SRO Proofreading Up-Conversion

This documents some of the semi-automated changes based on bodleian proofreading made to the SRO data after the initial conversion (and up-conversion of roman numerals, fees, dates, names, etc.) from abbreviated tei-corset schema.

1. Remove extraneous use of and

    a. Make “entry” the @type only where the contains either a <title> or a <fee> or both

    I changed any entry which did not have a title or a fee in it to be

    ** **

    This changed:

    Arber: Vol 1: 37, Vol 2: 46, Vol 3: 23, Vol 4: 1.

    b. should only be used where at least one immediately subordinate s has @type=”entry”

    **I changed any which did not have a (after above change) inside it to **

    This changed:

    Arber: Vol 1: 22, Vol 2: 65, Vol 3: 3, Vol 4: 0.

2. Relocation of apostrophized s from within to outside <surnames>s

    I moved all apostrophes inside surnames to just outside, but left them inside the <persName> element.

3. Italics: where <hi rend=”italics”> solely encloses text already tagged as something else, e.g. <persName>, move the @rend value into <persName>, and remove <hi>

    I checked any element which has only a <hi> child and merged its @rend with the parent's (if it had one).

4. Where smallcaps is the @rend value around <surname> (and not <forename>), still mark @role as “non-stationer”

    I checked for any persName (of @role='unknown') where its children had a @rend and rewrote the @role as stationer or non-stationer if it had bold or smallcaps.

5. Can we assumed “unknown” equals @role=”non-stationer”? If so, please change accordingly

    No, it is a useful distinction to preserve. The non-stationers have been explicitly marked (as smallcaps) as non-stationers but 'unknown' merely means I wasn't able to

tell from the encoding (or they didn't mark it as being smallcaps).

6. Check churches with their patronal saints are not captured as <persName>s
   I used an xpath to search for all instances of 'saint' inside lower-cased persName's and changed their markup manually. Some instances of place names (Saint Davids) changed to placenames, but I think there are more like that incorrectly marked as persName.

7. Check Monarch's names not tagged as surnames.
   As with 6 I used an xpath to find these and modify the markup manually.

8. Where women are referenced alongside their husbands, the same surname be deduced and added to the woman's tagging, e.g.
   <persName role="non-stationer" rend="smallcaps">
   <surname>Elizabeth</surname>
   </persName>, the widow of <persName role="non-stationer" rend="smallcaps">
   <forename>Robert</forename>
   <surname>Redman</surname>
   </persName>

   I decided against adding the woman's surname since it is not present in the text. (This is a transcript after all.) However, developed an xpath which allowed me to find these (where 'widow' or 'wife' mentioned) and correct some of those whose forenames are mistakenly marked as surnames.

9. Regularization of role-titles associated with <persName>s, where warden/wardyn/master appear sometime after the name, e.g.
   <persName role="unknown">
   <forename>John</forename>
   <surname>Cawod</surname>
   </persName> and
   <persName role="unknown">
   <forename>henry</forename>
   <surname>Cooke</surname>
   </persName> Wardyns</hi>

   There was not an easy way to do this for persNames outside of the <fw> or another containing element since the text 'Wardyns' or similar may be more than one text node away. There was too much risk of data corruption doing this across all of arber. Some instances manually corrected.

10. Moving of role-related names - "mr", "master", "king", "queen","mayor", "lord mayor", "lord", "duke", "earl", "alderman" inside the <persName> that follows them.

    This was non-trivial to do efficiently but able to be accomplished for lower-cased and initial-letter-upper-cased forms of these phrases. The case of the word was preserved while moving it inside the <persName>.

11. Addition of year field to <date>s where the year (often marked with "anno" precedes the date, e.g.
    <date when="--03-13">
    <hi rend="italics">Anno</hi> 1550. the 13 of march</date>

I replaced any missing year with the first instance of '1[0-9][0-9][0-9]' found inside the date element or, to cope with some where the year comes afterwards in a bit of text, appears in the first text sibling following the date. Not foolproof but an improvement.

This updated:
Arber: Vol 1: 419 , Vol 2: 187, Vol 3: 77, Vol 4: 20.

12. Roman numerals: interpret these in encoding, wherever they appear

    Only able to do this where they have been marked as numbers. Automatically determining whether they are numbers or words is non-trivial. I've changed as many that are marked as <num> that weren't already fees. I've attempted to determine whether it is an amount of money or just a number.

    This updated:
    Money: Arber: Vol 1: 1247 , Vol 2: 1946, Vol 3: 110, Vol 4: 6.
    Other roman numerals: Arber: Vol 1: 663 , Vol 2: 284, Vol 3: 15, Vol 4: 0.

13. Check numerals for processing errors

    Spot checked distinct-values list of found numbers... could provide for later manual checking.

14. <persName>s often appear at the end of book titles, but the <title> excludes them. Automatically add <persName> immediately following </title> to the <title>

    I've moved any persNames which instantly follow a </title> into the title. There is a danger of course, that this <persName> might be the person registering the title but quick spot checking didn't find this.

    This updated:
    Arber: Vol 1: 433 , Vol 2: 1133, Vol 3: 1817, Vol 4: 2221.

15. Check “van”, “de”, “wan de” etc. are not included in <forename>

    Updated many of these xpath searching for them and manually editing as attempts to script caused more errors.

16. Check dates are not inside <seg>s for fees

    Searched for these manually.

17. Escape “de”, “th” etc. after dates in Roman numerals, to enable machine-reading the dates;

    Dates need much more processing and checking that would take several more days of development; so generally have left them alone except as noted above (not in fees) and further above those which didn't have years but could have done.