Creating Workflows to Scale Out Open Access E-book Acquisitions at the Library of Congress

|  |  |  |
| --- | --- | --- |
| **Andrew Cassidy-Amstutz** | **Kristy Darby** | **Elizabeth Holdzkom** |
| *Library of Congress*  *United States of America*  *acas@loc.gov* | *Library of Congress*  *United States of America*  *kdar@loc.gov* | *Library of Congress*  *United States of America*  *eholdzkom@loc.gov* |
| **Camille Salas** | **Lauren Seroka** |  |
| *Library of Congress*  *United States of America*  *csala@loc.gov* | *Library of Congress*  *United States of America*  *lseroka@loc.gov* |  |

**Abstract – This poster illustrates the origins, exploration, and routinization of Open Access (OA) e-book acquisition, processing, and preservation at the Library of Congress (LC) led by the Digital Content Management section (DCM). This includes discussion of technical workflows, special issues and opportunities presented by expanded telework during the COVID-19 pandemic, ongoing challenges in expanding and routinizing OA e-book collecting, and iterative process development.**

**Keywords – Open Access, E-books, Metadata, Acquisitions**

**Conference Topics – Resilience**

In January 2017, the Library of Congress adopted a set of strategic steps related to its future acquisition of digital content entitled Collecting Digital Content at the Library of Congress. The first strategic objective in this plan is "expand and routinize acquisition and access of openly licensed and openly available digital works," identifying OA e-books as the target for year one. The Digital Content Management Section (DCM) engaged fully in this effort, collaborating with divisions throughout the Library and working with multiple acquisition streams to review and analyze the large amount of e-book content coming into the Library.

DCM led several pilot projects to test technical methods for obtaining e-books from various sources, transforming descriptive metadata, and processing the content for presentation on the Library's website, loc.gov. The first pilot explored the level of effort required to process ten OA e-books identified in the Directory of OA Books (DOAB) for which the Library already had print holdings. DOAB is a community-driven platform that provides data and e-book files, when available, for nearly 60,000 peer-reviewed OA e-books from more than 600 publishers. This resource offered a unique opportunity to systematically identify OA monographs that can be added to the Library’s collection. Acquiring the e-book files, putting them in managed, long-term preservation storage, and providing access on loc.gov rather than linking to the files on the open web required experimentation and iteration.

DCM staff took the lessons and workflows from this pilot of ten books and embarked upon another pilot to gauge how well the work would scale, this time working 100 titles through the process. Staff then began regularly processing DOAB e-books for which the Library already had print holdings.

This work required the creation, refinement, and evolution of workflows dealing with both metadata and digital content. DCM staff collaborated with staff from the Library of Congress Integrated Library Systems Program Office (ILSPO) to transform existing MARC bibliographic records for print books to correctly-formed bibliographic records for the corresponding e-book in batch using the MarcEdit software suite. Each e-book MARC record included the Creative Commons license information applied to the title in the 540 field, which was a new practice at the Library. Over time, DCM staff have developed and continue to develop Python scripts to pull and analyze MARC data as well as to create and enhance MARC bibliographic records in bulk. Processing the e-book content for online presentation required new workflows as well. The infrastructure of loc.gov requires thumbnail images be created and preserved alongside the e-book file for presentation. DCM staff established workflows to manage these files and generate derivatives using Bash and Python.

The COVID-19 pandemic prompted the Library to pivot to telework in March 2020. DCM and the Collection Development Office (CDO) developed a pilot project to enable Recommending Officers, the Library's subject matter experts, to select electronic books from DOAB remotely for inclusion in the Library's permanent collection. The project resulted in the development of a functional end-to-end process allowing titles from DOAB to be identified according to LC subject areas, selected, described, preserved, and made available on the Library's public website. From the beginning of the pandemic in March 2020 until September 2021, the Open Access Books collection on loc.gov grew from approximately 300 titles to over 3500 titles. The Library's shift to telework and the resilience and flexibility of staff across the organization created a unique opportunity to grow the Library's collection of OA e-books in spite of the challenges posed by the pandemic. This project helped raise awareness of the ongoing engagement with OA monographs at the Library as well as the possibilities of expanding and routinizing the work.

While DCM staff worked on developing and applying the workflows to process and make the e-books available to users of loc.gov, staff from many service units at the Library came together to contribute to the ongoing success of this program. The success of the telework project was only possible because of the strength of collaboration. The OA e-books endeavor thrives because of the partnerships between and among DCM, the CDO, ILSPO, General and International Collections Directorate, Special Collections Directorate, Office of General Counsel, Acquisitions and Bibliographic Access Directorate, and the Office of the Chief Information Officer. All of these entities continue to dedicate resources, time, and expertise to build and support the Library’s OA e-book program.

The work of routinizing, acquiring, processing, and making available OA e-books at the Library of Congress is not without challenges. These challenges arise in the technical work required to manage e-books and make them available for use as well as analyzing and repurposing descriptive metadata, created both at the Library as well as supplied by aggregators and publishers. The quality of files and metadata varies based on supplier; identifying, isolating, and determining the best ways to work with the corrupt, incorrect, or incomplete data is often time- and resource-consuming. As each OA e-book has an OA license which is reflected in the MARC bibliographic record, ensuring that licenses match the appropriate e-book is critical and sometimes difficult. Finally, a large collaborative project involving dozens of engaged staff members from many divisions and directorates requires transparency, flexibility, extensive documentation, and careful and clear communication at every point.

DCM launched a phased initiative in October 2021 to review the pilots' processes with the goal of creating mature and routinized workflows. The result was increased automation and scalability of the workflows through the development of various Python scripts. With improved efficiency and increased output, the Open Access Books Collection nearly doubled to 5600 titles as of August 2022. OA e-book processing is now routine at LC and we will repeat the workflow in the winter of 2022 and expect to continue to grow the collection on an annual basis.