IJDC | *General Article*

Developing Data Curation Protocols for Digital Projects at Vanderbilt: Une Micro-Histoire

Veronica A. Ikeshoji-Orlati
Vanderbilt University

Clifford B. Anderson
Vanderbilt University

Abstract

This paper examines the intersection of legacy digital humanities projects and the ongoing development of research data management services at Vanderbilt University's Jean and Alexander Heard Library. Future directions for data management and curation protocols are explored through the lens of a case study: the (re)curation of data from an early 2000s e-edition of Raymond Poggenburg's *Charles Baudelaire: Une Micro-histoire*. The vagaries of applying the Library of Congress Metadata Object Description Schema (MODS) to the data and metadata of the *Micro-histoire* will be addressed. In addition, the balance between curating data and metadata for preservation vs. curating it for (re)use by future researchers is considered in order to suggest future avenues for holistic research data management services at Vanderbilt.

Received 20 October 2016 ~ *Revision received* 23 January 2017 ~ *Accepted* 23 January 2017

Correspondence should be addressed to Veronica A. Ikeshoji-Orlati, Digital Scholarship and Scholarly Communications, Vanderbilt University Library, PMB 135, 230 Appleton Place, Nashville TN 37203. Email: v.ikeshoji-orlati@vanderbilt.edu

An earlier version of this paper was presented at the 12th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

http://dx.doi.org/10.2218/ijdc.v12i2.574
DOI: 10.2218/ijdc.v12i2.574

Introduction

The availability of research data management services (RDMS) at academic libraries has grown steadily in the past decade. One of the major reasons for the increased interest in RDMS, particularly at research universities, is data management requirements imposed by external funding agencies, such as the NSF and NIH in the US and Horizon2020 in the EU (Tenopir, Birch and Allard, 2012). As a result, RDMS often have taken the form of data management, curation, and publication consulting for ongoing, data-intensive research projects, primarily in the physical and life sciences.

The focus on curating data for scientific research currently in progress, however, has occluded the growing need for RDMS tailored to the data found in the arts, humanities, and social sciences.[1] As Flanders and Muñoz note in their introduction to digital humanities data curation (n.d.), humanities data are substantively different from scientific data, both in type (encompassing nearly any medium) and structure (sometimes based on analogue formats, other times entirely experimental). In addition, there is an ever-growing corpus of humanities datasets and digital humanities projects which, after substantial investments of time and resources, now lie fallow, their data and metadata becoming progressively less accessible with every system update and security patch. The proliferation of digital humanities projects at colleges and universities around the world, therefore, adds another critical dimension to the scope of academic RDMS: the (re)curation of legacy data and metadata for both long-term preservation and research or pedagogical (re)use.

In this paper, the intersection of digital humanities projects and RDMS at the Jean and Alexander Heard Library of Vanderbilt University will be considered through the lens of a case study: the re-curation of data from an early 2000s e-edition of Raymond Poggenburg's *Charles Baudelaire: Une Micro-histoire*. The solution applied to the *Micro-histoire* – marking up the data in XML using the Library of Congress Metadata Object Description Schema (MODS) – is not universally applicable to the plethora of legacy digital humanities projects under the stewardship of the Vanderbilt Libraries. Nevertheless, the balance struck between curating the *Micro-histoire* data and metadata for preservation and curating it for (re)use by students and scholars may prove informative in the ongoing development of a holistic, humanities-inclusive RDMS at the Vanderbilt University Libraries.

Digital Projects at Vanderbilt

At Vanderbilt, there has been a long, fruitful history of experimentation with, and online publication of, digital projects and resources in the humanities and social sciences. Growing out of rich physical archives and fruitful trans-institutional partnerships, the university web domain is rife with digital projects such as the TV News Archive[2], Ecclesiastical and Secular Sources for Slave Societies (ESSSS) project[3], Global Music

[1] In 2013, the Data Curation Centre in the UK hosted a one-day workshop on the topic of RDMS for the Arts and Humanities, highlighting the infrastructural shifts needed to actively curate diverse types of data: http://www.dcc.ac.uk/events/research-data-management-forum-rdmf/rdmf10-research-data-management-arts-and-humanities

[2] TV News Archive: https://tvnews.vanderbilt.edu/

[3] ESSSS (renamed the Slave Societies Digital Archive in 2017): http://www.vanderbilt.edu/esss/

Archive[4], and ETANA[5]. The Library also has a vigorous and vibrant exhibits program resulting in dozens of physical and online exhibitions showcasing diverse aspects of the collections.[6] The breadth of digital initiatives at Vanderbilt may be measured not only by the variety of subjects and topics, but also by the diverse forms of digital data represented, including texts, still images, films, and audio stored in numerous formats and accessed through myriad channels.

As at many other institutions, there has been robust technical and financial support for the creation of digital humanities projects and art exhibits, either through departmental and institutional research funds or external grants from private and governmental agencies. However, assets are limited for the requisite continual maintenance of the platforms which make the projects accessible (Kretzschmar and Potter, 2010). The dearth of human and fiscal resources is particularly noticeable in the case of projects and datasets designed to be accessed through visually-appealing, user-friendly web portals, since they require near-constant attention to ensure their security and functionality.

A comprehensive digital data curation initiative in the Vanderbilt Libraries is yet nascent, but a preliminary workflow for identifying digital projects' curation needs has been developed. Integral to identifying a digital project as a candidate for data curation is to establish its current stage in the development lifecycle. Is the project under active development? Is it in use by its target community but no longer under development? Or has the project seen declining or discontinued use for some reason? While any digital project may be a viable candidate for data curation efforts, the dozens of completed projects whose data and metadata are at risk of permanent loss are currently prioritized. Subsequently, the data curator considers what constitutes the primary intellectual content of the project. This is not a straightforward task, and determinations about what data and metadata will be curated may be based, in part, on the anticipated availability of time and resources to maintain or iteratively re-curate the project in the future.

Three main principles underlie Vanderbilt's digital data curation initiative as a whole; they may be outlined as follows:

- The data of a digital project ('content') are often distinct from the vehicle on which they are presented ('form'). The separation between form and content has long been used as a digital preservation strategy, particularly when libraries were first grappling with the proliferation of digital objects on a large scale (Hedstrom, 1998). While some projects integrate form and content in novel ways, in the majority of cases distinguishing between data and the software that serve them to end users remains one of the more sustainable models for preserving and enabling (re)use of digital projects' intellectual content.

- Digital project metadata are to be treated as data throughout the process of curation. This practice is based on the idea that metadata are themselves products of extensive, systematic intellectual labor, adding value to the primary data which they describe and contextualize.

- Digital project data *and* metadata should be curated not only for archival purposes, but also to facilitate (re)use thereof. The Library bears the responsibility of preserving and making accessible the data it stewards, which

[4] Global Music Archive: http://www.globalmusicarchive.org/

[5] ETANA: http://www.etana.org/

[6] Vanderbilt Library Exhibits: http://exhibits.library.vanderbilt.edu/

> requires considering both the stability and the (re)usability of the curated data in the long term.

The e-edition of *Charles Baudelaire: Une Micro-histoire* is an illustrative example of a legacy digital project which has been re-curated to preserve and facilitate access to its data and metadata by students, researchers, and the interested public. Clifford Anderson, Associate University Librarian for Research and Learning in the Vanderbilt Libraries, selected the *Micro-histoire* as a case study for establishing data curation protocols for legacy digital projects. In Fall 2016, Veronica Ikeshoji-Orlati, CLIR/DLF Postdoctoral Fellow for Data Curation in the Vanderbilt Libraries, executed the *Micro-histoire* data curation project, the details of which are documented below.

A History of the Micro-histoire: 1987-2016[7]

Before delving into the digital curation of the *Micro-histoire*, a brief introduction to the original work is warranted. In 1987, Raymond Poggenburg (1926-2004), then a professor of French literature at Vanderbilt University, published the first edition of the *Micro-histoire*. The aim of the project was to provide a detailed chronology of the life of Charles Baudelaire (1821-1867), the esteemed French poet and essayist.

The first edition of the *Micro-histoire* is divided into four parts: chronology, references, bibliography, and chronological index. The chronology, comprised of 4,155 dated entries, constitutes the majority of the tome. It begins with the birth of Baudelaire's paternal grandfather, Claude Baudelaire, in 1711 and ends shortly after Baudelaire's death in 1867. Subsequently, the references are presented as a series of endnotes, with up to ten citations for each entry. The bibliography, encompassing primary source documents and critical studies up to the date of publication, serves both as a works cited section and as a historiographical resource for scholars. Finally, the chronological index allows readers to look up individuals, places, or other entities and find where in the chronology they are referenced.

Some of the ways Poggenburg distinguished his *Micro-histoire* from a traditional biography of Baudelaire aided in its translation into a e-edition in the early 2000s. To emphasize the "factual" nature of the original work, for example, Poggenburg eschewed a single narrative format, instead crafting individual entries and organizing them chronologically to document specific data points related to the life and work of Baudelaire. Furthermore, he substantiated nearly every entry with at least one reference, leveraging and, in turn, highlighting the bibliographic depth of Baudelairian studies.

In the early 2000s, the second, e-edition of the *Micro-histoire* was published jointly by the Vanderbilt University Press and Jean and Alexander Heard Libraries.[8] Some data points were added, bringing the total number of chronology entries to 4,519 and extending the bibliography to 475 items. The switch from a physical to digital medium for the *Micro-histoire* was radical in many ways, but fundamentally it aligned with Poggenburg's desire to make the *Micro-histoire* "une sorte de chantier, un projet en développement, et non... un monument seulement" (1987).

The e-edition of the *Micro-histoire* may be considered the first curation of the data. The process began with the digitization of the book and ended with a user-friendly web

[7] Portions of this section have been adapted from the history of the project presented in the readme page of the GitHub repository for the *Micro-histoire* project, written by Veronica Ikeshoji-Orlati: https://github.com/HeardLibrary/Poggenburg

[8] See http://diglib.library.vanderbilt.edu/baud-search.pl

page for accessing the expanded chronology. Rather than preserve the layout of the first edition in an e-book format or scan and OCR a copy of the physical volume like a typical biographical work, it was decided that the contents of the *Micro-histoire* would be entered into two SQL databases – one for the chronology entries, one for the bibliography. Using a web portal built on the Perl framework, the chronology database could be queried by users from around the world, either by keyword or date range. The e-edition of the *Micro-histoire*, therefore, made clear decisions about the functionality of the work and sought to increase the use and accessibility of the publication by making it readily searchable online by students, researchers, and the interested public.

Re-curating the *Micro-histoire*: Fall 2016

In August 2016, due (in part) to the deterioration of the Perl framework on which the e-edition was built, the *Micro-histoire* data curation project was initiated. First, the e-edition chronology entry and bibliography databases were flattened and exported to CSV files. Subsequently, OpenRefine and an assortment of customized VBA macros were used to clean and standardize the data. In addition to basic data cleaning such as proofreading, the removal of duplicate entries, and the standardization of names and references, some of the substantive changes to the data were as follows:

- Dates were restructured and standardized to more accurately and consistently represent the original Micro-histoire entries.[9]
- Unfinished bibliographical references (e.g. missing page numbers, publication information, titles) were completed, where information was available.
- All bibliographical information (e.g. volume and issue numbers, article page numbers, etc.) was checked against the original publications or WorldCat records and corrected.

The Library of Congress Metadata Object Description Schema (MODS) version 3.6 was selected for structuring and marking up the *Micro-histoire* data and metadata in XML.[10] Some of the applicable benefits of MODS for the project include:

- Sufficient granularity for different types and formats of dates,
- Well-developed syntax and vocabularies for relating items and objects to one another,
- Adaptability for describing physical and digital objects,
- Integration with LC subject headings,
- Flexibility in the definition of top-level elements for non-bibliographic information.

It may seem curious to use MODS (a schema explicitly designed for bibliographical metadata) to organize data from a single publication. The unique structure of the *Micro-histoire*, however, lends to a facile blurring of the lines between data and metadata. Examination of how a couple of key MODS elements were applied to the *Micro-*

[9] On the dates in the *Micro-histoire*, see the 'Dates and LC Subject Headings' section below.
[10] Library of Congress MODS specification: http://www.loc.gov/standards/mods/mods-outline-3-6.html

histoire may elucidate how the re-curated dataset balances the needs of digital preservation and the desire to increase accessibility and (re)usability of the work.

References and Related Items

The structure and display of references in the re-curated *Micro-histoire* is an example of where preservation and (re)use concerns had to be weighed against each other. As outlined above, the original *Micro-histoire* had four distinct parts: chronology, references, bibliography, and chronological index. By appending the references to the appropriate entry for the chronology and replacing the chronological index with a keyword search, the early 2000s e-edition preserves much of the functionality of the original *Micro-histoire*. The bibliographical database, however, remains hidden to the end-user of the e-edition since there is no way to query it (directly or obliquely).

In the re-curated *Micro-histoire*, the MODS <relatedItem> element with "references" attribute is utilized to cite all of the sources associated with each chronology entry. For the purposes of digital preservation, each chronology entry is fully self-contained: there are no links between different sets of records, so the full bibliographical information of each work referenced appears in every chronology entry record. The structure of the e-edition (which mimics the original *Micro-histoire* in its use of unique identifiers linking the chronology entry and bibliography databases) may be more efficient in terms of computing power, but the MODS XML documents dramatically increase the visibility of the bibliography and facilitate historiographical searches, which were impossible with the e-edition.

Dates and LC Subject Headings

A more intractable problem for re-curating the *Micro-histoire* first arose when the book was translated from analogue to digital format: how to make machine-readable the recording and display of chronology entry dates. Though many of Poggenburg's entries have specific dates attached to them, the formatting is inconsistent and many are not absolute. Some entries, for example, date to "Spring" or "between October and November" of a specific year, whereas others may extend across multiple years. Even when the date is in the standard first edition format of day (one- or two-digit Arabic numeral), month (capital Roman numeral), and year (two- or four-digit Arabic numeral) separated by a space, a one- or two-letter abbreviation indicating the day of the week may be inserted between the month and year (e.g. 28 VIII v 35 for Friday, August 28, 1835).

In the early 2000s e-edition, the date issue was partially resolved by translating all dates into a MM/DD/YYYY format. The re-curated *Micro-histoire* MODS XML files go a step further by defining start and end dates for periods of time which Poggenburg had chosen to define verbally, rather than calendrically, in the first edition. The increased specificity of the calendrical date-recording system for chronology entries in the re-curated *Micro-histoire* substantially increases the sensitivity of date- or time-period-based search results. Nevertheless, in order to preserve the literary nuance of Poggenburg's original chronology entry dates, the verbal descriptions are recorded in a MODS <note> element as well.

Beyond the basic balance of (re)use-worthy (calendrical) and preservation-necessitated (verbal) date formats, the recording of chronology entry dates in the re-curated *Micro-histoire* files also serves to contextualize the work as a whole and define the nature of its contents. By utilizing the MODS <temporal> subelement of the

<subject> field to contain the chronology entry start and end dates, the contents of the entries are linked to the LC subject headings for the time periods in which they occurred. As a result, the bibliographical information about works referenced in each chronology entry also become linked to LC temporal subject headings. Such broader contextualization of the content and bibliography of the *Micro-histoire* facilitates deeper consideration of the historiography of Baudelairian studies as a whole. By treating data (Poggenburg's chronology entry date) as metadata (a point in time which defines an LC heading or subheading), the re-curated *Micro-histoire* also serves to affirm the content of the chronology entries as an account of substantiated historical events.

Defining 'Accessibility'

While the separation between content and form has been identified previously as a guiding principle to digital data curation initiatives in the Vanderbilt Libraries, it is necessary to consider what the change in delivery platform means for (re)use of the *Micro-histoire* data. The search function on the e-edition of the *Micro-histoire*, while riddled with problems, is relatively easy to use. It requires no specialized skills or familiarity with a programming language to execute basic keyword or date-range searches. Though the results are poorly optimized and the bibliography is all but inaccessible, the initial investment by the user to access the data is minimal. The re-curated *Micro-histoire*, however, has a much higher barrier to entry in terms of its usability. Its means of delivery – a repository on GitHub, a website whose primary audience is developers and programming enthusiasts – requires users to download the entire dataset. While the associated readme file for the *Micro-histoire* repository explains the main elements and attributes used in the dataset, a user still would have to refer to the MODS 3.6 documentation for a complete understanding of the structure of each file. Then, in order to sort through the chronology entries, one must be armed with some basic programming knowledge, ideally in the form of XQuery fluency.

While more of the data and metadata is accessible to the end user in the re-curated *Micro-histoire* MODS XML files, crafting the queries to extract it may be perceived as a great deal more effort than running a basic keyword search of the chronology entries. Could that mean that the data are, in fact, less accessible than in the e-edition of the *Micro-histoire*? Fundamentally, no. The increased familiarity with analytical tools required by datasets such as the re-curated *Micro-histoire*, however, does highlight the need for library RDMS to simultaneously curate data and invest in arming researchers with the skills necessary to successfully (re)use and build upon the burgeoning collection of datasets available to them.

From Praxis to Protocol: Lessons from the *Micro-histoire*

As indicated above, the *Micro-histoire* was selected as a case study for the budding digital data curation initiative at the Vanderbilt Libraries. The primary goal for the *Micro-histoire* re-curation project was to find an appropriate balance between archive-ready data and metadata and a format which would encourage increased (re)use of the materials. A few brief observations on the process may be useful at this point.

Having previously undergone some curation in its translation from analogue to digital format, the *Micro-histoire* presented a unique set of opportunities and challenges.

Due to its catalogue-like structure of research-based statements, the *Micro-histoire* data and metadata were well-suited to restructuring using the MODS 3.6 framework. Data (Poggenburg's chronology entry dates) could fruitfully act as metadata (LC temporally-defined subject headings), and metadata (bibliographical information about references) integrated smoothly into the data (as chronology entry related items). MODS may not be appropriate for other datasets, however, and its success in the case of the *Micro-histoire* is, at least in part, the result of the original structure and intent of the work itself.

The *Micro-histoire* MODS XML documents reflect a compromise between preservation and (re)use goals. Fortunately, the two often go hand-in-hand, since they both require maximal exposure of the underlying data and metadata of a digital project. In the case of the *Micro-histoire*, the re-curated files provide the user with all of the data and metadata necessary to learn about the historiography of the work, specific topics or subjects, and time periods referenced. Whether the structure is as efficient as a relational database, however, may be debated, and the extent of (re)use of the re-curated *Micro-histoire* may be circumscribed by the technical skills of its end users.

The question of (re)use raised by the re-curation of the *Micro-histoire* raises broader concerns about what happens after data and metadata are cleaned and sent out into the world and what sorts of infrastructural investments libraries must make to effectively disseminate the digital data they contain. In the case of the re-curated *Micro-histoire*, the resultant collection of MODS XML files may be seen to both increase and decrease access to the data and metadata. While data hidden in the e-edition are visible in the MODS XML files, the learning curve to query and use those data is much steeper than the original web portal. RDMS, therefore, need to extend beyond the practice of managing and curating data to integrate outreach and education efforts. Training researchers in the analytical tools, programs, and methods needed to fruitfully interact with the pared-down data and metadata platforms which libraries can feasibly sustain will be critical to the long-term success of data curation efforts. This is particularly important in the arts, humanities, and social sciences, where the diversity of project and data types will continue to evade standardized curation formats for the foreseeable future.

Conclusion

Curating legacy digital content often requires straddling the line between the realms of research data management and digital preservation. As libraries increasingly devote their resources to preserving and making accessible digital projects for the longue durée, properly curating those digital assets becomes ever more critical. Striking the balance between maintaining preservation and archiving standards while facilitating (re)use of the materials should remain at the fore in digital data curation and RDMS conversations. So, too, should considerations of how to arm researchers with the tools necessary to fruitfully engage with and (re)use the curated datasets being generated.

References

Flanders, J. & Muñoz, T. (n.d.). An introduction to humanities data curation. Digital Humanities Data Curation Guide. Retrieved from http://guide.dhcuration.org/contents/intro/

Hedstrom, M. (1998). Digital preservation: A time bomb for digital libraries. *Computers and Humanities 31*, pp 189-202.

Kretzschmar, W. & Potter, W.G. (2010). Library collaboration with large digital humanities projects. *Literary and Linguistic Computing 25*(4), pp 439-445.

Poggenburg, R.P. (1987). Charles Baudelaire: Une micro-histoire. Vanderbilt University Press: Nashville, TN.

Tenopir, C., Birch, B., & Allard, S. (2012). Academic libraries and research data services: Current practices and plans for the future. A White Paper of the Association of College and Research Libraries. Association of College and Research Libraries: Chicago, IL. Retrieved from http://www.ala.org/acrl/sites/ala.org.acrl/files/content/publications/whitepapers/Tenopir_Birch_Allard.pdf