14 June 2020
Current Best Advice on Hand-Cleaning XML

Note that new files are now provided in textbase/xml. I suggest you replenish your supply there, rather than using any you may have in your own directories. The new files should have many fewer "matches" presenting overlapping passages of the text. In them, if a file contains a search word multiple times within the space of 3,000 characters, in most cases it should only produce one match for that search word in that passage. This should save us having to read the same text many times over for the same search word. It will not change the number of matches from passages that contain muliple search words.

1. Immediately on starting work on a new file, assign it a new name: the old name in .xml, but ending in "_clean.xml". Put a copy of this file in the "textbase/xml/xml_clean" directory on Box, so that I'll know you're working on it and therefore won't upload any other versions that would confuse us all. You'll remove or replace this placeholder version later.

2. Insert some white space into the file to make it easier to read. Search for </match> and replace it with </match> plus a few <return> or <newline> characters—possibly represented as \n, probably best copied and pasted from the file itself. This will put a few lines of white space between the end of each retrieved passage and the start of the next.

3. Read through the whole xml file, and decide if there is anything worth keeping. If not, put the file in textbase/xml/done, delete the "_clean.xml" placeholder version from textbase/xml/xml_clean, and go on to the next file.

4. If you decide to keep any passages, fill in the standardizable XML tags for the whole file, using search-and-replace: replace <source_author></source_author> with <source_author>[author's name or names]</source_author>.
Replace <source_title></source_title> with <source_title>[book title]</source_title.>
Do the same for the "cleaner_name" and "date_added" tags.

5. For the passages you decide to keep, fill in as much of the missing data as you can, between the relevant XML tags—*we need the data for <source_page_number> and <place_name> (from the standardized list, HDW Paris Project/tools/place_index.csv). You'll also need to create the unique identifier and fill it in as follows: <uid>[base filename including bpt6k number]_x</uid> where x is a number you assign in order, starting with 1 for the first match in the XML that you decide to keep.* ~~Also, your name or initials, and the date: YYYY/MM/DD, so that if the dates ever alphabetize, they will do so in chronological order.~~

~~5. Finally, something I forgot to mention today: add a tag with a unique identifier number. This will be a line under the <match> tag (and indented like the other things under it) saying This means that every passage we keep will have a number that identifies it uniquely. In the future, code will add the tags automatically, and we will only have to fill in the file name and number.~~

6. Decide where each passage you want to keep should begin and end.

7. Do your cleanup on the text. Delete all newline characters, except for where they mark paragraph breaks. This should simplify formatting any paragraph breaks that the passage contains.

8. Save the text at several points as you work.

9. When done with a file, upload it to xml/xml_clean, replacing the placeholder file. Move the original to xml/done.