A Legacy of Publications

A Case Study of the NIST Technical Series Publications Digitization Project

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

Overview

The National Institute of Standards and Technology (NIST) Information Services Office (ISO) is currently conducting a multi-year project to digitize 24 000 legacy agency publications. These include a large number of scientific technical reports. This paper examines the evolution of the digitization and preservation process utilized by ISO, from its inception to its current workflow and future goals.

Background

The National Bureau of Standards (NBS) was established by the U.S. Congress in 1901 in response to a national need for standardized metrology. In 2016, it is the nation's oldest physical science laboratory, with a long history of science and engineering output. NBS became the National Institute of Standards and Technology (NIST) in 1989. NIST is a non-regulatory federal agency made up of about 3000 science and technology researchers. NIST promotes U.S. innovation and industrial competitiveness by advancing measurement science, standards, and technology. The Information Services Office (ISO) supports and enhances the research activities of the NIST scientific and programmatic community through a comprehensive program of knowledge management.

NIST Technical Series

NIST began producing agency publications, or gray literature, in 1902 with the first volume of the Circular of Information of the National Bureau of Standards. Since then, NIST has published numerous reports in 92 technical report series.

The 92 technical report series encompass a wide range of formats and subjects. Some reports were intended to be shorter in length, such as NIST Internal and Interagency Reports, while some were much longer, such as NBS Monographs and NIST Special Publications. Examples of

subjects include detailed information about materials used in low-cost housing construction during the New Deal era (the Technical Information on Building Materials for Use in the Design of Low-Cost Housing series) and investigations into the building failures that occurred following the attack on the World Trade Center on September 11, 2001 (National Construction Safety Team Act Reports series). Collectively, these technical reports are all known as the NBS/NIST Technical Series Publications.

Today, ISO occupies the unique position of serving NIST both as its research library and historical archive, as well as the publisher of the NIST publications. This role has enabled ISO to collect and preserve copies of all available NIST publications, many having been preserved in ISO's collection since the agency's earliest days.

ISO has approximately 37 000 NBS/NIST Technical Series publications in its collection, 24 000 of which will be digitized by the end of the project. The remaining 13 000 publications were not included in the planned digitization project because they were produced for other federal agencies, or they contained proprietary information. ISO chose to consider the digitization of these publications at a later date.

Digitization Team

Numerous members of the Information Services Office provided valuable work to the legacy digitization project. The project required expertise in digitization, digital preservation, metadata creation and transformation, cataloging, and data manipulation. The digitization team was made up of different members at different points during the project's inception and implementation. Among these were ISO's metadata librarians, digital services librarian, serials librarian, the integrated library system administrator, and several librarian contractors. Because the project had so many different aspects, one team member serves as the project coordinator.

Drivers

The Information Services Office began to consider digitizing NIST's legacy agency publications in the 2000s, as digitization projects became a prominent trend in libraries. The prospective endeavor had a number of relevant drivers:

1) Both digitization techniques and the quality of digitized materials greatly increased over time, and digitization hardware and software became more affordable. Libraries began to see a trend towards large-scale digitization projects. A digitization project would fulfill

ISO's strategic objective of increasing the visibility of and long term access to NIST scientific and technical research results.

2) ISO began receiving increased customer requests for digitized publications. Digital journal subscriptions and e-books also facilitated customers' interest. The NIST Technical Series publications also became born-digital, and were no longer produced in print. While NIST researchers continued to utilize ISO's physical collections, the digital collections became increasingly popular.
3) Many of the earliest NBS/NIST Technical Series publications were unavailable to customers because they were old, fragile volumes not in circulation. Digital copies of these publications would enable customers to use them, and would make more customers aware of their existence. Digitization would also create additional copies of these deteriorating volumes, ensuring their long-term preservation.
4) The Office of Science and Technology Policy's (OSTP) February 22, 2013 memo[1] on increasing access to results of federally funded scientific research provided additional incentive for undertaking the project. While ISO had begun to digitize the legacy Technical Series publications prior to the OSTP memo and the creation of NIST's public access plan, the policy's emphasis on providing increased access to agency publications enabled ISO to receive funding from NIST management to complete the digitization in a shorter timeframe than originally planned.

Initial Challenges

ISO faced several initial challenges when it began the digitization project.

1) Comprehensive documentation of the publication series.

When ISO began planning to digitize the NIST Technical Series publications on a large scale, it became apparent that further information about the publications was required. Over the years, multiple organizations within NIST had published different series within the NIST Technical Series. No complete index of all the publications existed. The problem was complicated by the lack of a clear date when series previously published in print became born digital. Between the 1990s and 2012, NIST Technical Series publications were published in print, born digital, and

[1] Office of Science and Technology Policy. "Memorandum for the Heads of Executive Departments and Agencies: Increasing Access to the Results of Federally Funded Scientific Research." Whitehouse.gov. February 22, 2013. Accessed October 1, 2016. https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf.

often published in both formats. In the early days of NIST agency publications, NIST maintained publications indexes listing all Technical Series publications.

2) Changes in past cataloging practices and ensuing issues with associated metadata.

Changes in cataloging practices throughout the years also provided a challenge for the project. Some legacy Technical Series had been cataloged at the item level, with one bibliographic record per individual publication title. Other series were cataloged as serials, with one bibliographic record per series title. In those cases, the individual publications were added as item records and attached to the bibliographic record in a long list. They contained very limited metadata beyond their publication numbers and titles. In other cases, some publications within larger series were missing from the catalog altogether.

These inconsistencies led to the decision to update the catalog's metadata as the digitization project progressed. Librarians now create a new record for each electronic publication as the publication is digitized. By the end of the digitization project there will a complete set of new MARC (or MAchine Readable Cataloging) records for all of the digitized Technical Series publications, in addition to the existing records for the Technical Series print volumes.

Scope and Selection

Using a variety of resources, the project coordinator compiled a list of the series to be digitized in order to estimate the total publication count, and to determine the order in which each series would be digitized. The librarians then examined this list to determine whether there were any restrictions on dissemination.

One series in particular had dissemination restrictions because the work they described had been carried out for other organizations. While those other organizations were able to release these publications, NIST could not release them at the time they were written without the permission of the NIST director. This series, the NBS Reports – commonly known as "graybacks" for their gray covers— are frequently requested by customers. The NIST Director granted permission to release a set of 1800 of these reports in October 2015. Up to that point, the NBS reports were copied and made available to customers upon request and on a case by case basis.

Because records did not exist for each publication, particularly those created in the early days of the National Bureau of Standards, there was no accurate number of publications in each series. Estimates were created based on the internal tracking system of individual publication numbers assigned to each technical report. Although individual publications had been marked

as withdrawn or superseded through the years and thus would not be scanned, this method created a fairly accurate estimate.

ISO uses an existing contract through Fedlink, a federal library consortium, to arrange the digitization of the NIST Technical Series publications through the Internet Archive.[2] The Internet Archive is a non-profit library of free books, software, movies, and other materials, with a mission of ensuring access to digitized materials in perpetuity. Internet Archive produces PDF copies of its digitized publications in addition to other derivative file types, including JPEG 2000, EPUB, and Kindle. All of the materials digitized by Internet Archive are available from their website.

Internet Archive's scanning facilities include high resolution scanning equipment and a dedicated staff. They were able to quickly digitize the NIST publications on a large scale and using specialized equipment.

ISO began sending its publications to Internet Archive scanning facility at the Library of Congress in July 2011. The publications were boxed in batches of 300 to 600 and shipped to Internet Archive's scanning facility, where they were digitized with overhead scanners, and then shipped back to ISO. Each returning shipment contained a manifest that included the digitized file names, which ISO then used to download its own copies of the digitized files.

Metadata

Internet Archive requires metadata for each print publication sent for digitization, so one of ISO's initial challenges for this project was to create or obtain complete metadata for each publication. Due to inconsistencies in cataloging practices over time, ISO could not simply download the MARC records from the catalog and transform them into Internet Archive's required format.

After taking these inconsistencies into account, ISO decided in early 2013 that it would be best to create new and accurate metadata for each publication. With this method, each digitized publication has its own electronic catalog record. During this time, the new RDA[3] cataloging standard was implemented by the Library of Congress, reinforcing ISO's decision to require one record per item format. Previous ISO catalogers had added URLs to existing print records in order to direct the customer to the digital item's location. However, the method of creating and maintaining a separate MARC record for each digitized publication has since allowed ISO to accumulate a complete set of metadata solely for its digitized collections, and to transform that metadata as required for a variety of different uses and formats.

---

[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 Resource Description and Access (RDA) cataloging standard was developed by the Joint Steering Committee for Development of RDA (JSC) as the successor to the Anglo-American Cataloguing Rules, 2nd Edition Revised (AACR2), which had been the cataloging standard for American libraries since 1978. RDA was fully implemented by the Library of Congress on March 31, 2013. For more information, about RDA, see http://www.rda-jsc.org/archivedsite/index.html.

The metadata creation process began with existing MARC print records. The MARC data was downloaded from ISO's online catalog in spreadsheet format. A library technician created the batches for digitization at Internet Archive, compared the volumes on shelf with the MARC metadata, and used that information to create metadata for each item to be digitized. This metadata then accompanied the batch of publications sent to Internet Archive. ISO also used this metadata to create MARC records for the digitized publications.

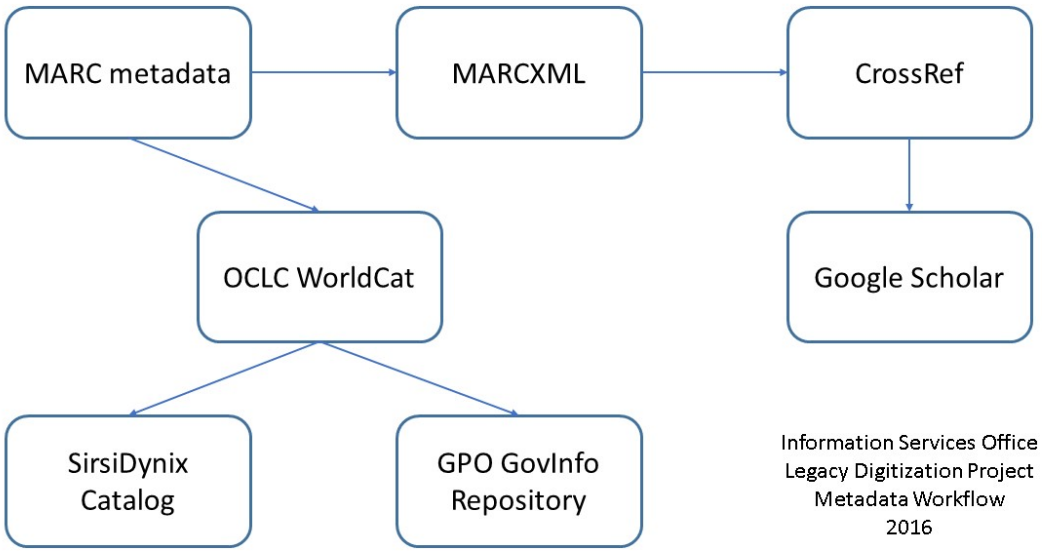
Transformations and Identifiers

The Technical Series publications metadata was transformed several times after it was created. The metadata began its life in a spreadsheet, and was then transformed into a MARC record using the MarcEdit tool. The MARC records were then transformed several times (see Figure 1).

First, the metadata was transformed to MARCXML, again using the MarcEdit tool. The MARCXML was then transformed into CrossRef's metadata schema. The metadata was then uploaded to CrossRef at which point the publication's Digital Object Identifier (DOI) was activated.

The CrossRef XML was then transformed into Google Scholar's schema and made available to be crawled by Google.

The MARC records were uploaded to WorldCat. Once the records received Online Computer Library Center (OCLC) numbers, they were then uploaded to ISO's online catalog, and also sent to the Government Publishing Office (GPO)'s govinfo database along with their corresponding publications.



NIST Technical Series Metadata Workflow (Figure 1)

GPO and FDsys

The OSTP memo of 2013 mandated that research publications and scientific research data that resulted from federal funds be made publicly accessible. Newly-published NIST Technical Series publications fell within this category, as did NIST's peer-reviewed Journal of Research NIST, scientific research data and external peer-reviewed publications in scholarly journals. As NIST began to prepare to meet the requirements of the OSTP memo, ISO was asked to investigate and recommend online repositories to house, provide access to, and track the use of NIST content. As a result of this effort, NIST selected the Government Publishing Office's FDsys to serve as the repository for the Technical Series Publications.

In 2011, ISO staff had viewed a presentation by FDsys that described the repository as a place to preserve and access Federal documents. GPO had designed an infrastructure that paid special attention to standards for metadata and digital preservation. The FDsys site considered itself "content agnostic," meaning it was designed to accept any file type. It had a standard default interface, but GPO could implement a custom interface suited to NIST's needs. GPO also had a roadmap in place to address digital preservation, provided authentication and digital signatures for all of the publications in its repository, and made each publication's MODS and PREMIS metadata available for customer download.

FDsys Implementation

During the implementation process, ISO ironed out many of the details of how its publications would be displayed in the repository. NIST was among the first executive agencies to deposit content into the new FDsys repository and, as such, ISO served as one of their "guinea pigs." During the uploading process, ISO worked with FDsys to determine ISO's publication requirements, and how to best address them within the system's capabilities.

GPO offers its customers the option of selecting a standard or custom collection in the FDsys repository. A standard collection was the default option. It provided a basic search, browse, and display of uploaded materials with minimal customization to the FDsys software. A custom collection allowed the customer to work with the FDsys development team to create a customized display of the customer's content. NIST chose to create a standard collection in FDsys.

FDsys accepted content in batch format. The batches consisted of publications and their accompanying metadata. Each publication was packaged individually, and each package consisted of a PDF of the publication, its metadata in the form of a MARC record, and the JPEG 2000 files of the digitized publication created by Internet Archive. The JPEG 2000 files were sent to FDsys for preservation in its repository, and are not made available to customers.

Initially, ISO uploaded a number of publications from the NIST Technical Note series, which was the first NIST Technical Series to be digitized. FDsys was not set up to ingest files directly from the agencies. Instead of depositing the test files directly, ISO uploaded the files to GPO for processing. The test batches allowed ISO and the FDsys team to optimize display of publications both in search results and on each publication's details page.

The FDsys team worked with ISO to customize the date display on each publication's details page. FDsys' system was set up to display full dates, including month, day, and year, and any alterations to that would mean a global change throughout the system. ISO's early publications did not always include the month and day of publication, and so the date display was incorrect. However, since many of ISO's customers search for publications by their publication date, ISO did not want to suppress the date field entirely. Instead, ISO worked with the FDsys team to create a workaround to display the date as January 1 if further information was unavailable. On ISO's end, this involved a minor change in the format of any uploaded batches of publications and metadata to FDsys. A note field was also added to the item display page informing the customer of the details should the date display as January 1.

The final version of the FDsys publication details page included the PDF publication file, MODS[4] and PREMIS[5] metadata, and a zip file available for download. In addition to the date information, the details page identified each publication as an Executive Agency publication; provided its Superintendent of

[4] Metadata Object Description Schema (MODS) is a bibliographic XML schema for descriptive metadata, such as title, author, and publisher information, subject headings, permanent identifiers, or URLs.
[5] The PREservation Metadata: Implementation Strategies (PREMIS) data dictionary is a framework for technical metadata, including intellectual, object, and event entities, individuals associated with a work, and rights information.

Documents (SuDocs) classification; identified its government author as the National Institute of Standards and Technology; and specified its particular Technical Series name.

ISO's Content in govinfo

In early 2016, GPO released the new version of FDsys, which was renamed govinfo. This new version included a number of additional changes to ISO's collection in the repository, including improvements to both access and the publications' details.

The site's navigation was streamlined and updated. Additionally, NIST was offered the option of creating a landing page for its publications. A landing page was something ISO had requested in the beginning of the implementation process, but FDsys had not been able to provide one for a standard collection. When browsing NIST's publications on FDsys, customers had previously navigated directly to NIST's collection, which could then be limited by series name, date published, and other criteria. As ISO had only uploaded several series at that time, this did not impact the publications' accessibility. However, ISO planned to digitize at least 25 series and upload over 24 000 files, and was concerned that the browsing function might not be sufficient for optimal discovery of such a large number of publications.

Govinfo's ability to include a landing page solved this problem. It will significantly improve the accessibility and discovery of NIST's publications. The landing page[6] included links to allow customers to search and browse the collection, but also includes each series title, as well as its date range and a description of its parameters. Each series title links directly to the publications in that series, which can then be limited by other criteria. The landing page allows govinfo's customers to get a more inclusive look at all of ISO's content available in govinfo's repository. The publication details pages were also updated in a subsequent govinfo iteration to include more metadata. The pages now include the authors, the accurate publication date, and subject headings.

In Media Res

As of October 1, 2016, ISO has digitized over 16 000 of the NIST Technical Series publications. Seventeen Technical Series have been digitized. As part of the digitization process, all of the Technical Series publications have been made available to the public in several places, including in ISO's collection on archive.org, through the ISO online catalog, and in GPO's govinfo repository.[7]

Each year, ISO assessed its progress and reviewed the order for digitizing the remaining publications. Changes were made based on a number of factors: customer response; requests for publications; how quickly a smaller series could be digitized in comparison to a larger one; and additional funding. At the outset of the project, ISO was able to process about 1500 to 2000 publications per year. With funding

[6] NIST's govinfo landing page is located at https://www.govinfo.gov/browse/nist.
[7] The NIST Technical Series publications are available through Internet Archive at https://archive.org/details/NISTresearchlibrary.

from NIST management, ISO hired two librarian contractors to help with preparing batches for shipment to Internet Archive and processing digitized publications, which considerably increased the speed of the project.

Customer Feedback

Customers' response to the digitization project has been very positive. ISO's main customers are NIST researchers, many of whom have commented favorably about the availability of the Technical Series publications digitized so far. ISO has publicized the digitization of each legacy series upon completion, and customers are also told that their newly published Technical Series publications are being sent to govinfo for preservation.

This positive response has resulted in increasing requests for particular legacy series to be digitized more quickly, as well as inquiries as to when different series will become available. In response to these inquiries and requests from external customers, ISO changed its digitization order to meet customer demands.

Next Steps

ISO expects to complete the digitization of the NIST legacy publications by 2018. Approximately 8000 Technical Series publications remain to be digitized as of October 1, 2016.

ISO will also create a similar index. That index will contain a comprehensive list of each series detailing all of its titles and their authors, as well as the DOI assigned to each title. The index will encompass both the digitized legacy Technical Series publications, as well as the ongoing Technical Series publications still published by ISO today. The comprehensive index of the Technical Series publications and their DOIs is expected to be made available on the NIST website in 2017, and will be updated as future titles are published and digitized.

ISO employs the Baldrige Excellence Framework[8] to improve its organizational performance, and in accordance with that, ISO continues to improve and streamline the many arms of the digitization project. The project's workflow has been very successful, and ISO continues to improve the process where possible.

The digitized publications continue to be a valuable resource for ISO's customers. ISO has begun to link their contents to other types of materials in its collections, including museum objects and archival collections. Several Technical Series publications are also linked to archival photos

[8] The Baldrige Performance Excellence Program's Baldrige Excellence Framework at https://www.nist.gov/news-events/news/2014/12/2015-2016-baldrige-excellence-framework-and-criteria-businessnonprofit.

in the NIST Digital Archives. The ability to make these connections between different types of materials within its collections will provide ISO's customers with a deeper, more dimensional understanding of NIST research and accomplishments.