docWORKS/METAe

Automated Conversion of Printed Material into METS/ALTO Objects
What is docWORKS/METAe?

- Production tool for conversion of printed documents into fully tagged digital objects
- The METAe edition of docWORKS is the result of the EU-funded project METAe, September 2000 to August 2003
- Clients: Harvard, LC, GDZ, Royal Library Denmark
The project group

1. Leopold-Franzens-Universität Innsbruck (Co-ordinator), Austria
2. Universität Linz, Institut für Angewandte Informatik, University of Linz, Austria
3. Mitcom Neue Medien GmbH (ABBYY Europe), Germany
4. CCS Compact Computer Systeme, Germany
5. Universidad de Alicante, Spain
6. Friedrich-Ebert-Stiftung, Germany
7. Cornell University Library. Department of Preservation and Conservation, USA
8. BNF, Bibliothèque Nationale de France
9. The National Library of Norway, Rana division, Norway
10. Biblioteca Statale A. Baldini, Italy
11. Dipartimento di Sistemi e Informatica, University of Florence, Italy
12. Karl-Franzens-Universität Graz, Universitätsbibliothek, Austria
13. Scuola Normale Superiore, Centro di Ricerche Informatiche per i Beni Culturali, Italy
14. Higher Education Digitisation Service HEDS, UK
Challenges

- Digitization and retro-conversion of printed or textual material is getting more and more important:
  - Keep knowledge and cultural heritage alive
  - Preserve the origin
  - Enable quick and enhanced access by high structured documents
  - Open up new dimensions of research
  - Provide standardized output formats
Goals

- Automate the conversion process
- Make digitization more effective and safer
- Go for mass digitization
- Increase the added value of digitized collections
- Provide a standardized output format
System Overview

Input

docWORKS engine

Output

Image Pre-Processing
Layout Analysis
Character Recognition
Structural Analysis

Rules DB

METS/ALTO
METS/TEI
PDF
TIFF, JP2
## Matching of Image Files and Page Numbers

<table>
<thead>
<tr>
<th>Image-file</th>
<th>Pagination</th>
<th>Page-Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>000001.tif</td>
<td>Not counted</td>
<td>Np</td>
</tr>
<tr>
<td>000002.tif</td>
<td>Not counted</td>
<td>Np</td>
</tr>
<tr>
<td>000003.tif</td>
<td>Counted</td>
<td>I</td>
</tr>
<tr>
<td>000004.tif</td>
<td>Counted</td>
<td>II</td>
</tr>
<tr>
<td>000005.tif</td>
<td>Counted</td>
<td>III</td>
</tr>
<tr>
<td>000006.tif</td>
<td>Counted</td>
<td>IV</td>
</tr>
<tr>
<td>000007.tif</td>
<td>Counted</td>
<td>V</td>
</tr>
<tr>
<td>000008.tif</td>
<td>Counted</td>
<td>VI</td>
</tr>
<tr>
<td>000009.tif</td>
<td>Counted</td>
<td>1</td>
</tr>
<tr>
<td>000010.tif</td>
<td>Counted, not paginated</td>
<td>(2)</td>
</tr>
<tr>
<td>000011.tif</td>
<td>Counted</td>
<td>3</td>
</tr>
<tr>
<td>000012.tif</td>
<td>Counted</td>
<td>4</td>
</tr>
<tr>
<td>placeholder</td>
<td>Missing page</td>
<td>5</td>
</tr>
<tr>
<td>placeholder</td>
<td>Missing page</td>
<td>6</td>
</tr>
<tr>
<td>000013.tif</td>
<td>Counted</td>
<td>7</td>
</tr>
<tr>
<td>000014.tif</td>
<td>Counted</td>
<td>8</td>
</tr>
</tbody>
</table>
THE
AMERICAN MISSIONARY.
Vo., XXXII  JANUARY, 1878  No. 1
American Missionary Association
1877 - 1888

The work of the American Missionary Association has in recent years been
increased. We have been able to extend our work into new fields, and to open
new schools and establish new missions. The Association has been
able to raise larger sums of money than ever before, and to make a
greater impact on the missions.

In the past year, the Association has made a number of important
achievements. One of the most notable has been the establishment of a
new school in the South. This school has been attended by children of all
ages, and has proved to be a great success.

Another important achievement has been the publication of a new
missionary guidebook. This guidebook has been well received, and has
proved to be an invaluable resource for missionaries.

In summary, the American Missionary Association has had a
successful year, and we look forward to continued growth and
success in the future.
ALONZO DE OJEDA,*

HIS FIRST VOYAGE,

IN WHICH HE WAS ACCOMPANIED BY

AMERIGO VESPUCCI.†

CHAPTER I

Some Account of Ojeda—of Juan de la Cosa—of Amerigo Vespucci—Preparations for the Voyage.—(1495.)

Those who have read the History of Columbus will, doubtless, remember the character and exploits of Alonzo de Ojeda; as some of the readers of the following pages, however, may not have perused that work, and as it is proposed at present to trace the subsequent fortunes of this youthful adventurer, a brief sketch of him may not be deemed superfluous.

Alonzo de Ojeda was a native of Cuenca, in New Castile, and of a respectable family. He was brought up as a page or esquire; to the service of Don Luis de Corda, Duke of Medina Celis, one of the most powerful nobles of Spain; the same who

* Ojeda is pronounced in Spanish Ojeda, with a strong aspiration of the k.
† Vespucci, pronounced Vespycci.

for some time patronized Columbus during his application to the Spanish court.*

* Various Imprints, por F. Fierro y Orillana, p. 41. Las Casas, Hist. Ind. L. i. c. 82.
Structural Analysis
Structural Analysis
Digitization of Books and Journals (METAe)
Digitization of Books and Journals (METAe)
Digitization of Scientific Documents
Manual Editing of Descriptive Metadata / Volume
Manual Editing of Descriptive Metadata / Illustration
Scalable Client / Server Architecture

- Server n
- Quality Control Client m
- Quality Control Client 1
- File Server

- Auto-Import
- Image Preprocessing
- Layout Analysis
- OCR
- Structural Analysis
- Export
In Practice: docWORKS in Digitization Projects

Input

Image Files

Output

METS/ALTO XML
(Digital Objects)

PDF
DAISY

Repository

Presentation System

Content Conversion
(OCR & Structure)
KBDK Markup Policy

Les verbes impliquant négation

Les verbes impliquant négation ont une caractéristique que le complément est fait quand la principale est positive et vice-versa. Ils obéissent aux postulats de sens (114) et (115):

\[
\begin{align*}
\text{114: } & x (G) \rightarrow \text{ être } \\
\text{115: } & x (G) \rightarrow \text{ être }
\end{align*}
\]

Ainsi, lorsqu’un locuteur énonce (116 a), il ne peut pas nier (116 b) sans se contredire. De même, quand il énonce (117 a), il ne peut pas nier (117 b) sans contradiction:

\[
\begin{align*}
\text{116: } & \text{ Pierre a cédé de parler.} \\
\text{117: } & \text{ Pierre n’a pas cédé de parler.}
\end{align*}
\]

Comme les verbes impliquant positivité, les verbes impliquant négation peuvent d’abord se diviser en deux groupes principaux exemplifiés par cause de et d’exclure ce de.

Le premier groupe peut encore se diviser en trois sous-groupes:

1. Le premier sous-groupe comprend:
   - cause de
   - cause de
   - cause de

2. Le deuxième sous-groupe comprend:
   - débourse à

3. Le troisième sous-groupe comprend:
   - s’interrompe du
   - se priver de
   - se priver de

20: Ce verbe se rencontre seulement dans des constructions négatives.
Presentation of Converted Material

7 Breakfast Corn Bread,

One tea-cup of rice boiled nice and soft, to one and a half tea-cupful of corn meal mixed together, then stir the whole until light; one teaspoonful of salt, one tablespoonful of lard or butter, three eggs, half tea-cup of sweet milk. The rice must be mixed into the meal while hot; can be baked either in muffin cups or a pan.

8 Corn Egg Bread.

Two eggs, one pint of meal, half pint of sour milk, one teaspoonful of soda,—beat eggs very light,—one tablespoonful of melted lard or butter, mix all together, well stirred or beaten. Bake in an ordinary pan.

9 Plantation Corn Bread or Hoe Cake.

Half tablespoonful of lard to a pint of meal, one tea-cup of boiling water; stir well and bake on a hot griddle. Sift in meal one teaspoonful of soda.

10 Light Bread.

Half yeast cake to two quarts of flour, teaspoonful of salt, one dessertspoonful of butter or lard. Dissolve yeast in warm water; make up over night at 10 o'clock; make dough soft and spongy, and set to rise in a warm place. Next morning work the dough over until it be-
Presentation of Converted Material

Illustrations
- Plate 1. Texas State Capitol Building, Austin. Built with red granite.
- Plate 2. Gray granite quarry operated by Gooch & Company, 5 mi.
- Plate 3. The Llano quarrymen experience considerable difficulty.
- Plate 4. Taking Mocks of red granite out of Granite Mountain.
- Plate 5. Darragh Brothers' quarry at Granite Mountain, Burnet Co.
METS / ALTO XML

ALTO – Analyzed Layout and Text Object
METS

- Header
- DC/MODS descriptive metadata
- NISO 39.087 (mix), technical metadata
- Structural Map: Physical Structure
- Structural Map: Logical Structure
ALTO

- Styles
  - Paragraph (alignment, linespacing, etc.)
  - Font (name, size, bold, italic, etc.)

- Layout
  - Printspace
  - TopMargin
  - InnerMargin
  - OuterMargin
  - BottomMargin

- Objects in 5 areas above:
  - Text block
  - Text lines
  - Strings [coordinates, string (as printed), substitution (hyphenation)]
  - Spaces
  - Composed block
  - Picture
  - Table
  - Formula
METS / physical structure

<table>
<thead>
<tr>
<th>ORDER</th>
<th>LABEL</th>
<th>ORDERLABEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>I</td>
<td>I</td>
</tr>
<tr>
<td>2</td>
<td>II</td>
<td>II</td>
</tr>
<tr>
<td>3</td>
<td>III</td>
<td>III</td>
</tr>
<tr>
<td>4</td>
<td>IV</td>
<td>IV</td>
</tr>
<tr>
<td>5</td>
<td>V</td>
<td>V</td>
</tr>
<tr>
<td>6</td>
<td>VI</td>
<td>VI</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
METS / physical structure
Those who have read the History of Columbus will, doubtless, remember the character and exploits ...
AMERIGO VESPUCCI.†

CHAPTER I.

Some Account of Ojeda—of Juan de la Cosa—of Amerigo Vespucci—Preparations for the Voyage.—(1499.)

Those who have read the History of Columbus will, doubtless, remember the character and exploits of Alonzo de Ojeda; as some of the readers of the following pages, however, may not have perused that work, and as it is proposed at present to trace the subsequent fortunes of this youthful adventurer, a brief sketch of him may not be deemed superfluous.

Alonzo de Ojeda was a native of Cuenca. in
Summary: Milestones

- **METAe Project Start** - September 2000
- **Product Launch** - March 2003
- **Royal Danish Library** - July 2003
- **University Library of Lower Saxony (Göttingen)** - March 2004
- **Stanford University Library Evaluates docWORKS** - July 2004
- **Harvard University Library** - September 2004
- **LC contracts CCS for research project (NDNP)** - April 2005
- **LC evaluates docWORKS** - May 2005
Thank you!

Claus Gravenhorst
Daniel Lanz

claus.gravenhorst@ccs-gmbh.de
daniel.lanz@ccs-gmbh.de

Content Conversion Specialists
www.ccs-gmbh.de

http://meta-e.uibk.ac.at/