Linking Heterogeneous Health Data Sources in Brazil Centered on Drug Leaflet Processing

Márcia Jacobina Andrade Martins[1], **Claudia Bauzer Medeiros**[1]

[1]Institute of Computing – University of (UNICAMP) Campinas – SP – Brazil

m905106@dac.unicamp.br, cmbm@ic.unicamp.br

Abstract. *Health Information Systems often include a medication Recommendation module that helps doctors find medications based on symptoms. Most such modules rely on simple AI engines, fed by rules that correlate symptoms, diseases and medications. This, however, presents research and practical problems - e.g., some of the medications may no longer be commercially available, or their components may have been updated. Moreover, studies conducted to design such modules are based on corpora and databases in the English language. This hinders an adaptation to the Brazilian context, not only because of the language, but also due to the lack of authoritative integrated bases. To help solve these issues, we have designed a framework based on automatically extracting and linking information from all drug leaflets of approved medications in Brazil to feed recommendation systems. We processed and linked heterogeneous official data sources of the Ministry of Health, symptoms and diseases. The ongoing implementation, described here, created an ontology from the extracted data to enable linkage and identified quality problems in official data.*

1. Introduction

Health information systems involve a large set of modules, each of which connected to different data sources. One challenge in the Brazilian context is the intrinsic heterogeneity of official data sources, even within the Anvisa[1] system, and the absence of integrated sources – as opposed to other countries, e.g., USA or Canada. This challenge is aggravated by the need for identifying relevant authoritative sources, analyzing their documentation, and identifying curation issues.

This paper describes our ongoing work towards creating curated, linked data sources to provide a data infrastructure that, in turn, can be used to support the work of health professionals in Brazil. In particular, we are concerned with leveraging publicly available official sources, to complement systems that are based on Electronic Health Records. To this purpose, this work is centered on extracting information from drug leaflets, and connecting it to databases on diseases, active ingredients, medications, and others.

Our study is based on extracting data from 4 data sources: Medications from two ANVISA databases, including the portal[2] and the Datavisa system[3]; symptoms in Portuguese from BIREME MESH [4]; and *International Classification of Diseases (ICD-10)* codes. The final result is an ontology that can be used by distinct kinds of health information systems.

[1]The Brazilian Health Regulatory Agency – https://www.gov.br/anvisa

[2]https://consultas.anvisa.gov.br/

[3]https://dados.gov.br/dados/conjuntos-dados/medicamentos-registrados-no-brasil

[4]https://decs.bvsalud.org/en/

This presents a series of research and implementation challenges, as discussed in the paper. For instance, there are different versions of drug leaflets for the same medication, with distinct dates and/or target users (patients or professionals), with non-compatible data. Besides, some medications have no leaflet registered, and vice-versa, and some leaflets are unreadable pdf files. Also, it is hard to precisely identify the indications associated with a medication, since drug leaflets frequently do not use standardized vocabularies for diseases and symptoms. Additionally, each data source is updated with varying periodicity, which complicates validating a linkage strategy.

Our main contributions are: (i) identification of relevant open data sources; (ii) identification of quality issues, leading to data curation and deduplication; (iii) deciding which information to extract and link them to create a unified data platform. This platform can serve as input for health information applications, in particular recommendation systems. All data repositories and schemas are documented in Unicamp's data repository an open repository as a DOI-citable reference [Martins and Medeiros 2023].

2. Related Work

Related work involves algorithms and systems that link medications, symptoms and diseases to support health practitioners and health-related research, and creation of integrated datasets. Overall, this kind of linkage appears in papers that process drug leaflets or electronic health records (EHR), and is almost entirely based on corpora, ontologies and curated and integrated data sources in English. An exception is the work of [Silva 2016], which processes Brazilian leaflets available at the ANVISA site, extracting their main elements. Unlike our work, however, it does not link such elements to other official data sources nor does it construct any associated ontology or integrated databases.

Our work focuses on drug leaflets as we currently do not have access to EHR. Research involving linkage using EHR includes [Sohn and Liu 2014, Li and Xiao 2019]. Their prototypes are based on processing large amounts of EHR data, while research using leaflets concentrates on smaller datasets for obvious reasons (no limit to number of EHR records vs a limited number of available drugs).

Work on extracting information from drug leaflets concentrates on identifying key elements (e.g., compounds, indications), often with help of the *Structured Drug Labels (SPL)*[5] standard using Natural Language Processing (NLP). SPL is an XML-based standard adopted by the USA *Food and Drug Administration (FDA)* for managing products. The main result of this kind of research is the creation of structured files containing, for instance pairs <medication, indication> ([K. W. Fung and Demner-Fushman 2013]). Alternatively, such papers propose APIs to query an SPL-encoded drug leaflet base – e.g., [A. Flynn and Boisvert 2021]. Ontologies may also be created for drug indication – e.g., [S. J. Nelson and Tuttle 2021], in which a hierarchical ontological structure connects medications to indications.

Another trend is data consolidation into a public database that allows linking medications to indications, using a variety of sources, such as MedlinePlus [6], or NDF-RT[7],

[5]https://www.fda.gov/industry/fda-data-standards-advisory-board/structured-product-labeling-resources

[6]https://medlineplus.gov/

[7]https://nciterms.nci.nih.gov/ncitbrowser/pages/vocabulary.jsf?dictionary=NDFRTversion=February2018

the US National Drug File Terminology - e.g., [R. Khare and Lu 2014]. This exemplifies the variety of well structured health-related corpora in English, thereby facilitating data processing for health applications. Brazilian drug leaflets are only available in PDF, and need to be retrieved one-by-one via a query interface of the ANVISA system, complicating the construction of data linkage structures.

3. Architecture of the Framework

As mentioned in the Introduction, this work is based on four data sources (leaflets, medications, symptoms and diseases in ICD). Whereas the medication data and ICD can be downloaded in CSV, drug leaflets are in PDF and must be requested one by one. Thus, we had to develop a web scraping tool to download all drug leaflets into a single repository. Drug leaflets follow a standardized model established by ANVISA with predefined sections specific to medication description, helping to subsequently process the pdf files.

Figure 1 gives an overview of the framework architecture. Arrows labeled with ”a”represent actions to create repositories, while those labeled ”b”indicate the execution of processes necessary for developing the Medication ontology.

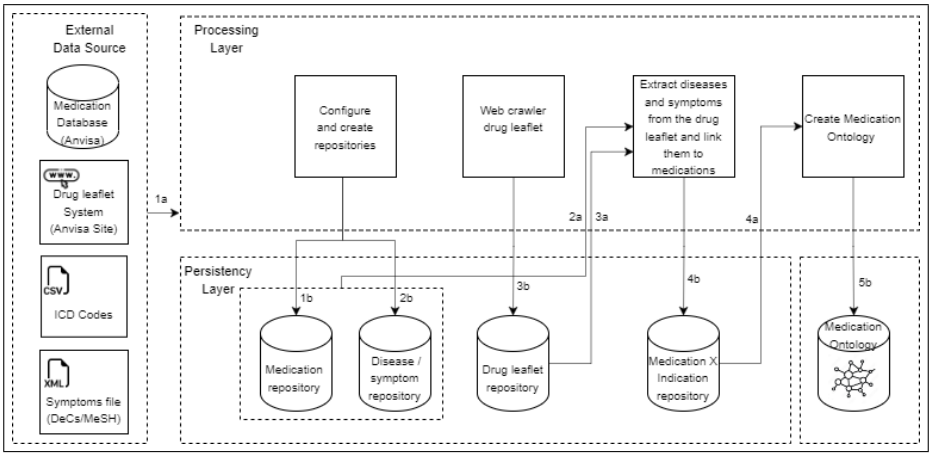

Figura 1. Architecture of the Framework.

Internal repositories are created at step ”Configure and create repositories”, which imports the Medication csv file (Anvisa) (1a) and stores it in our Medication repository (1b). The ICD Codes and the Symptoms files are imported to establish our Disease/Symptom repository (2b). The ”Web crawler”collects pdf drug leaflets from the Drug Leaflet System (Anvisa Site) (1a) and stores them in our Drug leaflet repository (3b).

Next, step ”Extract diseases and symptoms from the drug leaflet and link them to medications”leverages all three internal repositories. It identifies and extracts diseases and symptoms from the drug leaflets (3a) by comparing them with the data from the Disease/symptom repository (2b). It then links the extracted diseases and symptoms to the corresponding medications in the Medication repository (2a), storing the links in the Medication X Indication repository (4b).

Finally, data from the Medication X Indication repository (4a) is used to construct the Medication Ontology. The properties of the ontology reflect the relationships between medications, active ingredients. indications, diseases and symptoms.

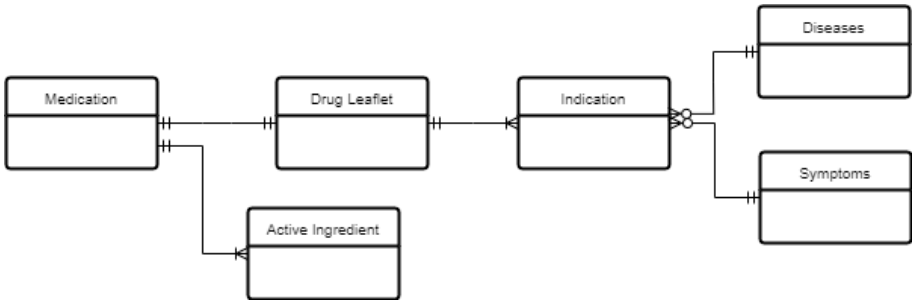

Figura 2. Simplified Data Model - main relations.

Figure 2 shows a simplified version of the data model, given space restrictions. It shows that Medications and Drug leaflets are interconnected, and that there are many Active Ingredients for a Medication. Indications are extracted from Leaflets and connected to Diseases and Symptoms – some indication texts mention diseases, other mention symptoms, or both.

4. Implementation

The bot and web scraping were developed in Python, repositories are sets of MySQL relations; and we use Protegé to manage the ontology. We first implemented a small prototype, using only 20 drug leaflets, processed manually, to investigate implementation alternatives and identify data quality and interoperability problems. We then proceeded to design the framework and implement all processes.

4.1. Configure and create repositories and curate data

The Datavisa medication database is a CSV file of 10,988 records (version Jan/2023), which we deduplicated to a total of 10,548 medications resulting in two relations in MySql: medication and active_ingredient.

The Disease repository was created from the ICD codes, available in csv file in the DATASUS site [8]. It contains 2,045 disease categories and 12,451 subcategories which we stored in two relations: disease_category and disease_subcategory. The Symptom repository, created from BIREME MESH (https://decs.bvsalud.org/por/), consists of 432 symptoms stored hierarchically in the symptoms_tree relation.

We downloaded 8,472 drug leaflets. Their deduplication eliminated older leaflet versions, resulting in 7,705 drug leaflets. Besides, thirteen leaflets were in doc format and were converted to pdf. Since there is a temporal gap between the medication database (Jan/2023) and Anvisa's leaflets (updated daily), there are medications in Datavisa without corresponding leaflets and vice-versa (e.g., medication *Admelog* has no leaflet). The final curated Leaflet Repository contains 7,476 drug leaflets, associated with Medications and Indications. This already points out to a problem in data quality – roughly 10% of the leaflets were discarded in this process.

Extract diseases and symptoms from leaflet and link to medications We used the SPACY NLP library to extract diseases and symptoms from the leaflets' section "Para que este medicamento é indicado"(For what is this medication indicated).

[8]http://www2.datasus.gov.br/cid10/V2008/descrcsv.htm

The Anvisa Drug Leaflet System offers two types of drug leaflets: for patients and for professionals. We compared both for several medications, discovering significant differences. For instance, the patient drug leaflet for AAS indicates its use for ”relief of mild to moderate pain, as well as symptomatic relief of pain and fever,”whereas the professional one highlights the “inhibition of platelet aggregation by blocking thromboxane A2 synthesis in platelets”. Since most symptoms in BIREME MESH use patient-related vocabulary, we only processed the patient leaflets.

4.2. Creation of the Medication ontology

The implemented ontology is still a prototype built manually in Protégé for validation. Ongoing work involves its automatic creation in a Turtle file to be exported to Protégé, for subsequent exploration in SPARQL.

Ontology classes are based on terms extracted from Medications and Drug Leaflets, which allow linking all data sources. Properties are based on the relationships between classes, e.g., a medication has multiple active ingredients and belongs to a specific category.

Many competency questions were considered to construct the ontology, through getting information from the linked sources, such as medications x symptoms, diseases x active ingredients, etc. This ontology has the following classes:

- *Medication* – commercial name, extracted from the drug leaflet.
- *Category* – classification of the registered medication type (new, similar, generic, biological, herbal, and others).
- *Active Ingredient* – pharmacologically active component in the medication
- *Company* – company holding the medication registration.
- *Indication* – disease or symptom that the medication treats.
- *Therapeutic Class* – classification of the medication based on the chemical function of the active ingredient or how it is used to treat a particular condition.

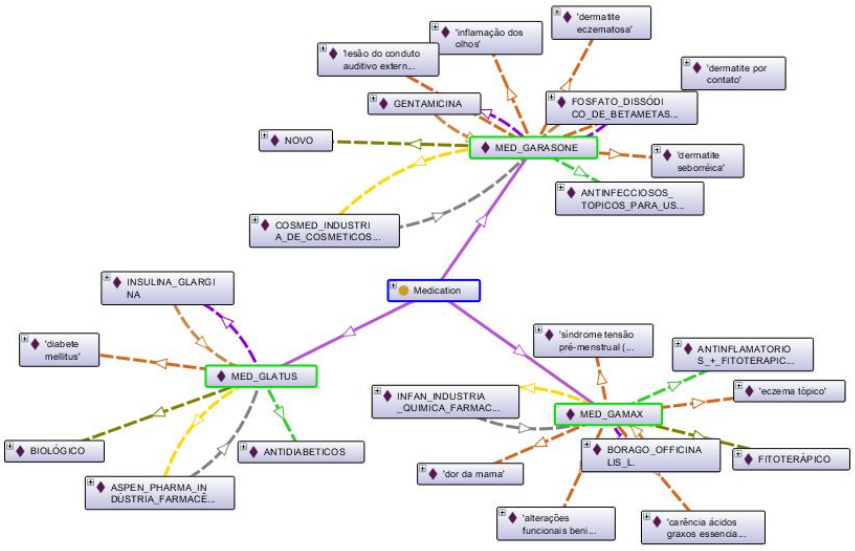

Figura 3. Medication Ontology - Relationships.

5. Conclusions and Ongoing Work

This paper presents the architecture and implementation of a framework to link heterogeneous medication-related Brazilian data sources centered on drug leaflets of officially approved drugs. Challenges include the identification of relevant data sources and vocabularies in Portuguese, definition of terms and content to extract, the extraction itself and the continuous update of interfaces and data structures within official sites. Another challenge is the intrinsic heterogeneity of the data sources, and of the interfaces provided to access them. Also, given the growing phenomenon of self-medication, we must consider the kinds of uses and accesses to the ontology. This is one of the reasons we limit ourselves to open data sources.

Ongoing work involves finalizing the NLP processing of leaflets and other data sources, and extending the ontology to symptoms and additional relevant information. The creation of the ontology itself, and its validation, is a major challenge – related work shows the need for expert validation, which is only feasible for small data volumes. We are also considering the concurrent use of graph databases.

Acknowledgements Work partially funded by projects FAPESP 2013/08293-7 and CNPq #308018/2021-4. We thank Dr. Nestor Andrade Piva for his help in checking our work on the drug leaflets.

Referências

A. Flynn, C. Huang, N. L. G. M. N. G. A. B. B. R. and Boisvert, P. (2021). An experiment to convert structured product labels into computable prescribing information. *2021 IEEE 9th International Conference on Healthcare Informatics (ICHI)*, pages 296–300.

K. W. Fung, C. S. J. and Demner-Fushman, D. (2013). Extracting drug indication information from structured product labels using natural language processing. *J Am Med Inform Assoc*, 20(3):482–488.

Li, Y. and Xiao, C. (2019). Developing a data-driven medication indication knowledge base using a large scale medical claims database). *AMIA Jt Summits Transl Sci Proc*, 2019:741–750.

Martins, M. J. A. and Medeiros, C. B. (2023). Medications, symptoms and drug leaflets extracted from public Brazilian sources. https://doi.org/10.25824/redu/JUHFWF, Repositório de Dados de Pesquisa da Unicamp, DRAFT VERSION.

R. Khare, J. L. and Lu, Z. (2014). Labeledin: cataloging labeled indications for human drugs. *J Biomed Inform*, 52:448–456.

S. J. Nelson, A. F. and Tuttle, M. S. (2021). A bottom-up approach to creating an ontology for medication indications). *Am Med Inform Assoc*, 28(4):753–758.

Silva, J. V. F. (2016). Facil Bula: Sistema que Estrutura o Bulario Eletronico da ANVISA.

Sohn, S. and Liu, H. (2014). Analysis of medication and indication occurrences in clinical notes. *AMIA Annu Symp Proc*, 2014:1046—-1055.