Sharon Rankin and Casey Lees

McGill Library Chapbook Project: A case study in TEI encoding

Published in:
OCLC Systems & Services: International digital library perspectives,
Vol. 31 Iss: 3, pp.134 - 143

doi: 10.1108/OCLC-07-2014-0030
http://www.emeraldinsight.com/doi/10.1108/OCLC-07-2014-0030

McGill Library Chapbook Project: A case study in TEI encoding

Introduction

McGill University Library is fortunate to have rich rare and special collections, so the curator's options for scholarly resources to include in a specially-funded digitization project were seemingly endless. It was decided to pull together all of the chapbooks existing in several named collections to create a *McGill Library Chapbook Collection*. As it turned out, the collections provided 933 chapbooks, published in the United Kingdom and the northeastern United States between 1780 and 1833. This collection is substantial in size and comparable to the chapbook collection at the National Art Library, at the Victoria and Albert Museum in South Kensington, England.

Glaister's *Glossary of the Book* defines a chapbook as "a paper-covered booklet costing a penny or so, as sold by travelling hawkers (chapmen) who included bundles of them with the buttons, threads, laces and so on which they carried from village to village. Chapbooks were usually about 6 in. by 4 in., had up to twenty-four pages illustrated with crude but lively woodcuts, and had a decorated cover title." (Glaister, 1979 p.92) Chapbooks are usually unbound, with their leaves stitched (hand sewn) making them a fragile work. Chapbooks are also sometimes called penny histories, as they were originally created as popular literature for adults, based upon medieval romance, English legends and folklore and many children's chapbooks were printed with nursery rhymes, fairy tales and natural history information.

The coloured paper cover of *The history of curious and wonderful birds*. New York: Printed and sold by Mahlon Day, [between 1825 and 1833]. Photo credit: Merika Ramundo

McGill Library Chapbook Project is not the first project to digitize chapbook holdings and make them available on the web. There are several sites where a sampling of individual chapbooks can be found fully digitized. For example, the *Literature for Children* digital library contains fifteen digitized chapbooks from the children's literature collections of Florida state's public university libraries and *Early English Books Online (EEBO*) contains nine chapbooks from the 17th century as fully searchable XML texts.

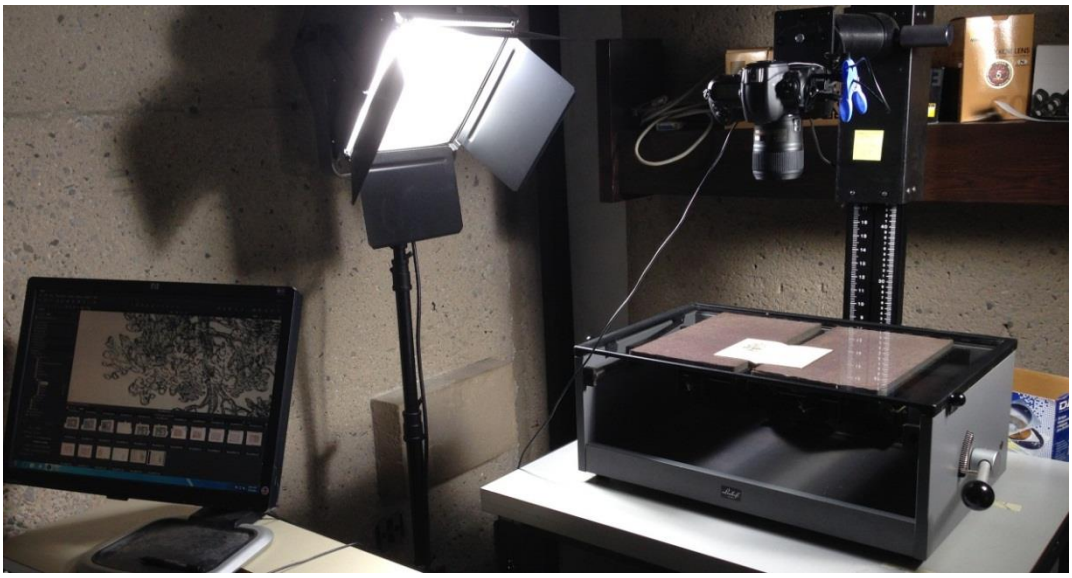
The McGill project is unique in that its objectives were to digitize every chapbook in the Rare Books and Special Collections Library, to make each work freely accessible, and to create an accurate text file to accompany each work.

The resulting website, *The McGill Library Chapbook Collection* (http://digital.library.mcgill.ca/chapbooks/index.php) is the product of the collaborative efforts of many McGill Library staff and part-time student assistants. A liaison librarian in Rare Books and Special Collections led the project and the TEI encoding work. The scanning, OCR processing and website creation activities took place in the Library's in-house digitization department in Library Technology Services (LTS). The cataloguing and digital metadata creation was completed by the Library's Collection Services department. The Interacting with

Print Research Group in the Faculty of Arts provided a collection of contextual essays on British chapbooks for the website.

Chapbook Digitization

The bulk of the digitization work was done in 2012 by students using a book easel and a Nikon D3X camera mounted on a camera copy stand. High resolution images as a two-page spread were captured for each chapbook opening, to ensure that the pages lined up correctly in the Internet Archive Book-Reader software. Post processing was performed using Adobe Photoshop and all of the texts have been loaded into the Internet Archive (https://archive.org/search.php?query=mcgill library chapbook collection) and will also be loaded into HathiTrust in the coming year.



The digital images were prepared in house using a linhof book cradle and a Nikon D3X camera mounted on a camera copy stand. Photo credit: Merika Ramundo

Several preservation issues arose that needed to be taken into consideration in the digitization and encoding of the chapbooks, due to the size, fragility and cheapness of the paper that printers of the day used. Students were instructed by the conservation librarian as to how to best handle such delicate materials. Most chapbooks have their leaves folded and pages sewn together with a single stitch, often making it difficult to keep them open for scanning and encoding. Cradles and

lead weights were used to support and hold the openings. Several of the chapbooks were bound by their previous owners and their tight bindings sometimes prevented the full scanning or encoding of the pages. Uncut sheets and foldout leaves presented scanning challenges. Some pre-existing damage was encountered - missing pages and water damaged pages reduced legibility in some cases.

From July 2012 to May 2013, students in LTS used the ABBYY FineReader OCR software to create and text proof all electronic texts against the page images. Once the text file was judged accurate, three final files were created for each chapbook: a text file, a PDF file, and an HTML file. The HTML file produced by this process was then transferred to the staff in Rare Books and Special Collections who used this file as an input file for TEI encoding. The encoding students used the actual chapbook during their encoding work, correcting the XML text to match the work in hand. This has resulted in the creation of a very accurate electronic text; reviewed by two individuals during two separate processes.

TEI Encoding Levels

At the basis of any digital humanities project is the requirement of a marked-up text file that can be read and manipulated by a computer program. Many library digitization projects provide access to a text file that has been created by an OCR (optical character recognition) scan of the digital image. Without text proofing, the resulting text file may contain errors (or unrecognizable characters) and is often not an accurate representation of the printed work. This text file also has no additional encoding that describes the structure of the page (i.e. does it have a header or page number?), the formatting of the text (i.e. is the text in italics?) or the structure of the work (i.e. is this the title page, where does the paragraph end?).

Since 1987, a literary text encoding markup language has been in development by scholars and other communities of practice called the Text Encoding Initiative (TEI). The resulting TEI Guidelines “have been an enormous success and today nearly every humanities textbase project anywhere in the world uses TEI” (Renear, 2004, p. 13). In 2002, the TEI Guidelines “P4” were

released for XML and updated in 2007 to "P5", the version used in the McGill Library Chapbook Project.

The TEI encoding of a text file creates a marked-up description of the text that can be used by programs specifically written to analyze text. The encoder uses defined TEI tags to recognize and indicate objects in the text and to signify what the object is. The TEI tags are defined in a set of DTDs (document type definitions) – XML markup declarations that define different document types and their associated characteristics. *The TEI Guidelines for Electronic Text Encoding and Interchange, P5: Guidelines for Electronic Text Encoding and Interchange* (2014) contains 18 modules or discrete tag sets that can be combined to provide a tag set relevant for one's text encoding project. There is a base tag set that contains elements (tags) found in all the modules and the other modules contain tag elements for specific types of literary works, for example dictionaries, verse, manuscripts.

McGill's project coordinator created and maintained a document to record all encoding decisions and this was based upon *Best Practices for TEI in Libraries* (2011) a set of guidelines prepared by TEI SIG on Libraries - Kevin Hawkins, Michelle Dalmau, and Syd Bauman. The guidelines specify five levels of TEI encoding complexity ranging from Level 1, a basic level to facilitate full text searching up to Level 5 that requires "substantial human intervention by encoders with subject knowledge" (p. 3). It was necessary to first decide on a level of TEI encoding that was appropriate for the project goals and the resources available.

The 4th encoding level was selected for the McGill project as this level of tagging supports sophisticated search and retrieval capabilities and textual analysis. Level 4 includes elements and attributes that describe the content of the text in addition to the appearance of text, and provides extensibility in case a higher level of encoding is desired in the future. The level of description and presentation at Level 4 is such that the finished text file can be read and understood by general readers even without the page images. Level 4 was judged suitable as both the project coordinator and the TEI encoders, graduate students in the Master of Library and Information Science program did not possess sufficient English literature subject expertise to code for Level 5.

McGill Library Chapbook Project used all of the 21 recommended elements for Level 4: <front>, <titlePage>, <back>, , <head>, <body>, <p>, <lg>, <l>, <figure>, <floatingText>, <note>, <hi>, <list>, <lb/>, <pb>, <facsimile>, <sic> <cor> or <choice> for error and typos, and <unclear>. Some of the chosen elements were changed and adapted several times as the project progressed in order to best reflect the structure and contents of the text in the chapbooks.

The TEI Schema

The graduate students used Oxygen, an XML editor for the TEI encoding and over the course of two years encoded 530 chapbooks. Using the XML editor, it was possible to check for the two necessary two conformance characteristics: "well-formed" and "valid". A well-formed XML file has all properly positioned start and end tags for each element used and the tags are nested (arranged) according to the XML syntax. A valid XML file contains only the tags defined in the XML schema for the project. The schema defines a subset of all of the elements and attributes that will be used to encode the text and is referenced in the header portion of the TEI encoded file.

The schema was created by the project coordinator using the ROMA interface (http://www.tei-c.org/Roma/), a customizing tool allows one to create an ODD file (for "One Document Does it all") that contains the selected tag set. ROMA also documents the schema in HTML or PDF. The schema was tailored for the project needs in two ways: unused tags were removed and two customizations were made to accommodate the coding of project specific data.

A subset of the TEI modules was selected; only those containing tags that were relevant for use during the encoding were placed into the schema. All other modules were removed from the schema, therefore reducing the tag set available for use by the encoders. This reduced the number of choices available to the encoders in the XML editor and helped simplify their tag selection and validation.

New chapbook tags

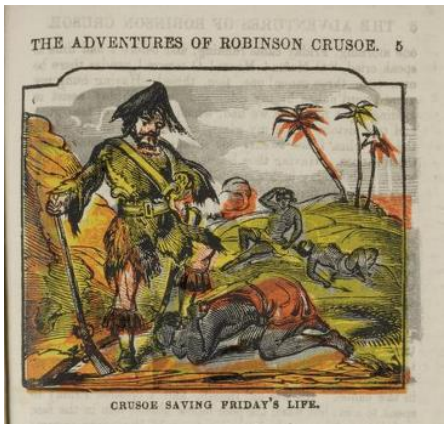
Scholars have classified chapbooks into broad subject categories, different from the Library of Congress Subject headings that were assigned during the cataloguing of the work. Neuberg's

(1964) classifications of chapbooks, is based upon Tillinghast's (1905) scheme in his *Catalogue of English and American chap-books and broadside ballads in Harvard College Library* and these categories are currently in use the *Lilly Library Chapbook Collection* at Indiana University, so this project needed to include a similar classification .

To insert these subject categories into the TEI schema, the type attribute in the <text> element was modified to allow one of the following 18 subject categories to be assigned: "Religious", "Household", "HistPolitBiograph", "GeographHistory", "TravelAdventure", "OddStrangeEvents", "ProseFiction", "RomancesFairyFolk", "NurseryRhymes", "Dramatic", "MetricalTales" (includes verse or poetry), "SongBooks", "JestBooks", "Occult", "Prophecies", "Crimes", "Instruction", and "Miscellaneous". Encoders found that there was some blurring between the "Instruction" and "Religious" categories in the case of very popular moral tales and treaties. It was common practice during the project to code works of this type as "Instruction" rather than "Religious" because generally the purpose of the chapbook is *instructing* how to be moral, though religion is most often a portrayed as significant part of morality in these works.

The woodcut illustrations in chapbooks are a subject of some study by history of printing scholars. "The woodcut illustrations that adorned chapbooks were designed to attract the reader by offering a visual component to the text, but they were rarely related to the subject matter and contributed little or nothing to the actual story" (Grand, 2012, Woodcuts, para 1). By encoding the subject of the woodcut image in the text where the image was printed, the project intends to provide an opportunity for some analysis of this thesis.

The Iconclass classification system was selected to ensure that the tag variable used by the coders was controlled vocabulary. Iconclass "is a classification system designed for art and iconography. It is the most widely accepted scientific tool for the description and retrieval of subjects represented in images (works of art, book illustrations, reproductions, photographs, etc.) and is used by museums and art institutions around the world." (Iconclass, 2012) Using the Iconclass 2100 Browser (http://www.iconclass.nl/iconclass-2100-browser) it was possible search online and copy the code and descriptive text into the <name> and <desc> tags. This new TEI element was defined for the schema using the TEI customization functionality of ROMA.




```
<fw type="pagenumber" rend="align(right)">5</fw>
<fw type="header" rend="align(centre)">THE ADVENTURES OF ROBINSON CRUSOE.</fw><lb/>
<figure rend="coloured">
    <ic:iconClass>
        <name>46A182</name>
        <desc>master ~ slave</desc>
    </ic:iconClass>
    <head>CRUSOE SAVING FRIDAY'S LIFE.</head>
    <lb/>
```

Woodcut from Defoe, Daniel, 1661?-1731. The life and adventures of Robinson Crusoe. [S.l. : s.n., 1840?], page 5 and accompanying TEI encoding snippet.

http://digital.library.mcgill.ca/chapbooks/fullrecord.php?ID=7194

It has yet to been decided how the Icon classification codes will be validated, but the task will be simplified due the freely available raw data set of the classification notations. Every notation in Iconclass is now linked open data and associated with a Uniform Resource Identifier (URI), which could be added during the validation process.

TEI XML File structure

This is a skeleton view of the TEI file structure without the encoded text showing the mandatory elements in each XML file:

```
<?xml version="1.0" encoding="UTF-8"?>

<?xml-model href="file:/U:\RBD\Chapbooks\Chapbook Project current schema/oddRBSC26.rnc"
type="application/relax-ng-compact-syntax"?>
```

```
<!DOCTYPE TEI [
<!ENTITY slong "&#x17f;">
<teiHeader>
<fileDesc></fileDesc>
<encodingDesc></encodingDesc>
<profileDesc></profileDesc>
<revisionDesc></revisionDesc>
</teiHeader>
<TEI>
<text>
<front>
    <titlePage></titlePage>
</front>
<body>
    [lines of text are entered here]</lb>
</body>
<back>
    [back matter on the text goes here, if any]
</back>
</text>
</TEI>
```

The opening XML declarations reference the Unicode character set and the location and name of the TEI schema – the ODD file. The <teiHeader> supplies the descriptive and declarative metadata for the digital file -- an electronic title page. The <teiHeader> is comprised of four main sections:<fileDesc> contains the bibliographic description of the digital text. <encodingDesc> preserves information about the digital text and the source it was derived from, <profileDesc> records information about who created the digital text, project name, institution,

funding, etc., and <revisionDesc> records the revisions made over time to the file. The encoders and reviewers added add their name, date and notes in the <change> element each time they touched the XML file. For the McGill Library Chapbook Project, the <fileDesc> data will be transferred from the corresponding MARC record in the Library catalogue.

The <front>, <body>, and <back> elements contain the text of the work and the TEI schema defines which tags are permitted within each element. For example the <titlePage> cannot be placed in the body or back of the work and has unique nested elements to describe the document title and publisher information. Using a division tag ( ) with attributes, it is possible to code the special sections encountered in a chapbook, for example, a cover, an epigraph or an alphabet in the front of the work. The <body> is where the majority of the paragraph and lines content goes with page breaks between each page. The <back> is usually reserved for any text on the back cover of the chapbook and sometimes includes a division for any advertisements or colophon. Form works are used throughout the text to code headers, footers, and page numbers. Notes with a descriptive attribute are also used to indicate any marginalia on the pages. The images of woodblock prints were coded in the front, header, and back sections.

Encoding challenges

The large number of elements and level of detail in the TEI encoding standard is a steep learning curve and involves time and practice to train students on the nuances of the elements and how to use them. The project coordinator created three documents that were constantly updated during the encoding phase of the project. One was a reference guide explaining the TEI structure, providing concrete coding examples and recording encoding decisions. The second document was a workflow checklist that each student used through the encoding of a chapbook. The third was an XML TEI template that contained the file structure and examples of encoding syntax in use for the project. Depending upon the complexity of the work and the expertise of the student, it often took several sessions/hours to complete the encoding and validation. For readers interested in consulting these project documents, McGill will be depositing copies into the new TEI Archiving, Publishing, and Access Service (TAPAS) repository currently in collaborative development.

Despite the numerous features of Level 4 encoding with TEI, there were several challenges in applying it to the literary form of chapbooks. The use of some of the TEI elements can be vague and uncertain. There are some features unique to individual chapbooks that are difficult to translate into the encoding syntax, for example tricky headers or footers that do not easily translate into form works. Some of the nested elements, specifically the title page could be frustrating – sorting out what elements area allowed and not allowed. Sometimes a feature in a chapbook appears in a part of the work that would not validate against the TEI structure, for example an alphabet division in the body of the text as opposed to the front, so an alternative coding solution needed to be found.

Most of the challenges with selecting a relevant Iconclass code revolved around general problems with controlled vocabulary. It was sometimes hard to select the code that best described the image and accurately reflect its content. Such decisions were dependent on the subjective interpretation of what is the primary focus in the woodblock image was.

Conclusion

Green (2014) examines five case studies of librarian and faculty collaborations on TEI encoding research projects in American universities and describes the importance of academic librarians' collaborative work with their faculty on research initiatives or encoding instruction. Specifically, the Text Creation Partnership (TCP) at the University of Michigan Library and the Etext Center, now the Scholar's Lab at the University of Virginia Library are profiled as successful librarian-faculty collaborations in the creation of TEI encoded literary texts. "… librarian support for research methodologies in digital scholarship, such as text encoding, not only enables scholars to learn scholarly skills but also is a critical conduit for inculcating scholars into the ethos of the communities of practice for digital humanities. (Green, p. 233)

Several digital humanities websites that provide researchers textual analysis based on TEI encoding. The Algernon Charles Swinburne Project at Indiana University, launched in March 2012 offers a number of custom features under its "Tools & Tactics" section including research visualizations, image and text analysis tools for specific use with the poetic texts of Algernon Charles Swineburne digitized by the project. The Women Writers Online at Brown University uses a vertical timeline bar feature for users to explore digitized texts by date. The

English Broadside Ballad Archive at the University of California Santa Barbara allows users to filter searches in the three categories of Bibliographic, Woodcut/Impression, and Tune with any simultaneous combination of metadata filters.

"With centuries of experience in reproducing, cataloguing, classifying and indexing documents, as well as in information design and retrieval, librarians are well positioned to take a role in text encoding -- to move beyond the scriptorium and beyond traditional library roles" (Sukovic, 2002, p. 6). The faculty partnerships with McGill librarians using the McGill Library Chapbook project are for the future. The TEI corpus of files is a laboratory awaiting teaching and research questions. Since the launch of the McGill Chapbook Project website in August of 2013, the site has generated interest and its access statistics are climbing each month. The decision to place the chapbook texts on in the Internet Archive will increase their visibility and use. The graduate students who participated in the TEI encoding over the two-year period have been fortunate to learn a new suite of skills involving XML file structure and the application of a markup language that requires interpretation. As McGill University Library's first involvement in the creation of TEI encoded digital editions, the work has developed project expertise and understanding in a practical sense about the complexities involved in a project of this nature.

References

Burnard, L. (2014). What is the text encoding initiative? Retrieved from http://books.openedition.org/oep/680

Glaister, G. A. (1979). Glaister's glossary of the book. (2nd ed.). London: George Allen &Unwin.

Grand, J. (2012). Street Print: A Brief History of English Chapbooks http://digital.library.mcgill.ca/chapbooks/nodes.php?p=004

Green, H. E. (2014). Facilitating Communities of Practice in Digital Humanities: Librarian Collaborations for Research and Training in Text Encoding. *The Library Quarterly*, *84*(2), 219-234. doi: 10.1086/675332

Iconclass. (2012). *Home.* Retrieved from http://www.Iconclass.nl/home

Indiana University. (2012). The Algernon Charles Swinburne project. Retrieved from http://webapp1.dlib.indiana.edu/swinburne/

Indiana University. (2012). User's guide to the Lilly Library chapbook index. Retrieved from http://www.indiana.edu/~liblilly/chapbook.shtml

Neuburg, V. E. (1964). Chapbooks: a bibliography of references to English and American chapbook literature of the eighteenth and nineteen centuries. London: Vine Press.

Renear, A. H. (2004). Text Encoding. In S. Schreibman, S. Raymond & U. John (Eds.), A companion to digital humanities. Blackwell Reference Online. Malden, Mass.: Blackwell Pub.

Sukovic, S. (2002). Beyond the scriptorium. *D-Lib Magazine.* Retrieved from doi: 10.1045/january2002-sukovic

TAPAS. (2014). TAPAS Project. Retrieved from http://tapasproject.org/about

TEI Consortium. (2011). Best Practices for TEI in Libraries. Retrieved from http://purl.oclc.org/NET/teiinlibraries

TEI Consortium. (2011, March 05). P5: Guidelines for electronic text encoding and interchange. Retrieved from http://www.tei-c.org/release/doc/tei-p5-doc/en/html/

University of California Santa Barbara. (2003) English Broadside Ballad Archive. Retrieved from http://ebba.english.ucsb.edu/

University of Florida George A. Smathers Libraries. (2011). *SobekCM.* Retrieved from http://ufdc.ufl.edu/sobekcm/

Welsh, C. and Tillinghast, W.H. (1968). Catalogue of English and American Chapbooks and Broadside Ballads in Harvard College Library, Singing Tree Press, Detroit.

Women Writers Online. (2012). Retrieved from http://textbase.wwp.brown.edu/WWO/search?browse-all=yes