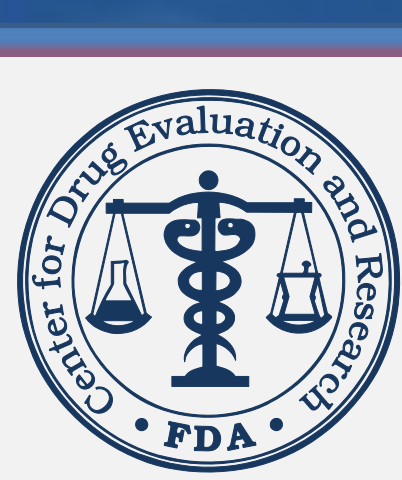




AskOCP: Semantically Enhanced Search applied to Clinical Review Documents



Eduard Porta Martin Moreno, Peter Lee

Division of Pharmacometrics, Office of Clinical Pharmacology, Office of Translational Sciences, Center for Drug and Research
US Food and Drug Administration, Silver Spring, MD, USA

Introduction

A wealth of text documents is archived at the FDA, with historical cases that may have certain characteristics similar to an ongoing NDA in terms of clinical outcomes or regulatory scenarios. Reviewers often find themselves in a situation where, they may want to go back and resort to exploring these archives. The amount of available textual data

is overwhelming, yet keyword based search often delivers a great number of documents that may contain the search terms, but are far from providing information relevant to the case.

We propose a search tool that strives to provide meaningful answers, by leveraging text analysis techniques.

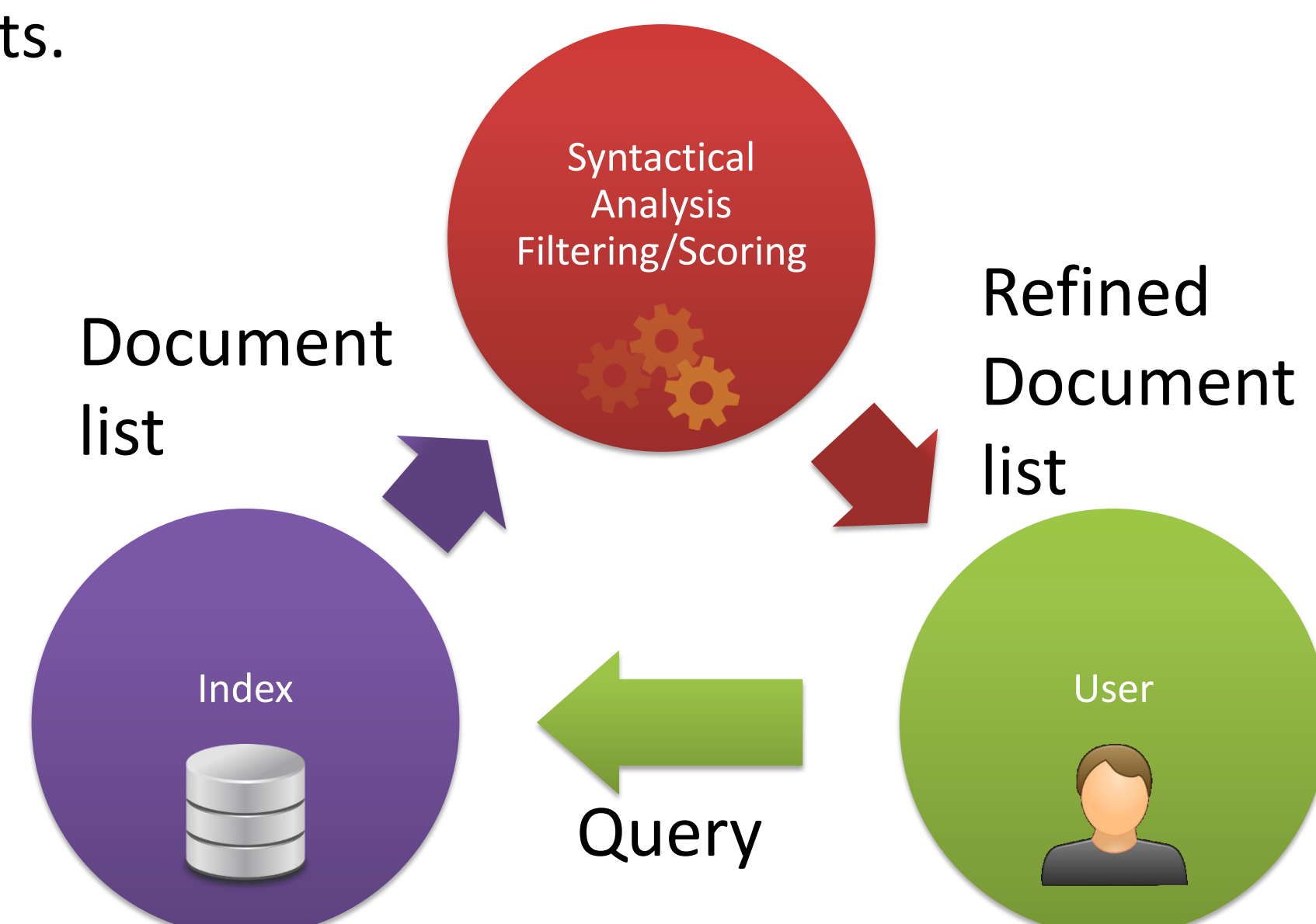
Goals

- Efficient access to archived regulatory decisions
- Help resolve current drug review issues
- Pinpoint numerical/categorical data
- Effortless and user-friendly experience

Methodology

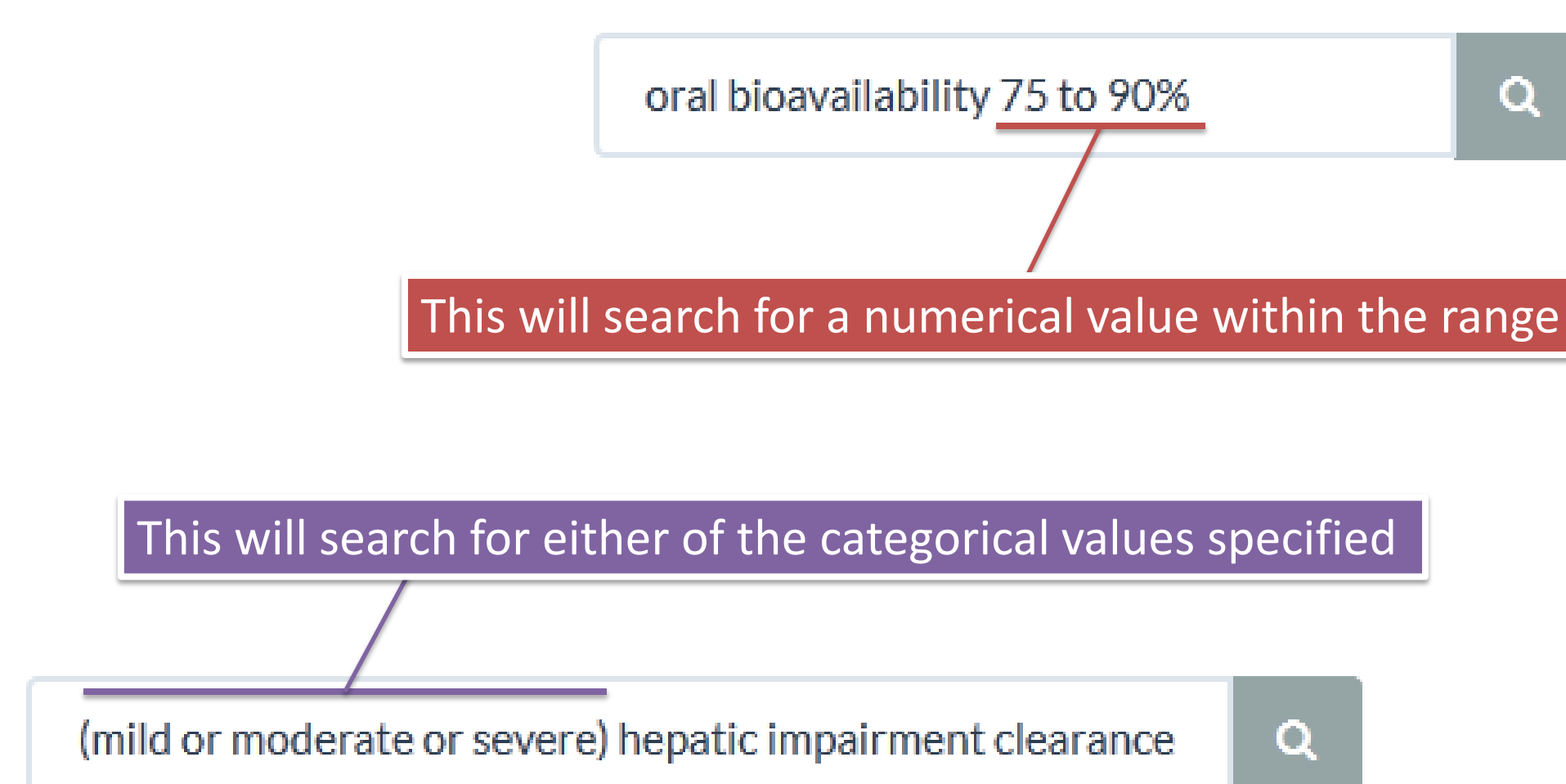
Search Method

Our search database consists of an inverted index of the stored documents, in which our query processor performs a first pass identifying text sections with keyword based matches that fall under a word proximity threshold. Then the system analyzes the syntactical relationships among the matched words to determine a relevancy score and presents a list of results.



Query Parameters

The proposed system allows the specification of numerical ranges and categorical variables in the query, with an approach that avoids the use of complicated syntax or cluttered user interfaces.



Multi-Query Search

The tool also provides the option to specify multiple search queries to be matched in different sections of the document. This provides the ability to better describe the case we are trying to locate.

×

metabolite increases adverse events

×

moderate renal impairment increase (auc or clearance)

×

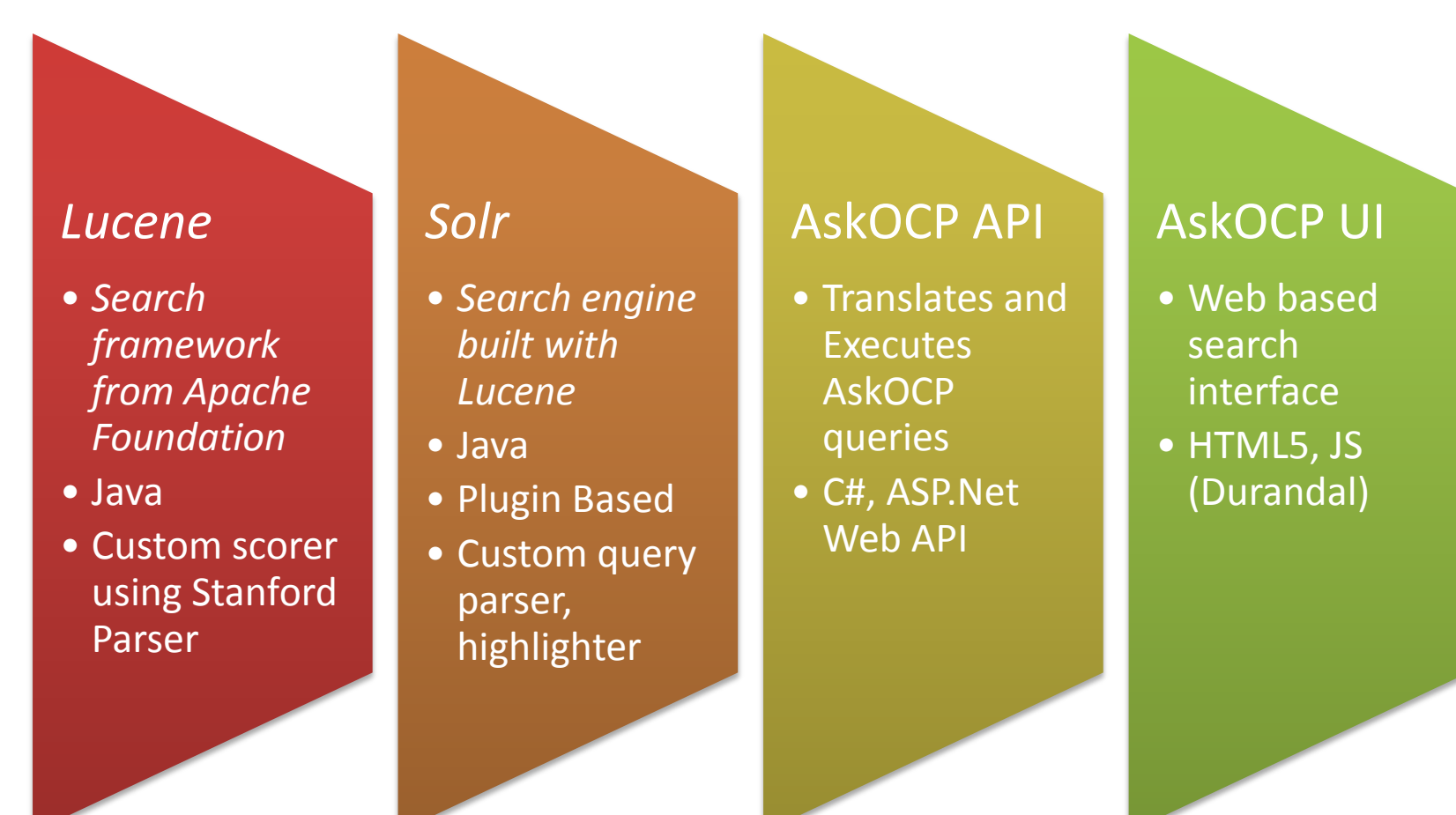
anticonvulsant

Add Query

System Overview

The System has been implemented as a series of plug-ins and modifications to the Apache Lucene-based Apache Solr search engine:

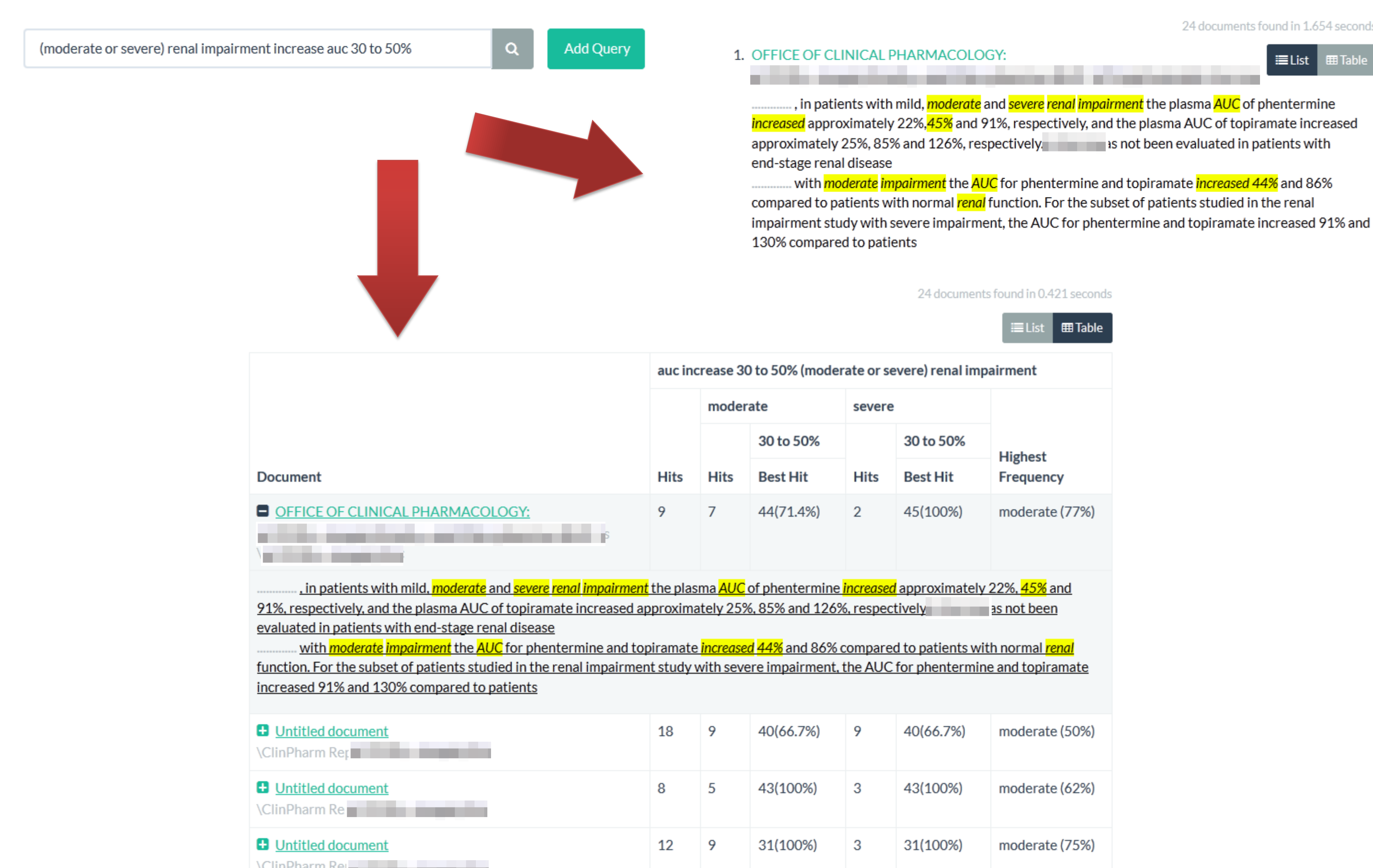
- The query parser converts the original query, including numerical and categorical parameters, to a format that Solr can understand.
- The document scorer is in charge of validating and re-scoring each of the potential document matches based on syntactical proximity. It uses the Stanford Parser from the Stanford NLP Group for syntactical analysis.
- The highlighter component has been implemented to provide meaningful and accurate fragments of highlighted findings for each document found.



The system also performs some word-wise preprocessing and postprocessing operations that allow the identification of plurals, derivative words and synonyms.

User Interface Workflow

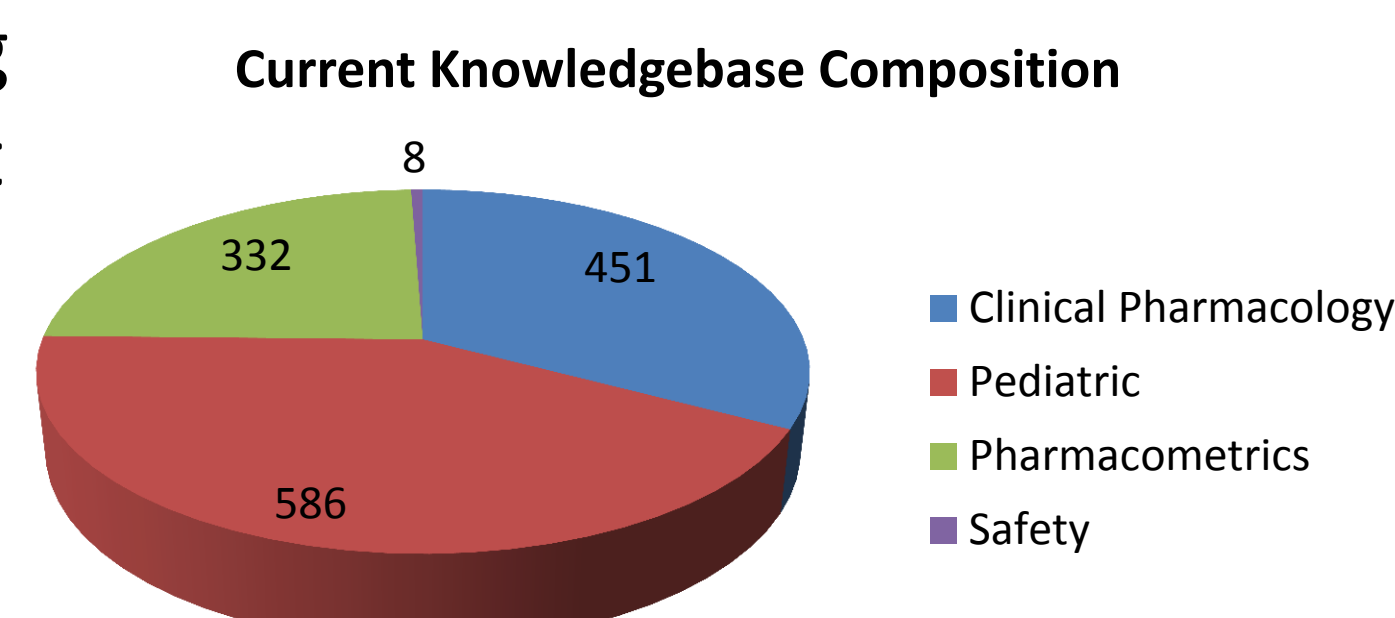
The user interface is designed to be simple and intuitive. It can present the search results in a tabulated format, summarizing numerical and text information in statistical metrics, which has proven valuable for regulatory research projects.



Conclusions

The current knowledgebase for AskOCP includes ~1500 documents from Clinical Pharmacology, Pharmacometrics, Pediatrics, and Drug Safety reviews. The tool has been released as a beta testing version to a small set of users to assess its utility and room for improvement. So far it has demonstrated its usefulness in a number of ongoing NDAs where the possibility to research archived reviews for similar cases provided valuable data.

The approach used in this tool can be applied to other similar text document databases. The architecture of the tool potentially allows it to be implemented on top of existing Lucene document databases and similar systems.



Contact

Eduard Porta Martin Moreno
FDA/CDER/OTS/OC
Eduard.PortaMartin-Moreno@fda.hhs.gov

Peter Lee
FDA/CDER/OTS/OC
Ider.Lee@fda.hhs.gov

Acknowledgments

This project was supported in part by an appointment to the Research Participation Program at the Center for Drug Evaluation and Research administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the U.S. Department of Energy and the U.S. Food And Drug Administration.