Lessons Learned

Takeaways from a Seven-Year Digitization Project at the National Institute of Standards and Technology

by Katelynd Bucher
Metadata Librarian
Information Services Office, National Institute of Standards and Technology

Overview

With the completion of a large-scale project comes the inevitable question: What have we learned, and how can we apply it in the future? The National Institute of Standards and Technology (NIST) Information Services Office (ISO) looks back on lessons learned during the completion of its project to digitize and preserve its collection of grey literature published by NIST and its predecessor, the National Bureau of Standards (NBS). NIST's legacy digitization project supported ISO's mission to create, maintain, and disseminate a knowledge base that supports NIST's scientific, engineering, and technical research. This paper examines that project over its seven-year duration, its goals and lessons learned, and the path it has laid for digitization and preservation of NIST publications in the future.

Background

The National Institute of Standards and Technology (NIST) is a non-regulatory federal agency with about 3000 science and technology researchers. NIST promotes U.S. innovation and industrial competitiveness by advancing measurement science, standards, and technology. The agency was established by the U.S. Congress in 1901 as the National Bureau of Standards (NBS) in response to a national need for standardized metrology. It is the nation's oldest physical science laboratory, with a long history of science and engineering output. NBS changed its name to the National Institute of Standards and Technology in 1989. The Information Services Office (ISO) supports and enhances NIST's research activities through a comprehensive program of knowledge management.

About the Project

In 2011, ISO began a large-scale project to digitize its legacy grey literature publications, mainly composed of technical reports. This body of grey literature is collectively known as the NBS/NIST Technical Series. It comprises 92 series titles, with publications from 1902 to the present. Because of its concurrent roles as NIST's research library, the custodian of NIST's historical archive, and the publisher of the *Journal of Research of NIST* and the NBS/NIST Technical Series publications, ISO is uniquely positioned to collect and preserve copies of NIST's legacy and current grey literature publications.

The legacy digitization project was initially conceived as a pilot project to digitize the *Journal of Research of the National Institute of Standards and Technology,* NIST's peer-reviewed journal publication—colloquially known as "JRes"—as well as its predecessors: the Bulletin of the National Bureau of

Standards; the Technologic Papers of the National Bureau of Standards; the Scientific Papers of the National Bureau of Standards; and the Journal of Research of the National Bureau of Standards. Upon the successful completion of the pilot, the project was expanded and became the NIST Technical Series legacy digitization project.

The scope of the NIST Technical Series legacy digitization project included the full collection of the NIST Technical Series publications. The NIST Research Library maintains a collection of approximately 37,000 NIST Technical Series publications. Approximately 24,000 publications were earmarked to be digitized, while the remaining publications were not included in the project planning, as they had been produced for or with funding from other federal agencies or contained proprietary information.

The project workflow[1]—which involved the batch processing of publications—included the following steps for each publication:

- metadata was created or updated;
- the publication was sent offsite to the Internet Archive[2] scanning facility for digitization. The digitized publications were made available in perpetuity on archive.org by Internet Archive.[3]
- ISO downloaded the publication from archive.org, assigned it a Digital Object Identifier (DOI) from Crossref.org, and added the metadata to the ISO library catalog;
- the publication and its accompanying metadata were deposited into the U.S. Government Publishing Office (GPO)'s preservation repository, govinfo.[4]

The initial project workflow was simple and involved metadata creation; digitization of the publications by Internet Archive; downloading of the digitized publications by ISO; and making the digitized publications available in the public domain on ISO's servers. The workflow became more complex as the project unfolded. As new tasks—such as assigning DOIs—were added to the workflow, team members strove to find ways to make more complex tasks more efficient. Most changes were added via trial and error and were intended to eliminate duplicate work or to automate time-consuming work done by hand. For example, team members created an XSLT transformation to convert Marc metadata to the Crossref XML schema, rather than completing each batch manually. Team members used tools such as Xcopy and Robocopy to transform, move, and manage large batches of files; and utilized MarcEdit's Marc Editor functions and regular expressions to make global changes to the metadata in those files. The workflow for downloading the large batches of digitized publications from Internet Archive also changed over time; team members used first wget and later Python to complete the task.

ISO sent the last batches of publications to Internet Archive at the end of 2017, and they were processed throughout 2018. By the end of 2018, the legacy digitization project had completed the digitization of 24,611 publications.

[1] For more details about ISO's legacy digitization project workflow, see (Bucher, 2016).

[2] The identification of any commercial product or trade name does not imply endorsement or recommendation by the National Institute of Standards and Technology.

[3] The NBS/NIST Technical Series publications are available through Internet Archive at https://archive.org/details/NISTresearchlibrary.

[4] The NBS/NIST Technical Series publications are available in the GPO govinfo repository at https://www.govinfo.gov/collection/nist

Project Goals Set and Surpassed

ISO set a number of initial goals for the legacy digitization project during the planning stage and added additional targets as the project progressed. Five goals were established:

(1) **Digitize the entirety of the NIST Technical Series publications, make the digitized publications publicly available, and deposit them in a secure preservation repository.** ISO met and, in fact, exceeded this goal, as all of the series that had been part of the project's initial plan were successfully digitized, as well as several other publication series that were digitized at staff and customer request.

(2) **Fulfill customer requests for digitized materials.** In the 2000s, as the initial digitization pilot was implemented, digital journal subscriptions and e-books piqued customers' interest in the availability of digital publications from the NIST Research Library. The current Technical Series publications also transitioned to born-digital during that period and ceased print production.

While the majority of the NIST Technical Series publications were available in both the NIST Research Library's circulating and reference collections prior to the beginning of the legacy digitization project, many print copies were not in circulation due to their age and fragile condition. Additionally, some of these smaller series—particularly those published in the early twentieth century—were unique to the NIST Research Library's collection. The digitization of the Technical Series publications made them widely available to the public.

(3) **Assign DOIs to all Technical Series publications.** The goal of assigning permanent identifiers to the digitized Technical Series publications was incorporated into the project workflow in 2014. Giving each publication a DOI was an additional method of ensuring its availability and increased its discoverability through CrossRef metadata and linked content. During this process, DOIs were also added to the publishing workflow for new Technical Series publications, and thus the full collection of NIST's grey literature has been assigned DOIs.

(4) **Metadata normalization.** This objective arose during the scanning of the first major technical report series, the NBS/NIST Technical Notes. Initially, the metadata created for the scanned publications was based on the print records, which were inconsistent, having been created by numerous catalogers over many decades. Because the digitized publications' metadata would be sent to different places and in different formats, ISO recognized the importance of uniform metadata for those publications.

The solution to this challenge was to create a new, normalized metadata record for each publication. ISO then successfully extended the metadata template used by the digitized publications beyond the legacy digitization project to include newly published Technical Series publications.

(5) **Streamline the project's workflow.** Numerous processes, such as the creation of batches of publications to be sent in a specific format to GPO, involved many steps that required completion by hand. ISO created several scripts and regular expression tools to automate those steps, which expedited the process.

Lessons Learned

There were several notable lessons learned during the completion of ISO's legacy digitization project. The most significant of those pertained to project expectations versus realities. In any project as large as this one, the application of Murphy's Law should always be expected. Surprises, however, can also be opportunities for growth and improvement, and despite unexpected developments, or perhaps because of them, the legacy digitization project settled into an effective workflow that facilitated the successful completion of its goals. Some of the lessons learned during the project's completion are as follows:

Project Duration

The most significant lesson learned was that ISO needed to be flexible with the project's time frame, which ended up stretching well beyond the initial plan. The project was intended to be completed in three years, but that target increased to five years, and then finally to seven.

The main factor that contributed to the project's extended timeframe was that the amount of time required for the completion of the scanning process was always in flux. ISO sent the Technical Series publications one series at a time to be digitized at the scanning center. The publications were sent, series by series, in batches of varying numbers of boxes of approximately one cubic foot in size. Sometimes those batches were comprised of large volumes containing numerous publications, or one large publication. Other times, they comprised hundreds of small, individual volumes. The size of the batch and the length of the volumes within that batch were the factors that most affected the duration of the project.

The scanning center's workflow (including the scanning process), the quality control inspection of the digital files, the placement of digital files on archive.org, and the return of physical volumes to ISO all took varying amounts of time. This was especially true when the scanning center processed a larger batch. As the legacy digitization project progressed, ISO eventually began to send only larger batches, as the time spent at the scanning center allowed ISO staff to complete the remainder of the project workflow on those materials that had been previously returned.

Team Members

Another lesson learned was a better understanding of how project staffing needed to respond to changing project conditions in order to complete the project in the desired time frame. Numerous ISO staff members were involved in the various processes during the project and included those with expertise in digitization; digital preservation; cataloging and metadata; and data curation. The number of team members fluctuated over time. Initially, one librarian was assigned to begin the project as it transitioned from a pilot to a project of much larger scale. As the scope of the project expanded, an estimate of the number of publications to be digitized was determined, and a realistic time frame for digitizing that number of publications was set, the project's team grew to involve a second librarian, with extra staff assisting as needed.

During the initial three years of the project, the speed of digitization varied depending on the speed of the scanning center and of the ISO staff members in processing the publications before and after scanning. Time was the project's greatest foe. When ISO had sufficient staff dedicated to the project, more publications could be sent for digitization at a time, but sending more publications meant that it took more time for those publications to be scanned. This left project members at loose ends as they waited for a previous batch to return before sending a new one, meaning that at times there was a month or more of no activity on the project. Alternately, when there was not sufficient staff dedicated to the project—a single project member, for example—smaller batches of publications were sent for scanning, and those batches were digitized and returned much more quickly. This frequently led to the single staff member becoming overwhelmed by the volume of processing required within the shorter timeframe.

Through trial and error, ISO determined that three full-time team members—one metadata librarian and two librarian contractors—enabled the greatest amount of productivity on the project, with staff neither overwhelmed nor left at loose ends, and the project progressed more quickly and evenly.

Retrospective Time Commitments

As stated previously, one of ISO's goals was to streamline the project's workflow. Team members made frequent adjustments and improvements to their processes, both to increase their efficiency and to maximize the quality of the digitized publications and their accompanying metadata. As these adjustments were made—particularly those that involved enhancing metadata content or reorganizing the digital materials for more efficient processing—they also required retroactive changes to materials that had already completed the project workflow.

The decisions to include a local tag in all metadata records for easier identification, to increase the granularity of subject heading fields for improved keyword searching, and to change parts of the naming schema for the digital files, all became tasks that required more time and more steps than were originally anticipated. Writing a new script to automate part of the process, which ultimately saved ISO a great amount of time, caused delays as it required staff to learn new skills to create the script and integrate it into the workflow.

Another complication that caused the project's extended duration was the thorough quality control check at the end of the project workflow, rather than at the end of each step in the digitization process. Though higher-level quality checks were performed as publications moved through the different stages of the project, making sure all items were present in the larger batches and discovering immediate errors, a more granular quality check of each publication and its metadata was not performed until near the end of the project workflow. This choice was made in order to reduce the time required for each item to move through the digitization and preservation workflow, and to eliminate duplicate work. This unfortunately resulted in the discovery of minor errors in the publications' metadata at the end of the project workflow that then required corrections to multiple files in different locations and formats. Had these errors been caught earlier in the workflow, the volume of edits required would have been much less extensive.

The lesson learned through these experiences was that in such a large-scale project, what can go wrong will always go wrong, and that edits and changes to the workflow will take longer than expected, particularly if they must be applied *retroactively*. For large-scale digitization projects, it is best to plan for things to take longer than anticipated, and it is always best to take the time to do the task thoroughly the first time around. It may not seem as though it will save time, but it will eliminate the need for many retrospective edits.

Project Outcomes

There have been numerous significant outcomes following the completion of the legacy digitization project. The most obvious of these is the increased availability and discoverability of the Technical Series publications. After the publications were digitized and made available in the public domain, their metadata was disseminated to numerous access points, including the NIST Research Library catalog, WorldCat, CrossRef, govinfo, the Federal Depository Library Program (FDLP), and more. The Technical Series publications are no longer sitting static on the library shelves, only accessible to the NIST Research Library customers as a physical resource but are now widely available via multiple channels.

Another outcome has been the increased customer engagement with legacy Technical Series publications. Customer response to the legacy digitization project was positive from the outset. Requests from customers for digital copies of individual Technical Series publications and for digital access to full series were the initial drivers for the legacy digitization project. As the digitization of each series was completed and the publications were made available online and deposited into govinfo, NIST's marketing of the publications' availability increased customer awareness of both the Technical Series as digital publications and of the digitization project in general. ISO began to receive requests for the digitization of specific volumes and series as the project began. Customers also reached out to ISO as they discovered older publications in their offices, often as they cleaned or prepared for retirement, and wondered if those publications had already been digitized. In many cases, these customer donations filled gaps in the Technical Series publications collection, and ISO was able to include them in the digitization workflow.

The legacy digitization project has also resulted in the implementation of an enriched, normalized metadata template for the Technical Series publications. As discussed previously, this has enabled the metadata to be more easily disseminated to various platforms, thereby increasing discoverability. This has applied both to the digitized legacy publications and to the new publications currently being published by ISO. The scripts and tools that ISO utilized during the legacy digitization project have also been implemented throughout ISO's metadata creation and management workflows. They have increased efficiency and normalization across the board and in numerous processes. The scripting and skills ISO staff acquired and enhanced during the legacy digitization project have also had applications in other workflows in the NIST Research Library. Staff have modified the scripts and transformations utilized in the legacy digitization project to automate several cataloging and publishing processes, for example, eliminating manual steps and duplicated work. Looking forward, enhancing these skills amongst ISO staff would allow for more flexibility between staff on other projects as well, encouraging a shared body of institutional knowledge and enabling staff to step in and out of projects, lending assistance and expertise as needed.

The creation of an enriched, normalized metadata template for the NIST Technical Series publications has also catalyzed ISO staff interest in cataloging and metadata creation workflows and how they are applied throughout the Library. Team members hope to build on this increased awareness and to facilitate a deeper understanding of how digitization and metadata processes work amongst staff who do not normally work with metadata or digital materials. This will lead to a better understanding of what metadata is, how it is created, and how it is used, with no magic wands involved.

Another successful outcome has been the increased findability of content within the digital publications themselves. The full texts of the digitized publications are searchable. This has enabled ISO staff and customers to find content within the digitized publications more easily and effectively than they could before the volumes were made available in their digital formats. The full text searchability within the digitized files has been very valuable and has enabled ISO to better serve its customers. ISO staff have been able to provide more information to customers on specific people and events in NIST's history by searching for mentions of them within the full text of a publication, rather than relying solely on searching its metadata. For example, this feature has been especially useful to historians seeking information about a person who worked in a particular research area at NIST but who may not have been listed as author in a specific publication. The individual might only have been mentioned in the publication's text. The ability to search that publication's full text via keyword for one such individual's name has been invaluable on numerous occasions.

The digitized publications have also provided an effective way to make connections with other NIST materials, including manuscripts and photographs from the NIST Historical Archives and artifacts in the NIST Museum. The DOIs assigned to the publications enable ISO to link publications to the archival collections and museum artifacts. The DOIs can be included in the museum and archival metadata in perpetuity to illustrate relationships, creating connections and enriching NIST's historical materials.

Gotta Catch 'Em All

Due to the expected amount of time it would take to digitize a large number of Technical Series publications, ISO decided during the planning stages that the project's implementation would focus solely on scanning the volumes already present in the NIST Research Library's collection. While those volumes were moving through the project's workflow, ISO would not concurrently seek out volumes missing from the library's collection.

Through the years, ISO endeavored to retain a copy of every Technical Series report published by NIST, but its collection was not complete. As stated previously, ISO's customers have begun to reach out to the library in increasing numbers as the legacy digitization project progressed to offer Technical Series publications to the library for digitization. In response to this, and with the digitization of the library's collection of Technical Series publications now complete, ISO has moved forward into an active role of reaching out to customers and to peer libraries to try to fill the remaining holes in the library's collection. The majority of the missing publications come from the NBS Reports collection, colloquially known as the greybacks because of their grey report covers. The library's collection included approximately 1900 greybacks, all digitized during the legacy digitization project. However, according to their publication numbers, ISO estimates that there are closer to 10,000 publications in the NBS Reports series and is actively seeking copies of the missing publications.

The Road Goes Ever On

ISO's legacy digitization project, begun in 2011 and completed in 2018, met its goals and its target numbers, and is now complete. However, ISO's efforts to seek out the Technical Series publications missing from its collection will continue, albeit on a smaller scale than that of the legacy digitization project. ISO intends to find new ways to connect to the digitized Technical Series content, and to use the experience gained and lessons learned during the legacy digitization project in future projects and endeavors. The knowledge that ISO has gained over the last seven years will have many applications, including finding new ways to help customers discover ISO's grey literature content, new methods of utilizing and enriching its metadata, and making NIST content accessible via new avenues. ISO has reaped enormous benefits from its digitization endeavor and hopes to continue doing so into the future.

References

Bucher, K. (2016). A Legacy of Publications. *Journal of Digital Media Management*, 378-385.