

Drug Label Explorer



Spring 2022

Software Engineering Capstone, CSCI E-599 Section 2

Group Members

Ken Brown, Agi Kajanaku, Leo Landau, Sam Negassi, Ky Nguyen

Customer

David Edelen

Teaching Staff

Peter Henstock, Roman Burdakov

Table of Contents

Drug Label Explorer	1
Table of Contents	2
Drug Label Explorer: An Exploratory Tool Utilizing Information Extraction to Analyze FDA and EMA Drug Labels	3
ABSTRACT	3
INTRODUCTION	3
LITERATURE REVIEW	5
RELATED WORK	5
DATA SOURCES	5
RESULTS	6
DRUG LABEL EXPLORER	6
DISCUSSION	9
METHODS	9
NEXT STEPS	10
CONCLUSION	11
REFERENCES	12
2. System Design	15
2.1 Tech Stack	15
2.2 Tool Suite	15
2.3 System Modules	16
2.4 Architectural Diagrams	18
3. Testing Results	20
4. Development Process and Lessons Learned	21
4.1 Meeting the Requirements	21
4.2 Estimates	22
4.3 Risks	22
4.4 Team Dynamic	24
5. Appendix	25
5.1 Technical Requirements	25
5.2 Wireframe	47
5.3 Unit Test Code Coverage Report	48
5.4 Performance Test Results	50

Drug Label Explorer: An Exploratory Tool Utilizing Information Extraction to Analyze FDA and EMA Drug Labels

ABSTRACT

The primary purpose of a drug label is to provide relevant information to patients, healthcare providers, medical professionals, and regulatory agencies regarding the safe and effective use of a medication. In addition, information contained in drug labels is a valuable source of information for public health, medical, and pharmaceutical researchers. Currently, the Food and Drug Administration (FDA), DailyMed, and European Medicines Agency (EMA) websites are the primary public sources of drug label information. This paper discusses Drug Label Explorer (DLE), a web-based exploratory tool that can help healthcare and pharmaceutical professionals quickly find a specific drug label, or find specific information contained within drug labels. This tool includes a search functionality that allows users to narrow their searches by section, brand or generic name, manufacturer, or approving agency of drug labels. Our tool also provides the ability to compare section texts between drug labels and highlight changes between versions of a drug label. These functionalities help professionals who are drafting language for a new drug label and healthcare providers who are trying to decide which drug is the right choice for their patients. DLE is backed by a MariaDB database where the drug label data from the FDA and EMA public data sources is loaded into a single structured database. We extracted 46,005 FDA-approved and 1,284 EMA-approved drug labels, and in the case of FDA drug labels, we reduced from over 950 unique section titles to 83 section titles by grouping similar titles under a single generalized title. DLE makes these drug labels' data accessible to users through an intuitive UI built using the Django web framework and backed by an Apache web server. So that the information in our database stays useful, DLE periodically updates its data sources as new and updated information becomes available in the authoritative public sources listed above.

Keywords: Drug labels, FDA, EMA, information extraction, regulatory information

INTRODUCTION

In this paper, "drug labeling" is used as an umbrella term to encapsulate all the information in the structured drug labels of both the FDA and EMA approved labels. Drug labels contain critical information on prescription medications such as dosage and administration, indications, contraindications, adverse reactions, warnings, clinical pharmacology, and pharmacokinetics, among other drug-relevant information [17, 18, 19]. The main purpose of a drug label is to inform healthcare providers of relevant information on the safe dispensing and administering of medication while also avoiding drug-related medical errors and other serious adverse reactions. Additionally, this information can help people make more informed decisions, such as

if a patient is on several medications and needs to ensure that a particular combination is appropriate for their treatment [30]. Moreover, by comparing and analyzing information extracted from drug labels, new perspectives can be gained to avoid adverse reactions and undesired drug interactions, as well as to help drug classification and facilitate precision medicine, which is personalized medicine based on genetics and biomarkers.

The findings relating to drug interactions and their associated adverse events in drug applications are summarized in a section of the drug labeling. Specific inquiries, such as which HIV medicines are known to interact with methadone and which pharmaceuticals will interact with disulfiram, are then queried against the labeling data [31]. Drug labels are known to include a significant number of pharmacogenomic biomarkers. These biomarkers are anticipated to influence the efficacy and adverse effects of medications in patients from specific population subgroups [29]. This information helps find new trends and the frequency of genetic differences that are linked to more risks to public health [31].

As is evident, information on drug labels can be a great source or starting point for a plethora of research specialties, including life sciences, medicine, and pharmacology. A prevalent challenge is that most of the available drug label data sources utilize an explanatory research approach to how their users can navigate their tools. Explanatory tools refer to systems that aim to characterize the causes of a well-defined problem. In the case of DailyMed, the problem is in finding a specific drug. Exploratory tools refer to systems that aim to characterize unknowns in a respective field, for example, discovering products based on a description of an ideal product instead of a name. While there are many benefits to explanatory tools, users must know exactly what information is to be found. Exploratory tools have the benefit of discovery, where data can be extrapolated in a way that permits users to gain deeper insight into analysis.

Currently, users who want to utilize drug labeling to further their research interests must navigate tools like DailyMed, FDALabel, and EMEA databases. These are the authoritative, publicly available sources that make US and EU approved drug labels available. Although there are other public domain web-based platforms that make drug label data available, their use and searchability are limited, and in some cases, restricted to paying members. Therefore, researchers who hope to further their research by making use of drug label data have to rely on the above-mentioned public data sources. They have to download the drug label data, store it on a third party machine, such as their local machine, and either manually compare and contrast differences or implement general-use analytical tools. They also have to keep monitoring the public data sources for the publication of any new drug labels or updates to existing drug labels. In this paper, we discuss the features and functionalities of our Drug Label Explorer (DLE) tool, which we believe will help to fill the gap and address the needs of patients, healthcare providers, pharmaceutical professionals, regulatory agents, as well as public health and medical researchers.

LITERATURE REVIEW

RELATED WORK

FDALabel [16] and DailyMed [17] contain drug label data for the US, while the EMA [18] contains drug label data for the EU. These three data sources are the main places where you can find labels for approved prescription drugs. These web-based platforms include basic search capabilities and the ability to extract available drug labels. The FDA maintains a database of pharmaceutical Safety-related Labeling Changes (SrLC) that enables users to search for drug label changes, but it lacks the key capability of searching for specific text inside a drug label document. Also, the information is only available in XML or PDF format, and there is no way to look for specific sections or compare labels.

Moreover, the user interfaces are not intuitive. We also looked at RxList [20], ReedTech [21], WizMed [22], the Cerner Website [23], and Drugs.com [24]. These commercially available websites host searchable databases of medical label information that differ in terms of content, search options, and accessibility. Some of these commercial databases are subscription-based [21, 22], while others are ad-supported [20, 24], and still others are restricted to healthcare enterprise use only [23]. The majority of these commercially available websites assert that the FDA provided the data for their databases, although others claim to have data from the EMA and other nations, with one claiming to have data from nine countries [23]. While we were unable to verify the functionality of the fee-based and private sites, the ad-supported sites' capacity to analyze the content of drug labels fell short of enabling us to quickly locate the drug labels that matched our query. RxList.com and Drugs.com, for example, allow users to search for the text of medicine labels. However, they do not support "exact match" searches, and it is occasionally unclear why the labels are included in the search results. Also, these services don't make it easy to compare the information on two drug labels to find text that is similar.

DATA SOURCES

FOOD AND DRUG ADMINISTRATION (DAILYMED)

The Food and Drug Administration (FDA) is a centralized US government agency charged with guaranteeing the safety, efficacy, and security of human pharmaceuticals, biological products, and medical devices [19]. DailyMed is the authorized source for FDA label information, which is frequently referred to as package inserts. It is a freely accessible medication labeling database resource that the National Library of Medicine (NLM) submits to the FDA and contains the most recent versions of drug labeling submitted to the FDA. These labels are defined in the Health Level Seven (HL7) Structured Product Labeling (SPL) standard, which defines the different sections of a medicine label [28]. It utilizes Logical Observation Identifiers Names and Codes (LOINC) to connect the various components and subsections of human prescription medicine and biological product labeling. The FDA

requires that medicine labels be "informative and accurate, without being promotional in tone or deceptive in content." About 51,000 human prescription drugs and biological products that have been approved in XML/PDF format can be used in the United States [25], and that number is quickly growing by a few hundred each year.

EUROPEAN MEDICINES AGENCY

The European Medicines Agency (EMA) is the agency in Europe that is responsible for the scientific evaluation, inspection, and safety monitoring of pharmaceuticals in the European Union, Iceland, Norway, and Liechtenstein [18]. The EMA provides information about medications available to the public. Their data includes information about the early stages of a drug's development, its first evaluation, its post-approval status (withdrawals, updates, and safety reviews), and its post-approval status (withdrawals, updates, and safety reviews). The EMA drug labels differ from the FDA in that they are PDF only and also have a formatted structure.

RESULTS

DRUG LABEL EXPLORER

Drug Label Explorer (DLE) is a tool that provides a robust search functionality that supports a wide variety of queries, including data filtering and aggregation using several different attributes. This allows the user to identify and compare desired labels across FDA and EMA agencies—which no publicly available solution has previously provided. In addition to this, DLE periodically updates itself with newly published drug label data for both the FDA and EMA. DLE users can view all versions of a drug label and note the respective changes. This is helpful given that approximately 450 drug label updates are published by the FDA every week [4]. These updates can originate from a variety of sources, such as spontaneous reports (52%), clinical trials (16%), and pharmacokinetic studies (11%) [3].

Users who want to create their own labels based on the structure and language of other drug labels can upload custom drug labels to their accounts and access their labels privately. This feature is especially useful to drug manufacturers to optimize unpublished or new versions of drug labels by comparing their labels with the rest of the database available on our website. To mitigate tedious processes, users can save queries and compare desired search results on the DLE platform without having to outsource the storing of a given label to a different system. From there, users can choose to compare labels from their saved searches or from search results as needed.

DLE accessed the previously described data sources and extracted 46,005 drug labels from the FDA, of which 9,202 were not yet approved and 36,803 were active labels. DLE was able to extract data from 1,284 drug labels from the EMA database. We discovered that 6.7 percent of the sections in these EMA files failed to map data to all of the anticipated sections. To address this, the sections containing these mapping

errors were omitted while maintaining the remainder of the files for a larger corpus and ensuring the accuracy of the remaining features. After parsing the various data sources, it was determined that nine areas of the EMA's medicine labels comprised an average of 93 percent of the label sections. These frequently encountered sections include Indications, Posology, Contraindications, Warnings, Interactions, Pregnancy, Driving Effects, and Overdose. In the case of FDA sections, we iterated from over 950 section titles to 83 section titles by grouping more specific, similar groups under a generalized title per grouping. These 83 sections contain 96 percent of all the information on drug labels, with the remaining 4% being saved in an "other" section.

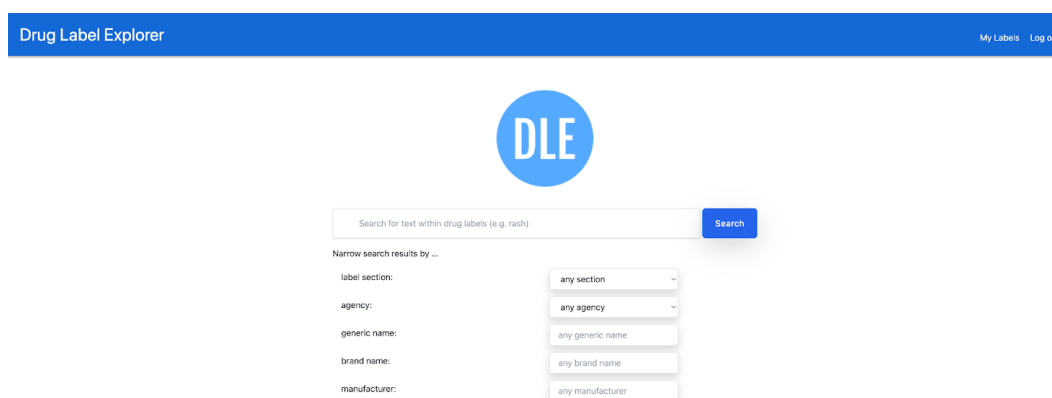


Figure 1.1: Drug Label Explorer Landing Page View: search can be specific text or limited to a given drug name or brand.

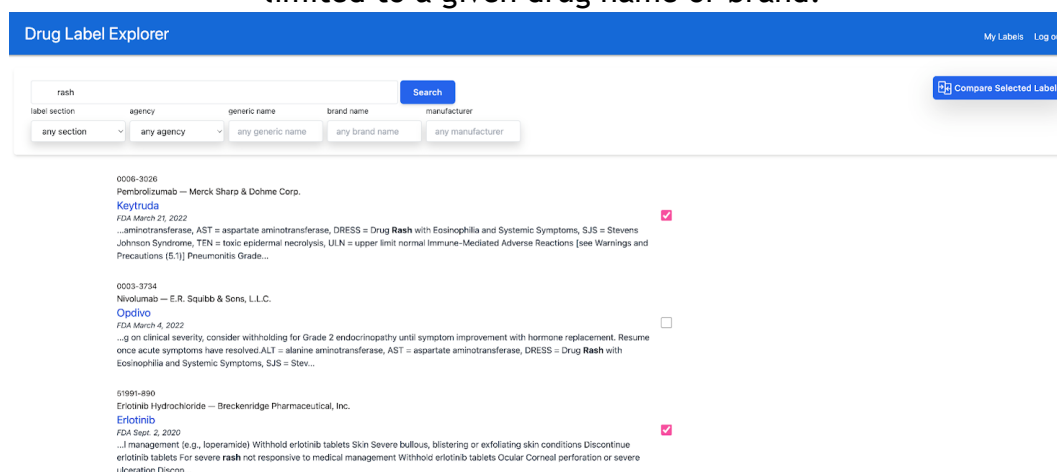


Figure 1.2: Drug Label Explorer Search Results View: shows an example of results for an exact search user query, where items can be selected for comparison and the compare button directs the user to the next step.

Drug Label Explorer

My Labels Log out

Comparing drugs Keytruda and Erlotinib:

Product Name: Keytruda Generic Name: Pembrolizumab Version Date: March 21, 2022 Product Number: 0006-3026 Marketer: Merck Sharp & Dohme Corp. Source: FDA	Product Name: Erlotinib Generic Name: Erlotinib Hydrochloride Version Date: Sept. 2, 2020 Product Number: 51991-890 Marketer: Breckenridge Pharmaceutical, Inc. Source: FDA
--	--

Select Section: [Adverse Reactions](#)

Hide All Sections

Adverse Reactions

The following clinically significant adverse reactions are described elsewhere in the labeling. Severe and fatal immune-mediated adverse reactions [see Warnings and Precautions (5.1)]. Infusion-related reactions [see Warnings and Precautions (5.2)]: Most common adverse reactions (reported in ≥20% of patients) were: KEYTRUDA as a single agent: fatigue, musculoskeletal pain, [rash](#), diarrhea, pyrexia, cough, decreased appetite, pruritus, dyspnea, constipation, pain, abdominal pain, nausea, and hypothyroidism. (6.1) KEYTRUDA in combination with chemotherapy: fatigue/asthenia, nausea, constipation, diarrhea, decreased appetite, [rash](#), vomiting, cough, dyspnea, pyrexia, alopecia, peripheral neuropathy, mucosal inflammation, stomatitis, headache, weight loss, abdominal pain, arthralgia, myalgia, and insomnia. (6.1) KEYTRUDA in combination with chemotherapy and bevacizumab: peripheral neuropathy, alopecia, anemia, fatigue/asthenia, nausea, neutropenia, diarrhea, hypertension, thrombocytopenia, constipation, arthralgia, vomiting, urinary tract infection, [rash](#), leukopenia, hypothyroidism, and decreased appetite. (6.1) KEYTRUDA in combination with axitinib: diarrhea, fatigue/asthenia, hypertension, hepatotoxicity, hypothyroidism, decreased appetite, palmar-plantar erythrodysesthesia, nausea, stomatitis/mucosal inflammation, dysphonia, [rash](#), cough, and constipation. (6.1) KEYTRUDA in combination with lenvatinib: hypothyroidism, hypertension, fatigue, diarrhea, musculoskeletal disorders, nausea, decreased appetite, vomiting, stomatitis, weight loss, abdominal pain, urinary tract infection, proteinuria, constipation, headache, hemorrhagic events, palmar-plantar erythrodysesthesia, dysphonia, [rash](#), hepatotoxicity, and acute kidney injury. (6.1) To report SUSPECTED ADVERSE REACTIONS, contact Merck Sharp & Dohme Corp., a subsidiary of Merck & Co., Inc., at 1-877-888-4231 or FDA at 1-800-FDA-1088 or [www.fda.gov/medwatch](#). Because clinical trials are conducted under widely varying conditions, adverse reaction rates observed in the clinical trials of a drug cannot be directly compared to rates in the clinical trials of another drug and may not reflect the rates observed in practice. The data described in the WARNINGS AND PRECAUTIONS reflect exposure to KEYTRUDA as a single agent in 2799 patients in three randomized, open-label, active-controlled trials (KEYNOTE-002, KEYNOTE-006, and KEYNOTE-010), which enrolled 912 patients with melanoma and 682 patients with NSCLC, and one single-arm trial (KEYNOTE-001), which enrolled 655 patients with melanoma and 550 patients with NSCLC. In addition to the 2799 patients, certain subpopulations in the WARNINGS AND PRECAUTIONS describe adverse reactions observed with exposure to KEYTRUDA as a single agent in a non-randomized, open-label, multi-cohort trial (KEYNOTE-019), a non-randomized, open-label, single-cohort trial (KEYNOTE-055), and two randomized, open-label, multi-cohort trials (KEYNOTE-021 and KEYNOTE-024).

The following serious adverse reactions, which may include fatalities, are discussed in greater detail in other sections of the labeling: Interstitial Lung Disease (ILD) [see Warnings and Precautions (5.1)] Renal Failure [see Warnings and Precautions (5.2)] Hepatotoxicity with or without Hepatic Impairment [see Warnings and Precautions (5.3)] Gastrointestinal Perforation [see Warnings and Precautions (5.4)] Bullous and Exfoliative Skin Disorders [see Warnings and Precautions (5.5)] Cerebrovascular Accident [see Warnings and Precautions (5.6)] Microangiopathic Hemolytic Anemia with Thrombocytopenia [see Warnings and Precautions (5.7)] Ocular Disorders [see Warnings and Precautions (5.8)] Hemorrhage in Patients Taking Warfarin [see Warnings and Precautions (5.9)] The most common adverse reactions (≥ 20%) with erlotinib tablets from a pooled analysis in patients with NSCLC across all approved lines of therapy, with and without EGFR mutations, and in patients with pancreatic cancer were [rash](#), diarrhea, anorexia, fatigue, dyspnea, cough, nausea, and vomiting. (6.1) To report SUSPECTED ADVERSE REACTIONS, contact Breckenridge Pharmaceutical, Inc. at 1-800-367-3395 or FDA at 1-800-FDA-1088 or [www.fda.gov/medwatch](#). Because clinical trials are conducted under widely varying conditions, adverse reaction rates observed in the clinical trials of a drug cannot be directly compared to rates in the clinical trials of another drug and may not reflect the rates observed in practice. Safety evaluation of erlotinib tablets is based on more than 1200 cancer patients who received erlotinib tablets as monotherapy, more than 300 patients who received erlotinib tablets 100 or 150 mg plus gemcitabine, and 1228 patients who received erlotinib tablets concurrently with other chemotherapies. The most common adverse reactions with erlotinib tablets are [rash](#) and diarrhea usually with onset during the first month of treatment. The incidences of [rash](#) and diarrhea from clinical studies of erlotinib tablets for the treatment of NSCLC and pancreatic cancer were 70% for [rash](#) and 42% for diarrhea. Non-Small Cell Lung Cancer First-Line Treatment of Patients with EGFR Mutations The most frequent (≥ 30%) adverse reactions in erlotinib tablets-treated patients were diarrhea, asthenia, [rash](#), cough, dyspnea, and decreased appetite. In erlotinib tablets-treated patients the median time to onset of [rash](#) was 15 days and the median time to onset of diarrhea was 32 days. The most frequent Grade 3-4 adverse reactions in erlotinib tablets-treated patients were [rash](#) and diarrhea. Dose interruptions or reductions due to adverse reactions occurred in 37% of erlotinib tablets-treated patients, and 14.3% of erlotinib tablets-treated patients discontinued therapy due to adverse reactions. In erlotinib tablets-treated patients, the most frequently reported adverse reactions leading to dose modification were [rash](#) (19%), diarrhea (19%), and asthenia (15.6%). Common adverse reactions in Study 1

Figure 1.3: Drug Label Explorer Comparison View: highlights the view for comparing two labels based on their section content, and users can also compare three search results in the same view.

Drug Label Explorer

My Labels Log out

Keytruda

Generic Name: [Pembrolizumab](#)
Version Date: [March 21, 2022](#)
Product Number (NDC Code): [0006-3026](#)
Marketing Authorization Holder (MAH): [Merck Sharp & Dohme Corp.](#)
Source: [FDA](#)
Go to this [link](#) for the original (FDA/EMA) source document.

List of all Drug Label versions with the same product name and marketer (select any two versions to compare)

☐ [Keytruda | 0006-3026 | FDA \(Dec. 28, 2018\)](#)
☐ [Keytruda | 0006-3026 | FDA \(March 21, 2022\)](#)

Compare Versions

Jump to a specific section: [All Sections](#)

Spl Product Data Elements

Go to the top

Recent Major Changes

Indications and Usage, Small Cell Lung Cancer – Accelerated Approval Indication Removed (1)	03/2021
Indications and Usage, Previously Treated Gastric Cancer – Accelerated Approval Indication Removed (1,9)	02/2022
Indications and Usage (1)	03/2022
Dosage and Administration (2)	03/2022
Warnings and Precautions (9)	07/2021

Go to the top

Indications And Usage

KEYTRUDA is a programmed death receptor-1 (PD-1) blocking antibody indicated:

Melanoma

- for the treatment of patients with unresectable or metastatic melanoma. (1,1)
- for the adjuvant treatment of adult and pediatric (12 years and older) patients with Stage III, IIC, or III melanoma following complete resection. (1,1)

Figure 1.4: Drug Label Explorer Selected Drug Label View: captures the readability of a single drug label by breaking it up into appropriate sections, which can be accessed directly or via comparison features.

DISCUSSION

METHODS

Drug Label Explorer is a website that acts as an interface to a database that contains FDA and EMA data. The workflow used to collect the drug labels from the FDA and EMA data sources can be broken down into four general steps. First, the data is downloaded, then it is parsed, then it is normalized, and finally, it is put into a database. The FDA data is in XML format, while the EMA data is in PDF format, in conjunction with the HTML data format from the website. The drug label data for both of these sources is semi-structured, making it challenging because there can be a sporadic nature to the structure. Given that the PDF data from the EMA can be difficult to parse, some of the data points for the EMA were extracted from the HTML files provided by their website.

The process of standardizing the data into the same structure allows for more efficient and optimal comparisons. So, after the process of downloading the files and parsing the raw data from each data source, the extracted data from different formats was mapped into a single data structure. Some limitations in data parsing include mapping. As we stated, 4% of FDA labels are put under another labeled section. This works for a majority of use cases, but in edge cases, we want to improve our ability to optimize identification of those sections.

One of the main areas we focused on for the data mapping was the section names. For the FDA, the section names do not always have the same title. So there could be multiple phrases for the same logical section, e.g. "Indications and Usage Section," "Indications," "Indications and Usage", "Indications and Usage Section" should all map to the same logical section name to allow for easy comparisons between different drug labels. Due to the fact that the formats of some of these FDA drug labels vary in terms of similar section titles with slight wording variations, DLE relied on domain experts to validate condensing available section names to standardized section names describing the same type of pertinent information about the particular drug. This was done to make it easier for users to navigate and to improve search results by putting information about drugs under a broader term.

For the search feature, MySQL Full-Text Search was used to match user-supplied queries via the InnoDB engine. By default, the DLE search uses the "Natural Language Mode" provided by InnoDB full-text search. The Natural-Language-Mode calculates a semantic score for each drug-label section-text based on the user provided query and returns the section-text in descending semantic score order. When a user provides a search query wrapped in double quotes, e.g. "Bone cancer", the DLE search processes the search in "Boolean Mode," which is also supplied by the Full-Text Search feature. In Boolean-Mode, only section text that has the exact search query contained within its text will be returned. There is no semantic score ordering when Boolean-Mode is used.

The individual drug label page displays the text of a drug label that has been entered into our database. Users can search for a specific drug and navigate to the individual drug label page from the search results page. The individual drug label pages include a link to the original document—i.e., a link to the drug label page on the DailyMed or EMA websites, depending on the agency that approved the drug. DLE provides two types of comparison views: *version comparison* and *label comparison*. The individual drug label pages list all older and current versions of a drug label. From this page, users can select any two versions of the label and launch a *version comparison* view to see how the label text has changed over time. The version comparison view lists the two versions' text aligned by section and highlights the text changes (the text *diff*) between the two versions. DLE uses a Python library named *diff-match-patch* to compare each section's text and highlight the differences between them. A drop-down filtering menu allows the user to filter the results to only display the sections with different text or the sections with matching text. Alternatively, a user can filter down to the comparison result of a specific section.

In addition to the version comparison discussed above, users have the ability to select two or three drug labels from the search results page and launch a *label comparison* view. The label comparison view displays the selected drug label's text aligned by section. This allows for comparing text across drug labels authored by different manufacturers and within a certain classification. Such a side-by-side display of drug label texts can help someone trying to draft a new drug label by providing a contrast of approved wording. It can also help a healthcare provider trying to prescribe a medicine decide between multiple prescription drug options. The label comparison view provides similar filtering features discussed above. However, unlike the version comparison view, the label comparison view does not highlight differences in the text. This is an intentional design decision because dissimilar drug labels are expected to have very different texts, and highlighting those differences will only lead to noisy outcomes.

NEXT STEPS

In next steps, we plan to improve comparison and search features by implementing the Bidirectional Encoder Representations from Transformers (BERT) model, which is a deep learning model that can improve section name mapping and label comparisons by improving the ability to utilize free text more closely aligned to natural language. Semantic search will be further improved by using the MedDRA ontology to account for biomedical terminology. DLE plans to expand to store drug label information for other agencies in different countries and in various languages in addition to English to further meet the needs of DLE users on a global scale. Finally, we plan to implement visualizations to iteratively improve the user experience. Data visualization is a widely known concept, and DLE would benefit from providing users with features that allow them to quickly identify patterns, make more insightful observations, and recognize potential trends.

CONCLUSION

In this paper, we developed software to extract, transform, and load (ETL) data dynamically from FDA/DailyMed and EMA. Our software also provides the ability for registered users to upload their private drug labels. DLE then searches for keywords or side effects through user input using search queries and compares labels by section title and section contents. Combining data from multiple agencies, allowing users to upload their own drug labels, allowing users to compare drug labels side by side, and having a section mapping process that allows diverse data to be queried and viewed in a standard way, Drug Label Explorer seeks to fill a gap in the analysis of drug label data. The software source code for this project is available at <https://github.com/DrugLabelExplorer/dle>.

REFERENCES

- [1] FDA: Code of Federal Regulations, Title 21, Vol.4, Chapter 1, Part 201-Labeling. Source: <https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfcfr/CFRSearch.cfm?CFRPart=201>
- [2] European Commission: A Guideline on Summary of Product Characteristics (SmPC), September 2009. Source: https://ec.europa.eu/health/system/files/2016-11/smpc_guideline_rev2_en_0.pdf
- [3] Lester, Jean, et al. (2013). Evaluation of FDA safety-related drug label changes in 2010. *Pharmacoepidemiology and Drug Safety*, vol. 22.3, p302-305. Source: <https://onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/pdfdirect/10.1002/pds.3395>
- [4] Fang, Hong, et al. (2016). FDA drug labeling: rich resources to facilitate precision medicine, drug safety, and regulatory science. *Drug Discovery Today*, vol 21.10, p1566-1570. Source: <https://www.sciencedirect-com.ezp-prod1.hul.harvard.edu/science/article/pii/S1359644616302240>
- [5] Moore, Thomas J., Sonal S., and Curt D. F. (2012). The FDA and new safety warnings. *Archives of Internal Medicine*, vol. 172.1, p78-80. Source: <https://jamanetwork-com.ezp-prod1.hul.harvard.edu/journals/jamainternalmedicine/fullarticle/1108624>
- [6] Dusetzina, Stacie B., et al. (2012). Impact of FDA drug risk communications on health care utilization and health behaviors: a systematic review. *Medical Care*, vol. 50.6, p466. Source: <https://oce-ovid-com.ezp-prod1.hul.harvard.edu/article/00005650-201206000-00002/HTML>
- [7] Seminerio, M. J., and M. J. Ratain. (2013). Are drug labels static or dynamic? *Clinical Pharmacology & Therapeutics*, vol 94.3, p302-304. Source: <https://ascpt-onlinelibrary-wiley-com.ezp-prod1.hul.harvard.edu/doi/full/10.1038/clpt.2013.109?sid=vendor%3Adatabase>
- [9] O. Nieminena, P. Kurkib, K. Nordstro. (2005). Differences in product information of biopharmaceuticals in the EU and the USA: implications for product development. *European Journal of Pharmaceutics and Biopharmaceutics*, vol. 60.3, p319-32 Source: <https://www.sciencedirect-com.ezp-prod1.hul.harvard.edu/science/article/pii/S0939641105000780>
- [10] Rodriguez, T., et al. (2021). Medical Error Reduction and Prevention. National Center for Biotechnology Information Source: <https://www.ncbi.nlm.nih.gov/books/NBK499956/>

[11] Tariq, R., et al. (2021). Medication Dispensing Errors And Prevention. National Center for Biotechnology Information Source:
<https://www.ncbi.nlm.nih.gov/books/NBK519065/>

[12] Delgado, N., etl al. (2019). Fast and accurate medication identification. npj Digital Medicine, vol. 2.10 Source:
<https://www.nature.com/articles/s41746-019-0086-0#Sec6>

[13] Jeetu, G., et al. (2010). Prescription Drug Labeling Medication Errors: A Big Deal for Pharmacists. Journal of Young Pharmacists, vol 2.1, p107-111 Source:
<https://www.sciencedirect.com/science/article/abs/pii/S097514831021021X>

[14] Davis, T. C., Federman, A. D., Bass, P. F., 3rd, Jackson, R. H., Middlebrooks, M., Parker, R. M., & Wolf, M. S. (2009). Improving Patient Understanding of Prescription Drug Label Instructions. Journal of General Internal Medicine, vol. 24.1, p57-62 Source: <https://link.springer.com/article/10.1007/s11606-008-0833-4>

[15] Shrank, W., Avorn, J., Rolon, C., & Shekelle, P. (2007). Effect of content and format of prescription drug labels on readability, understanding, and medication use: a systematic review. The Annals of pharmacotherapy, vol. 41.5, p783-801. Source: <https://journals-sagepub-com.ezp-prod1.hul.harvard.edu/doi/pdf/10.1345/aph.1H582>

[16] FDA Databases: Source (Orange Book):
<https://www.accessdata.fda.gov/scripts/cder/ob/index.cfm> and Source (Drugs@FDA):
<https://www.accessdata.fda.gov/scripts/cder/daf/index.cfm>

[17] NIH, DailyMed Database: Source:
<https://dailymed.nlm.nih.gov/dailymed/index.cfm>

[18] EMA, Medicines Database: Source:
<https://www.ema.europa.eu/en/medicines/what-we-publish-medicines-when-0>

[19] FDA, Drug Safety-related Labeling Changes (SrLC) Database: Source:
<https://www.accessdata.fda.gov/scripts/cder/safetylabelingchanges/>

[20] RxList Website: <https://www.rxlist.com> (a WebMD owned product)

[21] ReedTech Website: <https://www.reedtech.com>

[22] WizMed Website: <https://wizmed.com>

[23] Cerner Website: <https://www.cerner.com/solutions/drug-database> (an Oracle owned product)

[24] Drugs.com Website: <https://www.drugs.com>

[25] FDALabel: Full-Text Search of Drug Product Labeling: Source:

<https://www.fda.gov/science-research/bioinformatics-tools/fdalabel-full-text-search-drug-product-labeling#What%20is%20Included%20in%20Labeling>

[26] Krist Shingjergji, Remzi Celebi, Jan Scholtes, Michel Dumontier Relation extraction from DailyMed structured product labels by optimally combining crowd, experts and machines, Journal of Biomedical informatics, Vol. 112, Oct 2021 Source: <https://www.sciencedirect.com/science/article/pii/S1532046421002318>

[27] Chen M.J., Vijay V., Shi Q., Liu Z.C., Fang H., and Tong W.D. "FDA-Approved Drug Labeling For the Study of Drug-Induced Liver Injury." Drug Discovery Today, vol. 16, p697-703 Source: <https://www.sciencedirect.com/science/article/abs/pii/S1359644611001668?via%3Dihub>

[28] Shi, Y., Ren, P., et al. (2021). Information Extraction From FDA Drug Labeling to Enhance Product-Specific Guidance Assessment Using Natural Language Processing. Frontiers in Research Metrics and Analytics, vol 6. Source: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8222600/pdf/frma-06-670006.pdf>

[29] Hong, F., Harris, S., et al. (2016). FDA drug labeling: rich resources to facilitate precision medicine, drug safety, and regulatory science. Drug Discovery Today, vol 21. Source: <https://www.sciencedirect.com/science/article/pii/S1359644616302240?via%3Dihub>

[30] Lindquist, L., Lindquist, L., et al. (2014) Unnecessary Complexity of Home Medication Regimens among Seniors. Patient Education and Counseling, vol 96. Source: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4061206/>

[31] Wu, L., Liu, Z. (2019). Study of serious adverse drug reactions using FDA-approved drug labeling and MedDRA. BMC Bioinformatics 20(S2). Source: https://www.researchgate.net/publication/331748852_Study_of_serious_adverse_drug_reactions_using_FDA-approved_drug_labeling_and_MedDRA

[32] Fei, H., Ren, Y., Zhang, Y., Ji, D., Liang, X. (2021). Enriching contextualized language model from knowledge graph for biomedical information extraction. Briefing in Bioinformatics, vol 22. Source: <https://academic.oup.com/bib/article-abstract/22/3/bbaa110/5854405>

[33] MySQL. MySQL 8.0 Reference Manual. Source: <https://dev.mysql.com/doc/refman/8.0/en/>

[34] Stoeva, M. (2021). EVOLUTION OF WEBSITE LAYOUT TECHNIQUES. Source: https://www.researchgate.net/profile/Maya-Stoeva-2/publication/354675809_Evolution_of_Website_Layout_Techniques/links/61459d1c3c6cb3106977314d/Evolution-of-Website-Layout-Techniques.pdf

2. System Design

We chose to use the Django framework for this project, partially due to the popularity of Python. We are happy with this decision as it enabled a quick ramp-up time for any developers that were not yet familiar with the technology. We utilized a modular setup of "apps" in the Django project: data, search, compare, users. The data schema and ETL processes are in the data module. The search module contains the code for processing the user search requests, the search views, as well as the load testing and performance testing scripts. The compare module contains the code for the comparison pages. The user module contains the code for the MyLabels feature as well as the user login features. Overall, this choice of a modular approach worked out very well for the team, allowing concurrent development and a seamless integration of features.

2.1 Tech Stack

The project essentially uses a LAMP (Linux, Apache, MariaDB, and Python) technology stack, which ultimately revolves around a Python application being deployed on AWS. The backend of the application is written in Python utilizing the Django framework, with the front-end served via Django templates, effectively reducing most of the application logic within one single framework. The web server used for serving the web requests is Apache, which was ultimately chosen because one of the project members has many years of experience with the library, though many other alternative web servers can be a drop-in replacement. The Python application is served via the `mod_wsgi` Apache plugin. Lastly, the database the project team decided on is MariaDB. Due to the availability of its ColumnStore engine for fast analytics across large datasets as well as its familiar SQL syntax, MariaDB was chosen. Ultimately, the team did not use the ColumnStore engine; after testing its suitability for the project, it was decided to use the default InnoDB engine in MariaDB instead.

2.2 Tool Suite

The team is using Github as a primary tool suite for the project. Github Projects was chosen as the primary planning software. This ultimately leads us to use Github as the

hosted git Version Control System. Following this trend, the project uses Github Actions to orchestrate its CI pipeline. And Github Issues are used to track any bugs and action items that arise.

For testing, we are using the Django test harness which is an extension of Python's unittest module. The unit tests are executed using Github Actions on every pull request and on every merge into the main git branch. The results of the tests can be easily seen on Github with a green check mark indicating success and a red X indicating a test failure.

Utilizing all of Github's built-in tools reduces the amount of learning required with other existing 3rd party tools. Asynchronous communication is handled through Slack messages and email is used for coordinating meetings with Zoom conferencing for those who are not a part of the slack organization.

2.3 System Modules

The project requirements were broken up into modules to facilitate developers' working on different parts of the application at the same time in a remote environment, with team members in different time zones having different work schedules. The original technical requirements for this project, including estimates for when each feature was to be delivered, are included in the Appendix. For the requirements, they are categorized into what roughly equates to the base modules of the product.

- **Non-functional/DevOps:** We have a web application that is accessible on the internet that acts as a gateway to the features of the application. The web app supports encrypted traffic (https) and is easy for an admin to redeploy.
- **Data:** The website is backed by public data sources, which are cleaned, merged, and regularly updated. This information is stored in a database that makes it easy for the following features to make queries.
- **Users:** The website allows both anonymous and logged-in access, with some features only available to logged-in users.
- **MyLabels:** When logged in, users can upload their drug labels and have them be queryable by the system. These uploaded drug labels are parsed and inserted

into the Drug Label Explorer database. The user-uploaded drug labels are only accessible to the user who uploaded the drug label.

- **SearchForm:** One of the main features of the website is a search form that allows drug labels to be queried. Capabilities for searching include: by drug brand name; by manufacturer; by label section; by agency; and by generic drug name. This search form gives the user the ability to fine tune their results to get exactly what they need.
- **SearchResults:** The search results are displayed cleanly, and the search terms are highlighted in the results. From this page, the user is able to see blurbs from each search result with relevant keywords highlighted. From this page, the user will also be able to click into a result to get more information, or select multiple results to compare them side by side.
- **SingleLabelView:** The user can switch to a detailed view of a single drug label. From this view, the user will be able to see the entire content of the selected drug label.
- **CompareView:** The user can select two drug labels to view side by side, including separate drugs or different versions of the same drug. The view will clearly highlight areas that are similar and the differences.

2.4 Architectural Diagrams

Figure 2.1 Diagram showing the latest system architecture in AWS.

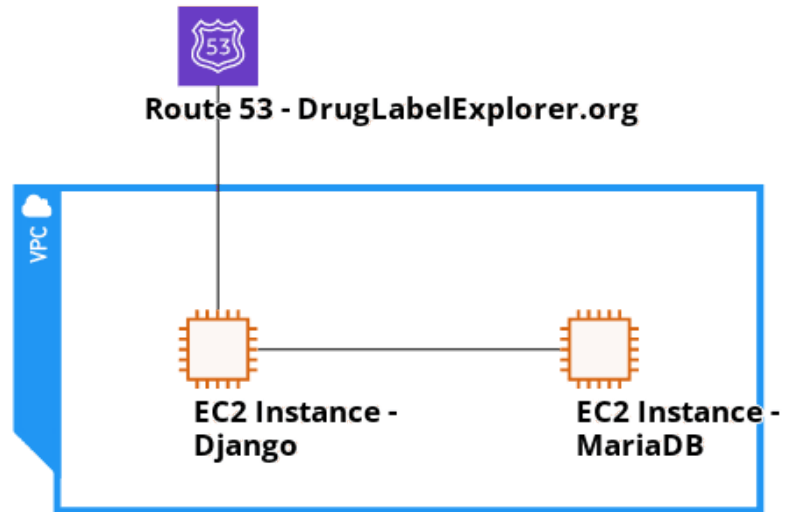
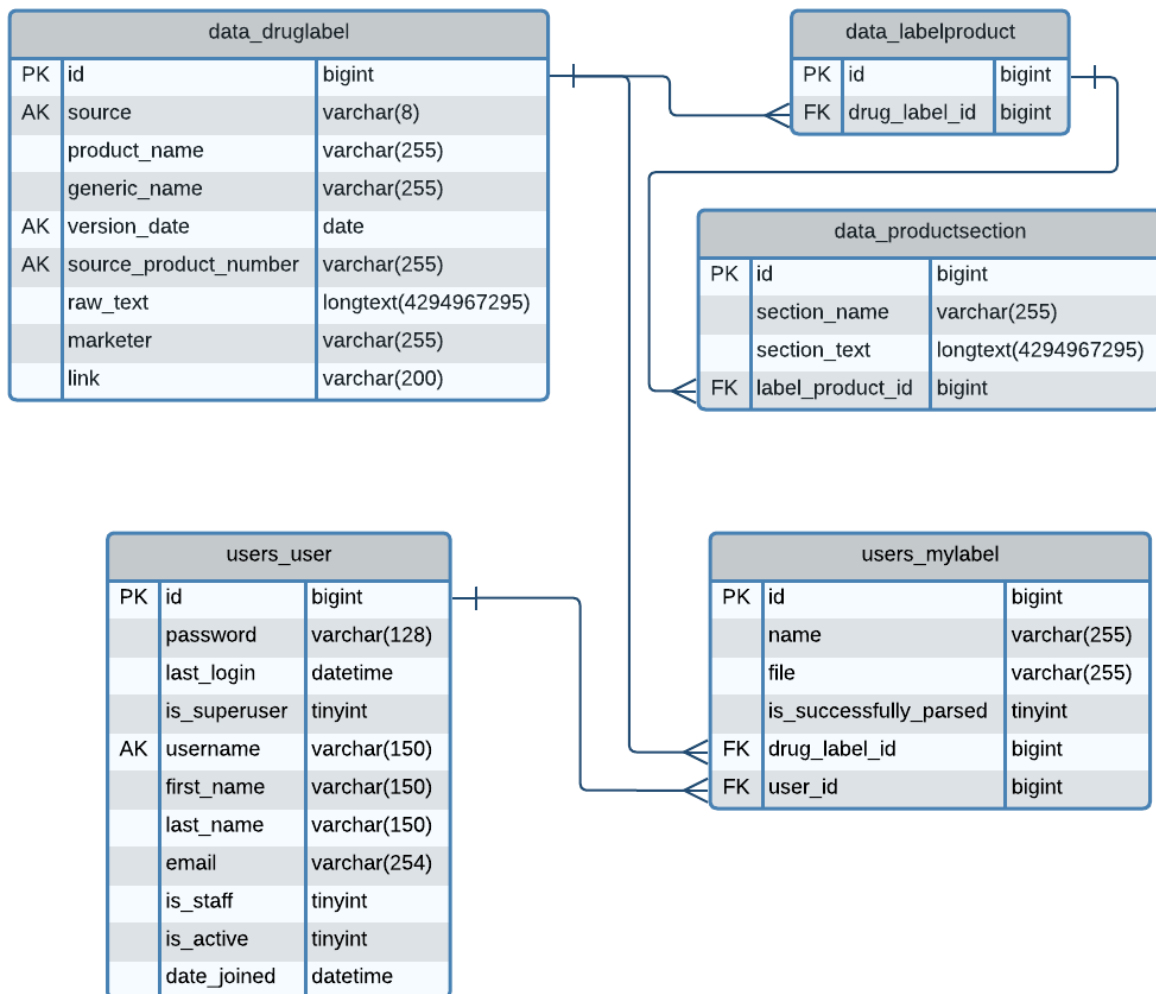


Figure 2.2 Diagram showing the latest class / database model for DrugLabelExplorer.



Above are the necessary components of the database for the Drug Label Explorer. An individual drug label will have one record in the data_druglabel table, and each subsection will have a record in the data_productsection table. When a user inserts a drug label, it is processed the same way but an additional record is inserted into the users_mylabel table to store relevant metadata. The fields marked PK are the primary keys, FK are foreign keys, and AK are alternate compound keys. AKs are set to ensure uniqueness.

3. Testing Results

For this project, our team conducted unit testing, performance testing, load testing, UI testing, and ad hoc testing. Unit testing is automatically performed on every pull request and on every code merge into the main git branch. Current unit test coverage is 63%. A report of the code coverage is included in the Appendix.

Performance testing was conducted periodically throughout development. The team was able to implement significant performance improvements, getting the average query time from 63 seconds per query to around 10 seconds per query. To facilitate performance testing, a script was developed that can run the performance tests automatically with the command `python manage.py performance_tests`. The test results are output to a CSV file and a png is created of the results. Some examples of the output are included in the Appendix.

There were some issues the team encountered with the database server in terms of handling the load of multiple queries at the same time. To ensure the reliability of the system, the team created a script that can perform load tests on the system. This is run with the command `python manage.py load_tests`. This helped the team ensure that the system could handle a reasonable load of queries concurrently.

Informal UI testing was performed with the customer at the team meetings. The team implemented many of the suggestions that occurred at these back and forth sessions. In this way, the team had an iterative approach to the User Interface implementation. The team provided the latest iteration of each feature to the customer as soon as it was ready. This allowed the customer to give the team valuable feedback on the progress of the work, even before the features were completed. Ad hoc testing was also conducted as a part of the development process. Developers would—in addition to the unit testing—perform ad hoc tests on the features under development. Developers would perform ad hoc testing on the features assigned to them, but would also test the features assigned to the other developers during each sprint. The team had the goal of having at least one other developer help test the code before a feature was merged into the main repository branch.

4. Development Process and Lessons Learned

4.1 Meeting the Requirements

The team was able to complete most of the original specified requirements. The requirements completed include parsing and loading of the data; being able to search the data; displaying the search results; being able to compare the drug labels; and being able to upload drug labels that are only available to the individual. Some of the features that we were not able to implement due to time constraints include: being able to search via MedDRA synonyms, being able to save search queries, and being able to export the search results. Through weekly meetings, the team and the customer worked together to change the order of the requirements.

Holding weekly meetings with the customer where we displayed the current (work in progress) status of the project was invaluable in gathering feedback from the customer. This allowed us to focus our efforts on the areas of the project that provided the most value to the customer. The team originally planned to meet weekly and have 1-week sprints. Partway through the project, we added a meeting, so we ended up having two meetings a week plus our customer meeting. Our team had to deal with some changes to our original plans as the project progressed. One of the issues was the performance of the database. The team originally planned to use a MariaDB database with the ColumnStore engine as the database for the project. This turned out not to be the best choice after the implementation failed to deliver the desired performance.

Another of the issues we encountered was that the loading of the data was more finicky than initially anticipated. We had in our model the concept of "sections" in the drug labels. But after working with the data and discussing it with the customer, there was another level in the hierarchy, "subsections." The team was able to make progress in handling the subsections in the drug label data by working closely with the customer to refine the project requirements as the project progressed.

4.2 Estimates

Our initial time estimates were somewhat inaccurate. In pretty much all cases, the amount of time spent working on a feature was about 2-3X that of the estimated time. This was mostly due to the time required to deal with "unforeseen" issues. So our initial estimates were somewhat of a "best case scenario" estimate. In the actual development process, there were bugs that were introduced that added time to debug and fix. Also, the amount of time iterating on a solution was not well captured by the estimates. For the features that we developed, they were improved over time, so in some cases, our estimates were "time estimates to deliver the first version" of a feature rather than "time estimates to deliver the final version" of a feature.

4.3 Risks

The Drug Label Explorer project enables users to quickly retrieve data on drug labels and their changes over time. This takes a substantial amount of data aggregation across many data stores and geographies in order to produce a dataset that is easily queryable. Two identifiable risks are presented when dealing with data ingestion in this scenario: data mining across varied data stores that change with the region, and modeling the mined data in such a way that querying the data becomes a trivial task.

Effective data mining across multiple regions will enable the project to procure data and provide an effective strategy for data transformation. Originally, the project had requirements that the data be extracted from many different countries' drug agencies. This presented a large risk because we would have to support parsing and translating an enumerable number of languages and label formats. The project team was able to de-risk this requirement significantly by scoping the project to two drug agencies, FDA and EMA. With only two agencies as data sources, the project team can mine data from each agency's datastore more effectively without having to program the mining tools to take into account different languages and label formats.

Once data is extracted from the drug agency datastore, it will need to enter a transformation process that will parse, clean, and store the data in a queryable format. The risk that comes with ETL on mined data from these datastores is that the parsing step will have to make tradeoffs between acquiring more data points or maintaining high accuracy within the extracted data. This risk is apparent when we

compare the FDA XML files against the EMA PDF files. Both of these agencies provide data that is loosely structured, and when parsing such data, the aforementioned risk is apparent. The project team was able to reduce the risk and complexity of this ETL step by reducing the scope of the features on the dataset when persisting the data. With the focus on high-accuracy and high-quality features on the dataset, parsing and storing the data is now more focused and easier to do.

In addition to the aforementioned risks, there are also the inherent risks associated with the limitations of web scraping tools available for working with PDFs. This risk is highlighted by the use of PDF templates that the implementation requires when expecting a PDF of a certain format to be parsed. Inevitably, there will be PDFs that don't align perfectly with the template and will reduce the accuracy of the parsed data. The project team has mitigated this risk by assuming that the accuracy level of parsed data may not be 100%, and any errors will be logged and categorized internally to further improve the system. The team planned to meet with the customer every week to talk about the actual results. This will help make sure the customer is happy with the project by the end of the time frame.

While we anticipated the risks in working with the PDF data from the EMA data source, in reality, parsing the XML data from the FDA data source turned out to be more challenging. It turned out that the FDA XML data is structured in a less standardized way than the EMA PDF data, which made parsing the FDA data more complex. One of the main risks that was unanticipated was the performance of the database. If we were not able to query the data in a reasonably timely manner, the whole project would be a bust. The team did a good amount of planning initially to come up with what we thought was a good solution for how to store and query the data, but testing showed that we needed to change course. As a result, we were able to deliver a project that has a reasonable query time. Unfortunately, this took more development cycles than anticipated. Another unanticipated risk is if a team member gets sick, takes a vacation, or is unable to work on the project for unforeseen personal reasons. This was not a major problem for our team. But there might have been one or more occasions where a developer's absence slowed down the project's progress a little bit.

4.4 Team Dynamic

The Drug Label Explorer Team has a shared goal of developing the best possible outcome to improve the solution for our client. To do so, we agreed to utilize open communication, which will be primarily conducted through Slack to account for differing time zones and personal work patterns, and we used Google Drive and GitHub for deliverables. We have set up a shared Google Drive and a GitHub Project to this end, and client meetings are recorded and shared throughout the project. All team members are encouraged to discuss issues and problems that may arise. DLE plans to use each member's unique skills to keep the project running smoothly and efficiently while still working together on all parts of it.

In terms of conflict resolution, DLE collectively prefers a minimal contact strategy, meaning issues will be raised between conflicting parties and only be brought to the entire group and/or the teaching staff if conflicts cannot be resolved or compromised internally. This strategy aligns with our shared value of open communication. Group decisions made by DLE will have a hybrid of majority rule and being guided by members with advanced expertise in a given topic (consensus decision-making).

DLE originally planned to meet twice per week: on Saturday (internally to discuss varying relevant agendas) and on Friday (with the customer to discuss relevant issues/agenda). We ended up adding a meeting on Tuesday to facilitate the project's progress. The team initially planned to have 9 weekly sprints, running from Thursday to Thursday, showing the latest progress with the customer on Friday. The goal is to have open communication with the customer. We sought and got useful feedback every week and talked about any problems as they came up to help the project succeed.

Having the weekly sprints did not really help us as a team. Instead, we integrated features into the main branch in an ad hoc manner whenever they were ready. The team initially decided that each team member would "self manage," meaning that we would create our own tickets and update our progress on the Github Project Workboard. This did not work out as well as originally hoped as some team members would forget to perform these management tasks. Not having a designated "project manager" might have also hurt the team a little in that there was no one to "keep team members accountable" if they were following through with their tasks in the

expected timeframe. Overall, this wasn't really an issue for our team. But if this was a larger project or if there were concrete deliverables that were more time-sensitive, this might have been a larger issue.

5. Appendix

5.1 Technical Requirements

For these requirements, Week 1 Starts on March 3rd, 2022, and the requirements are expected to be delivered on or before Milestone 2 (April 7th) or Milestone 3 (May 5th) as listed.

Deliverable 1:

There is a website is available on the public internet that allows people to run queries on drug labels

Category: Non-functional

Estimate (hours): 8

Start week: 1

Milestone delivery: 2

Deliverable 2:

Website is protected by industry standard TLS encryption

Category: Non-functional

Estimate (hours): 1

Start week: 1

Milestone delivery: 2

Deliverable 3:

Website can handle a small number of concurrent users - tens

Category: Non-functional

Estimate (hours): 0, because a small number of Users are handled automatically, and if we wanted to handle millions of users, that would take more time.

Start week: 1

Milestone delivery: 2

Deliverable 4:

Instructions provided to deploy / redeploy all System components on Amazon Web. AWS CloudFormation template(s) to assist with the deployment of System components Services (AWS)

Category: Non-functional

Estimate (hours): 5

Start week: 1

Milestone delivery: 3

Deliverable 5:

Database with security measures to protect the data stored in the database including encryption at rest and encryption in transit (via SSL)

Category: Non-functional

Estimate (hours): 3

Start week: 1

Milestone delivery: 2

Deliverable 6:

Database instance is configured to automatically run Snapshot backups daily, keeping a 30 day rolling window of backups

Category: Non-functional

Estimate (hours): 1

Start week: 6

Milestone delivery: 3

Deliverable 7:

Response times of the system are within reason

Category: Non-functional

Estimate (hours): 12

Start week: 6

Milestone delivery: 3

Deliverable 8:

System has access to FDA Drug Label data from DailyMed (SPL/XML), to include the latest versions for all approved prescription drug labels

Category: Data

Estimate (hours): 40

Start week: 1

Milestone delivery: 2

Deliverable 9:

System has access to FDA Drug Label data from DailyMed (SPL/XML), to include ALL historical versions for all prescription drug labels for the previous 3, 5, or 7 years (TBD)

Category: Data (History)

Estimate (hours): 20

Start week: 6

Milestone delivery: 3

Deliverable 10:

System automatically refreshes data from its data sources at specified cadence (daily, weekly, monthly)

Category: Data (Refresh)

Estimate (hours): 20

Start week: 6

Milestone delivery: 3

Deliverable 11:

System accesses data from EU data source (PDFs), to include the latest version for all prescription drug labels

Category: Data

Estimate (hours): 75

Start week: 3

Milestone delivery: 2

Deliverable 12:

Drug label data can be accessed using MedDRA terms

Category: Data

Estimate (hours): 35

Start week: 2

Milestone delivery: 3

Deliverable 13:

Data from all data sources is standardized using the Findable, Accessible, Interoperable, Reusable (FAIR) principles. This will be done with a uniform schema and search tools designed to directly interface with such.

Category: Data

Estimate (hours): 12

Start week: 2

Milestone delivery: 2

Deliverable 14:

Drug label search functionality is available to an unregistered / guest / null User

Category: Users

Estimate (hours): 0

Start week: 1

Milestone delivery: 2

Deliverable 15:

Basic user authentication including sign-up, sign-in, and password reset using email allows for additional features such as uploading labels, saving queries, etc.

Category: Users

Estimate (hours): 20

Start week: 1

Milestone delivery: 2

Deliverable 16:

Ability to upload labels conforming to a supported type: FDA/XML, EU/PDF; labels are only available to the single user (by default)

Category: MyLabels

Estimate (hours): 18

Start week: 4

Milestone delivery: 3

Deliverable 17:

Ability to share saved labels with other registered users; after selecting a label and choosing an email address, the system will send an email with a link to a page in the system that shows the label

Category: MyLabels

Estimate (hours): 5

Start week: 6

Milestone delivery: 3

Deliverable 18:

Sharing user-uploaded drug label, grants access to the registered user with the recipients email address

Category: MyLabels

Estimate (hours): 3

Start week: 6

Milestone delivery: 3

Deliverable 19:

User-uploaded drug labels show up in the user's search results along with other drug labels; only the user who uploaded the label or other users with whom the label was shared have access

Category: MyLabels

Estimate (hours): 4

Start week: 6

Milestone delivery: 3

Deliverable 20:

Ability to Save searches

Category: MyQueries

Estimate (hours): 6

Start week: 7

Milestone delivery: 3

Deliverable 21:

Main page of the application has a SearchForm area that includes the functionality for searching the Drug Labels. In general this can include drop-downs, checkboxes, text fields, etc.

Category: SearchForm

Estimate (hours): 35

Start week: 1

Milestone delivery: 2

Deliverable 22:

Ability to limit searches to FDA Drug Labels, EU Drug Labels or both

Category: SearchForm

Estimate (hours): 2

Start week: 1

Milestone delivery: 2

Deliverable 23:

Ability to Search by Product (Generic and/or Brand Name)

Category: SearchForm

Estimate (hours): 3

Start week: 1

Milestone delivery: 2

Deliverable 24:

Ability to Search by Application number, DEA schedule, NDC, UNI code, SET ID

Category: SearchForm

Estimate (hours): 3

Start week: 4

Milestone delivery: 2

Deliverable 25:

Ability to Search by Product Characteristics (color, imprint, shape, size, scoring, etc)

Category: SearchForm

Estimate (hours): 3

Start week: 4

Milestone delivery: 2

Deliverable 26:

Ability to Search by drug Marketer

Category: SearchForm

Estimate (hours): 3

Start week: 4

Milestone delivery: 2

Deliverable 27:

Ability to Search by Label Section

Category: SearchForm

Estimate (hours): 3

Start week: 4

Milestone delivery: 2

Deliverable 28:

Ability to Search by MedDRA terms

Category: SearchForm

Estimate (hours): 3

Start week: 4

Milestone delivery: 2

Deliverable 29:

Ability to perform wildcard search on drug label data when searching within drug label categories

Category: SearchForm

Estimate (hours): 4

Start week: 6

Milestone delivery: 3

Deliverable 30:

Ability to perform proximity search – a user should be allowed to search for drug label terms that are within a specified distance from each other (e.g. number of words apart, within the same paragraph, or within the same section)

Category: SearchForm

Estimate (hours): 15

Start week: 7

Milestone delivery: 3

Deliverable 31:

Ability to Filter Search Results - Pharmacologic class

Category: SearchForm

Estimate (hours): 1

Start week: 5

Milestone delivery: 2

Deliverable 32:

Ability to Filter Search Results - marketing categories

Category: SearchForm

Estimate (hours): 1

Start week: 5

Milestone delivery: 2

Deliverable 33:

User has some ability to adjust what data columns are displayed from the query results

Category: SearchForm-Results

Estimate (hours): 8

Start week: 3

Milestone delivery: 2

Deliverable 34:

Ability to Group Search Results by Generic Name

Category: SearchForm-Results

Estimate (hours): 2

Start week: 3

Milestone delivery: 2

Deliverable 35:

Ability to Group Search Results by Manufacturer

Category: SearchForm-Results

Estimate (hours): 1

Start week: 3

Milestone delivery: 2

Deliverable 36:

Ability to Group Search Results by Country

Category: SearchForm-Results

Estimate (hours): 1

Start week: 3

Milestone delivery: 2

Deliverable 37:

Ability to Group Search Results by Marketing Category (i.e. Application Type)

Category: SearchForm-Results

Estimate (hours): 1

Start week: 3

Milestone delivery: 2

Deliverable 38:

Ability to specify “latest version” or “all versions” for the drug labels in the search results. Drug label versions are derived from the date the document was last updated.

Category: SearchForm

Estimate (hours): 2

Start week: 2

Milestone delivery: 2

Deliverable 39:

Ability to have multiple search criteria. Ability to apply up to 5 search criteria with AND operators.

Category: SearchForm

Estimate (hours): 5

Start week: 2

Milestone delivery: 2

Deliverable 40:

After the search is executed, the search results are displayed to the user. The search results view should display a list of the matching drug labels. The search results may be paginated when they exceed a specified number of drug labels.

Category: SearchResults

Estimate (hours): 50

Start week: 2

Milestone delivery: 2

Deliverable 41:

The Search query parameters used in the search are highlighted in the SearchResults when present

Category: SearchResults

Estimate (hours): 3

Start week: 6

Milestone delivery: 3

Deliverable 42:

A details page for the drug label is shown after the user clicks on an item from the search results.

Category: SingleLabelView

Estimate (hours): 20

Start week: 4

Milestone delivery: 3

Deliverable 43:

The Search query parameters used in the search are highlighted in the SingleLabelView

Category: SingleLabelView

Estimate (hours): 3

Start week: 6

Milestone delivery: 3

Deliverable 44:

In the SearchResults there is the ability to select two labels. After selecting two labels, the user can then compare the labels.

Category: SearchResults - Compare

Estimate (hours): 4

Start week: 6

Milestone delivery: 3

Deliverable 45:

Side-by-Side Comparison with "Track Changes" View (Two labels). As a user scrolls through the page, both sides of the view should be in sync.

Category: Compare

Estimate (hours): 38

Start week: 3

Milestone delivery: 3

Deliverable 46:

Search results are automatically highlighted in the side by side comparison view of the drug labels

Category: Compare

Estimate (hours): 2

Start week: 5

Milestone delivery: 3

Deliverable 47:

Ability to navigate to the VersionHistoryView from the SearchResults view

Category: SearchResults - VersionHistory

Estimate (hours): 2

Start week: 4

Milestone delivery: 3

Deliverable 48:

Ability to navigate to the VersionHistoryView from the SingleLabelView

Category: SingleLabelView - VersionHistory

Estimate (hours): 2

Start week: 4

Milestone delivery: 3

Deliverable 49:

A Version History View page is displayed showing changes to a drug label over time

Category: VersionHistory

Estimate (hours): 32

Start week: 3

Milestone delivery: 3

Deliverable 50:

Search results automatically highlighted in the version history page

Category: VersionHistory

Estimate (hours): 2

Start week: 6

Milestone delivery: 3

Deliverable 51:

Ability to export selected columns from multiple labels from the SearchResults

Category: Export

Estimate (hours): 3

Start week: 8

Milestone delivery: 3

Deliverable 52:

Ability to export Label Comparison to HTML

Category: Export

Estimate (hours): 2

Start week: 8

Milestone delivery: 3

Deliverable 53:

Ability to export the Version History View to HTML

Category: Export

Estimate (hours): 2

Start week: 8

Milestone delivery: 3

Deliverable 54:

All export HTML pages include highlighting of the search parameters, when present

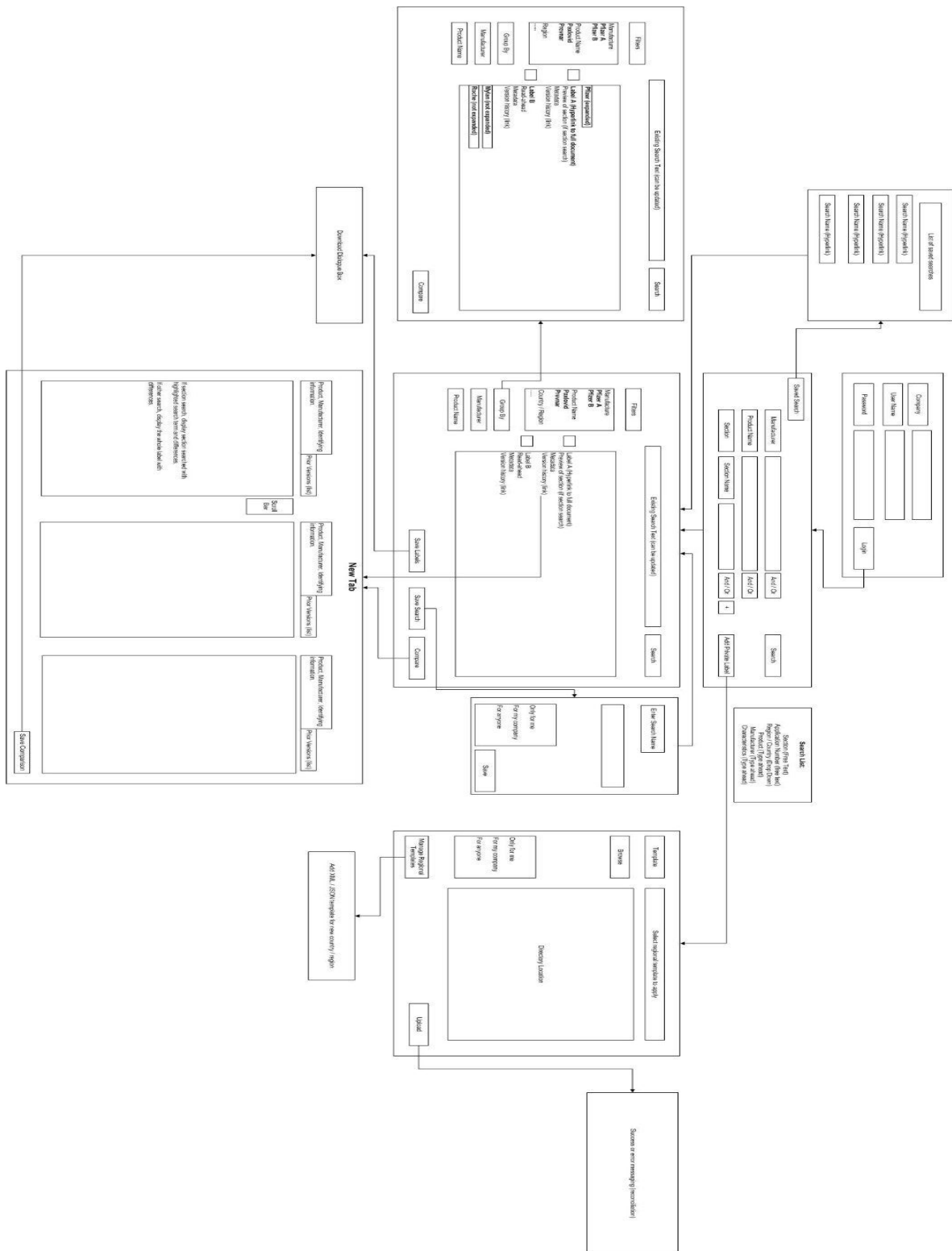
Category: Export

Estimate (hours): 1

Start week: 8

Milestone delivery: 3

5.2 Wireframe



Wireframes from our client, David Edelen

5.3 Unit Test Code Coverage Report

Name	Stmts	Miss	Cover		

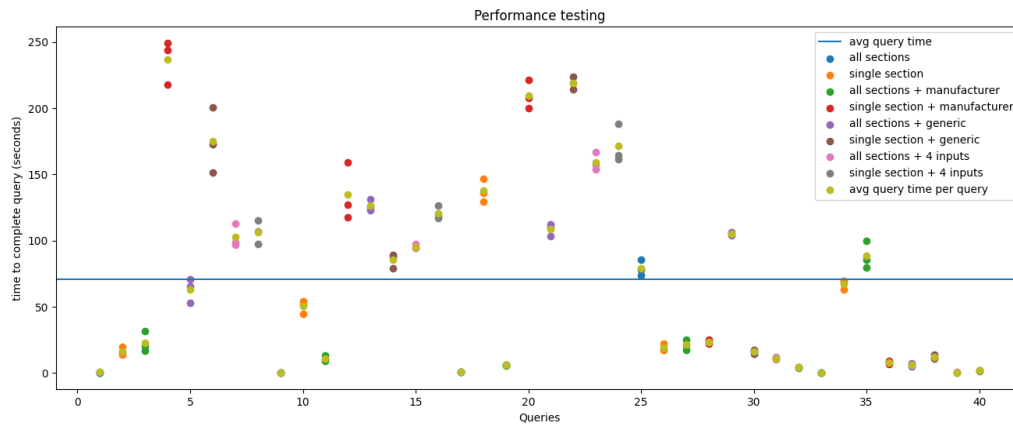
compare/apps.py	4	0	100%		
compare/models.py	2	0	100%		
compare/tests.py	1	0	100%		
compare/urls.py	4	0	100%		
compare/util.py	87	79	9%		
compare/views.py	130	120	8%		
data/apps.py	4	0	100%		
data/constants.py	2	0	100%		
data/management/commands/load_ema_data.py				185	44 76%
data/management/commands/load_fda_data.py				235	79 66%
data/management/commands/update_latest_drug_labels.py				27	2 93%
data/models.py	21	0	100%		
data/tests.py	62	0	100%		
data/urls.py	4	0	100%		
data/views.py	22	16	27%		
dle/settings.py	29	1	97%		
dle/urls.py	7	1	86%		
manage.py	12	2	83%		
search/apps.py	4	0	100%		
search/models.py	18	2	89%		
search/search_constants.py	2	0	100%		
search/services.py	84	36	57%		

search/tests.py	23	0	100%
search/urls.py	3	0	100%
search/views.py	25	16	36%
users/apps.py	4	0	100%
users/forms.py	10	0	100%
users/models.py	13	1	92%
users/tests.py	58	2	97%
users/urls.py	4	0	100%
users/views.py	80	35	56%

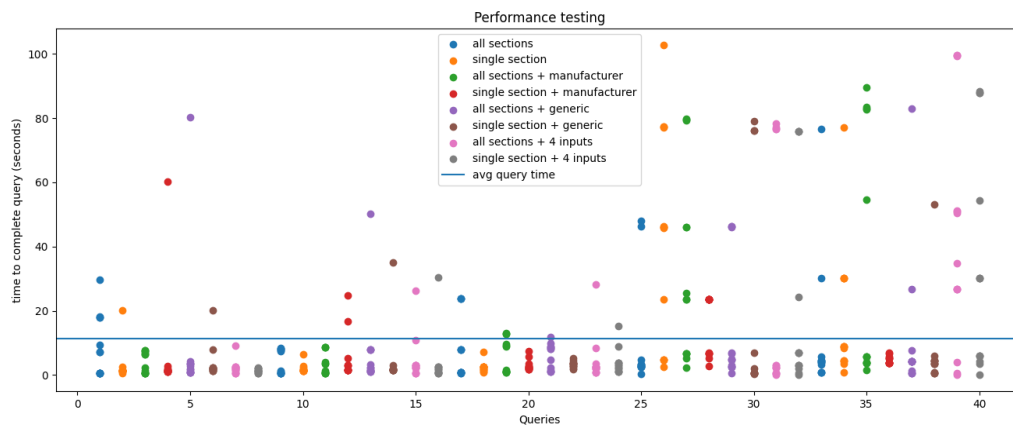
TOTAL	1190	436	63%

5.4 Performance Test Results

Initial performance testing indicated an average query time of 63 seconds per query.



Updated hardware in conjunction with query improvements led to an average query time of around 10 seconds per query.



After additional improvements the team was able to get this down to 2.7 seconds average query time per our performance test benchmark tool.