Conversion Process

2014-07-19

Notes after work of 2014-07-19, converting notes:

- Last conversion left out A LOT and was full of errors. Still used as a base after updating to P5.
- Files in hand do not match the printed copy I have. Was there a subsequent printing? The files seem to be a later version, e.g. have notes on sources where the print indicates none were found (yet).
- Need help with list of TEX character mnemonics, maybe. Depends on how many I come across.
- Some non-TEX command codes in there (eg \sc{}). Some special character encoding not matching TEX either.
- Need help with transliteration of greek. Have marked with `hi[@rend = 'greek']` for now.
- Have marked <title>s in notes as best possible. May or may not be able to expand ot the text following and mark full citations with <bibl>.
- Q: what is the logic to where the notes numbering is reset? Seems to happen right in the middle of a year?

Done:

- Source organized, reviewed, analyzed
- Template TEI file with header complete
- Old conversion code for both text and notes updated to P5 encoding
- Notes conversion pretty far along, producing output compliant with PTJRS TEI schema

To be done:

- Continue fleshing out process with more notes data
- Complete special character handling
- Finish notes conversion, open tasks for enhancements we can do later with marking persons, bibliographic references, abbreviations.

2014-07-27

Notes:

- Am in favor of embedding notes so that individual entries are portable. If not, we can still pull them out from the end of a file.
- Rec'd Laura's notes (scan) about some errors in the old conversion of the text
- Seems the logic for resetting notes numbers is that Script couldn't go past # 99 (lol). TBC.

Done

- Notes conversion iteratively updated. First round of processing complete through FN22.TEX.

Next:

- Finish first round of processing.
- Search and clean '/', '{' codes missed in conversion or too rare to bother programming for. Put in TEI files by year, validate.

2014-08-4

- Notes conversion complete for first phase. Will re-examine possible enrichments later.

2014-08-06

- Abbreviations captured. Those which have multiple expansions marked in the file.
- in the notes, 1867 # 60, there is some greek that is not converted. This section (documented prev.) does not match the hardcopy.
- character conversion errors sent by LB fixed in text conversion
- old conversion code for text discarded as unusable
- many character conversion errors repaired: em, en, thin spaces;
- em spaces used to indicate blanks, should be markup (fix later)
- need to investigate where line breaks may indicate markup needed
- editorial conventions marked. Supplied will need different display style sheet logic than the default ptjrs since it has to retain the question marks in it owing to things like [word? word?]
- special character markup well underway
- started markup of months, days, years, sdays, xdays, dates. Using @role so we can mark iso dates on granular elements later.
- tbd: xlists, daylists, and all macros listed in script file
- tbd: full sweep of spec. chars

(As of) 2014-08-27

- all macros in script file and data have conversion mappings
- all special characters mapped including Greek
- conversion robust after iterative development
- transactionography research done and documented
- 1767-1769 converted and loaded to DC PubMan
- estimate 16-26h to process remaining documents

(As of) 2014-09-22

- Notes have persons of interest to PFE tagged as <persName>. This is not ALL persons. We will extract capsule bios from notes containing them, generate citations (generated), and prepare them for loading into DARMA.
- References and bibliographic cites in the notes have been tagged, though not all of them. Letters to/from and internal references to MB were marked as <ref>. Bibliographic cites were marked as <bibl>. The following were tagged. In the future a pass can be made to complete the bibliographic cites tagging.
    - Letters to and from TJ. Further work may be required to break some of these into multiple references when there are multiple dates mentioned within a reference. Some were done on the cleanup pass. Some other Papers links were
    - internal references to the MB
    - bibliographic cites for the following: Farm Book, Garden Book, Sowerby, Library Catalogue, TJ Index
- Embedded tables have been moved into the listing they belong with. These will need manual review and intervention to improve their visual accuracy.
- TBD:
    - some minor embedded tables cleanup;
    - add dates to entries (as best programmatically possible),
    - linking of notes and reference,
    - generation of capsule bios w/cites for loading to DARMA
    - updates to PubMan xsl for use of <addSpan> and <delSpan>
    - Speak to David Sewell about how we might organize entries (grouping) when this is published

2014-09-24

- Remaining embedded tables cleanup done
- dates added - rather successfully - to the entire data set (excepting embedded tables, which by inference have the data of the row they are embedded in). Each row has an @n attribute storing an ISO date value. Dates that need checking - a very small percentage - are marked by the following processing instructions:
    - <?check day part of date?>
    - <?check month and day part of date?>
    - <?check entire date?>