# State of the Data

# Overview

In Aug. - Sept. of 2014 Infoset undertook the task of converting the existing digital form of the Jefferson Memorandum Book Volumes (MB) into XML conforming to TEI P5. The source files were in two formats: The text was encoded for the Scribe word processing system that was used for the typesetting of the volumes, and the textual notes were in TeX. Additional resources included a rapidly drafted conversion to TEI P4 from around 2000.

The programmatic part of the conversion has reached the end of its first phase. This document describes the data as it stands, gives background to some of the decisions taken and priorities set, and outlines future work that may take place to enrich the TEI form of the MB. Informal notes in the form of status updates made during the conversion are available [here](file:///C:\wiki\display\VIFH\Conversion+Process).

Primary priorities of this first phase of work were:

* Structure the data into a TEI P5 conforming state
* Preserve editorial conventions from the original documentary edition
* Preserve formatting from the original as possible and/or appropriate. Also preserve comments with formatting instructions (= “interventions”) from the Scribe and TeX files.
* Achieve 100% character, character formatting, and copy fidelity. Encode the data in Unicode, marking languages if possible (e.g. Greek)
* Segment the data into year divisions and mark all transactions
* Date the transactions
* Mark persons in the textual notes so that capsule biographies can be extracted (for PFE)
* Formally link references in the text to the notes, retaining original print numbering.
* Mark locations in the data where the conversion output will need review and possibly intervention to correct or improve.

Secondary priorities:

* Markup of references to documents to/from TJ
* Markup of bibliographic references.
* Markup measures, commodities, or anything else that might support future digital works emphasizing transactional research.

Early on the degree of difficulty and scope of the MB precluded the last two secondary priorities, though some important bibliographic references were marked. Details on those and future work that can be done to enrich the TEI version of the MB follow below.

# Data Structures

The existing TEI P5 customization used by the Papers of Thomas Jefferson: Retirement Series was chosen as the schema for the MB, as there was already substantial supporting infrastructure available for it and the MB are (obviously) a closely related work.

As is the case with many documentary editing works there is a mix of semantic and presentation markup done, i.e. the documents are structured and marked up for what they are, but presentational markup is embedded as well to preserve the flavor of the historical document.

## The Text

Using the print copies as a guide, the MB entries are laid out in tabular fashion though not slavishly so since the digital medium offers some flexibility print did not. The basic layout of entries is as follows:

# Columns

# 1 2 3 4

# Year

# Month Day Entry for the day with

# a hanging indent.

# Day X Entry for the day with an X, always includes a date. Has a

# hanging indent.

# Entry for the 'S day' (same day), sans date or 'X' goes here left aligned on entry column

# X An 'X day' looks like this, left aligned on entry column. Never has a date.

# X A 'day list' indents to align left with hanging indent above, always has an 'X' or a checkmark

# An 'X list' indents to align left with hanging indent above. No preceding character.

# Sublists indent another em and have a hanging indent

# Subsublists indent another em and have a hanging indent (by default, usu. these indicate a special alignment)

Where there four columns, the first three usu. being a component of the date or indicator thereof, and the fourth being a text area where transaction or multiple transactions are listed for it. Levels of indentation within that area have meaning, and there may be embedded tables within it as well.

The names of the roles the cells and entries have is shown above (and carried to the XML markup—see below); month, day, s day, x day, day list, x list, sublist, subsublist.

Of course, this IS Jefferson, and the MB were created over a long period of time, so there are many permutations on the above to be found. There are also standalone charts/tables for things such as weather.

Translated into XML markup, sans wrapping TEI document markup, the entries end up looking like:

. . .

<div1 type="lawyear">

<head>

<supplied>1768.</supplied>

</head>

<table>

. . .

<tr n="1768-10-18" valign="top" role="month">

<td width="6.25%">Octob.</td>

<td role="day" width="6.25%" align="right"><date>18.</date></td>

<td width="6.25%"/>

<td xhtml:style="margin-left:1em;text-indent:-1em;">Thomas Phelps (Buckingham) v. James Lyon (Chesterfeild).

A petn. for 400 acres on Wolf creek Buckingham. The petn. has been entered but the summs.

never served so send to the pl. Alias Summs. Recd. 20/.</td>

</tr>

<tr n="1768-10-21" valign="top" role="day">

<td width="6.25%"/>

<td width="6.25%" align="right"><date>21.</date></td>

<td width="6.25%"/>

<td xhtml:style="margin-left:1em;text-indent:-1em;">Anne Scott (Amelia) v. Joseph Egleston et ux.

(Amelia). Removed from Amelia to this Gen. court by Certiorari. Recd. 50/.</td>

</tr>

<tr n="1768-10-21" valign="top" role="sday">

<td colspan="3" width="18.75%"/>

<td xhtml:style="margin-left:1em;text-indent:-1em;">Tetter v. Paris.</td>

</tr>

<tr n="1768-10-21" valign="top" role="sday">

<td colspan="3" width="18.75%"/>

<td xhtml:style="margin-left:1em;text-indent:-1em;">Henderson v. Paris.</td>

</tr>

<?format The two lines above are to be offset to the left of those below, ?>

<?format with a bracket, see copy. special coding needed. JJL ?>

<tr n="1768-10-21" valign="top" role="sday">

<td colspan="3" width="18.75%"/>

<td xhtml:style="margin-left:1em;text-indent:-1em;"><?format .in +22?>I find that Tetter's

petn. is entered April 12. 1768. for 275 acres and Henderson's is entered April 14.

for 271 acres.</td>

</tr>

<tr n="1768-10-20" valign="top" role="day">

<td width="6.25%"/>

<td width="6.25%" align="right"><date>20.</date></td>

<td width="6.25%"/>

<td xhtml:style="margin-left:1em;text-indent:-1em;">Thomas Wood (Pr. Edward) v. Philemon

Halcomb (Pr. Edwd.) and Leander Hughes (Pr. Edwd.). Trespass. To appear for pl. Writ

retnble. to this court.</td>

</tr>

<tr n="1768-10-20" valign="top" role="sday">

<td colspan="3" width="18.75%"/>

<td xhtml:style="margin-left:1em;text-indent:-1em;">Grattan v. Waterson.</td>

</tr>

<tr n="1768-10-20" role="sday" valign="top">

<td colspan="3" width="18.75%"/>

<td>

<table>

<?format tabs 26 66?>

<?format Next line will align iff £ is one em wide. JJL ?>

<tr role="etbt">

<td>Judgmt. for</td>

<td>£6– 3–8</td>

</tr>

<tr role="etbt">

<td>Tobo. Cts. 127 ℔ 21/2 L.'s fee 15/</td>

<td> 1–16–2</td>

</tr>

<?format Change hyphens on line below to horizontal rule. ?>

<tr role="etbt">

<td/>

<td>———-</td>

</tr>

<tr role="etbt">

<td>to be recd. by Grattan</td>

<td> 4– 7–6</td>

</tr>

</table>

<?format tabs end-current-spec?>

</td>

</tr>

</table>

</div1>

. . .

Where you see the following conventions for the entries:

* @role is used on tr and td to denote the kind of information in the row or cell. Add etbt to this list of names, that denoting a table embedded within a table column (4).
* @n is used to store a date value for now. In the future we might consider adding one of the att.datable attribute classes to tr in the TEI customization to more accurately describe these, but that would be best done after they are reviewed. Date values that need checking - a very small percentage - can be located by running a report listing any of them that won't properly cast to an ISO date or are marked by the following processing instructions adjacent to the row:
  + <?check day part of date?>
  + <?check month and day part of date?>
  + <?check entire date?>
* @align, @valign, @width, and @xhtml:style are used to store presentation related values. Once the data is reviewed we could work on reducing dependence on embedded formatting markup, moving display logic into the presentation layer to the greatest extent possible.

While formatting instructions are discussed below, it's worth covering some related to tabular presentation here. The Scribe macro files did contain measurements for tabular settings and there are formatting instructions throughout that denote changes in them. Default width of the page in the print product was 80 units, maximum with allowed 120 units, and set of tab stops are contained in these instructions based on those units, modifying the default tab stops, returning to them, altering maximum line widths, etc. Efforts to employ them didn't bear much fruit since: the screen offers more real estate (and aspects of the display are user alterable such as width and zoom); and, generally the tables formatted well once divided into cells properly. With some additional tweaks to padding it appears that most of the tables requiring hand massaging to present well would need it anyway, and implementing measurements based on these print formatting instructions wouldn't nec. bring better results.

### Other Conventions

* Formatting instructions were converted to XML processing instructions.
* Character formatting changes use the <hi> element and the @rend attribute to store the format details, e.g. <hi rend="italic"> or <hi rend="Greek">
* Editorial conventions follow  the Retirement Series practice of using <supplied> for editorially supplied text and <del status="retain"> for restorations of deleted text. In addition, there is other markup used in the MB where spans of deletions or restorations were non-hierarchical: <addSpan/> and <delSpan/>. Noted below in Supporting Infrastructure Updates are some changes to the Retirement Series display stylesheet rules that will need to be made for them to work correctly with the MB, where sometimes multiple pieces of supplied text are in the same element, or use of '?' within them can cause display errors.
* Note references use the ref element, e.g. <ref type="textual-note" n="n36" target="d1e2274a1048964"/>, where @n is used to retain the original print reference number. Again (and listed below), there are changes needed to make the display stylesheets work properly with these since in the MB there can be more than one reference to the same note. In these instances the number is thrown off.

### Known Issue

The text is slightly incomplete at this time, in that the source file for 1798 was truncated. We are researching whether a non-corrupt version can be found.

## The Textual Notes

Textual notes, formerly travelling separate from the text, are now grouped into divisions as siblings to the year they belong to. Here's a sample (again sans TEI doc structure):

**Example Textual Note**

. . .

<note type="textual-note" xml:id="d1e1254a1048964" n="8">

<p>The architectural section of the <bibl>1767 MB</bibl>, which begins on the verso

of the page ending with the preceding entry, contains five pages of

memoranda followed by two blank pages. It is difficult to determine

exactly when TJ made these memoranda, and the chronology is further

complicated by additions and emendations he made over an extended

period. The earliest entries must date between 1767, when TJ first

began using the 1767 notebook, and 2 Oct. 1769, when the

miscellaneous memoranda resume. MB entries for May 1768 indicate

that TJ had by then determined to build at Monticello and by Feb.

1769 he had resolved on a plan for a dwelling, although it was

evidently smaller than the one to which these notes pertain (<ref pm:link-pointer-type="relatedDocument">MB 7

and 15 May 1768, cash accounts</ref>; <ref

pm:link-pointer-type="relatedDocument">TJ to Thomas Turpin, 5 Feb.

1769</ref>). The dimensions and brick calculations in these

memoranda correspond to a plan (Nichols, No. 40) for a house which

was to have a two-story central block and flanking wings, each

eighteen feet square with an attic above. This preliminary plan was

later modified by the addition of octagonal projections to the wings

and central block, as in Nichols, No. 49, which, with the elevations

in Nichols, Nos. 47 and 48, was the basis for the construction of

the first Monticello house, believed to have been begun in 1770. </p>

<p>The architectural memoranda are particularly important not only for

the information about the design and construction of the house but

also for the insight they provide into TJ's approach to solving

formal architectural problems and his reliance on two important

books, <persName>James Gibbs</persName>, <hi rend="italic">Rules for

Drawing the Several Parts of Architecture</hi> (<bibl>Sowerby, No. 4l84</bibl>)

and the Leoni edition of either 1715 or 1742 of <hi rend="italic"

>The Architecture of A. Palladio</hi> (<bibl>Sowerby, Nos. 4174, 4175</bibl>),

for the detailing of architectural elements. The <persName>Marquis

de Chastellux</persName>, visiting TJ in 1782, remarked that this

early house “resembles none of the others seen in this country” and

added that “Mr. Jefferson is the first American who has consulted

the Fine Arts to know how he should shelter himself from the

weather” (Chastellux, <hi rend="italic">Travels</hi>, <hi

rend="small-caps">ii</hi>, 391). </p>

</note>

. . .

The following content items are marked in the notes:

* ref is used for references to other TJ documents. It is also used to mark internal cross-references to other locations in the MB.  If by happenstance a reference to a letter or document from a non-TJ source could be marked, it was, but there is no consistency to it. Where possible references to multiple correspondences (as on a sequence of dates) were marked, but there are still instances that need to be split, e.g. change <ref pm:link-pointer-type="relatedDocument">MB 1 Aug. 1789, 4 Mch. 1790</ref> into <ref pm:link-pointer-type="relatedDocument">MB 1 Aug. 1789</ref>, <ref pm:link-pointer-type="relatedDocument">4 Mch. 1790</ref>.
* persName is used to mark person names. Not all person names were marked however, only those potentially of interest to PFE.
* bibl is used to mark citations to the following: Farm Book, Garden Book, Sowerby, Library Catalogue, TJ Index
* title was used to mark such bibliographic citation components where it could be done with certainty, and to make easier future marking of cites.

# Review and Intervention Work

Basic review processes:

* Slugging of digital files to manuscript copy. Review of character transliteration (esp. for Greek).
* Checking of dates, correction if nec.

More detailed review and intervention processes:

* Search and check all processing instructions named check.
* Formatting of tables with intervention instructions as nec. Search for XML processing instructions named format to get details.

# Supporting Infrastructure Updates

## Stylesheet Updates

See above. There are stylesheet updates needed to make the editorial conventions and not references display correctly.

# Future Work

* Analysis of whether the current segmentation of the data supports digital publication or whether that should be altered.
* Any programmatic support for the Review and Intervention process (above).
* In the course of work some time was taken to review proposed methods of handling transaction records in TEI. Most of that time was spent looking at the article authored by Catherine Tomasek and Syd Bauman, [Encoding Financial Records for Historical Research](http://jtei.revues.org/895). This is a promising approach, and one takeaway is that it may be possible to generate the “transactionography” markup from the MB TEI source if more work on it is completed. While not an exhaustive list, that would include marking in the text: run-in transactions, unique ids for all transactions, names, places, orgs, attestation, certainty, measures, units, currencies, and commodities. This would make the data suitable for a digital edition that mined it. See [Notes on Transactionography](file:///C:\wiki\display\VIFH\Notes+on+Transactionography) for some offhand thoughts recorded during the review.
* Complete markup of bibliographic references in the textual notes, and any references to publications from Jefferson's libraries in the text.
* References to TJ document (letters to/from) could be formally linked to digitally published versions of them. Split references to multiple dates where needed.
* Internal cross-references could be formally linked to their targets in the MB.
* Mine place identifications in the textual notes.
* Mark quotations in the textual notes.
* Reduce use of embedded formatting markup.
* Links between references and notes are one way, not bidirectional as they are in the Retirement Series documents. The notes have unique ids and the references point to them, but not the reverse. I don't see where this is anything that needs remedying as it can be affected in post-processing to prepare the data for publication if nec., but it's easy to do next time we're in the data if desired.