https://doi.org/10.2352/issn.2168-3204.2022.19.1.13

Mass Digitization with Smartsheet: Leveraging a Commercial Solution for Flexible Project Management

Emma Stanford; Hoover Institution Library & Archives, Stanford University; Stanford, CA

Abstract

The Hoover Institution Library & Archives (HILA) has implemented Smartsheet, a cloud-based project management tool, to manage tasks and cross-team handoffs for its new mass digitization program. By combining task-specific tools such as Capture One and LIMB Processing with the administrative flexibility of Smartsheet, HILA has succeeded in leveraging commercial project management functionality for cultural heritage purposes, resulting in improvements to our program's efficiency, flexibility, and reporting capabilities.

About the Digital First Initiative

The Hoover Institution Library & Archives at Stanford University (HILA) recently launched an ambitious Digital First Initiative (DFI), aiming to make full archival collections or "fonds" available online through mass digitization. This initiative is the latest step in a long history of reformatting and digitization activity at HILA, starting with the microfilming of entire archival collections and shifting in recent decades to targeted digitization of high-profile objects. To manage digitization tasks before DFI, HILA used a combination of Trello boards and individually-maintained spreadsheets. This worked because the team was small and the throughput low, but there was no visibility of progress outside the digitization team, and no easy way to track progress within a team. DFI represents a return to collection-level digitization, with an expanded team of photographers and new protocols for description, imaging and ingest, and an ambitious set of throughput goals.

Under DFI, each archival collection is approached as a separate digitization project, with distinct requirements depending on condition, material types, processing status, and donor obligations. The digitization workflow begins with an assessment of the collection's needs, with the outcomes of this assessment determining the timeline, sequence of work, and resources required. Subsequent stages in the workflow include archival processing, conservation, imaging, ingest package creation, and ingest into the HILA repository. For several of these stages, specific software tools are used (figure 1). The description team uses ArchivesSpace [1] for describing archival collections, the imaging team uses Capture One [2] for creating and exporting images, and the ingest team uses LIMB Processing [3] and LibSafe [4] for creating and validating ingest-ready information packages. HILA needed a tool that would join up these team-specific tasks into a unified workflow, where objects, instructions and questions could easily move from one team to another; and we needed a tool that was flexible enough to meet the unusual needs of archival digitization.

The central difficulty of setting up our digitization workflow was the concept of archival hierarchy. Like many archives, HILA follows the multilevel description model put forward by standards such as DACS [5] and ISAD(G) [6]. Each collection is described as a whole, and then—time and resources permitting—at the level of its parts, which include increasingly specific levels of series, subseries, boxes and folders. This hierarchical structure presents problems for the end-user's discovery of digitized archival collections, because most of the available solutions for displaying digitized objects online assumes that the most detailed description is available at the object level. However, it also presents logistical problems during the digitization process. At HILA, boxes are barcoded and tracked in a circulation system, but folders within boxes, and individual documents within each folder, are not. In most cases, each folder is digitized as a separate object, analogous to a book or artwork. To get to this point, however, much work needs to happen at the collection, series and box level, as the folder structure needs to be finalized and identifiers need to be applied. A linear, object-first digitization management tool such as Goobi [7] would not fully meet our needs. We needed a solution that would support the tracking of a branching, hierarchical set of tasks over the course of a project.

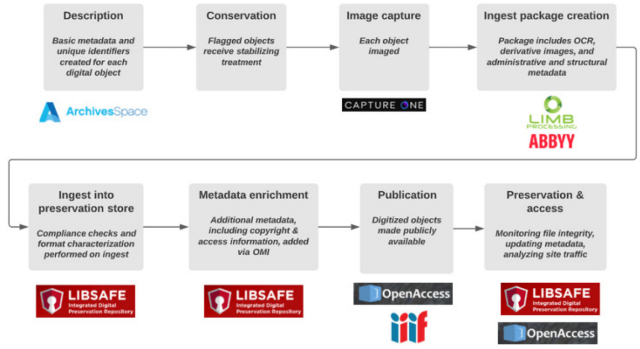


Figure 1. *Digital First Initiative workflow, showing software tools used in each step.*

Other requirements for a digitization management solution were informed by HILA's staffing needs and reporting obligations. Because of the unique needs of each collection and the relatively small size of the HILA digital team, we needed to be able to adjust our workflow quickly and transparently, without having to rely on a vendor or an in-house developer to make changes to a behind-the-scenes codebase. We also needed to be able to surface detailed metrics from both ongoing and completed projects, firstly in order to satisfy donor reporting requirements but also in order to identify efficiencies and bottlenecks in this brand-new digitization workflow. Finally, we needed a solution that would be robust and resilient, rather than adding to the burden of our overstretched system administrators.

To meet these needs, it was clear that we needed a tool more flexible than Goobi and more powerful than a spreadsheet, so we looked at project management tools such as Smartsheet [8], Trello [9] and Jira [10]. Two final factors made Smartsheet the preferred option of these: Stanford University has an active community of Smartsheet users in other departments, and Smartsheet was already in use at HILA to manage collection-level accessioning, making it a familiar tool to some of our users.

What Is Smartsheet?

Smartsheet is a commercial cloud-based tool for collaborative management, marketed to software companies and other sectors that have adopted agile project management. The central unit of any Smartsheet solution is the "sheet", a collection of tabular data that can be viewed as a spreadsheet, a Gantt chart, or a set of Trello-like cards. In addition to sheets, Smartsheet offers reports (filtered aggregations of rows and columns from different sheets) and dashboards (graphical displays of metrics, charts, links, and embedded reports). Alongside these basic elements, a set of premium integrations offer additional data management options, such as pivot tables, customizable calendars, and automatic data import and export workflows.

The basic spreadsheet structure of Smartsheet was familiar to many HILA staff members, some of whom who had experience tracking digitization and other tasks using a system of Google or Excel spreadsheets. A few key functionalities set Smartsheet's spreadsheet offerings apart, however:

1) Cross-sheet references. It is easy within Smartsheet to write a formula referencing a column or range of cells in another sheet, making it possible to aggregate and disperse data across sheets. This is crucial in order to, for example, count how many of the objects in box 45 have been imaged according to the object tracker and then display the total in the box list.
2) Filtered reports. A report in Smartsheet is a view of the data from one or many sheets, filtered and grouped according to the criteria you have selected. Reports allow our users to view the tasks assigned to them from across different collections, without seeing data that is irrelevant to their work.
3) Automations. Within each sheet, it's possible to set up automations to record a date, notify a person, update a cell value, or copy information to another sheet. With the premium apps DataMesh and Data Shuttle, it's also possible to set up automations to import or export data from or to an external spreadsheet, or to automatically update data in one sheet when data in another sheet changes.

These features were integral to HILA's implementation of Smartsheet for managing DFI.

The DFI Smartsheet Solution

Implementing the Solution

For 18 months during 2020 and 2021, the majority of HILA staff worked from home and were encouraged to focus on documentation and process improvements, including buildout of the DFI workflow. Two staff members, Emma Stanford and Erik Lunde, worked with a Smartsheet consultant procured through Stanford University to design and build the DFI Smartsheet solution. The consultant's involvement was minimal and mainly advisory, amounting to 30 hours over the course of five months.

We started by mapping out the project workflow with a flowchart and then breaking down the types of data that we would need to record and track at each stage of the workflow. Then, using a small collection (three manuscript boxes or about 100 archival objects) as a test case, we built out the steps to manage object-level tasks (image capture, QC, and ingest) before beginning to think about box-level and collection-level tasks. In building out the workflow, we focused primarily on tracking the points at which a task is completed or transferred from one person to another, while continuing to rely on task-specific tools (Capture One, ArchivesSpace, LIMB Processing, team-specific spreadsheets and documents) to track details about work completed within a task. This allowed us to take advantage of Smartsheet's process management tools without sacrificing the specialized functionality of task-specific software or duplicating an existing system for managing task-specific metadata.

Structure of the Solution

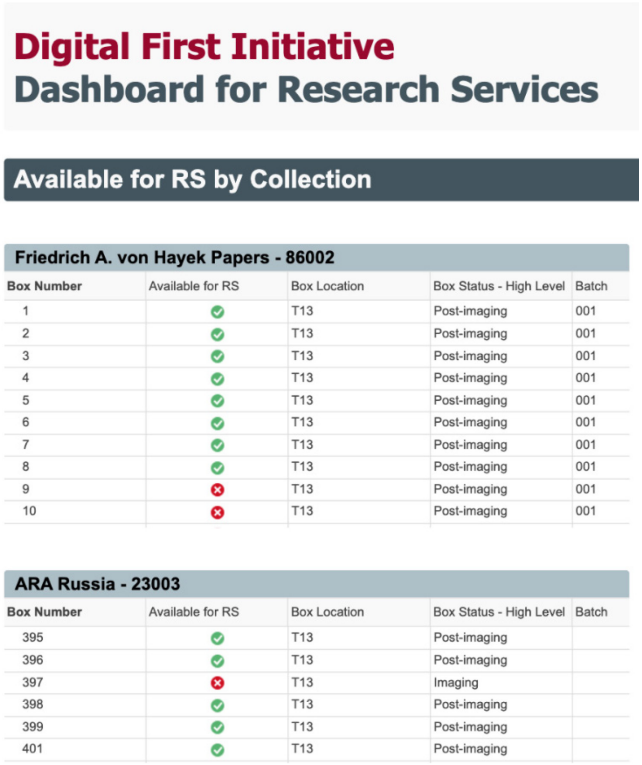
The DFI Smartsheet solution in its current form consists of five source sheets for each archival collection:

- An object tracker, with a row for each object in the collection
- A box tracker, with a row for each box in the collection
- A collection-level metrics sheet that pulls sums and averages from the object tracker and box tracker
- A time tracker sheet with a weekly fillable form
- A Gantt-style project plan

Rows in the object tracker and box tracker are populated via Smartsheet's Data Shuttle app, which allows the user to specify data to be added from an attached spreadsheet based on a unique identifier. Lists of box numbers or object identifiers are exported manually from ArchivesSpace and imported to Smartsheet using Data Shuttle. The box list and object tracker each contain between 20 and 100 columns to track progress and record project metadata, from a simple "Box Archival Processing Complete" checkbox column to an "Imaging Instructions" free-text column to a "Status" column that calculates the current workflow stage an object or box is at based on the values of other columns. Each sheet also contains a number of automations for tasks such as sending email notifications, recording dates, and copying a row to another sheet based on changes in specific columns. As an example, if a photographer working on an object checks off "Image Capture Complete", this triggers automations to record the current date in the "Capture Complete Date" column and send a notification to the person responsible for completing QC on that object. If an error is flagged during QC, this triggers automations to send a notification to the photographer and update the "Image Capture Complete" status to unchecked.

In addition to the collection-specific sheets, the solution contains "Task reports" that display filtered views of multiple sheets so that team members can view and edit the objects or boxes assigned to them. There are also several DFI-wide metrics sheets, which use cross-sheet formulas and Smartsheet's pivot app to calculate collection- and initiative-wide totals and averages. Finally, there is a growing set of dashboards that display key metrics, calendars and tasks, customized to specific audiences. For instance,

a dashboard designed for our lead photographer contains information about the team's productivity in terms of images captured, images QC-ed, and QC failure rate. A dashboard designed for HILA's Research Services team contains embedded reports from each collection being digitized, showing color-coded availability statuses for each box depending on where the box is in the digitization workflow (figure 2). The information on these dashboards is updated in real time, meaning there is no end-of-the-month or end-of-the-quarter scramble to gather data.



Digital First Initiative
Dashboard for Research Services

Available for RS by Collection

Friedrich A. von Hayek Papers - 86002

Box Number	Available for RS	Box Location	Box Status - High Level	Batch
1	✓	T13	Post-imaging	001
2	✓	T13	Post-imaging	001
3	✓	T13	Post-imaging	001
4	✓	T13	Post-imaging	001
5	✓	T13	Post-imaging	001
6	✓	T13	Post-imaging	001
7	✓	T13	Post-imaging	001
8	✓	T13	Post-imaging	001
9	✗	T13	Post-imaging	001
10	✗	T13	Post-imaging	001

ARA Russia - 23003

Box Number	Available for RS	Box Location	Box Status - High Level	Batch
395	✓	T13	Post-imaging	
396	✓	T13	Post-imaging	
397	✗	T13	Imaging	
398	✓	T13	Post-imaging	
399	✓	T13	Post-imaging	
401	✓	T13	Post-imaging	

Figure 2. *Portion of a dashboard produced for the Research Services team, showing boxes' availability for research use.*

As of writing, the Smartsheet solution is being used to manage four active collection digitization projects totaling over 6000 archival objects, with each project expected to last between two months and three years depending on the size of the collection. When a new project is started, a set of Smartsheet files is copied from the main template, with all cell links and cross-sheet formulas automatically updated to reference the new project files. After each project concludes, the project folder is archived and we complete a final analysis of project metrics before exporting the project data to .xlsx and .csv formats for preservation.

Limitations and Workarounds

The main difficulties encountered in implementing the Smartsheet system arose from the fact that Smartsheet does not have the searching and indexing power of a database tool or the specific functionality of a DAMS. In other words, it is designed for project metadata, not for project data. However, we needed to record some basic project data—identifiers, image counts, and box locations—in order to allocate resources, predict throughput, and satisfy stakeholders. Three data-management problems and solutions are presented here as examples of the ways in which Smartsheet's flexibility can be used to make up for its domain-specific shortcomings.

Identifiers

Unique identifiers are crucial to accurately tracking hundreds of thousands of objects and images. Each of HILA's archival objects is tracked in ArchivesSpace using a randomly-generated 128-bit identifier known as a refID (for example, "f946bedf973fdad7fc76d82495661e98"). The refID links the digitized object with the archival record and is used in the name of each file ingested into the repository. During the DFI pilot, however, it became clear that refIDs could not be used by imaging staff in naming image files and tracking task completion, because the values were not human-readable; they bore no relation to each other or to the physical objects being imaged. It was too easy to mix up objects and too time-consuming to find a specific object in a Smartsheet task report.

As a solution, we implemented a second identifier, which we call an object mask, and which is made up of the collection ID, box number, folder number, and sleeve number (if one exists). Because each folder or sleeve corresponds to an archival object, there is a one-to-one relationship between object mask and refID. Imaging staff use the object mask to name their files and track their work. Once images are ready to be ingested, a list of object masks and refIDs is exported from Smartsheet and a Python script is run to programmatically rename each image with the correct refID.

Image counts

The second problem we encountered was how to calculate and record the number of images captured over time. The approaches we had used in the past relied on error-prone data entry (for example, each photographer keeps track of their own totals and adds them to a shared spreadsheet at the end of the week) or insufficiently granular third-party tools (for example, post-processing tools such as Goobi track the total number of images prepared for ingest each month).

As with identifiers, the solution we implemented for recording image counts involved a small amount of coding in addition to Smartsheet tools. Each week, a Python script loops through all the object folders in our image QC drive and produces a spreadsheet listing the total number of images in each object folder. A staff member uploads this spreadsheet to Smartsheet to trigger a Data Shuttle workflow that imports these numbers to a standalone sheet with only two columns, Object Mask and Number Of Images. Within each collection's object tracker, the image count for each object mask is pulled in using a cross-sheet formula referencing that standalone sheet. If the image count for an object changes from week to week—for instance, if an object is reimaged because a page was accidentally skipped—the new number overwrites the old one. If someone forgets to upload the spreadsheet in a given week, they can upload it whenever they remember and the image counts in Smartsheet will be updated immediately.

Timesheets

Because our team members work on more than one project at once, a crucial element of calculating throughput is tracking how much time each person spends on each project per week. Smartsheet does not record full timestamps when tasks are completed, only dates, and there is no integrated timesheet app. To collect the data we needed, we built a time-tracker sheet for each project, with an integrated form that is automatically sent to each person to complete

each week. The person records how many days they spent on each stage of a project, and this number is tracked against the number of tasks they completed that week (using a cross-sheet formula to count the number of objects in the object tracker where the relevant task completion date matches the current week). This allows us to monitor each person's throughput week to week and to identify bottlenecks as they occur.

Success of the Solution

Over the course of developing and testing any software solution, one inevitably accumulates a list of its shortcomings. A small sampling: Smartsheet does not support barcode fonts, meaning that in order to print out object identifier sheets for our photographers to use, we had to build an automation to export Smartsheet data to Google Sheets. The pivot app is not as flexible or powerful as pivot tables in Excel, and pivot workflows set up by one person can't be viewed or edited by anyone else. While it's possible to view a log of changes to a sheet, there is no version control, and changes to the workflow template must be painstakingly copied to each active project. Finally, we have not yet stress-tested our solution's scalability. Over the next several years, Smartsheet's limits on sheet size and cell links per sheet may require us to create multiple object tracker sheets for larger collections, or to freeze some of our cross-collection tracking functions for completed collections.

There is also the question of usability. HILA's Smartsheet implementation has no glossy front-end user interface. Most users interact with it by editing values in a spreadsheet. Administrators edit formulas within large spreadsheets, whereas end users (including photographers) check off boxes or choose from dropdowns within smaller filtered reports, but some familiarity and patience with tabular data is required. Even the filtered reports are a double-edged sword; some users have reported that they find their limited view of the Smartsheet solution confusing, because they don't understand what happens to a task before or after they work on it. Further training and socialization of the entire DFI Smartsheet ecosystem is needed so that each team member can better understand (if they wish to) how their interaction with the Smartsheet solution fits in with the whole. Meanwhile, when it comes to reporting out to a general audience, the project dashboards we built for HILA's leadership team leave a great deal to be desired in terms of design and usability (figure 3). We plan to investigate options to export live Smartsheet data to a third-party data visualization tool, such as Tableau or Google Data Studio, although it is also likely that Smartsheet's dashboard interface will improve with time.

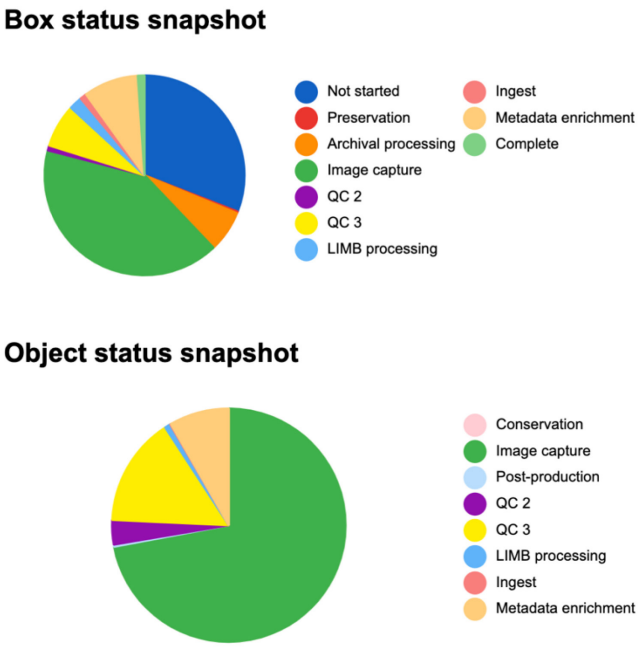


Figure 3. *Example of dashboard charts showing the current status of all boxes and objects in the DFI workflow. The charts are interactive, allowing the user to hover over to view more information or filter categories.*

For HILA's purposes, however, Smartsheet's advantages outweigh its shortcomings. Its flexibility is crucial for our evolving workflow, especially as each collection may have different needs. As long as column names are not changed, the Smartsheet project template can easily be adjusted to fit the needs of a particular project. For example, a preservation review step can be added prior to archival processing, or an additional QC step can be added after imaging so that a language specialist can verify the orientation of a set of Chinese-language pamphlets. We can also easily start tracking new metrics—for example, average hours of work per archival box—without having to hire a consultant or submit a support ticket. While our Smartsheet consultant was essential in setting up some of our cross-sheet formulas and premium app integrations, we have found that the advanced Excel users on our team are able to jump right into writing and editing formulas with the help of the Smartsheet community forum.

The community forum is one aspect of perhaps the most important advantage of Smartsheet over other solutions: its size and wide uptake. There is substantial online documentation, interactive video tutorials, and a robust forum. New features are added frequently, such as the recently launched Data Shuttle (which is integral to the DFI workflow, as discussed above) and Dynamic View. Smartsheet is used by several departments of Stanford University outside the cultural heritage space, meaning that institutional buy-in was relatively easy to obtain, and our contract with a Smartsheet consultant was handled centrally through Stanford. These advantages cannot be taken for granted in the cultural heritage space.

Conclusion

There is a lack of robust, user-friendly management software tools designed for the cultural heritage sector. Rather than build our own tools, or wrestle with the shortcomings of the tools that are readily available, HILA experimented with a commercial tool with a wide user base. Customizing the solution to our needs took time, and we have not finished evaluating the success of our experiment. Nevertheless, the results seem promising. By combining Smartsheet with task-specific tools such as Capture One and ArchivesSpace, we have been able to break out of cultural heritage digitization's underserved product landscape and engage with a larger community of practice, without compromising on the specific task management and reporting requirements of a mass digitization workflow.

References

[1] https://archivesspace.org/, retrieved 15 May 2022.

[2] https://www.captureone.com/, retrieved 15 May 2022.

[3] https://www.limbsuite.com/limb-processing, retrieved 15 May 2022.

[4] http://www.digitalpreservationsoftware.com/digital-preservation-solutions/libsafe-digital-preservation-software/, retrieved 15 May 2022.

[5] Society of American Archivists, Describing Archives: A Content Standard, version 2019.0.3, Society of American Archivists, 2020.

[6] International Council on Archives, ISAD(G): General International Standard Archival Description, Second Edition, 2000.

[7] https://www.intranda.com/en/digiverso/goobi/goobi-overview/, retrieved 15 May 2022.

[8] https://www.smartsheet.com/, retrieved 15 May 2022.

[9] https://trello.com/, retrieved 15 May 2022.

[10] https://www.atlassian.com/software/jira, retrieved 15 May 2022.

Author Biography

Emma Stanford is the Digital Services Manager at the Hoover Institution Library & Archives. Prior to her role at Hoover, she was Digital Curator at the Bodleian Libraries in Oxford. She holds a B.A. from Middlebury College (2012) and an MSc in Library Science from City, University of London (2017).