The Rosarium Project

Building a digital collection on the genus Rosa using <oXygen/> and the TEI

Julia Rachel Tryon
Phillips Memorial Library + Commons, Providence College, Providence, Rhode Island, USA

Received 13 January 2016
Revised 23 February 2016
Accepted 23 February 2016

Abstract

Purpose – This paper aims to describe the Rosarium Project, a digital humanities project being undertaken at the Phillips Memorial Library + Commons of Providence College in Providence, Rhode Island. The project focuses on a collection of English language non-fiction writings about the genus *Rosa*. The collection will comprise books, pamphlets, catalogs and articles from popular magazines, scholarly journals and newspapers written on the rose published before 1923. The source material is being encoded using the Text Encoding Initiative (TEI) Consortium's P5 guidelines and the extensible markup language (XML) editor software <oXygen/>.

Design/methodology/approach – This paper outlines the Rosarium Project and describes its workflow. This paper demonstrates how to create TEI-encoded files for digital curation using the XML editing software <oXygen/> and the TEI Archiving Publishing and Access Service (TAPAS) Project. The paper provides information on the purpose, scope, audience and phases of the project. It also identifies the resources – hardware, software and membership – needed for undertaking such a project.

Findings – This paper shows how straightforward it is to encode transcriptions of primary sources using the TEI and XML editing software and to make the resulting digital resources available on the Web.

Originality/value – This paper presents a case study of how a research project transitioned from traditional printed bibliography to a web-accessible resource by capitalizing on the tools in the TEI toolkit using specialized XML editing software. The details of the project can be a guide for librarians and researchers contemplating digitally curating primary resources and making them available on the Web.

Keywords Digital humanities, <oXygen/>, Digital collections, Roses, TEI, Text encoding

Paper type Case study

The Rosarium Project will collect, highlight, digitally curate and bring to the Web historical primary sources about the genus *Rosa*. The project focuses on non-fiction materials written about the rose in English from the sixteenth century up to 1923. These resources will include books, pamphlets, ephemera and articles from popular magazines, scholarly journals and newspapers. The digital texts – with their accompanying images, a rose glossary and a bibliography – will be fully searchable on the project's own website. Researchers will be able to browse the books and articles, search by subject, rose variety and color via drop-down menus or do a keyword search of the whole collection. Users will have freely available online access to historical documents about roses.

This paper discusses the project in its first phase, including the project's conception, workflow and required resources. Phase One is divided into three parts: research

Digital Library Perspectives
Vol. 32 No. 3, 2016
pp. 209-222

2059-5816
DOI 10.1108/DLP-01-2016-0001

undertaken and learning to encode, reevaluation of encoding strategy and reworking encoded texts and encoding the remaining texts in the first-phase group.

The principal researcher, a research librarian at the Phillips Memorial Library + Commons of Providence College, conceived the project as a way to collate, curate and provide easy access to everything written about the most revered flower in the garden. Hundreds upon hundreds of books and articles have been written about "the queen of flowers", but these primary sources are scattered in library collections and online repositories.

This project will fill a need for its intended audience, which includes researchers in the fields of garden history and horticulture, as well as popular culture scholars interested in leisure activities and suburban life. It is also expected to be of great value to gardeners, particularly rosarians, interested in learning tricks and techniques used before modern fertilizers and pesticides to those studying garden designs and styles of bygone eras and to those wishing to identify hardy old roses as that may still be available from specialty rose nurseries.

The idea to create an online collection of materials on the rose sprang from a much earlier bibliographic project, which had been started and shelved in the 1990s. Back then, there were and still are only two bibliographies on the rose written in English. They are both severely outdated and incomplete.

In early 2015, the principal researcher rekindled her interest in doing research on what was written on the rose. She wanted to collect information about the rose into an easy to use and access format, but a printed bibliography was out of the question; printed bibliographies are now passé. The notion of merely compiling a simple bibliography and uploading it to the internet, although potentially helpful to rosarians, was not adequately stimulating intellectually. Plus, current technology provides opportunities that just cannot be passed up; a mildly tech-savvy librarian can now do amazing things with just a little training and the right software. A major skill set of the principal researcher involves database searching and database management, and the idea of creating a fully searchable online database of writing on the rose was an exciting intellectual challenge.

There were some questions about how to go about creating a searchable database. The principal researcher consulted with the head of the library's Digital Publishing Services Department (DPS), Mark J. Caprio, because his team was responsible for creating a number of online collections. Mr Caprio suggested two possible methods based on his expertise in digital publishing. The choices were doing a text encoding project using the Text Encoding Initiative (TEI) Consortium's P5 Guidelines or creating a database using MySQL and PHP. Both options would require acquiring new skill sets, including learning new software.

Leveraging local experience and support possibilities was crucial and a very important part of the decision-making process. The principal researcher was primarily concerned with technical support for the project. DPS has expertise with the TEI, having recently completed a TEI-encoded project with members of the college faculty on the Dorr Rebellion (http://library.providence.edu:8080/xtf/index.html). Additionally, DPS has experience in presenting TEI projects using the eXtensible Text Framework (XTF), "which provides a powerful, flexible platform for providing access to digital content" (California Digital Library, 2015). The department thus could provide guidance and

support for the Rosarium Project in terms of the TEI encoding and an established means of hosting it on the Web with XTF.

As far as what had to be learned, the TEI option required learning to apply the guidelines in an extensible markup language (XML) file. XML is a fairly straightforward mark-up language, and learning it seemed doable considering the principal researcher's experience with HTML. The MySQL-PHP option, on the other hand, would require learning both a "relational database management system" (Wikipedia, 2016) and a programming language, both of which would be wholly new subjects to master. Choosing to work with the MySQL-PHP option would require taking classes to learn database creation and programming skills not otherwise used in the library. As a result, the learning curve for MySQL and PHP seemed much steeper and more problematical than learning to use the TEI.

With regard to the basic content of the database to be created, both options were viable. However, the TEI encoding option would allow the principal researcher to add contextual information that the MySQL-PHP option would not quite so easily do. The TEI framework encourages the incorporation of information that can assist the reader to better understand the work. This is done with "ographies", which are additional files containing lists of data that are linked to words in the text:

> These are like local authority lists that you create. These lists include the personography for information about people (prosopography), placeography for information about places (gazetteer), bibliography for citations, and orgography for information about organizations (Women Writers Project, 2016).

The principal researcher decided to go with an encoding project using the TEI because it would be easier, better supported and would provide more opportunity for adding value to the transcriptions of the rose writings. Particular selling points included the ability to standardize the text, while providing access to the original version and the capacity for interlinking the texts with other related TEI-encoded files. Encoding with the TEI would also permit the insertion of images into the body of the text and linking out to relevant information on the Web. Plus, encoded texts are machine-readable and manipulable, making them a great resource for researchers in the field.

A little bit about the TEI

The TEI is an internationally accepted standard for encoding text in the digital humanities. The TEI specifies:

> […] encoding methods for machine-readable texts, chiefly in the humanities, social sciences and linguistics. Since 1994, the TEI Guidelines have been widely used by libraries, museums, publishers, and individual scholars to present texts for online research, teaching, and preservation (Text Encoding Initiative Consortium, 2013).

The TEI has a few unbreakable rules but is mostly very adaptable so as to be useful in a wide assortment of encoding projects.

The TEI acknowledges:

> […] that each scholar must have the freedom of expressing their own theory of text by encoding the features they think important in the text. A wide array of possible solutions to encoding matters is demonstrated in the TEI Guidelines, which therefore should be considered a reference manual rather than a tutorial. Mastering the complete TEI encoding scheme

implies a steep learning curve, but few projects require a complete knowledge of the TEI (TEI by Example, 2014).

Ironically, this makes learning the TEI easy but means using it appropriately and effectively is more difficult.

Each TEI-encoded file must contain two sections – the TEI header and the text. The TEI header is where all the metadata for the electronic text is contained. The TEI header must include a description of the file, which perforce includes the title of the electronic record, details about the publication of the electronic text (such as the people involved and the project sponsor) and data on the source for the electronic text. The TEI header is also where a description of the project as a whole may be kept along with a taxonomy, a record of changes made to the file and statements on editorial policies. The text section is where the transcription of the written material is housed but can also include images and front and back matter, such as notes, bibliographies, glossaries and indexes.

Phase One Part A

The Rosarium Project kicked off in March of 2015. The principal researcher conducted initial research using online databases such as the American Antiquarian Society's *Historical Periodical Collection* and *The Reader's Guide Retrospective*, online catalogs such as Harvard's *Hollis+* catalog and *WorldCat*, digital repositories such as the Hathi Trust, Google Books and the Internet Archive and the two English language bibliographies written on roses. Bibliographic records for materials discovered were created and stored in the online bibliographic utility RefWorks. Over 900 potential sources were identified in this round of research.

The principal researcher decided to start with a manageable group of articles. To do this, the Reader's Guide Retrospective database, available through the library, was searched. The Reader's Guide is a well-respected index to journal articles and began publication in the early twentieth century. The materials it indexed were popular journals, widely circulated and routinely held by libraries; this is key as it means that the articles indexed would be easy to acquire. The Reader's Guide uses standardized subject headings, and the appropriate heading for this project's purpose was simply "roses". Next the search results were limited to materials written before 1923. Then the results of the search were reviewed to remove records for poetry and other literary genres. This resulted in a set of 163 non-fiction, English language articles, which were published between 1894 and 1922.

Fifty of the articles came from *The Garden Magazine*, which was published between 1905 and 1924. The magazine had 20 regular departments, one of which was for roses, thus explaining the high article count for the title. Twenty-six articles appeared in magazines dealing with country life, the most prominent being *Country Life in America*. A surprising number appeared in general interest periodicals, literary reviews and art journals. The subject matter of the 163 articles spans a variety of topics from basic pruning techniques to rose petal recipes.

Citations were downloaded from the Reader's Guide Retrospective database to RefWorks. A bibliography was created with these citations and a spate of interlibrary loan requests ensued.

Materials were received and processed in April and May of 2015. Most of the articles were received electronically as Portable Document Formats (PDFs). The principal researcher scanned photocopied articles and turned them into PDFs. She copied the

collection of source materials onto an external hard drive for safe keeping and then loaded copies of the PDFs into a Dropbox account for easy access because work on the project would be done at multiple locations.

In an effort to prepare for learning the TEI, the principal researcher undertook a basic course in HTML via lynda.com in April 2015. Although the TEI encoding is done in XML, learning basic coding practices proved useful and made learning the TEI much easier. At no time has she taken a course in XML and has not had a need to do so. A basic explanation of XML is provided on the TEI Consortium's website (www.tei-c.org/release/doc/tei-p5-doc/en/html/SG.html).

An introduction to the TEI came via materials for a basic TEI workshop that is given by the people behind the Women Writers Project (WWP) at Northeastern University (www.wwp.northeastern.edu/outreach/resources.html). All members of the DPS Department had attended the workshop and recommended reviewing the documentation since there were no TEI workshops slated for the near future.

The principal researcher learned to encode texts by studying and following examples of TEI-encoded text. The TEI is explained and documented in the guidelines on the Consortium's website (www.tei-c.org/Guidelines). Each TEI element has its own webpage, which includes a brief description, information on where element tags can appear, as well as examples of how it is used.

The first articles were encoded in April and May 2015 using the XML editor software <oXygen/> on a 27" iMac running OS X Yosemite. This software knows the complete encoding options for the TEI and provides templates with the basic file structure (Figure 1). <oXygen/> gives prompts for potential tags to be used based on the location of the cursor and surrounding tags. Writing well-formed code is easy with <oXygen/>

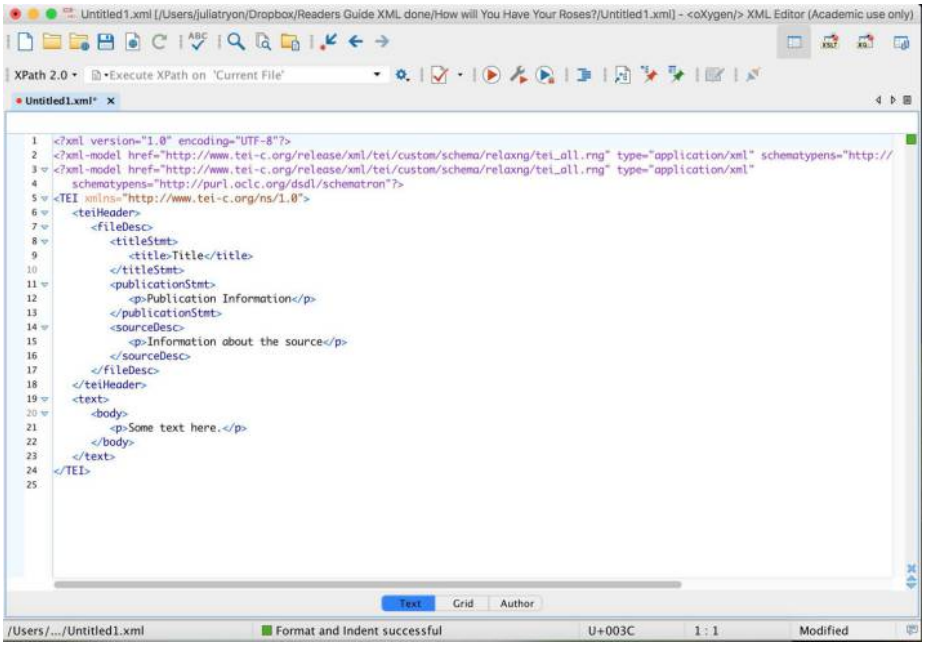

Figure 1.
TEI template

as it validates the code as it is typed. Unmatched tags or tags placed improperly are underlined in red and are hard to miss. Also an explanation of the error is provided at the bottom of the screen (Figure 2). There is an online user manual for <oXygen/> (www.oxygenxml.com/documentation.html); but using <oXygen/> is so straight forward, the principal researcher has never needed to consult it. The license for <oXygen/> for academics is only $99.00 and is available for systems running Windows and Linux, as well as OS X. In theory, one could use a basic text editor, such as TextWrangler, which is free for Mac users, but having the complete TEI schema in <oXygen/> is the better option especially for someone new to text encoding. There may be other software programs that do what <oXygen/> does, but it is the program the WWP uses.

As far as equipment required for the project goes, only a computer is needed; but a large screen is a must. The principal researcher purchased the 27" iMac because of the need to consult multiple documents simultaneously and a 24" iMac could not provide enough screen space. Typically, she consults the original PDF, multiple <oXygen/> files and the TEI website all at the same time. It might be possible for some to run a project such as this on a dual screened computer. However, the principal researcher has found that she needs to very closely examine transcriptions and encoded texts and that using a second monitor with a 24" iMac has not worked out for her. A 27" iMac currently starts at $1,799.00.

The workflow for the project starts with the transcription of an article from the PDF to a Pages document. Pages for Mac is used because it permits the keying of diacritics and special characters and because the principal investigator prefers it to Microsoft Word, but any word processing software would do. Transcribing the texts directly into <oXygen/> was also an option, but familiarity with the Pages software and a cleaner looking workspace made it a better choice.

For each article, a new XML file is created in <oXygen/> using the full TEI schema template. The transcription is then copied and pasted into the body portion of the XML

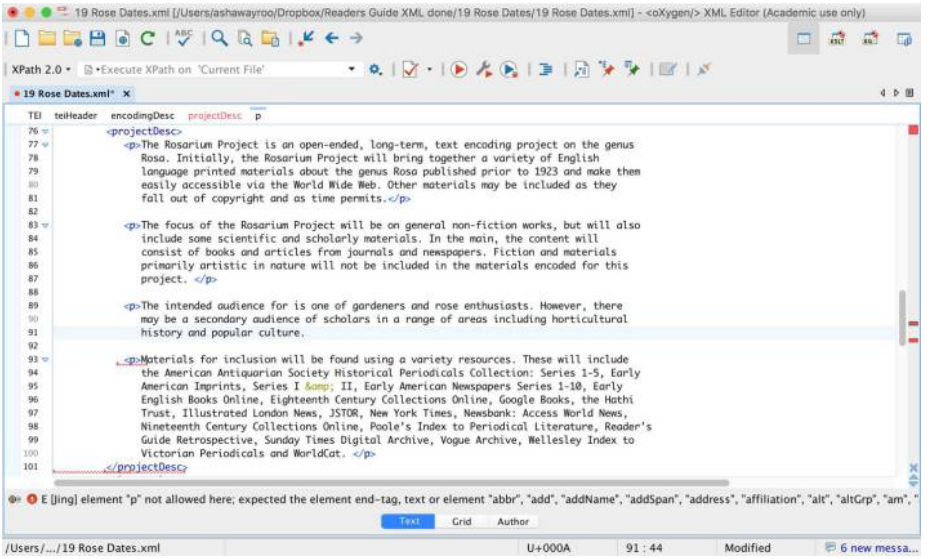

Figure 2.
<oxygen/> error message

file. Next the required elements of the TEI header are fleshed out. These are the <titleStmt>, <publicationStmt> and <sourceDesc> which fall within the <fileDesc>. Subject headings are also determined and added to the TEI header at this point.

Next the text of the transcription is encoded following the TEI (Figure 3). The principal researcher has chosen to markup names of people, places and roses so they can be linked to contextual information located in the personography and glossary files. Flower colors are also marked up as "by color" will be a drop-down search option. Names of people are tagged with the <persName> element. Varieties of roses are tagged with the <name> element. Colors of roses are marked with <seg type="color"> tags, and geographic locations are tagged with the <placeName> element.

The encoded file is compared to the original PDF for stylistic details because text transcribed in italics or in a bold font come over as plain text. Footnotes must also be added to the <oXygen/> file because they also do not come over with the transcription. Once that is complete, the fully encoded file is then uploaded to the TEI Archiving Publishing and Access Service (TAPAS) Project website (http://beta.tapasproject.org) to see how the code renders (Figure 4). At this time, the rendered file is compared with the original PDF, looking for typos and spacing issues and making sure any images have uploaded correctly.

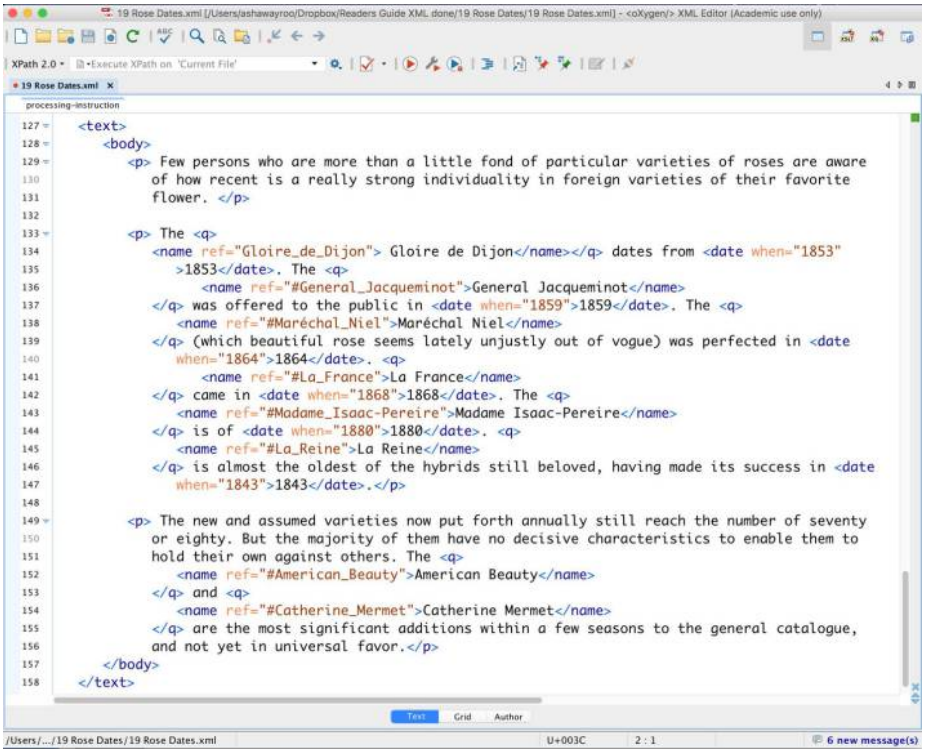

Figure 3.
Text

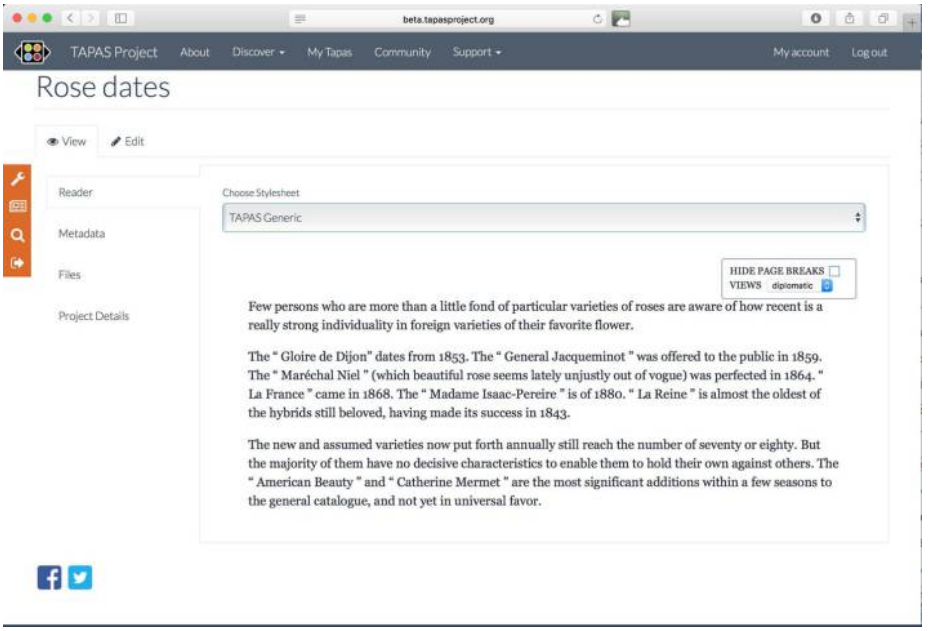

Figure 4.
Encoded file as it renders on the TAPAS Project website

The TAPAS Project provides:

> TEI publishing and repository services at low cost to those who lack institutional resources: faculty, students, librarians, archivists, teachers, and anyone else with TEI data who wants to store, share, and publish it. TAPAS seeks to achieve these goals in a collaborative, open, and community-driven way using open-source tools (TAPAS Project, 2015).

Membership in the TEI Consortium, which is only $50.00 per year, is required to take advantage of the TAPAS Project.

At the end of Phase One Part A, basic records for 15 articles were taking shape, but there were many unresolved issues stemming from the uniqueness of the project. The desire to make a fully searchable database on roses meant a need to be able to search by variety, color and subject. Learning the TEI in a vacuum, with no similar projects to crib off of, made it difficult to determine how to tag rose varieties and subjects properly. There were also questions about how to create a "glossary" of rose varieties because examples found were for glosses of the classic variety. Despite this, Phase One Part A was considered basically a success as the principal researcher was able to learn enough of the TEI to encode transcriptions and have them properly render on the TAPAS Project website.

Phase One Part B

In November of 2015, the principal researcher luckily was able to attend a workshop on basic TEI encoding held at Northeastern University by Julia Flanders and Syd Bauman, the principals behind the WWP. This workshop was crucial because it provided a general overview to the TEI mindset, explained basic encoding decision-making strategy and provided an opportunity to get detailed answers to specific questions from

TEI professionals. Materials from the workshop also provided links to additional resources, which have proven particularly helpful in moving the project forward (www.wwp.northeastern.edu/outreach/seminars/intro_2015-11/presentations/welcome_and_wrapup/wrapup_02.xhtml). Among the resources was a link to an especially helpful website – TEI by Example (http://teibyexample.org/TBE.htm).

As a result of the workshop and the newly discovered resources, the principal researcher decided to re-encode all the previously encoded files in December 2015. She edited each file to provide more and better metadata and to fully take advantage of the TEI header portion of the files. In particular, information on the electronic publication was expanded, a project description and notes on editorial policies were added and a changes log was created for each file. Tags for rose colors were corrected to conform to TEI tag usage standards. Graphic URLs were updated to reflect TEI tag formation standards.

In addition, a bibliography for the 163 articles in Phase One was encoded into its own file. Each entry was given an identifying tag. This and improved bibliographic data were inserted into the source description element of the TEI header of the appropriate files (Figure 5).

At the same time, the principal researcher also created a taxonomy for the project because existing taxonomies, such as the Library of Congress Subject Headings (LCSH), were not detailed enough. For example, LCSH include headings for gallicas and damasks but omitted other rose families, such as albas, portlands and noisettes. Subject headings previously assigned to the first encoded files were mixed with the common and botanical names for roses. Lists of botanical names came from sources, such as the lists of species roses on Wikipedia (https://en.wikipedia.org/wiki/List_of_Rosa_species) and

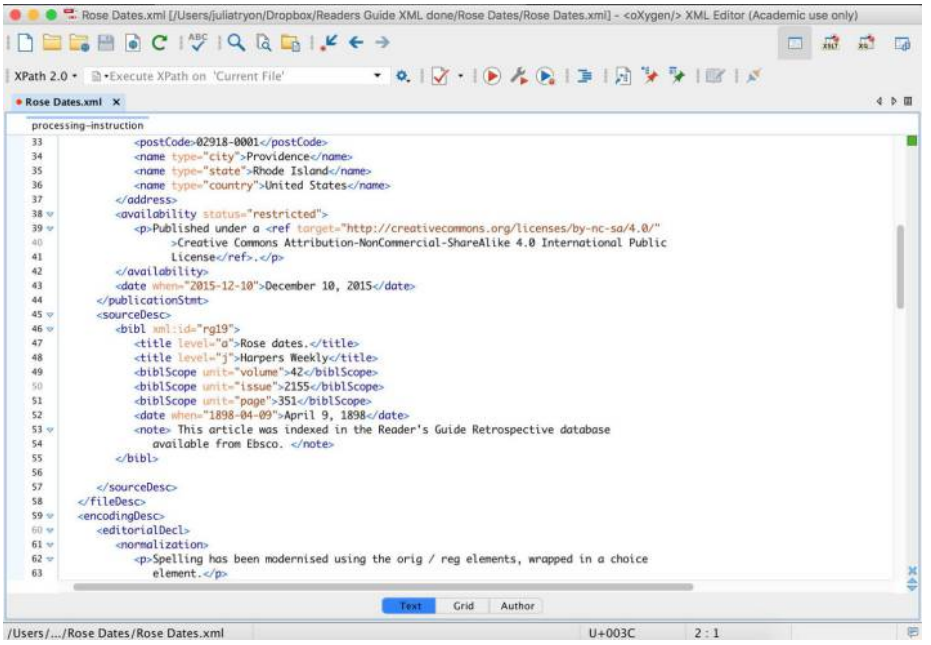

Figure 5.
Bibliography

on www.everyrose.com, and from books on roses in the personal collection of the principal researcher. The taxonomy currently has 216 entries. There is the expectation of further entries being added as the project proceeds and of current entries being revised with particular attention to synonyms. Appropriate entries from the newly created taxonomy were added to the class declaration element of the TEI header in all files (Figure 6).

The principal researcher also began a personography. Entries for authors and persons named in the articles, such as famous rose breeders, were created with the xml:id attribute in the personography file (Figure 7). These reference IDs link the personography entries to the <persName> tags in the text section of the appropriate files. At present, most entries are placeholders. It is expected that significant research will have to be done to provide useful details concerning the named persons. It is hoped that, at minimum, basic biographical data can be found for all entries. Images and links to relevant resources on the Web will also be added when available.

The primary researcher began encoding a glossary of rose varieties and rose related things (rose gardens, rose societies, etc.) as well (Figure 8). Again most entries are currently placeholders for more detailed entries. Information for each rose is expected to include type, color, habit, breeder, release date and other data that may be of interest. Entries will also include a color photo of the rose when one is available (Figure 9). Research for the glossary is expected to require a considerable amount of time, but the benefits of having this information linked to the transcriptions are significant.

Phase One Part B was very successful. As a result of the training and newly discovered resources, for example, the TEI by Example website, whose importance cannot be overstated, robust XML files were encoded for the first 15 articles that

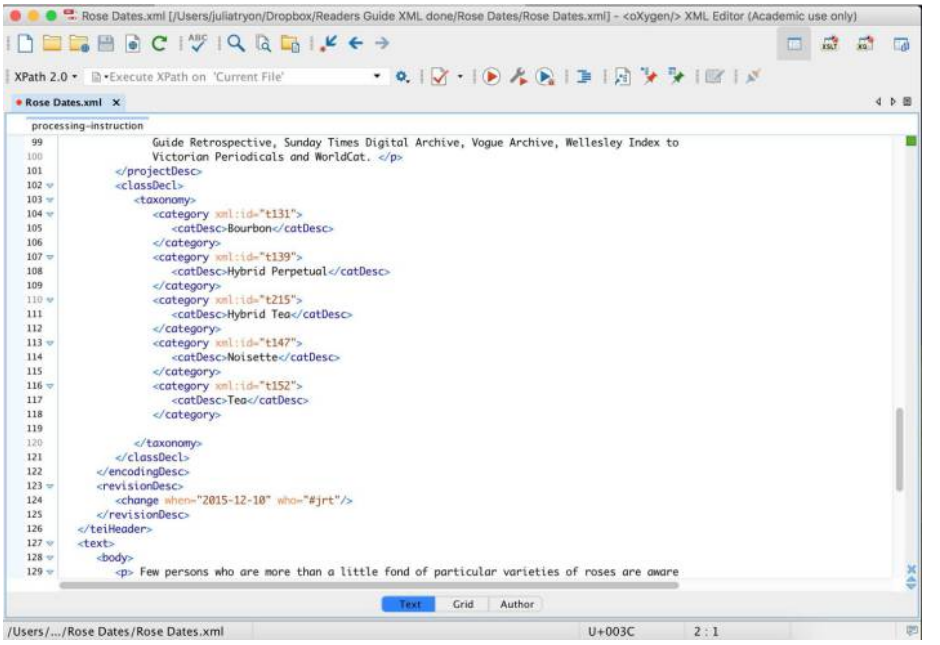

Figure 6.
Taxonomy

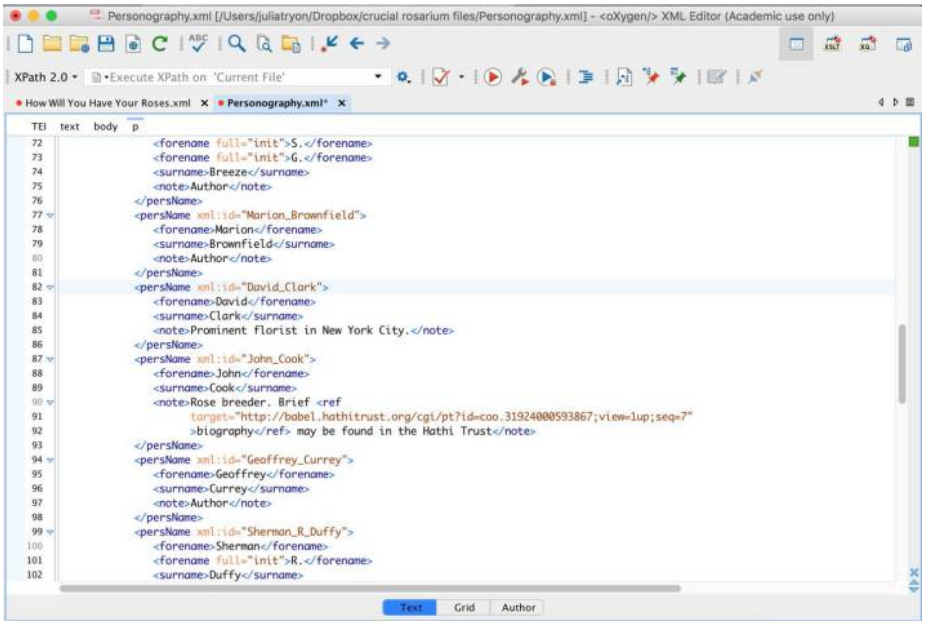

Figure 7.
Personography

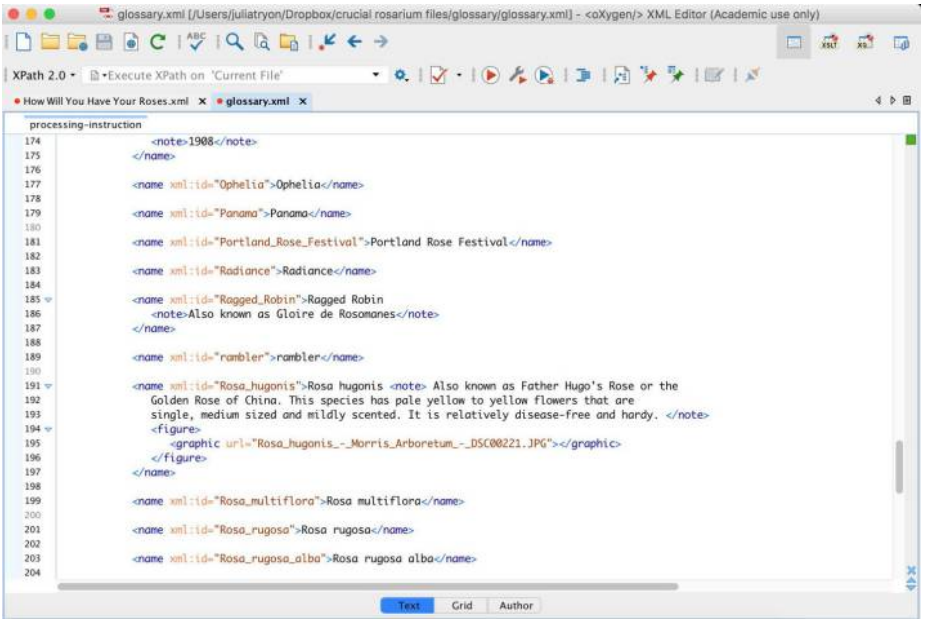

Figure 8.
Glossary encoded

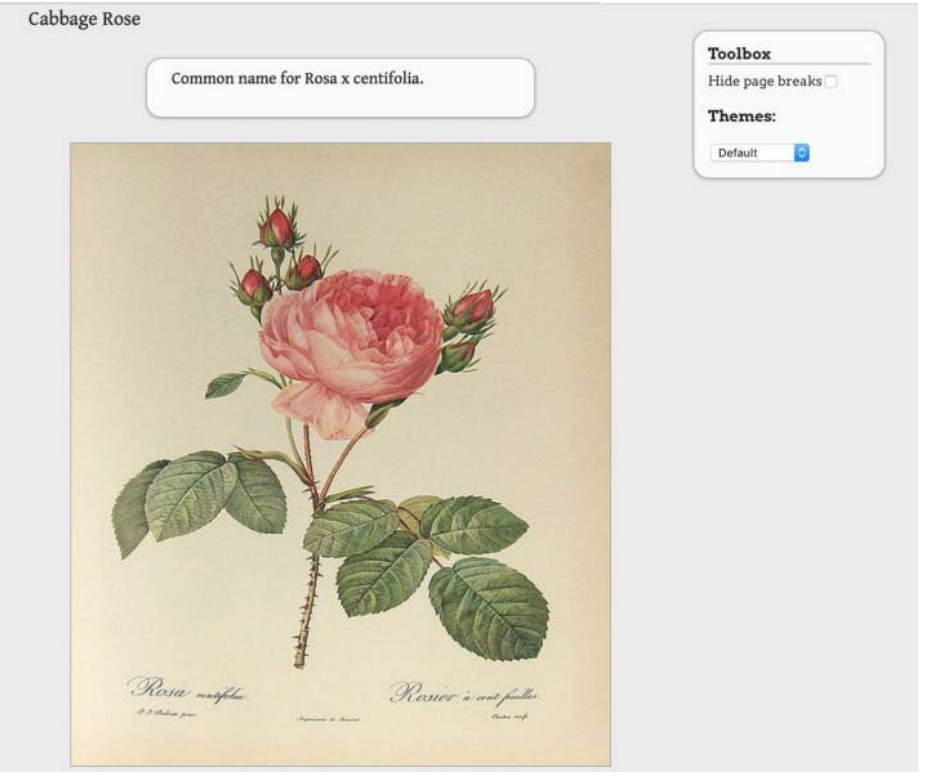

Figure 9.
Glossary as it is rendered

properly rendered on the TAPAS Project site. Also as a result of the TEI workshop, the principal researcher was able to make crucial decisions regarding what and how to tag transcriptions and set down encoding rules for the project. Growing encoding experience has ensured consistency in the application of those rules.

Phase One Part C

The last part of Phase One of the Rosarium Project has begun. It will be the longest to date because it will be the transcription and encoding of the remaining articles of the original 163 identified through the Reader's Guide Retrospective. It is expected to continue for over a year, assuming the daily responsibilities of the principal researcher remain constant. A major concern is the amount of research needed to provide contextual information for these articles. Indeed, this research may well consume more time than the encoding itself.

Concurrently, work has begun on the web presence for the project. The members of the DPS Department are tackling this work. They will create a website for the project that will enable users to fully search the database of encoded texts and easily browse through the resources. As yet, there has been no date determined as to when the project will be accessible to the public via the Web.

Future work: Phase Two and beyond

In Phase Two in the Rosarium Project, the principal researcher will return to the publications indexed by the Reader's Guide Retrospective. She has discovered, while working with some of the primary resources, that the Reader's Guide was not completely rigorous in its indexing of materials written on the rose. This will mean reviewing each issue of the journals indexed to identify further material. Some of this work may be done online using the Hathi Trust, Google Books and the Internet Archive, but some titles will require hands-on research at libraries with major horticultural and popular culture special collections.

After that, future phases will focus on transcriptions of books published between 1799 and 1923, articles written in other horticultural journals and newspapers from 1700-1923, book chapters relating to the rose in general texts published from the sixteenth century forward and materials issued by the two major rose societies: the Royal National Rose Society and the American Rose Society.

It is also possible that there may be collaboration in the future with other TEI encoded projects. Because the TEI is an internationally accepted standard, collaboration with researchers working on text-encoding projects in related areas of horticultural science, garden history, leisure studies and suburban life would be easy and welcome.

The principal researcher is pleased with the results of the project thus far and is prepared to continue the project in stages over the next 15 years. Once the first texts are made available on the Web, she will make the project known to the proper constituencies and seek feedback.

The principal researcher can highly recommend the practice of encoding texts with the TEI as a means of digitally preserving and curating historical material. The encoding process gets easier with practice, and there is plenty of opportunity for scholarly work in providing the contextual information. She feels that TEI projects of this sort are a natural fit for libraries and librarians.

List of resources

- American Antiquarian Society Historical Periodicals Collection: www.ebscohost.com/archives/aas-thematic-collection
- Dropbox: www.dropbox.com
- Google Books: https://books.google.com
- Hathi Trust: www.hathitrust.org
- Hollis+: http://hollis.harvard.edu/primo_library/libweb/action/search.do?vid=HVD
- Internet Archive: https://archive.org
- <oXygen/>: www.oxygenxml.com
- Pages for Mac: www.apple.com/mac/pages/?cid=wwa-us-kwg-mac
- RefWorks: www.proquest.com/products-services/refworks.html
- Reader's Guide Retrospective: www.ebscohost.com/academic/readers-guide-retrospective
- TAPAS project: http://beta.tapasproject.org
- TEI by example: http://teibyexample.org/TBE.htm
- Text encoding initiative consortium: www.tei-c.org/index.xml

- TextWrangler: www.barebones.com/products/textwrangler/features.html
- Women Writers Project: Resources for Teaching and Learning Text Encoding: www.wwp.northeastern.edu/outreach/resources.html
- WorldCat: www.worldcat.org

References

California Digital Library (2015), *eXtensible Text Framework (XTF)*, available at: www.cdlib.org/services/access_publishing/publishing/tools/xtf/ (accessed 10 February 2016).

TAPAS PROJECT (2015), *About*, available at: http://beta.tapasproject.org/about (accessed 19 December 2015).

TEI by Example (2014), "Module 0:Introduction to text encoding and the TEI", available at: http://teibyexample.org/modules/TBED00v00.htm (accessed 19 December 2015).

Text Encoding Initiative Consortium (2013), *TEI: The Text Encoding Initiative*, available at: www.tei-c.org/index.xml (accessed 19 December 2015).

Wikipedia (2016), *MySQL*, available at: https://en.wikipedia.org/wiki/MySQL (accessed 10 February 2016).

Women Writers Project (2016), *Encoding Contextual Information*, available at: www.wwp.northeastern.edu/outreach/seminars/_current/presentations/contextual_encoding/contextual_encoding.xml (accessed 5 February 2016).

Corresponding author
Julia Rachel Tryon can be contacted at: jtryon@providence.edu