An NLP-Based Approach for Metadata Extraction and Analysis of FDA Guidance Documents

Jayanti Das; Kylie Haskins; Laurie Muldowney; ShaAvhrée Buckman-Garner

Introduction

- The Center for Drug Evaluation and Research (CDER) generates approximately 100 new guidance documents each year that represent the Agency's current best thinking on a particular topic.
- There is currently no means for CDER staff to perform a full-text search of FDA guidance documents to learn more about the Agency's policies as they relate to a specific area of interest.
- The Office of Translational Sciences (OTS) developed an internal Pilot Text-Searchable Guidance Database that provides complete text-searchability of all PDF formatted FDA guidance documents to make it easier for CDER staff to identify the Agency's current thinking on specific topics, provide recommendations to sponsors, and harmonize policies across the Center.
- The database averages approximately 200 views per day, and OTS has received positive feedback that the database adds value and supports the efficient and accurate implementation of policies across the Center.

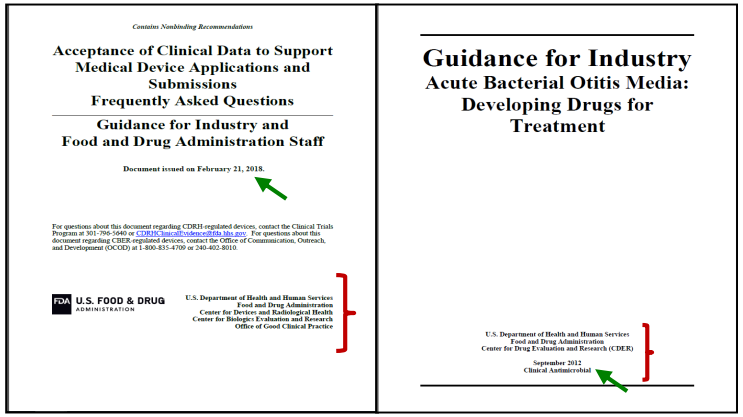
Challenges

- In an effort to improve the functionality of the database, OTS sought to add filter options, including filtering by Center and guidance type (i.e., draft vs final).
- Inconsistent metadata tagging of guidance documents and variability of guidance document formatting led to difficulties extracting the desired information to add this functionality.

Solution Method

- OTS developed an automated, natural language processing (NLP)-based approach with "R" software language that uses pattern-matching of metadata to extract fields and field-values for center and document type from all published guidance documents and then tags each document with this information with high accuracy and precision.

Examples of Different Document Formatting



Contains Nonbinding Recommendations

Acceptance of Clinical Data to Support Medical Device Applications and Submissions
Frequently Asked Questions

Guidance for Industry and Food and Drug Administration Staff

Document issued on February 21, 2018.

U.S. FOOD & DRUG ADMINISTRATION

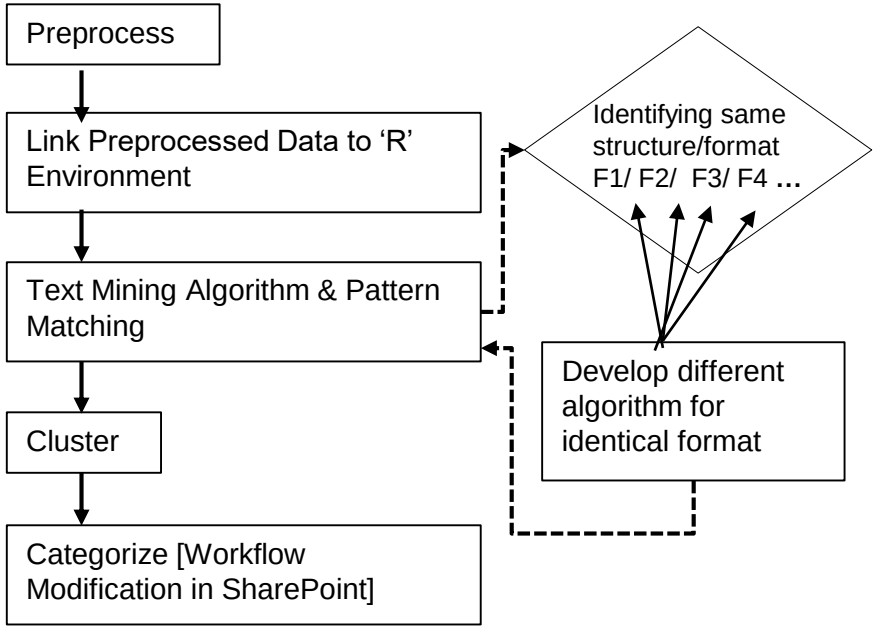
U.S. Department of Health and Human Services
Food and Drug Administration
Center for Devices and Radiological Health
Center for Biologics Evaluation and Research
Office of Good Clinical Practice

Guidance for Industry
Acute Bacterial Otitis Media: Developing Drugs for Treatment

U.S. Department of Health and Human Services
Food and Drug Administration
Center for Drug Evaluation and Research (CDER)

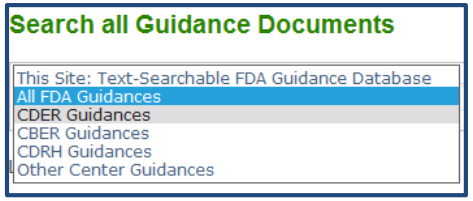
September 2012
Clinical Antimicrobial

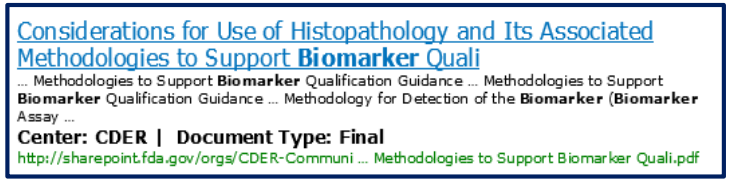
Method Flow Chart



Results

The NLP-based approach for metadata extraction and analysis allowed the addition of a center category filter to the Pilot Text-Searchable Guidance Database and also supported the inclusion of document type (i.e. draft, final) to be listed under the guidance title in the search results.





Conclusions

The NLP-based approach for metadata extraction and analysis successfully used pattern-matching of metadata to extract fields and field-values with high accuracy and precision. As the Agency continues to utilize more computational tools for organizing and sharing information on regulatory policies and reviews, potential applications of this NLP-based approach will likely increase to support accurate identification and extraction of requested information for dissemination across the and to the public. Agency

Acknowledgement

Jayanti Das is supported by a fellowship from the Oak Ridge Institute for Science and Education, administered through an Inter-Agency agreement between the U.S Department of Energy and the FDA.

Disclaimer

The poster is not a formal dissertation by FDA and does not represent Agency position or policy.