Automating Metadata Creation:
Enhanced Discovery and Description of Maternal Child Health Data

Saurabh Vashishtha[1,2], Amanda Harrigan[1], Sharon Farnel[1], Kendall Roark[3]
[1] University of Alberta Libraries [2] Department of Medicine, University of Alberta, [3]Purdue University

Introduction

Only a small fraction of data collected through clinical and health research is ever shared, while a considerable amount of data remains 'hidden' or undiscoverable. This limits the potential for secondary analysis and long-term value of the data.

We are in the process of developing a study catalogue which will describe and make discoverable selected data and research records of approximately 40 clinical trials and cohort studies by Women and Children's Health Research Institute (WCHRI) researchers.

This study catalogue will help researchers reduce duplication of research, make their research more visible, and promote collaborations among different research groups with similar research interests.

Our aim is to make study documentation and metadata creation as efficient and easy to integrate into researcher workflows as possible.

Question

How can we repurpose metadata embedded in online registries, published articles, and discipline specific research documents, like codebooks and protocols?

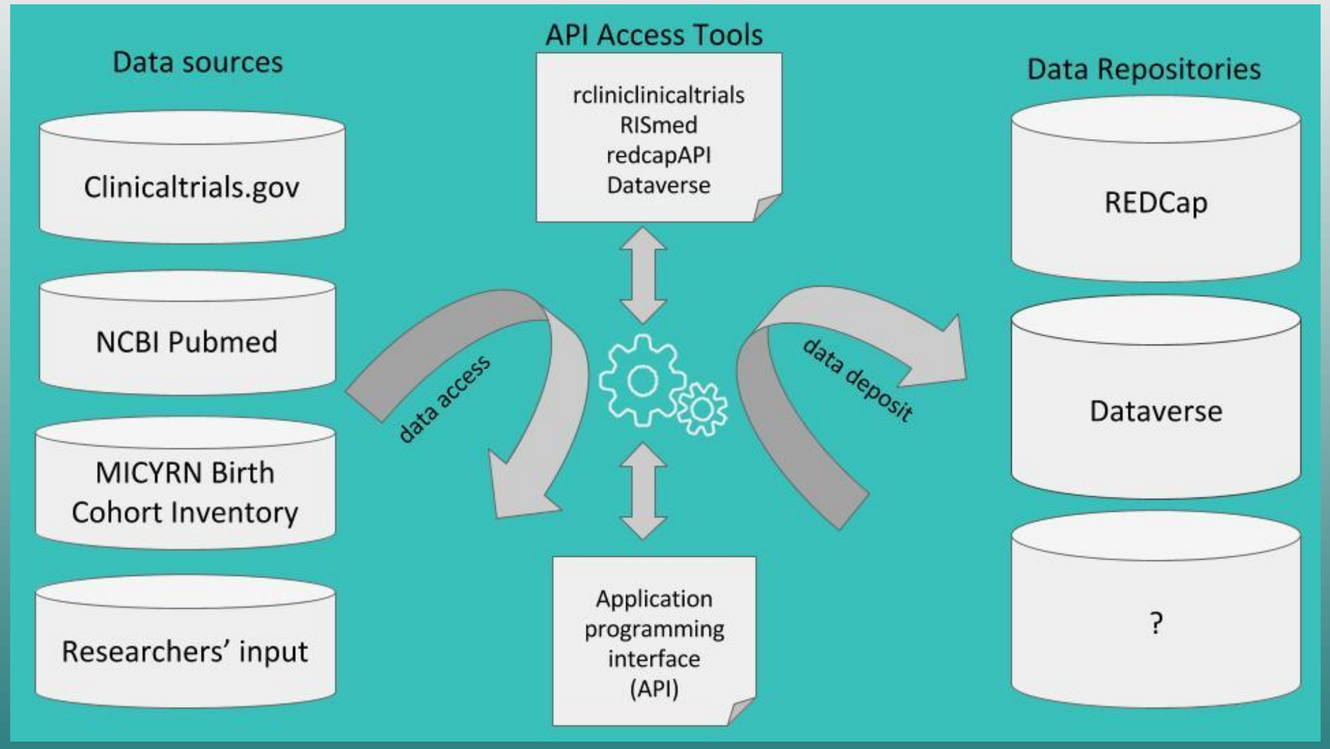
Methods

We implemented processes to pre-populate parts of our DDI-compatible metadata schema with publically and locally available information. We have developed in-house scripts to export and import the metadata from resources such as NCBI, clinicaltrials.gov, REDCap and Dataverse using APIs.

This is a semi automated process and various 'R' based packages are used to access the metadata using APIs.

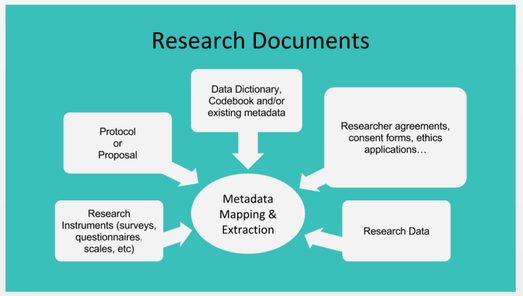
As well, publically unavailable data will be accessed from the research documents that researchers who agreed to participate in the study share with us. It will be processed and described using the DDI-compatible metadata schema.

Further, metadata may be mapped and deposited to different data repositories, as yet to be determined.



This project is supported by the CLIR Cataloging Hidden Archives and Special Collections Grant for 2015-2016, *Bridging the Research Data Divide: long-term value and access for historical and contemporary maternal, infant and child research data" (PIs: Gustainis E, HarvardU; Farnel S, UAlberta; Roark K, PurdueU)*

Next Steps



Useful metadata is stored in different sources, such as research protocols, publications or codebooks, and in different formats, such as access databases, csv files, pdfs, and word documents.

These unstructured, text heavy sources contain information useful to describe the studies and populate the metadata.

We will use the R-based 'text mining (™)' package to search for and extract useful metadata from these different sources and populate our structured metadata schema.

Corresponding Authors

Saurabh Vashishtha
Data Curation Assistant
vashisht@ualberta.ca

Kendall Roark
Corresponding Co-PI
roark6@purdue.edu