

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

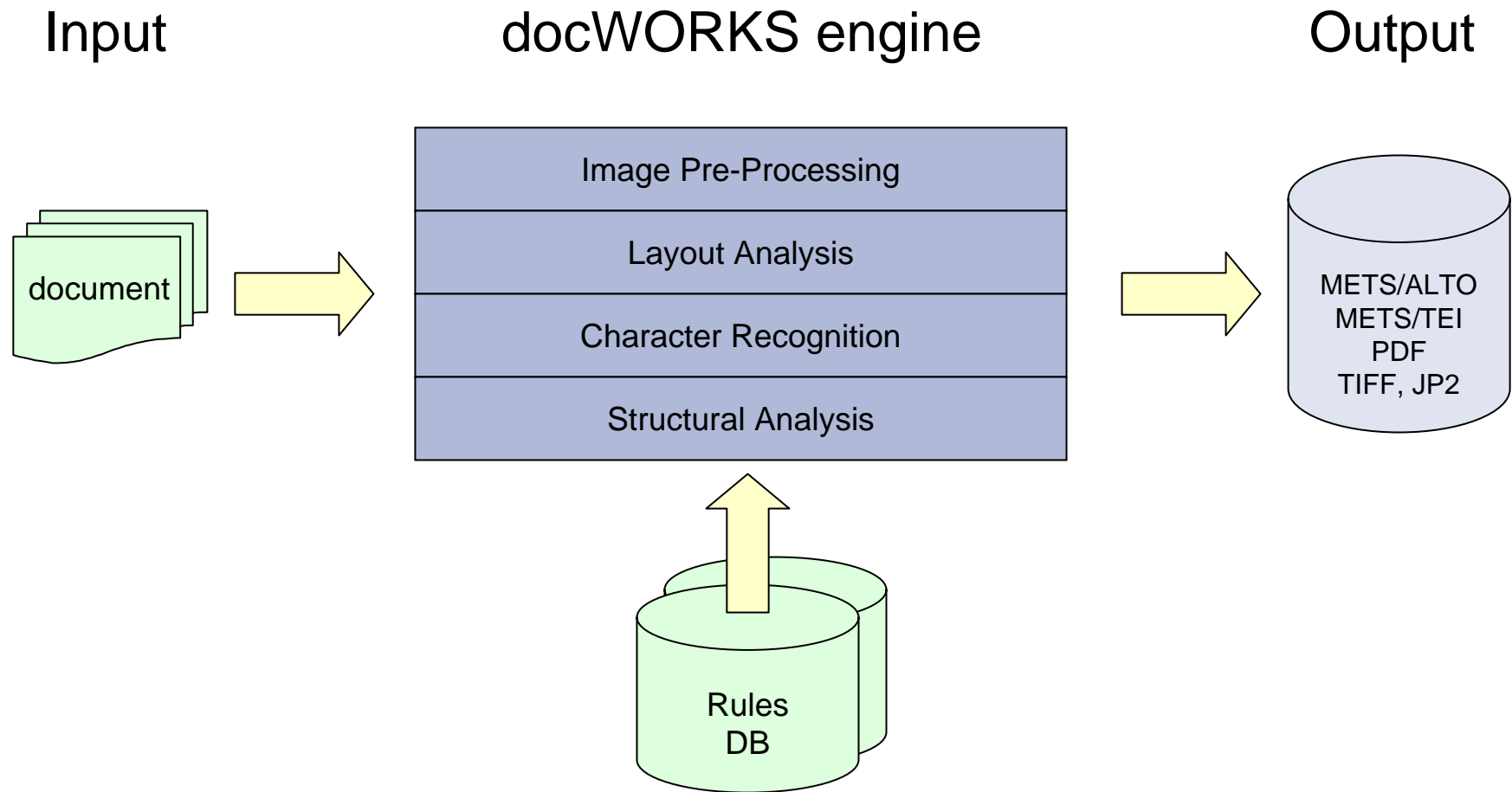
Challenges

- ☞ 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



Matching of Image Files and Page Numbers

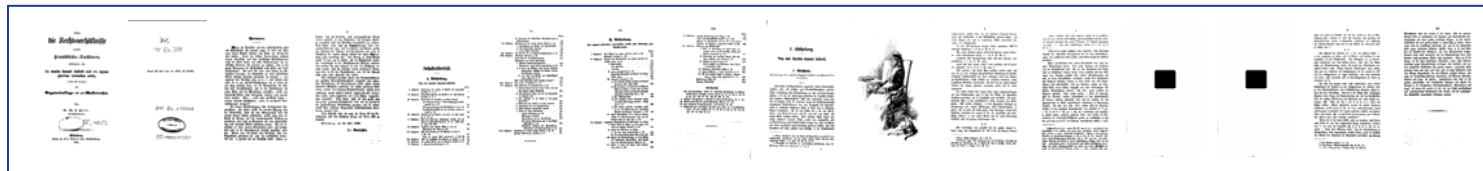
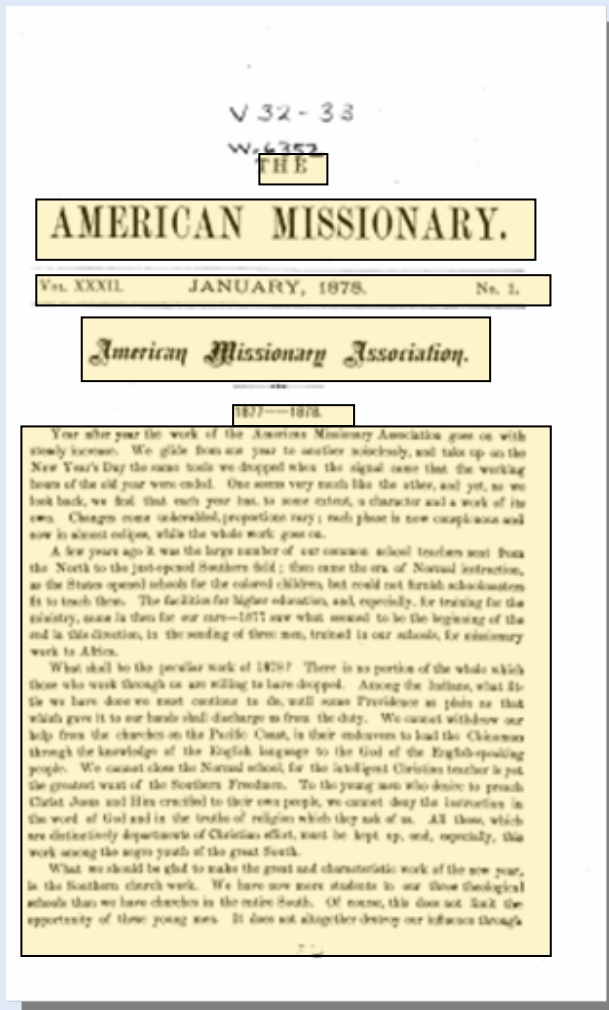


Image-file	Pagination	Page-Number
000001.tif	Not counted	Np
000002.tif	Not counted	Np
000003.tif	Counted	I
000004.tif	Counted	II
000005.tif	Counted	III
000006.tif	Counted	IV
000007.tif	Counted	V

000008.tif	Counted	VI
000009.tif	Counted	1
000010.tif	Counted, not paginated	(2)
000011.tif	Counted	3
000012.tif	Counted	4
placeholder	Missing page	5
placeholder	Missing page	6
000013.tif	Counted	7
000014.tif	Counted	8

Traditional OCR - Output



THE
AMERICAN MISSIONARY.

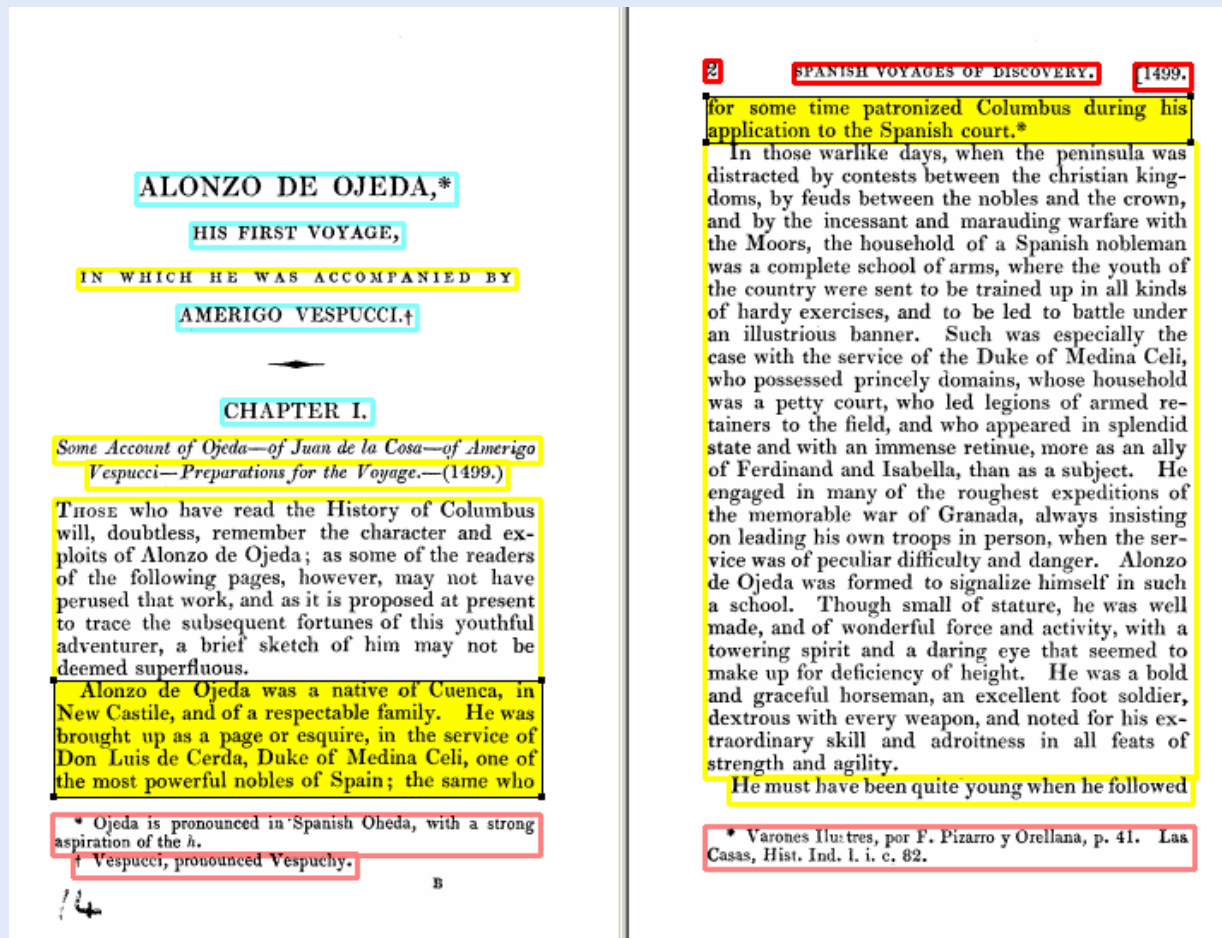
Vo.. XXXII JANUARY, 1878 No. 1

American Missionary Association

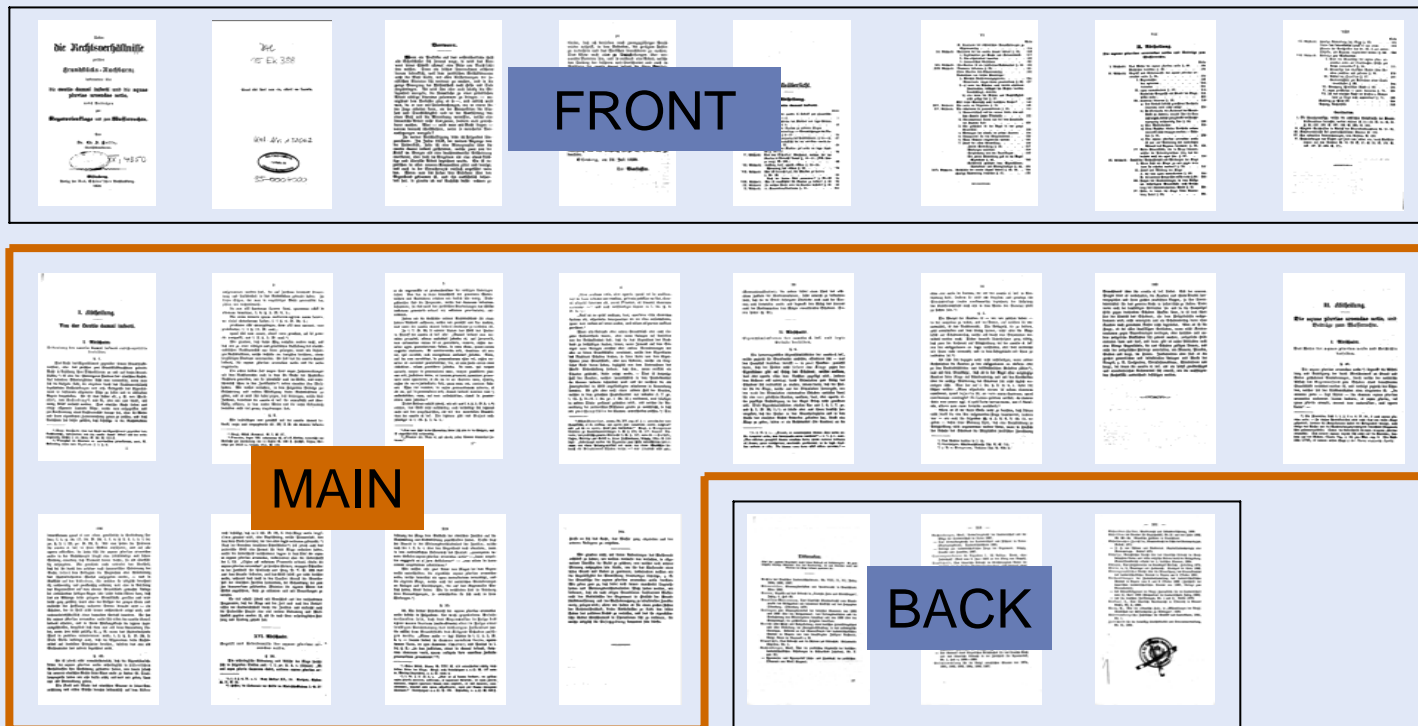
1877 - 1888

[illegible]

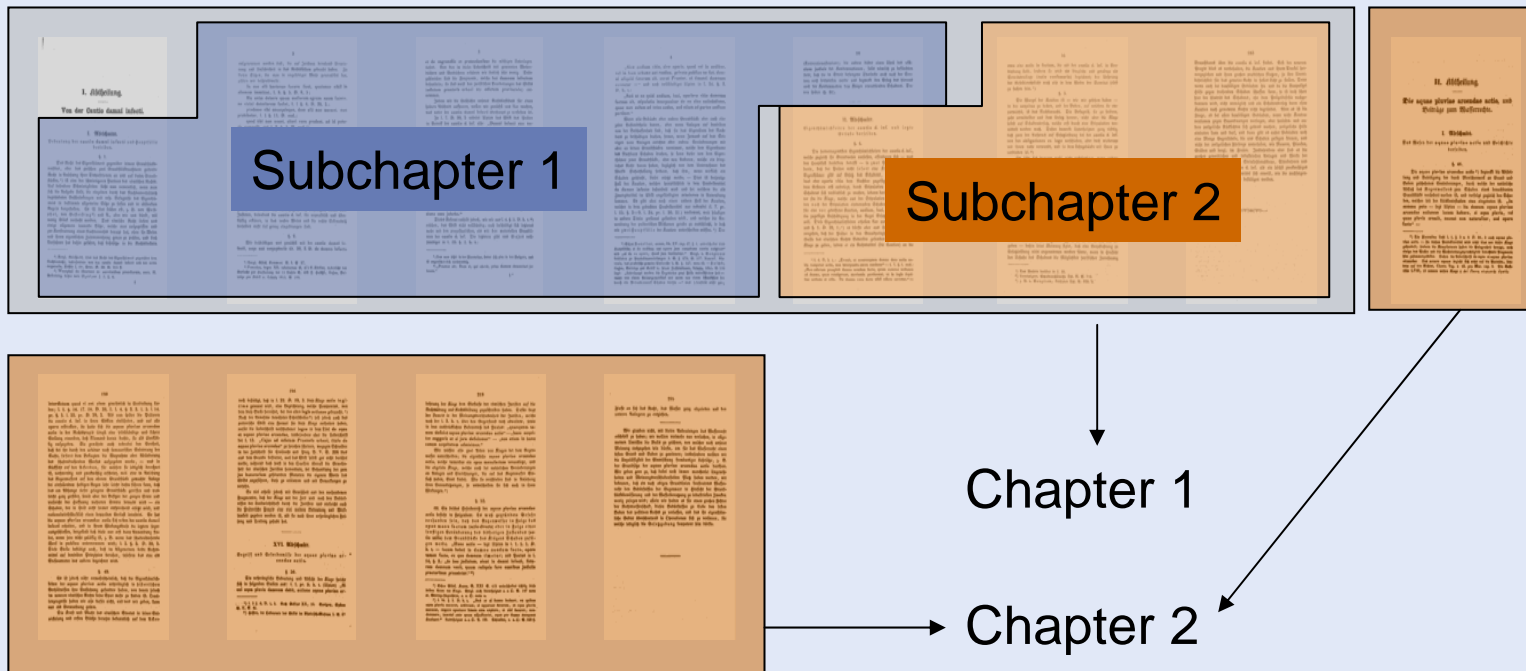
Physical and Logical Structure



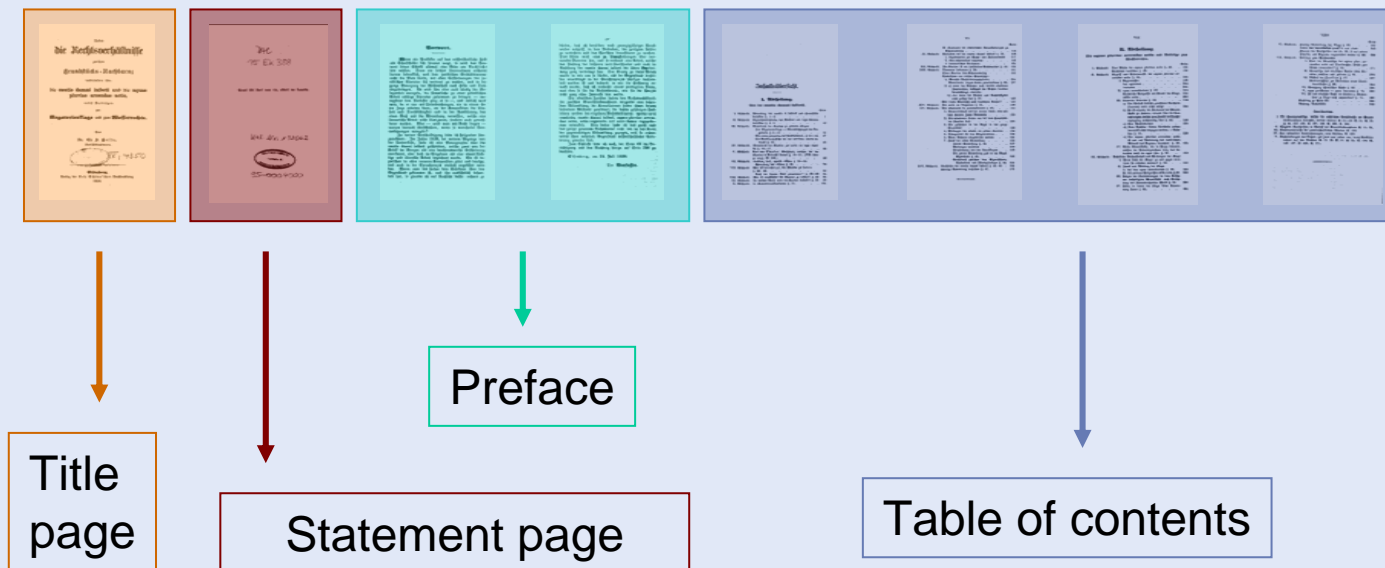
Structural Analysis



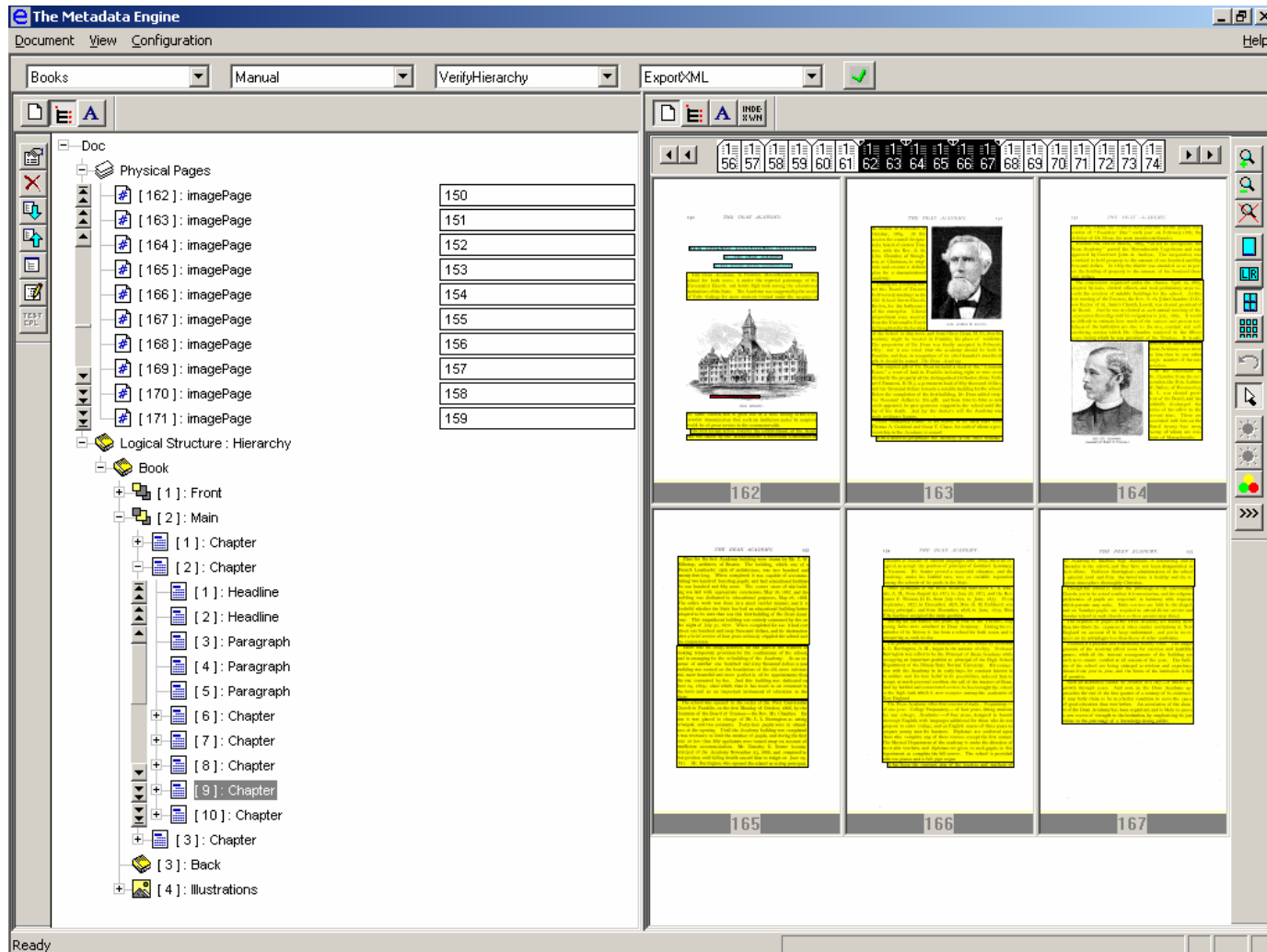
Structural Analysis



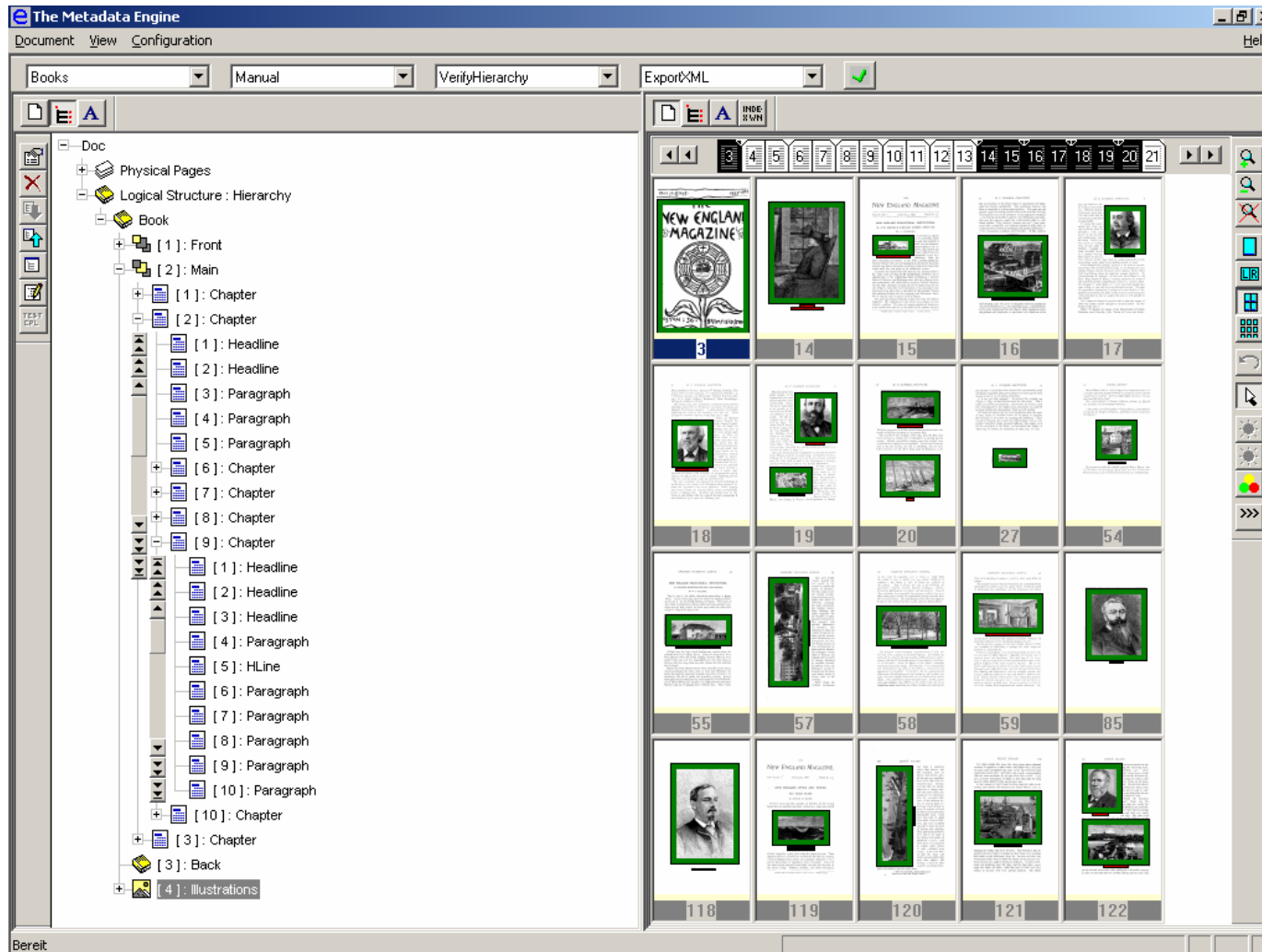
Structural Analysis



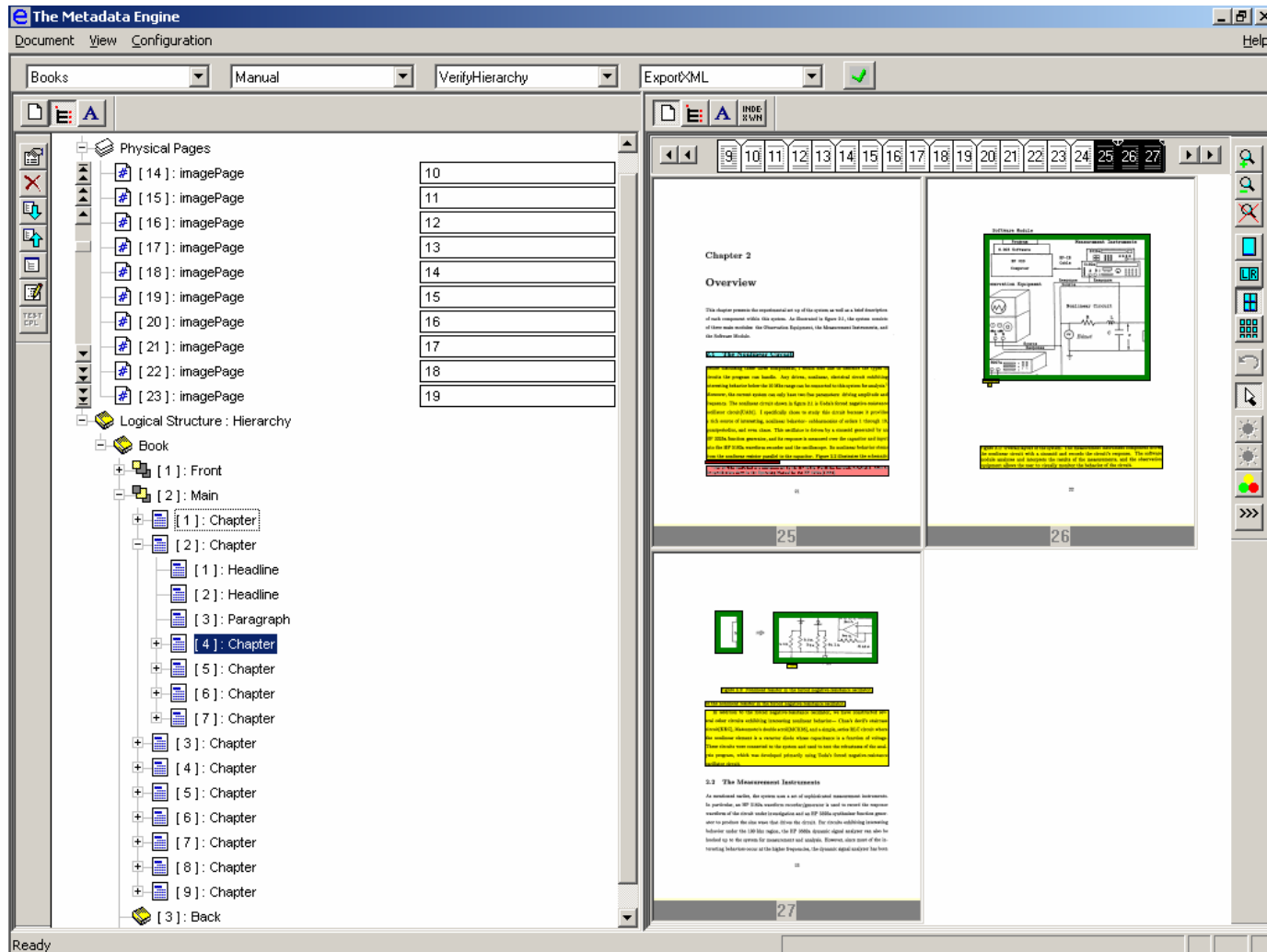
Digitization of Books and Journals (METAe)



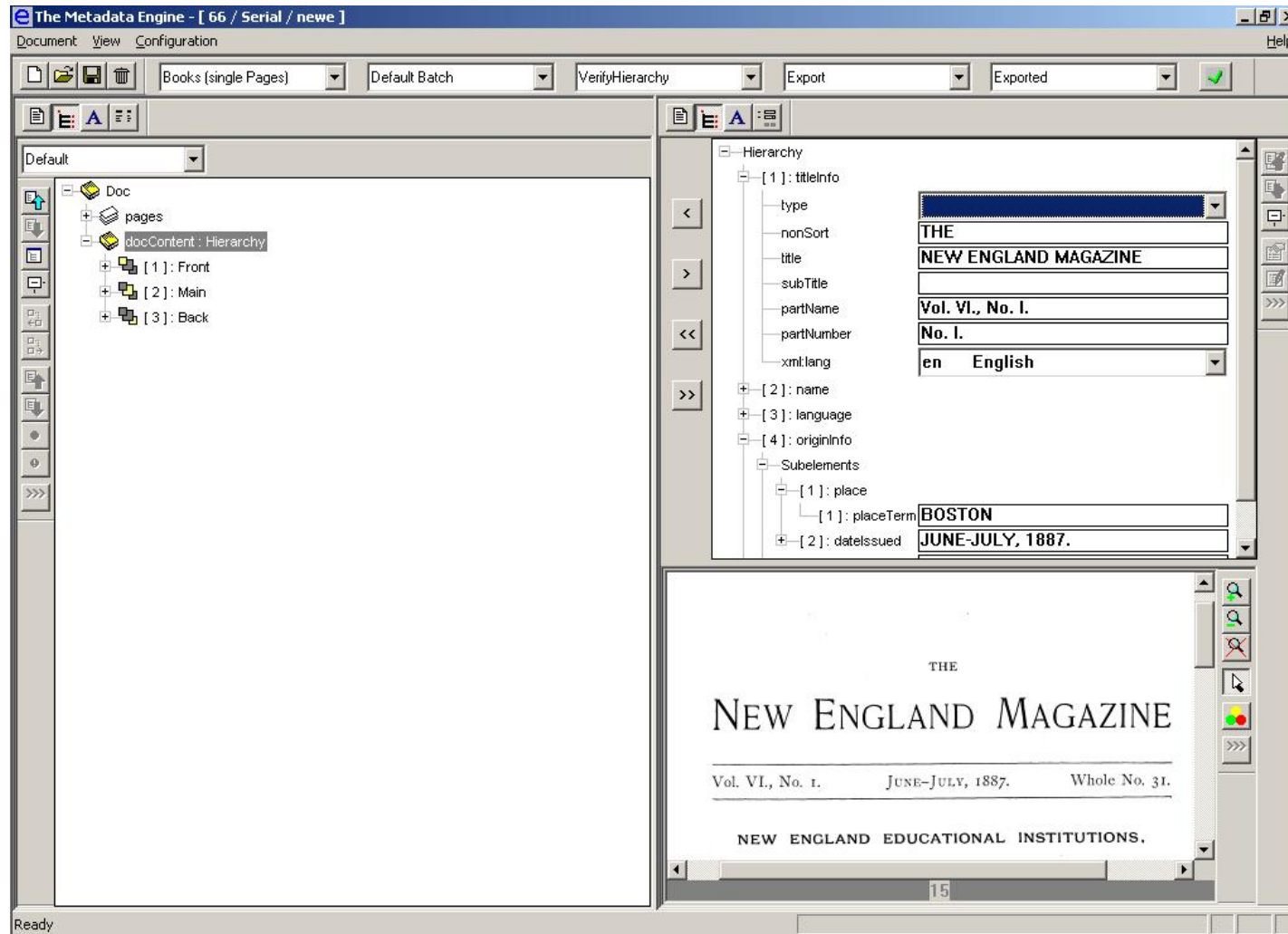
Digitization of Books and Journals (METAe)



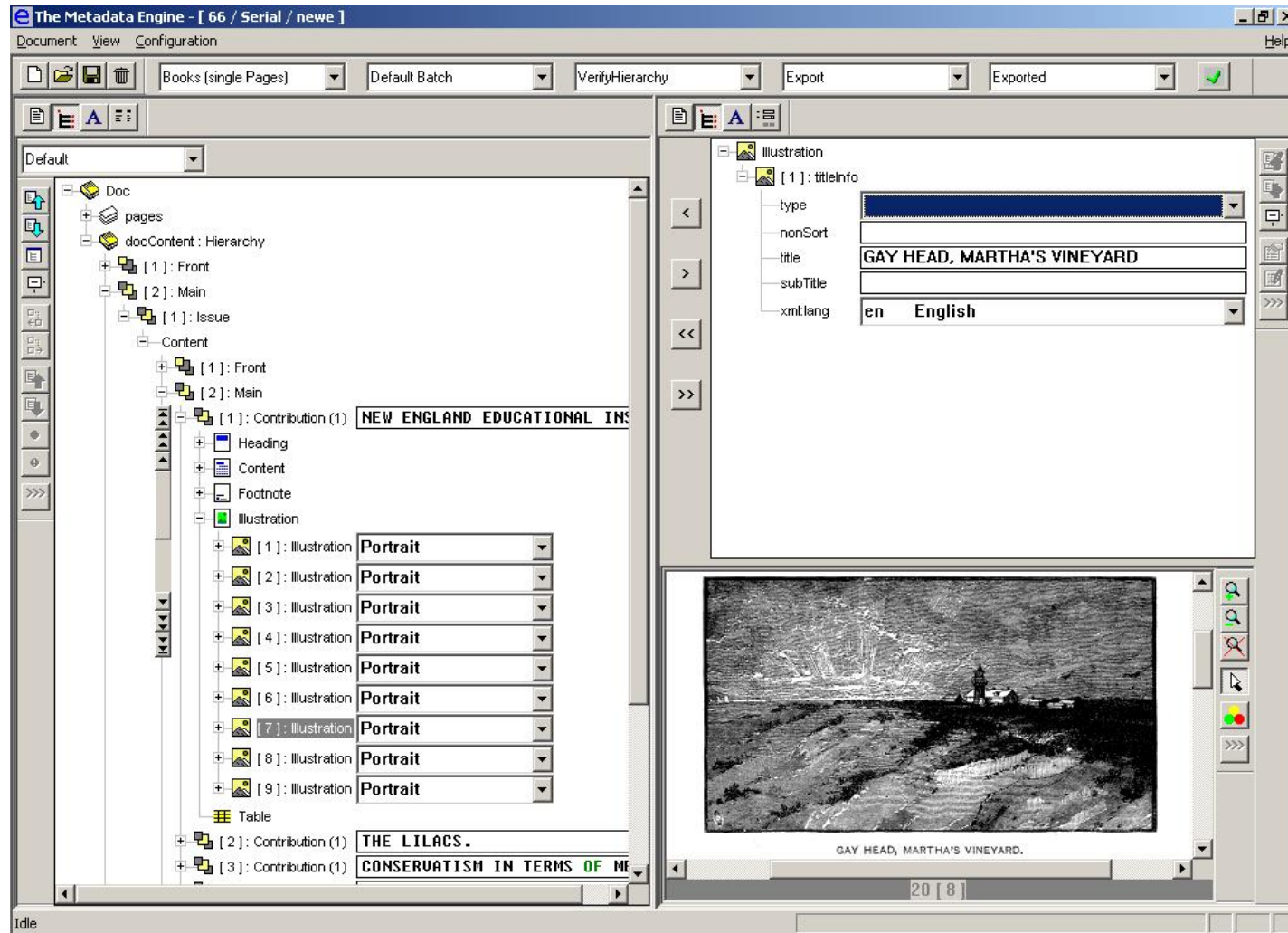
Digitization of Scientific Documents



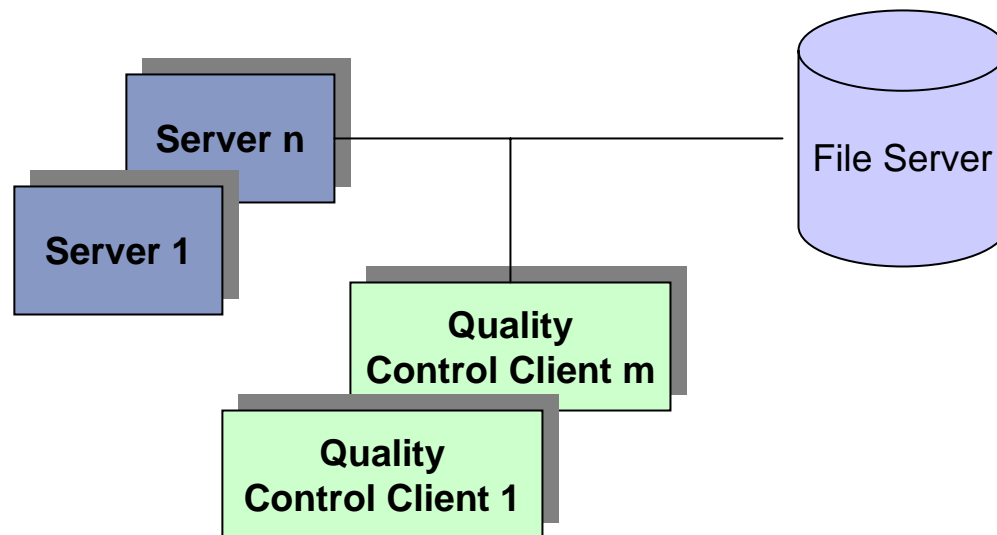
Manual Editing of Descriptive Metadata / Volume



Manual Editing of Descriptive Metadata / Illustration

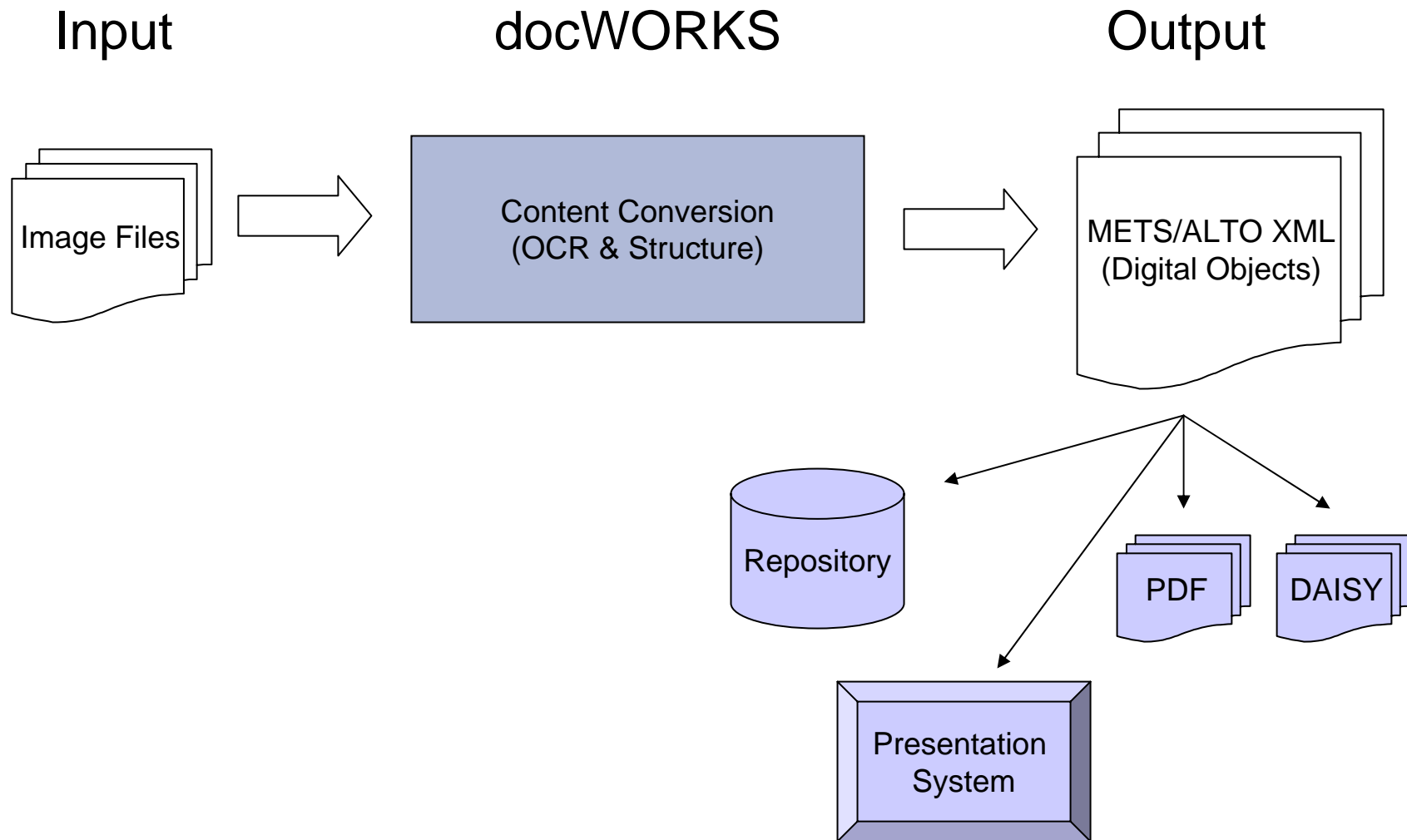


Scalable Client / Server Architecture



- Auto-Import
- Image Preprocessing
- Layout Analysis
- OCR
- Structural Analysis
- Export

In Practice: docWORKS in Digitization Projects



KBDK Markup Policy

148 Page no

012: Est-ce que Pierre a vu Marie à Paris?
013: Est-ce que Marie est partie?

Old Markup
Running title

chapter

Les verbes implicatifs négatifs

Les verbes implicatifs négatifs ont cette caractéristique que le complément est faux quand la principale est positive et vice-versa. Ils obéissent aux postulats de sens (114) et (115):

Formule

Marked as Formula because of signs.

114: $e(S) \Rightarrow \neg S$ $e(S)$ est une condition suffisante pour $\neg S$
115: $\neg e(S) \Rightarrow S$ $\neg e(S)$ est une condition nécessaire pour S

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:

116: (a) Pierre a cessé de parler.
(b) Pierre ne parle plus.
117: (a) Pierre n'a pas cessé de parler.
(b) Pierre parle.

Comme les verbes implicatifs positifs, les verbes implicatifs négatifs peuvent d'abord se diviser en deux groupes principaux exemplifiés par cesser de et douter de.

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

Table

1°. Le premier sous-groupe comprend:	cesser de	finir de
s'arrêter de	décourager de	s'interrompre de
2°. Le deuxième sous-groupe comprend:	douter de	
3°. Le troisième sous-groupe comprend:		
s'abstenir de	s'empêcher de	se priver de
s'excuser de	faillir (S)	se priver de
désigner de	ne pas lâcher de	refuser de
différer de	manquer à/de	renoncer à
se dispenser de	s'abstenir de	se priver de
épargner de	construire de	réfuter à
éviter de/que	oublier de	

28: Ce verbe se rencontre seulement dans des constructions négatives.

Footnote

Page part of Contribution in Main

Presentation of Converted Material

Adobe Reader - [mrs_fisher.PDF]

Datei Bearbeiten Anzeigen Dokument Werkzeuge Fenster Hilfe

Auswählen 87%

Digitale Fotos effizient organisieren

Leseseiten

Seiten

Anlagen

Kommentare

Optionen

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 grid-dle. 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-

17 von 84

Presentation of Converted Material

Adobe Reader - [virtual_landscapes.PDF]

Datei Bearbeiten Anzeige Dokument Werkzeuge Fenster Hilfe

Auswählen 100%

Adobe PDF vom Desktop aus erstellen

Lesezeichen

- "TEXAS GRANITES"
- Illustrations
 - Plate 1. Texas State Capitol Building, Austin. Built with red granite
 - Plate 2. Gray granite quarry operated by Gooch & Company, 6 miles west of Austin
 - Plate 3. The Llano quarrymen experience considerable difficulty getting their granite to the market. Rolling a 35-ton block eight miles to Llano for shipment.**
 - Plate 4. Taking Mocks of red granite out of Granite Mountain
 - Plate 5. Darragh Brothers' quarry at Granite Mountain, Burnet County
- Tables

Seiten

Anlagen

Kommentare

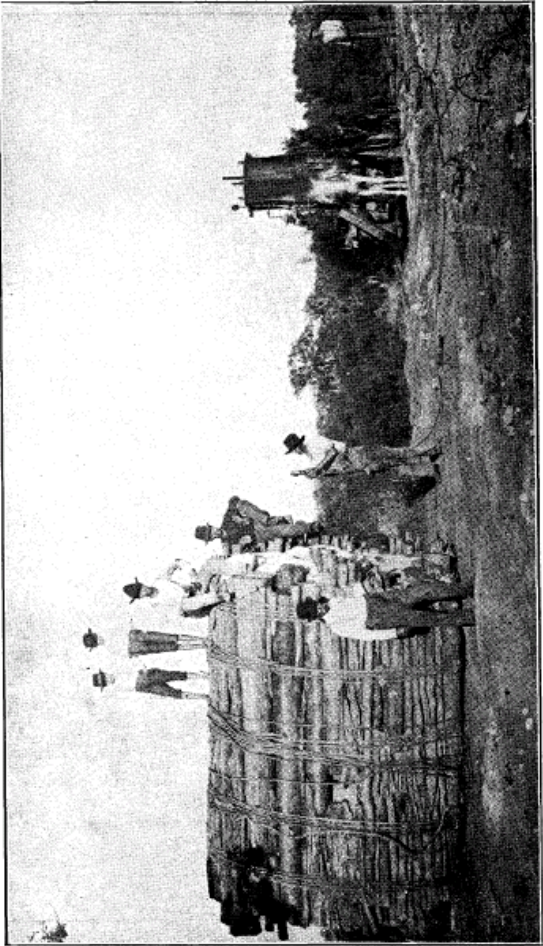


Plate 3. The Llano quarrymen experience considerable difficulty getting their granite to the market. Rolling a 35-ton block eight miles to Llano for shipment.

162,6 x 240,6 mm

12 von 14

Presentation of Converted Material

PDS Ladies' magazine. - Microsoft Internet Explorer

Datei Bearbeiten Ansicht Favoriten Extras ?

Zurück Suchen Favoriten Wechsels zu Links SnagIt

Adresse <http://pdstest.harvard.edu:8080/pdx/servlet/pds?op=f&id=410910&n=246&s=4>

Harvard University Library
Page Delivery Service

Ladies' magazine.

Search View Text Printable Version Related Links Help Copyright

Page Number: 241 Go to Sequence Number: 246 Go to

Ladies' magazine. [591 pages.]

- "THE LADIES' MAGAZINE."** [pp. unnumbered]
- [Blank Page] [pp. unnumbered page-unnumbered]
- [Blank Page] [pp. unnumbered page-unnumbered]
- [Picture Page] [pp. unnumbered page-unnumbered]
- [Title Page] [pp. unnumbered page-unnumbered]
- [Statement Page] [pp. unnumbered page-unnumbered]
- JANUARY. [pp. 1-49 (seq. 6-54)]
- FEBRUARY. [pp. 49-97 (seq. 54-102)]
- MARCH. [pp. 97-144 (seq. 102-149)]
- APRIL. [pp. 145-193 (seq. 150-198)]
- MAY. [pp. 193-240 (seq. 198-245)]
- JUNE. [pp. 241-289 (seq. 246-295)]**
- JULY. [pp. 289-337 (seq. 295-343)]
- AUGUST. [pp. 337-384 (seq. 343-390)]
- SEPTEMBER. [pp. 385-432 (seq. 391-438)]
- OCTOBER. [pp. 433-480 (seq. 439-486)]
- NOVEMBER. [pp. 481-529 (seq. 487-535)]
- DECEMBER. [pp. 529-576 (seq. 535-582)]
- INDEX. [pp. 3-8 (seq. 583-588)]
- [Blank Page] [pp. unnumbered page-unnumbered]
- [Blank Page] [pp. unnumbered page-unnumbered]
- [Blank Page] [pp. unnumbered page-unnumbered]

LADIES' MAGAZINE.

VOL. I. JUNE. No. VI.

DRESS.

"The world is still deceived with ornament." So said William Shakspeare, and two centuries have made, in this respect, little alteration. There seems to be, in mankind, a propensity to display, to prize outward show, to look with favor on the wearer of a fine suit, rather than on the merit of him who deserves one.

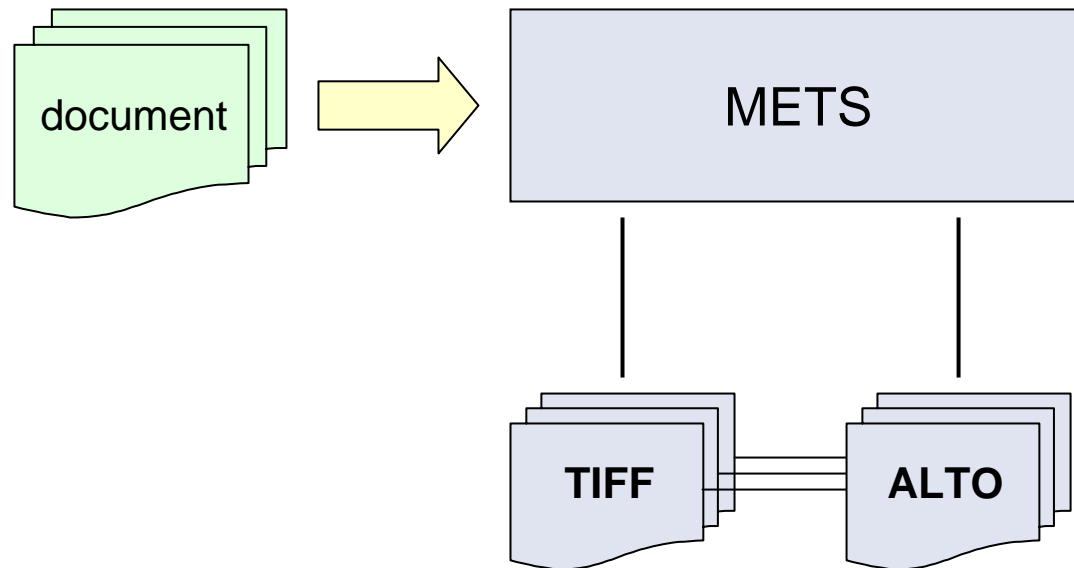
But such remarks have been made by every cynic since the days of Antisthenes, and I did not commence with the cynical intention of railing in "good set terms" against the modes and customs of the world. I believe, with the melancholy Jaques, that

— "Who cries out on fashion
That can therein tax any private party!—
Doth it not flow as hugely as the sea,
Till that the very, very means do ebb?"

This ebbing of the means is one of the most disagreeable drawbacks on a life of dissipation and display. Yet it is often salutary in its operation. When the tree is bowed by the fury of the storm, if, instead of sinking beneath the shock, the roots entwine and fix themselves more firmly during the agitation, then, when the tempest is over, that tree will rise again more healthy and vigorous. Just so

Internet

METS / ALTO XML



ALTO – **A**nalyzed **L**ayout and **T**ext **O**bject

METS

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

ALTO

■ Styles

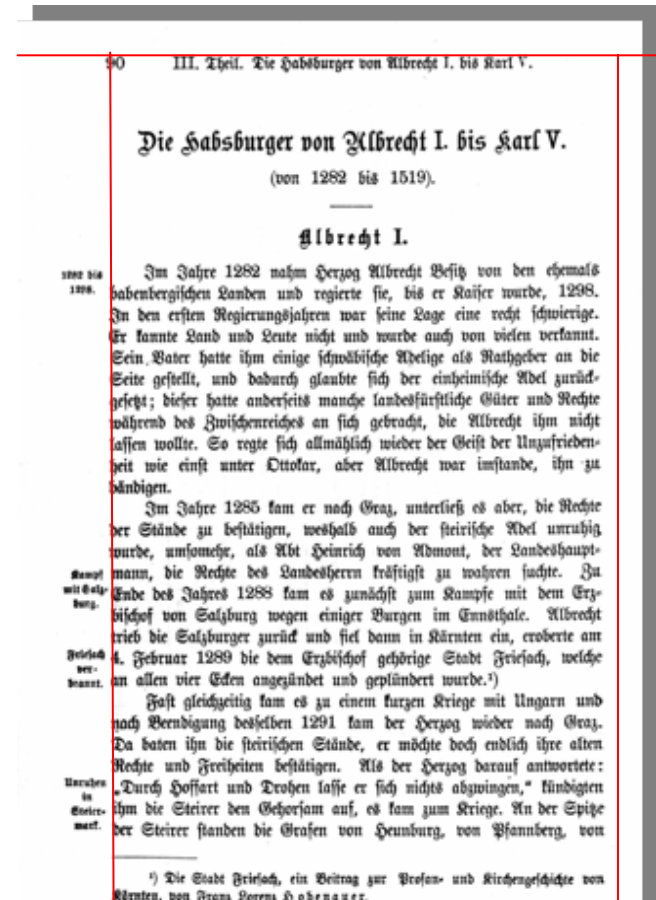
- Paragraph (alignment, linespacing, etc.)
- Font (name, size, bold, italic, etc.)

■ Layout

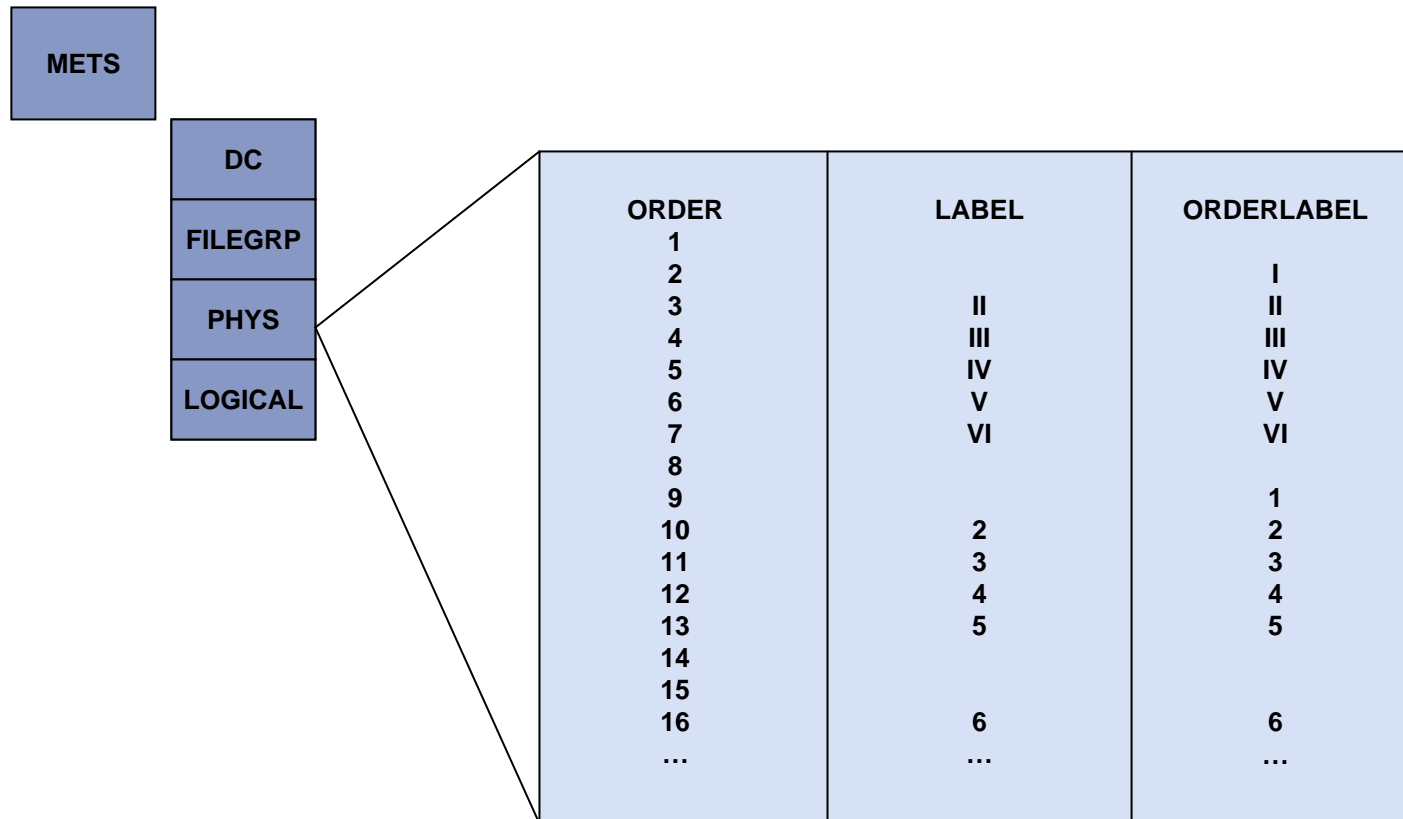
- Printspace
- TopMargin
- InnerMargin
- OuterMargin
- BottomMargin

■ Objects in 5 areas above:

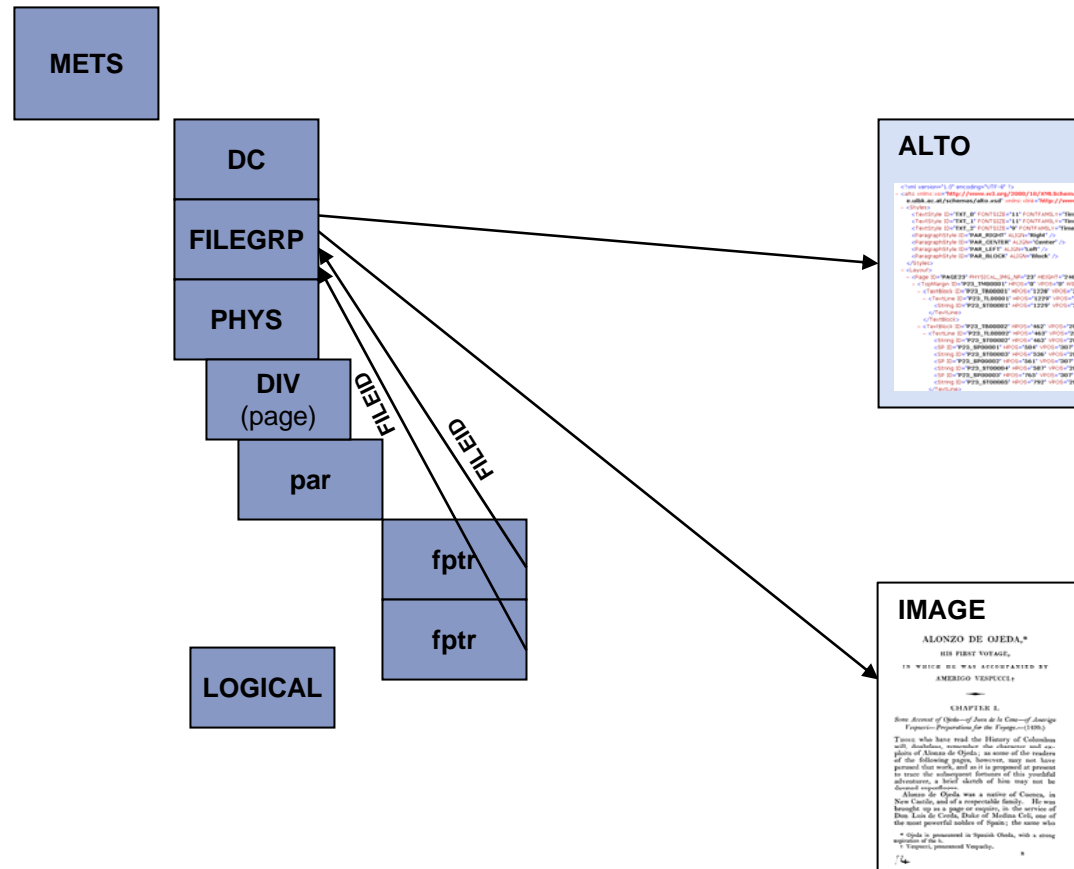
- Text block
 - Text lines
 - Strings [coordinates, string (as printed), substitution (hyphenation)]
 - Spaces
- Composed block
- Picture
- Table
- Formula



METS / physical structure



METS / physical structure



DLF spring forum 2005



ALTO / Page Layout and Text Content



```

D:\CCS\METAe\Metae\Export\EXP00099\EXP00099-ALTO00023.xml - Microsoft Internet Explorer
Datei Bearbeiten Ansicht Favoriten Extras ? Links »

<?xml version="1.0" encoding="UTF-8" ?>
- <alto xmlns:xsi="http://www.w3.org/2000/10/XMLSchema-instance" xsi:noNamespaceSchemaLocation="http://meta-
  e.uibk.ac.at/schemas/alto.xsd" xmlns:xlink="http://www.w3.org/TR/xlink">
- <Styles>
  <TextStyle ID="TXT_0" FONTSIZE="12" FONTFAMILY="Times New Roman" />
  <TextStyle ID="TXT_1" FONTSIZE="8" FONTFAMILY="Times New Roman" />
  <TextStyle ID="TXT_2" FONTSIZE="6" FONTFAMILY="Times New Roman" />
  <TextStyle ID="TXT_3" FONTSIZE="9" FONTFAMILY="Times New Roman" />
  <TextStyle ID="TXT_4" FONTSIZE="7" FONTFAMILY="Times New Roman" />
  <ParagraphStyle ID="PAR_LEFT" ALIGN="Left" />
  <ParagraphStyle ID="PAR_BLOCK" ALIGN="Block" />
</Styles>
- <Layout>
- <Page ID="PAGE23" PHYSICAL_IMG_NR="23" HEIGHT="3225" WIDTH="2104">
  <TopMargin ID="P23_TM00001" HPOS="0" VPOS="0" WIDTH="2104" HEIGHT="197" />
  <InnerMargin ID="P23_IM00001" HPOS="0" VPOS="197" WIDTH="163" HEIGHT="2790" />
  <OuterMargin ID="P23_OM00001" HPOS="1828" VPOS="197" WIDTH="276" HEIGHT="2790" />
  <BottomMargin ID="P23_BM00001" HPOS="0" VPOS="2987" WIDTH="2104" HEIGHT="238" />
- <PrintSpace ID="P23_PS00001" HPOS="163" VPOS="197" WIDTH="1665" HEIGHT="2790">
- <TextBlock ID="P23_TB00001" HPOS="455" VPOS="588" WIDTH="1076" HEIGHT="85" STYLEREFS="TXT_0 PAR_LEFT">
  - <TextLine ID="P23_TL00001" HPOS="193" VPOS="249" WIDTH="456" HEIGHT="36">
    <String ID="P23_ST00001" HPOS="193" VPOS="249" WIDTH="178" HEIGHT="30" CONTENT="ALONZO" />
    <SP ID="P23_SP00001" HPOS="371" VPOS="279" WIDTH="24" />
    <String ID="P23_ST00002" HPOS="395" VPOS="249" WIDTH="58" HEIGHT="30" CONTENT="DE" />
    <SP ID="P23_SP00002" HPOS="453" VPOS="279" WIDTH="25" />
    <String ID="P23_ST00003" HPOS="478" VPOS="249" WIDTH="171" HEIGHT="36" CONTENT="OJEDA,*" />
  </TextLine>
</TextBlock>
- <TextBlock ID="P23_TB00002" HPOS="643" VPOS="759" WIDTH="707" HEIGHT="61" STYLEREFS="TXT_1 PAR_LEFT">
- <TextLine ID="P23_TL00002" HPOS="272" VPOS="321" WIDTH="300" HEIGHT="27">
  <String ID="P23_ST00004" HPOS="272" VPOS="322" WIDTH="49" HEIGHT="20" CONTENT="HIS" />
  <SP ID="P23_SP00003" HPOS="321" VPOS="342" WIDTH="17" />
  <String ID="P23_ST00005" HPOS="338" VPOS="322" WIDTH="86" HEIGHT="20" CONTENT="FIRST" />
  <SP ID="P23_SP00004" HPOS="424" VPOS="342" WIDTH="16" />
  <String ID="P23_ST00006" HPOS="440" VPOS="321" WIDTH="132" HEIGHT="27" CONTENT="VOYAGE," />
</TextLine>
</TextBlock>
- <TextBlock ID="P23_TB00003" HPOS="255" VPOS="918" WIDTH="1484" HEIGHT="40" STYLEREFS="TXT_2 PAR_LEFT">
  - <TextLine ID="P23_TL00003" HPOS="108" VPOS="980" WIDTH="620" HEIGHT="17">

```

ALTO / Hyphenated Word

D:\CCS\METAe\Metae\Export\EXP00099\EXP00099-ALTO00023.xml - Microsoft Internet Explorer

```

</TextBlock>
- <TextBlock ID="P23_TB00007" HPOS="168" VPOS="1709" WIDTH="1651" HEIGHT="619" STYLEREFS="TXT_3 PAR_BLOCK">
- <TextLine ID="P23_TL00008" HPOS="72" VPOS="723" WIDTH="698" HEIGHT="32">
  <String ID="P23_ST00038" HPOS="72" VPOS="724" WIDTH="88" HEIGHT="24" CONTENT="THOSE" />
  <SP ID="P23_SP00023" HPOS="160" VPOS="748" WIDTH="19" />
  <String ID="P23_ST00039" HPOS="179" VPOS="724" WIDTH="55" HEIGHT="23" CONTENT="who" />
  <SP ID="P23_SP00024" HPOS="234" VPOS="748" WIDTH="17" />
  <String ID="P23_ST00040" HPOS="251" VPOS="724" WIDTH="62" HEIGHT="23" CONTENT="have" />
  <SP ID="P23_SP00025" HPOS="313" VPOS="748" WIDTH="18" />
  <String ID="P23_ST00041" HPOS="331" VPOS="723" WIDTH="60" HEIGHT="24" CONTENT="read" />
  <SP ID="P23_SP00026" HPOS="391" VPOS="748" WIDTH="17" />
  <String ID="P23_ST00042" HPOS="408" VPOS="724" WIDTH="40" HEIGHT="23" CONTENT="the" />
  <SP ID="P23_SP00027" HPOS="448" VPOS="748" WIDTH="19" />
  <String ID="P23_ST00043" HPOS="467" VPOS="723" WIDTH="105" HEIGHT="32" CONTENT="History" />
  <SP ID="P23_SP00028" HPOS="572" VPOS="755" WIDTH="18" />
  <String ID="P23_ST00044" HPOS="590" VPOS="724" WIDTH="30" HEIGHT="23" CONTENT="of" />
  <SP ID="P23_SP00029" HPOS="620" VPOS="755" WIDTH="13" />
  <String ID="P23_ST00045" HPOS="633" VPOS="724" WIDTH="137" HEIGHT="25" CONTENT="Columbus" />
</TextLine>
- <TextLine ID="P23_TL00009" HPOS="72" VPOS="756" WIDTH="697" HEIGHT="29">
  <String ID="P23_ST00046" HPOS="72" VPOS="757" WIDTH="55" HEIGHT="28" CONTENT="will," />
  <SP ID="P23_SP00030" HPOS="127" VPOS="785" WIDTH="20" />
  <String ID="P23_ST00047" HPOS="147" VPOS="757" WIDTH="134" HEIGHT="28" CONTENT="doubtless," />
  <SP ID="P23_SP00031" HPOS="281" VPOS="785" WIDTH="19" />
  <String ID="P23_ST00048" HPOS="300" VPOS="757" WIDTH="136" HEIGHT="23" CONTENT="remember" />
  <SP ID="P23_SP00032" HPOS="436" VPOS="785" WIDTH="22" />
  <String ID="P23_ST00049" HPOS="458" VPOS="756" WIDTH="41" HEIGHT="24" CONTENT="the" />
  <SP ID="P23_SP00033" HPOS="499" VPOS="785" WIDTH="18" />
  <String ID="P23_ST00050" HPOS="517" VPOS="757" WIDTH="124" HEIGHT="24" CONTENT="character" />
  <SP ID="P23_SP00034" HPOS="641" VPOS="785" WIDTH="18" />
  <String ID="P23_ST00051" HPOS="659" VPOS="758" WIDTH="48" HEIGHT="23" CONTENT="and" />
  <SP ID="P23_SP00035" HPOS="707" VPOS="785" WIDTH="18" />
  <String ID="P23_ST00052" HPOS="725" VPOS="767" WIDTH="44" HEIGHT="15" CONTENT="lex"
    SUBS_TYPE="HypPart1" SUBS_CONTENT="exploits" />
  <HYP CONTENT="-" />
</TextLine>
- <TextLine ID="P23_TL00010" HPOS="72" VPOS="789" WIDTH="697" HEIGHT="33">
  <String ID="P23_ST00053" HPOS="72" VPOS="790" WIDTH="73" HEIGHT="32" CONTENT="ploits"
    SUBS_TYPE="HypPart2" SUBS_CONTENT="exploits" />
  <SP ID="P23_SP00036" HPOS="145" VPOS="822" WIDTH="12" />
  <String ID="P23_ST00054" HPOS="157" VPOS="790" WIDTH="29" HEIGHT="23" CONTENT="of" />

```

ALTO / Hyphenated Word

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/>