Índice

List of Figures

Chapter 1

Especificação 1

1.1 Considerações preliminares

Para o escopo desse trabalho foi definida como fonte de dados a serem trabalhados o repositorio Data Sharing/BR da FAPESP, onde são disponibilizados dados de pacientes, exames e desfechos relativos a exames dentro do contexto da Covid-19, esecificado como dados abertos no combate à pandemia.

A parte de garantir a corretude das implementações sobretudo das a-box para aplicação por exemplo em sub-domínios nas ontologias clinicas / hospitalares específicas localiza-se no diretório DataSharingFAPESP HAIS, onde pode ser verificado o script relativo ao processamento do Stage Mapping, no caso utilizando mySQL.

Basicamente foram definidos dois estágios no processo de criação de ontologia referente aos dados fonte e consequente processamento dos mesmos. No primeiro estágio, a título de obtenção das rotinas de processamento e visualização gráfica da organização e conteúdo da fonte, foram desenvolvidos, em Python, utilizando a biblioteca RDFLib e métodos auxiliares de processamento de texto e dados tabulares, procedimentos para transformar os dados em si em um modelo processável a partir de uma ontologia mais simples conforme pode ser observado no arquivo 'ontology pacientes.ttl'.

Em conjunto com esse procedimento foi também utilizada a biblioteca CoW (CSV on the Web), que produz a conversão automática das colunas de uma tabela, no caso por exemplo a tabela de exames, em um arquivo em JSON-LD para descrição das classes e propriedades e posterior conversão e criação das A-box sob a formatação do "componente de terminologia" (T-box). Para essa conformidade os dados no diretório /DataSharingFAPESP HAISS/* estão disponibilizado os scrits e código fonte, bem como arquivos explicativos referentes às partes necessárias para o correta execução dos procedimentos de ETL direcionada a aplicação clínica.

1.1.1 Análise Preliminar dos dados

Sendo assim os dados tabulares utilizados conforme disponibilizado em /dados/tabelas/* foram processados segundo alguns procedimentos visando diferentes aplicações. Para analisar o processamento da ontologia segundo a do domínio de terminologias utilizado, alguns recortes mais simples foram retirados, por exemplo para iniciar o processo das tuplas em exames ou realização de consultas e possíveis inferências segundo a abox determinada foram retirados cortes numéricos sem determinação de filtros.

Por outro lado, para garantir a corrutude técnica da abox processada segunda a ontologia definida junto ao grupo HAIS foram realizadas uma série de normalizações para gerar tabela conforme disponível no arquivo em /Data SharingFAPESP HAISS, e para garantir a consistência dos dados escolheu-se processar os exames relativos a UREIA.
Utilizando o software White Rabbit verificou-se inicialmente os padrões, possíveis redundância de dados, amplitude e variância dos resultados, durante um primeiro estágio que serviu para o desenvolvimento de uma ontologia mais geral conforme pode ser verificado na primeira parte do projeto.

Termos distintos normalizados p/ o campo CD UNIDADE, de ambos os hospitais:
mil/mm³ fL U/L g/dL eritroblastos/100 leucócitos mg/dL pg milhões/mm³ g/L
/campo
UI/mL mEq/L
Segs µg/dL
mg/L
ng/mL
pg/mL
ng/mL FEU

Tamanho máximo em bytes dos atributos da tabela exames: Max length id paciente 32 id atendimento 32 dt coleta 10 de origem 4 de exame 59 de analito 46 de resultado 44 cd unidade 28 de valor referencia 306

A tabela exames é a que contem maior variância no tamanho e despadronização de nomenclatura entre os atributos, no caso da tabela desfechos os valores giram em torno dos menores valores da tabela exames,

isso é, sem considerar o valor de referencia (centenas de bytes) e contém menor variância, ou seja, os valores são bastante semelhantes ao que encontramos nos campos id paciente e atendimento, dt coleta e origem, e cd unidade.

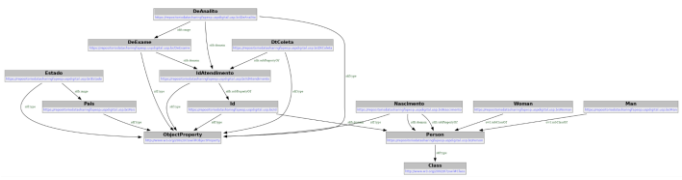

Figure 1.1: Visão mais ampla incluindo uma possível ontologia para integrar dados pessoais e georeferenciais

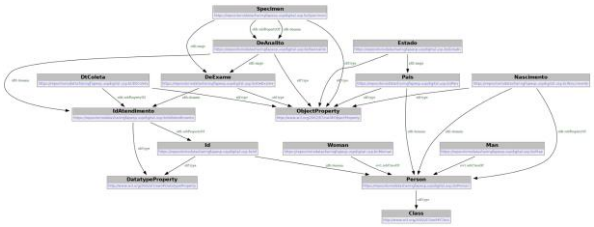

Figure 1.2: Visão especificando os dados relacionados a exames

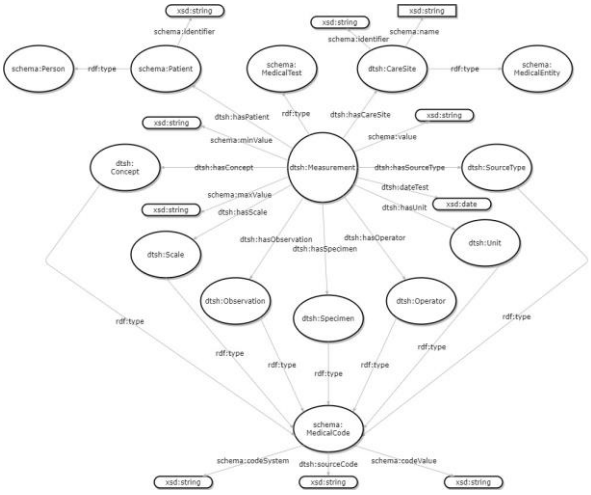

Figure 1.3: Visão da ontologia gerada c/ o grupo HAIS, no processamento conforme foi..

Chapter 2

Especificações 2 e 3

2.1 Criação da Ontologia

Foram definidos alguns processos definidos utilizando algumas ferramentas como a biblioteca RDFLib em Python, o software Open Refine e Inicialmente foram definidos alguns procedimentos utilizando as funções auxiliares em Python para processamento de texto para processar amostras das tuplas, a amostra em si é obtida utlizando-se parâmetros diagnosticados pela equipe especialista, no caso da resolução de mapeamento dos atributos dentro das tuplas tendo vista procedimentos e geração de formatos que venham a facilitar a inclusão dos dados tabulares em sistemas de inteligência p/ reconhecimento de padrões.

Foram geradas algumas amostras mais generalizadas (com relação as orientações de campo de conhecimento técnico entre as diferentes tabelas (por ex. bpsp exames 01.csv, bpsp pacientess.csv, etc.) com relação a leitura técnica especializada dos campos no caso georeferencial e social no caso por exemplo dos dados provenientes de registros de localidade e informação pessoal e dos campos de conhecimento/classificação hospitar, seja p/ os dados do tipo de atendimento e internação em si, da entidade ou hospital envolvida (CareSite), ou dos dados técnicos referentes aos exames em si, ou seja, tipo, medidas, referências, etc.

Conforme as definições iniciais para escopo da ontologia, ficou definido para o mapeamento dos dados que:
-O mapeamento automático das colunas utilizando, conforme script em /scripts/convert.py, ao gerar um arquivo de dados de descrição dos dados tabulares em JSON-LD possibilita um sequencial mapeamento das propriedades e classes, mudança de tipo de dados, definição de títulos e descrição, atribuição de sujeitos e predicados, conforme ficou definido em scripts/matadata format.py;
-Um mapeamento

-A criação de uma ontologia específica com análise técnica pertinente para os exames visando a inserção dos dados na formatação de ontologia clínica conforme entidades conhecidas (e a sub-sequente necessidade de 'tradução' entre modelos diferentes, no caso por exemplo das terminologias utilizadas em exames e analitos, bem como suas referências e resultado) ocorreu com o apoio dos modelos já bem definidos internacionalmente (SNOMED, UCUM, LOINC, SUS);

-Sob a perspectiva de ETL (do inglês Extract, Transform, Load), a possibilidade de realizar consultas realmente confiáveis para automação completa dos dados-alvo para o caso dos dados clínicos (tipo de exame e analito, perfil de atendimento e entidade, inclusão de classes a partir de concatenamento para formatação dos protocolos) dependeu de recortes para garantir a interoperabilidade.

2.1.1 Ontologia Clínica

Foi criado uma estrutra denominada Stage Mapping (estágio de mapeamento), onde o grupo HAIS trabalhou na disponibilização de mapeamentos específicos para cada um dos atributos mapeados dentro dos objetos, isso é, com a referência aos dados referentes a cada unidade trabalhada dentro das ontologias pré-selecionadas (LOINC, SNOMED, UCUM, etc.).

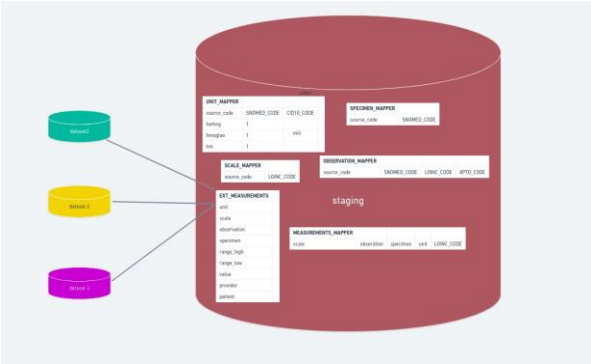

Figure 2.1: Visão geral do Stage Mapping

As classes encontram-se comentadas dentro do arquivo datasharingbr.ttl, <alterações>, conforme sua respectiva coluna nas tuplas dentro das tabelas:

-CareSite: unidade hospitar

-Measurement: id atendimento

-Observation:exame

-Source type:origem

-Specimen: analito

-Unit: unidade

Chapter 3

Especificações 4 e 5

3.1 Tabela DataSharing

A seguir a descrição sobre os parâmetros da abox:

-Para a correspondência das instituições relacionadas (Care Site) foi utilizada como sub-classe do domínio Medical Organization, schema (schema.org).

-O valor das medidas (referência mínima e máxima) e valor do exame foram processados a princípio como strings como propriedades equivalentes no mesmo domínio (schema:minValue, schema,maxValue, schema:Value).

Utilizando o mesmo domínio (schema) foi feito o mapeamento do id gerado como identificação do paciente;

-O processo de mapeamento gerou a partir do pré-processamento das tabelas, conforme pode ser visto no arquivo datasharingbr.ttl, os pares concept e vocabulary para identificar, respectivamente, o valor em si do atributo na ontologia especificada e o nome do domínio de vocabulário especifico. As correspodências podem ser verificadas na tabela measurements.csv;

- A inclusão das uris para as classes identificadas pelos pares para cada elemento mapeado foi feita utilizando a função hashlib, na ferramenta Open Refine e gerando ids únicos para cada par conceito e vocabulário mapeado, os ids são identificados na descrição das propriedades (hasScale, hasUnit, hasObservation, etc.).

Figure 3.1: P/ a amostra geral de exames (50056) centrada no conceito LP14288-2 - LOINC tem-se os specimens (analitos) conforme observado (2 analitos diferentes com o vocabulario SNOMED)

3.2 Processamento da amostragem

A construção da ontologia inicial para o projeto dependerá óbviamente de considerações como tipo de inteligência, limites e amplitudes da mesma, o que envolve a parte de filtragem técnica dos processos de leitura, construção de vocabulários, eliminação de redundâncias, detecção e correção de possíveis erros.

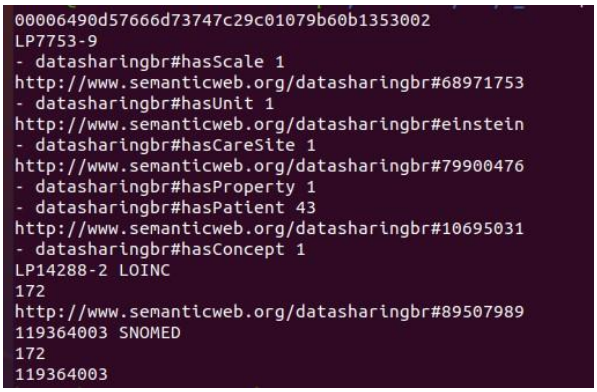

Figure 3.2: query dtsh Hais() e observation specimen dtsh Hais(), amostra (sample) p/ 43 pacientes

Sendo assim a parte de criação das amostras (arquivos sample em abox) envolveu a parte de teste (determinação de limites de instâncias nos arquivos (assertion) em abox).

Com isso surgem possibilidades de resolução nas dependências geradas para intercâmbio entre as possíveis ontologias, ou seja, determinar o formato de dados que resulte em melhor intercâmbio entre diferentes abox após o processamento em diferentes bases de conhecimento dos dados tabulares, o que também ocasiona por exemplo em recortes específicos (por exemplo redução e união de atributos), e determinação de procedimentos de concatenação para gerar ids de suporte para inclusão na parte de identificação das URIs.

Utilizar um tipo de estruturação conjuntiva, para adaptação em tabelas menos generalizadas, pode gerar resultados tanto maiores quanto menores em termos de adição de classes e propriedades (relacionando ambas os vocabulários por exemplo), a orientação dos procedimentos sub-sequentes depende da necessidade e orientação da inteligência (para determinar a possibilidade por exemplo de consulta mais rápida mensurada conforme adaptação e possibilidades de inferência útil.

A partir de uma ontologia conforme a criada (a partir da terminologia datasharingbr.ttl), o processamento de ontologias menos especificadas pode ser feito através de Group Patterns, por exemplo fazendo a correspondência via string do exame (Uréia) ou verificando quais pacientes possuem mesmo id (schema:identifier na ontologia criada, e id paciente). No caso na função query identify_exams urea(g) utiliza-se a estrutura de grafo conjuntivo p/ fazer as correspondências entre o arquivo test exames.ttl, amostra selecionada com um instância de uréia, e o arquivo EINSTEIN Exames_2 SAMPLE.csv.nq, gerado utlizando a biblioteca cow (csv on the web), que possuem correspondência de 1952 com o mesmo id conforme a ontologia base (data sharing - Hais) mas apenas 6 atendimentos distintos conforme pode ser observado executando a query.

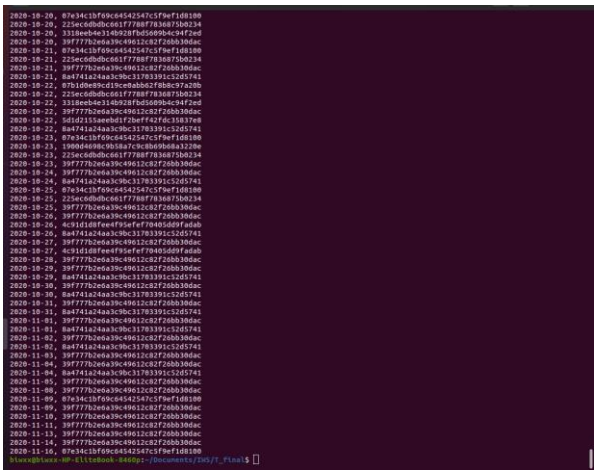

Figure 3.3: query pacientes id at()

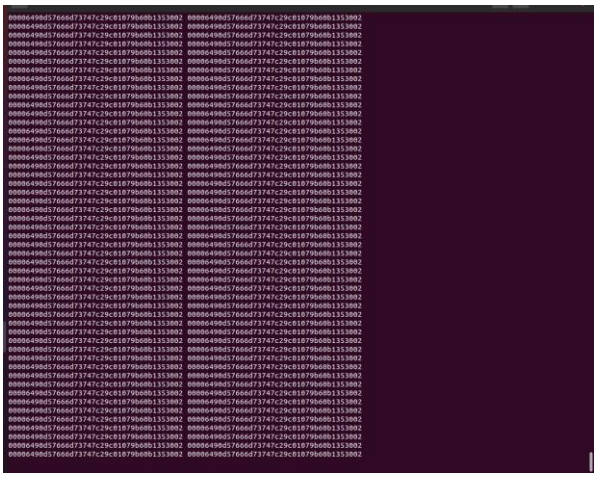

Figure 3.4: query identify exams urea()

O arquivo EINSTEIN Exames 2 SAMPLE.csv.nq foi gerado a partir de uma amostra do csv, para 5000 tuplas de exames.

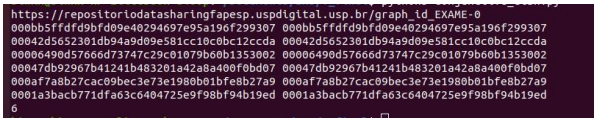

Figure 3.5: query identify exams urea(), correspondência de exames distintos entre as amostras

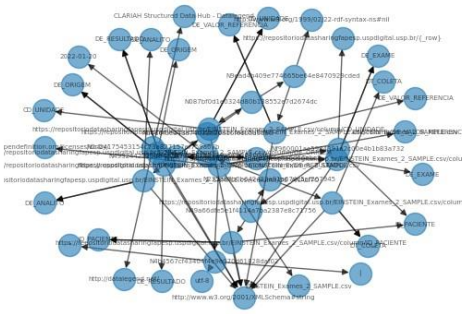

Figure 3.6: EINSTEIN Exames 2 SAMPLE.csv-metadata.json

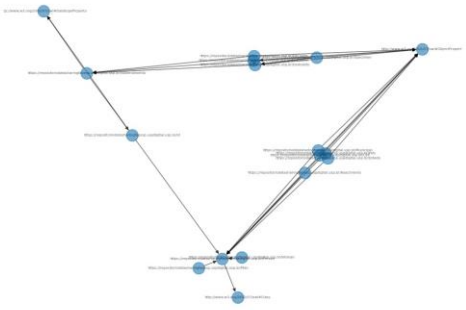

Figure 3.7: