Kicking and Screaming: Challenges and advantages of bringing TCP texts into line with the Text Encoding Initiative

James Cummings and Sebastian Rahtz

IT Services, University of Oxford

September 2012

Introduction

This paper[1] addresses some of the practical problems of working with the underlying digital files prepared by the Text Creation Partnership using tools developed for standard TEI files. This work at Oxford was driven primarily by a desire to experiment with useable ePub editions of the ECCO corpus released in 2010, and was undertaken at Oxford University Computing Services (now subsumed into the IT Services department at the University of Oxford), independently of the TCP team at the Bodleian Library or in Michigan.

The 40188 files of the TCP EEBO collection have been built up over the last decade using the markup technology of the 1990s, namely SGML and variations on the third edition of the Text Encoding Initiative Guidelines (http://www.tei-c.org/), to a gradually increasing standard of consistency. The delivery of the texts through the conventional web site alongside the facsimile page images works well, but problems arise if we want to take advantage of some of the tools now commonly used to process digital files, particularly those based on the current TEI recommendations. This involves transforming the SGML markup to XML, and then to the latest edition of the TEI (P5). We discuss some of the problems involved in this type of conversion, such as changes needed to the TEI Guidelines themselves to cover textual phenomena identified in EEBO

[1] We would like to express great thanks to Paul Schaffner for his very patient and understanding help in explaining decisions made around the TCP and for replying immediately to silly questions about the markup. We are also grateful to Martin Mueller, Stephen Ramsay and Brian Pytlik Zillig of the Monk and Abbott projects, who wrestled with some of the same dilemmas before and in parallel with us, for pleasurable discussions of minutiae and a shared urgency about the importance of exposing the TCP texts in current TEI markup.

which cannot adequately be described in current TEI recommendations, and decisions needed to map some of the variants adopted by TCP back onto canonical TEI P5 markup. The work of Brian Pytlik Zillig and Stephen Ramsay (see Pytlik Zillig 2009) testing whether conversions have lost any content along the way is beyond the scope of this paper, but is clearly an important part of the longer-term development.

We present some of the software we have developed for the TCP conversion. The exercise of transformation gives an interesting opportunity to examine some of the encoding of TCP texts, analyze the range of textual phenomena which are recorded, and predict which structures which will be amenable to discovery by future scholars.

The 40000-text corpus of TCP also provides a good test of general TEI tools. For this paper we describe some tools and the results we found when using them on TCP texts. As a case study we examine the generation of ebook editions (ePub format) of the TCP texts from the converted TEI. The results of such conversions can be discussed for their usefulness for contemporary readers and any failures in representing the intellectual content of the original text.

Converting the TCP files to TEI

The Text Creation Partnership digitization programme is a large and complex operation, with very detailed guidance and standards (http://www.textcreationpartnership.org/docs/) worked out over the last decade. The decision was taken when it started to use SGML markup as the archival form, and a variation on the Text Encoding Initiative Guidelines, version P3. Unfortunately, SGML-aware software is increasingly hard to come by, and the advantages of working in XML are large.[2] The TEI has also moved on a lot since P3, with the current TEI P5 release making many changes and improvements (some of them due to proposals arising from the TCP work). Interchange or comparison with other TEI texts, or use of TEI-aware software, dictates that we should have a way to transform the TCP texts into valid TEI P5 XML. This does not mean that once texts are in TEI that they are inherently interoperable, at least not without effort, but this should be easier with the TEI version of the EEBO TCP corpus owing to them being created by a single project under a single set of encoding guidelines.

Most of the TCP transform is relatively trivial, managed with a simple XSLT script. What is needed is to

[2] One disadvantage is that XML only supports Unicode; any character entities must resolve to a Unicode code point. The TCP SGML files can contain character entities which do not have to be resolved at the time of validation.

1. put all the elements in the TEI namespace (`http://www.tei-c.org/ns/1.0`)
2. convert the element and attribute names from all uppercase into the 'camelCase' used by TEI XML
3. change the name or structure of elements or attributes which have changed name
4. change some attribute values to fit stricter data types

Thus from

```
<DIV1 TYPE="title page">
 <PB REF="1"/>
 <P>THE THEORIKE AND PRACTIKE OF MODERNE WARRES, Discoursed in Dialogue vvise.</P>
 <P>VVHEREIN IS DECLARED THE NEGLECT OF Martiall discipline: the
   inconuenience thereof: the imperfections of manie training Captaines:
   a redresse by due regard had: the fittest weapons for our Moderne
   VVarre: the vse of the same: the parts of a perfect souldier in
   generall and in particular: the Officers in degrees, with their
   seuerall duties: the imbattailing of men in formes now most in vse:
   with figures and Tables to the same: with sundrie other martiall
   points. VVritten by ROBERT BARRET.</P>
 <P>Comprehended in sixe Bookes.</P>
 <Q>Ozar morir, da la vida.</Q>
 <P>
  <FIGURE>
   <HEAD>ANCHORA SPEI</HEAD>
   <FIGDESC>printer's or publisher's device</FIGDESC>
  </FIGURE>
 </P>
 <P>
  <HI>LONDON,</HI> Printed for VVilliam Ponsonby. 1598.</P>
</DIV1>
```

we create

```
<div type="title_page">
 <pb facs="1"/>
 <p>THE THEORIKE AND PRACTIKE OF MODERNE WARRES, Discoursed in Dialogue vvise.</p>
 <p>VVHEREIN IS DECLARED THE NEGLECT OF Martiall
   discipline: the inconuenience thereof: the imperfections
   of manie training Captaines: a redresse by due regard had:
   the fittest weapons for our Moderne VVarre: the vse of the
   same: the parts of a perfect souldier in generall and in
   particular: the Officers in degrees, with their seuerall
   duties: the imbattailing of men in formes now most in vse:
   with figures and Tables to the same: with sundrie other
```

```
   martiall points. VVritten by ROBERT BARRET.</p>
 <p>Comprehended in sixe Bookes.</p>
 <q>Ozar morir, da la vida.</q>
 <figure>
  <head>ANCHORA SPEI</head>
  <figDesc>printer's or publisher's device</figDesc>
 </figure>
 <p>
  <hi>LONDON,</hi> Printed for VVilliam Ponsonby. 1598.</p>
</div>
```

Note the change of the @*type* value from `title page` to `title_page`; the TEI defines the attribute as an enumeration of space-separated values (http://www.tei-c.org/release/doc/tei-p5-doc/en/html/ref-att.typed.html), so `title page` would mean that the `<div>` is of type `title` and `page`.

The TCP texts use the `<MILESTONE>` element to indicate a number of things. These have been standardised according to the following criteria.

1. If the parent of the `<MILESTONE>` is a TCP `<NOTE>` element but the `<MILESTONE>` does not have an @*N* attribute, then it is ignored.
2. If the `<MILESTONE>` has a @*UNIT* attribute but does not have an @*N* attribute (or it is empty) then it is turned into a marginal note.
3. Conversely, if the `<MILESTONE>` has no @*UNIT* but does have a @*N* attribute that is used in turning it into a marginal note.
4. If none of these conditions are matched, and it has a @*UNIT* that is in a list of editorial units (e.g. article, canon, chapter, commandment, date, day, folio, indulgence, leaf, line, monarch, motive, month, reason, verse, year) decided in correspondence with TCP and it is turned into a `<note>` with a @*subtype* attribute preserving this unit.
5. Otherwise a a marginal `<note>` with a `<label>` child containing a combination of the content of @*UNIT* and @*N* is created.

This conversion attempts to preserve the intention behind the TCP use of `<MILESTONE>`, but always runs the risk of mis-categorising some uses.

In some cases there are markup structures which TCP encoded containing a `<P>` or sometimes a `<HEAD>` element which do not need to be migrated the TEI equivalent (`<p>` and `<head>` respectively) because they will end up converted into a structure which does not need this. For example, when the resulting `<head>` should actually be an `<epigraph>` or the `<p>` would be inside an `<add>` element which contains paragraph content.

The majority of other elements, and indeed attributes, are able to be smoothly transformed to their modern TEI P5 equivalents. One of the benefits of undertaking such a conversion is the chance to correct human error in the original. In many cases this is too difficult to distinguish and so the conversion silently outputs the equivalent TEI P5, but in some cases, especially in attribute values, it makes sense to standardise and correct errors. The TCP @*PLACE* attribute, indicating where a note or addition was made, is a good example of this. Our investigations of this found a number of inconsistencies in the text of this attribute. In converting it the following corrections were made when processing a TCP @*PLACE* attribute:

1. If the attribute's value is equal to 'marg', 'marg;', 'marg)', 'marg='ma / rg', or '6marg' it is safely and reasonably convereted to 'margin'.
2. If instead the value claims that it is 'unspecified', the attribute is removed since it is assumed to be unspecified if not present.
3. If the attribute's value is 'foot', 'foot;', or 'foor;' then it is replaced with the recommended 'bottom'.
4. Occasionally TCP has provided characters such as: '‡', '†', '||', '6', '“', '1', or '*' instead of a location. It was thought best in converting these not to discard them, but instead preserve them as an @*n* attribute and not create the @*place* attribute.
5. Otherwise a TEI @*place* attribute is created and the current value passed through.

In general such corrections have only been undertaken where an attribute has a limited set of values, where the value provided by TCP is clearly a typo (e.g. has a semi-colon after it), and where it has been easily detected in the debugging of various other aspects of the conversion. A future conversion might be applied to the resulting TEI in order to correct other inconsistencies not yet discovered.

Validation of the TEI files

Having converted the TCP files to (what looks like) TEI XML, we can then validate it against the TEI current schemas (1.2.0 at the time of writing). Unfortunately, the TCP has marked up constructs which are not allowed for in the TEI, which occur in about 3000 texts. These fall into seven groups, which require extensions to the TEI as listed in Table 1.

	Element	**change**
1	`<signed>`	needs a looser content model (paraContent, not phraseSeq), so that it can contain `<list>`; and needs to be allowed to appear at top of `<div>` as well as bottom
2	`<stage>`	needs placement attributes (@*place*), to enable stage directions to be placed in the margin, and allowed in model.phrase
3	`<trailer>`	needs a looser content model to let it contain `<l>` and `<list>`
4	`<cell>`	needs a looser content model (specialPara)

5	<closer>	allowed to have <postscript> inside it
6	<label>	needs looser content model (specialPara)
7	<table>	needs to allow model.divBottomPart at the end, so that it can contain <trailer>

Table 1 Changes needed in TEI to accomodate EEBO conversions

Examples of each variation are given below.[3]

The extent to which these changes impact on the corpus is catalogued in Table 2, which shows that 1, 4, and 5 occur significantly often enough to need a solution — it may well be worth examining the actual instances of 2, 3, 6 and 7 by hand to see whether a different encoding could be used.

Divergence category	ECCO	EEBO	All	% of texts
from TEI	287	2775	3062	7.1920
1	68	1593	1661	3.9014
2	0	6	6	0.0141
3	3	36	39	0.0916
4	10	91	101	0.2372
5	246	1339	1585	3.7228
6	1	39	40	0.0940
7	0	5	5	0.0117

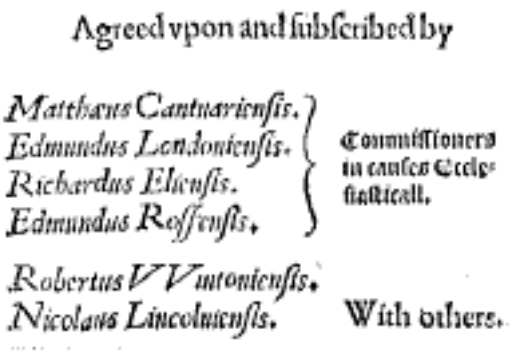
Table 2 Extent of variation in converted texts from the TEI schema

1. Example of <list> inside <signed> (A00033, STC 10029.5):

```
<closer>
 <signed>Agreed vpon and subscribed by
 <list>
   <item>Commissioners in causes
Eccle<lb rend="hidden" type="hyphenInWord"/>siasticall.<list>
     <item>Matthaeus Cantuariensis.</item>
     <item>Edmundus Londoniensis.</item>
     <item>Richardus Eliensis.</item>
     <item>Edmundus Roffensis.</item>
```

[3] We are grateful to ProQuest for permission to use fragments of images for illustrating the examples.

```
    </list>
   </item>
   <item>Robertus VVintoniensis.</item>
   <item>Nicolaus Lincolniensis.</item>
  </list>With others.</signed>
</closer>
```

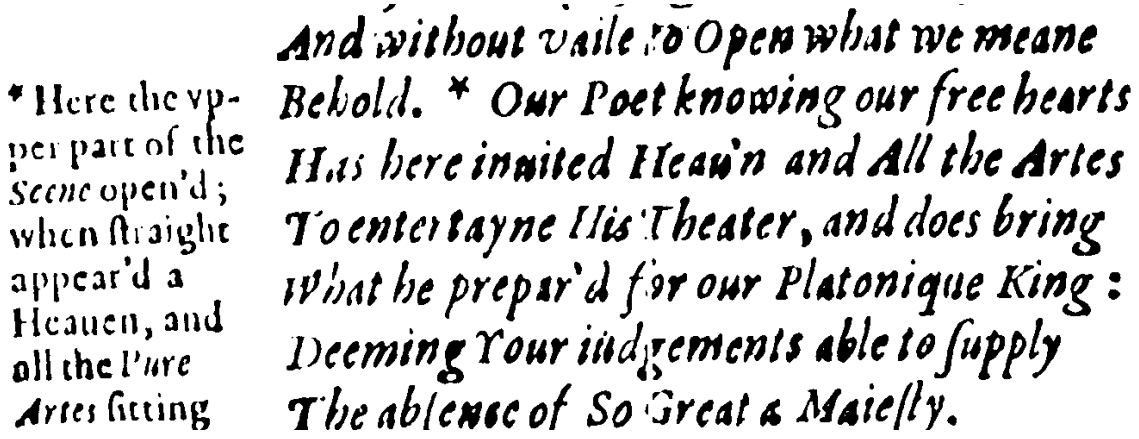

Agreed vpon and ſubſcribed by

Matthæus Cantuarienſis.
Edmundus Londonienſis.
Richardus Elienſis.
Edmundus Roffenſis.
} Commiſſioners in cauſes Eccleſiaſticall.

Robertus VVintonienſis.
Nicolaus Lincolnienſis. With others.

2. Example of <stage> with @*place* (A03496, STC 13617):

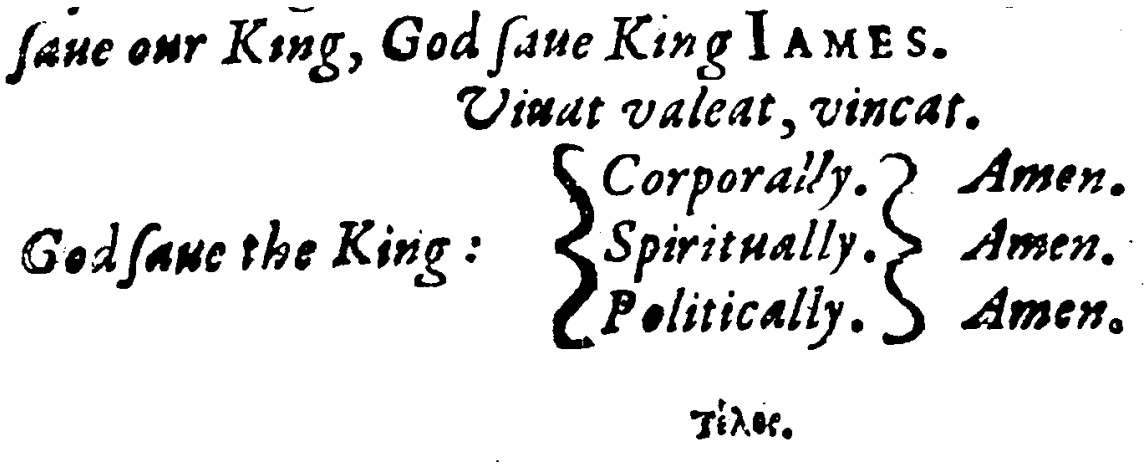
```
<l>Behold. <stage n="*" place="margin">Here the
vp<lb rend="hidden" type="hyphenInWord"/>per part of the <hi>Scene</hi> open'd; when
   straight appear'd a Heauen, and all the <hi>Pure Artes</hi> sitting on
   two semi<lb/>circular ben<lb/>ches, one a<lb/>boue another: who sate
   thus till the rest of the <hi>Prologue</hi> was spoken, which being
   ended, they descended in order within the <hi>Scene,</hi> whiles the
   Musicke plaid</stage> Our Poet knowing our free hearts</l>
```


* Here the vpper part of the Scene open'd; when straight appear'd a Heauen, and all the Pure Artes ſitting

And without vaile to Open what we meane
Behold. * Our Poet knowing our free hearts
Has here inuited Heau'n and All the Artes
To entertayne His Theater, and does bring
What he prepar'd for our Platonique King:
Deeming Your iudgements able to ſupply
The abſence of So Great a Maieſty.

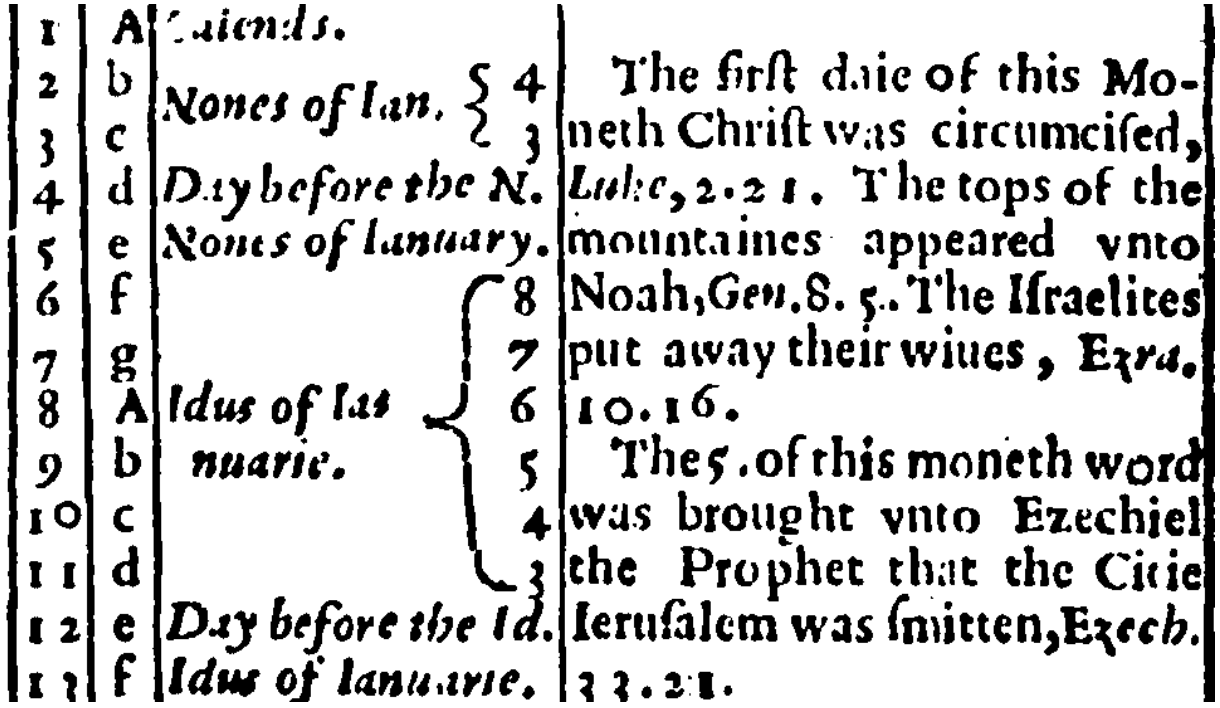
3. Example of <list> inside <trailer> (A01472, STC 11597):

```
<closer>Viuat valeat, vincat.</closer>
<trailer>God saue the King:
<list>
  <item>Corporally. Amen.</item>
  <item>Spiritually. Amen.</item>
  <item>Politically. Amen.</item>
 </list>
</trailer>
```



4. Example of <p> inside <cell> (A01629, STC 11761.5):

```
<cell>
 <p>The first daie of this Mo<lb/>neth
   Christ was circumcised,
 <hi>Luke,</hi> 2. 21. The tops of the
   mountaines appeared vnto
   Noah, <hi>Gen.</hi> 8. 5. The Israelites
   put away their wiues, <hi>Ezra.</hi>
   10. 16.</p>
 <p>The 5. of this moneth word
   was brought vnto Ezechiel
   the Prophet that the Citie
   Ierusalem was smitten, <hi>Ezech.</hi>
   33. 21.</p>
 <p>The sixt of this Moneth
   Christ was worshipped of the
   wise men, Mat. 2. 1. &amp;c. bap<lb/>tized
 <hi>Mat.</hi> 3. 15. turned wa<lb/>ter
   into wine, <hi>Ioh.</hi> 2. 1. &amp;c.
   as testifieth Epiphanius.</p>
</cell>
```



5. Example of <postscript> inside <closer> (A01160, STC 11275):

```
<closer>
 <signed>L'Amy de Coeur.</signed>
 <postscript>
  <p>Monsieur the Counte, shall finde his most affectionate
     commendations. His Highnesse shall see this word. L'Amy de
     Coeur.</p>
  <p>Come with speed.</p>
 </postscript>
</closer>
```

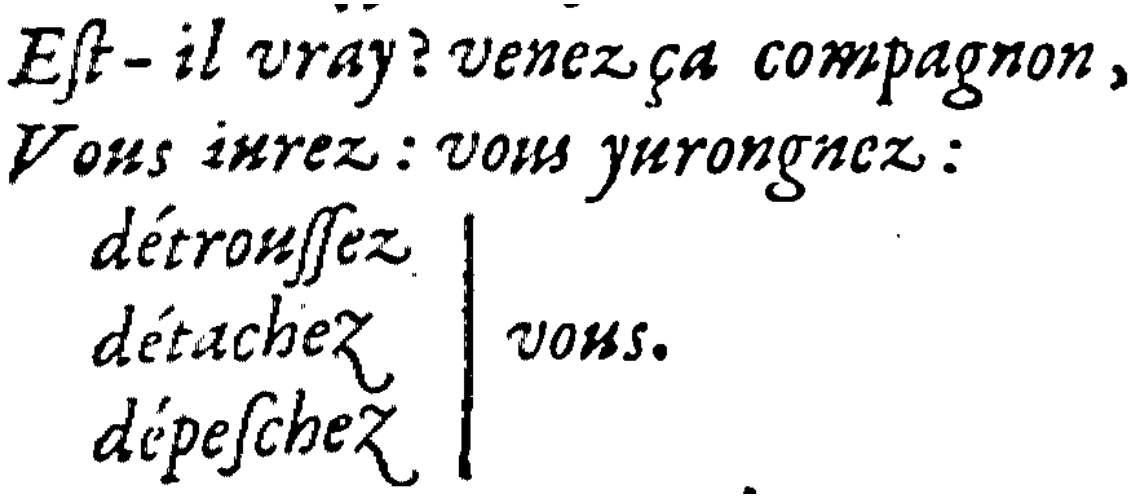
ſelues vnto you, this 12. of May. *L'Amy de Cœur.*

Monſieur *the Counte, ſhall finde his moſt affectionate commendations. His Highneſſe ſhall ſee this word.* *L'Amy de Cœur.*

Come with ſpeed.

6. Example of <list> inside <label> (A02026, STC 12173.3):

```
<p>
 <hi>Est-il vray? venez ça compagnon, Vous iurez: vous yurongnez:</hi>
 <list>
  <label>
   <list>
    <item>détroussez</item>
    <item>détachez</item>
    <item>dépeschez</item>
   </list>
  </label>
  <item>vous.</item>
 </list>
 <list>
  <label>Nicolas se mocque de</label>
  <item>
   <list>
    <item>moy,</item>
    <item>vous,</item>
    <item>elle,</item>
    <item>eux.</item>
   </list>
  </item>
 </list>
</p>
```

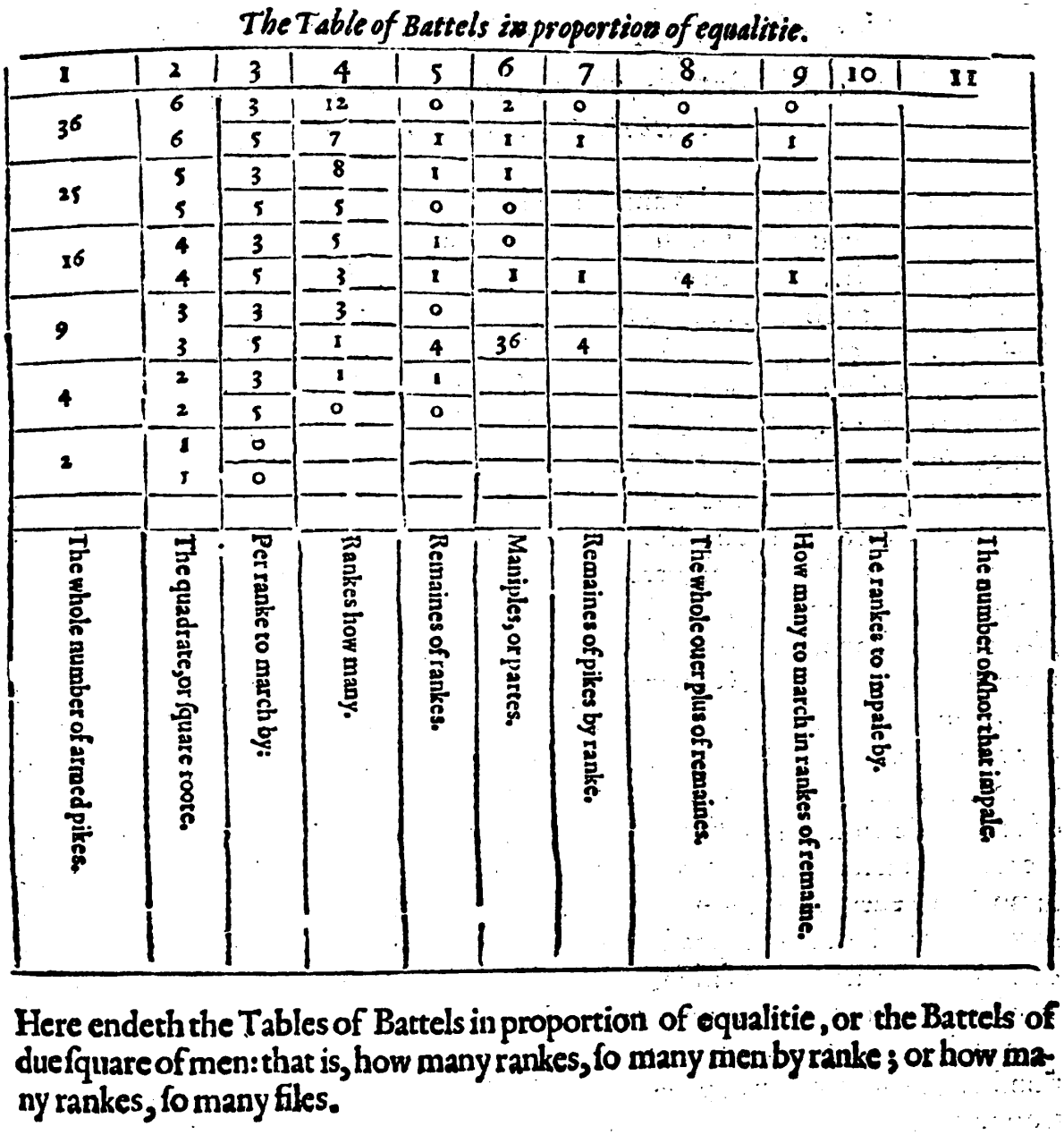


7. Example of <trailer> at the end of <table> (A04863, STC 1500):

```
<table>
 <head>The
   Table of Battels in proportion of
```

```
  equalitie.</head>
<row>
 <cell>1</cell>
 <cell>2</cell>
 <cell>3</cell>
 <cell>4</cell>
 <cell>5</cell>
 <cell>6</cell>
 <cell>7</cell>
 <cell>8</cell>
 <cell>9</cell>
 <cell>10</cell>
 <cell>11</cell>
</row>
<row>
 <cell rows="2">36</cell>
 <cell>6</cell>
 <cell>3</cell>
 <cell>12</cell>
 <cell>0</cell>
 <cell>2</cell>
 <cell>0</cell>
 <cell>0</cell>
 <cell>0</cell>
 <cell> </cell>
 <cell> </cell>
</row>
<row>
 <cell>6</cell>
 <cell>5</cell>
 <cell>7</cell>
 <cell>1</cell>
 <cell>1</cell>
 <cell>1</cell>
 <cell>6</cell>
 <cell>1</cell>
 <cell> </cell>
 <cell> </cell>
</row>
<row>
 <cell rows="2">25</cell>
 <cell>5</cell>
 <cell>3</cell>
 <cell>8</cell>
 <cell>1</cell>
 <cell>1</cell>
 <cell> </cell>
```

```
  <cell> </cell>
  <cell> </cell>
  <cell> </cell>
  <cell> </cell>
 </row>
 <row>
  <cell>5</cell>
  <cell>5</cell>
  <cell>5</cell>
  <cell>0</cell>
  <cell>0</cell>
  <cell> </cell>
  <cell> </cell>
  <cell> </cell>
  <cell> </cell>
  <cell> </cell>
 </row>
 <row>
  <cell rows="2">16</cell>
  <cell>4</cell>
  <cell>3</cell>
  <cell>5</cell>
  <cell>1</cell>
  <cell>0</cell>
  <cell> </cell>
  <cell> </cell>
  <cell> </cell>
  <cell> </cell>
  <cell> </cell>
 </row>...<row>
  <cell>The whole number of armed pikes.</cell>
  <cell>The quadrate, or square roote.</cell>
  <cell>Per ranke to march by:</cell>
  <cell>Rankes how many.</cell>
  <cell>Remaines of rankes.</cell>
  <cell>Maniples, or partes.</cell>
  <cell>Remaines of pikes by ranke.</cell>
  <cell>The whole ouer plus of remaines.</cell>
  <cell>How many to march in rankes of remaine.</cell>
  <cell>The rankes to impale by.</cell>
  <cell>The number of shot that impale.</cell>
 </row>
 <trailer>Here endeth the Tables of Battels in proportion of equalitie, or the Battels
of due square of men: that is, how many rankes, so many men by ranke; or how
ma<lb rend="hidden" type="hyphenInWord"/>ny rankes, so many files.</trailer>
</table>
```



The Table of Battels in proportion of equalitie.

1	2	3	4	5	6	7	8	9	10	11
36	6	3	12	0	2	0	0	0		
	6	5	7	1	1	1	6	1		
25	5	3	8	1	1					
	5	5	5	0	0					
16	4	3	5	1	0					
	4	5	3	1	1	1	4	1		
9	3	3	3	0						
	3	5	1	4	36	4				
4	2	3	1	1						
	2	5	0	0						
2	1	0								
	1	0								
The whole number of armed pikes.	The quadrate, or ſquare roote.	Per ranke to march by:	Rankes how many.	Remaines of rankes.	Maniples, or partes.	Remaines of pikes by ranke.	The whole ouer plus of remaines.	How many to march in rankes of remaine.	The rankes to impale by.	The number of hoot that impale.

Here endeth the Tables of Battels in proportion of equalitie, or the Battels of due ſquare of men: that is, how many rankes, ſo many men by ranke; or how many rankes, ſo many files.

Analysis of the converted TEI files

Once the files are converted to TEI P5 XML, then additional investigation and statistical analysis is possible. For example, Table 3 lists the frequency of all elements in the corpus. Fairly predictably the ten most frequent elements are <hi> (39911344), <lb> (21300024), <l> (8068582), <p>

(7640775), <gap> (4890337), <desc> (4890207), <note> (4231424), <pb> (3400530), <item> (2687551), and <cell> (2404536). This is fairly predictable because the top four elements like <hi>, <lb>, <l>, and <p> are either structural in nature or, like <hi> are used in an extremely general purpose manner. Indeed there is a gap of 2.75 million instances between the last of these <p> and the next closest most frequently used element, <gap>.

Element	Occurrences
ab	7
abbr	491349
add	1632
argument	53317
author	79594
availability	40188
back	17301
bibl	231745
biblFull	40188
body	65619
byline	5481
cell	2404536
choice	4013
closer	72814
date	117561
dateline	27200
desc	4890207
div	1040318
edition	770
editionStmt	770
editorialDecl	40188
encodingDesc	40188
epigraph	24488
expan	4013
extent	40165
figDesc	15338
figure	69257
fileDesc	40188
floatingText	22065
front	30221
fw	326
gap	4890337
group	1794

head	1233568
hi	39911344
idno	126674
item	2687551
keywords	37441
l	8068582
label	470464
language	11791
langUsage	11791
lb	21300024
lg	543524
list	277870
note	4231424
notesStmt	40150
opener	43463
p	7640775
pb	3400530
postscript	2724
profileDesc	37578
projectDesc	40187
ptr	307
publicationStmt	80376
publisher	80381
pubPlace	80470
q	411602
ref	3402
row	532734
salute	35951
seg	112225
seriesStmt	11793
signed	62800
sourceDesc	40188
sp	1106459
speaker	1101703
stage	143568
table	28516
TEI	40188
teiHeader	40188
term	80022
text	45345

textClass	37521
title	113037
titleStmt	80376
trailer	45797
unclear	4022

Table 3 Frequency of TEI elements in EEBO after conversion from TCP to TEI P5

The least frequent tags tell us more about the conversion: <choice> (4013), <expan> (4013), <ref> (3402), <postscript> (2724), <group> (1794), <add> (1632), <edition> (770), <editionStmt> (770), <fw> (326), <ptr> (307), and lastly <ab> (7). It is expected that <choice> and <expan> appear the same number of times because EEBO TCP always uses

```
<ABBR EXPAN="secundum">sm</ABBR>
```

form of markup for abbreviations which gets transformed into

```
<choice>
 <abbr>sm</abbr>
 <expan>secundum</expan>
</choice>
```

and does not use the TCP <EXPAN> element at all. This tells us that <choice> and <expan> will only appear as a result of the conversion to TEI P5, but this is to be expected since <choice> did not exist when TCP started. It is interesting to note that EEBO TCP does not use equivalents of <orig> or <reg> in their markup.

The density of markup can be crudely tested by plotting file size (in bytes on disk) against number of XML elements. Figure 1 shows that progression is not simply linear, though this may be distorted by outliers of large simple texts and short complex texts.

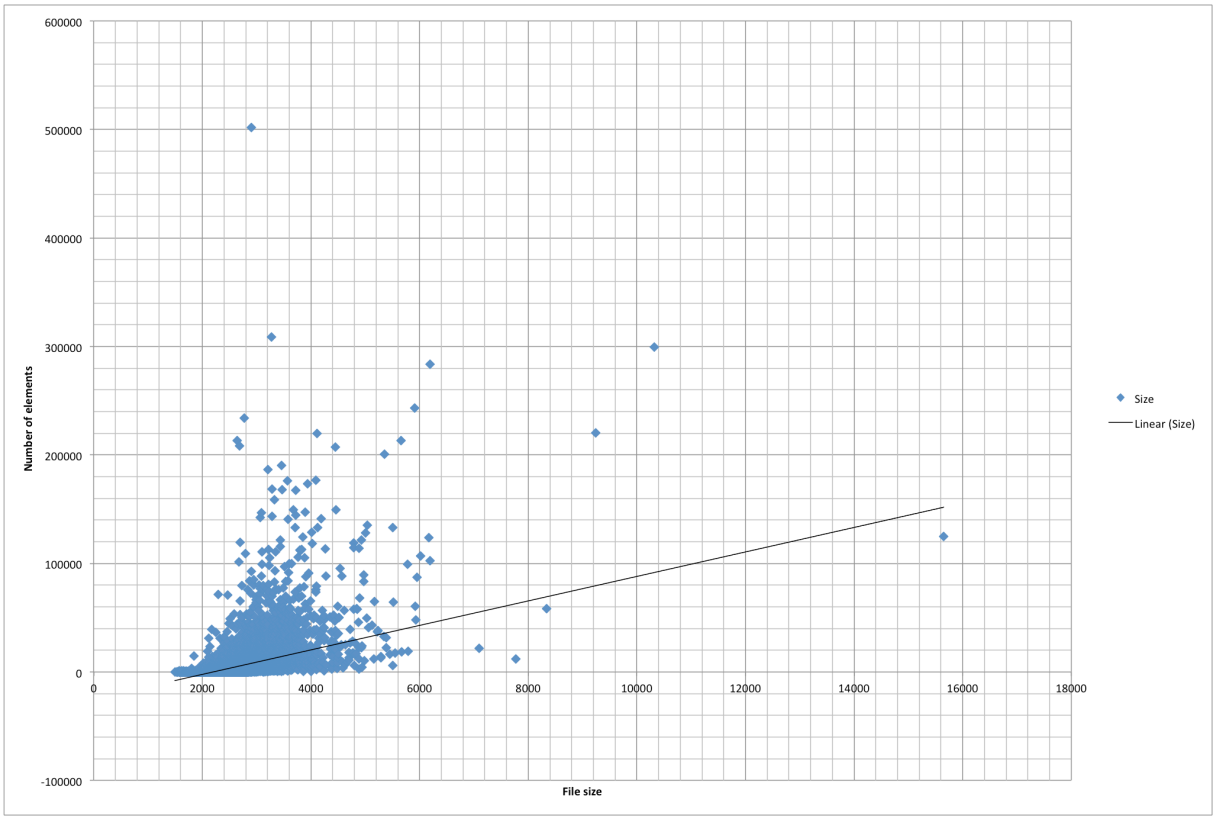


Figure 1 Plot of filesize vs number of elements

The uses of certain attributes, notably @*type* or @*rend* can also be routes into understanding the EEBO TCP corpus. There are over 8700 distinct values for the @*type* attribute, so they can't all be displayed here. Table 4 shows the top twenty @*type* values, its frequency, and a list of elements it was used on.

Value	Occurrences	Elements
hyphenInWord	21281924	lb
milestone	588137	note
chapter	155977	div
part	142597	div list head lg q
section	115562	div
STC	46298	idno
BIBNO	40188	idno
TCP	40188	idno
sub	39249	head q
poem	32693	div lg

letter	30280	div floatingText head
title_page	29921	div
entry	25507	div
subpart	19602	div lg q
text	19058	div
subsection	17449	div
recipe	14508	div
question	13677	div
verse	11750	div lg
Psalm	11543	div

Table 4 Twenty most frequent @*type* values

In this case we can see legacy marks of the conversion in the first two @*type* values: 'hyphenInWord' and 'milestone' as @*type* attribute values were added by the conversion. Values such as 'STC', 'BIBNO', and 'TCP' are used in the metadata and appear in every file. That 'chapter', 'part', and 'section' are the most frequent appearing on `<div>` is indicative of the nature of the corpus. That there are 155977 `<div>` with a @*type* of 'chapter' compared with only 14508 marked as 'recipe' is informative. (That there are this many marked as 'recipe' is also interesting.)

The @*rend* value, indicating how this part of the source text was rendered in the original, has only forty-five distinct values (Table 5).

Value	**Occurrences**	**Elements**
....	1	gap
.....	3	gap
.....	4	gap
.....	1	gap
.....	1	gap
above	1447	hi
AngloSaxonType	11	hi
below	83	hi
blackletterType	935	p hi head lg div
block	3	q
bold	767	hi
centerJustify	20	div lg
decorInit	112106	seg
deleted	2	seg

FrakturType	1	hi
fullStops	2	gap
greek-tai-lig	5	seg
greek-xi	7	seg
hidden	21281924	lb
indent	424	l p lg
inline	1843	q
invert	6	div
IrishType	3	hi
italic	1120	p l q lg div hi
large	37	p hi
ligature	32	seg
manuscript	42	hi
margAsterisks	5	q
margDblQuotes	30	hi q
margQuotes	18144	hi q div p sp
margSglQuotes	141	hi q
maze	1	lg
none	3400530	pb
onBlank	5	add seg
rightJustify	405	hi seg cell l
roman	464	p q hi head div opener
rotateClockwise	4	div
rotateCounterclockwise	2	div
small	16495	p hi cell l div
smallCaps	21	hi
stage	54	hi
strikethrough	10	cell
sub	1073	hi
sup	647987	hi
variantByzantineglyph	2	seg

Table 5 Values for @*rend* attribute

The most frequent @*rend* values are also indicative of the conversion process. A @*rend* value of 'hidden' appears 21281924 times and always on `<lb>` (which appears 21300024 in total). This is because the conversion process looks at the text nodes in the document and converts the TCP use of | for an in-word hyphen creating a linebreak, and replaces it with:

```
<lb rend="hidden" type="hyphenInWord"/>
```

thus creating many @*rend* values of 'hidden' and @*type* values of 'hyphenInWord'. The next most frequent @*rend* value is 'none' used on <pb> which is created by the conversion process when a TCP <PB> element had a @*REF* attribute. This is converted into a TEI @*facs* attribute and the @*rend* value set to 'none'. After these two the most frequent, and possibly more meaningful, @*rend* values are 'sup' (647987) and 'decorInit' (112106). The first of these is a direct correspondence to superscript text in the original EEBO TCP corpus because the TCP <SUP> element has been converted to the standard TEI use of <hi> with a @*rend* value of 'sup'. The 'decorInit' value is used to indicate the presence of a decorative initial.

Rendering the TCP texts

With 40000 TCP texts now in a form (more or less) compliant with the TEI P5 Guidelines, we now use existing TEI tools to process them, including analysis and display. We decided to experiment with rendering the texts in a form suitable for modern readers, using the family of XSL stylesheets (http://www.tei-c.org/Tools/Stylesheets/) developed largely at Oxford. We considered three alternatives:

1. Making ePub files which can be viewed using devices such as an Apple iPad, and converted to the form used by Kindle;
2. Transforming the XML to LaTeX code, and using the TeX typesetting system (in practice, the XeTeX variant, which works with Unicode internally) to generate PDF;
3. Making ISO/IEC 29500 Open Office XML documents (the format of Microsoft Word .docx files) from the XML texts.

The results are shown below. As a variant of the ePub version of the digital text, we also looked at generating fixed layout ePub in which the page facsimiles[4] and the rendered text are presented side by side in a double page spread.

[4] We are grateful to ProQuest for permission to use the facsimile images in this paper.

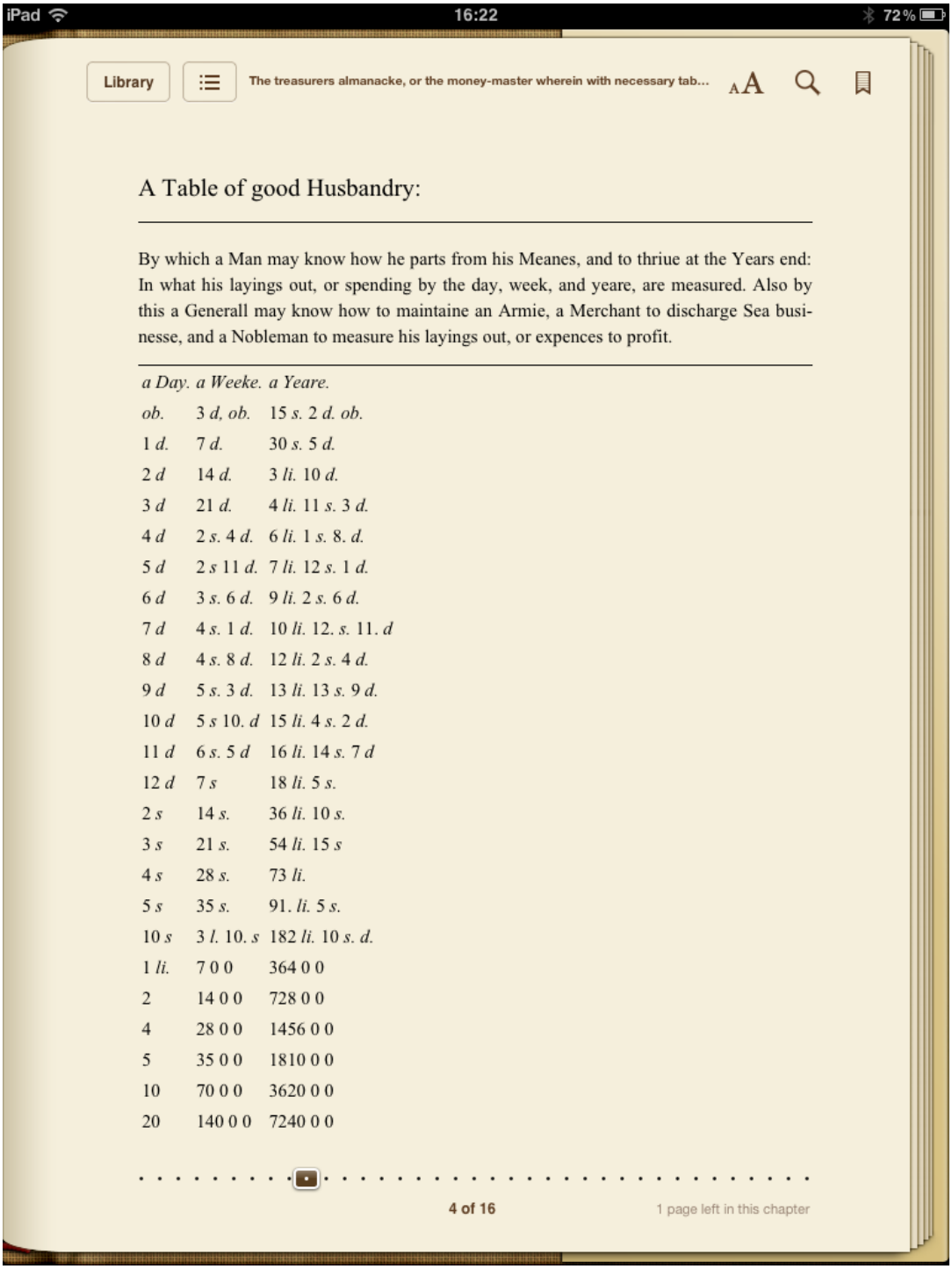


Figure 2 Document 1 as ePub (in Apple iBooks)

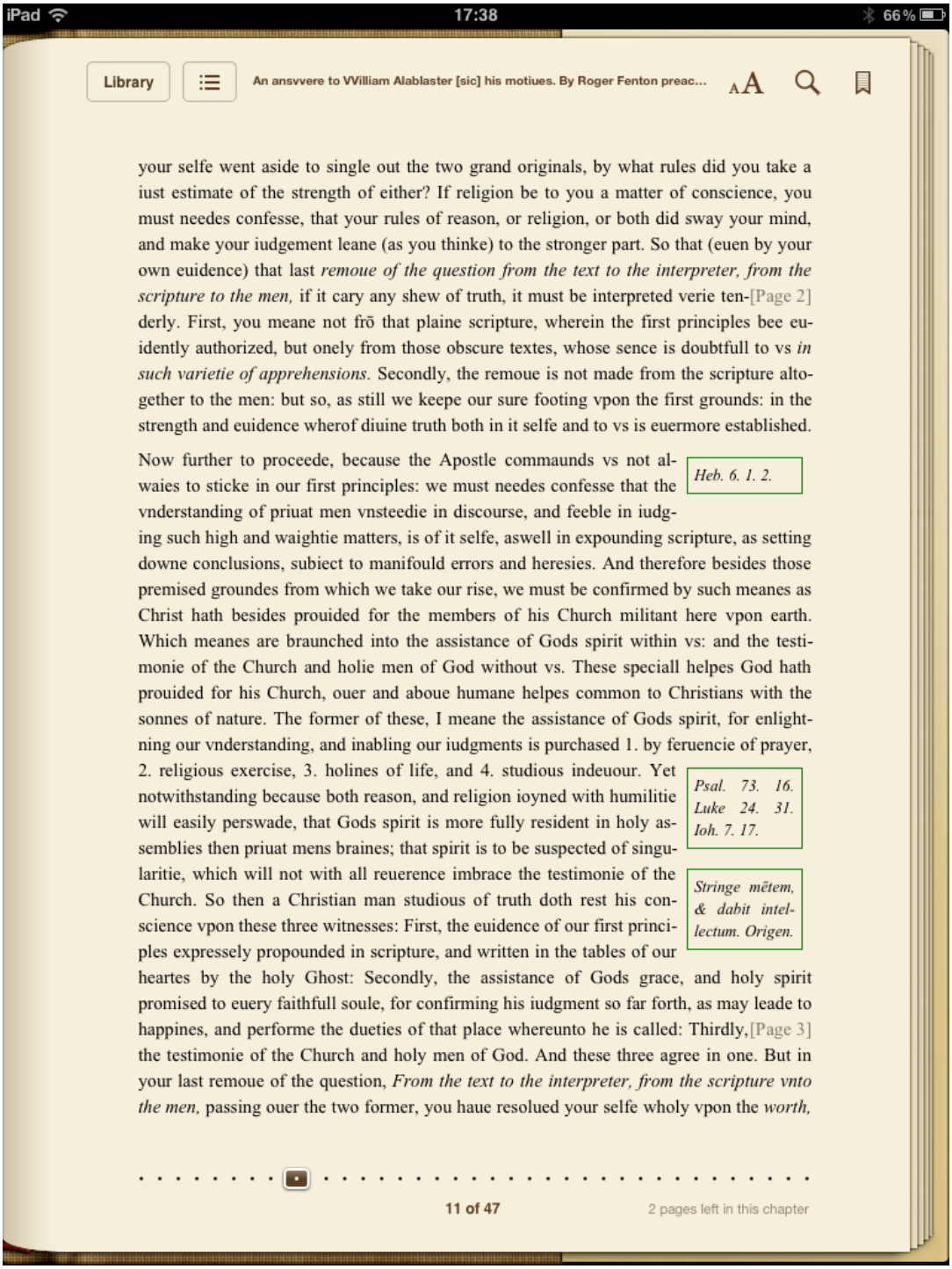


Figure 3 Document 2 as ePub (in Apple iBooks)

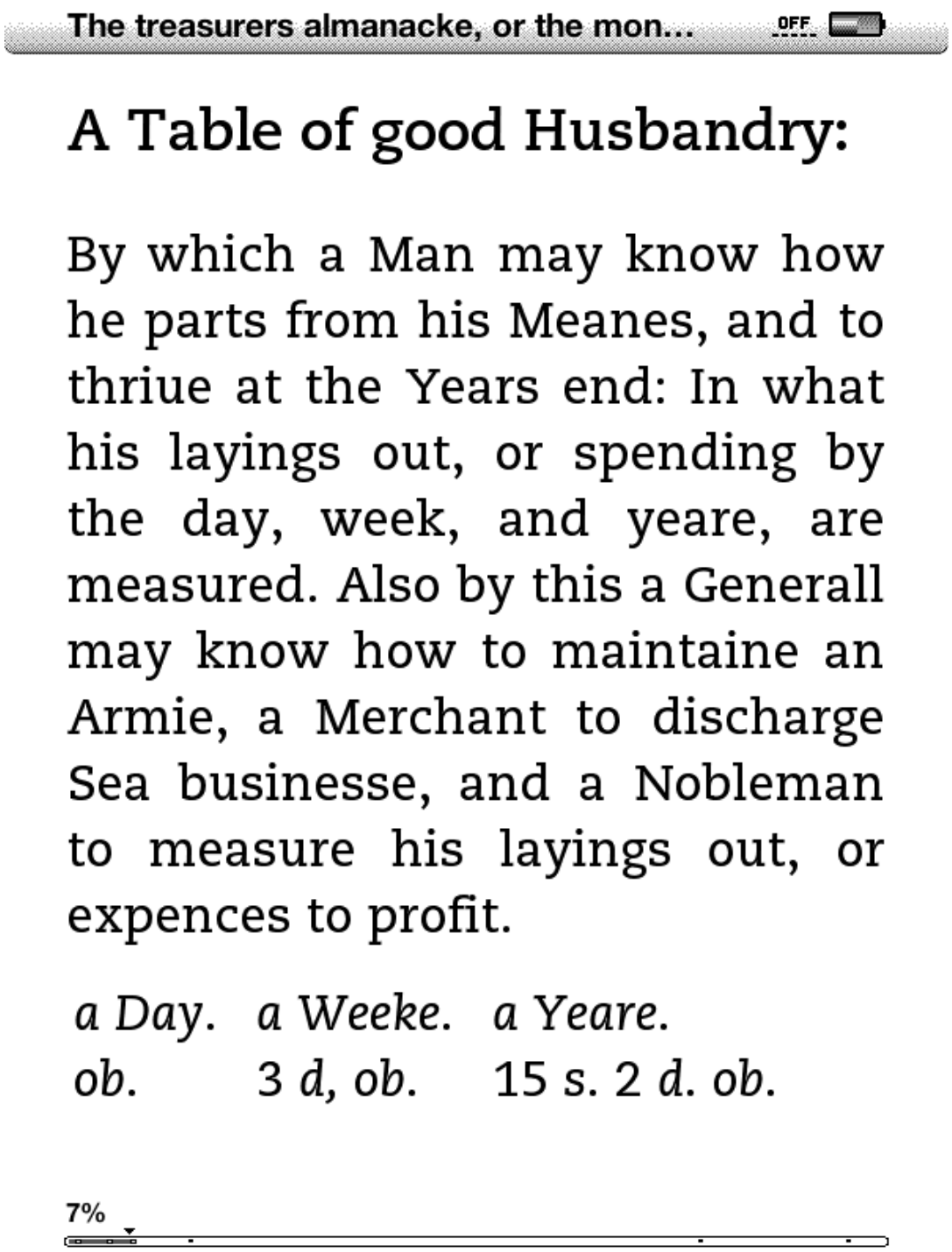


Figure 4 Document 1 on Kindle

to vs is euermore established.

Now further to proceede, because the Apostle commaunds

Heb. 6. 1. 2.

vs not alwaies to sticke in our first principles: we must needes confesse that the vnderstanding of priuat men vnsteedie in discourse, and feeble in iudging such high and waightie matters, is of it selfe, aswell in expounding scripture, as setting downe conclusions, subiect to manifould errors and heresies. And therefore besides those premised groundes from which

15%

Figure 5 Document 2 on Kindle

1 A Table of good Husbandry:

By which a Man may know how he parts from his Meanes, and to thriue at the Years end: In what his layings out, or spending by the day, week, and yeare, are measured. Also by this a Generall may know how to maintaine an Armie, a Merchant to discharge Sea businesse, and a Nobleman to measure his layings out, or expences to profit.

a Day.	*a Weeke.*	*a Yeare.*
ob.	3 *d, ob.*	15 *s.* 2 *d. ob.*
1 *d.*	7 *d.*	30 *s.* 5 *d.*
2 *d*	14 *d.*	3 *li.* 10 *d.*
3 *d*	21 *d.*	4 *li.* 11 *s.* 3 *d.*
4 *d*	2 *s.* 4 *d.*	6 *li.* 1 *s.* 8. *d.*
5 *d*	2 *s* 11 *d.*	7 *li.* 12 *s.* 1 *d.*
6 *d*	3 *s.* 6 *d.*	9 *li.* 2 *s.* 6 *d.*
7 *d*	4 *s.* 1 *d.*	10 *li.* 12. *s.* 11. *d*
8 *d*	4 *s.* 8 *d.*	12 *li.* 2 *s.* 4 *d.*
9 *d*	5 *s.* 3 *d.*	13 *li.* 13 *s.* 9 *d.*
10 *d*	5 *s* 10. *d*	15 *li.* 4 *s.* 2 *d.*
11 *d*	6 *s.* 5 *d*	16 *li.* 14 *s.* 7 *d*
12 *d*	7 *s*	18 *li.* 5 *s.*
2 *s*	14 *s.*	36 *li.* 10 *s.*
3 *s*	21 *s.*	54 *li.* 15 *s*
4 *s*	28 *s.*	73 *li.*
5 *s*	35 *s.*	91. *li.* 5 *s.*
10 *s*	3 *l.* 10. *s*	182 *li.* 10 *s. d.*
1 *li.*	7 0 0	364 0 0
2	14 0 0	728 0 0
4	28 0 0	1456 0 0
5	35 0 0	1810 0 0
10	70 0 0	3620 0 0
20	140 0 0	7240 0 0
30	210 0 0	10860 0 0
40	280 0 0	14480 0 0
50	350 0 0	18100 0 0
100	700 0 0	36200 0 0
100	3500 0 0	181000 0 0
1000	7000 0 0	362000 0 0

Account from 1 li. to 1000 li. by the Weeke, and not by the Day: For if you count by the Day, 1 li. a Day is 365 li. a Yeare, and 1000 li. a Day is 365000 li. a Yeare, and so of the rest from 1 li. to a 1000.

2 A Table of simple Interest, at 8 *li.* &c.

THis Table, which needs no Description; first, readily sheweth the Interest of any Summe from 2 *s.* 6 *d.* to 900 *li.* for any time within a Yeare. Secondly, it is most vsefull and expedient for the casting vp of Interest arere, or behind, vpon forfeited Bonds, or accruing vpon seuerall payments of money. Thirdly, is right necessary as well for the Borrower as the Lender, whereby the one may know how much Interest be should pay, and the other what to receiue; and consequently neither of them doe, nor suffer iniury therein.

Summes.	*A Yeare*	6. *Mon.*	3. *Mon.*	1. *Mon:*

Figure 6 Document 1 as PDF, generated from LaTeX

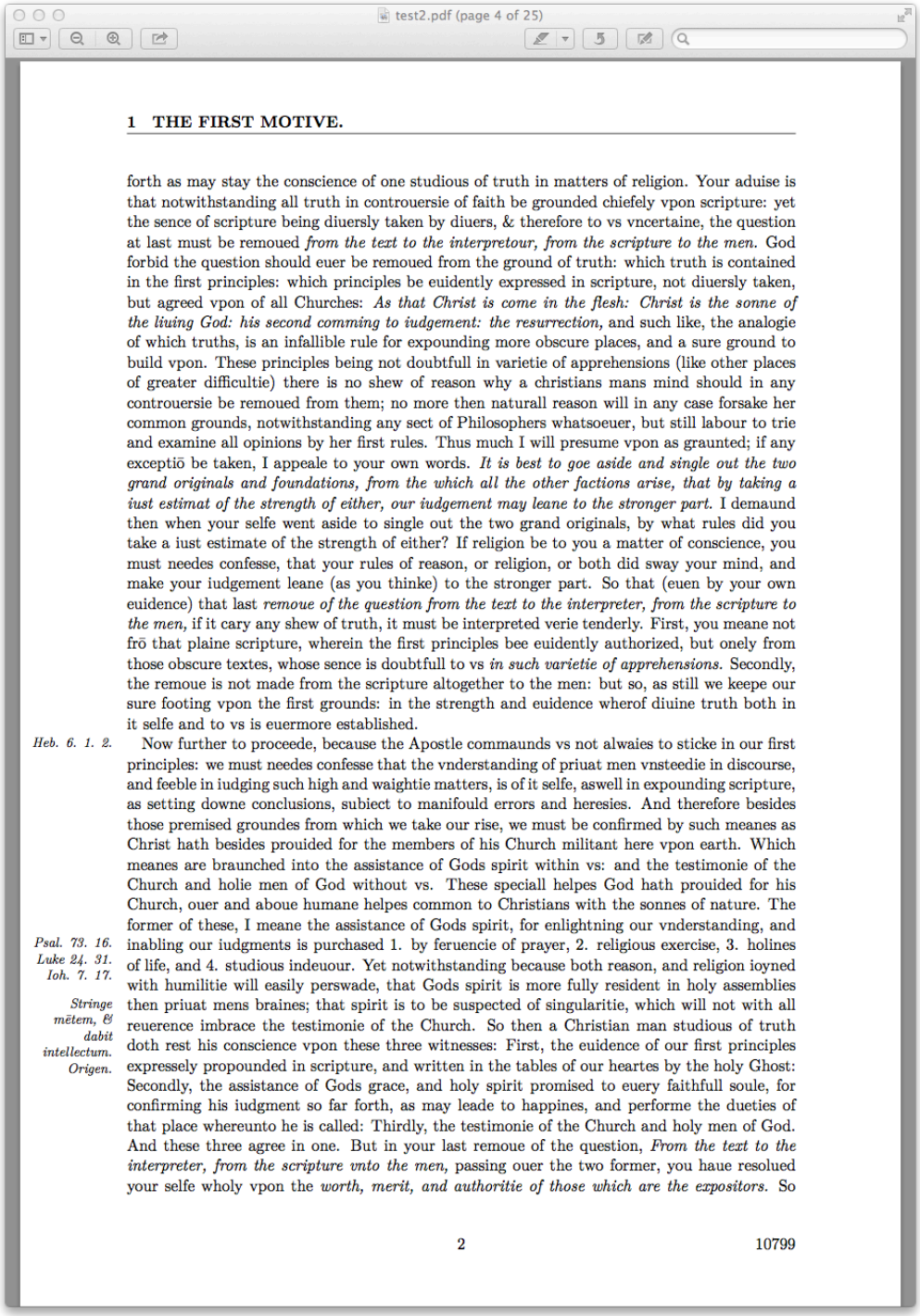
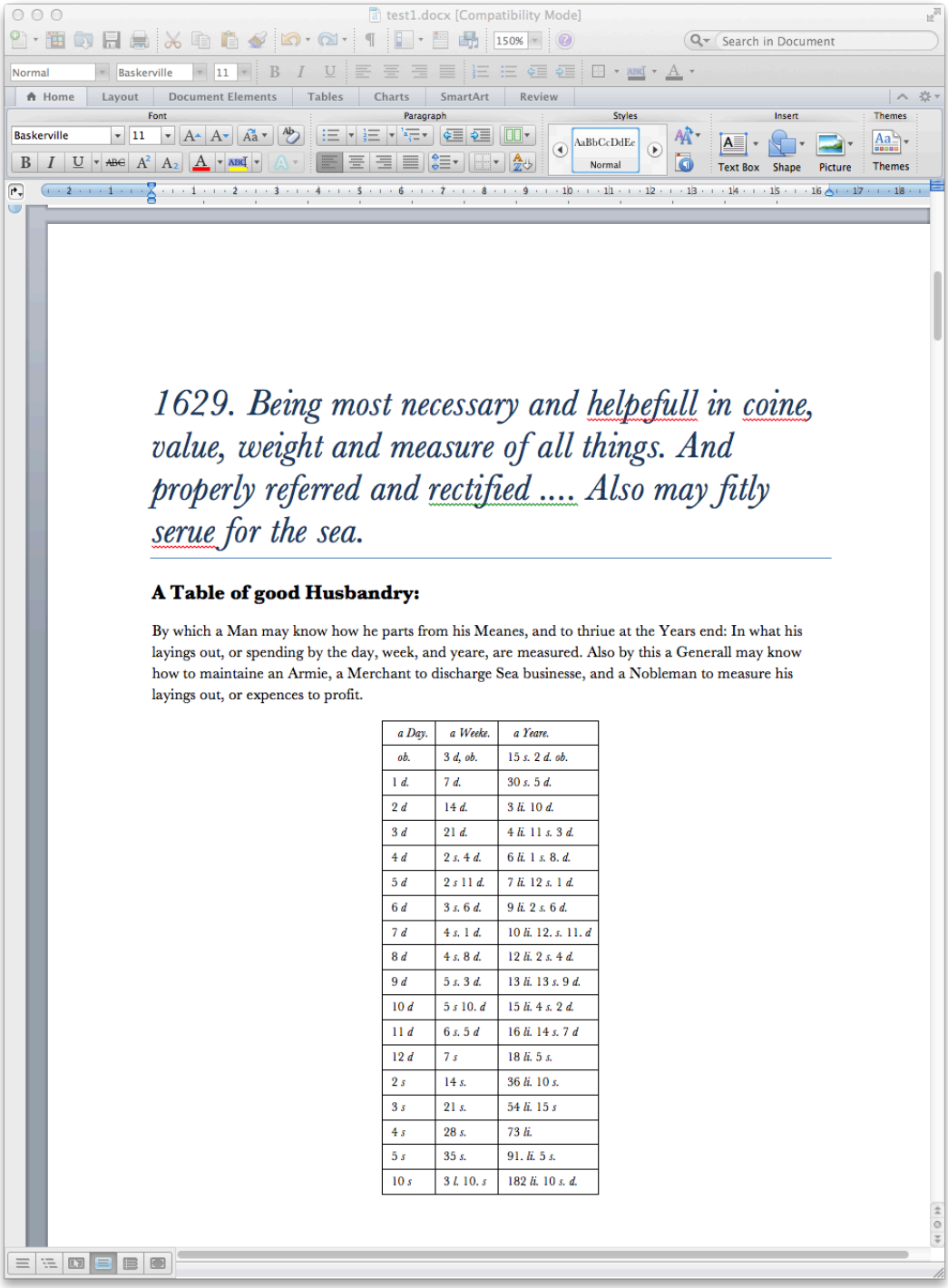


Figure 7 Document 2 as PDF, generated from LaTeX



1629. Being most necessary and helpefull in coine, value, weight and measure of all things. And properly referred and rectified .... Also may fitly serue for the sea.

A Table of good Husbandry:

By which a Man may know how he parts from his Meanes, and to thriue at the Years end: In what his layings out, or spending by the day, week, and yeare, are measured. Also by this a Generall may know how to maintaine an Armie, a Merchant to discharge Sea businesse, and a Nobleman to measure his layings out, or expences to profit.

a Day.	*a Weeke.*	*a Yeare.*
ob.	3 *d, ob.*	15 *s.* 2 *d. ob.*
1 *d.*	7 *d.*	30 *s.* 5 *d.*
2 *d*	14 *d.*	3 *li.* 10 *d.*
3 *d*	21 *d.*	4 *li.* 11 *s.* 3 *d.*
4 *d*	2 *s.* 4 *d.*	6 *li.* 1 *s.* 8. *d.*
5 *d*	2 *s* 11 *d.*	7 *li.* 12 *s.* 1 *d.*
6 *d*	3 *s.* 6 *d.*	9 *li.* 2 *s.* 6 *d.*
7 *d*	4 *s.* 1 *d.*	10 *li.* 12. *s.* 11. *d*
8 *d*	4 *s.* 8 *d.*	12 *li.* 2 *s.* 4 *d.*
9 *d*	5 *s.* 3 *d.*	13 *li.* 13 *s.* 9 *d.*
10 *d*	5 *s* 10. *d*	15 *li.* 4 *s.* 2 *d.*
11 *d*	6 *s.* 5 *d*	16 *li.* 14 *s.* 7 *d*
12 *d*	7 *s*	18 *li.* 5 *s.*
2 *s*	14 *s.*	36 *li.* 10 *s.*
3 *s*	21 *s.*	54 *li.* 15 *s*
4 *s*	28 *s.*	73 *li.*
5 *s*	35 *s.*	91. *li.* 5 *s.*
10 *s*	3 *l.* 10. *s*	182 *li.* 10 *s. d.*

Figure 8 Document 1 as Word document

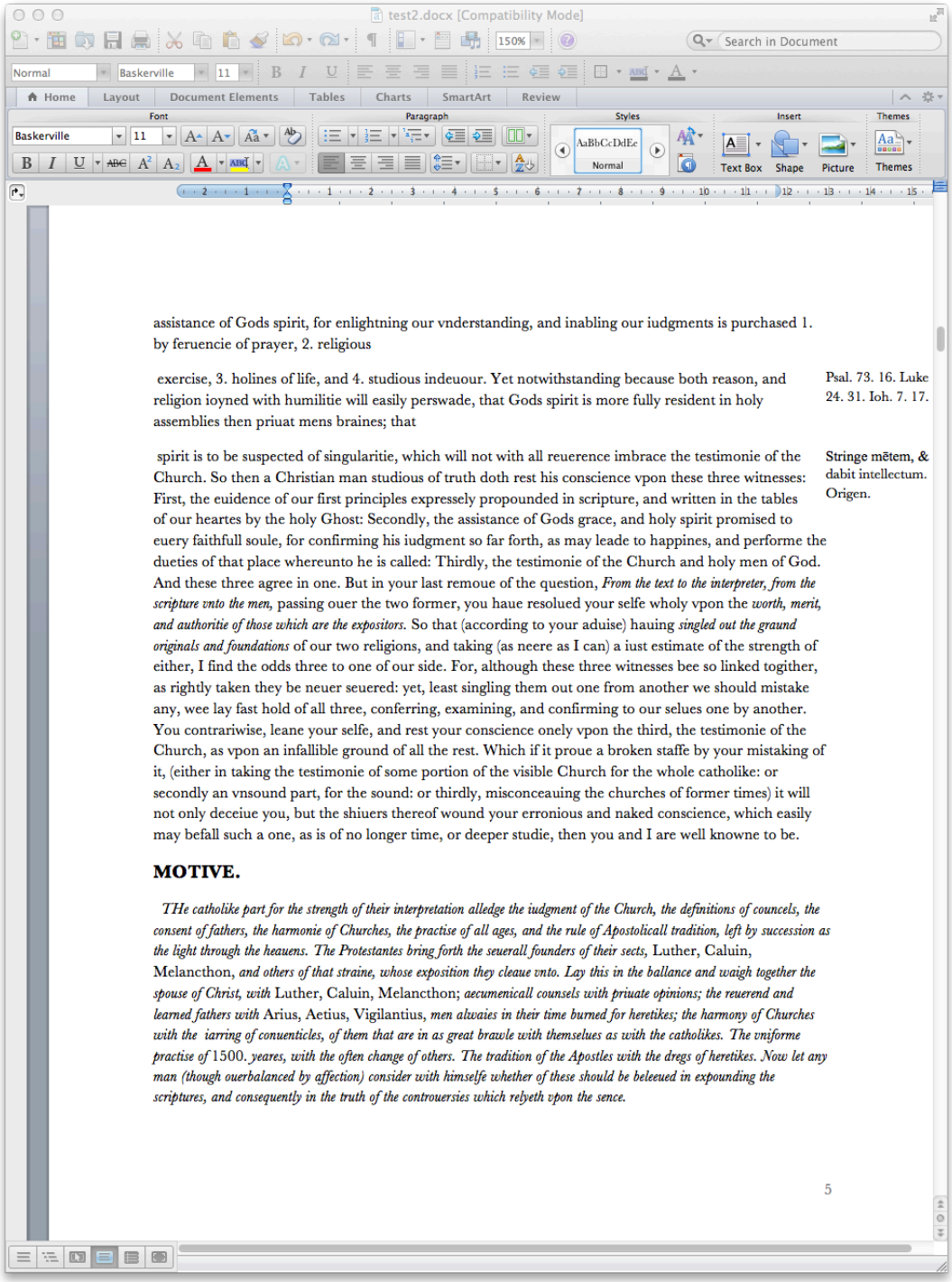


Figure 9 Document 2 as Word document

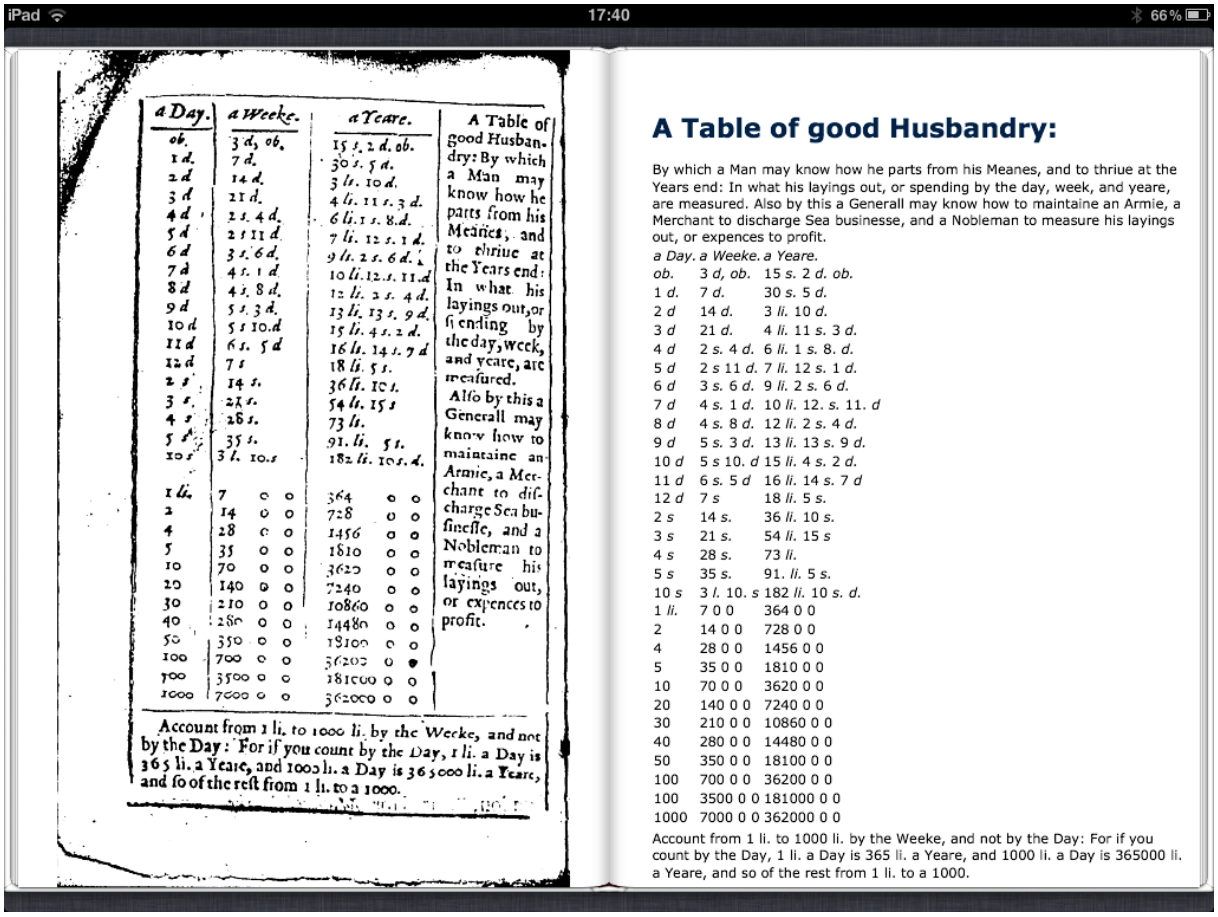


Figure 10 Document 1, fixed layout ePub with alternate facsimile and rendered text (in Apple iBooks)

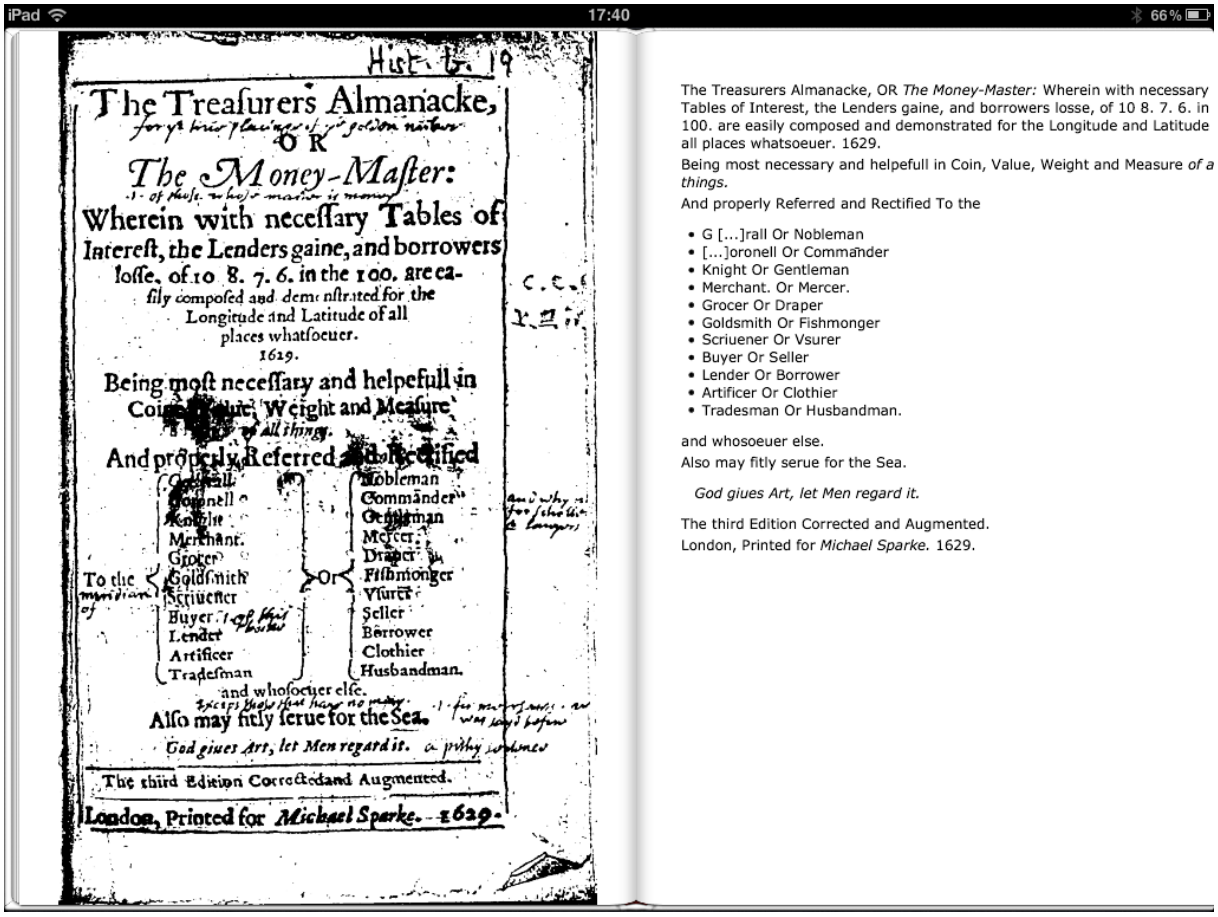


Figure 11 Document 1, fixed layout ePub with alternate facsimile and rendered text (in Apple iBooks)

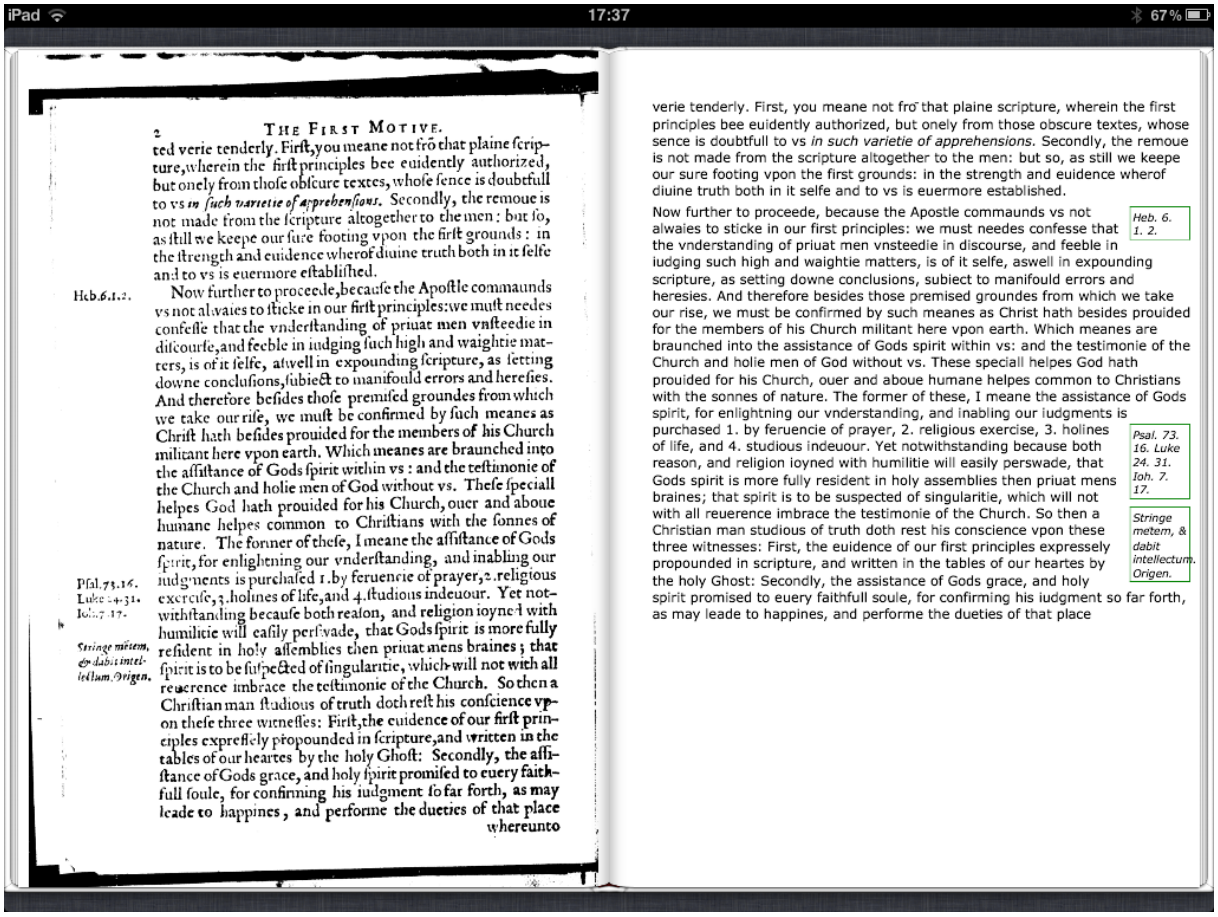


Figure 12 Document 2, fixed layout ePub with alternate facsimile and rendered text (in Apple iBooks)

Making fixed layout ePub?

How did we make fixed layout ePub automatically? The key principle is that we alternate pages of text with pages which contain nothing but a facsimile page image. Each page is a separate HTML file, and the ePub is instructed to use fixed layout (Elizabeth Castro's very helpful guides at http://www.elizabethcastro.com/epub/ explain the process). The page breaks in the TEI file are marked with a <pb> element), and these are used to divide the output into separate files; as they come from TCP, the breaks just have a @*REF* attribute containing a number; this has to expanded to be the name of a real file, (STC-24212-1261_07-p3b.png) and turned into a @*facs* attribute . Care has to be taken when a page break occurs inside a paragraph or line of verse, to ensure the context is closed in one file and reopened in another.

The page image facsimiles are taken from the EEBO web site (http://eebo.chadwyck.com/), which offers the facility to download a TIFF version of the page being viewed (obviously, these are copyrighted images, and permission must be obtained to re-publish them). These are usually images of double-page spreads, and are at quite a high resolution. For the purposes of fixed format epub, we want them at a fixed size of 1000x1700 (padded with white if necessary), and one image per page. For demonstration purposes, we cropped and divided the double page images by hand, and then automated the resizing with an ImageMagick (http://www.imagemagick.org/) script:

```
convert -resize 1000x1700
 -background white -gravity center -extent 1000x1700
```

Suggestions for future work

One area of future work is to improve the conversion. A possible benefit of having analysed the *@type* values, for example, is that we can notice that although a value of 'zodiac_sign' is used 23 times on `<div>` 'zodiacal_signs' is used only once and so is probably a typo. Some work has already gone into separating the wide variety of 'poem' types into a *@type* of 'poem' and a *@subtype* of whatever other classification was given. Other candidates similar separations into type and subtype could include 'abstract', 'account', 'act', 'advert', and hundreds more. In cases such as these there are values such as: 'advert_for_book', advert_for_books', 'advert_for_cider_etc.', which could be broken into a *@type* of 'advert' and a *@subtype* of 'for_books' fairly easily. Similarly there are 72 divisions with a type of 'advert' and 43 with a type of 'adverts' and it might usefully be explored whether these should be rationalised (or whether the plural is serving a useful grouping purpose).

Another area for future might include the marking of every orthographic word, with ID numbers, in order to facilitate pointing into the corpus by future scholars. Moreover, comparing these words against corpora of Early Modern English might allow automatic part of speech tagging (with the usual caveats of historical spellings).

The rendering into modern formats is well advanced, but there remain serious problems of consistency across different formats, and graceful degradation to the less capable formats. The very extensive use of the margin for adding extra material in these early books presents a notable challenge for some formats, as do some of the very complex table layouts. Such layout issues ignore, of course, the even more obvious problem of special characters and one-off sigla which defeat the optimistic assumptions of Unicode.

Already many projects are using EEBO texts, converted to TEI, as a starting point to undertake another layer of research (for example metrical analysis). The challenge is not just to make all these texts freely available to scholars as a consistent set of XML files (as foreseen by the TCP project), but to do so in a way which makes enhancement (extra tagging and analysis), addition (eg of mathematics and non-English script), and correction (filling in gaps in transcription) a process of contribution towards a single, evolving, community-maintained resource rather than a series of niche variants from a common source. To do this will need the cooperation of an international community, and openly and publicly developed tools and infrastructure. The University of Oxford has a long history in supporting the TEI through partnership and technological development, and as an EEBO TCP partner it hopes to contribute to efforts to exploit the EEBO TCP corpus.

Further Reading

[1] Burnard, Lou; O'Brien O'Keeffe, Katherine; Unsworth, John (eds.) *Electronic Textual Editing*. MLA: New York, 2006. http://www.tei-c.org/About/Archive_new/ETE/

[2] Cummings, James. 'The Text Encoding Initiative and the Study of Literature'. In *A Companion to Digital Literary Studies*; Siemens, Ray; Schreibman, Susan (Eds.) Blackwell Publishing: Oxford, 2007; 451-476. http://www.digitalhumanities.org/companionDLS/

[3] Pytlik Zillig, Brian. 'TEI Texts that Play Nicely: Lessons from the MONK Project'. *Journal of the Chicago Colloquium on Digital Humanities and Computer Science* , 1.3 (2011) https://letterpress.uchicago.edu/index.php/jdhcs/article/view/81

[4] Pytlik Zillig, Brian. 'TEI Analytics: converting documents into a TEI format for cross-collection text analysis'. *Literary and Linguistic Computing* 2009 24: 187-192. http://llc.oxfordjournals.org/content/24/2/187

[5] Welzenbach, Rebecca. "Making the Most of Free, Unrestricted Texts: A first look at the promise of the Text Creation Partnership" MPublishing [Online], (2011) http://hdl.handle.net/2027.42/87997

[6] Wittern, Christian; Ciula, Arianna; and Tuohy, Conal. 'The making of TEI P5', *Literary and Linguistic Computing* 2009 24: 281-296. http://llc.oxfordjournals.org/content/24/3/281