Universidade do Minho

INTEROPERABILIDADE DE CATÁLOGO DE METADADOS

-Projeto de LCC-

Realizado por:

- Eduardo Soares (a95917)
- Hugo Ferreira (a100082)
- Marco Silva (a97554)

Índice

Introdução

A gestão eficaz de dados implica não apenas o seu armazenamento, mas também a organização dos respetivos metadados. Ferramentas como o CKAN e o OpenMetadata são amplamente utilizadas para catalogar e descrever conjuntos de dados, embora utilizem modelos distintos de representação. Esta diversidade dificulta a interoperabilidade entre sistemas.

Este trabalho propõe uma solução para facilitar a troca de metadados entre CKAN e OpenMetadata. A abordagem implementada visa preservar a integridade da informação durante o processo de conversão, garantindo uma integração eficiente e sem perda de dados.

Metadados e Interoperabilidade

Metadados são dados sobre dados. Desempenham um papel essencial na organização, descoberta, compreensão e reutilização de conjuntos de dados, descrevendo aspectos como a sua origem, estrutura, conteúdo, responsáveis e condições de utilização. Em ambientes complexos e distribuídos, os metadados tornam-se indispensáveis para garantir a governança da informação e facilitar a interoperabilidade entre sistemas.

A interoperabilidade, neste contexto, refere-se à capacidade de diferentes plataformas de gestão de dados partilharem, compreenderem e reutilizarem metadados de forma automatizada e sem perda de significado. Para tal, é fundamental o uso de vocabulários comuns e normas padronizadas, que permitam alinhar modelos internos distintos.

CKAN

O CKAN (Comprehensive Knowledge Archive Network) é uma plataforma open source amplamente utilizada para a gestão e publicação de dados abertos. Desenvolvido inicialmente pela Open Knowledge Foundation, o CKAN é hoje adotado por governos, organizações públicas e privadas em todo o mundo, servindo como infraestrutura central para portais de dados abertos, como o data.gov nos Estados Unidos e o dados.gov.pt em Portugal.

A plataforma permite organizar, catalogar e disponibilizar conjuntos de dados de forma estruturada, através de uma interface web e de uma API RESTful. Os seus principais componentes incluem organizações, recursos e metadados associados, sendo estes últimos essenciais para a descoberta e reutilização dos dados. O modelo de metadados do CKAN é relativamente flexível, permitindo a adição de campos personalizados (extras), o que facilita a sua adaptação a diferentes domínios e requisitos.

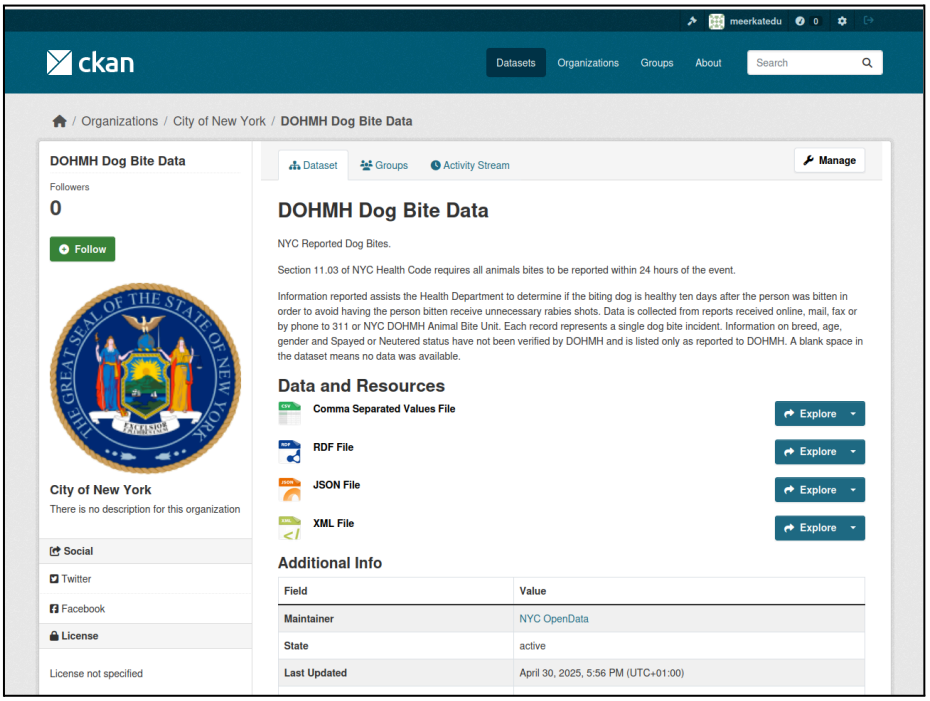

Figura 1 - CKAN

A API do CKAN suporta operações como criação, atualização, pesquisa e eliminação de datasets e recursos, sendo uma ferramenta fundamental para automatizar a gestão de grandes volumes de informação. No contexto deste projeto, o CKAN representa o ponto de origem ou destino de metadados que necessitam de ser convertidos para um formato interoperável, compatível com outras plataformas, como o OpenMetadata.

DCAT

O DCAT (Data Catalog Vocabulary) é um vocabulário RDF desenvolvido e recomendado pelo W3C(World Wide Web Consortium) com o objetivo de facilitar a interoperabilidade entre catálogos de dados na web. A sua função principal é fornecer um modelo comum para descrever conjuntos de dados, promovendo a partilha e reutilização de informação entre diferentes sistemas, organizações e países.

Este vocabulário define classes e propriedades para representar elementos como catálogos, datasets, distribuições, agentes responsáveis, palavras-chave, datas de publicação e outros aspetos relevantes. A sua estrutura modular permite também a extensão do modelo base com propriedades específicas para contextos particulares, sem comprometer a compatibilidade com implementações “standard”.


```
{
    "@context": {
      "dcat": "http://www.w3.org/ns/dcat#",
      "dct": "http://purl.org/dc/terms/",
      "xsd": "http://www.w3.org/2001/XMLSchema#"
    },
    "@graph": [
      {
        "@id": "http://my-ckan-site.org/dataset/425e361b-bad9-4a8f-8cc4-2e147c4e8c18",
        "@type": "dcat:Dataset",
        "dcat:distribution": {
          "@id": "http://my-ckan-site.org/dataset/425e361b-bad9-4a8f-8cc4-2e147c4e8c18/resource/df0fc449-fddf-41af-910a-f972b458956c"
        },
        "dct:identifier": "425e361b-bad9-4a8f-8cc4-2e147c4e8c18",
        "dct:temporal": {
          "@id": "_:N1c32ba52ad1641d086101a4a4bcbe8a5"
        },
        "dct:title": "An example CKAN dataset"
      },
      {
        "@id": "_:N1c32ba52ad1641d086101a4a4bcbe8a5",
        "@type": "dct:PeriodOfTime",
        "dcat:endDate": {
          "@type": "xsd:date",
          "@value": "2024-12-31"
        },
        "dcat:startDate": {
          "@type": "xsd:date",
          "@value": "2024-01-01"
        }
      },
      {
        "@id": "http://my-ckan-site.org/dataset/425e361b-bad9-4a8f-8cc4-2e147c4e8c18/resource/df0fc449-fddf-41af-910a-f972b458956c",
        "@type": "dcat:Distribution",
        "dcat:accessURL": {
          "@id": "http://my-ckan-site.org/dataset/425e361b-bad9-4a8f-8cc4-2e147c4e8c18/resource/df0fc449-fddf-41af-910a-f972b458956c/download/data.csv"
        },
        "dct:format": "CSV",
        "dct:title": "Some data in CSV format"
      }
    ]
}
```


Figura 2 - DCAT

O DCAT é amplamente utilizado em iniciativas nacionais e internacionais de dados abertos, como o data.europa.eu, e serve como base para perfis especializados como o DCAT-AP (Application Profile for European Data Portals). No contexto deste projeto, o DCAT é adotado como referência ao formato intermédio, permitindo a conversão bidirecional de metadados entre CKAN e OpenMetadata de forma estruturada, normalizada e sem perda de informação.

OpenMetadata

O OpenMetadata é uma plataforma open source de gestão e governação de metadados concebida para integrar e organizar ativos de dados provenientes de múltiplas fontes, como bases de dados, pipelines de dados, dashboards, entre outros. Destaca-se pela sua arquitetura moderna, altamente extensível, e pelo foco na automatização de processos de catalogação e observabilidade dos dados.

A plataforma adota um modelo de entidades fortemente tipificado, onde cada ativo (como tabelas, esquemas, serviços ou utilizadores) é representado de forma explícita e relacionável. Além disso, o OpenMetadata oferece uma API RESTful robusta, o que permite a sua integração com outras ferramentas e o desenvolvimento de funcionalidades automatizadas, como ingestão de metadados, classificação de dados e cálculo de estatísticas de uso.

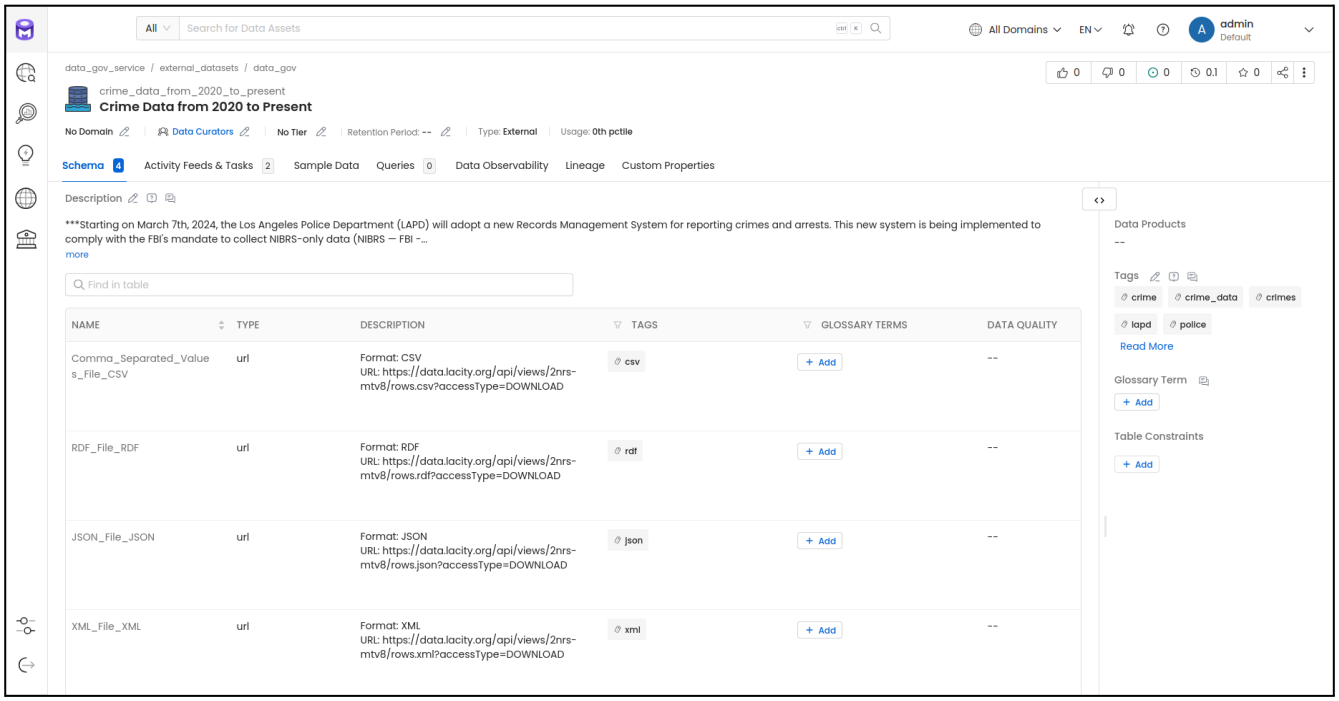

Figura 3 - OpenMetadata

No âmbito deste projeto, o OpenMetadata é tratado como um dos sistemas-alvo para interoperabilidade. Dada a sua estrutura mais rica e semântica em comparação com o CKAN, foi necessário adotar um modelo intermédio capaz de acomodar os metadados provenientes do OpenMetadata sem perda de informação. Foram utilizadas extensões do vocabulário DCAT para assegurar a conservação de campos específicos, como estatísticas de uso, perfis de colunas e relações com serviços.

Estratégia de Interoperabilidade

A interoperabilidade entre plataformas de catalogação de dados com modelos internos distintos requer uma abordagem que permita a troca de metadados sem comprometer a sua integridade, semântica e estrutura. Neste projeto, a interoperabilidade entre o CKAN e o OpenMetadata foi concretizada através do modelo intermédio DCAT.

A utilização do DCAT como formato *pivot* permitiu evitar a necessidade de criar conversões diretas e específicas entre os dois sistemas. Em vez disso, foram desenvolvidas funções de conversão bidirecional: CKAN <-> DCAT e OpenMetadata <-> DCAT. Esta abordagem modular traz duas grandes vantagens: por um lado, simplifica a arquitetura da solução e, por outro, permite escalabilidade futura para incluir outras plataformas compatíveis com o DCAT.

A interoperabilidade foi operacionalizada através do desenvolvimento de scripts Python, com recurso às APIs de ambas as plataformas. Estes scripts permitem extrair metadados, convertê-los para DCAT, e posteriormente reimportar para o sistema de destino, garantindo um fluxo de informação coeso, reversível.

Implementação da Solução

A solução proposta foi implementada em Python, com base numa arquitetura modular composta por quatro componentes principais: extração de metadados, conversão para DCAT, reconstrução no sistema de destino e controlo de fluxo. Cada componente corresponde a um conjunto de funções que interagem com as APIs das plataformas CKAN e OpenMetadata.

A extração de metadados foi realizada através da API REST do CKAN (/api/3/action/) e da API do OpenMetadata (/api/v1/). Para o CKAN, os metadados foram recolhidos por organização e por dataset, incluindo informações como título, descrição, recursos associados, autor, tags e campos personalizados. No caso do OpenMetadata, foram recolhidos por serviço, base de dados, esquema e tabela para obter os detalhes completos de cada tabela, incluindo colunas, tipo de tabela e tags.

A conversão para o modelo DCAT foi feita com base no vocabulário publicado pelo W3C. Foram definidos mapeamentos explícitos entre os campos dos sistemas fonte e as propriedades do DCAT. Estes mapeamentos foram documentados de forma sistemática para garantir rastreabilidade e coerência.

Na fase de reconstrução, os objetos DCAT foram reinterpretados e transformados novamente em estruturas compatíveis com os modelos de dados dos sistemas de destino. A criação e atualização de datasets no CKAN foi feita através de endpoints “package_create” e “package_update”. No OpenMetadata, a criação de tabelas exigiu a criação ou obtenção caso exista do serviço,base de dados e esquema, seguindo-se a chamada de um post ao endpoint de criação de entidades ”tables”.

A coordenação de todo o fluxo foi realizada por um script principal (main.py), que oferece ao utilizador um menu interativo com quatro opções principais:

- Syncronize ckan with openmetadata (import from openmetadata to ckan);
- Syncronize openmetadata with ckan (import from ckan to openmetadata);
- Update datasets to ckan from datagov;
- Update datasets to openmetadata from datagov.

Problemas na solução

- A estrutura do código está desorganizada e há repetição de código;
- Na extração de um dataset do openmetadata, ou seja, de uma tabela, há perda de informação, inclusive na organização a quem pertence o dataset, como meio de obter o melhor do pior, ao exportar para o ckan, é criada uma organização chamada “openmetadatadatasets”, para além disso perdemos argumentos como license e publisher.

Conclusão

A interoperabilidade entre plataformas de catalogação de metadados é um desafio cada vez mais relevante num contexto em que os dados são produzidos, armazenados e geridos por múltiplos sistemas heterogéneos. Este projeto demonstrou, de forma prática, que é possível estabelecer um fluxo bidirecional de conversão de metadados entre o CKAN e o OpenMetadata através da utilização do DCAT como formato intermédio.

A solução desenvolvida assegura, na maior parte, a integridade e a fidelidade dos metadados durante o processo de conversão, através do modelo DCAT. A abordagem modular e extensível adotada permite, não só a integração entre os dois sistemas em questão, mas também a futura expansão para outras plataformas compatíveis com o mesmo vocabulário.

Em suma, este projeto apresenta uma boa base para promover a interoperabilidade semântica entre sistemas de gestão de dados e contribui para um ecossistema de dados mais coeso, acessível e reutilizável.

Bibliografia

- CKAN: https://docs.ckan.org/
- DCAT: https://www.w3.org/TR/vocab-dcat-2/
- CKAN extensions: https://extensions.ckan.org/
- OpenMetadata: https://docs.open-metadata.org/
- DATA.GOV: https://data.gov/
- Docker: https://www.docker.com/