) n'est autre chose que la synthèse théosophique et symbolique du dogme pri- ; mitif des Religions en même temps qu'une méthode simplifiée d'Astrologie, retrouvées par le Mage Hermès (1) surnommé trismégiste qui étalt hiérophante des Temples de Thebes; 2000 ans avant J .- C, où li'se a servalt du Tarot comme liistrument augural ; il: étaits: alors grave sur as lames d'or, qui portaient en plus o alon) zues full een home toe en els les leng el e. Lauteur de la Table d'Emergade et du Pymandre resti

ALTO tags Version: 23.01.2014 31 / 53



Numbered titles <StructureTag TYPE="Structural" LABEL="FullTitle"/> L'ART D'ÊTRE GRAND-PÈRE <StructureTag TYPE="Structural" LABEL="Title1" DESCRIPTION="I"/> <StructureTag GUERNESBY TYPE="Structural" LABEL="Title2" L'aren de la piùé ner les posples penalele, Jo le vido mon creso et le l'emple tenjenne. Most je process peur alri l'embre des grands tais accede. DESCRIPTION="I"/> L'BRILÉ SATISPAIT Oh! I'm ve de si prin des fontes minfrables. Los ests, les closes, l'affrant exc. tôtes vénérables. Test do litches grandis par les troubles civils. Schladel allaceol shif to discui me testr.
L'imo s'epulse th, pinkyament contents; Lh d'on ne sels quelle ombre on se sent l'échireur. Je vale dann les ferèles chercher la vague berreur; Des jages qu'on sut du joger, des pritess vile Servatt et confilmt Dieu, prochant peur, processat confer. Pai tent va la laideur que notre boanté montre. La souvago épidenses des lexiscles me promin-Una serte de Joie et d'éponsante obsence, Bene autre blen le mal, dans nours vrai le fore. <StructureTag Et I'v freeve un eath's presque titel au tembres. Et le némit passent pous nes area triomplants, Mala jo ne m'ételus par; on peut rostur finabous Ful tant ve ce qui mort, on qui fuit, ce qui ploto, TYPE="Structural" Con, vison, fallis et valeen, l'el désecusió por joie De siver tamohila sa quelças sombre lien; Lb, seignant, je raddin; et, lors salvas qu'on dien Dans Possime, et, nous le ciel, sous le crypte sacrán, Sun, friesamer an vent profund de l'essercie. LABEL="Title2" Itim what discland date Photonic poor avelo DESCRIPTION="II"/> M'effricult geor reuteur dann bes villes in gioles . Just in made up lead Washings, on durati. La jemenus, l'uneur, le fores, le victoire, Je trouve ben d'avair un tree dans les facilits, Out will de faux, volt blux; qui velt de lois, veit josts. Le sensificate mid qu'ens cralemans augunts Est possible peur elle, et un les les heats lioux Car fe no sele per trop of je otoventirefo. Reputator et grandle, icia da saunda estitleura. Done ja vata su debert, umia span quiller la monde. 11 Paren marks empress viset, dans to their verticals On sur l'escurpament des falciers, s'associé Tranquille et sabilitent l'esmessité du soir, Qu'est-ou que vette terro? Une tempito d'Auset. Duta delle-makes, oh, reschera errente, gousn'abordence Il ne c'irole point de la terre su ross renemie Jameis qu'h des écoule, les pessent pour des perfs ; Dues l'erage des cris, ées détirs, des tressports, No senten-volus dens pre-qu'hymit va benneune d'homanes. On a besein de fair cous des grices épuis, Don amoure, des douleurs, des verses, has de motor; Et que toutes les soils de virité, de paix, Dans les façouis belons de cas prostituées Que nous monument fortune, architées, succès; D'équité, de raisse et de lavelire, magnessions. As food d'une tone, après trat de cizons qui mentant? Devent Job, quit, noullment, dit : Qu'ent-en que ju salu? Et Parest qui, treathine, dit : Qu'ait-ce que je peane? Rome cette momentageme et férene dépense Max Orlean and temperary level name cover, at, lebuschi-Malé précent, je regarde et juge le donin : Je Mises, pour primphées l'imm houselou dinochée, Do yapur, de bésent, de rels, que fail finant; Ha présence de sert tournest seu exhosite.

ALTO tags Version: 23.01.2014 32 / 53



| Footnote references | <pre> <structuretag description="1" type="FootnoteReference"></structuretag> </pre> | Le Tarot, du sanscrit TAR-O, deoile fixe (probablement la polaire qui dans l'astrologie antique symbolisait la tradition immuable) n'est autre chose que la synthèse théosophique et symbolique du dogme primitif des Religions en même temps qu'une méthode simplifiée d'Astrologie, retrouvées par le Mage Herramès (1) surnommé trismégiste qui était hiérophante des Temples de Thèbes, 2000 ans avant JC, où il se il servait du Tarot comme instrument augural; il était da alors gravé sur 22 lames d'or, qui portaient en plus ou de la Table d'Emergand et du Pymandic acot (1). L'auteur de la Table d'Emergand et du Pymandic acot (1). |
|------------------------|---|--|
| Footnotes | <pre><structuretag description="1" label="Footnote" type="Functional"></structuretag></pre> | Le Tarot, du sanscrit TAR-O, étoile fixe (probablement la polaire qui dans l'astrologie antique symbolisait la tradition immuable) n'est autre chose que la synthèse théosophique et symbolique du dogme primitif des Religions en même temps qu'une méthode simplifiée d'Astrologie, retrouvées par le Mage Hermés (1) surnommé trismégiste qui était hiérophante des Temples de Thèbes, 2000 ans avant JC, où il se il servait du Tarot comme instrument augural; il étaits alors gravé sur 22 lames d'or, qui portaient en plus de La Table d'Emerande et du Pymandre. |
| References to footnote | <pre><structuretag description="1" type="ReferenceToFootnote"></structuretag></pre> | |

ALTO tags Version: 23.01.2014 33 / 53



| Marginalias | <pre><structuretag label="Marginalia" type="Functional"></structuretag></pre> | DE LUMINE. riores Planetæ pro eâdem ratione non magis à Sole illuminati fint, quàm superiores, neque etiam quàm Cometæ, qui citra omnem controversiam ab illo remotiores sunt. Simile quid in vero etiam Mundo obtinere experientia evincit, & tamen non credo, rationem ejus reddi posse, supposito lumen aliud quid esse in objectis, quàm ejusmodi actionem aut dispositionem, qualem exposui. Actionem dico, aut dispositionem; observato enim illo quod jam dedi demonstratum, quòd * vacuo existente spatio, in quo Solest, cœli ejusdem partes versus oculos spectantium tenderent eâdem ratione, quàm cùm ab illius materià impelluntur, & æquali ferme cum impetu; percipi quidem poterit, haud opus esse, ut aliquam in se actionem habeat, neque etiam ut quid aliud sit, quàm merum spatium, ut qualem ipsum videmus ap- |
|-----------------|--|---|
| Figure captions | <pre><structuretag label="FigureCaption" type="Functional"></structuretag></pre> | State platition risk after the state of the |

ALTO tags Version: 23.01.2014 34 / 53



| Figure references | <pre> <structuretag description="9" label="FigureReference" type="Functional"></structuretag> </pre> | Strongulation in deal in biological growth and share demand some forms (true in the state of the |
|----------------------|--|--|
| References to figure | <pre><structuretag description="9" label="ReferenceToFigure" type="Reference"></structuretag></pre> | |

ALTO tags Version: 23.01.2014 35 / 53



| Table captions | <pre><structuretag label="TableCaption" type="Functional"></structuretag></pre> | 2. — Régime fiscal applicable aux fonds d'Etats étrangers et aux sociétés átrangères non abonnées. Nous avons indiqué les raisons pour lesquelles le régime de l'abonnement 1 pas paru applicable aux titres de sociétés étrangers. Par ailleurs, l'abonnement laisse en debrar de son domaine d'application les titres de sociétés et soilectivités étrangères dont l'introduction en France n'est pas le fait personné destités sociétés et collectivités etrangères dont l'introduction en France n'est pas le fait personné destités sociétes et collectivités, il convient dono d'étudier quel est le régime liscal auquel sont assujetits les titres de cest deux catégories circulant en Prance. 4. — Fonds d'Etats étrangers ne sont pas assujetits à la formatité préalable de l'abonnement pour pouvoir étre négociés en France. Mais ils sont assujetits à l'impôt du timbre. Les fonds d'Etats étrangers ne sont pas assujetits à la formatité préalable de l'abonnement pour pouvoir étre négociés en France. Mais ils sont assujetits à l'impôt du timbre et à l'impôt sur le revenu. 5. — Impôt du timbre. En vertu des dispositions législatives rappelées par l'article 19 du décret de codification du 29 décembre 1926, les titres de rentes, caprunts et autres effets publics des gouvernements étrangères sont sounis à un droit de tumbre, l'acchet en principe sur leur la cours de la cours au complant, relacide en principe sur leur la la cours de l'acchet en principe sur leur la la cours de la cours moyen pendant l'année prédécate est tembé an-dessous de la moité du pair, la perception s'effectue sur la valeur négociable déterminée par ce cours. TAUX DU INDUT. — Les taux successifs du droit de timbre sont indiquée par le fableau ci-dessous : Loi du 12 mai 1863 |
|---------------------|--|---|
| Table references | <pre><structuretag description="1" label="TableReference" type="Functional"></structuretag></pre> | |
| References to table | <pre><structuretag description="1" label="ReferenceToTable" type="Reference"></structuretag></pre> | |

ALTO tags Version: 23.01.2014 36 / 53



| Page numbers | <pre><structuretag description="937" label="PageNumber" type="Functional"></structuretag></pre> | The potation of year is Adeligate, over a cut of the forest and th |
|-------------------------------|--|--|
| Reference to page | <pre><structuretag description="8" label="ReferenceToPage" type="Reference"></structuretag></pre> | Nous sortions de chez le boulanger, quana nous vimes devant nous, une femme qui nous regardait fixement. (Page 8), |
| Bullets lists, numbered lists | <pre><structuretag label="UL" type="Functional"></structuretag> <structuretag label="OL" type="Functional"></structuretag></pre> | |
| | | |

ALTO tags Version: 23.01.2014 37 / 53





ALTO tags Version: 23.01.2014 38 / 53



Section headings, rubrics, column titles, etc.

<StructureTag
TYPE="Structural"
LABEL="SectionHeading"
DESCRIPTION="SOCIETY"/>



http://chroniclingamerica .loc.gov/lccn/sn86069133/ 1910-10-31/ed-1/seq-3/

ALTO tags Version: 23.01.2014 39 / 53





ALTO tags Version: 23.01.2014 40 / 53





ALTO tags Version: 23.01.2014 41 / 53



| Location, spatial information, place name | <pre><structuretag description="CLAY CITY, Ky" label="Place" type="Functional"></structuretag></pre> | Democratic Nominee For Congress and R. R. Friend. of Irvine, Speak to Pewell County Voters. CLAY CITY, Ky., Oct. 31.—Hon. A. Floyd Byrd, Democratic nominee for Congress from the Tenth Congressional District and Robert R. Friend, of Irvine, Ky., Saturday admired and the sign but said he did not the feet the child's apper. Mr. Norris, cattle last week and made among them 350 her Bigstaff, average weight 1.40 cents, and 60 of Hen age weight 1.450 por These cattle were but timore market, as we Morris purchased. A purchased of Ben F. | Coverage | note | about | a |
|---|--|--|----------|------|-------|---|
| Date, dateline | <structuretag description="Oct. 31." label="Date" type="Functional"></structuretag> | Democratic Nominee For Congress and R. R. Friend. of Irvine, Speak to Pcwell County Voters. CLAY CITY, Ky., Oct. 31.—Hon. A. Ployd Byrd, Democratic nominee for Congress from the Tenth Congressional District and Robert R. Friend, of Irvine, Ky., Saturday ad- | Coverage | note | about | a |

ALTO tags Version: 23.01.2014 42 / 53



http://gallica.bnf.fr/ark :/12148/bpt6k297360g/f1.i mage Authors could also be LES ECLIPSES
DE LA POLITESSE
United by the second of the s tagged with a <RoleTag> <StructureTag Article authors, (see section 5.). TYPE="Functional" copyright note, LABEL="ArticleAuthor" etc. DESCRIPTION="Léon Lafage"/> Evils malleureux - « les déliteits LA COUR coef call-bearent; — dies at titidesse de voer s'encamine tan propte english menourle, vous à la conocie de vivré de la conocie de conocie de la conocie del la conocie de la conocie del la conocie de la conocie del la c morce asserption of fragelli die la real Proport, ier manistere et it 100 to it is benne.

A Lenders, in possible ratio
issue, de det entre et al.

B article print in de perimeter que
in politicus sort ing art de jossie, que glier
perimeter mondad qui, dans puisances <StructureTag http://chroniclingamerica Roosevelt For Langley .loc.gov/lccn/sn86069133/ TYPE="Functional" 1910-10-31/ed-1/seg-6/ LABEL="Illustration" DESCRIPTION="John W. Langley"> John W. Langly <XmlData> If the intellectual value <-- bibliographic of an illustration is metadatas of the photography high, it can be captured with a structural tag. Illustration fall . Unfortunately it is </XmlData> to the well being of all <StructureTag/> I am your friend . Theodore Rosselt NAME AND ADDRESS OF TAXABLE PARTY ADDRESS OF TAXABLE PA <StructureTag TYPE="Functional" FARMS FOR SALE CONGRESSMAN LANGLEY LABEL="Caption"/> AS SEEN BY THE PRESS

ALTO tags Version: 23.01.2014 43 / 53



ALTO tags Version: 23.01.2014 44 / 53



4. Named Entities tagging

Named-entity recognition (NER) is a subtask of information extraction that seeks to locate and classify atomic textual elements into predefined categories (names of persons, organizations, locations, expressions of times, quantities, etc).

Representing the result of a NER processing in ALTO has several pros:

- Simplicity: no need to use another format for NE storage, no need to make links between this format and the ALTO block coordinates; no need to use a METS-ALTO approach.
- Autonomy: contents and NE are stored in the same file.



ALTO tag element: <NamedEntityTag>

Tag attributes:

- ID: Tag id
- LABEL: Named entity category (Person, Organization, Location, etc.)
- TYPE (optional): Suppertype or category of the named entity type
- DESCRIPTION (optional): Named entity textual content
- XmlData (optional): Any XML encoded metadata helping to describe the tag or the vocabulary used to describe the tag
- URI (optional): Any authority URI relevant for the tag element: NE authority repositories, gazetteers, etc.

Good pratices, limitations

- Multiple authorities can be assigned to a named entity (see example bellow).
- Nested tags can be described (see example bellow).
- Occurrences of the same NE can't clearly be described if they are near one of the other. The first two strings
 reference the "JS Bach" NE and the last three reference the same "JS Bach" NE, but the mechanism can't say
 where is the frontier.



ALTO tags Version: 23.01.2014 45 / 53



| Text tags, to be applied on <composedblock>, <textblock>, <textline> or <string> elements</string></textline></textblock></composedblock> | | | |
|---|---|---|---|
| Use Cases | ALTO tag | Examples | Remarks |
| Locations | Simple tagging <namedentitytag description="Lexington" label="Location"></namedentitytag> URI Authority <namedentitytag description="Lexington" label="Location" uri="http://www.geonames.org/4941935"></namedentitytag> Extra attributes <namedentitytag description="Kentucky" label="Location"> <namedentitytag description="Kentucky" label="Location"> <xmldata> <nec>1.0</nec> <!-- NE confidence value--> <variants>Kenekuke Kentaki Kentákii Hahoodzo Kèntòki Kentórki Kentuki Kentukia Kentukio Kentukis Kentuki Kentukia Kentukki Shitati Khén-thap-kî よここと 「といて、ことでは、「Kentaku Kentaku Kentaku Kentyki Kentaku Kentaku Kentaku Kentyki Kentaku Kuntaaku IIIrat Kentyki Uhunnuhh 306の330 帝司之命 帝之命 命之命 命之命 (Multiple Authorities> <</variants></xmldata></namedentitytag></namedentitytag> | Louisville, which owns eight of the twelve similar houses doing business in that city and which is known as the largest warehouse company in | http://chroniclingamerica.loc.gov/lccn/sn86069133/1910-10-31/ed-1/seq-1/http://chroniclingamerica.loc.gov/lccn/sn86069133/1910-10-31/ed-1/seq-1/ocr.xml |

ALTO tags Version: 23.01.2014 46 / 53



```
<NamedEntityTag ID="15a"</pre>
                 LABEL="Location"
                 DESCRIPTION="Louisville"
                 URI="http://www.geonames.org/4299276"
                  <NamedEntityTag ID="15b"
                 LABEL="Location"
                 DESCRIPTION="Louisville"
                 URI="mygeonames:louisville" />
                 <String CONTENT="Louisville" WC="1.0"</pre>
                 TAGREFS="15a 15b">
                 <!--Simple tagging -->
                                                                      her injuries was made, Dr. Reynolds
                 <NamedEntityTag LABEL="Person"</pre>
                                                                      an eve specialist, after the examina-
                 DESCRIPTION="Dr. Reynolds"/>
                                                                      tion, announced the eye permanently
                                                                      injured and the sight forever gone
                 <!-- URI Authority -->
                                                                      but said he did not think it would at-
                 <NamedEntityTag LABEL="Person"</pre>
                                                                      feet the child's appearance.
                 DESCRIPTION="Mr. Norris"
                 URI="http://catalogue.bnf.fr/servlet/autori
                                                                       Mr. Norris, entile buyer, was here
                                                                      hist week and made many purchases.
                 te?ID=11916925"/>
                                                                      among them 350 head of James M.
                 <!-- embedded authority description -->
                                                                      Bigstaff average weight 1,500 pounds
                                                                      at 6 cents; 150 of J. T. Highland,
                 <NamedEntityTag LABEL="Person"</pre>
                                                                      overage weight 1 400 pound, at 6
                 DESCRIPTION="James M Bigstaff ">
Persons
                                                                      cents, and 60 of Henry C. Hall, aver-
                  <XmlData>
                                                                      age weight 1,450 pounds at 6 cents.
                  <MADS
                                                                      These cattle were bought for the Bal-
                  xmlns="http://www.loc.gov/mads/">
                                                                      Morris purchased. Mr. W. B. Greene
                   <authority>
                                                type="personal"
                    <name
                                                                      purchased of Ben F. Mark 40 cattle.
                 authority="naf">
                                                                      average 1,400 pounds, at the prevail-
                      <namePart>Bigstaff,
                                                                     ing prices. There are yet about 1,000
                                                             James
                                                                     cattle fo ribe export market unsold
                 M</namePart>
                                              type="date">1835-
                                                                     in this county.
                      <namePart
                 1882</namePart>
                    </name>
                    </authority>
```

ALTO tags Version: 23.01.2014 47 / 53



| | <pre> </pre> | | |
|---------------|--|---|---|
| Organizations | -Simple tagging <namedentitytag description="Central Kentucky Tobacco Warehouse Company" id="020" label="Organization"></namedentitytag> | LEXINGTON, Ky., Oct. 31.—The biggest business transaction engineer ed 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 derl involves as exchange of from \$95,000 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. | Nested tags ar implemented with multipl tag refs. |

ALTO tags Version: 23.01.2014 48 / 53



5. Role tagging

Role tagging allows to describe the people involved in the content creation: author, editor, etc. This can be done with an open vocabulary or a dedicated vocabulary, like the MARC Code List for Relators.



ALTO tag element: <RoleTag>

Tag attributes:

- ID: Tag id
- LABEL: tag type or category
- TYPE (optional): Suppertype or category of the type
- DESCRIPTION (optional): Description of the tag, description of the tag content, or the content itself
- XmlData (optional): Any XML encoded metadata helping to describe the tag or the vocabulary used to describe the tag.

Version: 23.01.2014

URI (optional): Any authority URI relevant for the tag element.

Good pratices, limitations

• A specific name tagged with a Role tag can also be identify with a NE tag.



| Text tags, to be applied on <composedblock>, <textblock>, <textline> or <string> elements</string></textline></textblock></composedblock> | | |
|---|---|--|
| Use Cases | ALTO tag | Examples |
| Title pages | <roletag description="book's author" label="Author"></roletag> <roletag description="drawings creator" label="Illustrator"></roletag> | R. FALCONNIER DISTANCE HERMÉTIQUES DISTANCE TAROT Divinatoire Exaclenient reconstituées d'après les textes sacrés et selon la tradition des Mages du l'ancienne Égyple MESSINS DE MCE OTTO WEGENER |
| | <pre><roletag label="Publisher"></roletag> < Description of the vocabulary used -></pre> | PARIS LIBRAIRIE DE L'ART INDÉPENDANT 12, RUE DE LA CHAUSSÉE D'ANTIN, 11 1896 Teus droits rétorvés |
| | <roletag description="drawings creator" label="ill"> <xmldata> <scheme>marc:relators</scheme> </xmldata> </roletag> | |

ALTO tagsALTO_tags_v1_0.docx

Version: 23.01.2014

50 / 53



II LE TAROT

Tont le monde, grâce à vous, pourra désormais comprendre, admirer et au besoin consulter ces arcanes mystérieux de la Haule-Science des Mages, car vos Tarots sont d'une lucidité psychique incomparable et leurs maximes si consolantes à tous dgards, peuvent également servir à nous diriger dans telle ou telle voie favorable à notre réussite dans la bataille que nous livrons chaque jour à l'inconnu.

Tous mes compliments de sincère admiration pour le savoir et la persévérance que vous avez apportés dans vos recherches et croyez-moi bien, cher Monsieur, cordialement à vous.

A. DR THÉBES.

Version: 23.01.2014



6. Other tags

Any other kind of tags can be described.



ALTO tag element: <OtherTag>

Tag attributes:

- ID: Tag id
- LABEL: Tag type or category
- TYPE (optional): Suppertype or category of the type
- DESCRIPTION (optional): Description of the tag, description of the tag content, or the content itself
- XmlData (optional): Any XML encoded metadata helping to describe the tag or the vocabulary used to describe the tag

Version: 23.01.2014

• URI (optional): Any authority URI relevant for the tag element.



7. Reference documents

EPUB 3 Structural Semantics Vocabulary: http://www.idpf.org/epub/vocab/structure/

PAGE format: http://schema.primaresearch.org/PAGE

MARC Code List for Relators: http://id.loc.gov/vocabulary/relators.html

Europeana Newspapers:

- METS ALTO Profile (ENMAP), version 1.0, January 8th 2014
- Options for the representation of recognized named entities, version 0.1, December 12th, 2012

Version: 23.01.2014