

## CSV Guidelines: Mixed Matter

Welcome to the Library of Digital Latin Texts project. These guidelines are intended to give you a general overview of how the CSVs you produce will be used, and to provide you with the specific encoding procedures for the CSV version of the critical apparatus for a mixed matter text.

### General overview

The goal of our project is to produce an XML-encoded version of a Latin critical edition. When finished, these digital critical editions will be open source and freely available to the public. XML is a markup language which encodes information about a text within the text itself. XML encoding by hand is tedious and time-consuming, so we use Python scripts to automate the process. These scripts take as input a plain text file of the editor's text and a CSV (comma separated values) file with critical apparatus entries. Your job is to produce this CSV file (basically just a spreadsheet) from a printed critical edition. Because you are creating the input for an automated script, it is important to follow the punctuation and formatting guidelines presented in this document as exactly as possible.

### CSV Overview

A CSV (comma separated values) is, essentially, a very basic spreadsheet. It is usually easiest to edit your CSV in spreadsheet software and then export it as a text CSV. You should use software other than Microsoft Excel, because Excel will insert some Microsoft-specific markup that is hard to process. I like LibreOffice (free and available here: <https://www.libreoffice.org/download/download/>) but you can use whatever works for you.

### Specifications for the CSV file

Your work should be saved as a text CSV file, with UTF-8 encoding. The column delimiter should be the comma (,), and the text delimiter should be double quotes ("). Both of these are handled in the document settings, so you won't have to type them. If you're using LibreOffice, the correct document settings look like this:

**Export Text File**

**Field Options**

Character set: Unicode (UTF-8)

Field delimiter: ,

Text delimiter: "

☒ Save cell content as shown

☐ Save cell formulas instead of calculated values

☐ Quote all text cells

☐ Fixed column width

Help OK Cancel

## LDLT CSV Format

The required CSV format for drama has a total of 26 columns. The column headings are as follows: Section, Paragraph, Sentence, Poem, Line, Lemma, Lemma\_Witnesses, Lemma\_Sources, Lemma\_Annotations, General\_Comment, Reading\_1, Reading\_1\_Witnesses, Reading\_1\_Sources, Reading\_1\_Annotations, Reading\_2, Reading\_2\_Witnesses, Reading\_2\_Sources, Reading\_2\_Annotations, Reading\_3, Reading\_3\_Witnesses, Reading\_3\_Sources, Reading\_3\_Annotations, Reading\_4, Reading\_4\_Witnesses, Reading\_4\_Sources, Reading\_4\_Annotations.

It is important that the columns have these names, and absolutely essential that they appear in this exact order. The number of readings is flexible, so if you have more than 4 readings for any lemma, you can just add a Reading\_5 (or as many additional reading columns as necessary). If you add additional columns, you must add all 4 columns (Reading\_X, Reading\_X\_Witnesses, Reading\_X\_Sources, Reading\_X\_Annotations) for each additional reading, regardless of whether they are used.

## Column-by-column formatting guidelines

This section will discuss each column of the drama CSV in detail, including what goes in that column and the formatting and punctuation guidelines for each type of data.

### **Column 1: Section**

Many texts are separated into chapters, books, or sections. Put the section numbers in this column. If the work does not have section divisions, you can just put 1 in every row of this column.

### **Column 2: Paragraph**

When the lemma is in a prose paragraph, put the paragraph number in this column. If the lemma is in a poem or song, leave this column blank.

### **Column 3: Sentence**

When the lemma is in a prose paragraph, put the sentence number in this column. If the lemma is in a poem or song, leave this column blank.

### **Column 4: Poem**

When the lemma is in a poem or other verse chunk, put the number of the poem within the work in this column. If the lemma is in a prose paragraph, leave this column blank.

### **Column 5: Line**

When the lemma is in a poem or other verse chunk, put the line number within the poem in this column. If the lemma is in a prose paragraph, leave this column blank. The script can handle transposed line numbers (see .txt guidelines), and your line numbers should match those in your plain text file.

### **Column 6: Lemma**

The lemma for the critical apparatus entry goes here. This is reading that the editor believes to be correct, and therefore the reading that they print in their text. It is important that the text you enter in the lemma column exactly matches the lemma as it appears in the base text, so that the script can find and replace it. For example, if the base text says “iuuenis”, your lemma entry should say “iuuenis” and not “iuvenis”. There are 4 special cases you are likely to see with lemmas:

## Lacunae

In the case that there is a lacuna in the lemma (i.e. the editor believes text has been lost, but other witnesses or editors disagree), you should put “” where the lacuna is in the lemma. This tag can be surrounded with text on either side, if necessary. For example: this is my <gap reason=“lost”/> lemma text

## Multiple occurrences of lemma text

In the case where the full lemma text occurs multiple times in the sentence/section, it is necessary to explicitly specify to which occurrence the apparatus entry applies. The markup for this is: “word(1)” for the first occurrence, “word(2)” for the second, and so on. Note that there is no space between the word and the parentheses. You are most likely to see this for conjunctions and prepositions like et, sed, in, ab, ad, etc.

## Editorial additions

In cases where the editor has added a word or words, they should be surrounded with angle brackets, <>. For example: these are some <added> words

## Uncertain Speakers

Some texts (or parts of texts) have speakers. In many cases, there is uncertainty about who the speaker of a certain line is or where the speaker changes during the line. If there is uncertainty about the speaker for the whole line, the Lemma column should contain the speaker that the editor believes to be correct, in parentheses: “(Gaius)”. If there is uncertainty about where the speaker changes, the Lemma column should contain “word (Speaker) other words”. The reading columns can then contain variations with other words, a speaker change in a different place, or a different speaker, as appropriate.

## Column 7: Lemma\_\_Witnesses

Witnesses are manuscripts which contain the reading in question (in this case, the lemma). Each witness is designated with a siglum, which is usually a capital letter, but can also be designated with a Greek letter. Manuscript sigla should be encoded separated by spaces, e.g. “A B C”. There are some special cases for encoding witnesses, which are as follows:

## Corrected or changed manuscripts

In many cases, manuscripts have been changed or corrected many times. Our script currently handles 5 types of manuscript correction: *ante correctione* (ac), *post correctionem* (pc), *supra lineam* (spl), *sub lineam* (sbl), *in margin* (inmg).

In these five cases, the abbreviation should go in parentheses immediately after the siglum, with no space in between. For example, if the lemma comes from the corrected version of manuscript A and the pre-correction version of manuscript B, then you should enter “A(pc) B(ac)” column 7. The script can only handle one of these abbreviations per siglum, so choose the most specific one for the correction in question, and if necessary, put the Latin (not the abbreviation) for the others in column 9, Lemma\_\_Annotations.

### **‘mr’ (Manus Recentior)**

‘mr’ stands for manus recentior, and means a later hand writing in the same manuscript. It should be encoded with the siglum immediately followed by ‘mr’. For example, a later hand in manuscript A would be encoded “Amr”. Please note that this is different from corrected manuscripts (above) because there are no parentheses. Also note that there is no whitespace between the siglum and the ‘mr’.

### **Multiple hands in a manuscript**

For multiple hands in the same manuscript, use the manuscript siglum, immediately followed by “h” and then the number. For example, use “Ah1” for the first hand in manuscript A, or “Bh2” for the second hand in manuscript B. Note that there are no spaces between the siglum, the “h”, and the number.

### **Multiple manuscripts from one collection**

For numbered manuscripts from the same collection, use the siglum, immediately followed by “ms” and then the number. For example, use “Ams1” for the first manuscript in collection A, or “Bms2” for the second manuscript in B. Note that there are no spaces between the siglum, the “ms”, and the number.

### **Ed. Pr.**

Occasionally, you will see “ed. pr.” among the sigla for a lemma. This means the first printed edition, and is considered a witness, not a source. It should be encoded “ed.pr.”, with no space in between.

### **Other annotations**

In general, other types of subscripts and superscripts should be entered in line with the siglum, not separated from the siglum by whitespace.

### **Column 8: Lemma\_Sources**

Sources are modern editors who believe the lemma is the correct reading. Enter their last names, separated by one space, e.g. “Name AnotherName”. Sometimes, you will see the phrase “Name teste AnotherName”. This means that AnotherName agrees with Name’s reading. In this case, “AnotherName” goes in Column 8, and “Name teste” should go in Column 9, Lemma\_Annotations.

### **Column 9: Lemma\_Annotations**

These are the editor’s comments about the lemma. Usually, they are references to similar wording in other works. Most of the time, you can just copy and paste these notes into column 9. There are 2 special cases with notes:

#### **Alternate spellings**

Occasionally, the editor will want to note an alternate spelling of the lemma in one manuscript, without considering it as a totally separate reading. In this case, the critical apparatus would say something like “hae [heae M] MUTV”. Here, “hae” is the lemma, “M U T V” are the witnesses, and “[heae M]” is a note. The note should be encoded “[M heae]”. It is important that a note which contains an alternate spelling is in square brackets, and that the siglum of the witness is immediately after the opening bracket and is followed by a space.

#### **Multiple Notes**

If there are multiple notes on a lemma, they should be separated with a forward slash preceded and followed by spaces, i.e. “this is a note / another note”. If an alternate spelling note is necessary, it should be separated from other notes using this markup, e.g.: this is a note / [X spelling]

### **Column 10: General Comments**

This column contains general comments about the critical apparatus entry. These notes can be copy-pasted directly into the CSV. There is no special markup for this column.

### **Columns 11, 15, 19, 23: Reading\_X**

In general, the guidelines for the reading columns are the same as the guidelines for the corresponding lemma columns. Therefore, these sections will only cover differences and additions to the lemma guidelines.

Readings are encoded in the same way as lemmas, and the special cases in described for Column 6 still apply. There are three additional special cases which pertain to readings:

### **Omissions**

An omission means that a word or words are completely omitted from a witness. In this case, all of the omitted words should be the lemma, and the reading should simply be “om.” with no spaces.

### **Editorial Deletions**

An editorial deletion means that text is present in a manuscript, but an editor believes it shouldn’t be. This is usually denoted with “secl.” in the critical apparatus. An editorial deletion is encoded by enclosing putting “secl.” in the reading column.

### **Alternate spellings**

Sometimes, an editor will include two alternate spellings of the word in one reading. Such a reading will look like: “religion(e/i)s”. In this case, it can be encoded in the Reading\_X column as is, with the parentheses and forward slash included. The editor may also provide a note about which witness has which spelling. In this case, they should be encoded according to the instructions for alternate spellings under Column 9: Lemma\_Annotations.

### **Columns 12, 16, 20, 24: Reading\_X\_Witnesses**

These are encoded exactly the same way as the lemma witnesses are. See Column 7 above.

### **Columns 13, 17, 21, 25: Reading\_X\_Sources.**

Again, reading sources are encoded exactly the same way as the lemma sources are. See Column 8 above.

### **Columns 14, 18, 22, 26: Reading\_X\_Annotations**

The same encoding guidelines apply as in Column 9 above. The only additional thing is notes that go before readings. Usually, these are just “an” or “vel” and are used to signify the editor’s tentative conjecture. These should be the first note listed in the Reading\_X\_Annotations column, and should be separated

from other notes with " / " as described above. Other than that, no special markup is required.

## Helpful tips

1. I find it easiest to work in a spreadsheet editor, and to save my work in a spreadsheet format until the CSV is finished. This allows me to highlight in yellow the entries or cells I am unsure about, so I know what I need to ask questions about. Once my questions are answered and I am confident that everything is correct, I save it as a text CSV, with the encoding and setup described under “Specifications for the CSV file”.
2. Try to avoid using semi-colons ; in the annotation columns, as this seems to cause notes to split across multiple columns. I’m still working on this issue.
3. If you have questions, I am happy to help. I’ve been doing this encoding work for around a year now, so I can probably answer a lot of your questions. You can email me at [katy.felkner@ou.edu](mailto:katy.felkner@ou.edu).