Preserving Antebellum Print Culture

Jason Peak

May 3, 2013

Abstract

Poe's Magazine World, a digital humanities project centered around the literary culture of the 1840s in America, seeks to accumulate and present digital artifacts that re-contextualize this period in history for future readers, students and researchers. The project team is composed of a cross-disciplinary mix of scholars and students committed to establishing a research archive for the equally diverse audience of scholarly researchers, students and the general public. At its broadest, the scope of the project includes all American Antebellum authors and would take shape as part of a consortium of linked archives. As a first step, however, the project will focus on Poe's involvement with four periodicals, the *The Broadway Journal*, *Southern Literary Messenger*, *Burton's Gentleman's Magazine*and *Graham's Magazine*. The main tasks involved are summarized as follows:

1 Project Description

Understanding the project domain, background, goals and roadmap are essential to asserting the way forward. The domain of the project is hierarchical, and at its most general, it includes everything that can be defined as *Antebellum Print Culture*. Recognizing that small steps begin a great journey, the APC group have wisely decided to begin with a narrower scope, specifically, the writings of Edgar Allen Poe. As the first step, they have chosen to focus on his work as editor of periodicals. That effort is the subject of this discussion.

1.1 Domain

Poe is famous for his tales and poetry, indeed, every grade school child has been thrilled by *The Raven*, for instance. It is probably less well-known that in his capacity as an editor of several periodicals, he took the opportunity to offer commentary on the events and cultural themes that mark the American Antebellum period. The Poe's Magazine Worldproject will draw attention to this aspect of the author and his perspective by creating a digital collection of four of the periodicals in which he played a key editorial role: *Burton's Gentleman's Magazine*, *The Broadway Journal*, *Graham's Magazine*, and the

Southern Literary Messenger. This phase of the project will focus on Poe's involvement with these periodicals and will render each page as scanned image with fully searchable text.

1.2 Goals

As with any project, finding balance between the conceivable and the possible is essential to success. This team has chosen an incremental approach to its larger goals, and we will differentiate first steps from larger outcomes with the notions of *core* versus *stretch*.

1.2.1 Core

Immediate project goals include digitizing each page of every periodical volume of which Poe took part. Considering only the years in which he was actively employed by these editorials, the corpus contains just less than 4,000 pages. The primary deliverable for the next stage of this project will be a permanent, web-accessible repository of a scan for each page and fully searchable text. The tasks, explored below, required to achieve core goals include scanning, text extraction, curation, and preservation. Facilities for research will include, at a minimum, document retrieval and full-text search.

1.2.2 Stretch

With core foundation blocks firmly laid, other services may be built that further enhance the value of this resource for research including scholarly annotation, semantic curation, and a range of APIs that expose the archive to other repositories and applications as Linked Open Data. Annotation will consist of moderated scholarly annotations made by local and distributed domain experts. Semantic curation entails the identification and markup of the occurence of significant entities within the text, including people, places, corporate bodies. Client APIs will be developed and exposed to allow others to reuse the archived dataset for their own, as yet unimagined, purposes. In light of the current trend [citation]in archives towards Linked Open Data (LOD) and the aspirations of the project team, it is reasonable to consider the APC project well positioned to be one of the next notable contributors to this burgeoning community of archives.

1.3 Background

[needs much work for inclusion]The Antebellum Print Cultureproject has been underway for a few years, and in that time, there has been much discussion and converging general agreement over exactly what the content of the repository should be and how to best present it. In the last year, the project has focused its efforts around presenting a proof-of-concept prototype exploring three Poe tales. From a technical perspective, this has been achieved through adoption

and customization of the Omeka web publishing platform. More will be said about the strengths and weaknesses of Omeka later on, but for now, suffice it to say that the Omeka prototyping effort has exposed important requirements of the project, specifically, needs for:

1. a user-friendly management interface to the repository
2. flexible content presentation options
3. fine-grained metadata schemes
4. semantic relationships between entities, especially at the sub-document level

2 Requirements

Archival Preservation The APC project requires trusted archival preservation of its primary artifacts and derivative scholarship and interactions.

Exposure Artifacts and their metadata must be exposed to client applications including search engines, browsers, and visualization applications

Granular markup Text documents will be encoded in the TEI at the various levels enumerated in [1]

3 Early Investigations

Early investigations into software solutions for this project have centered around the popular Omeka publishing platform. We have used Omeka as a holistic solution to the multiple requirements: archival storage, ingest, access. Omeka is a web platform written in PHP and based on the industry standard Zend framework. The application relies on a database, in our case, MySQL, to maintain state, application configuration, user access control, record content, record metadata, and RDF-like inter-record relationship data.

3.1 Archival Storage

Using Omeka as an archive has a number of advantages:

open source, open standards The full stack of technologies underlying Omeka as archive are open source and based on open standards. While the platform can be deployed on the Windows OS, it is more commonly hosted on servers running open source linux. While our institutional infrastructure has decreed that we use Red Hat linux, a commercialized distribution, we could have just as easily chosen Ubuntu or any of the myriad free distribution. MySQL is a well-known, widely-used database made available by Oracle.

4 The Way Forward

Moving towards accomplishment of the core and stretch goals outlined previously, it is evident that the project should be ready to adopt new technologies into its existing information architecture. At the base, their must be a robust and trusted repository in which born-digital artifacts are deposited. The time and expense required to capture this corpus deserves no less, and we imagine that generations of future scholars and students will benefit from this effort, provided that the materials are well-preserved in accordance with the best practices currently known to the archives community. The term *Trusted Digital Repository* is loaded with implications and requirements that we will endeavor to satisfy. Standards exist [citation][DRAMBORA, TRAC] by which to measure and guide these efforts.

Further, in an increasingly interconnected and socially interlinked world, the digital archive assumes a certain burden to provide its content to patrons (researchers, students, the public) in increasingly compelling ways. One of the most promising technologies for achieving this end is embodied in the concept of a Semantic Web [citation][TBL]. The semantic web represents an inductive step beyond the character of the web we've grown to know in the last several decades. In this new web, hyperlinks are not merely a means by which to refer to some other, arbitrary, web resource, rather, the reference itself is imbued with an additional dimension that describes the nature of the relationship between the thing and what it points to. [citation][Barthes?] Employing such technology in an archives setting improves discoverability by expanding the role of the archive from simple document retrieval to something much richer.

The foregoing discussion connotes that access to the contents of the repository is a key driver of the effort. Indeed, our core goals will lead us to a stable and sustainable repository of the corpus, but opening the contents of that repository to outside research is where the project gains traction, value, and the sustainable funding required for permanence. In the following sections, we will explore the issues associated with both the core and stretch goals and recommend tools, standards and workflows for achieving them.

5 Formats and Standards

5.1 File Formats

Image TIFF image format will be required

Text [all non0-image files will be XML]

5.2 XML Formats

TEI

METS

MODS

MADS

DC

SKOS

EAC-CPF

EAD

RDF

5.3 Standards

OAIS

DRAMBORA

TRAC

6 Workflow

Workflow is loosely defined here. Inevitably, refinements will be made as issues are discovered and resolved. The acquisition process, from capture to ingest, can be imagined as a pipeline starting from the physical original and ending with an archival Submission Information Package (SIP). Stages in this pipeline include:

1. Scan
2. OCR
3. OCR correction
4. TEI markup
5. packaging

While each stage will require human intervention and oversight, the OCR correction stage is expected to incur the most significant outlay of human resources.

6.1 Scan

We intend to use facilities in the Hill Memorial Library for the bulk of the scans. This equipment is designed to scan a full page spread, two facing pages, at a time, but each page is saved as a discrete file. Scanning will be applied to each page in each issue of each periodical in the project scope.

6.2 OCR to Text

The ability to read the text contained in these pages is the basis of any scholarship and research to be done. The ability to perform computer-assisted search and retrieval can only be accomplished with machine readable text.

The scanning process will operate at the page level and will require manual operation. It will take as input physical original peridical pages and return as output a simgle high resolution TIFF image per page and text extracted through OCR. The products of this process are the basis of the archive.

periodical physical original magazine

page leaf of a periodical

scan digitization of leaf to TIFF

TIFF image of leaf

OCR process by which text is extracted from TIFF

raw text output of OCR

OCR correction manual process by which machine interpretation errors are corrected by humans

TEI-1 baseline format of text for archival storage

SIP METS-encoded record of a single periodical issue containing TIFF + TEI of each leaf contained in the original

7 Archives Software

A number of software packages exist for managing digital collections, and more are currently under development. Notable among these are ContentDM, Fedora, D-Space

8 Information Structure

8.1 METS Records

METS is a structural description and container format that can wrap other formats including TEI, RDF, MODS, EAD, etc. It enjoys widespread use and the endorsement of leading archival institutions, including the Library of Congress. As a structural description format, it is an ideal choice for storing records of multipage documents [citation]. Fedora uses METS.

9 Questions

1. What examples of RDF + Fedora exist?
2. describe the mapping from DC -¿ METS; is this automated in Fedora
3.

References

[1] TEI SIG on Libraries. Best practices for TEI in libraries. http://www.tei-c.org/SIG/Libraries/teiinlibraries/main-driver.html, October 2011.