Documentation

Narrative

Transcriptional Policies

Characters

When transcribing documents, efforts should be made to record all characters as they appear on the page, whenever possible. This includes ligatures for which there are unicode characters (æ, œ), characters with accent marks (aigu, grave, circumflex, umlaut, etc.), and other special characters (manicules, daggers, etc.). Note that we do not record characters for which there is no unicode reference. So, for example, c-t ligatures are recorded as two separate characters ("ct"). Additionally, we regularize all dash lenghts to either hyphen, en-dash, em-dash, or superdash. See also the Unicode cheat sheet[link here when you've written that]

Page, line and column breaks

We record all page breaks using the <pb/> element and all line breaks using the <lb/> element. If a word crosses a line (with a hyphen), the <lb/> element should be given the @break attribute with the value of "no". This indicates that the line break does not constitute a break in the word.

<cb/> should be used in instances of column breaks.

Errors and Corrections

For *typographical* errors that occur in the text, we use <sic> to record the spelling that is erroneous. We then correct the error using <corr>. Both <sic> and <corr> should be nested in <choice>. Please note that we *do not* use <sic> and <corr> for variant spellings. Please check the OED or some other resource to see if the spelling you encounter was used in texts that are contemporary to the one you are encoding.

In cases of *semantic* errors, you should use <orig> (for the original word or phrase) and <reg> (for the corrected word or phrase) within <choice>. These types of errors could include the mixing of gender pronouns, accidentally referring to one character by another's name, or (in cases where it is *obviously* a mistake) using the wrong word.

Typographical and semantic errors can be distinguished by the person probably responsible for the error: typographical errors are those that were probably made by the typesetter ("nnder" for "under," "taht" for "that"), and semantic errors were probably made by the author. If you are unsure or encounter an edge case, err on the side of using <orig> and <reg>.

All corrections should be made at the *word or phrase level*. So you should never have a choice that sits inside of a word.

Abbreviations

Abbreviated words should be recorded using the <abbr> tag. This should be nested within <choice> with the expanded word in <expan> *only* if the abbreviation is non-standard. So, for example, if you see the name “Mr. Jones” the “Mr.” *does not* need to be expanded to “Mister.” However, if you see the name “Alex’r” it *should* be expanded to “Alexander.”

Transcriptions of quoted material and speech

When recording quoted material, <quote> element; this includes material quoted from other texts, quotes from within the current text, and common phrases that are renditionally marked as such. This records the *semantic* information that “this is a quote.” You also must record the renditional or typographical cues that are present in the text. Transcribe quote marks *outside* the <said> or <quote> element, whenever possible (some instances in which this may not be possible are instances of quoted poetry). Remember that XML editors generally default to straight quotes when typed, so you must enter the unicode for either left- or right-facing curly quotes (U+201C or U+201D for double quotes and U+2018 or U+2019 for single quotes).

Additionally, quoted material may be marked by italicization or some other renditional marker. In these cases you use the @rend attribute on either <quote> to specify what the rendition is.

We do not record speech with the TEI's <said> element. Instead we simply record the speech with the appropriate quote marks, or with <seg>, and the appropriate @style attribute (if speech is marked by rendition).

Unclear or illegible text

The AAS has three different ways of transcribing illegible text: <gap/>, <supplied/> and <unclear/>.

If you cannot read the text and there is no way of guessing what the text says or supplementing your reading with another edition, you should use the <gap/> element. The <gap/> element requires the @extent attribute, which contains a short description of the extent of the damage (i.e. 1-2 characters).

<supplied> is for bits of text that are illegible in the source you are transcribing from, but whose content can be extrapolated from context, or supplied from another edition. [detail alt sourceDesc here?]

<unclear> is for texts whose content you can somewhat make out, but cannot be fully sure of the reading. This element is mostly used to state that the editor or transcriber is unsure of the reading they have provided.

Rendition

For the Just Teach One-EAAP project, we use two different attributes to record rendition: @style and @rend.

@style has snippets of CSS as its value. We use it to record casing, slant, alignment. *We do not use it to record font size, color, font, or other pieces of rendition*.

The syntax of the value of @style looks like this:

```
style="rendition:description ; second-rendition:description"
```

And the vocabulary for CSS is as follows:

- case
    - property: text-transform[1]
    - values: uppercase (for allcaps); none (for text to render as-is)
- slant
    - property: font-style
    - values: italic (for italicization); normal (for upright)
- alignment
    - property: text-align
    - values: left, center, right, justify (self-explanatory)

We also occasionally use the “white-space:nowrap” value on @style to undo elements that are normally defaulted to break. This should only ever be the case with the <head> element. So, for example, if you have a heading and a subheading on the same line that need to be recorded in separate <head> elements, you would use “white-space:nowrap” on the *second* <head> to indicate that there should not be a line break between the two headings.

The other attribute we use to record rendition is the @rend attribute. We only use this to record fonts, since we are describing fonts that do not necessarily have modern correlates. We only describe the broadest categories of fonts: blackletter, roman, and script.

“Blackletter” is the a font that is frequently used in early modern texts. For the purposes of this project, we will be using “blackletter” to describe fonts that look like the fonts you would see on mastheads of newspapers. “Roman” font is every other kind of typeface. “Script” is used for printed representations of signatures (*not* handwritten additions). For handwritten additions, you would use the <add> element. An example of the rend="script" are the signatures that are

[1] Note that this is indicating that the text should be *transformed* from its current state. This means, if a heading is in all capitals, you do not have to transcribe in allcaps in order to get the desired effect! When transcribing, please *use standard casing* for all transcriptions. This means that titles should typically have an initial capital with following lowercase letters. Roman numerals should usually be transcribed as capitals.

printed at the end of divisions in “Autographs for Freedom,” but you may find printed signatures in other places.

What to encode

For the JTO:EAAP project, we are primarily interested in recording:

1. Accurate Transcriptions
2. Selected rendition (see the rendition section)
3. Structural and logical divisions of the text (chapters, paragraphs, nested subdivisions, poetic lines and groups of lines, etc)
4. Some selected phrase level features
5. Physical descriptions of the printed text

Phrase level features that we care about include dates, people’s names, the names of places, the names of organizations, quotes, rhetorical emphasis (only if marked with rendition), and several other features that are only

Elements

Division-level elements

<back>

This element encloses all backmatter, such as epilogues, indices, errata lists, and any other features that occur after the main body of the work.

A marks a logical division of the text (such as a chapter, a section, or a book). elements can be nested inside of each other.

<floatingText>

The TEI does not allow elements to float within a bunch of <p> elements. However, sometimes there are divisions that “float” in a text without having other correlating divisions surrounding it. So the TEI came up with <floatingText> to address this problem.

<floatingText> is usually used for divisions like letters or narratives that have been nested within the larger division (like a chapter). So, for example, if the heroine of our text receives a letter in

the middle of a chapter, it is reproduced in full, and the larger narrative resumes immediately after, this would be <floatingText>.

The template for <floatingText> is as follows:

```
… text from the larger division</p>
<floatingText>
      <body>
            <head>A Nested Narrative</head>
            <p>text</p>
      </body>
</floatingText>
```

Note that <floatingText> must always have <body>! Beyond that, <floatingText> can have any of the features you would normally put in a division: headers, footers, even nested divisions!

<front>

This element records front matter, such as prefaces, tables of contents, title pages, editorial notes, and other features that occur before the main body of the work.

<group>

This element is used to group together a series of related texts. It contains multiple <text> elements. It is usually used to group together serialized novels that appear in newspapers.

<jto:hyperDiv>

This element is used to record hypertextual features, such as footnotes and marginal notes. It should always come within <text>, and before <front>. At this time, <jto:hyperDiv> can only contain <notes> (which itself contains the <note> element.

<text>

This element contains <hyperDiv>, <front>, <body>, and <back>. It may be contained by <group>, if more than one text needs to be considered together (this is only usually the case with serialized newspapers).

Chunk-level elements

<ab>

(anonymous block) This element is used to record paragraph-like chunks of text that are not technically paragraphs. You will most commonly find them as captions on figures and drawings.

Anonymous blocks are not used for lines or stanzas of poetry (see <l> or <lg>). They are actually pretty rare, and you will usually only use them for captions.

<l>

This element contains poetic lines, and is encapsulated by an <lg> (or a line group)

<lg>

(line group) This element groups together lines—like in a poem. We only use <lg> to encode the *outermost* group of lines. Any stanzas or other groupings are not recorded, unless they are indicated by whitespace (i.e. an extra line of space in between poetic lines).

<p>

This element is used for marking paragraphs.

Phrase-level elements

<bibl>

This element is used to record information about texts that are referenced within the text you are encoding. Use it to record: book and poem titles, *structured* information about chapters, page numbers or line numbers, the names of books of the bible (along with verse and chapter numbers, if applicable).

When using <bibl>, if recording an author's name, you will need to use <author> around <persName>. However, you *should not* use <bibl> simply to record an author's name. So for example, if you were to encounter the phrase "Shakespeare's works are often performed at the local theater," you *would not* use the <author> tag or the <bibl>. <persName> will suffice in those instances.

<date>

This element is used to record dates. We record dates when the year is provided, or can be reliably established from context. Dates without a year (e.g. April 14th) are not recorded.

<date> should always have the @when attribute, which follows a YYYY-MM-DD, YYYY-MM, or YYYY pattern.

<emph>

<emph> records rhetorical emphasis, such as: "I simply *won't* do what you're asking me to" or "He *said* he would be here by 8 o'clock." It should always have some sort of rendition indicated

with @style (most usually italic). **Only use <emph> for instances in which *rhetorical* emphasis is marked.** If rendition is being used for some other reason, record it with one of the elements listed in the phrase-level elements section or the “elements used to anchor rendition” section.

<persName>

<persName> is used to record the names of any who is both personified and has a name. It is used to encode personal names of *individuals* and *family names.* This means, if the text refers to “the Smiths,” you should encode “Smiths” with <persName>. You should also use <persName> to record names that are used as proper adjectives, as long as the name isn’t changed in order to achieve adjectival form. So, for example, if you encountered “the Hawthorne house,” you would encode “Hawthorne” with <persName>. However, if you saw the phrase “Dickensian name,” you would not encode “Dickensian” with <persName>.

<placeName>

<placeName> should be used to encode the names of places, whether geographical, political, or local. So anything from a street name to the name of a river to the name of a country would all be encoded with <placeName>. When you encounter a place that is also an organization, look to context to decide if it should be encoded with <placeName> or <orgName>. If it is being referred to as a location, you should probably use <placeName> (e.g. “We visited Harvard University”). However, if the reference is primarily unrelated to physical location, use <orgName> (e.g. “He is an alumnus of Harvard University”).

<orgName>

<orgName> is used to record references to specific organizations of various sizes. These could be large international organizations with many local chapters, like the Catholic Church, or they could be smaller, local organizations like the Rochester Ladies’ Antislavery Society. See the entry on <placeName> for information on the distinction between <placeName> and <orgName>

<quote>

<quote> includes material quoted from other texts, quotes from within the current text, and common phrases that are renditionally marked as such. If a phrase is specifically referred to by the author as a common phrase, you should mark it with quote, even if it is not a phrase you are personally familiar with. We *do not* record transcribed speech with the TEI <said> element or with <quote>. For more information on how to record the quote marks and rendition for quotes and speech, please see the *Transcriptions of quoted material and speech* section.

Elements used to anchor rendition

Certain elements will only be used as a placeholder for rendition. For example, you should use the **<hi>** element in instances where there is no *rhetorical* emphasis (<emph>), but you still feel as though you ought to record the rendition.

You should use the **<term>** element if an element is renditionally distinct because it is being defined, or because it is distinguished as jargon or technical language.

You should use **<soCalled>** if the rendition is used to indicate sarcasm or ironic distance.

<add>

The <add> element is used to record handwritten additions to the text. <add> may use attributes to record the place of the addition (“margin,” “inline” (for things that are above the printed lines in the text” or “overwritten” (if it is written over the printed text)), and the person who added the addition (if known). So, for example, if we knew that someone named John Smith wrote a marginal note, we would record it as:

```
<add place="margin" hand="#john-smith">John was here.</add>
```

The @hand attribute points to an element in the <teiHeader> called <handNote> (which is recorded inside <handHotes> inside the larger <profileDesc>. In this case, <handNote> would look something like this:

```
<profileDesc>
     <handNotes>
          <handNote xml:id="john-smith">This hand is John
          Smith’s</handNote>
     </handNotes>
</profileDesc>
```

You can record multiple <handNote> elements within <handNote> if there is more than one person’s handwriting in the text.

If there is no information on who wrote in the text, don’t worry about recording @hand or the <handNote> element. You can read more about the <handNote> element in the metadata documentation.

<del>

The <del> element is used to record passages that have been crossed out. If something has been obscured by a human hand, surround it with the <del> element. You can use it in

conjunction with <supplied>, <unclear>, or <gap>, if the crossout resulted in obscuring of the text.

Empty Elements

<pb>

<pb/> is used to mark page breaks in a text. You should place it wherever there is a break across a page. If there are multiple breaks between two sections you are trying to encode (especially if you are encoding a newspaper, where there will often be other stories between two sections of text), just record one <pb/>.

<cb>

<cb/> is used to record column breaks in a text.

<lb>

<lb/> is used to record line breaks. You should use the @break attribute to indicate when the line break *does not constitute a break in a word.* So, for example, if you encounter:

> The woman was especi-
> ally tired that day.

You would mark that line break with break=”no”. However, you *should not* use this attribute when hyphenated compound words cross lines. So, for example, “self-<lb/>assured” would *not* have the break attribute.

The default value of @break is assumed to be “yes,” so you don’t need to use this attribute unless there is a word being divided by a line break.

<anchor>

This attribute is used for anchoring notes to the text they annotate. <anchor/> will usually (although not always) be placed next to an asterisk, superscript, or some other character marking the annotation.

<anchor/> has two attributes: @xml:id and @corresp.

@xml:id is a unique identifier that allows the note to point to the anchor
@corresp is the attribute that points to the note. It should always have a “#” to indicate that it is pointing to something. For more information on notes and pointing, see the “Notes” section below.

Special Cases

Title Pages

The following is a template for a title page:

```
<titlePage type="main"> <!-- values on @type also include “sub” and “desc” -->
        <titlePart>Title
          <lb/>of
          <lb/>Work
        </titlePart>
        <docRole type="author"> <!-- values on @type also include “editor” “publisher” or “printer” -->
          by
          <persName>Name of Author here</persName>.
        </docRole>
        <epigraph>
          <quote>
            <p>Epigraphs within <quote> are also allowed in title pages
            </p>
          </quote>
        </epigraph>
        <docImprint> <!-- This section records information about the publication and printing of the document -->
          <pubPlace><placeName>Auburn</placeName>:</pubPlace>
          <publisher>
            <orgName><persName>Alden</persName>, <persName>Beardsley</persName> &amp;amp; co.</orgName>
          </publisher>
          <pubPlace><placeName>Rochester</placeName>:</pubPlace>
          <publisher>
            <orgName><persName>Wanzer</persName>, <persName>Beardsley</persName> &amp;amp;
co.</orgName>
          </publisher>
          <docDate><date when="1854">1854</date></docDate>
        </docImprint>
      </titlePage>
```

All of the elements listed in this template are allowed within the <titlePage>, but they are not all required, and they do not necessarily need to be listed in the order that they appear in the template.

Occasionally, you may find imprimaturs on the title page (see section on “imprimatur”), but you should only record these within <titlePage> if they are recorded on the same face of the same page as the rest of the title-page information. If it is on the verso side, or some other page, record colophon on its own.

Tables of Contents

Tables of contents are encoded within a division with the type value of “contents”.

They are then further encoded within a <list>, with <item> elements representing each entry in the contents table. Titles should be encoded within <rs> (with no further encoding happening within the <rs> element, with the exception of <hi> for rendition). <ref> should be used to encode the page numbers. Occasionally, there will be more information, like author names, represented in the table of contents. Encode these with the appropriate element.

Usually, tables of contents will have list headings for each component of the table of contents. use the <head> element to represent these.

```
<list>
        <head>Subject.</head> <head rend="break(no)">Author</head> <head rend="break(no)">Page</head>
        <item><rs>Introduction (The Colored People's
          <lb/>Industrial College</rs>) <persName style="font-style:italic">Prof C.L. Reason</persName>
<ref>11</ref></item>
        <item><rs>Massacre at Blount's Fort</rs> <persName style="font-style:italic">Hon. J.R.
Giddings</persName> <ref>14</ref></item>
        <item><rs>The Fugitive Slave Act</rs> <persName style="font-style:italic">Hon. Wm. Jay</persName>
<ref>27</ref></item>
        <item><rs>The Size of Souls</rs> <persName style="font-style:italic">Antoinette L. Brown</persName>
<ref>41</ref></item>
        <item><rs>Vincent Ogé</rs> <persName style="font-style:italic">George B. Vashon</persName>
<ref></ref>44</item>
        <item><rs>The Law of Liberty</rs> <persName style="font-style:italic">Rev. Dr. Wm. Marsh</persName>
<ref>61</ref></item>
        <item><rs>Swiftness of Time in God</rs> <persName style="font-style:italic">Theodore Parker</persName>
<ref>63</ref></item>
        <item><rs>Visit of a Fugitive Slave to the Grave
          <lb/>of Wilberforce</rs> <persName style="font-style:italic">Wm. Wells Brown</persName>
<ref>70</ref></item>
</list>
```

Imprimatur

An imprimatur is a statement about the authorization of the work. In the case of our texts, these are usually formal statements about the copyright status of the work. They usually appear on the verso of the title page. They usually begin with: “Entered according to an Act of Congress”

If it is on the same page as the title page, put <imprimatur> within the <titlePage> element. If it is on the verso, or some other page, encode <imprimatur on its own.

Letters

Letters should be recorded with . They have optional <opener> and <closer> sections, which can contain <salute> (for salutations), <signed> (for signatures), and <dateline> (for information about the place and time when the letter was written). In epistolary novels, you will always use .

However, if you encounter a letter that is nested within a larger chapter, or some other division, it is likely that you will use <floatingText>. This element allows you to place a division-like chunk of text inside a bunch of paragraphs. For more information, see the entry on <floatingText>.

Notes

You will occasionally encounter notes while encoding your text. These need to be linked to the annotated text through xml linking. For more information on this topic, see the Women Writers Project Internal Documentation.

We use linking on notes in a very similar manner to the WWP. We assign both the note and the annotated text an @xml:id. This must be a unique number within your document. I recommend using “n001, n002, n003” and so on for the notes, and “a001, a002, a003” for the annotations. Note that xml:ids *cannot start with numeric characters*.

You then connect them to each other using @target and @corresp. The note element is assigned @target which points to the annotation. So if the annotation had an xml:id of “a001,” the @target attribute on <note> would be target="#a001". The sharp sign indicates pointing.

Annotated text will usually be indicated by an * (asterisk), † (dagger), superscript, or some other character. In these cases, put a <seg> around the annotation marker, and apply the xml:id and @corresp attribute (which points to the note) to the <seg> element. In cases where the annotation is not marked, use the empty <anchor/> element.

Serialized Newspapers

When encoding serialized newspapers, you should always use the <group> element inside of <text>. This allows you to create multiple <text> elements within a single document. Each <text> element within <group> allows for the usual contents of <jto:hyperDiv>, <front>, <body>, and <back>.

Each <text> element within group should correspond to a <sourceDesc>, which contains information about that specific issue and volume of the newspaper. For more information on setting up serialized texts, see the text-serial-template xml file, and the metadata encoding documentation.