Using SharePoint workflows and InfoPath forms to manage a large-scale digitization project: A case study of the Kissinger Papers Project

Kevin L. Glick, Head of Systems and University Archives, Manuscripts and Archives, Yale University Library, New Haven, CT
Rebecca Hirsch, Kissinger Project Archivist, Manuscripts and Archives, Yale University Library, New Haven, CT

Abstract

This paper will describe the designing and implementation of a workflow management system tracking the mass digitization of archival papers using off-the-shelf applications (SharePoint and InfoPath) and an outside vendor service (Qdabra). This paper is a case study of the Kissinger Papers Digitization Project at Yale University Library. In addition to describing the process of development, the paper will provide an in-depth description of one form's functionality and provide suggestions for future use of the system at other institutions.

Background

When the staff of Manuscripts and Archives and Yale University Library began planning the mass digitization and paper duplication of the two collections of the Henry A. Kissinger papers held at Yale University, they encountered a number of complicating factors that led the group to the creation of an automated workflow management tool to track all aspects of the project, from shipping boxes to vendors to the ingest of the digitized images into the digital repository. These complicating factors included: the shared ownership of one of the collections with the Library of Congress Manuscript Division, the large size of the two collections, and a relatively short time frame in which the very large high profile project could take place.

Yale University Library has two Kissinger papers collections, the larger of which is jointly owned with the Library of Congress Manuscript Division. When they were in Dr. Kissinger's possession, the papers formed an integrated collection. To maintain that integrity, as well as increase access to these materials, Yale and the Library of Congress decided to undertake a joint project to arrange, describe and digitize all of the archival materials together, with the work undertaken by Yale staff and contractors. At the end of the project, each institution will hold both paper and digital copies of the jointly owned collection. In order to create a second paper copy of the collection, the digitized images were printed and interfiled with the original documents, creating two copies of the collection identical in box and folder numbers. One of the collections contains all of the originals owned by Yale University and copies of the documents owned by the Library of Congress, while the other consists of all originals owned by the Library of Congress and copies of those documents owned by Yale University. This was necessary to allow each institution to retain its originals and copies of the originals owned by the other institution, while not disrupting the intellectual original order of the collection. The work necessary to create these two digital and paper copies of the collection has been very time consuming and labor intensive, but also has been undertaken with sensitivity to the nature of the materials and a very sharp attention to detail.

One of the most important factors influencing the need for a robust project tracking system was a wish for very high quality. The high profile of the collection and the two institutions' very low threshold for error necessitated stricter quality assurance procedures on a much larger scale than anything the Library had attempted in the past. The directive of the team was, as much as possible, to produce two absolutely identical collections.

The second major factor motivating the creation of the workflow tool was the large size of the two Kissinger papers collections held by Yale University. The collections span 15,728 folders in 1,162 boxes and consists of paper archival records like correspondence, memoranda, writings, speeches, photographs and other material. These materials document the career of the diplomat, author and foreign policy expert and scholar Henry A. Kissinger, who served as United States Secretary of State from 1973 to 1977 and as assistant to the President for national security affairs (National Security Advisor) from 1969 to 1975. Digitization of the two collections will create over 1.5 million high resolution master tif images, pdf use copies, and OCR text files.

A third major factor was that there were funder and other stakeholder requirements to undertake this mass digitization and to provide online access in a relatively short period of time given the scale of the project. The time allotted for the digitization, quality assurance, interfiling and ingest of these approximately 1.5 million page images was approximately 15 months.

All of these factors led the production team to seek a robust digitization project tracking system to allow many different staff to work on many different tasks in different places simultaneously and to have them supervised and managed by very few staff, while still adhering to stringent quality standards.

Because the specific department tasked with undertaking the work, Manuscripts and Archives, did not have any tools to manage this large and complex project, and other library technology resources and expertise were limited, the team decided to work with external consultants and vendors research possible solutions and eventually to build a workflow system to manage all aspects of this digitization project.

Initial Functional Specification and Landscape Review

The project began with the definition of the functional and system requirements of the project, an evaluation of possible existing solutions and of the procedures and tools employed by other similar projects at different institutions. All of the potential workflows of the project were discussed in-depth, diagrammed,

and analyzed to understand the potential system requirements implications, and the necessary tracking and quality assurance data.

The Library conducted an in depth analysis of the landscape of four different possible different solutions: the Taverna open source and domain-independent workflow management system designed for scientists; Activiti, a light-weight workflow and business process management platform targeted at business people, developers and system administrators; Digital Assets Factory from Bibliotheca Alexandrina, a relational database application that is not based on a workflow engine, but is a desktop Java application on top of a MySQL database; and finally various versions of SQL Server and SharePoint database and workflow services, a technology stack that was already supported at Yale and specifically inside the Library.

Eventually, after considering a number of possible options, the project team decided to use a custom-built combination of Microsoft InfoPath forms and workflows stored and managed by SharePoint 2013 rules, tasks, and workflows, while simultaneously storing the much of the data needed for interaction with other systems or complex reporting in an external SQL Server database. InfoPath Forms Services enables a browser-enabled InfoPath form to be hosted on a SharePoint installation and rendered as an HTML page with client-side script and post back behaviors similar to an ASP.NET page.

Some of the other existing solutions fulfilled some or even many of the functional requirements. However, the decision keyed largely on the existing staff technical skill level and ability to build a robust, highly customized solution with only minimal time from programmers and central IT systems support.

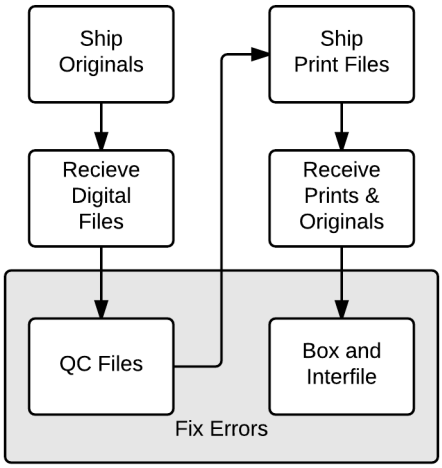
There were several key factors that lead to the decision to go with the Microsoft solution stack. Yale and the Library already had an existing commitment and staff familiarity with Office and SharePoint. SharePoint workflows allowed for robust design, equal to other solutions, but with tools that eased development, required less programmer time; and reduced development time. InfoPath Forms were determined to be dramatically easier to develop and deploy than other options, like Java Swing forms. Technically savvy non-developers could modify them.

However, while technically savvy non-developers could modify InfoPath forms, the initial development process can be complicated, particularly for feature like sorting tables, populating repeating tables from secondary sources or submitting InfoPath form data to a SharePoint list. In addition, there are some specific characteristics of InfoPath and SharePoint that can make customization of forms with the pushing of data to external SQL tables a bit slower and more complicated. We discovered a third-party vendor that had built two different products to mitigate these issues. Qdabra had produced qrules, an InfoPath add-on that makes it much simpler to add functionality to InfoPath forms. There is no need to write code or even understand our code to utilize this tool. In addition, Qdabra had created a Database Accelerator (DBXL), a single web service that operates in between InfoPath, SharePoint, and SQL Server making it much easier to maintain a complete set of data table relationships at the same time. In the end, due to severe time constraints and limited system support staff time, we decided to enter into a consulting contract with Qdabra to collaborate with Yale staff on design and to build the forms and SharePoint site; as well as a maintenance contract to provide ongoing support for the entire site for the life of the Kissinger digitization project.

In-depth Workflow Analysis and Form Design

Once the Microsoft software stack of InfoPath and SharePoint had been selected, Manuscripts and Archives staff began the design process by envisioning and diagramming the workflows that would need to be handled by the system. Surprisingly, this turned out to be one of the most challenging and time-consuming parts of the design process. The two staff members responsible did not have previous experience with large scale archival paper digitization projects and the expertise that did exist elsewhere in the organization mostly came from mass digitization of books, a process that proved to be quite different in a few key aspects. In particular, the team struggled with the initial conceptualization of quality, what quality could possibly be achieved in a large collection of heterogeneous paper records, and what impact this had on the process. Regardless of the time spent on this conceptualization, the effort was vital to the design process. Without it, the forms and site design process would have been much more difficult and may have not achieved all of the intended outcomes.

Figure 1. *High-level workflow diagram*



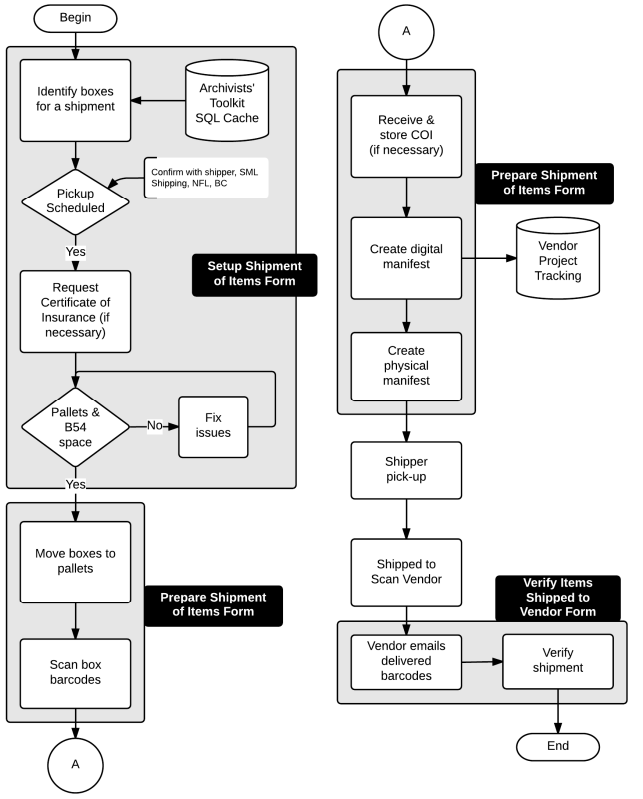
During the conceptualization and design work, a separate flow chart was created for each sub-workflow, to assist with form design and database construction. At a macro level, the project's workflow can be understood as seven separate sub-workflows (illustrated in Figure 1):

1. In the Ship Originals workflow, archival materials are selected for shipment to the vendor; prepared and packaged for shipment; and successful delivery of all containers is verified.
2. In the Receive Digital Files workflow, digital images of originals are received from the vendor.
3. The QC Files workflow describes a two-level quality assurance process. A series of automated checks are performed on the received drives, including file naming, file counts, expected file types, and simple validations. If a drive or entire folder on a drive is failed, the vendor is informed and required to fix the

errors and re-submit that portion of the drive in a future shipment. If an entire drive, or a substantial portion of it, passes the automated check, the workflow tool guides in the selection of a random sample of digital folders from the drive, and the creation of tasks for manual quality control checks. This was an important feature as almost 16,000 archival folders will result in over 1,600 separate quality control tasks once the project is completed this summer. Once these tasks are completed, the system manages the process of supervisor review, failing quality control and sending notice to the vendor, or passing quality control and advancing on to the subsequent workflows.

4. In the Ship Print Files workflow, batches of folders that have passed quality control are aggregated and identified to be included on drives for vendor printing, with file manifests for the drives being automatically generated.
5. The Receive Prints and Originals workflows are really two distinct processes that utilize similar logic. Folders of the printed copy are delivered from the print vendor. Separately, boxes of archival originals are delivered from the digitization vendor. A set of workflow triggers in the system automatically audits the approval of such transfers. The system helps track the shipments and verify the delivery of the containers for auditing purposes.
6. In the Box and Interfile workflow, the print copies that were received in temporary boxes in the previous step are reboxed into their permanent archival boxes. After reboxing, the print copies are interfiled with the original collection to create two copies of the collection. The workflow system helps the supervisor to create tasks for staff from candidate boxes that have passed the previous workflows, and helps manage the process of staff work, approval, failure, and/or supervisor review of each one. By the end of the project there will have been about 760 reboxing tasks and over 11,000 interfiling tasks.
7. Any errors that are identified in the Box and Interfile workflow, must be dealt with in the Fix Errors workflow. The system documents the decision to fix errors in-house or to instead send the originals back to the digitization vendor to restart the process. The system also helps with the creation of tasks for staff to undertake the in-house rescanning, reprinting, renaming, and refiling.

Once the workflows were conceptualized, the InfoPath forms were designed, the connecting workflows were created, and the SharePoint site was created. The form design process began with Yale staff creating wireframes outlining what each form should look like and describing how they would interact in the conceptual workflows. This included narrative descriptions about each form's functionality and what SharePoint and/or SQL database fields the form would draw from and update. Because project time was severely limited and funding was available, the consulting vendor Qdabra was then contracted to build the forms and the underlying SQL database in which all the information generated by the forms was stored. This system allowed project staff with an intimate knowledge of project workflows, but only basic knowledge of database design and wireframe software to create a robust and dynamic system. A screenshot of the resulting system home screen



is displayed as Figure 5 below. It shows some of the different workflow grouping, with active task to allow for the project manager and student staff to quickly see and understand the work at hand.

Example Sub-workflow: Ship Originals

Figure 2. *Detailed workflow diagram for Ship Originals*

To further illustrate the system, this paper will examine one particular sub-workflow in-depth. The first process in a digitization project involving an outside vendor is the shipment of items selected for digitization to the digitization vendor. The major steps involved in this workflow are the selection of items for shipment; the preparation of items for shipment and the verification from the vendor that what was shipped was what they received. After outlining the process, it became clear that each of these steps necessitated their own form: Setup Shipment of Items, Prepare Shipment of Items, and Verify Items Shipped to Vendor. (See Figure 2. Forms are indicated by gray areas. Black boxes indicate form names.)

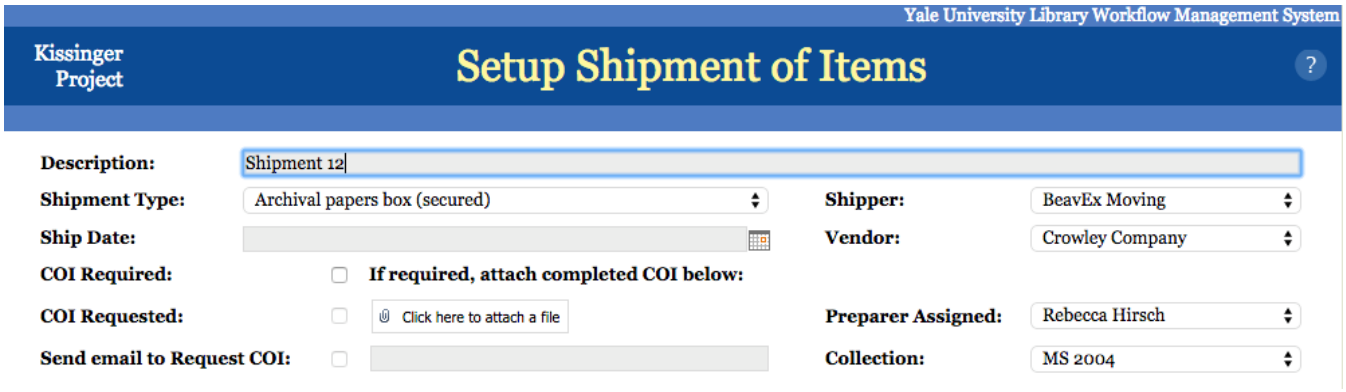


Figure 3. *Shipment creation*

The Setup Shipment of Items Form is the first form of three in the Shipment of Materials to Vendor workflow. This form only has one view and aims to trigger a shipment workflow in SharePoint. When the form is loaded, it automatically queries items from the database to populate the drop-down field values and call numbers and sets the form's persistent ID and creator. If the form is newly created, it will then query the potential container or item candidates. In our case, this data comes from Archivists' Toolkit.

On a newly opened form, the shipment creator is prompted to: provide a description; choose the boxes to be shipped; schedule the shipment; obtain and load the certificate of insurance, or COI (if required); and define the shipper, vendor, shipment type, preparer assigned and collection (see Figure 3). Selecting a collection from the drop-down will populate the table with a list of available candidates.

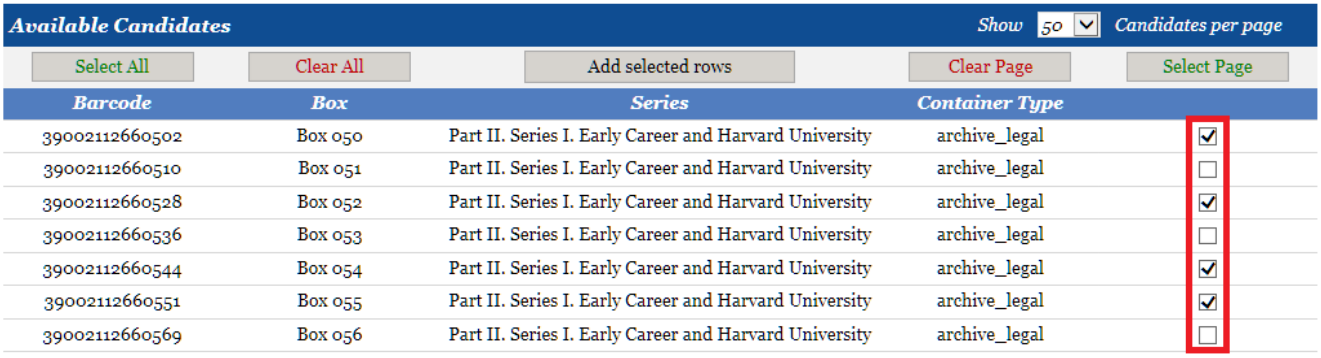


Figure 4. *Selection of candidates*

Once added, the form will display a table with a list of the selected candidates, which gives the shipment creator the option to remove candidates from the selection. An initial Save submits the form to the SQL database, with any new entry updating the tables. Simultaneously, the data is submitted to the SharePoint library. If an entry is saved with a request for a COI, submitting to SharePoint will trigger a workflow action to send an email to the person in charge of the COI, along with a request to attach the COI and reply to the person who initiated the request. The COI request is then logged to the workflow history. Before submitting, the shipment creator first has to confirm authorization from the building shipping manager and the department shipping manager. Once both checkboxes are ticked, the Submit button is automatically enabled. The Submit button on the other hand submits the form to the database and SharePoint, or updates the items if the data is already existing. This action simultaneously triggers a SharePoint workflow to send an email notification and assigns a task to the preparer assigned, which is logged in the workflow history. The completion of this leads to the next step in the process, the preparation of items to be shipped.

Further use of System and Conclusions

In the end, the development of this workflow management system can be seen as a significant success, one of the most successful aspects of the entire Kissinger Papers digitization project. The system is currently being utilized to manage the very complicated workflow of the ongoing digitization project, with completion anticipated by mid-summer. At beginning of April 2015, over 1.1 million pages had been scanned and passed initial quality control from 11,059 archival folders in 700 archival boxes. Of that total, 8,598 folders have been printed by the vendor, with 5,515 folders re-boxed and interfiled into two separate collections (one for Yale and the other for Library of Congress). All of this work has been completed in less than a year.

In assessing the success of the project, it is important to consider some of the issues encountered during the work, as well as some aspects unique to this project. Due to the organizational nature of the Library, the project was very slow to coalesce and move forward at its inception. There was no single point of contact or consultation on digitization projects or on systems development. This led to a somewhat inefficient process of committee meetings

and group work that can be common in very large organization like Yale University Library. Also, while it was valuable to consider a full range of possible workflow management systems, the group spent much too much time getting to our final decision to build a custom system in InfoPath and SharePoint with Qdabra consulting. The rapid development that occurred after this decision showed that we could have saved several months if we had reached such a decision much earlier. Also, because we hoped that the system would not be developed as a one-off solution, much energy was expended considering many different types of digitization workflows, some of which were not yet fully conceived and most which did not face the time pressures of this project. This somewhat slowed down the development of the functionality necessary to deal with the unique nature of this project.

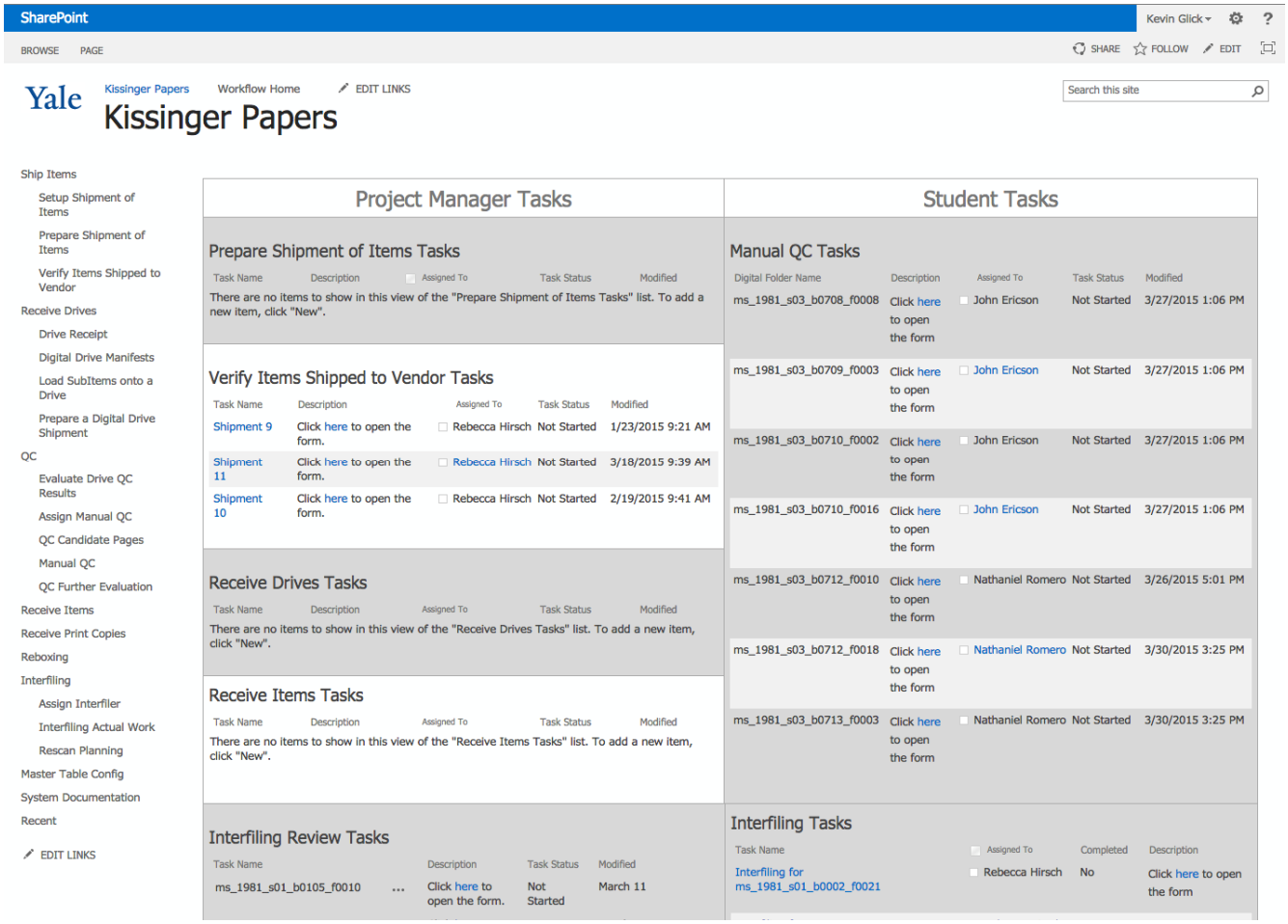


Figure 5. *Workflow tool homepage*

It is important to also note that Yale University Library's reality when faced with a project like this one, is likely different than many other institutions. The scale, complexity and high-profile nature of this project is unusual. This allowed us to focus significant attention to this project, allocate significant staff time, and attract the attention of participants across the Library. This project also came with its own funding source and a very firm and accelerated timeframe for completion. With time and available systems staff the only two limiting factors, the project was able to move ahead faster than might normally be possible by spending extra money. This lead to perhaps a disproportionate resource allocation devoted to consultants versus internal staff as well as a custom-built system with proprietary software versus a collaboratively built open source system. These may not be the same choices made by another institution, depending on their own unique situation.

We believe, however, that this work can be re-purposed. In the Library the technology stack has been built upon and implemented in a second library unit, to manage the production workflow for the digitization of brittle books. This success leads us to believe that this work may be to be taken up by other institutions trying to build similar systems. However, there are a few issues that any institution needs to know before proceeding. One needs to fully understand one's own workflows before attempting to work on the forms or the site design. This can take time. Don't short-change this step. Don't try to make each form more complicated than necessary. Imagine how workflows can be streamlined and simplified. It is easier to simplify a workflow than to make any particular form contain too much logic. This can lead

to unforeseen errors that are difficult to test before implementing the production system.

There are also a few key things that an institution needs to have in order to implement a system like this. The organization needs to have some level of commitment to, and familiarity with, the Microsoft Office, SQL Server, and SharePoint technology stack. There should be some pre-existing data that can be extracted about the containers or items that are to be managed in the workflow. In Yale's case, the data was stored in the Archivist's Toolkit. Other institutions may not have such data. Perhaps most importantly, there must be either a production system support team or a funding source to purchase such services directly from the vendor. In Yale's case, we were unable to allocate internal support for the production system. We instead contracted with the vendor to support the system. This is an important aspect as the more complicated and larger the system gets, the more likely unforeseen complications may occur during the heat of production work. One does not want to be hampered by minor issues that can slow down the work that this system is meant to support.

We plan to share all of the workflow diagrams, wireframes and narrative specs, InfoPath forms with description of qrules, SharePoint workflows and design description, and database design in a GitHub repository project to allow others to build on this work.

Author Biography

Kevin L. Glick is the Head of Digital Information Systems and University Archives at Manuscripts & Archives, Yale University Library. He has been at working in various positions related to archives and technology since 2002. Before Yale, he was a project administrator and researcher for the International Research on Permanent Authentic Records in Electronic Systems (InterPARES). He has Master in Library Science from the University of Albany, SUNY and a Master of Arts from Western Michigan University.

Rebecca Hirsch is the Kissinger Project Archivist at Manuscripts & Archives, Yale University Library. Formerly an archivist at the University of Southern California, she has a Master of Science in Library and Information Science from Long Island University and a Master of Arts in History from New York University. She is a member of the Society of American Archivists.