Reviving the VPP: an initial report

Author: Lou Burnard

Current position: Independent Researcher

Affiliation: Unaffiliated

Host institution: NUIG

Mentor(s): Justin Tonra

Period of stay: 26/3 to 13/4; 20/7 to 27/7

The VPP

As noted in the project proposal, the original Victorian Plays Project (VPP) was one of the few serious attempts to digitize in a systematic way a significant area of 19th century British culture: the Victorian Theatre. The VPP collected a database of information describing various aspects of a major collection of theatrical source materials, Thomas Hailes Lacy's "Acting Editions", originally published between 1850 and 1873. The project also produced benchmark digital editions of a substantial number of these texts in PDF format. Initially funded by the UK's AHRC for two years (2005-2007), the project came to a premature end following the untimely decease of its PI, Professor Richard Pearson, though the bulk of the material produced was archived at NUIG, and remains available there.

Research Questions and Methodology

I initially drew up a series of long term goals for the revival of the project, as follows:

- collate available information about the Lacy Acting edition (bibliographic details, provenance, performances, etc.)
- set up and publish an XML database of this information
- explore alternative methods of creating a substantial set of TEI XML transcriptions, by automatic or semi-automatic conversion from existing file formats (PDF, TIFF, HTML etc.)
- create a pilot set of 100-200 titles, conforming to Dracor standards
- document methods and tools used
- promote engagement and experimentation with the datasets by the scholarly community
- review the history of the project as a case study in DH

Not all of this could be achieved within the timescale of 8-12 weeks originally proposed for the CLS Infra fellowship, of course. My initial focus was on the following:

- identifying and locating whatever resources remained available from the original VPP
- discussing methods used in the original VPP with the researchers involved
- reviewing and enhancing existing online bibliographic information about the Lacy Acting Edition to create a new online database
- experimenting with methods of creating TEI XML versions for a sample of the existing materials

Research Visits and Outcomes

My first visit to Galway provided the context for this initial work. I was able to discuss technical aspects of the web-based CMS which hosts the surviving version of the VPP and its associated database with local support staff at NUIG, and ascertain that no other related digital resources were available there. My host at NUIG and I discussed (via Zoom) the initial project's *modus operandi* with Dr Kate Mattacks, who had worked on it as a research assistant from the beginning. Dr Mattacks also put me in touch with the digitization experts at Birmingham Library who had been responsible for creating the original page images of the whole collection. I subsequently obtained from Dr Mattacks copies of the bulk of the TIFF files used as sources for the PDF files archived at NUIG; only about 20 files are missing.

During my visit I was able to profit from excellent working conditions at the Moore Institute, accessing many online resources, and experimenting with a number of relevant tools. I also used the NUIG library, and attended a couple of interesting seminars on related topics. I am particularly grateful to my host and other colleagues at the Moore Institute for their warm welcome.

Outcomes included the following:

A substantially enlarged and revised version of the Lacy database.

I corrected numerous typos and supplied a few omissions. I added links to titles available in digital form elsewhere on the Internet. I also added links to titles for which the Archive of the Lord Chamberlain's Office at the BL holds a manuscript version; digitized versions of these are included in Gale's "19th Century Collections Online" series. The database is currently maintained as an XML file in a Github repository. I plan to continue to enhance it, and to publish it using CETEICEAN or TEI Publisher. A proof-of-concept implementation using CETEICEAN is currently online at https://lb42.github.io/Lacy/lacyCatalogue.html

Identifying occurrences of the same title in different sources was not always straightforward: titles appearing in the Lacy Acting Edition might be cited simply using (some words from) the title, by the volume number in which they appeared (in roman or arabic numerals), by their number within the volume, by their number within the whole Edition, or some combination of these. I experimented with the idea of using a "fingerprint" derived from the words of the title as a means of matching titles automatically, but with only mixed results. I allocated each title

a simple identifier, numbering them in sequence of their first appearance in the Lacy Acting Editions.

At the time of writing, the database contains 1498 titles, corresponding with 100 volumes published between 1850 and 1873. 684 titles have at least one digitized version. 344 titles have a VPP PDF version, backed up by original page images for all but 25 titles. 340 titles are linked to manuscripts preserved in the Lord Chamberlains Office archive. I have so far identified 158 digitized titles available from other sources such as the Internet Archive (these include mostly but not exclusively titles from the Hall Collection of Prompt Books), the HathiTrust, and elsewhere. Only 34 titles of these titles are already available in the VPP. I hope to add more links for digitized versions in other collections as data becomes available.

Bibliographies of Victorian theatre such as Nicoll or Mullin have for the most part followed earlier catalogues in deriving their data primarily from records of performance (playbills etc.) rather than from bibliographic data. Where such data is machine tractable, I plan to include it in the database as well, to provide (for example) an indication of whether or not Lacy was reprinting a play previously published in an older collection such as Dicks. This remains a work in progress, since I have been unable to find any complete online bibliography corresponding with the publications of Nicoll or Mullin.

Experiments in auto-conversion

The texts which the VPP makes available were originally scanned and saved as TIFF files. These were then processed by Abbyy Fine Reader or similar OCR software to produce a plain transcribed version, which was proof read and then saved in PDF format. My understanding is that this PDF was corrected directly using Adobe software, so that the versions now available online probably represent the best available texts as regards accuracy, though they are still not entirely error-free. On the other hand, these corrections in some cases apparently introduced some variation in formatting (for example font size or case changes not present in the original).

I identified a few other projects which are attempting to create TEI-XML versions more or less automatically from PDF sources. Sophisticated machine-learning approaches such as GroBid have been used, for example in the Europarl or RDA projects. But these projects are for the most part dealing with very large quantities – thousands – of fairly consistently formatted documents. Almost every PDF I looked at in the VPP has a different format. There is no consistency in the way in which (for example) speaker prefixes or stage directions are formatted. Variations in type size and format are unpredictable. Moreover, the structure of a printed play text is complex, necessitating many more distinctions than are typically made in digitizing archival sources or patent documents. I approached a few colleagues with expertise in this area, and their reactions confirmed my suspicions that it would be impractical tto adopt this approach until a reasonably large number (in the hundreds) of correctly tagged play texts was available to serve as a training set. Such a model might be built on the basis of existing TEI corpora – for example those provided by the ECCO collection -- but these are from a different historical period.

In the absence of such a model, I carried out some experiments with a variety of freely available tools claiming to convert PDF to HTML, or XML. Such tools analyse the information stored in a PDF file and attempt to represent the way the characters recognised are to be displayed on the page, in lines, blocks, pages etc. There are two widely used open XML standards for representing such information: PAGE and ALTO; an interesting project called SegmOnto is developing generic tools for constructing a conversion pipeline based on ALTO, for example. I began my experiments with a much older tool, called *pdftohtml* derived from the venerable *xpdf* library which was originally developed in order to produce printed output from a PDF file independently of Adobe software. This can be used to generate XML

files conforming to an ALTO-like schema (in which, for example, each line of text is an element, bearing attributes indicating its position and formatting properties). A series of XSLT scripts (for example, those provided by Dario Kampkaskar in his pdftotei library) can then be applied to transform this to something more tractable and TEI-like.

I also carried out some experiments with Abby FineReader, probably the current market leader in commercial OCR software. The French research agency Huma-Num provides access to Abbyy and also to Tesseract, its leading open source competitor, which enabled me to do some comparison between the two. Neither of these was able to provide a complete solution for a typical VPP PDF, but a pipeline in which Abbyy was used to generate Word, which was then further refined by a series of XSLT stylesheets, starting with the TEI-provided “docxtotei” conversion, seemed to offer the most promise.

The devil is, as ever, in the details. Of the systems I experimented with, all failed one or more of the following hurdles:

- white space between individual words must be preserved
- order of substrings within the document must be preserved
- font distinctions (italic vs roman) should be preserved
- adjacent structural units (e.g. a series of <hi> elements with the same values for @rend) should be merged correctly (almost impossible when for example an italicized stage direction contained non-italicized words)

Also, because of the way the VPP PDFs were re-edited, distinctions based on size of characters or font were sometimes misleading or mistaken. Similarly, the formatting of (for example) parenthesized stage-directions was unreliably indicated: sometimes the opening or closing parenthesis was inside the stretch of italics, and sometimes not.

Recognition of the gross structure (division into acts and scenes) was partially successful for most texts I examined, because the TEI stylesheets mostly correctly distinguished heading lines from body text. Similarly, it proved possible to automatically identify individual speeches by pattern matching (e.g. looking for blocks of text beginning with an italicized word followed by a full stop); to identify stage directions (individual blocks beginning with a [character, or substrings given between parentheses) and so on. However, since these typographic patterns varied considerably across texts, my experiments so far suggest that a degree of manual intervention is likely always to be necessary, if only to customize my generic ”abby2tei” stylesheet to meet some unanticipated typographic eccentricity.

Future work will need to test the generalisability of this pattern-matching approach, by attempting to produce a more substantial set of TEI-conformant texts. At the same time, I would like to explore the idea of producing a training set based on ECCO dramatic texts, which might lead to a more efficient conversion pipeline.

Conclusion

During my second visit to Galway, we agreed that it would be desirable to promote the work done on enhancing the VPP database and to ensure continued access to the associated resources. We agreed that this should be considered as a “relaunch” of the VPP project rather than a replacement for it. This relaunch could not however be undertaken until some infrastructural issues relating to server provision at Galway had been resolved. The work will however continue, and I am very grateful to the CLS INFRA scheme for having “kickstarted” the project.