Transcription, Proofing, and Coding Protocols for Primary Works in the

Charles Brockden Brown Electronic Archive[1]

By Mark L.Kamrath, Philip Barnard, Will Dorner, and Amy Giroux

April 21, 2016

I.Textual reproduction, OCR, and accuracy

In December 2006, the Brown project contracted with Aptara to keyboard and encode in XML nearly 1000 texts believed to be Brown's—the number of texts we have acquired to date for keyboarding. Their process combined use of Optical Character recognition (OCR) software and double-keyboarding. The accuracy guaranteed was a minimum of 99.95% (1 error every 2,000 characters), but closer in actuality to 99.99% (or 1 error every 10,000 characters).

Initial transcriptions are based upon print originals (letters) or high resolution TIFF, JPEG, or PDF images. The bulk of the print images were made available to us through ProQuest American Periodical Series Online (see Appendices). The base-level encoding of Brown's texts encompasses fundamental markup, including the requisite tags for header metadata and page-layout tags such as <pb> for page breaks, <p> for paragraphs, and <lb> for line breaks.

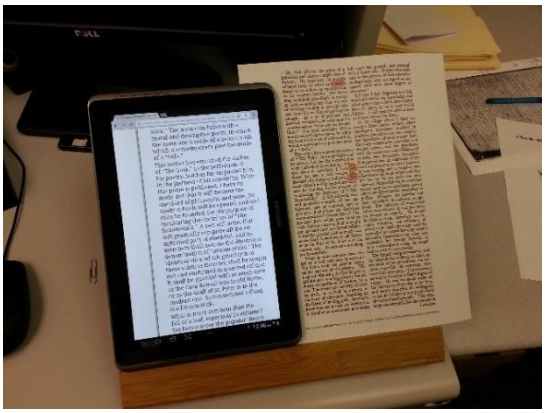
We display texts without correction of errors or regularization, except for obvious date or pagination elements. Original line breaks and hyphenation in both holograph and print materials are retained. In the case of Brown's letters, we retain strikethroughs, underlining, and superscript insertions.

Next, once the initial keyboarding and coding was completed, through a multi-tiered proofing process each of Charles Brockden Brown's individual works (original publication and manuscript texts) are compared side by side with the electronic facsimile renditions by multiple readers. We use a team of graduate students to proofread the first round, and a team of undergraduate students to proof the second round, each noting their findings. Project editors spot-checked proofing for quality control.

The proofing method we have selected and worked to refine involves the following steps:

1. A Samsung Galaxy Tab 2 tablet (project-owned), supported on an angled book stand, is used to display digital renditions from the archive side-by-side with original hardcopy printouts of PDFs or manuscript transcriptions as appropriate (see images below).

[1] For a complete historical description of the editorial and coding steps we took, see "The Charles Brockden Brown Electronic Archive: Mapping Archival Access and Metadata" in *Archive Journal* (2013)




2. A line-by-line comparison of the two texts is intended to reveal any errors or inconsistencies evident in the digital display rendition relative to the original text.

3. Errors or inconsistencies between the two texts are documented in pencil (on the first round) by the discovering reviewer directly on the hardcopy printout of the original. Subsequent runs employ different colored pens or pencils. The use of multiple reviewers affords the highest level of accuracy possible.

Once we have completely proofed the primary texts, a team of graduate students, supervised by project faculty members integrated the changes using TEI standards. Both universal stylistic inconsistencies and text-specific coding errors were addressed on a case-by-case basis directly within the XML by at least two XML-proficient graduate students, referencing the documented printouts one by one. Following a subsequent review to ensure all issues were addressed consistently and completely resolved, amended files were re-indexed into the archive.

II. TEI coding standards and specifications

We initially used TEI Lite (P4) to describe the structure of a text, which allows for greater granularity for users once they have searched within our database of Brown's primary texts. The TEI system also provides for metadata to be attached to the transcription of a text, which allows for faceted searching of our database

After extensive consultation with Syd Bauman and some trial and error with XTF rendering, we decided on the following (header and body) schema for our files:

A TEI <teiHeader> element:

```
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title type="main">Character of Dr. Franklin</title>
      <title type="version">TEI Document</title>
      <author cbb:cert="hybrid">Brown, Charles Brockden</author>
      <respStmt>
```

```
      <resp>transcription and initial TEI coding</resp>
      <orgName>Aptara</orgName>
    </respStmt>
    <respStmt>
      <resp>TEI encoding enhancement</resp>
      <name>Mark L. Kamrath</name>
      <name>Philip Barnard</name>
    </respStmt>
  </titleStmt>
  <editionStmt>
    <edition>TEI P5</edition>
  </editionStmt>
  <publicationStmt>
    <publisher>The Charles Brockden Brown Electronic Archive and Scholarly
Edition</publisher>
    <address>
      <postBox>P.O. 161346</postBox>
      <orgName>University of Central Florida</orgName>
      <settlement>Orlando</settlement>
      <region>FL</region>
      <postCode>32816</postCode>
    </address>
    <idno type="accession">1807-02150</idno>
    <date when="2009-01"/>
    <availability status="restricted">
      <p>Copyright 2009 The Charles Brockden Brown Electronic Archive and Scholarly
Edition.</p>
    </availability>
  </publicationStmt>
  <sourceDesc>
    <biblStruct>
      <analytic>
        <author>Brown, Charles Brockden</author>
        <title>Character of Dr. Franklin</title>
      </analytic>
      <monogr>
        <title level="j">American Register</title>
        <imprint>
          <biblScope type="vol">I</biblScope>
          <date when="1807"/>
          <biblScope type="pp">150-159</biblScope>
        </imprint>
      </monogr>
    </biblStruct>
  </sourceDesc>
</fileDesc>
```

```
    <xi:include href="./cbb_encodingDesc.xml"/>
    <profileDesc>
      <textClass>
        <keywords scheme="#CBB_keywords"><term>essay</term></keywords>
      </textClass>
    </profileDesc>
  </teiHeader>
```

A TEI <text> element:

```
<text type="magazine">
    <body>
      <div>
        <pb n="150" facs="1807-02150-150.jpg"/>
        <head>CHARACTER OF DR. FRANKLIN.</head>
        <p>A just view of the character<lb/>
of Dr. Franklin has probably ne-<lb/>
ver been given by any of his<lb/>
countrymen. While living, the<lb/>
world was divided into passion-<lb/>
ate friends and rancorous ene-<lb/>
mies, and since his death a kind<lb/>

…

those studies which the learned<lb/>
have generally turned from in dis-<lb/>
dain. Respect is due to scholar-<lb/>
ship and science; but the value of<lb/>
these instruments is apt to be over-<lb/>
rated by their possessors; and it<lb/>
is a wholesome mortification, to<lb/>
show them that the work may be<lb/>
done without them.</p>
      </div>
    </body>
  </text>
```

Within the <teiHeader>, the <titleStmt> contains information including the title of the TEI document as well as the various entities responsible for creating it, which could include an original author in addition to a coder or transcriber, indicated by <respStmt>. The <publicationStmt> contains metadata pertaining specifically to the XML document, and might include information about the location in which the XML file is stored and its availability or copyright status. The <sourceDesc>, on the other hand, details the bibliographic information of the source document around which the XML was encoded. The <keywords> element indicates the genre of the text (“essay” in the above example), allowing the user to find it and other texts classified as an “essay” by entering that term as a search query in XTF. Other genre categories

we use include “novel,” “poem,” “letter,” and more. As a number of Brown’s writings may be classified as being of more than one genre, e.g., both “letter” and “poem,” we identify as such in the metadata.

In the <body> element is contained the entirety of the textual content. Features we have marked up for representation include structural elements such as page, column, and line breaks, paragraphs, and features specific to individual genres—stanzas, letter openers and closers, and footnotes, as examples. One of the difficulties we faced was the decision to render the text in a single column (as opposed to the double-column format demonstrated in Brown’s periodical writings). While column breaks have been preserved and indicated within the markup, XTF currently displays the textual flow in a single column due to its default rendering of text in an HTML <table> element.

An additional difficulty occurred when keyboarders marked text as undecipherable or as a <gap> during the initial transcription and encoding process; such text will need to be re-examined against the original copy of the periodical. In other cases, transposed letters in a word or an error on the part of the Aptara keyboarder concerning pagination, for instance, will need close scrutiny and correction.

In all cases, though, original capitalization, spelling, punctuation, and end-line hyphenation in Brown’s letter manuscripts and print texts are left unchanged—we undertake a conservative editorial approach and avoid modernizing any aspect of the text except for typography, since we will not replicate original manuscript font type.

Each entry in the primary bibliographic list of publications by Brown is therefore represented in TEI by a <biblStruct> element, which uniquely identified by the accession number. Each <biblStruct> in turn contains an <analytic> element, in which is found <author> and <title> information about an article, poem, essay, or other portion within a journal itself. Contained in <monogr> is the title of the publication, the editors of the publication if applicable, and the <imprint>, which houses the date and place of publication, as well as volume, issue, and pagination information. A typical entry in the secondary bibliography follows a pattern much like that of the primary bibliography in terms of publications within collections. For books published as individual items, the <analytic> element is excluded.

Each full text by Brown will be contained within an XML file (named by accession number) that adheres to the general TEI P5 infrastructure, beginning with the <teiHeader> element that lays out the bibliographic metadata about the TEI file itself in <fileDesc> as well as about the encoding of the text in <encodingDesc>. The file description statement lays out information about the file's title, edition, and source text, including statements regarding distribution, access, and responsibility. Following the header, the <text> element contains the transcription of the text. Each file's <text> contains at least fundamental page-layout tags, including <p> for paragraph, <lb/> for line break, and <pb> for page break.