Blue Mountain Reference

Cliff Wulfman

2013-01-22 Tue

Contents

1 Identifiers and Naming Conventions

Blue Mountain assigns a URI to each magazine, magazine issue, METS, and MODS record it maintains:

```
<BMTNPREFIX> ::= "urn:PUL:bluemountain"
<DATESTRING> ::= CCYY-MM-DD | CCYY-MM | CCYY
<ISSUEINDEX> ::= <0-9><0-9>
<BMTNID>     ::= "bmtn" <a-z><a-z><a-z>
<ISSUID>     ::= <BMTNID> "_" <DATESTRING> "_" <ISSUEINDEX>

<TITLEURI>    ::= <BMTNPREFIX> ":" <BMTNID>
<ISSUEURI>    ::= <BMTNPREFIX> ":" <ISSUEID>



<TITLEMETSURI> ::= <BMTNPREFIX> ":td:" <BMTNID>
<ISSUEMETSURI> ::= <BMTNPREFIX> ":td:" <ISSUEID>

<TITLEMODSURI> ::= <BMTNPREFIX> ":dmd:" <BMTNID>
<ISSUEMODSURI> ::= <BMTNPREFIX> ":dmd:" <ISSUEID>
```

This syntax is explained below.

1.1 Blue Mountain Identifiers (bmtnids)

Blue Mountain is a *database* of magazines: that is, it is a system that links information together. In order to link information about specific objects – magazine titles, magazine issues, ec. – Blue Mountain must assign a *unique identifier* to each object. (This is a well-known feature of information science.).

Blue Mountain adopts the Universal Resource Name conventions (URN) to compose unique identifiers for *titles* and *issues*, as well as for the metadata records (METS and MODS) used to encode information about them.

A *Journal Title* is the journal or magazine as a whole, as opposed to discrete volumes or issues – the entire run of the magazine. Blue Mountain assigns a sequential *blue mountain identifier*, or *bmtnid* to each title. The bmtnid will take the form *bmtnNNN*, where NNN is a hexavigesimal number (e.g., *aaa*, *aab*, *aac*, etc.).[1] Blue Mountain will maintain bmtnids in a *project registry*, a file maintained with other administrative files.

[1] This convention has been adopted to suport the naming conventions in Veridian, which prohibit the use of integers in identifiers.

1.2 Issuance Strings

Each journal *issue* will be assigned an *issue identifier* or *issueid* of the form *bmtnid_ issuanceString*, where *issuanceString* corresponds to the date of issuance and takes the form *CCYY-MM-DD_ II*, defined as follows:

CCYY A four-digit number representing the year of publication (e.g., 1912)

MM A two-digit number representing the month of publication, where January = 01, February = 02, etc.

DD A two-digit number representing the day of publication (e.g., 01, 02 .. 30, 31).

II A two-digit index of daily issuance (e.g., the first issue of the day is 01, the second is 02, and so on). This convention is adopted from the issuance of newspapers, which not infrequently issued a morning edition and an evening edition on the same day. Blue Mountain's adoption allows us to distinguish among magazine texts that were published on the same day: a regular issue and a supplement, for example.

1.2.1 How to Compose an Issuance String

Issueids, like ISO 8601 dates, are organized from most significant to least significant chronological unit (i.e., from year to day). This format has two advantages: it allows ids to be sorted naturally, and it enables *variable precision*: the representation of daily, monthly, or yearly issuance.

- If you know the year, month, and day of publication (e.g., you know that the issue was published on January 5th, 1912) :: then the issuance string os 1912-01-05_01.

- If you have two issues published on the same day (e.g., an issue and a special supplement were both issued on January 5th, 1912) :: then the issuance strings are 1912-01-05_01 and 1912-01-05_02.

- If you know only the year and month of publication (e.g., you know it was published in January, 1912, but you do not know on what day: 1912-01_01.

- If you know that the magazine published two issues monthly, but you do not know the dates of publication (e.g., you have two issues published in January, 1912): The issuance strings are 1912-01_01 and 1912-01_02.

- If you know only the year of publication (e.g., you know that the issue was published in 1912, but you do not know the month or, therefore, the day of publication): The issuance string is 1912_01.
- If you have several issues published in the same year, but you know neither the month nor the day of publication (e.g., you know the journal published two issues in 1912, but you do not know the months or days of publication): The issuance strings are 1912_01 and 1912_02.

A Real Example The journal *le coeur à barbe* has the Blue Mountain identifier **bmtnaad**. Only one issue was published, in April of 1920.

titleid	urn:PUL:bluemountain:bmtnaad
title METS id	urn:PUL:bluemountain:td:bmtnaad
title MODS id	urn:PUL:bluemountain:dmd:bmtnaad
issueid	urn:PUL:bluemountain:bmtnaad$_1$920 $-$ 04$_0$1
issue METS id	urn:PUL:bluemountain:td:bmtnaad$_1$920 $-$ 04$_0$1
issue MODS id	urn:PUL:bluemountain:dmd:bmtnaad$_1$920 $-$ 04$_0$1

1.3 File Names

Names of Blue Mountain files will be constructed using the naming convention described above.

1.3.1 Image File Names

```
<EXTENSION> ::= "tif" | "jp2"
<IMGINDEX>  ::= <0-9><0-9><0-9>
<FILENAME>  ::= <ISSUEID> "_" <IMGINDEX> "." <EXTENSION>
```

Image files shall be named *issueid_nnn.jp2* or *issueid_nnn.tif*, where

- *issuid* is the identifier of the issue;
- *nnn* is a three-digit number indicating the location of the image file in the sequence of image files (not necessarily the number printed on the page that has been photographed);
- *jp2* is the conventional file extension for JPEG2000 files.
- *tif* is the conventional file extension for TIFF files.

For example,

```
bmtnaad_1925-06-03_01_001.jp2
bmtnaad_1925-06-03_01_002.jp2
...
```

1.3.2 ALTO File Names

```
<EXTENSION> ::= "alto.xml"
<IMGINDEX>  ::= <0-9><0-9><0-9>
<FILENAME>  ::= <ISSUEID> "_" <IMGINDEX> "." <EXTENSION>
```

ALTO files shall be named *issueid_ nnn.alto.xml*, where

- *issuid* is the identifier of the issue
- *nnn* is a three-digit number corresponding to the sequence number of the image file to which this ALTO file corresponds
- *alto* indicates the schema used to encode the document
- *xml* indicates the format of the file.

For example,

```
bmtnaad_1925-06-03_01_001.alto.xml
bmtnaad_1925-06-03_01_002.alto.xml
...
```

1.3.3 METS File Names

```
<EXTENSION> ::= "mets.xml"
<FILENAME>  ::= <ISSUEID>  "." <EXTENSION>
```

METS files shall be named *issueid.mets.xml*, where

- *issueid* is the identifier of the issue
- *mets* indicates the schema used to encode the document
- *xml* indicates the format of the file.

For example,

```
bmtnaad_1925-06-03_01.mets.xml
```

1.3.4 PDF File Names

```
<EXTENSION> ::= "pdf"
<FILENAME>  ::= <ISSUEID> "." <EXTENSION>
```

PDF files shall be named *issueid.pdf*, where

- *issueid* is the identifier of the issue
- *pdf* indicates the format of the file.

For example,

```
bmtnaad_1925-06-03_01.pdf
```

2 Journal Objects

A *journal object* will comprise the following elements:

title-level descriptive metadata A detailed, machine-readable description of the periodical as a whole. Encoded in MODS for compatibility with library systems, but translatable into other formats (e.g., TEI).

title-level bibliography An article-level prose description. (*bmtnid.tei.xml*)

title-level metadata wrapper Pulls together the title-level metadata, the bibliography, and the issue-level metadata (*bmtnid.mets.xml*)

issues one or more issues, each of which entails the following:

> **preservation-quality images** high-quality TIFF files ('master TIFFs'), produced according to local best practices and in conformance with the FADGI standards (http://www.digitizationguidelines.gov/guidelines/digitize-technical.html).
>
> **generative image derivatives** more manageable forms of the master TIFFs, meant to serve as the source for online deliverables, etc. Encoded in the JPEG2000 format, according to specifications described below.
>
> **delivery derivatives** images optimized for delivery over the World Wide Web.
>
> **issue-level descriptive metadata** a MODS document (see below).

text encodings Initially these will be in the form of corrected OCR for each page, encoded in the ALTO schema (output by ABBYY via docWORKS). Future encodings will likely include TEI representations, derived from the ALTO documents, for detailed textual analysis.

deliverable text-under-image PDF another ABBYY output format.

issue-level metadata wrapper a METS document. The METS half of METS/ALTO, the structMap of this document links constituent-level items to the regions identified in the ALTO documents, and to the page image. (See below for detailed specification.)

3 Arrangement (Blue Mountain Directory Structure)

The components of the journal object have different storage and access requirements. Master TIFF files are very large binary files that will seldom be accessed but must be carefully preserved (they are expensive or impossible to replace). Image derivatives, too, are large binary files, but they can be regenerated from the master TIFFs and therefore require less care, but they will be accessed from a variety of sources (primarily the web). PDF files are hybrids: they are large binary files, composites of image derivatives and OCR output that cannot easily be recreated and so must be preserved more carefully than image derivatives while still being accessible. Metadata files are relatively small but very expensive to replace, and so must be curated carefully. They are also liable to updating, so version tracking is important.

The Blue Mountain Project will manage these assets separately. The non-binary data and metadata will be stored and managed in a distributed version control system (DVCS), which will enable change management, collaborative development among PUL and its METS/ALTO vendor, and resource sharing, as stipulated in the grant.

Master TIFF files and text-under-image PDFs will be maintained in a *preservation store*; image derivatives, and copies of the PDFs and the metadata, will be kept in an *access store.*

3.1 The Metadata Store

Metadata will be organized as a hierarchy of files and directories, like this:

```
- periodicals/
  - bmtnID/
    - bmtnID.mets.xml
    - bmtnID.mods.xml
    - bmtnID.tei.xml
    - issues/
```

The issues/ directory will be organized by publication date, following the same convention as that used for constructing identifiers. So, for example,

```
- bmtnabi/
  - issues/
    - 1859/
      - 01/
        - 05_01/
          - bmtnid_issueid.mets.xml
          - bmtnid_issueid.mods.xml
          - bmtnid_issueid.tei.xml
          - alto/
            - bmtnid_issueid-001.alto.xml
            - bmtnid_issueid-002.alto.xml
```

3.2 The Preservation Store

The Preservation Store will be arranged as a filesystem mirroring the structure of the metadata tree and rooted at /usr/share/BlueMountain/pstore/periodicals.

```
- pstore/
  - periodicals/
    - bmtnid/
      - issues/
        - CCYY/
          - MM/
            - DD_II/
              - bmtnid_issueid.pdf
              - bmtnid_issueid_001.tif
              - bmtnid_issueid_002.tif
```

3.3 The Access Store

Like the Preservation Store, the Access store will be arranged as a filesystem mirroring the structure of the metadata tree; it will be rooted at

/usr/share/BlueMountain/astore/periodicals.

```
- astore/
  - periodicals/
    - bmtnid/
      - issues/
        - CCYY/
          - MM/
            - DD_II/
              - bmtnid_issueid.pdf
              - generative/
                - bmtnid_issueid_001.jp2
                - bmtnid_issueid_002.jp2
                - bmtnid_issueid_003.jp2
              - delivery/
                - bmtnid_issueid_001.jp2
                - bmtnid_issueid_002.jp2
                - bmtnid_issueid_003.jp2
```

4 Profiles

4.1 METS Profile

There are two kinds of METS records in Blue Mountain:

1. **Title-Level METS** – A METS document encapsulating information about the magazine title as a whole.
2. **Issue-Level METS** – A METS document encapsulating information about an individual issue of a magazine.

These are described in greater detail below.

4.1.1 Title-Level METS

(Greater detail to come.)

The metadata for the title will be encapsulated in a title-level METS record: the title-level descriptive metadata (either as an embedded MODS record or pointed to), a pointer to the bibliographic history, and (possibly) pointers to issue-level metadata.

4.1.2 Issue-Level METS

The metadata for each issue shall be encapsulated in a METS record. A skeleton sample of such a record is the following:

```
<?xml version="1.0" encoding="UTF-8"?>
<mets xmlns="http://www.loc.gov/METS/"
      xmlns:xlink="http://www.w3.org/1999/xlink"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xsi:schemaLocation="http://www.loc.gov/METS/ http://www.loc.gov/standards/mets/me
      TYPE="Magazine"
      OBJID="urn:PUL:bluemountain:bmtnaad_1925-06-03_01"
      LABEL="vozrozhdenie_1925-06-03_01">
  <metsHdr>
    <agent ROLE="CREATOR" TYPE="ORGANIZATION">
      <name>Princeton University Library, Digital Initiatives</name>
    </agent>
    <metsDocumentID TYPE="URN">urn:PUL:bluemountain:td:bmtnaad_1925-06-03_01</metsDocu
  </metsHdr>
  <dmdSec ID="dmd1">
    <mdWrap MDTYPE="MODS">
      <xmlData>
        <!-- MODS record goes here -->
      </xmlData>
    </mdWrap>
  </dmdSec>

    <!--Use a single administrative section (<amdSec>) as a
        wrapper for the technical metadata for all the images-->
  <amdSec>
    <techMD ID="techmd1">
      <!-- technical metadata (MIX) for first image -->
      <mdWrap MDTYPE="NISOIMG">
        <!-- The technical metadata docWorks provides goes here -->
      </mdWrap>
    </techMD>
    <techMD ID="techmd2">
      <!-- technical metadata for the second image -->
      <mdWrap MDTYPE="NISOIMG"/>
    </techMD>
    <!-- <techMD> elements for remaining image files -->
  </amdSec>

  <fileSec>
    <fileGrp ID="IMGGRP" USE="Images">

      <!-- Note that the AMDID attribute contains the ID of the
      <techMD> element corresponding to the file. Note, too,
      the use of the GROUPID attribute, which groups together
      the image file and its corresponding ALTO file. -->

      <file ID="IMG001" AMDID="techmd1" GROUPID="page1" MIMETYPE="image/jp2" CHECKSUM="
        <FLocat LOCTYPE="URL" xlink:href="file://.bmtnaad_1925-06-03_01_001.jp2"/>
      </file>
      <file ID="IMG002" AMDID="techmd2" GROUPID="page2" MIMETYPE="image/jp2" CHECKSUM="
        <FLocat LOCTYPE="URL" xlink:href="file://.bmtnaad_1925-06-03_01_002.jp2"/>
      </file>
    </fileGrp>

    <fileGrp ID="ALTOGRP" USE="OCR">
      <file ID="ALTO001" GROUPID="page1" MIMETYPE="text/xml" CHECKSUM="xxxx" CHECKSUMTY
        <FLocat LOCTYPE="URL" xlink:href="file://.bmtnaad_1925-06-03_01_001.alto.xml"/
      </file>
      <file ID="ALTO002" GROUPID="page2" MIMETYPE="text/xml" CHECKSUM="xxxx" CHECKSUMTY
        <FLocat LOCTYPE="URL" xlink:href="file://.bmtnaad_1925-06-03_01_002.alto.xml"/
      </file>
    </fileGrp>
  </fileSec>
  <structMap TYPE="PHYSICAL">
    <div/>
  </structMap>
  <structMap TYPE="LOGICAL">
    <div/>
  </structMap>
</mets>
```

The root element <mets> contains these attributes:

TYPE the fixed value *Magazine*

OBJID the URN for the issue

LABEL the *issueid*

<metsHdr> The <metsHdr> element shall contain two elements:

<agent> A constant value for all records:

```
<agent ROLE="CREATOR" TYPE="ORGANIZATION">
 <name>Princeton University Library, Digital Initiatives</name>
</agent>
```

<metsDocumentID TYPE="URN"> Contains a string whose contents is composed as follows:

```
PREFIX:ISSUID
```

Where *PREFIX* is the following fixed value:

```
urn:PUL:bluemountain:td:
```

And *ISSUEID* is the issue identifier, computed using the rules above.

<dmdSec> The record contains a single <dmdSec> element with an ID attribute of "dmd1"' it contains an embedded MODS record for the issue (described below).

<amdSec> The <amdSec> contains a *<techMD>* element for each image file (a <mix> record).

<fileSec> The fileSec comprises two *<fileGrp>* elements: one for the images and one for the ALTO records.

<fileGrp ID="IMGGRP" USE="Images"> The *IMGGRP* file group contains <file> elements that indicate the location of each image file, with attributes linking the file to the corresponding technical metadata and to the corresponding ALTO file.

- <file>

  ID a unique XML id

AMDID the ID of the <*techmd*> element corresponding to the image file

GROUPID an ID that links an image file to an ALTO file. The image file for a page and the ALTO file containing the OCR output for that page share an id (conventionally named *pageN*, where N is a sequence number).

MIMETYPE the constant "image/jp2" for jpeg2000 images

CHECKSUM the checksum of the file, according to the algorithm specified in *CHECKSUMTYPE*

CHECKSUMTYPE the algorithm used to compute the checksum; usually SHA-1.

<fileGrp ID="ALTOGRP" USE="OCR"> Like the <*fileGrp*> for images, but corresponding to the ALTO files. (The ALTO files do not have technical metadata, so there is no AMDID attribute.)

<structMap> The <structMap> element describes a hierarchical arrangement of the parts (<div>s) making up the digital object described by the METS. For this project, there are two kinds: a *physical structMap*, which delineates the pages of the newspaper issue in reading order, and a *logical structMap*, which functions as an outline of the newspaper's contents. Both of these are assembled by docWorks, using configuration rules.

<structMap type="PHYSICAL">

<structMap type="LOGICAL">

- The <div> hierarchy
  The outlines below show the hierarchical relationship among the <div> elements in the logical structMap. Each div is described more fully below.
  - Magazine
    - Volume+
      - Issue+
      - Contents
      - { Article* | Illustration* | Section* }
      - Advertisements

        - { SponsoredAd+ | Section* }
- Article
    - Header*
        - Head+
        - Byline*
    - Body
        - { Paragraph* | Section* }
- Illustration
    - Graphic+
    - Caption?
        - Paragraph+
- SponsoredAd
    - { Graphic* | Paragraph* }
- Section
    - Header?
    - Body
        - { Article* | Illustration* | SponsoredAd* | Section* }
- Paragraph
    - TextBlock+

- <div TYPE="Magazine">
  The root <div> of the logical structMap is <div TYPE="Magazine">. It must contain one or more <div TYPE="Volume"> elements (in practice it will contain only one).

  Attributes:

  TYPE must be “Magazine”

  LABEL The name of the magazine, equivalent to the top-level <mods:titleInfo> element.

- <div TYPE=”Volume”>?
  A <div> representing a (possibly) bound volume of issues. In most cases, we are representing each issue of a magazine as a separate digital object, so the <div TYPE=”Volume”> element will in practice contain only one <div TYPE=”Issue”>.

  Attributes:

  TYPE must be “VOLUME”

  LABEL The volume caption, if present

- <div TYPE=”Issue”>
  A <div> representing the actual issue. It contains the “contents” of the paper: the editorial content and the advertisements.

  Attributes:

  TYPE must be “ISSUE”

  LABEL The issue number and the date of publication

  DMDID the ID of the <dmdSec> for the object (in practice, always “dmd1”)

  The Issue <div> contains, in most cases, three sub-<div>s: <div TYPE=”PublicationInfo”>, <div TYPE=”EditorialContent”> and <div TYPE=”SponsoredAdvertisements”>, described below.

- <div TYPE=”PublicationInfo”>
  Contains <div>s corresponding to the metadata about the magazine printed in the issue itself: mastheads, nameplates, folio lines, page numbers, etc.

- <div TYPE=”EditorialContent” LABEL=”Contents”>
  Contains <div>s corresponding to the TextContent and Illustration elements, in publication order. These elements have DMDID attributes whose values link them to the corresponding <relatedItem> elements in the <mods> record.

- <div TYPE=”SponsoredAdvertisements” LABEL=”Advertisements”>
  Contains <div>s corresponding to the SponsoredAdvertisement

elements, in publication order. These elements have DMDID attributes whose values link them to the corresponding <relatedItem> elements in the <mods> record.

- <div TYPE=”TextContent”>

  A <div> representing a piece of editorial content: an article, a review, a letter, a poem, etc.

  Editorial content takes a number of forms: it may or may not have a headline; it may or may not have a byline; it may have subsections, each with its own headline (subhead).

  A TextContent <div> MAY contain a <div TYPE=”Header”>; it will always have a <div TYPE=”Body”>.

  Attributes:

  TYPE must be “TextContent”

  DMDID the ID of the <mods:relatedItem type=”constituent”> element corresponding to this piece in the newspaper.

  LABEL SHOULD be equivalent to the contents of the mods:relatedItem/mods:titleInfo/mo element

- <div TYPE=”Header”>

  A <div> containing the component’s (the TextContent, SponsoredAd, or Section) heading information: a combination of headline and byline. The Header may contain one or more Head elements (encompassing, for example, a headline and a subhead); it may also contain one or more Byline elements (which may not necessarily be physically contiguous in the physical layout of the page).

  Attributes:

  TYPE must be “Header”

- <div TYPE=”Head”>

  A <div> designating the region associated with a head of some kind: a headline, a subhead, etc.

  Attributes:

  TYPE must be “Head”

- <div TYPE=”Byline”>
  A <div> designating one or more regions associated with the writer of an article: usually the writer’s name, but sometimes also the writer’s position or other biographical information.

  Attributes:

  TYPE must be “Byline”

- <div TYPE=”Body”>
  A container <div> for the body of an article or section. A BODY may contain paragraphs, illustrations, or sections, in any order.

- <div TYPE=”Paragraph”>
  A <div> that contains one or more text blocks representing the contents of a logical paragraph. Paragraphs have a sequential order within their containing article, caption, or sponsored ad.

  Attributes:

  TYPE must be “Paragraph”

  ORDER the index of the paragraph in its containing div (1, 2, etc.).

- <div TYPE=”Section”>
  A section is a container <div> of other <div>s. It may or may not have a Header; it will contain some combination of articles, illustrations, SponsoredAds, and other sections.

- <div TYPE=”Illustration”>

- <div TYPE=”Graphic”>
  A div designating the location of a graphic on the page.

- <div TYPE=”Caption”>

- <div TYPE=”SponsoredAd”>

- <div TYPE=”TextBlock”>
  A div designating the region of a block of text on a page.

4.2 MODS Profile

There are two kinds of MODS records in Blue Mountain:

1. Title-level MODS
2. Issue-level MODS

4.2.1 Title-Level MODS

The descriptive metadata for most, if not all, of the Blue Mountain titles has been taken from MARC records retrieved from Princeton's OPAC and machine-converted, then edited and enhanced by hand. Here is a sample:

```
<?xml version="1.0" encoding="UTF-8"?>
<mods xmlns="http://www.loc.gov/mods/v3">

  <identifier type="bmtn">urn:PUL:bluemountain:bmtnaad</identifier> <!-- (identif

  <recordInfo>
    <recordIdentifier>urn:PUL:bluemountain:dmd:bmtnaad</recordIdentifier> <!-- (r
    <recordContentSource authority="marcorg">NjP</recordContentSource>
    <recordOrigin>http://catalog.princeton.edu/cgi-bin/Pwebrecon.cgi?BBID=4939605
    <languageOfCataloging>
      <languageTerm authority="iso639-2b" type="code">eng</languageTerm>
    </languageOfCataloging>
  </recordInfo>

  <titleInfo>
    <nonSort>Le</nonSort>
    <title>coeur à barbe</title>
    <subTitle>journal transparent</subTitle>
  </titleInfo>

  <name type="personal" authority="viaf" valueURI="http://viaf.org/viaf/73848255"
    <namePart type="family">Eluard</namePart>
    <namePart type="given">Paul</namePart>
    <namePart type="date">1895-1952</namePart>
  </name>
  <name type="personal" authority="viaf" valueURI="http://viaf.org/viaf/96123513"
    <namePart type="family">Ribemont-Dessaignes</namePart>
    <namePart type="given">Georges</namePart>
    <namePart type="date">1884-1974</namePart>
  </name>
  <name type="personal" authority="viaf" valueURI="http://viaf.org/viaf/27072443">
    <namePart type="family">Tzara</namePart>
    <namePart type="given">Tristan</namePart>
    <namePart type="date">1896-1963</namePart>
  </name>

  <typeOfResource>text</typeOfResource>

  <originInfo script="Latn">
    <place>
      <placeTerm type="text">Paris</placeTerm>
    </place>
    <dateIssued>1922</dateIssued>
    <dateIssued encoding="iso8601" point="start">1922</dateIssued>
    <dateIssued encoding="iso8601" point="end">1922</dateIssued>
  </originInfo>
  <language>
    <languageTerm authority="iso639-2b" type="code">fre</languageTerm>
  </language>
  <subject authority="lcsh">
    <topic>Dadaism</topic>
    <genre>Periodicals</genre>
  </subject>
  <subject authority="lcsh">
    <topic>Dadaism</topic>
    <geographic>France</geographic>
    <genre>Periodicals</genre>
  </subject>
</mods>
```

- The MODS record contains an <identifier> element whose type is *bmtn*. Its value is a URN for the title, which is of the form

  ```
  urn:PUL:bluemountain:BMTNID
  ```

  where the string *urn:PUL:bluemountain* is constant (for all Blue Mountain URNs) and *BMTNID* is the Blue Mountain project identifier of the periodical.

- The MODS record also contains a <recordInfo> element, which provides a link back to the original OPAC record, as well as a <recordIdentifier> uniquely identifying the record itself; it is simply the journal URN with *dmd* inserted into the identifier:

  `urn:PUL:bluemountain:dmd:BMTNID`

- The <name> elements are associated with authorities to enhance search and broaden the interconnectedness of the data. http://viaf.org is the preferred authority; http://id.loc.gov should be consulted when a name is not found in viaf.org; if a name is found in neither, a local authority will be created (To be determined later).

- Dates are encoded in ISO standard 8601 format (see http://www.iso.org/iso/catalogue_detail?csnu for an overview see http://en.wikipedia.org/wiki/ISO_8601). The extended form of the representation is preferred.

- Subject headings will conform with existing standards in a manner yet to be determined.

4.2.2 Issue-Level MODS

Blue Mountain encodes descriptive metadata for the contents of each magazine issue, so the issues may be searched and analyzed.

```
<mods xmlns="http://www.loc.gov/mods/v3">
  <recordInfo>
    <recordIdentifier>urn:PUL:bluemountain:dmd:bmtnabg_1911-01-05_01</recordIdentif
  </recordInfo>
  <identifier type="bmtn">urn:PUL:bluemountain:bmtnabg_1911-01-05_01</identifier>
  <typeOfResource>text</typeOfResource>
  <genre>Periodicals-Issue</genre>
  <titleInfo>
    <title>Der Sturm</title>
    <subTitle>Wochenschrift für Kultur und die Künste</subTitle>
  </titleInfo>
  <part type="issue">
    <detail type="volume">
      <number>1</number>
      <caption>Jahrgang 1911</caption>
    </detail>
  <detail type="number">
    <number>45</number>
    <caption>Nummer 45</caption>
  </detail>
</part>
<originInfo>
  <dateIssued>5. Januar 1911</dateIssued>
  <dateIssued keyDate="yes" encoding="w3cdtf">1911-01-05</dateIssued>
</originInfo>
<relatedItem type="host" xlink:type="simple" xlink:href="urn:PUL:bluemountain:bmtn
  <recordInfo>
    <recordIdentifier>urn:PUL:bluemountain:dmd:bmtnabg</recordIdentifier>
  </recordInfo>
</relatedItem>
<relatedItem type="constituent" ID="c001">
  <titleInfo lang="ger">
    <title>INHALT:</title>
  </titleInfo>
  <typeOfResource>text</typeOfResource>
  <language>
    <languageTerm authority="iso639-2b">ger</languageTerm>
  </language>
  <part>
    <extent unit="page">
      <start>1</start>
    </extent>
  </part>
  <genre type="CCS">TextContent</genre>
</relatedItem>
<relatedItem type="constituent" ID="c002">
  <titleInfo lang="ger">
    <title>Vorüber</title>
  </titleInfo>
  <name type="personal">
    <displayForm>Zeichnung von Oskar Kokoschka</displayForm>
    <role>
      <roleTerm type="code" authority="marcrelator">cre</roleTerm>
    </role>
  </name>
  <typeOfResource>still image</typeOfResource>
```

```
       <part>
         <extent unit="page">
           <start>1</start>
         </extent>
       </part>
       <genre type="CCS">Illustration</genre>
     </relatedItem>

     <!-- The remaining constituents go here -->
</mods>
```

<mods> The root element of the document. When output as a standalone document, it has fixed attributes, as illustrated below:

```
<mods xmlns="http://www.loc.gov/mods/v3"
      xmlns:xlink="http://www.w3.org/1999/xlink"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xsi:schemaLocation="http://www.loc.gov/mods/v3
        http://www.loc.gov/mods/v3/mods-3-4.xsd">
```

<recordInfo> The <recordInfo> element contains information about the MODS record itself. It shall contain a *<recordIdentifier>* element, as below.

<recordIdentifier> The <recordIdentifier> element contains the ISSUEMODSURI, the unique identifier for the MODS record itself, as described above.

<identifier type=”PUL”> The <identifier> element is used to identify the resource the MODS record describes (the magazine issue). Its value is the resource’s ISSUEURI, as described above.

```
<identifier type="PUL">urn:PUL:bluemountain:bmtnaad_1925-06-03_01</identifier>
```

<relatedItem type=”host”> Each issue-level MODS record is related to the title-level record *via* a <relatedItem type=’host’> element.

```
<relatedItem type="host" xlink:type="simple" xlink:href="urn:PUL:bluemountain:PUBID">
  <recordInfo>
    <recordIdentifier>urn:PUL:bluemountain:dmd:PUBID</recordIdentifier>
  </recordInfo>
</relatedItem>
```

where *PUBID* is the publication identifier of the title.

The xlink:href shows the semantic relation between the issue and its host; the <recordIdentifier> is a specific key to the title-level record.

<language> The <language> element(s) indicates the language of the resource (the magazine issue). If the issue contains material written in several languages, the record should include a <language> element for each one. The value of the <languageTerm> element must be drawn from iso639-2b. For example:

```
<language>
  <languageTerm type="code" authority="iso639-2b">rus</languageTerm>
 </language>
```

<titleInfo> The <MODS:titleInfo> element shall be determined by standard cataloging rules.

```
<titleInfo>
  <nonSort>Le</nonSort>
  <title>coeur à barbe</title>
  <subTitle>journal transparent</subTitle>
</titleInfo>
```

<MODS:part> The <MODS:part> element shall take the following form:

```
<MODS:part>
 <MODS:detail type="volume">...</MODS:detail>
 <MODS:detail type="issue">...</MODS:detail>
</MODS:part>
```

<MODS:detail type=”volume”>

```
 <MODS:detail type="volume">
  <MODS:number>ARABICVOL</MODS:number>
  <MODS:caption>Vol. MASTHEADVOL</MODS:caption>
</MODS:detail>
```

Where

- ARABICVOL is the volume number expressed as a non-formatted arabic numeral (e.g., 1, 2, 3, . . . 10, 11, . . .)
- MASTHEADVOL is the volume number as it appears in the masthead.

<MODS:detail type=”issue”> The <MODS:detail type=”issue”> element shall take one of two possible forms:

- For “normal” issues (i.e., those following the recorded sequence of publication), record both the sequential number of the issue as an arabic numeral and the issue number as it appears in the masthead:

  ```
  <MODS:detail type="issue">
   <MODS:number>ARABICISSUE</MODS:number>
   <MODS:caption>No. MASTHEADISSUE</MODS:caption>
  </MODS:detail>
  ```

Where

- ARABICISSUE is the issue number expressed as a non-formatted arabic numeral (e.g., 1, 2, 3, . . . , 10, 11, . . .)
- MASTHEADISSUE is the volume number as it appears in the masthead.
- For “special” issues (e.g., supplements, etc.), for which there is no sequential number for the

issue, the <MODS:detail type=”issue”> element should take the following form:

```
<MODS:detail type="issue">
 <MODS:caption>CAPTIONTEXT</MODS:caption>
</MODS:detail>
```

Where *CAPTIONTEXT* is determined using standard cataloging rules. ORG-LIST-END-MARKER

<MODS:originInfo> The <MODS:originInfo> element shall be used to record the date of issuance, as follows:

```
<MODS:originInfo>
 <MODS:dateIssued>PRINTEDDATE</MODS:dateIssued>
 <MODS:dateIssued encoding="iso8601" keyDate="yes">ISODATE</MODS:dateIssued>
</MODS:originInfo>
```

Where

- *PRINTEDDATE* is the date as it appears in the cover page FolioLine, or in the Masthead.

- *ISODATE* is the value of the date in the masthead, expressed in iso8601 format (YYYY-MM-DD) – see http://www.w3.org/TR/NOTE-datetime for details.

4.2.3 <relatedItem type="constituent">

Traditional library cataloging does not extend to the contents of periodicals, yet this level of description is precisely what is required by scholars of periodicals, and the Blue Mountain Project is committed to providing it, as well as to formulating guidelines, in cooperation with scholars and librarians, for this level of description. The specifications for this description, therefore, must be considered work in progress, work that will necessarily evolve over the course of the Project.

That being said, the Project will, at the outset, capture information about the following sorts of constituents:

- traditional editorial content (articles, features, letters to the editor, etc.)

- significant illustrations (figures, tip-ins, etc.)

- advertisements

The last sort – advertisements – is the most controversial, and the most difficult for librarians to understand, although advertisements are among the most heavily studied parts of historical periodicals. There are at present no established rules for describing advertisements, and their variety and abundance pose serious practical challenges to projects with limited resources. This version of the specification, therefore, provides little guidance on the description of periodicals, other than providing a framework for this level of detail to be created at a future date, by scholars, researchers, and other students of the material who wish to advance scholarship by enhancing the data provided here.

- TextContent

- Illustration

- SponsoredAd

- Section

They are described in greater detail below.

Here is a hypothetical example, in English:

```
<relatedItem type="constituent" ID="c17">
  <titleInfo>
   <nonSort>A</nonSort>
   <title>Modest Proposal</title>
  </titleInfo>
  <name type="personal" authority="viaf" valueURI="http://viaf.org/viaf/14777110">
   <displayForm>Jonathan Swift</displayForm>
   <role>
    <roleTerm type="code">cre</roleTerm>
   </role>
  </name>
  <language>
   <languageTerm type="code" authority="iso639-2b">eng</languageTerm>
  </language>
  <part>
   <extent unit="page">
    <start>25</start>
    <end>29</end>
   </extent>
  </part>
  <genre type="CCS">TextContent</genre>
 </relatedItem>
```

- The *type* attribute has the value *constituent*, because this related item is a constituent (a part) of the newspaper.

- The *ID* attribute may be any valid XML ID (it must begin with a character). By convention, the ID will begin with the letter *c* followed by a sequential number. (The docWorks processing flow seems to generate the *ID* attributes.) This attribute links the description to a <div> element in the METS logical structMap.

<titleInfo> The <title> is transcribed as it appears on the page, using standard cataloging rules.

<name>

- The <name> elements (there may none, or there may be more than one) are used to record the names of the people or organizations who are responsible for the constituent – usually, this is simply the author of the piece.

  The *<name>* is transcribed as it appears on the page and is encoded in the *<displayForm>* element. All *<name>* elements shall include a *<role>* element, which shall designate the generic role, *cre*, in the *<roleTerm>* subelement.

  When possible, encoders should supply a link to a name authority, preferably http://viaf.org.

<language>

- The language used in the text. If more than one language is used, there should be a <language> element for each.
- The <language> element shall contain the subelement *<languageTerm>*, a three-letter code derived from the ISO639-2 standard, found at http://www.loc.gov/standards/iso639-2/. The code form should be used.

<part> Contains a single <extent> element.

<extent unit=”page”>

- The <extent> records the page or pages on which the constituent appears:

  when the item appears on a single page encode the page number as a solitary *<start>* element.

  ```
  <extent unit="page">
   <start>3</start>
  </extent>
  ```

when the item appears on multiple sequential pages encode the first page in a *<start>* element and the last page in an *<end>* element.

```
<extent unit="page">
 <start>3</start>
 <end>4</end>
</extent>
```

when the item appears on non-sequential pages encode the pages as a series in a *<list>* element, as in

```
<extent unit="page">
 <list>3; 5</list>
</extent>
```

when the item appears on a mix of sequential and non-sequential pages `<extent unit="pa`

```
 <list>1-2; 5</list>
</extent>
```

For an article that starts on page 1, continues on page 2, and then skips to page 5. ORG-LIST-END-MARKER

<genre type=”CCS”> The *<genre type=”CCS”>* is determined from the docWorks configuration: for articles and other editorial content, it will be *TextContent*; for photographs, cartoons, and other illustrations, it will be *Illustration*; for advertisements, is will be *SponsoredAd*. Those <relatedItem type=”constituent”> elements that contain *other* constituents will have the genre *Section*.

Kinds of Constituents

TextContent These are the most common sorts of constituents: articles, notices, poems, stories, sports scores – all textual editorial content.

A TextContent constituent may contain *other constituents*: in particular, an article may contain illustrations.

- Examples
  - A basic article

```
<relatedItem type="constituent">
 <titleInfo>
  <title>  i</title>
 </titleInfo>
 <name type="personal">
  <displayForm>. .</displayForm>
  <role>
   <roleTerm type="code">cre</roleTerm>
  </role>
 </name>
 <language>
  <languageTerm type="code" authority="iso639-2b">rus</languageTerm>
 </language>
 <part>
  <extent unit="page">
   <start>2</start>
  </extent>
 </part>
 <genre type="CCS">TextContent</genre>
</relatedItem>
```

– An article with an embedded illustration

```
       <relatedItem type="constituent">
 <titleInfo>
  <title>  </title>
 </titleInfo>

 <language>
  <languageTerm type="code" authority="iso639-2b">rus</languageTerm>
 </language>
 <part>
  <extent unit="page">
   <start>4</start>
  </extent>
 </part>
 <genre type="CCS">TextContent</genre>

 <relatedItem type="constituent">
```

```
       <titleInfo>
        <title>. . </title>
       </titleInfo>
       <part>
        <extent unit="page">
         <start>4</start>
        </extent>
       </part>
       <genre type="CCS">Illustration</genre>
      </relatedItem>
     </relatedItem>
```

Illustration We use *Illustration* to refer to all kinds of graphic “art”: photographs, cartoons, charts, etc. Most illustrations in *Vozrozhdenie* (but not all) are accompanied by some sort of *caption*: a line or two of text, usually beneath the graphic, that describes the illustration, or names the creator of the illustration, or both.

For docWorks processing, the caption should be used as the <title>.

- Examples

```
<relatedItem type="constituent" ID="c2">
  <titleInfo>
   <title> </title>
  </titleInfo>
  <language>
   <languageTerm type="code" authority="iso639-2b">rus</languageTerm>
  </language>
  <part>
   <extent unit="page">
    <start>1</start>
   </extent>
  </part>
  <genre type="CCS">Illustration</genre>
 </relatedItem>
```

SponsoredAd Advertisements are an important and plentiful constituent of *Vozrozhdenie*. They usually appear in blocks or sections (see below); we do not attempt to assign them titles or creators.

- Examples

```
<relatedItem type="constituent" ID="c3">
 <titleInfo>
  <title>[Advertisement]</title>
 </titleInfo>
 <language>
  <languageTerm type="code" authority="iso639-2b">rus</languageTerm>
 </language>
 <part>
  <extent unit="page">
  <start>3</start>
 </extent>
 </part>
 <genre type="CCS">SponsoredAd</genre>
</relatedItem>
```

Section A section is a “composite constituent”: it contains other constituents. Advertising blocks are encoded as sections; named sections of the paper are encoded as sections, too.

- Examples

```
<relatedItem type="constituent">
 <titleInfo>
  <title>,   </title>
 </titleInfo>
 <part>
  <extent unit="page">
   <start>3</start>
  </extent>
 </part>
 <genre type="CCS">Section</genre>

 <relatedItem type="constituent">
  <titleInfo>
   <title>""</title>
  </titleInfo>
```

```
       <part>
        <extent unit="page">
         <start>3</start>
        </extent>
       </part>
       <genre type="CCS">TextContent</genre>
      </relatedItem>

      <relatedItem type="constituent">
       <titleInfo>
        <title>" "</title>
       </titleInfo>
       <part>
        <extent unit="page">
         <start>3</start>
        </extent>
       </part>
       <genre type="CCS">TextContent</genre>
      </relatedItem>
     </relatedItem>
```

4.3 ALTO Profile

For each page, an encoded representation of the layout and the machine-readable text on the page shall be provided, using the ALTO schema, version 2.0 or higher, with the following specifications, adopted from the NDNP:

- The text shall be encoded in the natural reading order of the language in which the text is written;

- Point size and font data to at least the word level shall be included;

- The ALTO file shall include bounding-box coordinates to at least the word level;

- Non-rectangular blocks shall not be used. Some illustrations may format as "tight" in the document.

4.4 Image Profiles

4.4.1 TIFF: Image Description

In general, Princeton University Library adheres to the standards elaborated by the Federal Agencies Digitization Guidelines Initiative (FADGI), whose Still Image Working Group produced a document entitled *Technical Guidelines for Digitizing Cultural Heritage Materials* in 2010. Archival images will be captured in 24-bit RGB and digitally rendered at varying resolutions to produce a uniform long dimension of 7200 pixels, then stored as uncompressed TIFF files with a large, non-proprietary color profile (Pro Photo RGB). The homogenization of the archival files to a long dimension of 7200 pixels allows us to produce uniform derivative images rapidly and estimate our storage needs more accurately.

4.4.2 JPEG2000: Image Description

Generative Image Derivatives Derived from the Master TIFF files with the following formula:

```
kdu_compress -i YOURINPUT.tif -o YOUROUTPUT.jp2 Creversible=yes -rate -,1,0.5,0.25 \
-jp2_space sRGB \
-double_buffering 10 \
-num_threads 4 \
-no_weights \
-quiet
```

Delivery Derivatives To generate a JP2000 using Kakadu, use the following recipe (taken from *The National Digital Newspaper Program (NDNP) Technical Guidelines for Applicants*):

```
kdu_compress -i YOURINPUT.tif -o YOUROUTPUT.jp2 -rate
1,0.84,0.7,0.6,0.5,0.4,0.35,0.3,0.25,0.21,0.18,0.15,0.125,0.1,0.088,0.0
75,0.0625,0.05,0.04419,0.03716,0.03125,0.025,0.0221,0.01858,0.015625
Clevels=6 Stiles={1024,1024} Corder=RLCP
```

5 Blue Mountain DocWorks Configuration

5.1 Containers

5.1.1 Editorial Content Container

Article Copy

5.1.2 PubInfo Container

5.1.3 SponsoredAd Container

6 General Markup Policy for Blue Mountain Magazines

6.1 Preliminary

For each title, Princeton will provide the following:

1. A "sample copy markup" (PDF): a full issue marked up according to the General Markup Policy;
2. A title-specific markup addendum clarifying unusual zoning and structural features; and
3. A marked up copy of any issue(s) deviating substantially from either the general or title-specific markup policies.

6.2 VerifyPageFrames (VPF)

Straighten and deskew all pages. When the input files include two-page spreads, do not split them into separate files unless instructed.

6.3 VerifyLayoutElements (VLE)

The VLE step is skipped for serials and books, so zoning is performed in VPN.

6.4 VerifyPageNumbers (VPN)

It is important to refer to any sample copy markup and any title-specific markup addenda in addition to this document, as the location and ways of identifying zones will vary by publication.

6.4.1 Zone Types to Tag

PubInfo Container

Page Number

Zone Name (in docWorks) PageNumber

Description Used when the publication includes page numbers

Location and Features Varies by publication; consult addenda.

Nameplate

Zone Name (docWorks) Nameplate

Description The space at the top of the magazine cover, usually containing the name (the logo), location, volume, issue, and sometimes taglines. Designs range from simple to intricate, depending on the design of the publication.

Location and Features The nameplate often contains the same information as the *masthead* but is a different layout element. Its purpose is to make the magazine identifiable when seen from a distance (as at a magazine stand, for example). The typeface is often a custom display type.

Not every publication design includes a recognizable nameplate.

Masthead

Zone Name (docWorks) Masthead

Description A section inside the magazine containing various pieces of publication information, such as the editor's names, contact information, and subscription rates.

Location and Features Mastheads vary in size and amount of information included. Some may take up an entire page. In newspapers, the masthead often occurs on the editorial page; in magazines, it may appear in various locations (consult title-specfic mark-up addenda).

Folio Line (Inside)

Zone Name InsideFolioLine

Description A folio line is an identification line, appearing (when it does appear) on each page of a newspaper or magazine. The front-page folio line is different from those on inside pages.

Location and Features Not all magazine layouts feature a folio line. It may run at the top or bottom of each page. It can also run as part of the logo on special pages or within the masthead. It often consists of the publication date in one corner, the name of the publication (centered), and the page number in the other corner. It is sometimes separated from the rest of the page by a cutoff rule.

Front Folio Line(s)

Zone Name FrontFolioLine

Description An identification line or lines on cover of the magazine often containing one or more of the following elements: the volume number (or the number of years the publication has been in print); the issue number; the place of publication; and date of publication.

Location and Features Usually joins the nameplate. Often separated from the logo by borders or cutoff rules.

Not every publication layout includes a recognizable front folio line. Many of the Blue Mountain designs omit it, or scatter the informaiton traditionally contained in a front folio line across the front page.

General Publication Information

Zone Name GenericPubInfo

Description A catch-all category for other "metadata" about the magazine not identifiable as nameplate, masthead, or folio lines.

Location and Features Highly variable, but usually found on outside or inside covers. Consult title-specific addenda.

Editorial Content Container

Head

Zone Name Headline

Description A phrase at the beginning of a magazine section, usually indicating the subject of the following copy text.

Location and Features A newspaper *headline* is a *kind* of head; magazines usually have many other kinds of heads. *Heads* are usually recognizable by weight, size, or position of type: they are often heavier or larger, and are often (though not always) centered above one or more columns of type with space before and after. A single magazine constituent may have several *levels* of heading: most have a top-level head (usually the title of the piece), to distinguish the piece as a whole from other magazine content, but many have typographically distinct *subheads* that subdivide the constituent itself into sections (see *Subhead* below).

Subhead

Zone Name Subheadline

Description A phrase demarcating a part of a constituent or section of a magazine. A constituent or section cannot have a subhead without a head.

Location and Features Subheads break up a chunk (a constituent, a section) into smaller units. Subheads are usually distinguished typographically from heads by smaller type sizes and different spacing.

Article Copy

Zone Name Textblock

Description A block of body text. *Article Copy* is a misnomer; *Editorial Copy* would be better. It is usually, though not always, divided into paragraphs; Blue Mountain zones tables and lists as Article Copy.

Location and Features May be preceded by a Head (though a Head is not mandatory). May be subdivided into smaller units through the use of rules and/or subheads.

Tagging Notes

- Blue Mountain zones tables and lists as Article Copy.
- Blue Mountain also tags tables of contents as Article Copy.

Byline

Zone Name Author

Description A zone containing one or more regions associated with the writer of an article, musical composition, letter, or other editorial content: usually the writer's name, but sometimes also the writer's position or other biographical information.

Location and Features The byline is usually located either at the beginning of the content, between the head and the first zone of copy, or at the end of the copy text. Bylines can be difficult to distinguish. They are sometimes set in italic or boldfaced type or in all-caps; they are sometimes a full name (first name, last name), sometimes a last name only, or one or two initials, or a name with a term of address in the language of the magazine (Mr., Dr., Madame, M., Herr, etc.)

Art

Zone Name Illustration

Description A drawing, photograph, a reproduction of a print or a painting; a graphic poem. Sometimes has a caption above, below, or to the side of it.

Location and Features May occur anywhere; may occupy a single column, several columns, or be set outside the column grid of the page; or may take up a full page or a multi-page spread.

The Avant-garde saw the invention of "graphical text" like concrete poetry, shape poetry, Futurist "words in liberty," etc. While these are in fact hybrid forms, both graphic and text, Blue Mountain zones them as art.

Caption

Zone Name Caption

Description One or more textual blocks associated with an Illustration.

Location and Features Captions may appear above, below, or (occasionally) to the side of the illustration to which they belong. They may contain several lines and/or zones of information: the identity of the subject; the name of the artist/photographer/poet; the title of the artwork being printed or reproduced. Consult title-specific addenda.

Music

Zone Name Music

Description A zone of musical notation (in Western music, most often one or more staves, perhaps with clefs and key signatures, and notes; sometimes heavily annotated with text.)

Location and Features Size and extent varies. May be a small musical example occupying a few lines of a column, or a multi-page musical compostion. Some zones of music may be accompanied by heads or captions.

Footnote

Zone Name Footnote

Description A zone of type, usually at the foot of the page, serving as an annotation to some text on the page.

NOTE This zone is under consideration.

SponsoredAd Container

Sponsored Ad (Advertisement Copy)

Zone Name Advertisement

Description An often-heterogenous zone of mixed typography, graphics, and layout whose function is to draw attention to some event, product, or service. *Sponsored* advertisements are those advertisements that appear in the magazine because an outside agency (a company, a theater, a publisher) has paid for it to appear.

Location and Description Ads may appear anywhere, but most often they appear in groups at the end of a magazine issue. They are often boxed. They may be any size – from a few inches to a partial or full page.

6.4.2 Zone Types to Ignore

Rules Do not zone horizontal and vertical rules – the (usually thin) lines used to separate regions of the page.

Ornamental Artwork Typographer's ornaments, ornamental borders, anonymous graphics at the heads of sections. (Consult title-specific addenda.)

6.5 VerifyPageHierarchy (VPH)

Add the following metadata, if available:

- Volume Number
- Issue Number
- Issue Date
- Publication Title

This information will always be supplied by Princeton in the form of *preliminary issue-level MODS* (see below).

6.6 VerifyHiearchy (VH)

Identify all components as either PubInfo, EditorialContent, or SponsoredAds.

- Classify all zones in PubInfo or Editorial Content containers manually. (All ad zones will be grouped automatically into SponsoredAd containers.)
- Composition of magazine constituents is highly variable. Some general guidelines:
  - A constituent very often is composed of a head, a byline, and one or more sections of body copy. Look for a typographically distinct head to begin the constituent; a byline will sometimes follow immediately after the head, but it may also appear at the very end.
  - Use the head as the constituent's title.

  - Many of the Blue Mountain titles are art magazines; the illustrations often have heads, captions, and bylines. If a head is present, treat it as the title.

6.7 OCR Correction

- Correct all heads, subheads, bylines, and captions.
- Delete empty zones.
- Ignore hand-written text blocks.