**Documenting the Dataset Catalog**

Kristi Torp

NLM Associate Fellow 2023-2024

June 20, 2024

Project Sponsor: Alvin Stockdale

**Table of Contents**

[Acknowledgments 3](#_Toc169864532)

[Abstract 4](#_Toc169864533)

[Objective 5](#_Toc169864534)

[Method 5](#_Toc169864535)

[Outcomes 6](#_Toc169864536)

[Conclusion 6](#_Toc169864537)

# Acknowledgments

I want to thank my project sponsor, Alvin Stockdale, and the Dataset Catalog owner, Peter Seibert, for their encouragement and teaching me many valuable skills. They gave me an understanding of what goes into creating a new NLM product and how much can be accomplished with initiative and creativity. Another thanks to Mike Davidson for teaching OpenRefine to the Associate Fellows, as it helped save a lot of time with data cleanup. I also want to thank Teresa Zayas Caban for the excellent policy discussions and direction.

I want to give special thanks to my mentor, Kathel Dunn, for being a listening ear, cheering me on, and helping me navigate my projects this year.

# **Abstract**

**Objective**: This project's objective was to create new documentation detailing the environmental scan methodology and further develop collection development guidelines for the NLM Dataset Catalog, which is currently in beta launch.

**Methods:** The Associate Fellowcontinued an environmental scan to find new medical dataset repositories for possible inclusion in the Dataset Catalog.  She documented the process, including API calls and questions that arose, and formatted the document similarly to other NLM standard operating procedures. She met with policy experts to discuss best practices for developing inclusion criteria and began revising the current external collection development guidelines.

**Outcomes:** This project resulted in a new document detailing the procedures for the environmental scan methodology for the NLM Dataset Catalog. In addition, it contains new recommendations for the collection development guidelines based on feedback from multiple stakeholders at NLM.

**Conclusion:** The developed process documentation will allow others to follow and adapt the methodology more easily in the future. The documented methodology follows the structure of other NLM standard operating procedures, and the updated inclusion criteria recommendations better reflect NLM collection development guidelines.

# **Objective**

The Dataset Catalog is a new product from the National Library of Medicine currently in beta launch. This catalog, often described as the "PubMed of datasets," includes descriptive information about datasets from repositories managed and hosted by various institutions made available through a single, easy-to-use interface. The Dataset Metadata Model (DatMM) facilitates this interface. DatMM gives the Dataset Catalog standardized metadata for biomedical datasets in a linked data model that allows users to search and retrieve datasets across multiple repositories. New products require new documentation to understand workflows better and record procedures for future use. The objective of this project was to create new documentation detailing the environmental scan methodology and further develop collection development guidelines for the Dataset Catalog.

# **Method**

I began the project by continuing an environmental scan to find new potential medical dataset repositories for possible inclusion in the Dataset Catalog. This scan used the repository finder Re3data.org to discover new potential repositories. During working meetings, my project sponsor and I worked together to develop the best way to discover, retrieve, and evaluate potential repositories. I copied all relevant information, including search terms and API calls, as we moved through the process. I also noted any difficulties or lessons we learned in my documentation. For example, we observed the importance of checking subject hierarchies, as “medicine” is both a primary and secondary heading, and we wanted the primary heading.

My project sponsor used the API calls I copied to develop a Python script to download all the required repository metadata into a spreadsheet. I then copied all the previously reviewed repositories into the new spreadsheet and deduplicated the potential list from the existing list. I used Excel and OpenRefine to deduplicate and clean the data. I then began evaluating the new repositories based on the current collection development requirements given to me by my project sponsors, documenting the process and questions that arose to help inform the edits I would make to develop the collection development guidelines further.

At this time, I also met with a policy expert at NLM to discuss what the inclusion criteria should look like and include. During this meeting, I learned how NLM shares inclusion criteria in PubMedCentral (PMC) and MEDLINE on public-facing websites so that the public and potential journals know what NLM expects. These products used a table to organize their inclusion criteria, so I adopted a similar approach. We also discussed some of the current policy issues with the current Dataset Catalog collection development guidelines. Because the inclusion criteria were part of these guidelines, I included the guidelines in my update. I again looked to PMC and MEDLINE to inform how I drafted the text. Once my project sponsor and I reviewed the drafts for both the guidelines and the environmental scan process, I had them reviewed once more by NLM Policy. We discussed further changes, and I connected with NLM staff from PMC and MEDLINE to view examples of their procedural documentation. I changed my process documentation from a narrative format into one that reflected standard operating procedures (SOPs) with numbering and bullets for easier readability.

I attended informational sessions about the Dataset Catalog to see if I missed important information about the scope and inclusion criteria. I participated in the practice with NLM staff, the actual office hours, and the technology showcase at the annual Medical Library Association conference. I noted the questions at each meeting and found that there were often questions about inclusion criteria, the linked data model (DatMM), and clarification about whether the Dataset Catalog was a repository. These questions, plus the ones I had while evaluating potential repositories, helped inform how I would write the process documentation and revise the inclusion criteria guidelines. For example, I included definitions of uncommon words such as "knowledgebase" and "domain-specific." Once I finished evaluating, I indexed the new repositories using the categorization scheme created by an Associate Fellow the previous year. I recorded my process and had a fully documented environmental scan methodology developed into standard operating procedures.

# **Outcomes**

The Dataset Catalog team now has a record of the internal process of discovering and evaluating potential repositories. The procedures are numbered in an easy-to-follow format and include information about the software and technical skills needed to follow them. I also gave recommendations for updating the collection development guidelines on the website, particularly the inclusion criteria, based on feedback from the project sponsor, NLM policy experts, and my experience reviewing potential repositories. Policymakers place great importance on a fully developed policy before a product launch in order to protect NLM. However, this is challenging to achieve with a new product as it develops and changes during and after the beta phase. Unresolved issues outside this project's scope remain between the Dataset Catalog product owner and the policy team, so the updated guidelines are considered recommendations only. If this project is adapted or expanded in the future, I recommend studying existing NLM product policies and standard operating procedures and consulting with owners of other NLM products. Despite the unresolved issues, this project helped clarify the evaluation process, how naming metadata would be recorded, and which inclusion criteria were necessary and understandable for the new Dataset Catalog.

# **Conclusion**

The Dataset Catalog is in beta phase as the product owner gathers data about its usage. The product is subject to change as it grows during this phase, and its guidelines and methods may also need revising in the future. If so, the developed process documentation will allow others to follow and edit the process more efficiently. The documented methodology follows the structure of other NLM standard operating procedures, and the updated inclusion criteria recommendations incorporate feedback from the product owner and NLM policy experts and better reflect NLM collection development guidelines.