Capítulo

5

Harmonização Global de Dados de Saúde: O Papel dos Vocabulários Padronizados OHDSI

Maria Tereza Fernandes Abrahão (HIAE, OHDSI Latam), Pablo Jorge Madril (OHDSI Latam), Mateus de Lima Freitas (HIAE, OHDSI Latam)

Abstract

The Observational Health Data Sciences and Informatics (OHDSI) has emerged as a leading force in large-scale health data analysis, and its use of the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) plays a central role in this success. The standardization provided by the OMOP CDM is essential to enable large-scale observational analysis, allowing data from different sources to be harmonized and analyzed in a consistent and reliable manner. A vital component of this process is the integration of the OMOP CDM with a comprehensive and effective vocabulary system. This vocabulary plays a crucial role in structuring and organizing data, providing a solid foundation for integrating diverse data sources. By ensuring semantic consistency and interoperability between data, the ontology facilitates comprehensive and reliable analysis, even when dealing with complex and heterogeneous data sets.

Resumo

A Observational Health Data Sciences and Informatics (OHDSI) emergiu como uma força líder na análise de dados de saúde em larga escala, e sua utilização do Common Data Model (CDM) do Observational Medical Outcomes Partnership (OMOP) desempenha um papel central nesse sucesso. A padronização proporcionada pelo CDM OMOP é fundamental para viabilizar a análise observacional em grande escala, permitindo que dados de diferentes fontes sejam harmonizados e analisados de maneira consistente e confiável. Um componente vital desse processo é a integração do CDM OMOP com um sistema de vocabulário abrangente e eficaz. Esse vocabulário desempenha um papel crucial na estruturação e organização dos dados, fornecendo uma base sólida para a integração de diversas fontes de dados. Ao garantir a consistência semântica e a interoperabilidade entre os dados, o vocabulário facilita uma análise abrangente e confiável, mesmo quando lidando com conjuntos de dados complexos e heterogêneos.

5.1 OHDSI – Introdução

A OHDSI (Observational Health Data Sciences and Informatics)[1], é uma colaboração global que redefiniu a área de pesquisa observacional em dados de saúde trazendo a possibilidade de realizar análises sistemáticas em grandes massas de dados provindas de diversas fontes. Fundada em 2014, sua missão é capacitar uma comunidade de ciência aberta para gerar evidências médicas, promovendo decisões de saúde baseadas em dados sólidos. A iniciativa desenvolveu um Modelo Comum de Dados (Common Data Model CDM) padronizado, chamado CDM OMOP (Observational Medical Outcomes Partnership), que facilita a integração de dados de saúde de diferentes fontes. Além disso, a OHDSI fornece vocabulários padronizados e ferramentas de software livre que possibilitam a geração sistemática de evidências em larga escala para análise avançada de dados de saúde.

O Vocabulário Padronizado OHDSI é um dos componentes fundamentais desta iniciativa. A OHDSI desenvolve e mantém um vocabulário de referência centralizada em grande escala para harmonização internacional de dados de saúde. Esse vocabulário inclui uma ampla gama de termos e conceitos relacionados à saúde, abrangendo diferentes áreas, como doenças, procedimentos médicos, medicamentos, resultados clínicos e características demográficas dos pacientes. Ao padronizar os termos usados para descrever esses elementos, o Vocabulário Padronizado OHDSI permite que pesquisadores e profissionais de saúde compartilhem e comparem os resultados de análises estatísticas realizadas nos dados de diferentes fontes e sistemas de saúde de forma consistente e interoperável.

A padronização dos vocabulários de dados de saúde facilita a realização de estudos multicêntricos e meta-análises, além de promover uma melhor compreensão e interpretação dos resultados. Isso também é crucial para o desenvolvimento e aprimoramento de métodos analíticos em epidemiologia e pesquisa clínica.

O capítulo está estruturado como se segue. Primeiro, a seção 5.1 oferece uma breve fundamentação sobre a iniciativa OHDSI, desafios e soluções. A seção 5.2 apresenta o modelo comum de dados (CDM) e a seção 5.3 conceitos gerais dos vocabulários. A seção 5.4 os principais vocabulários, a seção 5.5 apresenta a estrutura dos vocabulários OHDSI. A seção 5.6, o Athena (consulta, seleção e download de vocabulários), a seção 5.7 a ferramenta Usagi (mapeamento dos vocabulários). A seção 5.8, o Atlas, a seção 5.9, apresenta as considerações finais e conclusões e a seção 5.10 as referências consultadas.

Na elaboração deste capítulo, seguiu-se o livro da OHDSI, The book of OHDSI[2], de domínio público, sob a licença Creative Commons Zero v1.0 Universal, (16/04/2020). O livro é um documento vivo, mantido pela comunidade por meio de ferramentas de desenvolvimento de código aberto e evolui continuamente. A versão

[1] http://www.ohdsi.org

[2] The book OHDSI http://book.ohdsi.org

online, disponível gratuitamente, sempre representa a versão mais recente. O texto do livro foi traduzido e complementado na escrita deste capítulo. As figuras e tabelas que ilustram os estudos, foram adaptadas do livro e de apresentações dos tutoriais da OHDSI. Os exemplos foram elaborados a partir dos conhecimentos adquiridos pelos autores na participação em eventos OHDSI.

5.1.1 Uma breve história da OHDSI, desafios e soluções

A iniciativa OHDSI (Observational Health Data Sciences and Informatics), fundada em 2014, tem sido fundamental na transformação da pesquisa observacional em dados de saúde. Seu modelo comum de dados (CDM OMOP) e suas ferramentas de software livre abriram caminho para análises sistemáticas em grandes volumes de dados de saúde, provenientes de diversas fontes. Isso não apenas facilitou a manipulação e análise desses dados, mas também promoveu a interoperabilidade e reprodutibilidade na geração de evidências médicas.

Ao estabelecer uma comunidade global de pesquisadores e bancos de dados observacionais de saúde, com um centro de coordenação na Universidade de Columbia, a OHDSI facilitou a colaboração em escala internacional. Com centenas de pesquisadores em mais de 30 países e registros de saúde de cerca de 600 milhões de pacientes únicos em todo o mundo, a OHDSI está capacitando a comunidade a gerar evidências que promovam melhores decisões e cuidados de saúde. Essa abordagem colaborativa e sistemática está mudando a forma como a pesquisa médica é conduzida, com o objetivo final de melhorar a saúde globalmente [Abrahão 2019].

A adesão dos pesquisadores às práticas de ciência aberta, como as promovidas pela iniciativa OHDSI, desencadeia uma série de benefícios significativos para a pesquisa e a sociedade em geral. Ao abraçar os valores de transparência, colaboração e acessibilidade, a ciência aberta amplia o alcance e o impacto do conhecimento científico de várias maneiras:

1. **Eficiência na pesquisa:** Ao compartilhar dados, materiais e resultados de pesquisa de forma aberta, os pesquisadores podem evitar a duplicação de esforços e acelerar a progressão do conhecimento, aumentando a eficiência da pesquisa.
2. **Confiabilidade dos resultados:** A transparência e a replicabilidade são fundamentais para a confiabilidade dos resultados científicos. Ao permitir que outros pesquisadores examinem e reproduzam os estudos, a ciência aberta promove a confiabilidade e a validade dos achados.
3. **Criatividade e inovação:** O acesso aberto ao conhecimento científico inspira novas ideias e abordagens, incentivando a criatividade e a inovação. Ao permitir que uma ampla gama de indivíduos acesse e contribua para a pesquisa, a ciência aberta pode levar a descobertas inesperadas e avanços significativos.

4. **Colaboração global:** A ciência aberta transcende fronteiras geográficas e disciplinares, facilitando a colaboração entre pesquisadores de diferentes países e áreas de especialização. Isso pode levar a parcerias produtivas e abordagens interdisciplinares para resolver desafios complexos.
5. **Benefícios para a sociedade:** Ao tornar o conhecimento científico amplamente acessível, a ciência aberta beneficia não apenas a comunidade acadêmica, mas também a sociedade em geral. Isso pode levar a avanços em saúde, tecnologia, meio ambiente e outras áreas que impactam diretamente a qualidade de vida das pessoas.

A adoção generalizada das práticas de ciência aberta é fundamental para maximizar o potencial da pesquisa científica e garantir que seus benefícios sejam compartilhados de forma ampla e equitativa.

5.1.2 Componentes da OHDSI

A interoperabilidade entre sistemas depende de 2 fatores [Mucheroni 2011]:

- Interoperabilidade sintática (forma)
- Interoperabilidade semântica (conteúdo)

Para ser possível efetuar comparações e aplicar métodos estatísticos em conjuntos de dados de fontes diversas, a OHDSI se fundamenta em 3 componentes principais:

- Modelo comum de dados: CDM OMOP
- Vocabulários Padronizados
- Ferramentas para preparação do CDM, bibliotecas para análises estatísticas e definição de estudos

As ferramentas disponibilizadas pela OHDSI operam de maneira coesa e padronizada para apoiar uma variedade de análises de dados observacionais no nível do paciente. Os principais pontos em destaque são:

1. **Interoperabilidade com o modelo CDM OMOP:** Todas as ferramentas da OHDSI são projetadas para interagir com bancos de dados estruturados no modelo CDM OMOP. Isso permite que diferentes ferramentas acessem e analisem os dados de forma consistente, independentemente do banco de dados subjacente.
2. **Padronização das análises:** As ferramentas da OHDSI padronizam as análises para vários casos de uso, garantindo consistência e uniformidade nos resultados. Por exemplo, ao calcular uma taxa de incidência, as ferramentas oferecem opções padronizadas para especificar os parâmetros necessários, garantindo que os cálculos sejam realizados de maneira consistente em diferentes projetos e por diferentes usuários.

3. **Facilidade de execução:** Ao padronizar as análises e fornecer uma interface amigável, as ferramentas da OHDSI tornam mais fácil a execução de análises complexas de dados observacionais.
4. **Melhoria da reprodutibilidade e transparência:** Ao garantir que as análises sejam padronizadas e bem documentadas, as ferramentas da OHDSI promovem a reprodutibilidade dos resultados. Isso significa que outros pesquisadores podem reproduzir os mesmos resultados usando as mesmas ferramentas e configurações. Além disso, a transparência é aprimorada, pois as escolhas metodológicas são claramente definidas e documentadas.

No geral, as ferramentas da OHDSI desempenham um papel crucial na promoção da colaboração, reprodutibilidade e transparência na análise de dados observacionais de saúde, permitindo avanços significativos na pesquisa e na prática clínica. O conjunto de ferramentas disponibilizadas pela OHDSI auxiliam a preparação da base no modelo CDM OMOP, na validação do processo de mapeamento, na verificação da qualidade dos dados que compõem o CDM, na elaboração e análise dos diferentes tipos de estudos, facilitando a exploração dos dados e a geração de evidências [Abrahão 2022]. Podemos citar:

- Ferramentas para geração do banco CDM OMOP
    - ETL: White Rabbit, Rabbit-In-A-Hat[3]
    - Vocabulários: Athena[4]/Usagi[5]
    - Qualidade: Achilles[6] e *Data Quality Dashboard* (DQD)[7]
- Ferramentas de análise
    - Geração de Coortes/estudos - Atlas[8]
    - Análises estatísticas: HADES[9]

5.1.3 Onde achar: principais referências

As informações a respeito dos componentes da OHDSI podem ser classificadas em:

➢ Informações gerais: http://www.ohdsi.org - Site principal;

➢ Código e instalações: https://github.com/OHDSI/ Código fonte de todas as ferramentas. Em particular destacamos: *Common Data Model* (https://github.com/OHDSI/CommonDataModel), com a definição completa do modelo e as implementações para os diversos bancos suportados;

[3] White Rabbit, Rabbit-In-A-Hat http://ohdsi.github.io/WhiteRabbit/index.html
[4] ATHENA http://athena.ohdsi.org
[5] Usagi http://ohdsi.github.io/Usagi/
[6] ACHILLES http://www.ohdsi.org/web/achilles
[7] DQD https://ohdsi.github.io/DataQualityDashboard/articles/DataQualityDashboard.html
[8] Atlas https://github.com/OHDSI/Atlas/wiki/Atlas-Setup-Guide
[9] HADES https://github.com/OHDSI/Hades

- ➢ Tutoriais e vídeos: (https://www.google.com/search?q=youtube+ohdsi), documentação e tutoriais em vídeo dos eventos anuais do grupo;
- ➢ Fórum: http://forums.ohdsi.org/
- ➢ Livro OHDSI: The book of OHDSI

5.2 Modelo Comum de Dados (CDM OMOP): Estrutura e Domínios

Um modelo de dados comum desempenha um papel fundamental em várias áreas cruciais da saúde, como:

1. **Colaboração:** Ao adotar um modelo de dados comum, diferentes instituições de saúde, pesquisadores e profissionais médicos podem compartilhar e entender os dados de forma consistente. Isso facilita a colaboração entre diferentes partes interessadas, permitindo a troca de informações e o trabalho conjunto em projetos e iniciativas de saúde.

2. **Pesquisa:** Um modelo de dados comum fornece uma estrutura padronizada para organizar e analisar dados de saúde. Isso simplifica o processo de pesquisa, permitindo que os pesquisadores combinem e analisem grandes conjuntos de dados de forma eficiente. Como resultado, os estudos de pesquisa podem ser realizados de maneira mais eficaz, levando a descobertas significativas e avanços na área da saúde.

3. **Qualidade do cuidado:** Com acesso a dados mais abrangentes e consistentes, os provedores de saúde podem tomar decisões mais informadas sobre o tratamento e a gestão da saúde de seus pacientes. Um modelo de dados comum ajuda a garantir que as informações relevantes estejam disponíveis quando necessário, o que pode levar a uma melhoria na qualidade do cuidado e melhores resultados para os pacientes.

4. **Abordagem integrada:** Ao padronizar a maneira como os dados são coletados, armazenados e compartilhados, um modelo de dados comum promove uma abordagem mais integrada para o gerenciamento da saúde. Isso significa que os dados podem fluir mais facilmente entre diferentes sistemas e partes interessadas, permitindo uma visão mais abrangente e holística da saúde de um indivíduo ou população.

O CDM OMOP é uma estrutura de dados padronizada que facilita o armazenamento, a padronização, integração e análise de informações de saúde. Compreender a estrutura do modelo OMOP é o primeiro passo para qualquer organização que deseje adaptar suas bases de dados para este padrão. Essa compreensão

envolve estudar a organização das tabelas do modelo, os tipos de dados que cada uma armazena e como esses dados são inter-relacionados. Alguns pontos-chave sobre o CDM OMOP são:

1. Estrutura padronizada: estabelece uma estrutura comum para representar uma variedade de dados clínicos, com informações centradas no paciente, procedimentos médicos, condições médicas, medicamentos prescritos e resultados de exames. Isso ajuda a organizar os dados de forma consistente e compreensível.

2. Interoperabilidade: ao adotar o CDM OMOP, diferentes sistemas de saúde e pesquisadores podem compartilhar e comparar dados de saúde de maneira mais eficiente. Isso facilita a integração e análise de grandes conjuntos de dados de diversas fontes, promovendo a colaboração e a troca de informações.

3. Flexibilidade: projetado para ser flexível e abrangente, permitindo a adaptação para atender às necessidades específicas de diferentes estudos e projetos de pesquisa. Ele pode acomodar uma ampla gama de dados clínicos e de pesquisa, proporcionando uma estrutura versátil para análise e investigação.

4. Suporte à pesquisa clínica: amplamente utilizado em estudos de pesquisa clínica e epidemiológica para investigar uma variedade de questões relacionadas à saúde, como segurança de medicamentos, eficácia de tratamentos e padrões de doenças. Ele fornece uma base sólida para análise de dados e geração de insights significativos.

5. Comunidade ativa: OMOP conta com uma comunidade ativa de pesquisadores, desenvolvedores de software e organizações de saúde que trabalham em conjunto para melhorar e expandir o modelo. Isso garante que o CDM do OMOP esteja sempre atualizado e alinhado com as melhores práticas da indústria, promovendo sua relevância e eficácia contínuas.

Esses pontos destacam a importância do CDM do OMOP como uma ferramenta valiosa para padronizar e harmonizar dados observacionais de saúde, facilitando a pesquisa e promovendo avanços na área da saúde.
O CDM OMOP define uma estrutura comum para representar informações clínicas, que facilita a análise comparativa e colaborativa. A Figura 5.1 apresenta o esquema do modelo CDM OMOP.

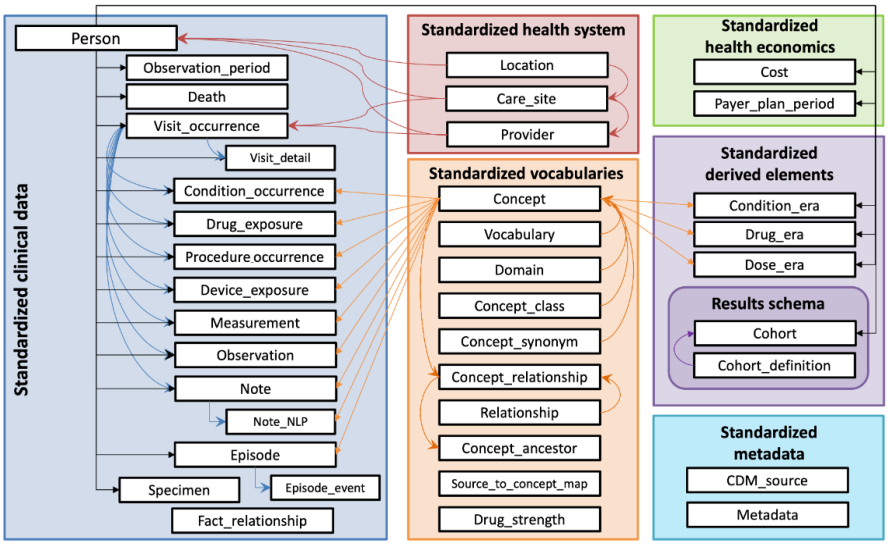


Figura 5.1 Visão geral da estrutura do *CDM OMOP V5.4* [10]

Os componentes do CDM OMOP são:

- Esquema – (tabelas onde são colocados os dados) é formado por:
  - 39 tabelas com 433 colunas
    - 17 tabelas com dados clínicos (centrados na pessoa)
    - 10 tabelas de vocabulários
    - 3 tabelas com informações de saúde, 2 com dados econômicos, 3 com dados derivados, 2 de resultados e 2 com metadados
- Vocabulário:
  - 153 vocabulários distribuídos em 41 domínios
    - Padrões: SNOMED, RxNorm, LOINC
  - 9 milhões de conceitos
    - >3,3 milhões de conceitos padrão
    - >5,1 milhões de códigos-fonte
  - 629.000 conceitos de classificação
    - >55 milhões de relacionamentos conceituais
    - >84 milhões de relacionamentos ancestrais
- Convenções – (regras de como os dados devem ser armazenados).

O CDM é otimizado para fins típicos de pesquisa observacional para identificar populações de pacientes com determinadas intervenções de saúde e resultados; na caracterização dessas populações de pacientes para vários parâmetros, como informações demográficas, história natural da doença, prestação de cuidados de saúde,

[10] Fonte: https://ohdsi.github.io/CommonDataModel/

morbidades, tratamentos; prever a ocorrência destes resultados em pacientes individuais; estimar o efeito que estas intervenções têm na população.

Para atingir este objetivo, o desenvolvimento do CDM segue os seguintes elementos de design:

- **Adequação à finalidade:** O CDM visa fornecer dados organizados de uma forma ideal para análise;
- **Proteção de dados:** Todos os dados que possam comprometer a identidade e a proteção dos pacientes, como nomes, datas de nascimento são limitados.
- **Design de domínios:** Os domínios são modelados em um modelo de dados relacionais centrado na pessoa, onde para cada registro a identidade da pessoa e uma data são capturadas no mínimo. Aqui, um modelo de dados relacional é aquele em que os dados são representados como uma coleção de tabelas vinculadas por chaves primárias e estrangeiras.
- **Justificativa para domínios:** Os domínios são identificados e definidos separadamente em um modelo de relacionamento entre entidades se tiverem um caso de uso de análise (condições, por exemplo) e o domínio tiver atributos específicos que não são aplicáveis de outra forma. Todos os outros dados podem ser preservados como uma observação na tabela de observação em uma estrutura entidade-atributo-valor.
- **Vocabulários Padronizados:** Para padronizar o conteúdo desses registros, o CDM conta com os Vocabulários Padronizados contendo todos os conceitos de saúde padrão correspondentes, necessários e apropriados.
- **Reutilização de vocabulários existentes:** Se possível, esses conceitos são aproveitados de organizações ou iniciativas nacionais ou industriais de padronização ou definição de vocabulário, como a Biblioteca Nacional de Medicina, o Departamento de Assuntos de Veteranos, o Centro de Controle e Prevenção de Doenças, etc.
- **Manutenção de códigos-fonte:** embora todos os códigos sejam mapeados para os vocabulários padronizados, o modelo também armazena o código-fonte original para garantir que nenhuma informação seja perdida.
- **Neutralidade tecnológica:** O CDM não exige uma tecnologia específica. Pode ser realizado em qualquer banco de dados relacional, como Oracle, SQL Server etc., ou como conjuntos de dados analíticos SAS.
- **Escalabilidade:** O CDM é otimizado para processamento de dados e análise computacional para acomodar fontes de dados que variam em tamanho, incluindo bancos de dados com até centenas de milhões de pessoas e bilhões de observações clínicas.
- **Compatibilidade com versões anteriores:** todas as alterações dos CDMs anteriores estão claramente delineadas no repositório github[11]. Versões mais

[11] https://github.com/OHDSI/CommonDataModel

antigas do CDM podem ser facilmente criadas a partir da versão atual e nenhuma informação que estava presente anteriormente será perdida.

A Figura 5.2 apresenta uma análise do esquema[12] do CDM OMOP com a documentação das tabelas, colunas, tipos de dados, tamanho, relacionamentos, um guia para o usuário, índices e convenções para o ETL dos dados. A Figura 5.3.1 e 5.3.2 ilustrão o detalhamento da tabela DEATH com sua documentação.

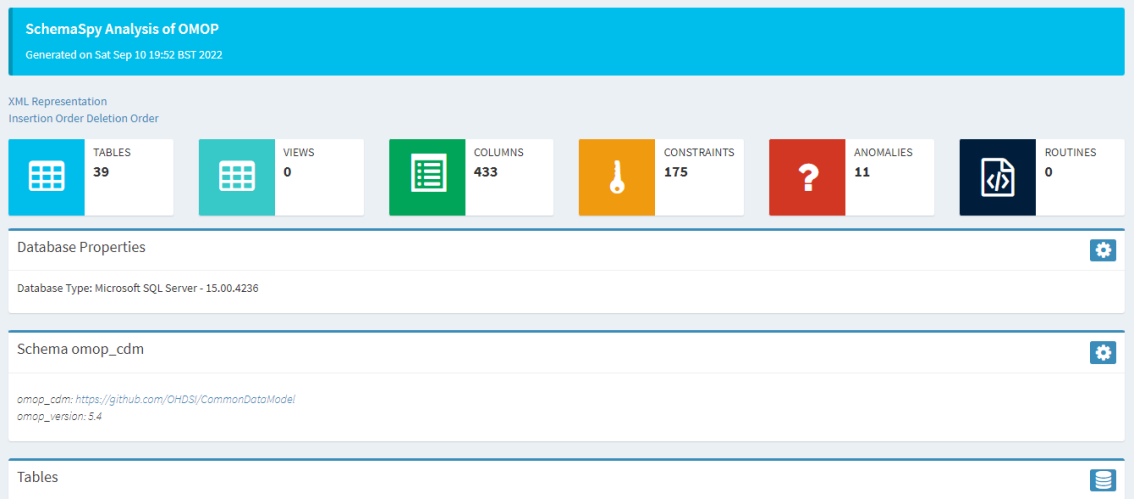


Figura 5.2 Análise do esquema do *CDM OMOP*

DEATH - OMOP Database

Column	Type	Size	Nulls	Auto	Default	Children	Parents	Comments
person_id	bigint	8			null		PERSON.person_id fpk_DEATH_person_idR	
death_date	date	6			null			User Guidance: The date the person was deceased. ETL Conventions: If the precise date include day or month is not known or not allowed, December is used as the default month, and the last day of the month the default day.
death_datetime	datetime	16,3	√		null			ETL Conventions: If not available set time to midnight (00:00:00)
death_type_concept_id	int	4	√		null		CONCEPT.concept_id fpk_DEATH_death_type_concept_idR	User Guidance: This is the provenance of the death record, i.e., where it came from. It is possible that an administrative claims database would source death information from a government file so do not assume the Death Type is the same as the Visit Type, etc. ETL Conventions: Use the type concept that be reflects the source of the death record. Accepted Concepts. A more detailed explanation of each Type Concept can be found on the vocabulary wiki.
cause_concept_id	int	4	√		null		CONCEPT.concept_id fpk_DEATH_cause_concept_idR	User Guidance: This is the Standard Concept representing the Person's cause of death, if available. ETL Conventions: There is no specified domain for this concept, just choose the Standard Concept Id that best represents the person's cause of death.
cause_source_value	varchar	50	√		null			ETL Conventions: If available, put the source code representing the cause of death here.
cause_source_concept_id	int	4	√		null		CONCEPT.concept_id fpk_DEATH_cause_source_concept_idR	ETL Conventions: If the cause of death was coded using a Vocabulary present in the OMOP Vocabularies put the CONCEPT_ID representing the cause of death here.

Figura 5.3.1 Detalhes da tabela DEATH

[12] Esquema OMOP Análise https://omop-erd.surge.sh/index.html

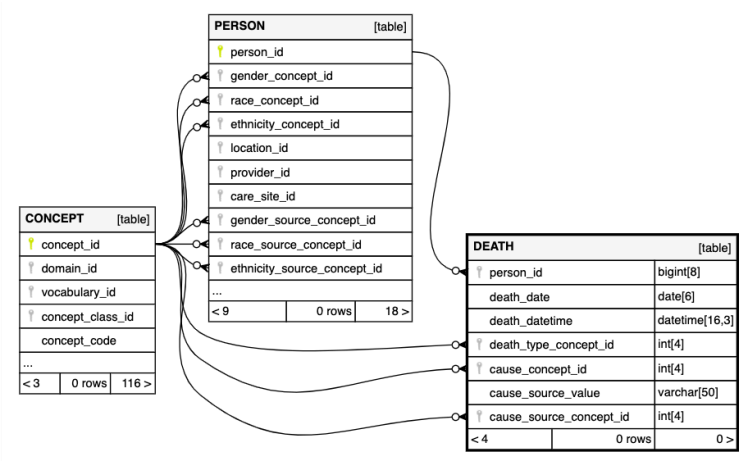


Figura 5.3.2 Fluxo de relacionamento da tabela DEATH

Essa padronização em um conjunto definido de tabelas e relacionamentos fornece um contexto comum para os elementos de dados clínicos, o que é necessário para criar métodos e algoritmos analíticos e de garantia de qualidade unificados que possam ser executados em toda a rede.

A harmonização do conteúdo de dados médicos é alcançada por meio do uso de vocabulários médicos ou esquemas de codificação, que são desenvolvidos e mantidos por diversas organizações e sociedades profissionais. Esses vocabulários garantem uma comunicação precisa e consistente sobre cuidados e tratamentos de pacientes. Eles variam desde conjuntos simples de códigos ou termos até hierarquias complexas ou vocabulários, frequentemente com cobertura cruzada de diversos domínios de saúde.

Para qualquer domínio específico, os membros de redes de dados distribuídas podem empregar diferentes vocabulários, versões distintas do mesmo vocabulário, vocabulários não públicos ou, em alguns casos, não utilizar vocabulário padronizado em seus dados. Essa diversidade pode complicar a integração e a análise dos dados, tornando a harmonização uma tarefa essencial para assegurar a interoperabilidade e a eficácia das trocas de informações na área da saúde.

5.3 Conceitos gerais dos vocabulários

Neste item vamos revisar, através de uma revisão bibliográfica, os principais conceitos relacionados com vocabulários como auxílio na compreensão da estrutura dos vocabulários da OHDSI.

Para isso reproduzimos partes das mais importantes fontes da literatura:

- O relatório do standard Guidelines for the construction, format, and management of monolingual controlled vocabularies (ANSI/NISO Z39.19-2005) [National 2005], fonte com as definições mais referenciadas da área
- Health Concept and Knowledge Management: Twenty-five Years of Evolution, uma atualização do trabalho seminal de Cornet e Chute [Cornet 2016]

- Desiderata for Controlled Medical Vocabularies in the Twenty-First Century, de Cimino [Cimino 2016]
- The MMI Guides: Navigating the World of Marine Metadata [Stocks 2010], de diversos autores, cada autor é indicado no correspondente parágrafo. Esta obra é um compêndio definitivo dos termos relacionados com a classificação de conceitos. É de destacar que este grupo apresenta um dos maiores trabalhos de criação de ferramentas gratuitas para criação e manutenção de vocabulários
- Outros autores são referenciados nos respectivos parágrafos.

A linguagem é uma das principais características que definem o ser humano. Única entre todas as formas de comunicação animal, se distingue por ser composicional, ela expressa pensamentos em sentenças abrangendo sujeito, verbo e objeto e reconhece passado, presente e futuro. Esta característica permite que a linguagem humana tenha uma capacidade ilimitada para gerar novas frases combinando conjuntos limitados de palavras, por exemplo, com 25 palavras de cada tipo podemos formular 15.000 frases diferentes. Outra característica é ser referencial, o que permite aos agentes falantes trocar informações específicas entre eles a respeito de outras pessoas ou objetos e as suas localizações ou ações [Pagel 2017].

À medida que cresce a quantidade de informação acessível globalmente, aumenta também a inerente dificuldade em encontrar itens de informação desejados. Dentro do contexto de arquivos documentais, uma das técnicas mais aplicadas para facilitar a descoberta de itens, tanto em sistemas manuais tradicionais quanto em sistemas mais recentes informatizados, tem sido a indexação.

A indexação consiste na atribuição de valores a atributos predefinidos para servir como base para pesquisa e descoberta de recursos. A combinação desses atributos e valores devem constituir informações suficientes para caracterizar com sucesso o conteúdo de um documento e permitir a futura recuperação deste documento apenas observando essas informações. Exemplos dos atributos comumente encontrados são: autor, título, assunto, resumo, etc. Estes são geralmente chamados de metadados [Ferreira 2005].

Mas esta flexibilidade também cria dificuldades resumidas na seguinte lista:

- Padronização: termos diferentes que representam o mesmo conceito (sinônimos)
- Ambiguidade: termos iguais para ideias diferentes. Exemplo:
    - Mercúrio: planeta
    - Mercúrio: metal
    - Mercúrio: Deus grego
- Pesquisa: dificuldade para achar o termo certo numa lista grande de termos
- Classificação: dificuldade para separar e agrupar conceitos em classes que facilitem a pesquisa

- Representação de conhecimento: Necessidade de transmitir informações a respeito da informação.

Assim aparecem as primeiras iniciativas de organização, no começo, a partir da definição de listas de termos para auxiliar na identificação unívoca de objetos ou conceitos.

Um termo é definido como uma ou mais palavras usadas para representar um conceito.

Estas listas de termos começam a crescer e agora precisamos achar os termos, distinguir sinônimos, evitar erros ortográficos; começamos assim a descrever o que hoje são chamados de vocabulários controlados.

Um vocabulário controlado é uma forma de inserir uma camada interpretativa de semântica entre o termo inserido pelo usuário e o banco de dados subjacente para melhor representar a intenção original do usuário [Karl 2002].

Isto permite que na hora da busca, possa ser apresentada uma lista de opções que capturem a intenção do usuário e facilitem a ele achar o que procura.

Um vocabulário controlado é uma lista de termos que foram enumerados explicitamente. Esta lista é controlada por uma autoridade de registro que se constitui na fonte autoritativa da verdade. Todos os termos de um vocabulário controlado devem ter uma definição inequívoca e não redundante, o que depende de quão rigorosa é a autoridade controladora em relação ao registro de novos termos [National 2005].

Um Authority File ou arquivo de referência ou autorizativo, é um tipo de vocabulário controlado que consiste em uma lista de rótulos e valores que estabelecem os valores aceitáveis que podem ser inseridos em um parâmetro específico. Nenhuma explicação ou informação adicional é fornecida sobre os valores aceitáveis.
Duas regras devem ser aplicadas:

- Se o mesmo termo é comumente usado para significar conceitos diferentes, então o seu nome é explicitamente qualificado para resolver esta ambiguidade. NOTA: Esta regra não se aplica a anéis de sinônimos.
- Se vários termos foram usados para significar a mesma coisa, um dos termos é identificado como o termo preferido no vocabulário controlado e os outros termos são listados como sinônimos ou pseudônimos.

A Figura 5.4 mostra os diversos tipos de vocabulários controlados em função da capacidade de representar conhecimento e dos problemas abordados [Heather 2022].

Lista de Termos - Glossários (Term List)	**Anel de Sinônimos (Synonym Ring)**	**Fonte autoritativa de verdade (Authority File)**	**Taxonomia (Taxonomy)**	**Dicionário de sinônimos (Thesaurus)**	**Ontologia (Ontology)**
Controle da ambiguidade	Controle de sinônimos	Controle de ambiguidade e sinônimos (Atributos)	Controle da ambiguidade Controle de sinônimos Relacionamentos hierárquicos	Controle da ambiguidade Controle de sinônimos Relacionamentos hierárquicos Relacionamentos associativos	Controle da ambiguidade Relacionamentos semânticos Classes Atributos

Figura 5.4 Classificação dos vocabulários

5.3.1 Classificação dos vocabulários controlados pela forma

Para permitir a gestão formal, um vocabulário controlado pode ser organizado estruturalmente de modo que se encaixe em uma dessas amplas categorias [Stocks 2010]:

- **Plana**: fornece um conjunto de termos obrigatórios que podem ser usados. Alguns vocabulários planos controlados fornecerão informações adicionais sobre cada termo.
- **Multinível**: baseia-se em um vocabulário plano e controlado, atribuindo cada termo a uma categoria.
- **Relacional**: fornece um conjunto de termos e captura como eles estão associados entre si.

A Tabela 5.1 apresenta esta organização.

Tabela 5.1 Classificação dos vocabulários pela forma

Categoría de Vocabulário Controlado	**Tipo de vocabulário controlado**	**Descrição**
Plano (Flat)	Arquivo de autoridade (Authority File)	Lista de Termos preestabelecidos
	Glossário	Lista de termos e definições dentro de um domínio específico
	Dicionário	Lista de termos, definições, e informações adicionais
	Lista de códigos	Lista de códigos (ex. abreviaturas) e definições
Multinível	Taxonomia	Termos classificados em categorias
	Subject Heading (Cabeçalho do Assunto)	Termos classificados em categorias, que podem ser classes abrangentes
Relacionais	Tesauros	Conjunto de termos e os relacionamentos entre valores individuais
	Redes Semânticas	Conjunto de termos/conceitos e relacionamentos direcionados
	Ontologias	Conjunto de termos e relações entre termos, aprimorado por informações adicionais fornecidas por regras e axiomas

5.3.1.1 Vocabulários planos: Arquivo de autoridade (Authority Files), Glossários, Dicionários, Diários, Lista de códigos

Todos os vocabulários simples contém um rótulo e um valor. Alguns vocabulários planos baseiam-se nesta base, adicionando uma definição ou informações adicionais sobre cada valor. Nenhum relacionamento é estabelecido, nenhuma hierarquia é estabelecida e nenhuma matriz complicada é criada.

5.3.1.2 Vocabulários multiníveis: Taxonomias, Título do assunto (Subject Headings)

Um vocabulário multinível é essencialmente uma forma de agrupar termos em classes com hierarquia. Uma classificação diz mais sobre os "termos" do que um vocabulário plano, colocando-os em subcategorias bem pensadas.

Em um vocabulário multinível, você pode examinar a qual subcategoria um termo pertence e também examinar as relações entre as subcategorias. Em alguns vocabulários multiníveis (taxonomias), a única conexão entre as subcategorias é uma

comparação "mais abrangente"/"mais específica" (BT-Broader Term/NT-Narrower Term).

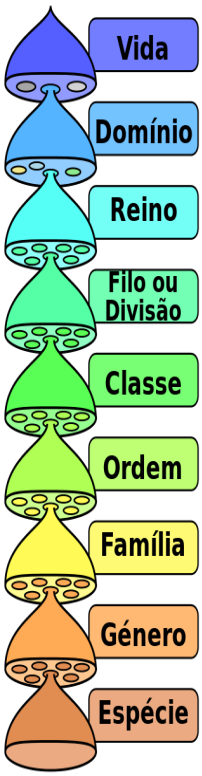
Em outros, você pode comparar categorias semelhantes em categorias mais amplas (cabeçalho de assunto- Subject Headings).

5.3.1.2.1 Taxonomia

É um vocabulário controlado multinível no qual os valores de metadados são agrupados de acordo com classes específicas de assunto, geralmente hierárquicas.
Táxon é derivado do grego taxis, ordenamento, por tanto taxonomia é a ciência que lida com a descrição, identificação e classificação de conceitos.

Talvez a taxonomia mais conhecida seja a taxonomia de Lineu, que classifica de forma única os seres vivos, extensamente usada nas ciências biológicas. Ela foi desenvolvida por Carolus Linnaeus (Conhecido normalmente como Carl von Linné ou em português como Carlos Lineu, botânico e médico sueco) no Século XVIII durante a grande expansão da história natural. A Tabela 5.2 mostra a classificação do ser humano.

Tabela 5.2 Classificação do ser humano na taxonomia de Linnean

 (figura: Peter Halasz)	**Domínio**: Eukarya Característica: seres vivos compostos por células com um núcleo celular organizado **Reino**: Animalia Característica: células eucarióticas com membrana celular, mas sem parede celular, multicelulares, heterotróficas **Filo**: Chordata Característica: notocorda, cordão nervoso dorsal e fendas branquiais faríngeas **Classe**: Mamíferos Característica: glândulas endotérmicas, capilares e mamárias (que servem para nutrir os filhotes) **Ordem**: Primatas Característica: clavícula, olhos voltados para frente, agarrar as mãos com os dedos, dois tipos de dentes **Família**: Hominidae Característica: postura ereta, cérebro grande, visão estereoscópica, rosto achatado, mãos e pés com diferentes funções **Gênero**: Homo Característica: coluna curvada em S **Espécie**: sapiens Característica: testa alta, queixo bem desenvolvido, ossos do crânio finos

Observe que, como seres humanos, recebemos um identificador único (Homo sapiens) e exibimos todas as características listadas (em outras palavras, como nossa

classificação está no final de uma lista aninhada, podemos herdar todas as características das “superclasses”). Para classificar completamente humanos nesta taxonomia, precisamos usar o termo "Homo sapiens", mas também pode ser usado "primatas". Esta não seria a classificação mais restrita, mas é uma classificação precisa.

5.3.1.3 Vocabulários Relacionais: Tesauros, Rede Semânticas, Ontologias

Os Vocabulários Relacionais, também chamados de listas de relacionamentos, contém um mecanismo para conectar termos. As relações são descritas por vários padrões e protocolos, como para tesauros no padrão ANSI [233]/NISO [109] Z39.19 - 2005, incluindo "mais abrangente"/"mais específico", "use" (*USE*, no sentido: "use no lugar", apontando para o termo principal), "usado para" (*UF-USE FOR*: apontado para os termos semelhantes) e "relacionado" (RT- Related Term).

5.3.1.3.1 Tesauro

No contexto de metadados, um tesauro é um tipo de vocabulário relacional controlado que fornece uma lista de termos de metadados com relações específicas entre os termos. De acordo com o padrão ANSI/NISO Z39.19 - 2003 [National 2005]:

Um tesauro é um vocabulário controlado organizado em uma ordem conhecida e estruturada de modo que relações de equivalência, homografia, hierarquia e associação entre valores são exibidas claramente e identificadas por indicadores de relacionamento padronizados que são empregados reciprocamente.

Os objetivos principais de um tesauro são:

(a) facilitar a recuperação de documentos

(b) alcançar consistência na indexação de documentos escritos ou registrados de outra forma e outros itens, principalmente para sistemas pós-coordenados de armazenamento e recuperação de informações.

Os tesauros basicamente pegam as taxonomias descritas acima e as estendem para torná-las mais capazes de descrever o mundo [Garshol 2004], não apenas permitindo que os assuntos sejam organizados em uma hierarquia, mas também permitindo que outras declarações sejam feitas sobre os assuntos. O padrão ISO 2788 [Garshol 1986] fornece as seguintes propriedades para descrever assuntos:

BT (Broader Term)

Abreviação de “termo mais amplo”, refere-se ao termo acima deste na hierarquia; esse termo deve ter um significado mais amplo ou menos específico. Na prática, alguns sistemas permitem múltiplos BTs para um período, enquanto outros não. (Existe uma propriedade inversa conhecida como NT (Narrower Term), para "termo mais restrito", que está implícita no BT.) Pode-se dizer que as taxonomias conforme descritas acima são tesauros que usam apenas as propriedades BT/NT para construir uma hierarquia, e

sem fazermos uso de nenhuma das propriedades descritas abaixo, então pode-se dizer que todo tesauro contém uma taxonomia.

SN (Scope Note)
Esta é uma string anexada ao termo explicando seu significado no dicionário de sinônimos. Isto pode ser útil nos casos em que o significado preciso do termo não é óbvio no contexto. "SN" significa "nota de escopo".

USE (e o complemento: UF Use For)
Refere-se a outro termo que deve ser preferido em vez deste termo; implica que os termos são sinônimos. (Existe uma propriedade inversa conhecida como UF: Use For, "use para"). Por exemplo, na figura Figura 5.5 do Tesauro da UNESCO, o termo "Dreams" aponta para "Sleep", representando o "USE" graficamente com uma seta. Na definição do termo "Sleep" vemos explicitamente a referência .

RT (Related Term / Related Concept)
Abreviação de "termo relacionado", refere-se a um termo que está relacionado a este termo, sem ser um sinônimo dele ou um termo mais amplo/restrito. No nosso exemplo: "Suggestopaedia"

Em suma, os tesauros fornecem um vocabulário muito mais rico para descrever os termos do que as taxonomias e, portanto, são ferramentas muito mais poderosas. Como pode ser visto, usar um tesauro em vez de uma taxonomia resolveria vários problemas práticos na classificação de objetos e também na busca por eles.

Na Figura 5.5 vemos o Tesauro da UNESCO[13], apresentando os relacionamentos de "*Dreams*" para o termo principal "*Sleep*" (*use* preferred term).

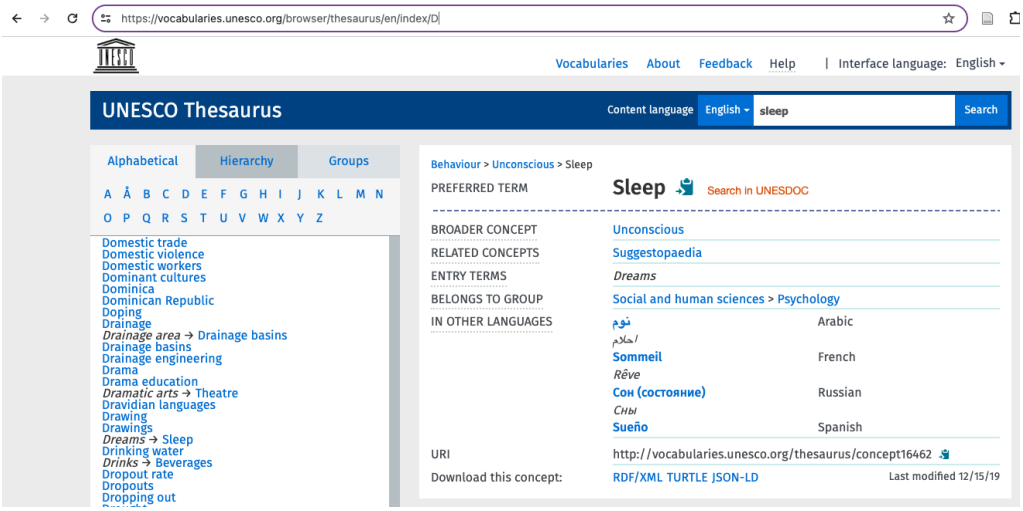


Figura 5.5 Tesauro da Unesco

[13] Unesco https://vocabularies.unesco.org

Por sua vez, a definição de "*Sleep*" aponta para "*Dreams*" (*use for*) mostrado como "*Entry Terms*".

Entry terms (pontos de entrada), no exemplo "*Dreams*", são os vocábulos que permitem que uma busca consiga recuperar mais facilmente o termo principal (preferred term) da definição de um termo (*recall*) [DeMars 2022].

O "MMI Guides: Navigating the World of Marine Metadata"[Stocks 2010] e o padrão (ANSI/NISO Z39.19-2005) [National 2005] contém definições para todos os elementos destas classificações.

5.3.1.3.2 Ontologias

Segundo Tom Gruber (1993):

"An ontology is a specification of a conceptualization"

(Uma ontologia é a especificação de uma conceitualização)

Ele descreve uma ontologia da seguinte maneira:

Uma ontologia especifica um vocabulário com o qual fazer afirmações, que podem ser entradas ou saídas de agentes de conhecimento (como um programa de software). Uma ontologia deve ser formulada em alguma linguagem de representação.

A demanda por uma linguagem de representação restringe o conceito, ou seja, a definição de Gruber (1995) descreve o conceito de ontologia formal [Madsen 2009].

No seu artigo de revisão, "Health Concept and Knowledge Management: Twenty-five Years of Evolution" [Cornet, R, and C G Chute 2016], os autores discursam sobre os conceitos já aqui descritos e definem a principal característica das classificações estatísticas:

Terminologia: Um sistema de conceitos com identificadores atribuídos e termos de linguagem humana, normalmente envolvendo algum tipo de hierarquia semântica. Alguns sistemas podem suportar a atribuição de múltiplos termos, ou sinónimos, a um determinado conceito; estes podem incluir termos em vários idiomas naturais, como inglês ou holandês.

Ontologia: Uma terminologia que invoca relações semânticas formais entre conceitos, normalmente manifestadas como um tipo de Lógica de Descrição.

Classificação: Um sistema terminológico destinado a descrever exaustivamente um domínio ou tópico, normalmente invocando a colocação criteriosa de categorias residuais, como Não especificado ou Não classificado em outro lugar, para alcançar a abrangência.

Classificações Estatísticas:

Uma classificação onde todos os conceitos são mutuamente exclusivos para evitar contar as coisas duas vezes. Isso normalmente é conseguido usando uma mono-hierarquia, onde cada conceito tem um e apenas um pai.

É importante distinguir as classificações estatísticas das terminologias, uma vez que servem propósitos diferentes. A classificação estatística mais conhecida é Classificação Internacional de Doenças[14] (CID).

O CID é importante porque fornece uma linguagem comum para registar, notificar e monitorar doenças. Isto permite ao mundo comparar e partilhar dados de uma forma consistente e padronizada – entre hospitais, regiões e países e ao longo de períodos de tempo. Facilita a recolha e armazenamento de dados para análise e tomada de decisões baseadas em evidências.

O objetivo mais importante do CID é manter estatísticas mundiais de morbidade e mortalidade através do tempo. Por isso nada pode ser contado duas vezes.
Isso obriga que a classificação tenha categorias do tipo:

- NEC: Not Elsewhere Classified (Sem classificação)
- NOS: Not Otherwise Specified (Sem especificação)

Estas são chamadas de categorias residuais. Elas, junto com a mono-hierarquia, são imprescindíveis para a construção de uma classificação estatística.

5.3.2 Exemplificando as diferenças entre os tipos de vocabulários

As seguintes figuras exemplificam os diferentes tipos [Stocks 2010].

5.3.2.1 Dicionário

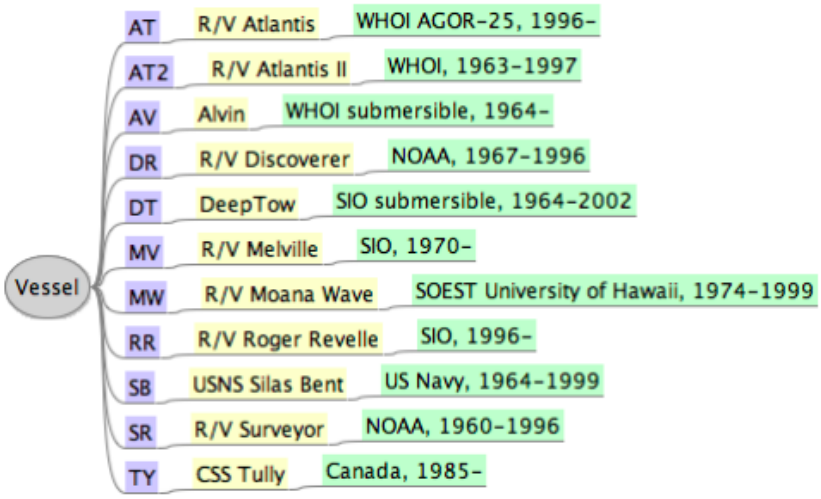


Figura 5.6 Dicionário

Cada termo é articulado com uma sigla. (1ª entrada, azul)
As siglas estão explicadas na descrição. (2ª entrada, amarelo)

[14] World Health Organization. (2004). ICD-10 : international statistical classification of diseases and related health problems : tenth revision, 2nd ed. World Health Organization. https://iris.who.int/handle/10665/42980

Informações adicionais sobre como cada termo surgiu estão incluídas na etimologia. (3ª entrada, verde).
Isto forma a definição clássica da relação objeto/metainformação onde a metainformação tem 3 campos: sigla, descrição, etimologia.

5.3.2.2 Taxonomia

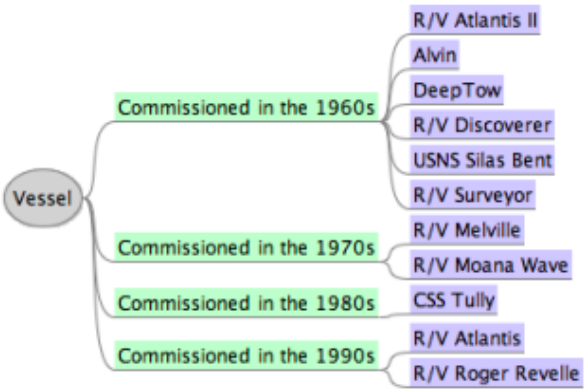


Figura 5.7 Taxonomia

Os termos reais (2ª entrada, azul) são colocados numa estrutura, de acordo com a década em que foram comissionados (1ª entrada, verde).

5.3.2.3 Ontologia

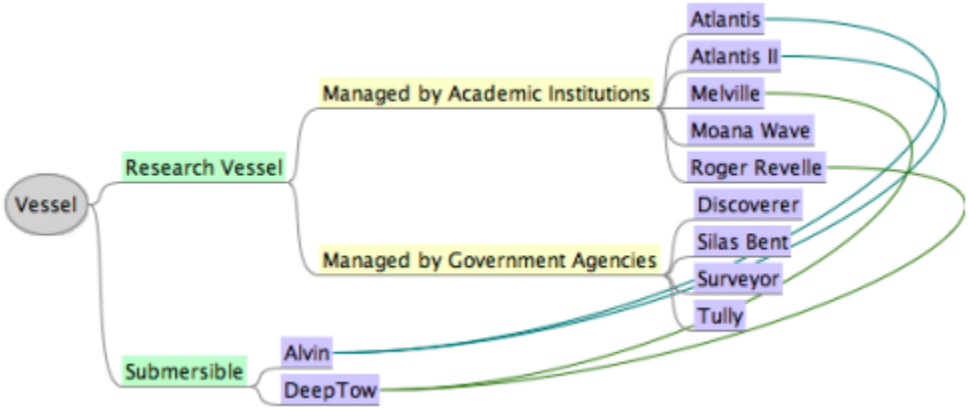


Figura 5.8 Ontologia

Os termos reais (3ª entrada, azul) são classificados em duas classes principais (1ª entrada, verde) e uma subclasse (2ª entrada, amarelo).
Observe que as embarcações estão conectadas a submersíveis, de acordo com a instituição operadora. Esta é uma inter-relação complexa que aumenta a hierarquia de classes.

Cada um desses vocabulários controlados representa a mesma lista de objetos do mundo real (ou seja, embarcações ou submarinos). Eles são apresentados com diferentes tipos de vocabulários controlados, usando termos diferentes para representar os mesmos objetos do mundo real e com informações ligeiramente distintas.

5.3.3 Importância dos vocabulários controlados

Temos então agora, um conjunto limitado de conceitos associados a termos específicos e com variabilidade controlada por uma autoridade que faz manutenção contínua deste vocabulário controlado.

A construção de vocabulários controlados é um processo demorado e trabalhoso, especialmente se o domínio a ser coberto é amplo e a terminologia utilizada é rica e complexa. O trabalho envolvido é justificado porque o uso de vocabulários controlados ajuda a garantir consistência na indexação e promove uma recuperação mais satisfatória.

Vocabulários controlados são importantes para pesquisadores por vários motivos [Stocks 2010]:

- Consistência
- Precisão
- Automação
- Simplificação de entrada
- Interoperabilidade
- Aprimoramento de pesquisas e descobertas
- Completude
- Gestão de longo e curto prazo
- Uso eficiente do tempo

Em muitos casos, termos de vocabulário controlado definem completamente o conteúdo permitido para um determinado elemento de metadados.

Além disso, um vocabulário controlado pode ser facilmente incorporado em procedimentos automatizados. Em um sistema de dados, um vocabulário controlado pode simplificar a entrada do sistema e contribuir para o controle de qualidade, fornecendo aos usuários ou outros sistemas uma lista de entradas permitidas e podem ser usados para verificar descrições de metadados existentes ou importadas quanto à consistência e correção, incluindo ortográfica e hifenização.

A necessidade de interoperabilidade surgiu logo após o desenvolvimento dos primeiros vocabulários controlados. Um trabalho considerável, tanto prático quanto acadêmico, foi feito para desenvolver métodos que permitam que vocabulários controlados sejam usados em vários bancos de dados e sistemas e compartilhados entre indexadores e pesquisadores.

Dentre os mecanismos que facilitam a interoperabilidade podemos citar os descritores e as palavras-chaves[15].

Descritores (subject headings): São conjuntos padronizados e formalmente atribuídos de termos (também chamados de descritores) em um banco de dados para identificar os tópicos principais de um livro ou artigo. Os dois exemplos mais conhecidos de descritores são:

- Library of Congress Subject Headings (LCSH) [Library of Congress 2004]
- MeSH (Medical Subject Headings)[16]

Palavras-chave são palavras ou frases que podem aparecer em qualquer lugar do item (dados de citação ou texto completo de um livro ou artigo). A pesquisa por palavra-chave é a forma normal de pesquisa nos mecanismos de pesquisa na web e é um bom começo para encontrar títulos de assuntos relevantes em bancos de dados.

[15] CEU Library, The Central European University Library, Vienna, Austria https://ceu.libguides.com/databasesearchtips/subjects

[16] MeSH (Medical Subject Headings) (2004) Bethesda (MD): National Library of Medicine. Available from: http://www.nlm.nih.gov/mesh

5.3.4 Knowledge organization system (KOS)

É um termo genérico usado para se referir a uma ampla gama de itens (por exemplo, títulos de assuntos, tesauros, esquemas de classificação e ontologias), que foram concebidos com relação a diferentes propósitos, em momentos históricos distintos. São caracterizados por diferentes estruturas e funções específicas, formas variadas de se relacionar com a tecnologia e são utilizados numa pluralidade de contextos por diversas comunidades [Mazzocchi 2018]. No entanto, o que todos têm em comum é que foram concebidos para apoiar a organização do conhecimento e da informação, de forma a facilitar a sua gestão e recuperação.

Uma das tipologias mais abrangentes tenha sido fornecida por Souza et al. [Souza 2012], que ainda identificam a estrutura como o principal critério de divisão, embora também esteja incluída uma divisão secundária, que leva em conta numerosos domínios de aplicação e casos de uso. A Figura a seguir apresenta este detalhamento.

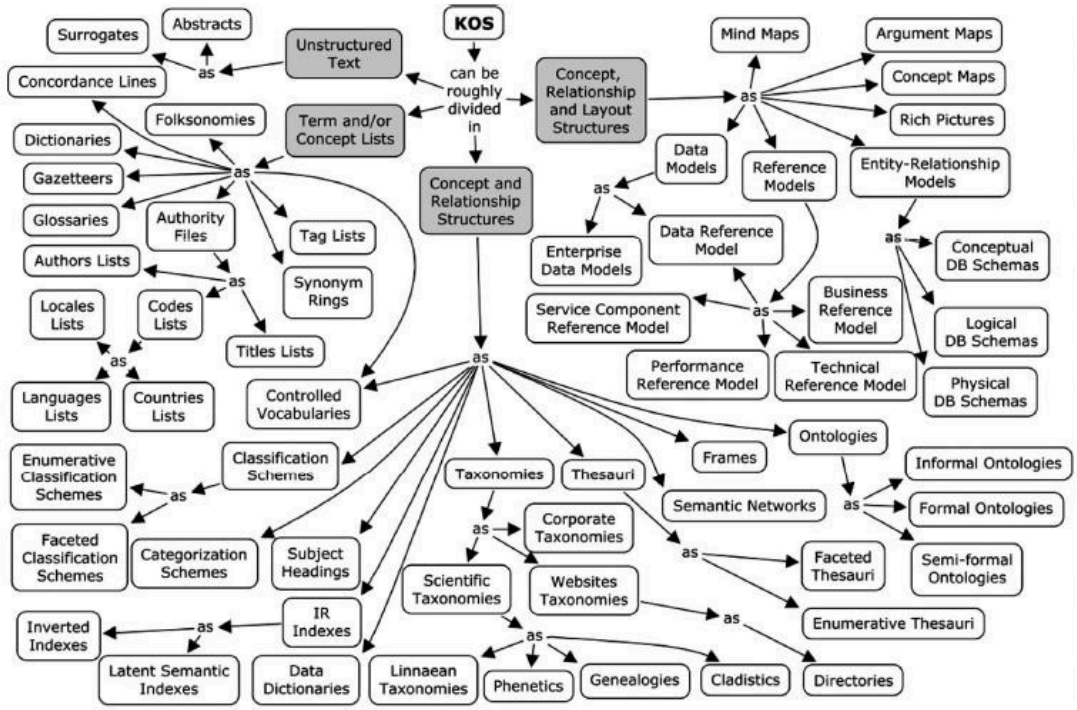


Figura 5.9 Souza et al.'s (2012) classification of KOSs

5.3.4.1 SKOS

O Simple Knowledge Organization System[17] é uma recomendação do W3C [Mazzocchi 2018] projetada para representação de tesauros, esquemas de classificação, taxonomias,

[17] Introduction to SKOS https://www.w3.org/2004/02/skos/intro

sistemas de controle de autoridade ou qualquer outro tipo de vocabulário controlado estruturado.

É uma área de trabalho que desenvolve especificações e padrões para apoiar o uso de sistemas de organização do conhecimento (KOS), como tesauros, esquemas de classificação, sistemas de cabeçalhos de assuntos e taxonomias no âmbito da Web Semântica.

A seguinte Figura descreve o avanço das terminologias no tempo [Lomax 2021]:

Figura 5.10 Terminologias no tempo (1960 a 2017)

5.4 Principais Vocabulários Padrão OHDSI

SNOMED CT[18] é uma abreviação para "Systematized Nomenclature of Medicine Clinical Terms", que em português significa "Nomenclatura Sistematizada de Termos Clínicos em Medicina". É um sistema de terminologia clínica abrangente usado globalmente na área da saúde e pesquisa clínica. SNOMED CT é organizado em torno de conceitos em vez de termos individuais, possui uma estrutura hierárquica e é amplamente utilizado para codificação clínica em registros eletrônicos de saúde, faturamento, suporte a decisões e gerenciamento de saúde da população. É mantido pela IHTSDO e continuamente atualizado para refletir novos conhecimentos médicos e mudanças na prática clínica. Em resumo, SNOMED CT desempenha um papel fundamental na padronização e interoperabilidade das informações clínicas, melhorando a qualidade e eficiência dos cuidados de saúde.

LOINC[19], ou Logical Observation Identifiers Names and Codes, é um sistema de codificação universalmente reconhecido para identificar testes laboratoriais e observações clínicas. Ele fornece códigos numéricos e alfanuméricos para uma ampla variedade de observações clínicas, como medidas de laboratório, observações clínicas, questionários e escalas de avaliação. LOINC é usado para padronizar a comunicação e intercâmbio de dados entre sistemas de saúde, facilitando a interoperabilidade e o compartilhamento de informações clínicas. Ele é mantido pelo Regenstrief Institute, Inc. e continua a evoluir para abranger novas áreas de observação clínica e laboratorial. Em resumo, LOINC desempenha um papel crucial na integração de dados clínicos e na promoção da qualidade e eficiência dos cuidados de saúde.

[18] https://www.snomed.org/ e https://www.nlm.nih.gov/healthit/snomedct/index.html
[19] https://loinc.org/

O **RxNorm**[20] é um sistema de terminologia (tesauro) desenvolvido pelo National Library of Medicine (NLM) dos Estados Unidos. Ele fornece nomes padronizados para medicamentos e produtos relacionados à saúde, bem como seus ingredientes ativos, formas farmacêuticas, dosagens e rotas de administração. RxNorm é amplamente utilizado em sistemas de informação de saúde para facilitar a interoperabilidade e a troca de dados sobre medicamentos entre diferentes instituições e sistemas de informação. Ele ajuda a evitar ambiguidades na comunicação sobre medicamentos, promove a segurança do paciente e auxilia na pesquisa clínica e farmacêutica.

O RxNorm vincula nomes normalizados para medicamentos clínicos a muitos vocabulários de medicamentos comumente usados em software de gerenciamento de farmácia e interação medicamentosa, incluindo os do First Databank, Micromedex, Multum e Gold Standard Drug Database. Ao fornecer links entre esses vocabulários, o RxNorm pode mediar mensagens entre sistemas que não usam o mesmo software e vocabulário.

No entanto, o principal desafio envolve as limitações do RxNorm, um modelo originalmente concebido para representar medicamentos no mercado dos EUA. Seu foco está nos medicamentos atuais, capturando detalhes como conteúdo e dosagem de forma eficaz. No entanto, é insuficiente em dois aspectos críticos: não pode representar conceitos internacionais sobre medicamentos e não inclui dados sobre medicamentos que já não estão disponíveis. Isso coloca obstáculos adicionais ao trabalho dos usuários, como:

- Necessidade de acesso a um amplo espectro de dados sobre medicamentos, incluindo informações internacionais e anteriores sobre medicamentos para estudos abrangentes.
- Exigir informações precisas e completas sobre medicamentos de vários mercados internacionais para melhor atendimento ao paciente e planejamento de tratamento.
- Companhias farmacêuticas que buscam dados extensos sobre medicamentos para análise e pesquisa do mercado global.

  Para resolver essas lacunas, o sistema RxNorm Extension[21] foi desenvolvido.

Esta solução inovadora foi concebida para expandir o âmbito do RxNorm, incorporando diferentes dados sobre medicamentos de diferentes países, ampliando significativamente a sua aplicabilidade para além do mercado dos EUA. Além disso, incorpora dados sobre medicamentos descontinuados, cruciais para estudos longitudinais e para a compreensão da evolução dos medicamentos e dos perfis de

[20] https://www.nlm.nih.gov/research/umls/rxnorm/index.html
https://www.ohdsi.org/web/wiki/doku.php?id=documentation:international_drugs

[21] https://medium.com/sciforce/rxnorm-extension-tool-to-standardize-source-drug-data-using-omop-cdm-71fd87eddaa2

segurança, tornando o RxNorm Extension uma ferramenta abrangente para análise global de medicamentos.

5.5 Vocabulários OMOP/OHDSI

Os Vocabulários Padronizados OHDSI são uma coleção de vocabulários públicos consolidados na estrutura da tabela CDM. Este processo envolve a atribuição de identificadores estáveis a códigos individuais, tornando-os únicos em todo o sistema, adicionando atributos e estabelecendo relações para integrar os vocabulários em uma estrutura ontológica global. Além disso, a OHDSI cria seus próprios vocabulários e relacionamentos para referência interna e padronização semântica. O conteúdo de um banco de dados CDM OMOP é um conjunto de eventos clínicos representados por um código que identifica o quê aconteceu ao paciente e em que data. Essa representação é definida num sistema de codificação.

Com o tempo, os sistemas desenvolvidos para descrever os eventos clínicos expandiram-se enormemente em tamanho e complexidade e espalharam-se por outros aspectos dos cuidados de saúde, tais como procedimentos e serviços, medicamentos, dispositivos médicos, etc. Os princípios fundamentais permaneceram os mesmos: são vocabulários controlados, terminologias, hierarquias ou ontologias com as quais algumas comunidades de saúde concordam com a finalidade de capturar, classificar e analisar dados de pacientes. Muitos destes vocabulários são mantidos por agências públicas e governamentais com um mandato de longo prazo para o fazer. Por exemplo, a Organização Mundial da Saúde (OMS) produz a Classificação Internacional de Doenças (CID).

Como resultado, cada país, região, sistema de saúde e instituição tende a ter as suas próprias classificações que muito provavelmente só seriam relevantes onde fossem utilizadas. Esta miríade de vocabulários impede a interoperabilidade dos sistemas em que são utilizados. A padronização é a chave que permite a troca de dados dos pacientes, desbloqueia a análise de dados de saúde a nível global e permite a investigação sistemática e padronizada, incluindo a caracterização do desempenho e a avaliação da qualidade. Para resolver esse problema, surgiram organizações multinacionais que começaram a criar padrões amplos, como a OMS mencionada acima, a Nomenclatura Padrão de Medicina (SNOMED) e Nomes e Códigos de Identificadores de Observação Lógica (LOINC). Nos EUA, o Comitê de Padrões de TI em Saúde (HITAC) recomenda o uso de SNOMED, LOINC e o vocabulário de medicamentos RxNorm como padrões ao Coordenador Nacional de TI em Saúde (ONC), para uso em uma plataforma comum para troca nacional de informações de saúde entre diversas entidades .

Normalmente, encontrar e interpretar o conteúdo dos dados observacionais de saúde, sejam dados estruturados usando esquemas de codificação ou dispostos em texto

livre, é passado até o pesquisador, que se depara com uma infinidade de maneiras diferentes de descrever eventos clínicos.

A OHDSI exige harmonização não apenas com um formato padronizado, mas também com um conteúdo que segue um padrão rigoroso.

A OHDSI desenvolveu o OMOP CDM, um padrão global para pesquisa observacional. Como parte do CDM, os Vocabulários Padronizados OMOP estão disponíveis para dois propósitos principais:

- Repositório comum de todos os vocabulários usados na comunidade
- Padronização e mapeamento para uso em pesquisa

Os Vocabulários Padronizados estão disponíveis gratuitamente à comunidade e devem ser utilizados para a instância do OMOP CDM como sua tabela de referência obrigatória.

Todos os vocabulários dos Vocabulários Padronizados são consolidados no mesmo formato comum. Isso evita que os pesquisadores tenham que compreender e lidar com vários formatos e convenções de ciclo de vida diferentes dos vocabulários originários. Ele é construído e administrado pela Equipe de Vocabulário OHDSI, que faz parte do Grupo de Trabalho CDM geral da OMOP.

A descrição do quê aconteceu num evento clínico dentro de um banco CDM é representada por um CONCEPT. As tabelas dos vocabulários expressam todos os possíveis CONCEPT que podem ser utilizados para descrever um evento clínico. Este vocabulário representa um arquivo de autoridade (Authority File) no sentido que todos os eventos clínicos presentes num banco CDM-OMOP precisam ser identificados com esta codificação.

5.5.1 Representação de Conteúdo através de Conceitos

Nas tabelas de dados do CDM o conteúdo de cada registro é totalmente normalizado e representado por meio de Conceitos (CONCEPT). Os conceitos são armazenados em tabelas de eventos com seus valores CONCEPT_ID, que são chaves estrangeiras para a tabela CONCEPT, que serve como tabela de referência geral. Todas as instâncias do CDM utilizam a mesma tabela CONCEPT como referência dos Conceitos, que juntamente com o Modelo Comum de Dados é um mecanismo chave de interoperabilidade e a base da rede de pesquisa OHDSI. Se um Conceito Padrão não existir ou não puder ser identificado, o valor do CONCEPT_ID é definido como 0, representando um conceito inexistente, um valor desconhecido ou não mapeável.

Os registros da tabela CONCEPT contêm informações detalhadas sobre cada conceito (nome, domínio, classe etc.). Conceitos, Relacionamentos de Conceitos, Ancestrais de Conceitos e outras informações relativas aos Conceitos estão contidas nas tabelas dos Vocabulários Padronizados. Existem apenas três tipos de conceitos, que são:

STANDARD CONCEPT - Conceito Padrão:
De todas as possíveis codificações disponíveis para codificar um conceito, apenas uma é escolhida como STANDARD (padrão) para representar o significado de cada evento clínico.
Por exemplo, o código MESH D001281, o código CIEL 148203, o código SNOMED 49436004, o código ICD9CM 427.31 e o READ CODE G573000 definem “fibrilação atrial” no domínio da condição, mas apenas o conceito SNOMED é padrão e representa a fibrilação atrial nas tabelas de eventos do paciente. Os demais são designados ou como conceitos NON-STANDARD ou como SOURCE-CONCEPT e mapeados para os Padrões. Podemos pensar neles como "Entry Points" ou "Entradas" dos vocabulários controlados, usadas para ajudar na procura de um termo. Os Conceitos Padrão são indicados através de um “S” no campo STANDARD_CONCEPT. E apenas esses STANDARD_CONCEPT são usados para registrar dados nos campos do CDM que terminam em "_CONCEPT_ID".

NON-STANDARD CONCEPT - Conceitos Fora do Padrão:
Conceitos não padronizados não são utilizados para representar os eventos clínicos, mas ainda fazem parte dos Vocabulários Padronizados e são frequentemente encontrados nos dados de origem. Por esse motivo, também são chamados de “conceitos fonte” (SOURCE CONCEPT). A conversão de conceitos fonte em Conceitos Padrão é um processo denominado “mapeamento”. Conceitos não padronizados não possuem valor no campo STANDARD_CONCEPT (são deixados como NULL).

CLASSIFICATION CONCEPT - Conceitos de Classificação:
Esses conceitos não são padrão e, portanto, não podem ser usados para representar os dados. Mas eles participam da hierarquia com os Conceitos Padrão e podem, portanto, ser usados para realizar consultas hierárquicas. Por exemplo, a consulta de todos os descendentes do código MedDRA 10037908 irá recuperar o conceito SNOMED Padrão para Fibrilhação Auricular.
A seguinte figura exemplifica os vocabulários envolvidos nos 3 tipos de conceitos:

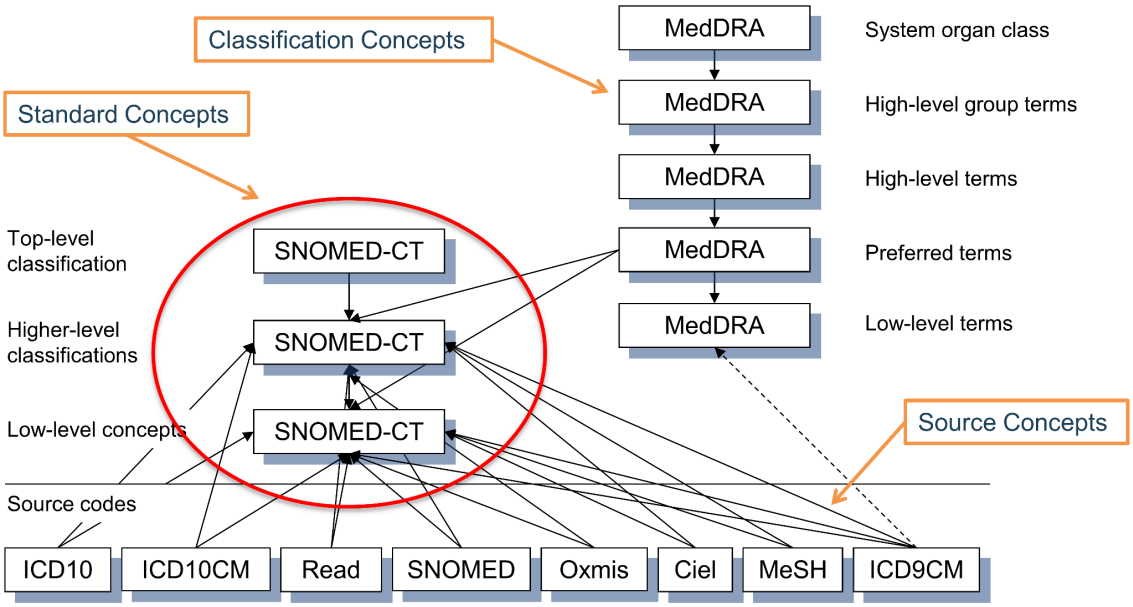


Figura 5.11 Exemplo dos tipos de conceitos

5.5.2 Estrutura

Nos Vocabulários Padronizados não há distinção por que uma informação não está disponível; pode ser devido a uma retirada ativa de informações pelo paciente, a um valor ausente, a um valor que não está definido ou padronizado de alguma forma ou à ausência de um registro de mapeamento em CONCEPT_RELATIONSHIP. Qualquer conceito deste tipo não está mapeado, o que corresponde por default, a uma mapeamento para o STANDARD_CONCEPT com o ID_CONCEPT = 0.

O conjunto dos códigos padrão (STANDARD CODES) forma um vocabulário controlado plano, em particular uma classificação estatística, dado que o ID_CONCEPT = 0 representa a categoria residual (NEC ou NOS). Isso se deve a que para poder montar análises estatísticas precisamos garantir que nada é contado duas vezes.

A função dos Conceitos classificatórios (CLASSIFICATION CONCEPTS) é subsanar os defeitos de uma estrutura plana, auxiliando a definir hierarquias classificatórias dos termos. Porém, devemos ressaltar que nas tabelas de eventos de um banco CDM-OMOP apenas existem STANDARD CODES.

As hierarquias classificatórias são herdadas dos vocabulários que deram origem aos STANDARD CODES, como SNOMED e RxNorm.

5.5.3 Hierarquia

Dentro de um domínio, os conceitos padrão e de classificação são organizados em uma estrutura hierárquica e armazenados na tabela CONCEPT_ANCESTOR. Isso permite consultar e recuperar conceitos e todos os seus descendentes hierárquicos. Esses

descendentes possuem os mesmos atributos de seu ancestral, mas também atributos adicionais ou mais definidos.

A tabela CONCEPT_ANCESTOR é construída automaticamente a partir da tabela CONCEPT_RELATIONSHIP percorrendo todos os conceitos possíveis conectados através de relacionamentos hierárquicos. Estes são os pares “É um” - “Inclui” (ver Figura 5.6) e outros relacionamentos que conectam hierarquias entre vocabulários. A escolha se um relacionamento participa do construtor de hierarquia é definida para cada ID de relacionamento pelo sinalizador DEFINES_ANCESTRY na tabela de referência RELATIONSHIP. A Figura 5.12 apresenta estes relacionamentos.

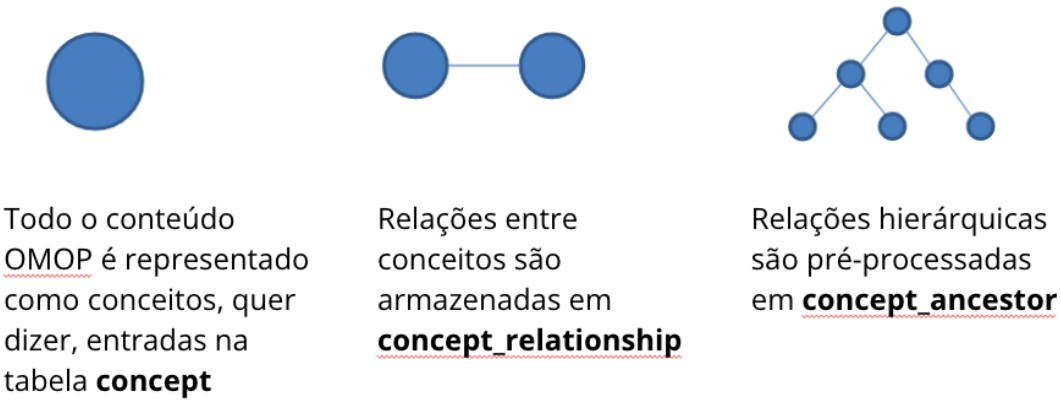


Figura 5.12 Representação dos conceitos e seus relacionamentos

A Figura 5.13 exemplifica os diversos relacionamentos que podem acontecer para o caso de "Fibrilação Atrial".

A ancestralidade de primeiro grau é definida através dos relacionamentos “É um” e “Inclui” (no sentido Paciente "é uma" Pessoa, Pessoa "inclui" Paciente) enquanto todas as relações de grau superior são inferidas e armazenadas na tabela CONCEPT_ANCESTOR. Cada conceito também é seu próprio descendente com ambos os níveis de separação iguais a 0.

O grau ancestral, ou número de passos entre ancestral e descendente, é capturado nos campos MIN_LEVELS_OF_SEPARATION e MAX_LEVELS_OF_SEPARATION, definindo a conexão mais curta ou mais longa possível. Nem todas as relações hierárquicas contribuem igualmente para o cálculo dos níveis de separação. Uma etapa contada para o grau é determinada pelo sinalizador IS_HIERARCHICAL na tabela de referência RELACIONAMENTO para cada ID de relacionamento.

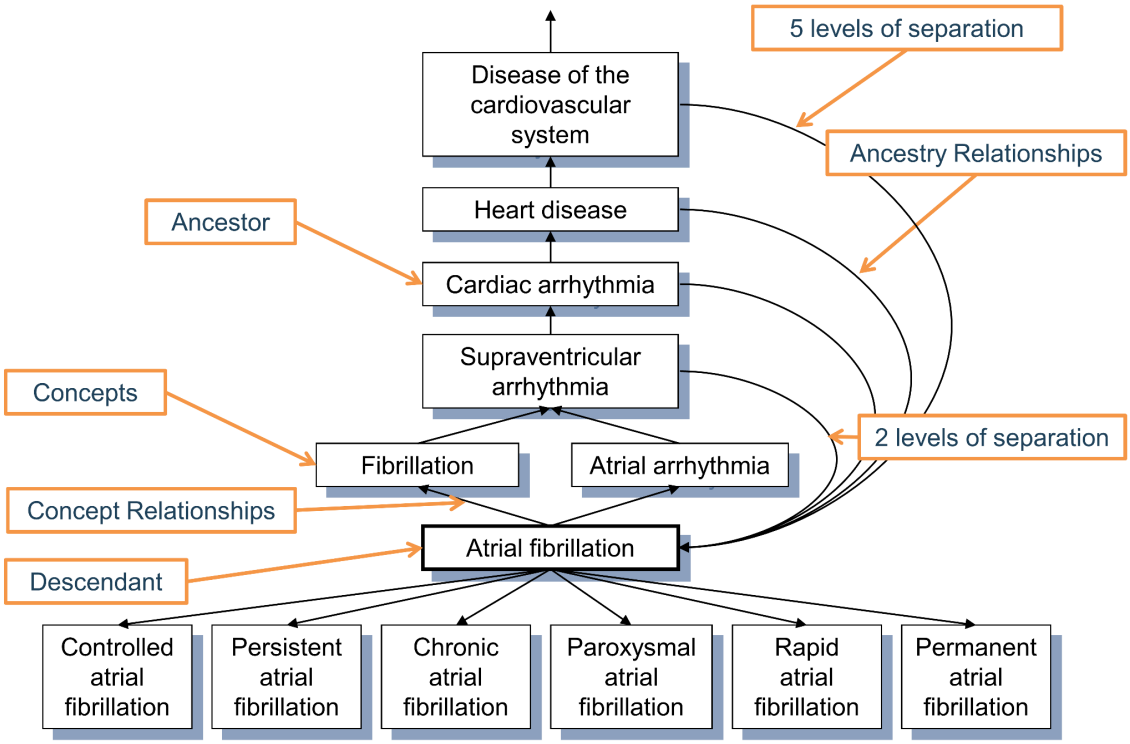


Figura 5.13 Exemplo de hierarquia da condição “Fibrilação atrial”

Neste momento, existe uma hierarquia abrangente de alta qualidade apenas para dois domínios: medicamentos (RxNorm) e doenças (SNOMED). Os domínios de procedimento, medição e observação estão apenas parcialmente cobertos e em processo de construção. A ancestralidade é particularmente útil para o domínio de medicamentos, pois permite navegar por todos os medicamentos com um determinado ingrediente ou membros de classes de medicamentos, independentemente do país de origem, marca ou outros atributos.

5.5.4 Domínios e conceitos

Eventos de diversas naturezas são organizados em domínios. Esses Eventos são armazenados em tabelas e campos específicos em cada domínio e são representados por conceitos padrão que também são específicos do domínio, conforme definido nos Vocabulários Padronizados. Cada Conceito Padrão tem uma atribuição de Domínio única, que define em qual tabela eles são registrados. Embora a atribuição correta de domínio seja objeto de debate na comunidade, esta regra estrita de correspondência domínio-tabela-campo garante que sempre haja um local inequívoco para qualquer código ou conceito. Por exemplo, sinais, sintomas e conceitos de diagnóstico são do Domínio de Diagnóstico (CONDITION), e são registrados no CONDITION_CONCEPT_ID da tabela CONDITION_OCCURRENCE.

Os Vocabulários Padronizados OHDSI são uma coleção de vocabulários públicos consolidados na estrutura da tabela CDM. Este processo envolve a atribuição de identificadores estáveis a códigos individuais, tornando-os únicos em todo o sistema, adicionando atributos e estabelecendo relações para integrar os vocabulários em uma estrutura ontológica global. Além disso, a OHDSI cria seus próprios vocabulários e relacionamentos para referência interna e padronização semântica.

Após essa preparação, os elementos individuais dos vocabulários são chamados de conceitos. Cada conceito possui um nome (descrição) e vários sinônimos, embora não haja uma tentativa de cobertura lexical abrangente para apoiar o processamento de linguagem natural ou a recuperação de informações. Todos os nomes dos conceitos estão em inglês, mas os sinônimos podem estar em qualquer idioma. Esses nomes e relacionamentos formam a estrutura da ontologia.

Os vocabulários padronizados são uma ontologia de referência comum obrigatória para todos os sites de dados na rede OHDSI. Compreende ontologias importadas e geradas de novo contendo conceitos e relacionamentos entre eles, e a práxis de converter os dados de origem para o CDM OMOP com base neles. Permite a harmonização através de domínios atribuídos de acordo com categorias clínicas, cobertura abrangente de entidades dentro de cada domínio, suporte para esquemas de codificação internacionais comumente usados e padronização de conceitos semanticamente equivalentes.

Os Vocabulários Padronizados OMOP, são uma parte fundamental da rede de pesquisa OHDSI e parte integrante do Common Data Model (CDM). Eles permitem a padronização de métodos, definições e resultados definindo o conteúdo dos dados, para pesquisa e análise de rede. Normalmente, encontrar e interpretar o conteúdo dos dados observacionais de saúde, sejam dados estruturados usando esquemas de codificação ou dispostos em texto livre, é repassado ao pesquisador, que se depara com uma infinidade de maneiras diferentes de descrever eventos clínicos. A OHDSI exige a harmonização não apenas de um formato padronizado, mas também de um conteúdo padrão rigoroso.

Os Vocabulários estão em constante evolução e compreendem hoje mais de 11 milhões de conceitos de 142 vocabulários, distribuídos em 44 domínios. É atualizado regularmente para incluir novos termos, refletir mudanças na terminologia médica e atender às necessidades emergentes dos pesquisadores e profissionais de saúde em todo o mundo. Com cerca de 8.600 usuários, já foram realizados mais de 50.000 downloads do sistema. Essa abordagem dinâmica e colaborativa garante que a ontologia OHDSI continue sendo uma ferramenta eficaz para promover a harmonização internacional de dados de saúde[22].

[22] https://www.ohdsi.org/wp-content/uploads/2023/11/OHDSI-Book2023.pdf

5.5.5 Convenções das tabelas do vocabulário OMOP

As tabelas que compõem o vocabulários padrão, bem como as convenções para a criação dessas tabelas estão detalhadas em um repositório github[23].

Dez tabelas compõem os vocabulários, são elas:

Concepts (Conceitos): Os conceitos do Common Data Model são derivados de uma série de terminologias públicas ou proprietárias, como SNOMED-CT, LOINC e RxNorm, ou gerados de forma personalizada para padronizar aspectos dos dados observacionais. Ambos os tipos de Conceitos são integrados com base nas seguintes regras:

- Todos os conceitos são mantidos centralmente pelo Grupo de Trabalho CDM e Vocabulários. Conceitos adicionais podem ser adicionados, conforme necessário, mediante solicitação;
- Todos os Conceitos é atribuído um identificador numérico exclusivo concept_id que é usado como chave para vincular todos os dados observacionais aos dados de referência do Conceito correspondente;
- O concept_id de um Conceito é persistente, ou seja, permanece o mesmo para o mesmo Conceito entre versões dos Vocabulários Padronizados;
- Um nome descritivo é armazenado para cada Conceito. Nomes e descrições adicionais para o Conceito são armazenados como Sinônimos na tabela CONCEPT_SYNONYM.
- Cada Conceito é atribuído a um Domínio e uma classe;
- O campo concept_code é usado para referenciar o vocabulário fonte;
- Os Conceitos Padrão (designados como 'S' no campo standard_concept) podem aparecer nas tabelas do CDM em todos os campos *_concept_id, enquanto os Conceitos de Classificação ('C') não devem aparecer nos dados do CDM, mas participar da construção da tabela CONCEPT_ANCESTOR e pode ser usado para identificar descendentes que podem aparecer nos dados.
- A vida útil de um Conceito é registrada através de seus campos valid_start_date, valid_end_date e invalid_ reason;
- Os valores para concept_ids gerados como parte dos Vocabulários Padronizados serão reservados de 0 a 2.000.000.000. Acima desta faixa, concept_ids estão disponíveis para uso local e garantem que não entrarão em conflito com versões futuras dos Vocabulários Padronizados;

Vocabularies (vocabulários):

- Existe um registro para cada vocabulário. O campo vocabulário_id contém um identificador alfanumérico;
- O registro com vocabulary_id = 'None' é reservado para conter informações sobre a versão atual de todos os vocabulários padronizados;

[23] https://ohdsi.github.io/CommonDataModel/dataModelConventions.html#Data_Model_Conventions

- O campo vocabulary_name contém o nome oficial completo do Vocabulário, bem como a fonte ou fornecedor entre parênteses;
- Cada Vocabulário possui uma entrada na tabela CONCEPT, que é registrada no campo vocabulário_concept_id;

Domains (Domínios):

- Existe um registro para cada Domínio. Os domínios são definidos pelas tabelas e campos do CDM OMOP que podem conter conceitos que descrevem os aspectos de saúde de um paciente:
- O campo domain_id contém um identificador alfanumérico, que também pode ser utilizado como abreviatura do Domínio;
- O campo domain_name contém os nomes não abreviados do Domínio;
- Cada Domínio também possui uma entrada na tabela Conceito, que é registrada no campo domain_concept_id;

Concept Classes (Classes de Conceitos):

- Há um registro para cada classe conceitual. Classes de Conceito são usadas para criar estrutura adicional para os Conceitos dentro de cada vocabulário;
- O campo concept_class_id contém um identificador alfanumérico, que também pode ser utilizado como abreviatura da Classe Conceito;
- O campo concept_class_name contém os nomes não abreviados da classe conceitual.
- Cada Classe Conceito também possui uma entrada na tabela Conceito, que é registrada no campo concept_class_concept_id;

Concept Relationships (Relacionamento de Conceitos):

- Os relacionamentos geralmente podem ser classificados como hierárquicos (pai-filho) ou não hierárquicos (laterais);
- Todos os Relacionamentos são direcionais e cada Relacionamento é representado duas vezes simetricamente na tabela CONCEPT_RELATIONSHIP;
- Há um registro para cada relacionamento de conceito;
- Relacionamentos de conceito definem relacionamentos diretos entre conceitos;

Relationship Table (Tabela de Relacionamentos):

- Existe um registro para cada Relacionamento;
- Os relacionamentos são classificados como hierárquicos (pai-filho) ou não hierárquicos (laterais);
- Eles são usados para determinar quais registros de relacionamento de conceito devem ser incluídos no cálculo da tabela CONCEPT_ANCESTOR;
- O campo relacionamento_id contém um identificador alfanumérico, que também pode ser utilizado como abreviação do Relacionamento;
- O campo Relationship_name contém os nomes não abreviados do Relacionamento;

- Cada Relacionamento também possui uma entrada equivalente na tabela Conceito, que é registrada no campo relacionamento_conceito_id;
- Relacionamentos hierárquicos são usados para construir uma árvore hierárquica a partir dos conceitos, que é registrada na tabela CONCEPT_ANCESTOR;
- As relações, também hierárquicas, podem ser entre Conceitos dentro de um mesmo Vocabulário ou aqueles adotados de diferentes fontes de Vocabulários;

Concept Synonyms (Sinônimos de Conceitos):

- O campo concept_synonym_name contém um sinônimo válido de um conceito, incluindo a descrição no próprio concept_name. Cada conceito possui pelo menos um Sinônimo na tabela CONCEPT_SYNONYM;
- Somente sinônimos ativos, atuais em inglês, são armazenados na tabela CONCEPT_SYNONYM;

Concept Ancestor (Conceito Ancestral):

- Cada conceito também é registrado como um ancestral de si mesmo.
- Somente conceitos válidos e padrão participam da tabela CONCEPT_ANCESTOR;
- Normalmente, apenas Conceitos de um mesmo Domínio são conectados através de registros da tabela CONCEPT_ANCESTOR, mas pode haver exceções.

Source to Concept Map (Mapa de Conceitos):

- Esta tabela não é mais usada para distribuir informações de mapeamento entre códigos-fonte e Conceitos Padrão para os Vocabulários Padrão. Em vez disso, a tabela CONCEPT_RELATIONSHIP é usada para esse propósito, usando o relacionamento_id='Maps to';
- No entanto, esta tabela ainda pode ser usada para a tradução de códigos-fonte locais em Conceitos Padrão;
- Os campos valid_start_date, valid_end_date e invalid_reason são utilizados para definir o ciclo de vida das informações de mapeamento;

Drug Strength (Força da droga):

- A tabela DRUG_STRENGTH contém informações para cada conceito de medicamento padrão ativo;
- Um medicamento que contém vários ingredientes ativos resultará em vários registros DRUG_STRENGTH, um para cada ingrediente ativo;
- As informações sobre a concentração do ingrediente são fornecidas como quantidade absoluta;
- Todos os vocabulários de Medicamentos contendo Conceitos Padrão possuem entradas na tabela DRUG_STRENGTH;

5.6 Athena: consulta, seleção e download de vocabulários

Athena[24] é um aplicativo de software de código-fonte aberto, baseado na web, desenvolvido pela comunidade OHDSI, gratuito e disponível publicamente. Permite a pesquisa e navegação pelos códigos que compõem os diferentes vocabulários com o objetivo de selecionar aqueles que melhor representem os conteúdos da nossa base de origem. No site selecionamos o conjunto dos vocabulários para download. Esta seleção será usada junto ao CDM OMOP e na ferramenta USAGI para mapeamento dos códigos dos vocabulários locais para o vocabulário padrão. A Figura 5.14 apresenta a interface da ferramenta. No primeiro acesso, fazer cadastro e login (simples e rápido com um clique na opção mais à direita do aplicativo.

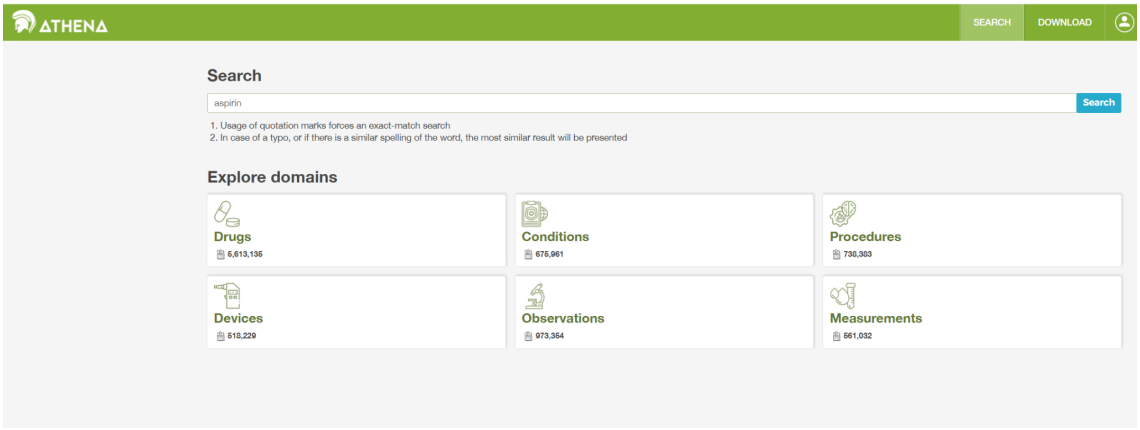


Figura 5.14 Interface do site do ATHENA

A seguir, um breve descritivo das funcionalidades da ferramenta:

A opção SEARCH, permite navegar nos:

- **Domínios**, cada conceito é atribuído a um domínio, que agrupam códigos dos vocabulários, tipo: condições, procedimentos, visitas, dispositivos, medidas, etc. Os domínios também direcionam para qual tabela e campo do CDM OMOP um evento clínico ou atributo de evento deve ser registrado;
- **Conceitos**, representam o significado de cada evento clínico e se subdividem em:
    - Padrão, apenas os conceitos SNOMED são padrão e representam a condição nos dados. Os demais são designados como conceitos não padronizados ou de origem e mapeados para os padrões. Os Conceitos Padrão são indicados por um "S" no campo STANDARD_CONCEPT. E somente esses Conceitos Padrão são usados para registrar dados nos campos do CDM OMOP;

[24] https://athena.ohdsi.org/search-terms/start

    - Não padrão, os conceitos não padronizados não são usados para representar os eventos clínicos, mas ainda fazem parte dos Vocabulários Padronizados e são frequentemente encontrados nos dados de origem. Por essa razão, eles também são chamados de "conceitos de origem". A conversão de conceitos de origem em Conceitos Padrão é um processo chamado "mapeamento";
    - Classificação, esses conceitos não são padrão e, portanto, não podem ser usados para representar os dados. Mas eles estão participando da hierarquia com os Conceitos Padrão e, portanto, podem ser usados para realizar consultas hierárquicas
- **Classes**, alguns vocabulários classificam seus códigos ou conceitos univocamente. Por exemplo, o SNOMED tem 33 dessas classes de conceito, que o SNOMED chama de "tags semânticas": achado clínico, contexto social, estrutura corporal, etc. Estas são divisões verticais dos conceitos. Outros, como MedDRA ou RxNorm, possuem classes de conceito que classificam níveis horizontais em suas hierarquias estratificadas;
- **Vocabulários**, existem 111 vocabulários atualmente suportados pelo OHDSI, dos quais 78 são adotados de fontes externas, enquanto o restante são vocabulários internos do OMOP. Esses vocabulários são normalmente atualizados trimestralmente. A fonte e a versão dos vocabulários são definidas no arquivo de referência VOCABULARY.
- **Validade**, se o conceito ainda está válido ou não para um vocabulário.

Os vocabulários necessários para o CDM OMOP:

Use a opção DOWNLOAD e selecione todos os vocabulários necessários para o CDM OMOP de uso local. Vocabulários com conceitos padrão e uso muito comum são pré-selecionados. Adicione vocabulários que são usados em seus dados de origem. Os vocabulários proprietários não têm botão de seleção. Clique no botão "Licença necessária" para incorporar tal vocabulário em sua lista. O time OHDSI responsável pelos vocabulários entrará em contato e solicitará que justifique a necessidade da sua licença ou ajudará a se conectar com as pessoas certas para obter uma. A Figura 5.15 apresenta o detalhamento de uma seleção de vocabulários.

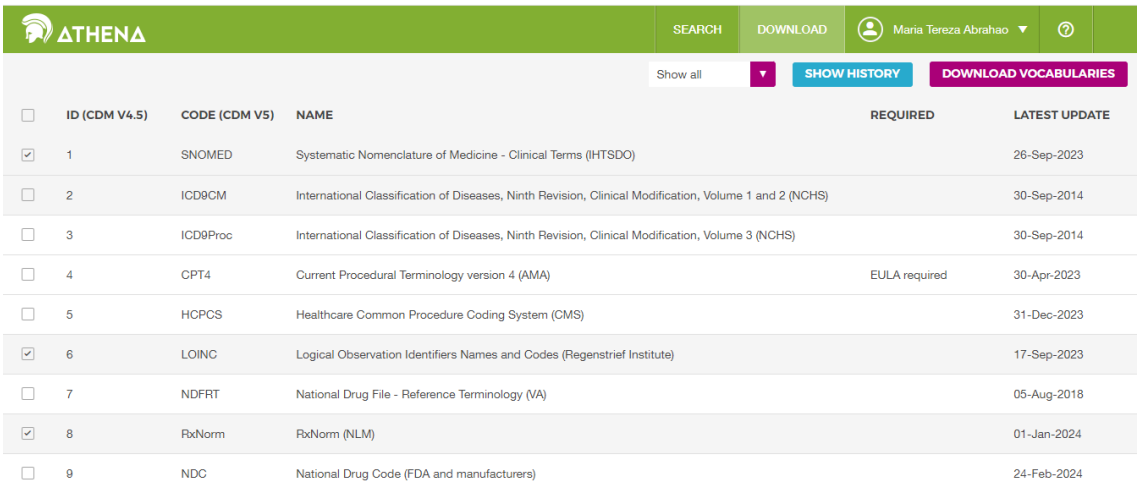


Figura 5.15 Rol de vocabulários selecionados

Ao terminar a seleção, a ferramenta apresenta um sumário dos vocabulários que foram selecionados e solicita um nome para o arquivo. Confirmando, dá-se início ao processo de geração das tabelas (pode demorar horas). No término, é enviado um e-mail de aviso e o arquivo estará pronto. A Figura 5.16 ilustra um exemplo do sumário.

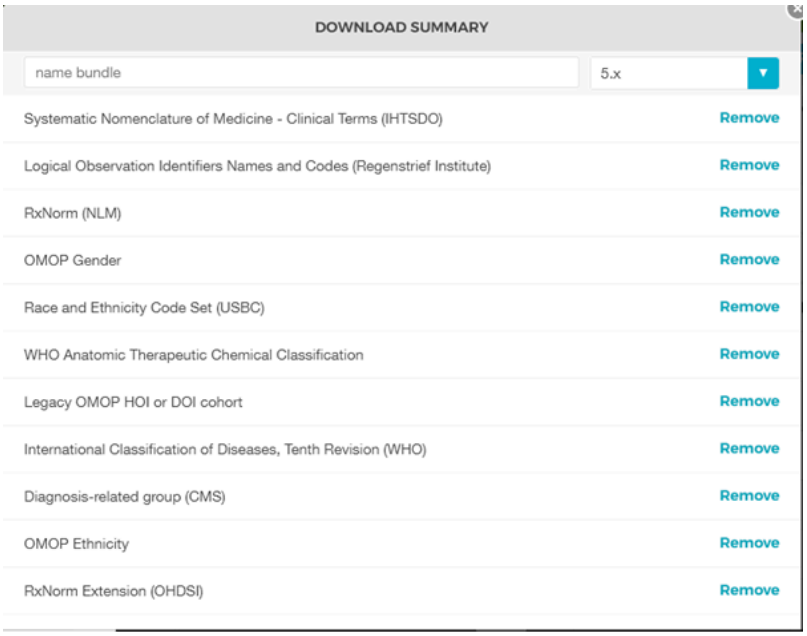


Figura 5.16 Sumário dos vocabulários selecionados

Baixe o arquivo zip com todas as tabelas de vocabulários padronizados e carregue em seu banco CDM local e na ferramenta USAGI.

A continuação, segue um breve resumo:

- Todos os eventos e fatos administrativos são representados nos Vocabulários Padronizados do OMOP como conceitos, relacionamentos de conceito e hierarquia de conceitos ancestral
- A maioria é adotada a partir de esquemas de codificação ou vocabulários existentes, enquanto alguns deles são selecionados de novo pela equipe de vocabulário OHDSI
- Todos os conceitos são atribuídos a um domínio, que controla onde o fato representado pelo conceito é armazenado no CDM OMOP
- Conceitos de significado equivalente em diferentes vocabulários são mapeados para um deles, que é designado por Conceito Padrão. Os outros são conceitos de origem.
- O mapeamento é feito através das relações conceituais "Maps to" e "Maps to value"
- Há uma classe adicional de conceitos chamados conceitos de classificação, que não são padronizados, mas, em contraste com os conceitos de origem, eles participam da hierarquia
- Os conceitos têm um ciclo de vida ao longo do tempo.
- Os conceitos dentro de um domínio são organizados em hierarquias. A qualidade da hierarquia difere entre os domínios, e a conclusão do sistema de hierarquia é uma tarefa contínua.

5.7 Usagi: Mapeando vocabulários locais para o Padrão

Usagi [Schuemie 2023] é uma ferramenta de mapeamento utilizada no contexto da terminologia médica, especialmente em projetos que envolvem a implementação de padrões como SNOMED CT. Desenvolvido pela *International Health Terminology Standards Development Organisation* (IHTSDO), o Usagi auxilia na criação de mapas entre terminologias médicas, facilitando a integração e interoperabilidade entre sistemas de saúde que utilizam diferentes conjuntos de códigos ou terminologias.

O Usagi automatiza parte do processo de mapeamento, sugerindo correspondências entre conceitos de diferentes terminologias com base em suas características semânticas e hierárquicas, garantindo uma integração mais eficiente e precisa entre os sistemas de informação de saúde.

Usando as funções do aplicativo, para importar códigos-fonte para Usagi, primeiro exporte os códigos-fonte do sistema de origem para um arquivo CSV ou Excel (.xlsx). Esse arquivo deve ter pelo menos as colunas contendo o código-fonte e uma descrição do código-fonte em inglês, porém informações adicionais sobre os códigos também podem ser trazidas (descrição no idioma original). Além das informações sobre os códigos-fonte, a frequência do código deve ser preferencialmente trazida, pois isso pode ajudar a priorizar quais códigos devem receber mais esforço no

mapeamento. Se alguma informação do código-fonte precisa ser traduzida para o inglês, o Google Translate pode ser usado.

Os extratos do código-fonte devem ser divididos por domínios (isto é, medicamentos, procedimentos, condições, observações) e não agrupados em um arquivo grande. Os códigos-fonte são carregados no Usagi no menu Arquivo -> Importar códigos. A partir daqui, será exibido “Códigos de importação…”, conforme visto na Figura 5.17.

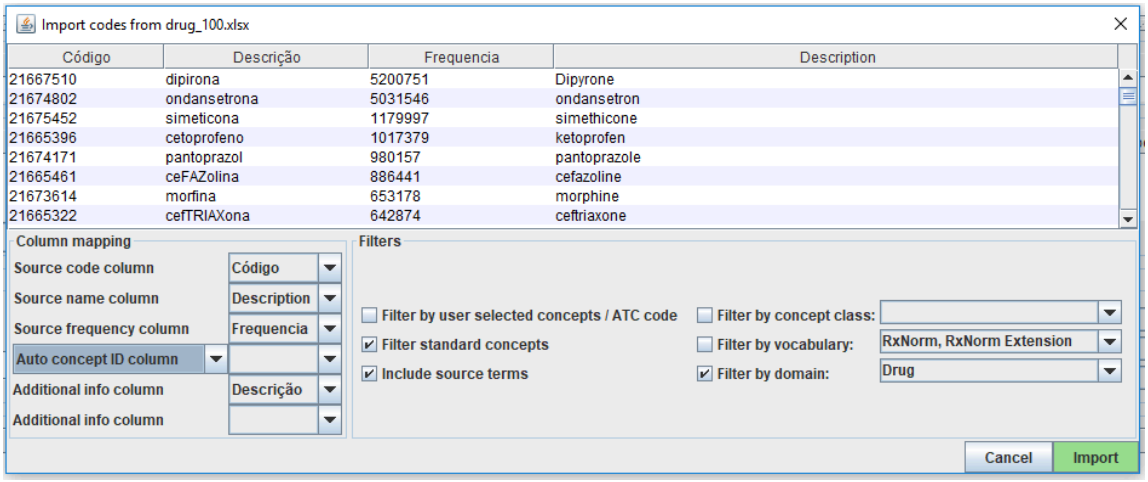


Figura 5.17 Tela de entrada dos códigos-fonte no Usagi

A figura apresenta os códigos (Code), os códigos-fonte que foram traduzidos para o inglês (English term), a frequência e os termos locais que foram adicionados. Usagi irá alavancar as traduções em inglês para mapear para o vocabulário padrão.
Pode-se definir algumas restrições para o Usagi ao mapear. Por padrão, o Usagi mapeia apenas para os conceitos padrão, mas se a opção 'Filtrar conceitos padrão' estiver desligada, o Usagi também irá considerar os conceitos de classificação. Assim que todas as suas configurações forem finalizadas, clique no botão “Importar” para importar o arquivo. A importação do arquivo levará alguns minutos, pois está executando o algoritmo de similaridade para mapear os códigos-fonte para os códigos padrão. Ao concluir a importação dá-se início ao processo de mapeamento. O Usagi é composto por 3 seções principais: uma tabela que permite a visão geral do mapeamento, a seção de mapeamento que foi selecionado e o local para realizar as pesquisas e alterar o mapeamento, se necessário. A Figura 5.18 ilustra essas seções.

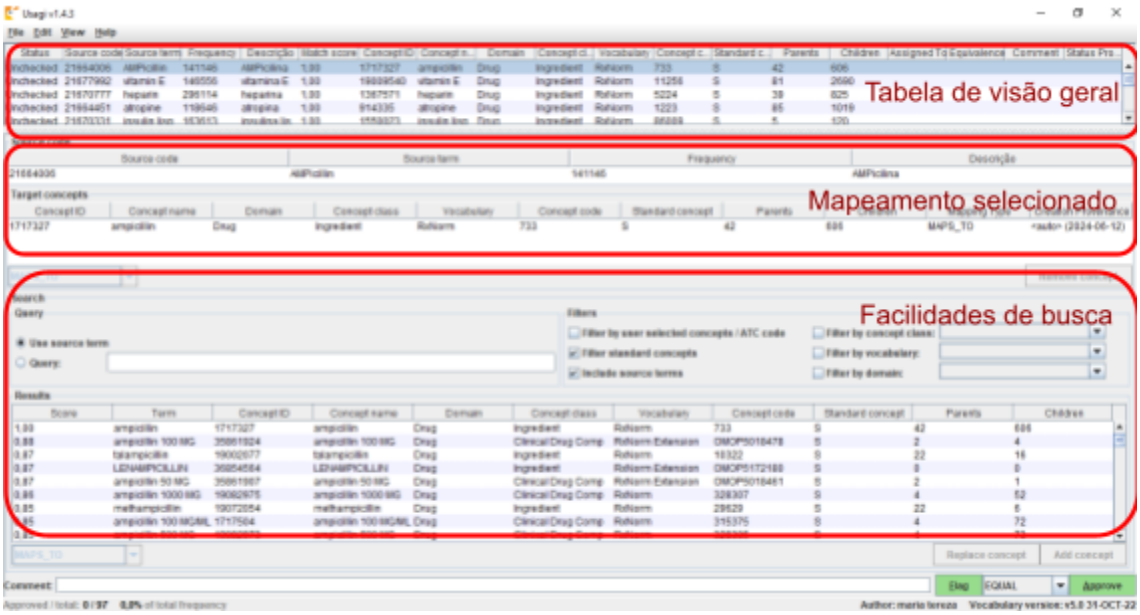


Figura 5.18 Tela de correspondência do Usagi

Logo após a importação dos códigos-fonte, são apresentados os mapeamentos sugeridos gerados automaticamente com base na similaridade de termos e nas opções selecionadas. O Usagi compara as descrições do código-fonte com nomes de conceitos e sinônimos para encontrar a melhor correspondência. Se for selecionado 'Incluir termos de origem', o Usagi considera os nomes e sinônimos de todos os conceitos de origem no vocabulário que mapeiam para um conceito específico. Se o Usagi não puder fazer um mapeamento, ele será mapeado para CONCEPT_ID = 0.

É necessário experiência em sistemas de codificação para mapear os códigos-fonte para seu vocabulário padrão associado. Código por código na aba de visão geral deverá ser visto para aceitar o mapeamento que Usagi sugeriu ou escolher um novo mapeamento.

É possível adicionar comentários aos mapeamentos, que podem ser usados para documentar o porquê uma determinada decisão de mapeamento foi feita.

Melhores Práticas:

- O mapeamento sempre deve ser feito por alguém que tenha experiência com esquemas de codificação.
- Ao clicar no nome de uma coluna, você pode classificar as colunas na aba de visão geral. Pode ser valioso classificar por pontuação (Match score); revisar os códigos nos quais Usagi tem mais confiança primeiro pode eliminar rapidamente um pedaço significativo de códigos. Também é valioso classificar por “Frequência”, gastar mais esforço em códigos frequentes do que em códigos não frequentes.

- É normal mapear alguns códigos para CONCEPT_ID = 0, alguns códigos podem não valer a pena encontrar um bom mapa e outros podem apenas não ter um mapa adequado.
- É importante considerar o contexto de um conceito, especificamente seus ascendentes e descendentes.

Depois de criar um mapa no USAGI, para sua utilização é preciso exportá-lo e anexá-lo à tabela SOURCE_TO_CONCEPT_MAP do Vocabulário OMOP.

Requisitos de instalação:

- Requer Java 1.8 ou superior http://www.java.com
- Obtenha a versão mais recente do vocabulário em Athena http://athena.ohdsi.org/
    - É necessária uma conta
    - Quando seu pacote de vocabulário estiver pronto, baixe-o e descompacte
- Instalação Usagi http://ohdsi.github.io/Usagi/installation.html
- Versão mais recente https://github.com/OHDSI/Usagi/releases/tag/v1.4.3
- Após o download, o Usagi pode ser iniciado simplesmente clicando duas vezes no arquivo jar

Instalação e suporte: Todo o código-fonte e instruções de instalação estão disponíveis no site GitHub da Usagi: https://github.com/OHDSI/Usagi

Configuração única:
Na primeira vez que você iniciar o Usagi, será solicitado que indexe os vocabulários. O Usagi não vem com os índices, você deve fornecer os arquivos de vocabulário. Para fazer isso, siga as seguintes etapas:

1. Obtenha a versão mais recente do vocabulário de Athena. É necessária uma conta. Você pode selecionar qualquer vocabulário que precisar.
2. Quando seu pacote de vocabulário estiver pronto, baixe-o e descompacte o pacote.
3. Acesse o Usagi conforme descrito acima.
4. Quando solicitado, especifique a localização dos arquivos de vocabulário baixados para criar o índice. A criação do índice de vocabulário é um processo caro do ponto de vista computacional e pode levar horas para ser concluído.
5. Ao terminar, a versão do vocabulário deve ser exibida no canto inferior direito do Usagi. As estatísticas do seu índice podem ser visualizadas em Ajuda -> Mostrar estatísticas do índice

5.8 Atlas: consulta, seleção e grupos de conceitos de vocabulários

A OHDSI oferece uma ampla variedade de ferramentas de código aberto para dar suporte a vários casos de uso de análise em dados observacionais no nível do paciente, que permitem interagir com um ou mais bancos de dados usando o modelo comum de dados, o CDM OMOP.

Existem três abordagens principais para a implementação de um estudo. A primeira é escrever código personalizado que não faça uso de nenhuma das ferramentas que a OHDSI tem a oferecer. A segunda abordagem envolve desenvolver a análise em R e fazer uso dos pacotes da Biblioteca de Métodos OHDSI. A terceira abordagem baseia-se na plataforma de análise interativa ATLAS, uma ferramenta que permite que não programadores realizem uma ampla variedade de análises com eficiência. O ATLAS faz uso das Bibliotecas de Métodos, mas fornece uma interface gráfica simples para projetar e executar análises.

O ATLAS é uma ferramenta gratuita, publicamente disponível e baseada na web, desenvolvida pela comunidade OHDSI que facilita o projeto e execução de análises em dados observacionais em nível de paciente, padronizados no modelo CDM OMOP. É a ferramenta usada para fenotipagem baseada em regras ou definição de coorte baseada em regras.

O ATLAS é implantado como um aplicativo da web em combinação com o OHDSI WebAPI. Para o desempenho de análises em tempo real requerer acesso a dados de nível de paciente no CDM. Para executar o Atlas, você deve estar atrás do Firewall de sua instituição.

É a ferramenta principal para a preparação das coortes e montagem dos estudos. Nela temos tópicos que nos permitem escolher os termos que vão formar, junto com outros critérios, a definição da nossa coorte (seleção dos pacientes que participam do estudo). Podemos navegar entre os conceitos, escolher os relacionamentos hierárquicos para incluir ou excluir determinados códigos, agrupar conceitos, exportá-los e importá-los.

À esquerda da ferramenta, temos a barra de navegação mostrando as várias funções fornecidas pelo ATLAS, sendo:

Fontes de dados: Fornecem relatórios padronizados e descritivos de caracterização das fontes de dados configuradas na plataforma Atlas. Esse recurso usa a estratégia de análise em larga escala, pré-computados ao término do mapeamento dos dados para o CDM OMOP.

Pesquisa de vocabulário: Para explorar o vocabulário padronizado OMOP para entender quais conceitos estão disponíveis e como aplicá-los em suas análises.

Conjuntos de Conceitos: Permite criar coleções de expressões lógicas que serão usados nas análises padronizadas. Um conjunto de conceitos é composto por vários conceitos do vocabulário padronizado em combinação com indicadores lógicos que permitem ao usuário incluir ou excluir conceitos relacionados na hierarquia do

vocabulário. Portanto, pesquisar o vocabulário, identificar os conceitos e especificar a lógica a ser usada para resolver um conjunto de conceitos, fornece um mecanismo poderoso para definir a linguagem médica usada nos planos de análise.

Definições de coorte: Para selecionar um conjunto de pessoas que satisfaçam um ou mais critérios por um período e são a base de entrada para todas as análises subsequentes.

Caracterizações: É um recurso analítico que permite examinar uma ou mais coortes e resumir características sobre essas populações de pacientes. Visa, através do uso de estatísticas descritivas, gerar hipóteses sobre os determinantes da saúde e da doença e para compreender os resultados clínicos de grupos específicos da população.

Caminhos de coorte: Possibilita observar a sequência de eventos clínicos que ocorrem em uma ou mais populações.

Taxas de incidência: Para estimar a incidência de desfechos em populações-alvo de interesse.

Perfis: Permite explorar dados observacionais longitudinais de pacientes para resumir o que está acontecendo com um determinado indivíduo.

Estimativa: Analisa os efeitos causais das exposições (por exemplo, intervenções médicas, tais como exposições a medicamentos ou procedimentos) sobre resultados de saúde específicos de interesse (por exemplo, a segurança ou eficácia de medicamentos ou outros tratamentos).

Predição: Realiza a previsão de resultados de saúde futuros a partir de dados existentes ao nível do paciente, a partir da aplicação de algoritmos de aprendizado de máquina, para apoiar a tomada de decisões clínicas, a avaliação de riscos e a validação de tais modelos de previsão.

Jobs: Essa opção é usada para acompanhar o estado dos processos que estão sendo executados. Os trabalhos geralmente são processos de longa execução, como gerar uma coorte ou computar relatórios de caracterização de coorte.

Configuração e Feedback: Ferramentas utilizadas pelo administrador do site.

A Figura 5.19 apresenta a interface da ferramenta Atlas.

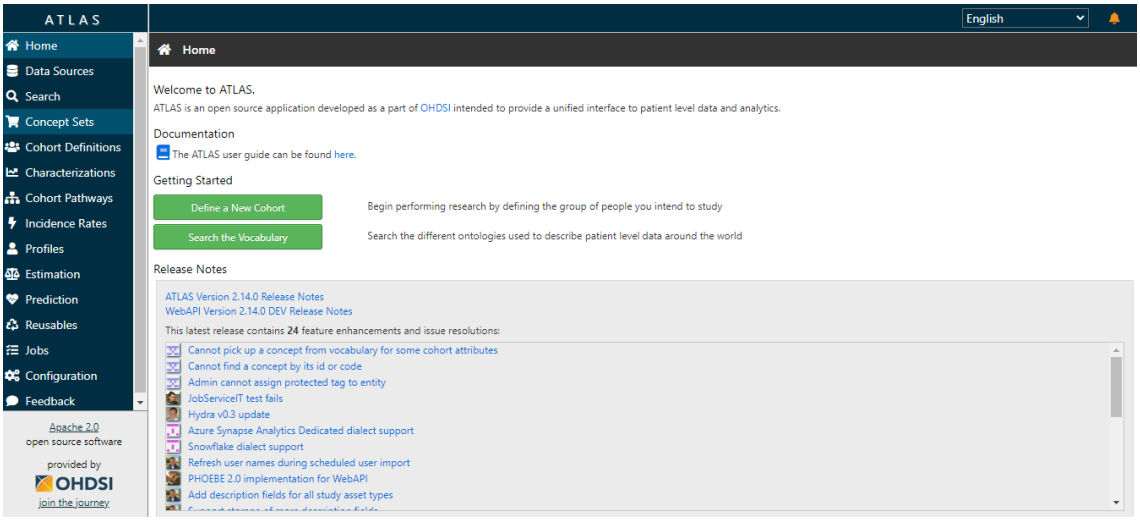


Figura 5.19 Interface web do Atlas

5.8.1 Conjunto de Conceitos

Os Conjuntos de Conceitos no contexto do CDM OMOP, são ferramentas essenciais para criar coleções de expressões lógicas utilizadas em análises padronizadas. Eles permitem a inclusão e exclusão de conceitos relacionados, baseando-se na hierarquia do vocabulário padronizado.

Funcionalidade e Utilidade:

1. Pesquisa do Vocabulário: Permite que os usuários pesquisem no vocabulário padronizado OMOP para encontrar os conceitos relevantes.
2. Identificação de Conceitos: Uma vez encontrados, os conceitos são identificados e selecionados para formar o conjunto de conceitos.
3. Especificação de Lógica: Os usuários podem definir a lógica de inclusão e exclusão, determinando como os conceitos relacionados serão tratados com base em suas relações hierárquicas.

Poder e Flexibilidade:

- Definição Precisa: Ao usar indicadores lógicos, como "incluir" ou "excluir", os usuários podem definir de maneira precisa e flexível os conjuntos de conceitos que serão aplicados em suas análises.
- Hierarquia do Vocabulário: A hierarquia do vocabulário OMOP facilita a inclusão ou exclusão de conceitos relacionados, garantindo que todas as variações relevantes de um termo sejam consideradas ou descartadas conforme necessário.
- Planos de Análise: Fornece um mecanismo poderoso para definir a linguagem médica usada nos planos de análise, permitindo uma padronização e consistência nas pesquisas.

Exemplo de Uso:

- Definição de Coorte: Um pesquisador pode criar um conjunto de conceitos para definir uma coorte de pacientes com diabetes, incluindo conceitos específicos de diabetes tipo 1 e tipo 2, enquanto exclui conceitos relacionados a outras formas de diabetes.
- Construção de Covariáveis: Pode ser usado para construir covariáveis específicas para análises, como incluir apenas medicamentos específicos dentro de uma classe terapêutica.

Benefícios:

- Eficácia na Pesquisa: Facilita a fenotipagem eficiente e a construção de covariáveis, melhorando a qualidade e a consistência dos dados utilizados em pesquisas observacionais.
- Interoperabilidade: A utilização de um vocabulário padronizado garante que diferentes grupos de pesquisa possam colaborar e compartilhar resultados de maneira eficiente.

Os Conjuntos de Conceitos OMOP são, portanto, uma ferramenta fundamental para pesquisadores que buscam gerar evidências robustas e confiáveis a partir de dados observacionais de saúde.

A Figura 5.20 apresenta o Atlas na aba de definições de conceitos, um exemplo de um conjunto de conceitos para COVID 19.

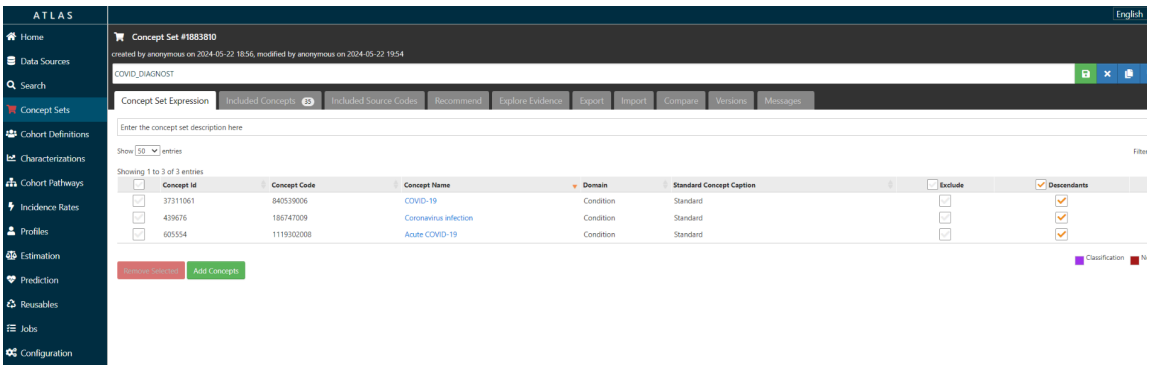


Figura 5.20 Concept set para diagnóstico de COVID 19

5.8.2 Coortes na ferramenta Atlas

As coortes são usadas (e reutilizadas) nas ferramentas de análise OHDSI para definir, por exemplo, as exposições e os resultados de interesse. A estratégia para construir uma coorte dependerá do rigor clínico de como seu consenso de especialistas define a doença. Isso quer dizer que o design de coorte certo dependerá da pergunta que está tentando responder. Pode-se optar por criar uma definição de coorte que use tudo o que

puder obter, use o menor denominador comum para compartilhá-la com outras instituições. Em última análise, fica a critério do pesquisador qual limite de rigor é necessário para estudar adequadamente a coorte de interesse.

Uma definição de coorte é uma tentativa de inferir algo que gostaríamos de observar a partir dos dados registrados. Em geral, a validação de uma definição de coorte baseada em regras ou algoritmo probabilístico, pode ser pensada como um teste da coorte proposta em comparação com alguma forma de referência "padrão ouro" (por exemplo, revisão manual de gráficos de casos).

As coortes parametrizadas vão ser geradas e depois visualizadas em relatórios pré formatados no Atlas. Toda a definição, conceitos e parametrização de uma coorte pode ser exportada no formato JSON para ser importada por outra entidade que faça uso do CDM OMOP. Essa coorte pode ser gerada e analisada em parceiros de estudos e seus resultados podem ser agrupados e analisados em conjunto, sem compartilhamento de dados, somente dos scripts das análises.

A Figura 5.21 apresenta a parametrização de uma coorte para COVID 19 na ferramenta Atlas. A Figura 5.22 mostra um template para exportação dos parâmetros e definições da coorte.

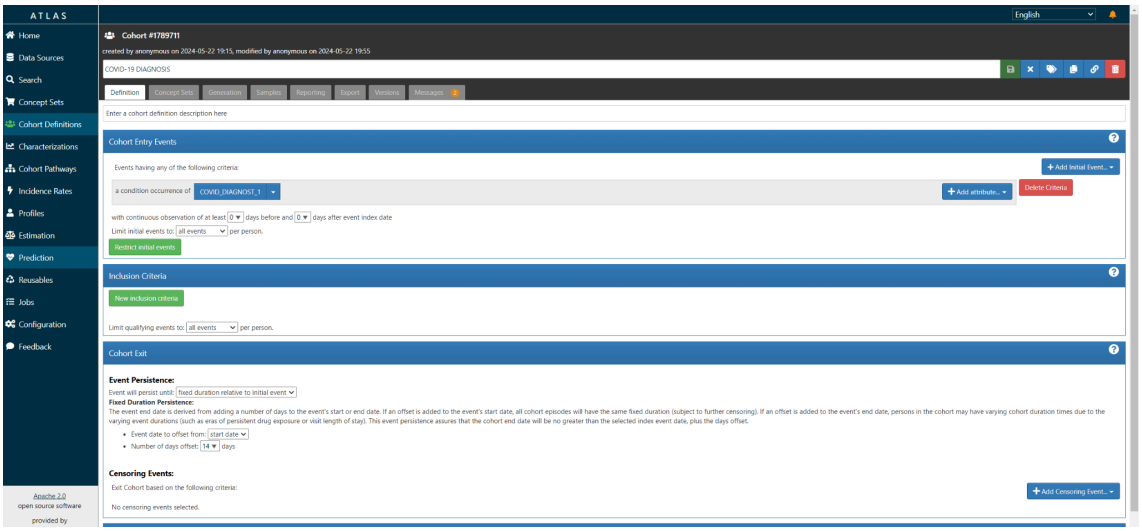


Figura 5.21 Exemplo de uma coorte de COVID 19

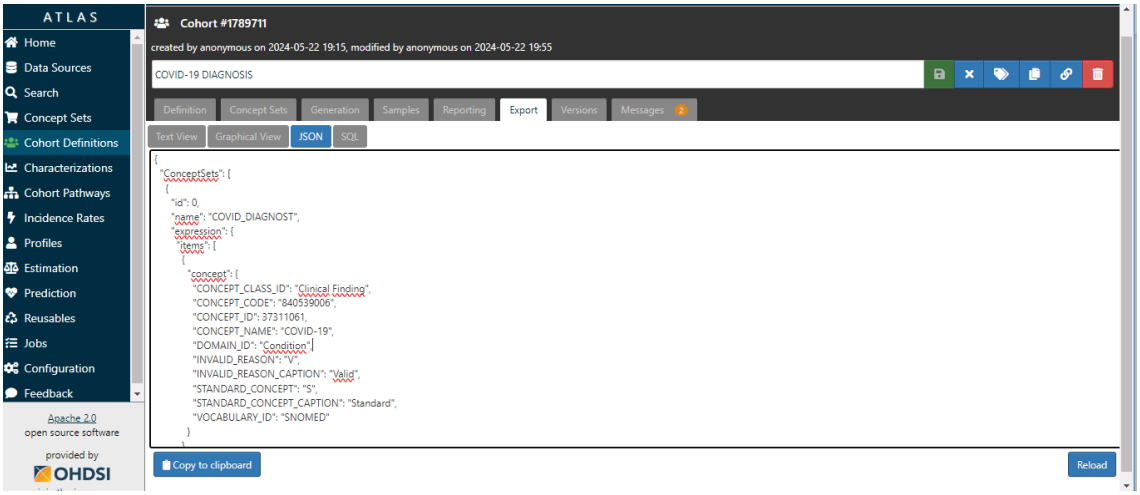


Figura 5.22 Exportação do JSON de uma coorte de COVID 19

Construir uma coorte é o bloco mais importante para responder a uma pergunta de pesquisa nas ferramentas OHDSI. As coortes são a base para realização de estudos de caracterização de populações, estimativas a nível populacional (*Population-Level Estimation* - PLE)[25] e predições a nível de paciente (*Patient-Level Prediction* - PLP)[26] [Abrahão 2022].

5.9 Considerações finais e conclusões

O vocabulário da OHDSI, definido como o conjunto dos códigos padrão (standard codes), cumpre um papel fundamental em limitar as variações possíveis na definição da seleção de pacientes que compõem o estudo (coorte) limitando o conjunto de termos e códigos plausíveis de serem utilizados nela. Com isso, cumpre a função de "*authority file*" dando consistência às análises estatísticas feitas em cima da seleção de pacientes definida pelo fenótipo.

A sua estrutura é de um vocabulário controlado plano, em particular uma classificação estatística mono-hierárquica e exaustiva. É uma lista de códigos e termos, dado que os termos de classificação (*classification codes*) não fazem parte da seleção da coorte, sendo apenas utilizados para pesquisa dos termos padrão pela ferramentas (ATLAS, ATHENA, etc).

O motivo disto é a necessidade de compatibilizar diversos sistemas de codificação de diferentes domínios (diagnósticos, drogas, procedimentos, etc), alguns dos quais ontologias (SNOMED), outros classificações (CID) ou tesauros (RxNorm), muitos dos quais se sobrepõem, complementam ou contradizem e evitar contagens duplicadas.

[25] https://ohdsi.github.io/TheBookOfOhdsi/PopulationLevelEstimation.html
[26] https://ohdsi.github.io/TheBookOfOhdsi/PatientLevelPrediction.html

Os vocabulários são a pedra angular desta padronização. Compreender a sua estrutura e vivenciar o seu uso em situações práticas permite ganhar critérios de discernimento e uma visão crítica na análise de outras ferramentas e de resultados de estudos.

Quais bases estamos comparando? A seleção de termos é condizente com o seu uso no conteúdo das bases? A comparação está semanticamente correta? Houve vieses de seleção pela diferença entre os vocabulários que descrevem os conteúdos das bases usadas no estudo? Estas são perguntas que devem ser sempre feitas ao analisar o resultado de qualquer estudo observacional.

Para ser possível mapear os diversos sistemas de registro de informações médicas com o objetivo de realizar comparações válidas entre fontes de dados diversas, foi necessário sacrificar a expressividade de conhecimento do vocabulário e escolher o mínimo denominador comum.

Como consequência, o trabalho da resolução das ambiguidades, compatibilização das terminologias e mapeamento dos diversos vocabulários entre si, foi deixada na mão de dois agentes: o processo de tradução e mapeamento do vocabulário local do sistema de registro eletrônico para o dicionário padrão durante a carga do banco CDM OMOP, e do próprio pesquisador durante a montagem do fenótipo, quer dizer, da definição dos critérios que vão selecionar a coorte do seu estudo.

Por isso, o mapeamento de vocabulário local para os termos padrão é o mais complexo de todo o processo de criação e carga de um banco CDM OMOP.

Não podemos deixar de mencionar as diversas iniciativas que estão em andamento para melhorar esta situação, compreendendo entre elas, a definição de fenótipos através de um processo probabilístico [Banda 2017] e o mapeamento do dicionário CDM-OMOP para estruturas ontológicas [Callahan, T J et al 2023] que suportem esquemas de inferências e raciocínio lógico na montagem dos fenótipos.

As decisões adotadas no sentido de abranger a maior quantidade possível de codificações e terminologias com uma estrutura plana pode parecer limitante, porém junto com um modelo de dados comum e análises estatísticas padronizadas, está permitindo pela primeira vez, fazer comparações entre fontes de dados diversas com segurança e confiabilidade sem compartilhar dados brutos.

Através dos exemplos do CID, MeSH e outros, vemos a importância de manter compatibilidade e estabilidade para que análises feitas através do tempo se mantenham válidas e consigam mostrar tendências que ultrapassam o período de vida dos sistemas de informação que as geraram. A definição dos vocabulários da OHDSI faz parte essencial deste processo.

A padronização do conteúdo das informações médicas é uma tarefa árdua e trabalhosa, porém, absolutamente vital para poder comparar resultados de estudos realizados em fontes diversas e obter evidências reprodutíveis e confiáveis do mundo real.

Referências bibliográficas

Abrahão, M T., Nobre M R C. e Madril, P J. (2019) "O estado da arte em pesquisa observacional de dados de saúde: A iniciativa OHDSI". In: Artur Ziviani; Natalia Castro Fernandes; Débora Christina Muchaluat Saade. (Org.). Livro de Minicursos do 19o Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS 2019). 19o ed. Porto Alegre: Sociedade Brasileira de Computação-SBC, (2019), I SBN-13 (15) 978-85-7669-472-4, v. 1, p. 141-189.

Abrahão, M T e Madril, P J. (2022) “Fenótipos no contexto da pesquisa observacional: OHDSI Phenotype Phebruary 2022”. In: Sociedade Brasileira de Computação. SBC 2022. (Org.). Minicursos do XXII Simpósio Brasileiro de Computação Aplicada à Saúde. 22 ed.Porto Alegre: SBC, 2022, v. 6, p. 219-261.

Banda, Juan., Halpern, Yoni., Sontag, David and Shah, Nigam. (2017) “Electronic phenotyping with APHRODITE and the Observational Health Sciences and Informatics (OHDSI) data network”. AMIA Summits on Translational Science Proceedings. 2017. 48-57.

Callahan, T J et al. (2023) “Ontologizing Health Systems Data at Scale: Making Translational Discovery a Reality”, https://arxiv.org/abs/2307.05727v2

Cornet, R, and C G Chute. (2016) “Health Concept and Knowledge Management: Twenty-five Years of Evolution.” Yearbook of medical informatics vol. Suppl 1,Suppl 1 S32-41. 2 Aug. 2016, doi: 10.15265/IYS-2016-s037 https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5171511/

Cimino, J J. (1998) “Desiderata for controlled medical vocabularies in the twenty-first century.” Methods of information in medicine vol. 37,4-5: 394-403. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3415631/

DeMars, M M and Perruso C. (2022) "MeSH and text-word search strategies: precision, recall, and their implications for library instruction". J Med Libr Assoc. 2022 Jan 1;110(1):23-33. doi: 10.5195/jmla.2022.1283. PMID: 35210959; PMCID: PMC8830400.

Ferreira, Miguel and Baptista, Ana Alice. (2005) "The use of taxonomies as a way to achieve interoperability and improved resource discovery in DSpace-based repositories". http://repositorium.sdum.uminho.pt/handle/1822/873

Gruber, Thomas R. (1993) "A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition", 5(2):199-220, 1993 https://tomgruber.org/writing/ontolingua-kaj-1993.pdf

Gruber, Thomas R. (1995). "Toward principles for the design of ontologies used for knowledge sharing". Originally in N. Guarino and R. Poli, (Eds.), International Workshop on Formal Ontology, Padova, Italy. Revised August 1993. Published in International Journal of Human-Computer Studies, Volume 43 , Issue 5-6 Nov./Dec. 1995, Pages: 907-928, special issue on the role of formal ontology in information technology. https://tomgruber.org/writing/onto-design.pdf

Heather Hedden. (2022) "The Accidental Taxonomist", Third Edition, Information Today, Inc. Medford, N.J. (November 8, 2022), ISBN: 978-157387-586-8

Karl Fast, Fred Leise and Mike Steckel. M. (2002) “What Is a Controlled Vocabulary?” http://web.archive.org/web/20030811115443/http://www.boxesandarrows.com/archives/what_is_a_controlled_vocabulary.php

Library of Congress (2004) “Library of Congress Subject Headings (LCSH)” 27th edition. Washington (DC). Five volumes. Hardbound. Published annually. ISSN 1048-9711

Garshol L. M. (2004) "Metadata? Thesauri? Taxonomies? Topic Maps!" Ontopia.

Garshol L. M. (1986) “Guidelines for the establishment and development of monolingual thesauri”, International Organization for Standardization (ISO)

Lomax, Jane and Wolf, Elizabeth. (2021) “The Evolution and Importance of Biomedical Ontologies for Scientific Literature” https://www.copyright.com/wp-content/uploads/2021/02/White-Paper-Evolution-Importance-of-Biomedical-Ontologies.pdf

Madsen, B. N. and Erdman Thomsen, H. (2009) “Ontologies vs. Classification Systems”. NEALT (Northern European Association of Language Technology) Proceedings Series, (4), 27-32. http://dspace.utlib.ee/dspace/bitstream/10062/9840/1/13-4-Final-Ontologies_vs_classification_systems_NODALIDA-2009.pdf

Mazzocchi, Fulvio. (2018) “Knowledge organization system (KOS)”. Knowledge Organization 45, no.1: 54-78. Also available in ISKO Encyclopedia of

Knowledge Organization, eds. Birger Hjørland and Claudio Gnoli, https://www.isko.org/cyclo/kos

Mucheroni, Marcos and Modesto, Fernando. (2011) “A interoperabilidade dos sistemas de informação sob o enfoque da análise sintática e semântica de dados na web” DOI: 10.9771/1981-6766rpa.v5i1.3622.

National Information Standards Organization. (2005) “Guidelines for the construction, format, and management of monolingual controlled vocabularies” (ANSI/NISO Z39.19-2005). Bethesda, MD: NISO Press. https://groups.niso.org/higherlogic/ws/public/download/12591/z39-19-2005r2010.pdf

Pagel, M. "Q&A. What is human language, when did it evolve and why should we care?" BMC Biol 15, 64 (2017). https://doi.org/10.1186/s12915-017-0405-3 https://bmcbiol.biomedcentral.com/articles/10.1186/s12915-017-0405-3#citeas

Schuemie M. (2023) “Usagi” https://github.com/OHDSI/Usagi

Souza, Renato Rocha, Douglas Tudhope and Mauricio B. Almeida (2012). "Towards a Taxonomy of KOS: Dimensions for Classifying Knowledge Organization Systems". Knowledge Organization 39, no. 3: 179–192.

Stocks, K.I., Neiswender, C., Isenor, A.W., Graybeal, J., Galbraith, N., Montgomery, E.T., Alexander, P., Watson, S., Bermudez, L., Gale, A., Hogrefe, K. (2010) "The MMI Guides: Navigating the World of Marine Metadata", https://uop.whoi.edu/techdocs/presentations/MMI_Guides.pdf

