CAPSTONE PROJECT CHARTER

Project Information

Project Title:
Automatic Keywording Investigation

Abstract:
Librarians manually keyword large datasets to enhance information findability in IHME's catalog. Automating this process could significantly streamline the cataloging workflow. This project documents data characteristics of IHME's collection to inform an automatic solution. Research revealed that an auto-keywording tool which could process ALL data types would require large investment into a sophisticated computational linguistics solution. This project delivers a practical solution focused on a few specific data types to immediately save Librarians hours of work per data set. The project also provides documentation as lasting evidence for future development and/or funding of a larger Language Learning Model project.

Team Member Names:
Tina Nowak

Sponsoring Organization:
Institute for Health Metrics and Evaluation

Project Tile

Team Information

Student Information

Tina Nowak
tnowak@uw.edu

Research IHME data types and identify candidates for automatic keywording. Develop a Proof-of-Concept recommending whether or not a Python program could successfully automatically keyword IHME data. Potentially develop the program itself.

Sponsor Information

Lyla Medeiros, Data Library Curator II, Taxonomy Specialist
Institute for Health Metrics & Evaluation
lylam@uw.edu
https://www.healthdata.org/

Provide training and orientation to IHME's data catalog and keyword taxonomy. Supervise research and project strategy development.

THE INFORMATION PROBLEM

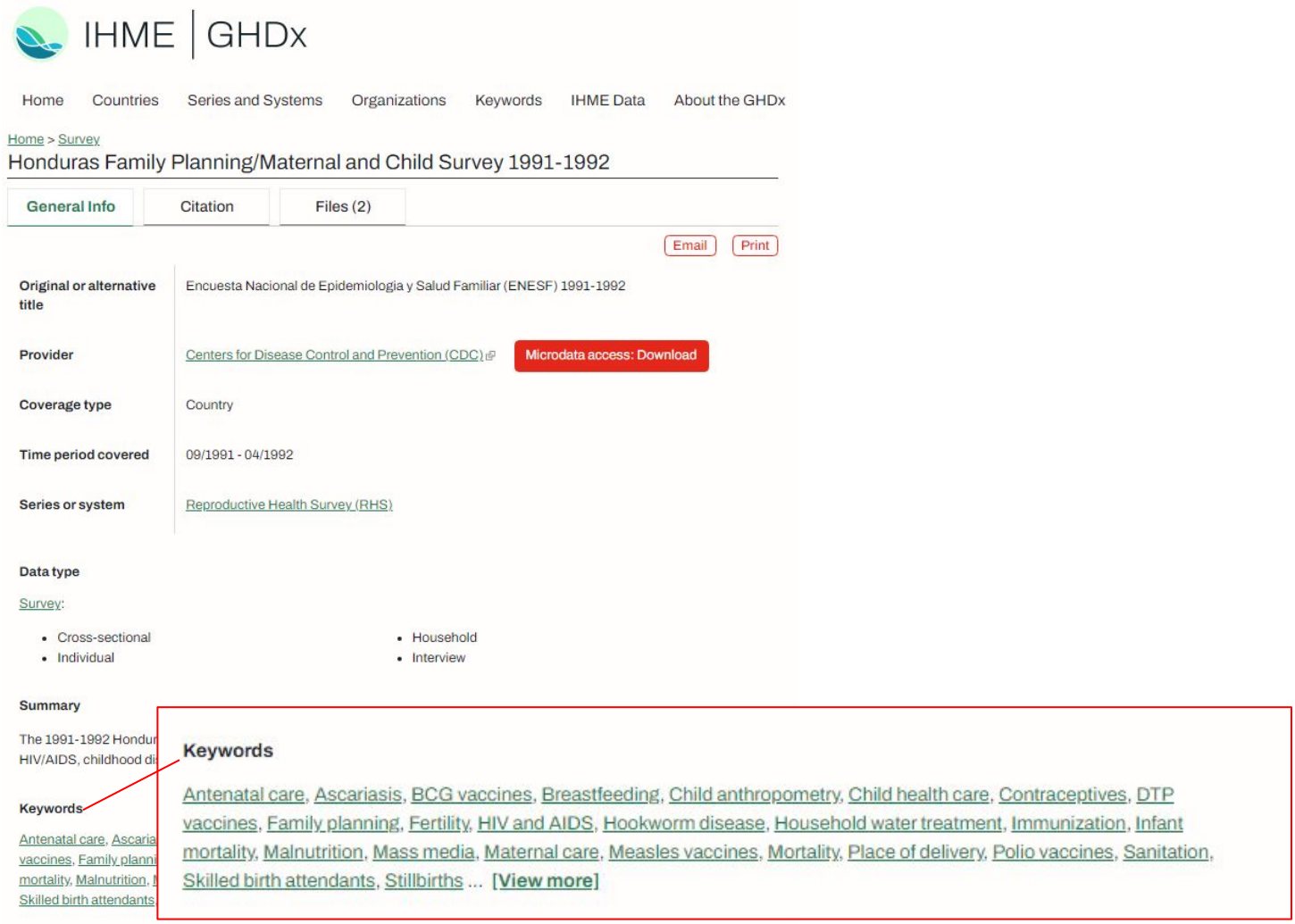
The Institute for Health Metrics & Evaluation (IHME) aims to improve the quality and longevity of life worldwide by publishing data to inform health policy and practice.

IHME's prominent work is the Global Burden of Disease (GBD) meta study which aggregates a large number of data sources to determine how diseases, illnesses, and risk factors are affecting mortality and disability of various demographic groups.

Librarians obtain primary data and organize it to make it useful and reusable for IHME researchers. They organize data according to its contents using a custom keyword taxonomy. Keyworded datasets enable researchers to utilize tools which manage their workflows.

Data librarians catalog a large volume of incoming datasets. Keywording is the most time consuming part of the processing workflow. They could intake and process more data sources if they could find ways to save time in the processing workflow.

Is it possible to automate part of the dataset keywording process to help Data Library Services spend less time keywording?



Example of keywords that are each manually applied to a datasource by IHME Data Librarians.

PROJECT OBJECTIVES

- Investigate and understand the information problem
- Establish Proof-of-Concept for an automatic keywording program
    - Create documentation which describes characteristics of IHME data types that should be considered in the design of a Python program
    - Create documentation which articulates applications of IHME's Keyword Taxonomy to incoming data
    - Articulate scope of the program based on research findings
- Write an automatic keywording program and test it with example datasets

KEY PROJECT & RESEARCH INSIGHTS

Documenting data type characteristics

- Data types are complex in **content** and **carrier**
- **Content**
  - Data is expressed in a variety of ways in varying levels of complexity
    - Survey
    - Report
    - Article
    - Data table
    - Etc
  - Ranges from simple numeric data tables to hundreds of pages of natural language
- **Carrier**
  - Variety of ways public health data is formatted
    - PDF
    - STATA file
    - DO file
    - Text Document
    - CSV / XSLX
    - Etc

Understanding applications of the keyword taxonomy to content

- The keyword taxonomy is used to tag keyword search terms to research data content
- Keywords can be literal or contextual
  - **Literal**
    - See the word in the data, tag the corresponding keyword
  - **Contextual**
    - See the idea/concept in the data, tag the corresponding keyword
- It is possible for catalog items to be tagged with both literal and contextual keywords which complicates the development of a automatic keywording program

APPROACH TOWARDS SOLUTION / KEY FEATURES / SURPRISES

IHME's mix of literal and contextual custom keywords means that an automatic program should be capable of handling both cases

- Literal keywords are applied using straightforward text matching
- Contextual keywords require a more complex solution that can process the concepts & ideas within natural language
    - Natural Language Processing, Large Language Models, Machine Learning, etc will be necessary

A program that could keyword ALL IHME data is currently unfeasible due to the large resource investment required for sophisticated natural language processing models.

A short term solution will still help librarians speed up their cataloging workflow – enabling data to be published for public health research benefit even faster. A Python program designed to keyword just 1 or 2 currently-feasible data types could actually save librarians hours per dataset.

- Scope! Design a Python script with very specific parameters that can process data that is lengthy (i.e. time-consuming) but which has straightforward matches to keywords in the taxonomy
- Provide a flow chart helps Data Librarians know what is or is not eligible to run through the script

APPROACH TOWARDS SOLUTION

ICD-10-3	Geschlecht	Insgesamt	davon im Alter von … bis … Jahre												davon im Alter von ..
			unter 1 Jahr	1-4	5-9	10-14	15-17	18-19	20-24	25-29	30-34	35-39	40-44	45-49	50-54
		Berechnungs- und Belegungstage													
A02	m	264	10												
A02	w	194													
A04	m	5567									4		20	63	37
A04	w	5526		3						4		2		11	32
A05	m	41													
A05	w	19													
A06	w	2													
A07	w	4										4			
A08	m	658							7						4
A08	w	491													
A09	m	3692								1		2	8	20	112
A09	w	5004			1		2	10		11		2		14	1
A15	m	2267							22		29		133	201	398
A15	w	662							10			20	9	12	7
A16	m	71													15
A16	w	138											42	4	
A17	w	129											33		
A18	m	125													
A18	w	67													
A19	m	359						5		45					156
A19	w	227					11				6				16
A26	m	29													

Example of a currently feasible data type for auto-keywording. This type of dataset is a good candidate because it contains International Classification of Diseases (ICD) codes which could be used as Keys to easily tag corresponding keyword taxonomy Values.

Name	Date modified	Type	Size
DEU_INPATIENT_ADMISSIONS_2000_Y20...	3/20/2023 11:50 AM	Microsoft Excel 97...	20,639 KB
DEU_INPATIENT_ADMISSIONS_2001_Y20...	2/24/2023 6:57 AM	Microsoft Excel 97...	21,047 KB
DEU_INPATIENT_ADMISSIONS_2002_Y20...	2/24/2023 6:57 AM	Microsoft Excel 97...	16,865 KB
DEU_INPATIENT_ADMISSIONS_2003_Y20...	2/24/2023 6:57 AM	Microsoft Excel 97...	10,897 KB
DEU_INPATIENT_ADMISSIONS_2004_Y20...	2/24/2023 6:58 AM	Microsoft Excel 97...	11,954 KB
DEU_INPATIENT_ADMISSIONS_2005_Y20...	2/24/2023 6:57 AM	Microsoft Excel 97...	11,948 KB
DEU_INPATIENT_ADMISSIONS_2006_Y20...	2/24/2023 6:58 AM	Microsoft Excel 97...	15,465 KB
DEU_INPATIENT_ADMISSIONS_2007_Y20...	2/24/2023 6:58 AM	Microsoft Excel 97...	15,897 KB
DEU_INPATIENT_ADMISSIONS_2008_Y20...	2/24/2023 6:58 AM	Microsoft Excel 97...	12,757 KB
DEU_INPATIENT_ADMISSIONS_2009_Y20...	2/24/2023 6:58 AM	Microsoft Excel 97...	11,743 KB
DEU_INPATIENT_ADMISSIONS_2010_Y20...	2/24/2023 6:58 AM	Microsoft Excel 97...	10,094 KB
DEU_INPATIENT_ADMISSIONS_2011_Y20...	2/24/2023 6:58 AM	Microsoft Excel 97...	13,580 KB
DEU_INPATIENT_ADMISSIONS_2012_Y20...	2/24/2023 6:58 AM	Microsoft Excel W...	8,847 KB
DEU_INPATIENT_ADMISSIONS_2013_Y20...	2/24/2023 6:57 AM	Microsoft Excel W...	4,295 KB
DEU_INPATIENT_ADMISSIONS_2014_Y20...	2/24/2023 6:57 AM	Microsoft Excel W...	4,277 KB
DEU_INPATIENT_ADMISSIONS_2015_Y20...	2/24/2023 6:58 AM	Microsoft Excel W...	4,238 KB
DEU_INPATIENT_ADMISSIONS_2016_Y20...	2/24/2023 6:58 AM	Microsoft Excel W...	4,255 KB
DEU_INPATIENT_ADMISSIONS_2017_Y20...	2/24/2023 6:58 AM	Microsoft Excel W...	4,169 KB
DEU_INPATIENT_ADMISSIONS_2018_Y20...	2/24/2023 6:58 AM	Microsoft Excel W...	4,156 KB

The dataset example above is just 1 of many of the same type of dataset in a single data series.

Many data series exist which contain this type of data.

An automatic tool could save a librarian hours of work **per dataset** which ultimately will save them days of cataloging work.

NEXT STEPS / RECOMMENDATIONS

Documentation provided by this project may be used as:

- Proof-of-Concept for a future computer science project (taken on by an IHME employee or Capstone student)
- Evidence towards future investment in Large Language Model (LLM) development
    - Ideally IHME invests resources into an LLM trained on the body of IHME public health data
    - This is a much bigger undertaking than text matching, but complete keywording of all IHME data types will never be possible without it
    - Even with a fully functional automatic keywording program, IHME Data Librarians will likely need to be in the loop to verify correctness of keywords and to complete other steps of the cataloging workflow

The range of data at IHME is highly contextual and their research products are very specific. Humans will likely always be in the loop of maintaining this catalog. Future automatic solutions shouldn't be seen as a replacement of Data Librarians rather they should be seen as a tool that affords Librarians time and energy towards other curation activities that preserve information context.