COLLECTIONS

A Journal For Museum and Archives Professionals

From the Practical to the Philosophical

volume 12 ◇ number 2 ◇ spring 2016

FOCUS ISSUE

EXPLORING THE SMITHSONIAN INSTITUTION TRANSCRIPTION CENTER

Articles

Cover: Making Time for Moynihan in the Morning: Volunteering with the Transcription Center. Courtesy of Christine Rosenfeld.

The Creation and Evolution of the Transcription Center, Smithsonian Institution's Digital Volunteer Platform

Andrew Gunther

Lead Application Developer, Enterprise Digital Asset Network, Office of the Chief Information Officer, Smithsonian Institution, gunthera@si.edu

Michael Schall

Senior Consultant, Quotient, Inc.
Contractor for the Office of the Chief Information Officer, Smithsonian Institution, mschall@quotient-inc.com

Ching-hsien Wang

Branch Manager, Library and Archives Systems Support Branch, Office of the Chief Information Officer, Smithsonian Institution, Transcription Center Project Manager, WangCH@si.edu

Abstract This article discusses the technical design considerations in creating and evolving a digital volunteer platform for transcribing historic documents and collection records. We outline the thought process of our technical team in attempting to architect and build a system that could achieve a mission of collecting knowledge to promote discovery as well as a platform that was extensible, versatile, able to be integrated, and adaptable to future needs. A unique and unexpected aspect of our project is that the digital volunteers not only contributed data but also shaped (and continue to shape) the technical product, user interface, and user experience.

As the nation's largest museum system, the Smithsonian Institution (SI) maintains an enormous (and growing) collection of historical documents: 138 million museum objects and specimens, 2 million library volumes, and 153,000 cubic feet

Collections: A Journal for Museum and Archives Professionals, Volume 12, Number 2,
Spring 2016, pp. 87–96.

of archival material.[1] The majority of these collections remains largely inaccessible—not yet digitized and held in boxes and folders in storage facilities—and thus unstudied. Initial efforts to digitize the museums' archives and objects yielded many new images electronically available to experts and the public. However, because images are the main yield of these efforts, much of this material's intellectual content remains locked in pixels. Furthermore, whether locked away in storage or digitized, discovery remains difficult because sheer volume alone keeps people from being able to find what they need.

In 2012, SI began an initiative to build an infrastructure that would analyze and transcribe these documents. The core goal of this project was to increase access to its collections by soliciting help from volunteers to transcribe its digitized materials. The yield would be intellectual data that the Smithsonian could use to enhance its collection records, further research, and feed digital data repositories accessible to the public for discovery and exploration. With these goals and the aforementioned challenges in mind, the launch of the Transcription Center (TC) was meant to be not a temporary fix but rather a long-term solution that could grow as collections expanded. So the TC continues to grow as digitized collections are selected for transcription. The TC additionally enables information about collections to grow through more complete archives, data, and metadata. On June 15, 2013, the TC (http://transcription.si.edu) was released.

Project Planning and Goals

In preparation for the creation and launch of the TC, a committee of Smithsonian archivists and librarians comprising key stakeholders first addressed project scope and identified the range of document types to be supported by the TC. They evaluated multiple ongoing crowdsourced projects in the industry and compared these with Smithsonian organization structures and document types. The group's final product was a functional requirements document outlining the following goals:

1. Create a simple and easy user interface for the public.
2. Create a system that is flexible enough to handle many document types (diaries, manuscripts, formatted logbooks, biological specimens, and labels).
3. Support different data formats that are required as finished products (single long field for full text, variable number of CSV fields, and PDF).
4. Allow a scalable system to support high traffic, that is, large number of users and large number of projects.
5. Support both transcribing materials and reviewing for quality control of the works.

6. Enable the transcribed text to be immediately searchable and downloadable by the public online.
7. Enable easy data export of transcribed data to other systems for repurpose and reuse.
8. Create a sustainable system by including an administration module for the internal staff to manage a large number of transcription projects (import, export, edit, review, delete and approval processes, and statistical report for analytics).

While creating our list of requirements, we looked at other digitization tools and methodologies, including Zooniverse, the Drupal Transcribe Distribution project (by the National Archives), as well as engagement practices, such as gamification and measured accomplishment (as developed by Credly). We also carefully considered the diversity of SI collections for transcription. Many existing transcription tools were designed so that participants could capture data from a specific selection site or divide the overall data into chunks, whereas we wanted participants to record all elements in the same field (a single site of data). Additionally, we believed we wanted to avoid extensive mousing and data selection by drawing boxes since our initial material was narrative in nature; we wanted to minimize user-interface interruptions to the actual act of transcribing (e.g., we wanted the user to be able to keep his or her fingers on the keyboard for the majority of the time). The research into the ways other digitization tools operated helped shape our service-oriented approach and allowed us to maintain a high degree of extensibility in our development.

While rapid development was not a stated goal of the TC, being able to develop and launch the site quickly allowed us to enter an interactive life cycle sooner. The TC came to life in a little under three months from initial project kickoff on April 4, 2013, to the initial public release on June 15, 2013. The quick timeline demanded an agile, iterative approach to make the system a success.

Project Personnel: Dividing Technical Teams

Technical tasks were divided into three main concentrations: data architecture, data interactions, and user interface/user experience (UI/UX). These three concentrations necessitated the creation of two distinct technical but complementary teams: the UI/UX design team and the transcription service data team. Following release, project administrators and digital volunteers additionally contributed to further and shape technical direction in significant ways.

The project was initiated by the Smithsonian Office of the Chief Information Officer, Library and Archives System Support Branch, and managed from

that office by Ching-hsien Wang. Lead software developers were Andrew Gunther (architecture/transcription) and Michael Schall (UI/UX). Dividing the labor between these two software teams enabled the rapid development of the TC.

Using two "tech leads" allowed us to move forward with twice the bandwidth if only one lead were in place: a clean separation of domain knowledge allowed the project to develop from two ends and meet in the middle. The project development was in large part completed with two dedicated resources, supported as needed by a team of four; these additional team members helped with specific tasks and in testing.

Just after launch of the TC in June 2013, two Presidential Innovation Fellows (PIFs) from the White House joined our efforts. As a third team, they contributed to additional focus on engagement and interaction with internal and external groups. The PIFs brought an iterative, agile methodology to the project to enable continued growth. They helped to bring the confirmation of our intent that our model could support different input formats; these were realized as a simple implementation of asset templates to the existing infrastructure. PIFs brought a focus on messaging and engagement to the public, a piece of critical importance to the continued success of the TC.

With the addition of this third cohort, the three teams involved in this initiative helped set the stage for the growth of the TC and remained in communication with one another for the duration of the six-month PIF tenure through December 2013. Key developments during this period included the creation of fielded data entry via a template, workflow and navigation improvements, and implementation of introductory instructions on first encounter with a project.

Leading into the final months of the beta, or trial, period from January 2014 to July 2014, the TC experienced iterative development. A number of features were introduced during this time period as a condensed development team addressed volunteer needs and staff preferences that surfaced via feedback e-mails and social media. These improvements included integrating a TC project coordinator, attention to the presentation and performance of the logbook template, an expanded My Work section of the site, more robust metrics for staff use, the Administrative Dashboard for staff project management, and new self-service import pathways.

Developing an Underlying System Architecture

Initial technical planning addressed the structure of the data repository. Records to be transcribed would be added as "assets," and each asset would be grouped under a "project." Projects could be grouped together into a hierarchy of parents and children—a decision that proved to be of value later in helping to break up large efforts into more manageable chunks. These hierarchies often reflect real-world groupings of individual volumes in a series of books or materials grouped into organizational categories (such as specimens by state). The model of assets and

projects allowed transcription efforts to be integrated into SI's many collection information systems (CISs) by synchronizing metadata between systems. Each asset consists of a media reference (typically a link to the source image) and is assigned a template that describes the way in which the asset is to be transcribed and how that asset should be displayed on the site.

To create a system for these records and new data that would maintain relationships and support fluid export operations, a relational database management system (RDBMS) was employed. RDBMSs are under the hood of most CISs used by SI; these include popular off-the-shelf solutions, such as Mimsy XG, EMu, and TMS (The Museum System). Relational databases include MySQL, SQL Server, Oracle, Postgres, and Sybase. These are popularly used as data stores for all types of applications. Relying on an RDBMS for TC minimizes the knowledge that a new development contributor would need to know in order to enhance the TC in the future. Based on the traditional and well-tested model of RDBMS, the platform was designed with an emphasis on simplicity to permit later customization for a variety of users and purposes and, ultimately, to facilitate even larger crowdsourcing efforts.

The most challenging design aspect was supporting the variety of content and optimizing presentation of the interface to volunteers. For example, handwritten letters might require simple transcriptions in a single text field, while specimen cards might require identifying the plant, serial number, and botanist. Data collection had to be extremely flexible but structured enough to be harvested in a meaningful way for later—and multiple—uses. To accomplish this, transcribed content is submitted back to the database in the form of JavaScript Object Notation (JSON), a lightweight data-interchange format. This makes it easy to extract content back into the individual CIS tools used throughout SI.

In addition to the JSON containing the information that is transcribed for each asset, we also assign an asset template to each record. These templates allow us to do three things: render a form to collection input, collapse the input into a form for storage, and render the final page for a completely transcribed asset. This flexibility allows us to collect some elements in different ways (such as collecting the discrete components of a latitude or longitude in the UI but collapsing them into a single value for storage). To help bring order and structure to the JSON input for each transcription, the asset templates dictate the intended feel of the form, providing instructions on what form fields to use, what type of help text to attach, and what auto-completion text and validation options to include. These template requirements are defined before a project is submitted and allow us to maximize efforts in transcribing materials from handwritten letters, botanical specimens, and banknotes to tabular ledgers. Modifications can be made to the template at any time to suppress data or solicit volunteers for more details without major structural intervention or system downtime.

From a data architecture standpoint, the introduction of JSON-stored data yielded a simplistic table design that necessitated and allowed any type of data to be stored in one database field instead of many specific tables with ever-changing

fields and data types. Embracing the flexibility of NoSQL methodologies with the proven, transactional support of an RDBMS, the simplicity of our JSON-based model allows us to interact with XML-centric systems as needed through traditional Data Access Objects.

With this model in mind, the technical transcription service data team began building the necessary systems, while the UI/UX design team designed the user experience. Both teams worked independently while maintaining an ongoing dialogue. Issues facing one team were quickly resolved with input from the other. The transcription service data team focused on implementing data architecture and interaction services, while the UI/UX design team focused on Drupal and the presentation of the TC.

Implementing Services to Make Transcription Happen

The transcription service data team selected a service-oriented architecture as a means of providing an infrastructure in which client applications could build graphical interfaces on top of basic functions that interacted with transcription tasks. These tasks were roughly defined in the functional requirements document but were fully realized during the development process in which the UI/UX design team used wireframes to convey how the interface looked and interacted. The idea to decouple responsibilities resonated through many parts of the project. In terms of the transcription service, this meant providing simple HTTP end points to provide actions.

An example of an action is returning metadata about a project along with statistical information (number of transcribed assets, number of contributors, etc.), as well as all attached project assets. The client application could then make an HTTP call to a specific end point passing in required parameters to tailor the returned metadata, stats, and assets further. The end point responds with a JSON object which is ingested by the client application to produce user interface seen by volunteers. Simple HTTP end points afforded the flexibility of tweaking the technical components behind the service, perhaps changing database structure or implementing new technology while continuing to yield a constant response format.

This opens the door to utilizing new technology without having to rebuild the entire TC. As an example, we use an RDBMS in a very NoSQL-like way; we could move to a true No-SQL solution in the future and leave our entire front end intact. This service-oriented approach also gives us the ability to offer simultaneous transcription experiences on different websites, both populating our system.

Designing the UI/UX and Drupal Platform

Drupal is an open-source Web-based content management system (CMS) that was gaining traction within SI as more and more Drupal websites went live. A degree

of institutional knowledge around Drupal made it easier for more people to use and contribute to the project. Drupal also makes for a rich, extensible application framework providing virtually everything needed to deploy a dynamic, accessible UI to facilitate transcription activities. The materials hosted in the TC cover a wide range of different formats, each needing different UI components to minimize friction with users (the TC volunteers) who are transcribing the materials. As a flexible, open-source system, Drupal is easily extended and customized to support that range of materials within the TC.

It was clear that we could leverage much of what is included within the core of Drupal to facilitate account registration and maintenance, menu management, page layout customization, and system administration. Custom modules would allow Drupal to use the transcription services for sending and receiving information to produce the displays. Drupal modules could be distributed as stand-alone pieces to Smithsonian units desiring a custom look and feel, perhaps catering to another audience other than the flagship volunteer transcription platform. Drupal also provided a low barrier of entry for system administrators, who could easily manipulate themes and menus.

Most of the responsibility of addressing diverse functional requirements fell to the UI/UX team. Although providing an interface that was inviting and that could solicit digital volunteers to contribute was critical, equally important was creating a workflow that could manage review and acceptance of transcribed text. Furthermore, the system had to be flexible enough to allow for multiple workflows to exist because units would have different requirements for approving submitted text. Both staffing and submitted content type also influenced the process by which completed data were approved. The team developed different workflows for digital volunteers to transcribe and review data and for administrators to perform final review and approval. These workflows were continually revisited during the iterative development process.

The Final Product

By the time of project launch in mid-June 2013, the technical teams—transcription service data and UI/UX design—yielded a decoupled HTTP-based transcription service with end points to facilitate user tracking, asset transcription, and the early stages of administrative controls, such as self-service project management, asset final review, and export controls. These services were utilized primarily by the TC but could be adopted by any Smithsonian unit with similar needs. In addition, an extremely modular platform was built using Drupal Core components. This further extended the value of the product, lowering the technical entry barrier and fitting into the current Smithsonian Web ecosystem.

At launch, 20 transcription projects were contributed by Smithsonian libraries, archives, and museums, consisting of a single material type (handwritten manuscripts, diaries, vocabulary cards, and field notebooks). In the following two

years, the system grew quickly in project number and project complexity and supported numerous material types. As of July 2016, the system supports 1,606 projects and 192,677 available assets with 190,317 pages transcribed and 6,671 active volunteer participants.

Discussion

In the end, the technical team yielded a platform that is extensible, versatile, and adaptable to change and that integrates with existing Smithsonian technological infrastructure. The final product also provided a centralized place for disparate projects to coexist, but such a system could also be used by an individual museum or archive. This flagship site allowed the digital volunteer to witness in one place the depth and breadth of the Smithsonian's holdings, from bumblebee specimens to astrophysical logbooks and artist diaries. The platform's design provided a vehicle to share these collections while furthering the institution's mission "to increase and diffuse knowledge." Success of the project was due to careful project planning, thoughtful system design and project management, and the strong partnership with museum staff and digital volunteers.

Since inception, additions and modifications have been made to the initial platform, but the data architecture, service-oriented data layer, and UI layer have remained constant. We believe this speaks to the strength of the initial design, and we highlight the following aspects of project development that likely contributed to its success.

1. Generic, iterative design: A major challenge was balancing functional requirements dictated by multiple clients (i.e., individual museums, libraries, and archives) with the greater SI mission. Institution-wide technical tools are largely precluded, as data exist in disparate silos created by different requirements. It was important to take a generic approach to the design process, identifying shared needs while permitting variation in individual workflows. Iterative design proved to be beneficial in addressing detailed needs as they arose. This allowed us to move through the development process without detouring from shared, initial goals.
2. Choosing the right tools: While the overall approach had to be generic, it was critical to become fluent with the technical specifics. Evaluating and learning from similar technical products was an important part of the process. Choosing the right technical tools would also allow growth and flexibility throughout the future life of the product. While resource and time limitations influenced technical decisions, careful tool selection following an honest appraisal of current capacities allowed us to build a successful custom product within our guidelines.

3. Public engagement: The rich and active public dialogue spawned from this technical product was never anticipated by the technical team. Highly visible on social media with accounts and activity starting in January 2014, such conversations continue to address all aspects of system involvement, from technical issues to content meaning and interpretation. Today, there are more than 7,100 followers for the @TranscribeSI Twitter feed. Although not implemented as part of the original development process, this robust communication with audiences creates opportunity for continued technical development. The TC project coordinator works with social media teams from the participating SI groups through cross-promotion to connect the public with projects and swiftly address technical issues that arise during transcription. In fact, the success of the product—in terms of visibility, adoption, a forum for help, a space for community, and camaraderie—would have never been achieved if the technical team had worked wholly within the platform and without this input.

Summary and Future Directions

Despite SI being a (very) large organization, we were able to pull off a technically challenging project with a small and dedicated team. We selected technologies and methodologies that facilitated rapid, iterative development and minimized task overlap. Nothing we did was dependent on SI's infrastructure. Anyone with knowledge of CMS workflows could implement a similar system using similar tools. What makes the TC special is that our thoughtful, small-team approach met the requirements of a (very) large organization. This approach can be adopted by teams at small, medium-size, and large organizations to maximize the realities of resources and time scales for projects.

The reality of the TC is that it continues to grow as collections are added in the guise of new projects that are uploaded every week. Today, the TC supports a narrow set of transcribed assets, but the flexibility of the system means that new asset types can be supported in the future, including audio and video transcription as well as guided, or decision tree, transcription. The design of the TC is flexible and modular to adapt to the needs of internal and external participants.

The success of the TC is guided not necessarily by a set of technical principles but rather by having been designed with a holistic approach. Building a platform does not work unless you have material to put there; likewise, having material and a platform does not help engage and retain volunteers. Crowdsourcing needs to be about the crowd: both the internal crowd of staff hoping to experience success and the external crowd of participants who want to learn, explore, and help SI improve. By considering the range of needs of internal participants, creating simple and flexible UI for best UX, and listening to the needs of those using the system, the

small teams tasked with developing the TC were able to create a reliable model for crowdsourcing and, ultimately, engagement at every level.

Note

1. See Smithsonian Digitization Dashboard, http://dashboard.si.edu/digitization.