https://doi.org/10.2352/issn.2168-3204.2023.20.1.12


# How it all came together: Building a local app to keep track of the digitization workflow

***Hernández Mejías, Natalia; San Juan, Puerto Rico - Ayala González, Hilda Teresa; San Juan, Puerto Rico - Ramírez, Víctor; San Juan, Puerto Rico***

## Abstract

*Thanks to an Andrew W. Mellon Foundation grant, the General Archives of Puerto Rico started a mass digitization project in 2020. The goal was to establish a digitization center and implement FADGI guidelines. As the project developed and the volume of work grew, a fast and simple way to track the items through their different stages was needed. Although several software options were available, they required more resources than we had on hand at the time. Understanding our needs and goals, our team's IT technician built an app tailored to the project's requirements. In the past year, we have not only successfully kept track of the objects through the digitization workflow, but the app also proved effective for maintaining team communication, collecting technical metadata, and recording relationships between objects and their collections.*

## Introduction

Thanks to an Andrew W. Mellon Foundation grant, the General Archives of Puerto Rico started a mass digitization project in 2020. The goal was to establish a digitization center and implement the standards specified in the Federal Agencies Digital Guidelines Initiative (FADGI). As the project developed and the volume of work grew, a fast and simple way to track the items through their different stages was needed. Although several software options were available, they required more resources than we had on hand at the time. Understanding our needs and goals, our team's IT technician built an app tailored to the project's requirements. In the past year, we have not only successfully kept track of the objects through the digitization workflow, but the app also proved effective for maintaining team communication, collecting technical metadata, and recording relationships between objects and their collections. Through this paper, we will share our experience using this locally developed tracking app.

The General Archives of Puerto Rico (GAPR) was established by Law No. 5 of December 8th, 1955, known as the Law for Records Administration in Puerto Rico (amended). It was placed under the purview of the Instituto de Cultura Puertorriqueña (ICP), the state arts agency of Puerto Rico. GAPR's mission is to receive, safeguard, preserve, and disseminate the cultural heritage documents of Puerto Rico. It is the largest repository on the Island, preserving over 90,000 cubic feet of files from government agencies, public corporations, municipalities, and private collections dating from 1730 until the present. The types of materials held by the institution include text, newsprint, maps, photographs, books, audio, and video collections.

In order to reach a wider group of patrons, and to provide access to the GAPR collections, various attempts to establish a digitization center had been made. It was not until 2020 when it all took shape. The institution embarked on a three year initiative supported with a grant from the Mellon Foundation. The project was pursuing to establish a digitization program. In the proposal we planned to achieve the following: create a state-of-the-art lab to facilitate digitization in mass, upgrade the conservation laboratory with new tools, and implement a communication system with storage solutions (servers). Other important components of the plan were: hire and train a new workgroup, design and implement a digitization workflow following FADGI recommendations, create an online repository, implement digital preservation strategies, and make 500,000 images accessible to the public.

The proposal for the digitization program was fixed in the use of FADGI guidelines. This element was very important because we wanted to implement a project capable of producing high quality images of the documents of the GAPR. The goal was to generate preservation copies to avoid manipulation of very fragile documents, prevent repetition of the digitization job in the future, and provide access copies to be shared through the online repository. We applied the Digital Imaging Conformance Evaluation program (DICE) with image targets (device and object) and the GoldenThread/NXT analysis software, all developed by Image Science Associates.

In terms of the collections, we have been able to work with different types of documents that hold a very significant historical and cultural importance for the archipelago. We are working with the historical Newspapers Collection, the Puerto Rico Police's incidents logbook, the foundational records of the city of San Juan known as "Minutes of the Council of San Juan" (these are bound volumes dated from 1722 to the 1900's), sheet music, a part of the Public Works Collection related to the Puerto Rico railroad, newspaper clippings and the first female mayor of San Juan's photographic collection.

Although we have been very lucky be able to work with valuable collections as we detail previously, it is important to recognize that in the process of developing our digitization program we have encountered multiple challenges. One of the main obstacles we faced was the lack of proper organization and description of collections. There was not a standardized way for naming objects or creating archival relationships. Also, the institution did not have a collection management system that could help us design the digitization workflows, keep track of the items that went into production or obtain the needed metadata for processing, preservation, and access.

The tracking app brings the project together and helps us overcome challenges regarding: collection relationships, digitization workflow, and collecting of technical metadata. Also, it helps us keep track of the progress of the initiative. It is important to mention that with Covid 19 it was key for us to have the information in a centralized system so everybody on the team could access the data needed to continue the project in the case of someone's absence. Fortunately, the app was implemented successfully, and we continue to improve it, adding new features to meet the project's needs.

## Background

At the start of our project, we evaluated the system Goobi. According to the webpage, "Goobi is an open-source software suite for controlling and presenting digitization projects. It consists of two core components, Goobi workflow and Goobi viewer, as well as a broad ecosystem of complementary tools and plug-ins." However, despite the benefits that this platform has brought to other digitization projects, we found that in order to successfully implement and adapt this application to our necessities, we would need more time than we had available. We also lacked a proper server, which was required if we wanted to reap the benefits it provides for things like image ingest. Another situation was that Goobi's default metadata schema was METS/MODS and we could not apply it to the complex archival collections we were working with. It is known that each institution has its own idiosyncrasy, and to develop a mass digitization project from zero, occasionally it is necessary to take alternative paths with the hope of meeting best practices and using time effectively.

After evaluating and discarding working with Goobi, and while trying to identify other software, our workflow information was being organized across several spreadsheets. One was used for tracking objects, another to enter capture metadata, another for technical metadata collected by the Quality Control (QC) team, and an additional one containing unique identifiers and collection hierarchies. This worked for the first few months. However, as the volume of objects to be digitized increased, it became clear that we needed a better system to keep track of the progress of our work and gather the documentation in a centralized location.

Although the need came from a project management view, we wanted the app to support the following activities: (1) maintaining objects' relationships with their respective archival hierarchies, (2) providing precise information regarding location of objects through the entire workflow, (3) generating consistent and uniform technical metadata for digitized objects, (4) giving cohesion to the digitization workflow by connecting the different stages: preservation (rehousing and cleaning), conservation, digitization, QC, and metadata creation, and (5) obtaining statistical data for digitized objects to evaluate project progress.

But for the app to work effectively, we first needed to work with a few in-house idiosyncrasies. The most challenging aspect of keeping track of the objects was that we decided for this first phase to digitize collections or sections of collections as a whole (e.g. series or sub-series), and they lacked finding aids. In most cases, only an inventory of the objects was available. Therefore, we were missing essential contextual information about the content and relationships of collections and their items. We also wanted to maintain a record of the entire archival hierarchies, from the collection to the item level, and in some occasions, a collection could have more than one name. Standardizing descriptive content was essential for the success of the project itself, but more importantly, we had to be able to display and relate items in the digital repository where collections and items would later be accessed.

As a first task to mitigate this situation, the metadata team and the project manager conducted an evaluation of the available inventories and compiled all the data they contained related to collection hierarchies. We then proceeded to edit the hierarchies to make them uniform. Around 300 inventories were examined.

This initial work was also crucial for the creation of the tracking application, since the app is based on the deeply nested collection hierarchies. As shown in Figure 1, each collection level is related to its parents.


***Figure 1**. Hierarchical relationships of collections*

The next step was to create a unique code for each level of the collection's hierarchies (Collection – Sub collection – Series – Subseries). This would help provide a short and consistent name, identify quickly where an item belongs, and create the file names of the new digital objects. The code is a simple autogenerated five-digit number that is assigned based on the order the collections are entered into the spreadsheet.

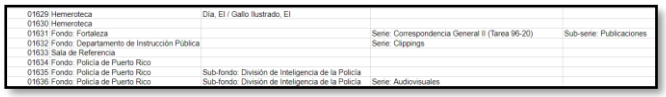


***Figure 2.** Codes for each level of the Collection's hierarchy*

Once the unique codes are assigned, the project manager designs the file naming convention for the digital items, which include: Code_Box#_Folder#. With this information, items are ready to enter the digitization workflow and start their journey through the tracking application.

## Workflow

To begin the digitization workflow, three steps must be completed by the project manager. The first step is to either look up or create the item's hierarchy. With hierarchies clearly established, we are able to provide a unique identifying code that functions as the backbone to maintain items related and connected through the whole process. Next, to ensure relationships are established between the items that belong together, we create a container which is equal to the physical box. The containers are then described in terms of their original physical location and the process they will be part of. Lastly, objects are created inside each container. Depending on the collection, objects can be a folder, an envelope, a publication or an image. For each object, data is provided about its physical dimensions, title, permanent location, location, workflow stage, and the person responsible for the object. Figure 3 presents the metadata compiled at this level.


***Figure 3**. Object administrative and basic descriptive metadata*

Once this data is provided, the object is ready to move through the workflow. Accordingly, the project manager is the person who assigns the container with its objects to the appropriate stage and the person who will be responsible for it. The steps include: pending processing, preservation evaluation/cleaning or rehousing, waiting for conservation treatment, object in treatment, treatment completed, revision/foliation, ready for digitization, in digitization, ready for

QC, in QC, QC completed, pending technical metadata, technical metadata completed, pending descriptive metadata, descriptive metadata completed, final revision, upload to CollectiveAccess, and process completed.

As team members carry out each task (preservation, conservation, digitization, QC, and metadata) they can add more information and update the status. For example, the digitization and quality control team members can record technical details of the digitized object. Data input includes the number of images, object types, software and equipment information, date of capture, color space, resolution, format, and megabyte extension. Figure 4 presents the additional technical and administrative metadata recorded.


***Figure 4**. Objects' additional administrative and technical metadata*

The application provides a filter to easily identify which objects are assigned to the different team members. It also maintains a log of the workflow steps with dates that reflects how the objects move through it. This is a great feature that can help estimate the time it takes to complete each step and plan for other projects.

Because the GAPR does not have a content (or digital assets) management system and has not generated Encoded Archival Descriptions (EAD), all the technical metadata for the digitized items is entered into the application and then exported and saved with the objects as part of the documentation process. Part of this technical metadata is later entered into the spreadsheet where the descriptive metadata is maintained for each item.

## Software Development, Difficulties and Achievements.

The tracking application was developed in-house using Linux and the following elements:

- A MySQL database (Language: SQL)
- Java Server REST back-end using the Spring Boot library (Language: Java)
- Next JS + React JS Web front-end graphic interface (Language: Javascript)

The app was created in three months with the programmer working part-time. In mid-December, the beta version was ready to be launched, and it was put into production in mid-January 2022. We must emphasize that feedback from the workgroup was key for further development of the application. Several meetings were held to gather thoughts and recommendations. We also created a Microsoft Teams channel exclusively for providing a space where everyone had the opportunity to report problems and request changes. Our IT technician constantly monitored this channel and implemented the requested changes. By following this strategy, our team was able to polish the app's performance in just a few weeks. Figure 5 presents a request that was resolved.


***Figure 5**. Capture of the Microsoft Teams channel*

One of the challenges reflected on the Teams channel was the time needed to complete the technical metadata on the app. The digitization and quality control team members shared that it takes around five minutes to fill out all the elements. To mitigate this situation, the programmer added an option to copy the record of the previous item's capture. With this in place, the digitization technician saves time when entering the data of the imaged items. The quality control team has a similar situation because they have to add information about the derivative. In the near future, the programmer will develop the same solution for the quality control team.

Another important element of the application is the report tool. The project manager uses this feature to follow the progress of assigned tasks. Two of the most important reports are: total images captured, and objects assigned to the different team members. The data detailing images captured can be exported as a spreadsheet. This allows us to evaluate the number of digitized items by category (bound, photographs, loose documents, and oversized). It also shows statistics by collection, capture dates, digitization technician, and other useful data for project management. Figure 6 presents the images captured statistics.


***Figure 6**. One of the app's reporting outputs showing captured images*

The report detailing object assignments to each team member helps us visualize what workflow tasks have been completed and which are pending. This is very useful, as it allows us to quickly adjust and redirect efforts when certain team members are in need of support. Figure 7 shows a report of objects assigned to each team member.


***Figure 7**. Another of the app's reporting outputs showing team tasks*

## Conclusions

Through a Mellon Foundation grant, the General Archives of Puerto Rico was able to establish a digitization center and implement FADGI guidelines to create its first digital collection. As part of the project management process, it was quickly learned that a tool was needed in order to keep track of the objects throughout the digitization workflow. The first tool evaluated was Goobi. Although it was very useful for generating metadata records, many of the materials we were working with

lacked finding aids or had flawed inventories. Due to these lingering issues, Goobi was not the best option for us.

The project IT tech suggested that we develop an in-house tool tailored to our needs. To make this a reality, in terms of project management, the action plan was: (1) standardizing the archival hierarchies, (2) revising collection inventories in order to correct the information contained in them, (3) correcting and standardizing the information provided for objects to include names, brief descriptions, storage information, and permanent location. By doing this, we were able to incorporate the objects into a tracking system in an agile way, without interrupting the digitization process. At the same time, we securely maintained objects' relationships within their collections. At the moment, we have entered 9,779 objects in the tracking tool, and soon we will be entering data for close to 20,000 negatives and 5,000 photographs.

This tracking application is a work in progress, and one of the aspects we are looking to improve is the visualization of the captures report within the app. For now, the best way to filter the data and obtain statistics is downloading the report spreadsheet. Another improvement we hope to implement is developing options to facilitate mass migration of data. Currently, the IT team is in charge of performing data migrations. However, we will develop strategies so that the metadata team can eventually take over this task.

We believe the development of the app has been truly successful and has helped us achieve our goals. Although the app was specifically created to address the needs of our institution, we will publish a version with generic elements for institutions like us who need a basic and simple solution for the management of a digitization project in an archive. In Latin America specifically, we maintain a strong archival tradition of respecting records relationships, one of the main aspects facilitated by the app. Therefore, we believe the app will be well-received by this population. We also hope that the tool will be useful to Spanish-speaking institutions, as it was developed completely in this language. The source code of our local app will be freely available to the general public, and only basic IT skills will be needed for its installation.

## References


[1] Steffen Hankiewicz & Jan Vonde, "End-to-End Digitization Workflow: Goobi to go for Newbies", short course, annual IS&T Archiving Conference, University of Lisbon (2019).

[2] intranda GmbH, goobi-workflow, https://github.com/intranda/goobi-workflow, accessed: March 2023.

[3] Emily Symonds & Cinda May "Documenting Local Procedures: The Development of Standard Digitization Processes Through the Dear Comrade Project", Journal of Library Metadata, 9:3-4, pp. 305-323 (2009), DOI: 10.1080/19386380903405207

[4] Larisa K. Miller "All Text Considered: A Perspective on Mass Digitizing and Archival Processing." The American Archivist, 76:2, pp. 521–41 (2013), http://www.jstor.org/stable/43490366. Accessed March 2023.


## Author Biography

*Natalia Hernandez Mejías holds two Master's degree in History and in Library and Information Sciences from the University of Puerto Rico, Río Piedras Campus. She has worked in a wide variety of librarianship roles in academic, public, and special libraries, from reference services, to workshop instructor, to project and grant management. Since 2020, she has worked as the project manager for the Mellon digitization initiative at the General Archives of Puerto Rico.*

*Hilda Teresa Ayala González has a Master's in Archival Studies from the University of British Columbia in Vancouver, Canada, and a master's in Library Science from the University of Puerto Rico. She has worked in various professional ettings as a librarian, archivist, consultant, and professor. In 2020 she was designated the General Archivist of Puerto Rico and the director of the National Library of Puerto Rico at the Institute of Puerto Rican Culture.*

*Víctor Ramírez has a Bachelor's in Computer and Information Science and Engineering from the University of Florida and also studied Urban Planning at the University of Puerto Rico. He has been designing and developing government information systems for 20 years and has undertaken civic projects such as Tren Urbano App, an application for tracking public transport. He is currently in charge of IT and software development for the General Archive of Puerto Rico's digitization project.*