Uma Proposta baseada no Dublin Core para Catalogação de Metadados de Fontes de Dados Governamentais

Hudson A. B. da Silva[1], João V. dos Santos[2], José Eduardo M. Jochem[3], Ana Fleck[4] Ronaldo dos Santos Mello[5], Carina Friedrich Dorneles[6], Renato Fileto[7]

[1]Departamento de Informática e Estatística - UFSC - Florianópolis - SC - Brasil

hudson.silva@ifpa.edu.br,santosjoao301@gmail.com,
joseeduardomj@gmail.com,anakfleck@gmail.com,
r.mello@ufsc.br,carina.dorneles@ufsc.br,r.fileto@ufsc.br

Abstract. *This work emphasizes the application of the Dublin Core (DC) standard to facilitate the retrieval, sharing, information extraction, and metric analysis of Brazilian governmental data sources. A total of 31 useful metadata elements were identified, analyzed, selected, and specified for the cataloging and interoperability of datasets. These elements were divided into two levels of granularity: general and specific. Unlike related works, this proposal extends 11 standard DC elements with 20 additional descriptors specific to the context of metadata from Brazilian governmental data sources and compares current metadata management tools.*

Resumo. *Este trabalho enfatiza a aplicação do padrão Dublin Core visando facilitar consulta, compartilhamento, bem como extração de informações e análise de métricas de fontes de dados governamentais brasileiras. Foram identificados, analisados, selecionados e especificados 31 elementos de metadados úteis para catalogação e interoperabilidade de conjuntos de dados. Estes elementos foram divididos em dois níveis de granularidade: geral e específica. Diferente de trabalhos relacionados, esta proposta estende 11 elementos padrão DC com mais 20 descritores específicos para aplicação no contexto de metadados de fontes de dados governamentais brasileiras e compara ferramentas de gerenciamento de metadados atuais.*

1. Introdução

A crescente disponibilidade de fontes de dados governamentais abertos no Brasil representa uma oportunidade única para a análise de dados e auxílio na tomada de decisões por órgãos públicos. Entretanto, a falta de padronização e gestão eficiente de seus metadados pode limitar seu potencial. Neste contexto, o projeto *Céos*[1], uma iniciativa em parceria com o Ministério Público, visa combater a corrupção em licitações utilizando inteligência artificial, e enfrenta esses desafios na gestão de seus conjuntos de dados. O projeto lida com um grande volume de dados oriundos de diversos órgãos de controle e fiscalização, sendo assim necessário catalogar seus metadados de forma eficiente para permitir, por exemplo, análises de proveniência e qualidade dos dados coletados.

A efetiva aplicação de um padrão de metadados facilita a uniformidade e a interoperabilidade, permitindo uma gestão eficaz dos dados, pois define um conjunto de diretrizes predefinidas que ditam sua estrutura e formato, garantindo consistência na descrição

[1]*https://ceos.ufsc.br/*

e gerenciamento dos dados. No contexto deste trabalho, o padrão *Dublin Core (DC)* foi escolhido pela sua simplicidade e ampla adoção [Core 2024]. Entretanto, é crucial investigar sua adequação ao contexto específico de dados governamentais brasileiros.

Desta forma, esse trabalho apresenta uma proposta de catalogação de metadados de fontes de dados governamentais brasileiras baseada no padrão DC, ou seja, uma proposta de extensão do DC com elementos específicos do domínio governamental. Esta proposta contribui para o desenvolvimento de uma solução de gestão de metadados dentro do projeto *Céos*, solução esta que esteja disponível à comunidade em geral e forneça *insights* valiosos sobre dados em contextos governamentais, impulsionando a inovação, a transparência e a participação cidadã.

O artigo está organizado em mais 4 seções. A Seção 2 discute trabalhos relacionados. A Seção 3 apresenta a proposta e a Seção 4 analisa ferramentas de gestão de metadados. Por fim, a Seção 5 é dedicada à conclusão.

2. Trabalhos Relacionados

No Brasil, o *Comitê Executivo de Governo Eletrônico (CEGE)* definiu um padrão de metadados [CEGE 2014] derivado dos 15 elementos do DC mais 5 elementos para o domínio governamental brasileiro. No entanto, esse padrão serve para descrever recursos informacionais e serviços específicos disponíveis na Internet. O trabalho de [Maali et al. 2010], por sua vez, avalia a disponibilidade e confiabilidade de 7 catálogos de metadados de domínio governamental de 5 países com base em testes de integridade e consistência. A partir desta avaliação, os autores propõem um vocabulário para a catalogação de metadados derivado de vocabulários existentes, como o DC.

Ainda no contexto de trabalhos que consideram o padrão DC, o trabalho de [da Silva et al. 2019] propõe uma comparação do *DCMES (DC Metadata Element Set)*, do *DCTERMS (DC Terms)* e do *DCAP (DC Application Profile)* com o objetivo de selecionar descritores DC que melhor atendem às necessidades de um projeto de pesquisa. A escolha foi pelos descritores DCTERMS, que foram considerados na extensão de uma ferramenta que captura descritores de metadados e pontua cada um deles. O trabalho de [Koshman 2010] apresenta uma descrição adaptada de elementos base do DC com foco em conjuntos de dados. Ele também propõe um modelo que aprimora a recuperação e visualização de conjuntos de dados, promovendo um contexto de metadados, com foco em parametrizar e comparar informações de conjuntos de dados.

Por fim, o trabalho de [Alasem 2009] propõe um processo para definição de metadados governamentais voltado a alguns países membros da *Commonwealth*. Ele é dividido em 4 fases: *(i)* estabelecimento de um grupo de trabalho de metadados; *(ii)* identificação de requisitos dos provedores de recursos governamentais, necessidades dos usuários e recursos governamentais a serem descritos por metadados; *(iii)* estudo de *Websites* governamentais e; *(iv)* determinação dos elementos de metadados apropriados.

Diferente desses trabalhos relacionados, esta proposta estende o padrão DC com 20 descritores adicionais que caracterizam fontes de dados quanto à qualidade, disponibilidade, modelos heterogêneos, versionamento, integração e proveniência. A única proposta que também lida com dados governamentais brasileiros (CEGE) define padrões de metadados para fontes de dados específicas. Além disso, esta proposta compara ferramentas de catalogação de metadados atuais.

3. Proposta

O projeto *Céos* manipula diversas fontes de dados de órgãos governamentais diferentes. O DC foi escolhido por apresentar uma semântica de fácil entendimento quando comparado com outros modelos de catalogação de metadados, como o *MARC21* [Alves and Souza 2007], além de contar com a ISO 15836-1:2017 [Alasem 2009], sendo assim um padrão internacional mantido pela *Dublin Core Metadata Initiative (DCMI)*. Dentre os princípios da DCMI, destacam-se a neutralidade e o foco interdisciplinar, garantindo que os padrões de metadados sejam acessíveis e adaptáveis a diferentes propósitos e tecnologias [DCMI 2024].

O DC possui 15 elementos básicos de descrição de metadados [DCMI 2024] [Weibel 1995] e este trabalho considerou 11 deles. Entretanto, para suprir as necessidades do domínio de fontes de dados governamentais no cenário brasileiro, após uma análise de diversas fontes de dados consideradas pelo projeto, 20 elementos foram definidos (em negrito na Figura 1), resultando em 31 elementos, conforme ilustra a Figura 1.

Elementos Catalogados	Granularidade Conjunto de Dados	Granularidade Tabela	Elementos Catalogados	Granularidade Conjunto de Dados	Granularidade Tabela	Elementos Catalogados	Granularidade Conjunto de Dados	Granularidade Tabela
Versão	X	X	Formato	X	X	Data de Criação		X
Direitos	X	X	**Tamanho**	X	X	**Data de Modificação**	X	X
Modelo de Dados	X	X	**Referências**		X	Criador	X	
Identificador	X	X	**Colunas**		X	Fonte	X	
Contribuidor	X		**Amostra de Dados**		X	**Licença**	X	
Título	X		**Data de Submissão**	X	X	**Utilizador**	X	X
Resumo	X	X	**Completude**	X	X	**Repositórios de Código**		X
Tema	X	X	**Limitações**	X	X	**Fóruns de Discussão**	X	
Relação	X	X	**Validade**	X	X	**Confiabilidade**	X	X
Tratamento de dados ausentes		X	**Periodicidade de Atualização**	X	X	**Dados Semelhantes de Outras Fontes**	X	
			Ferramentas de Análise		X			

Figure 1. Elementos X Granularidade —— Fonte: Do autor.

Os elementos propostos estão classificados em 2 níveis de granularidade (*específica* ou *geral*) conforme a sua ocorrência mais frequente em um desses níveis nas fontes de dados do domínio. Esses níveis consideram as fontes de dados (ou conjuntos de dados) como sendo bancos de dados (BDs) relacionais. A *granularidade específica* é a catalogação de metadados para tabelas relacionais específicas, como orçamentos públicos, licitações, contratos e receitas tributárias que pertencem a um conjunto de dados. A *granularidade geral* é o conjunto de dados (uma ou mais tabelas) oriundo de órgãos governamentais, como dados de portais de transparência de um município ou o IBGE. Os elementos são detalhados nas seções a seguir.

3.1. Elementos de ambas as Granularidades

O elemento **Versão** considera que um recurso pode ser versionado. Ele permite a rastreabilidade de modificações que ocorreram em ambos. Ele possui granularidade dupla (conjunto de dados e tabela), pois os conjuntos de dados e tabelas estão em constante atualização. O elemento **Direitos** descreve os direitos de acesso sobre um recurso. Ele permite o mapeamento de equipes que podem usufruir de um determinado conjunto de dados ou tabela. Ele possui granularidade dupla, pois o projeto conta com vários times de pesquisa e times podem utilizar dados de tabelas diferentes do mesmo conjunto de dados.

O elemento **Modelo de Dados** detalha a forma de persistência dos dados, como BD relacional, documento ou grafo. Ele estabelece o modelo de dados utilizado para o recurso. Ele possui dupla granularidade pois as informações do conjunto de dados como um todo ou de tabelas são persistidas em um certo modelo de dados.

O elemento **Identificador** identifica formalmente um recurso. Ele permite evitar qualquer ambiguidade relacionado aos nomes das tabelas e conjuntos de dados. Ele possui as duas granularidades, pois o projeto possui muitas tabelas em cada conjunto de dados, e vários dos mesmos. O descritor **Resumo** provê uma descrição breve sobre o conteúdo do recurso. Ele possui granularidade dupla, pois é importante ter informações mais detalhadas sobre tabelas e conjuntos de dados. Ainda, o descritor **Tema** indica o tema do recurso. Ele permite refinar as informações do conjunto de dados e tabelas em grupos padrões. Ele possui granularidade dupla, pois os conjuntos de dados e tabelas possuem características que podem ser englobadas em termos semelhantes. O descritor **Relação** mostra sobre relacionamentos entre recursos. Ele permite identificar a relação entre conjuntos de dados, tabelas, e entre conjuntos de dados e tabelas, por isso a necessidade de granularidade dupla.

O elemento **Data de submissão** indica a data de inserção do recurso no BD. Ele permite avaliar características temporais do conjunto de dados ou tabela. Ele possui granularidade dupla, pois ambas são inseridas em um banco de dados. O elemento **Data de modificação** indica a última data em que o recurso foi modificado. Ele permite avaliar características temporais do conjunto de dados ou tabela. Ele possui granularidade dupla, pois tabelas precisam ser atualizadas constantemente e, por conseguinte, os conjuntos de dados também.

O elemento **Formato** indica o formato do recurso. Ele possui granularidade dupla, pois conjuntos de dados e tabelas podem ter formatos de arquivos diferentes. O elemento **Tamanho** indica o tamanho do recurso. Ele permite conhecer o tamanho do conjunto de dados ou Tabela. Por esse motivo possui granularidade dupla.

O elemento **Completude** mostra a porcentagem de preenchimento geral dos campos em relação aos nulos. Ele permite *insights* sobre a qualidade dos dados. Ele possui granularidade dupla, pois as tabelas podem ter campos nulos, e por conseguinte os conjuntos de dados podem possuir tabelas com campos nulos. O elemento **Validade** indica a validade de um recurso, relacionado à diferença entre a periodicidade e a data de modificação do recurso. Ele permite saber se os dados de uma tabela são confiáveis diante a atualidade dos mesmos. Ele possui granularidade dupla, pois uma tabela fora da validade implica em um conjunto de dados fora da validade. Ainda, o elemento **Periodicidade de Atualização** indica o ciclo de tempo que o recurso precisa ser atualizado. Ele permite a constante atualização de um recurso, influenciando na sua confiabilidade. Ele possui granularidade dupla, pois as tabelas precisam estar em constante atualização, devido à novas etapas de tratamento e inserção de dados e, por conseguinte, os conjuntos de dados que possuem essas tabelas também.

O descritor **Limitações** mostra qualquer limitação que um recurso possa ter. Ele permite que as equipes saibam das restrições de um recurso e possam trabalhar nelas. Ele possui granularidade dupla, pois conjuntos de dados e tabelas podem apresentar limitações. O descritor **Utilizador** indica as equipes que utilizam o recurso nas linhas

de pesquisas. Ele permite a identificação de equipes relacionadas a um recurso a fim de tirar dúvidas, estabelecer um possível contato mais pessoal e mapear as mesmas. Ele possui granularidade dupla, pois as equipes podem trabalhar com os conjuntos de dados, tabelas ou ambos.

Por fim, o elemento **Confiabilidade** informa o grau de confiabilidade do recurso. Ele possui granularidade dupla, pois as tabelas podem ter certo grau de confiança e, por conseguinte, os conjuntos de dados também.

3.2. Elementos de Granularidade Conjunto de Dados

O elemento **Contribuidor** indica o órgão que disponibiliza o conjunto de dados. Ele pertence apenas aos conjuntos de dados. O elemento **Título** é o título do conjunto de dados. Ele permite identificar o mesmo de forma textual. Ele possui granularidade geral, pois há necessidade de descrever textualmente o conjunto de dados, uma vez que as tabelas já possuem um nome que as definem.

O descritor **Criador** mostra o criador do conjunto de dados. Ele permite saber a pessoa ou departamento responsável pela criação do conjunto de dados. O descritor **Fonte** mostra a origem do conjunto de dados. Ele permite localizar geograficamente onde o conjunto de dados foi coletado. Por esse motivo possui granularidade geral. O descritor **Licença** mostra a licença que rege o uso do conjunto de dados. Ele permite identificar licenças específicas de uso. Por esse motivo possui granularidade geral.

O elemento **Fórum de Discussão** contém o *link* para o fórum de dúvidas sobre o conjunto de dados. Ele permite um canal centralizado onde as equipes podem se comunicar e resolver incertezas sobre um conjunto de dados. Ele possui granularidade geral, pois engloba dúvidas referentes ao conjunto de dado e seu conteúdo. Por fim, o descritor **Dados Semelhantes de Outras Fontes** indica outras fontes com os mesmos dados. Ele permite identificar outros conjuntos de dados que contenham os mesmos dados, porém com uma confiabilidade melhor, por exemplo. Ele possui granularidade geral, pois um mesmo dado pode estar presente em diferentes conjuntos de dados.

3.3. Elementos de Granularidade Tabela

O descritor **Tratamento de Dados Ausentes** indica algoritmos utilizados para tratar os dados. Ele permite que as equipes possam saber como o tratamento foi realizado. Ele possui granularidade específica, pois muitos dos dados de uma tabela nem sempre estão em um mesmo padrão, sendo necessário um tratamento prévio anterior ao seu uso.

O elemento **Referências** indica o tipo de documento que descreve as colunas de uma tabela. Ele permite referenciar o dicionário de dados, o que garante um melhor entendimento da tabela. Ele possui granularidade específica, pois diz respeito a descrição de colunas de uma tabela. Já o descritor **Colunas** mostra as colunas pertencentes a uma tabela. Ele permite identificar que tipo de informações a tabela contém. Ele possui granularidade específica, pois conjuntos de dados não possuem colunas. O descritor **Amostra de Dados** mostra como o preenchimento das colunas de uma tabela é feito. Ele permite enxergar como os dados são dispostos em uma coluna e o tipo deles.

O descritor **Ferramentas de Análise** mostra os links das ferramentas e métodos utilizados. Ele permite a centralização das informações referentes às ferramentas de

análise. Ele possui granularidade específica, pois se refere a ferramentas e métodos relativos a tabelas e seus dados. O elemento **Repositório de Código** é onde todos os algoritmos e códigos referentes a uma tabela estarão. Ele permite a centralização e o compartilhamento de informações sobre o projeto de forma eficaz. Ele possui granularidade específica, pois são informações relacionadas a dados de tabelas.

Por fim, o elemento **Data de criação** permite avaliar características temporais da tabela. Ele possui granularidade específica, pois muitos conjuntos de dados são antigos e o conhecimento dessa informação não está disponível.

4. Análise de Ferramentas de Gestão de Metadados

Este trabalho analisou e comparou 2 ferramentas populares de código aberto: *OpenMetadata*[2] e *DataHub*[3]. Para determinar a mais adequada ao Projeto *Céos*, foram utilizadas pequenas amostras de dados governamentais. Os principais critérios para escolha da ferramenta incluíram flexibilidade na aplicação dos elementos selecionados, integração com ferramentas de dados, documentação detalhada, qualidade dos dados, recursos colaborativos e suporte da comunidade.

A *OpenMetadata* se destacou pela flexibilidade e documentação, pois a *DataHub* apresentou obstáculos devido à documentação insuficiente. As funcionalidades de análise de qualidade de dados do *OpenMetadata* se mostraram superiores, provendo métricas como estatísticas descritivas, distribuição de valores, integridade dos dados e detecção de anomalias. Recursos colaborativos como registro de ações e designação de tarefas com uma interface amigável também foram grandes diferencias dela.

5. Conclusão

Este trabalho apresenta uma proposta de definição de metadados no âmbito de dados governamentais e a aplicabilidade de um padrão de metadados para a catalogação de conjuntos de dados neste domínio. O propósito é facilitar o compartilhamento, organização, extração de informações e métricas relevantes ao Projeto *Céos*. A pesquisa inicial levou em consideração uma série de diferentes padrões de metadados e decidiu por adotar o DC, que é reconhecido e bastante aplicado como padrão para representação de metadados.

Os resultados parciais até o momento indicam que a aplicação do DC é satisfatória para a catalogação de metadados de conjuntos de dados governamentais no domínio desejado. A utilização desse padrão e a adoção de elementos específicos disponibiliza informações relevantes sobre os dados, facilita a compreensão dos metadados por diferentes usuários e permite a interoperabilidade entre os grupos e instituições envolvidas no projeto. Em termos técnico-científicos, a contribuição desta pesquisa é auxiliar significativamente processos de integração e análise de dados de diferentes fontes.

Entretanto, cabe ressaltar que, como esta pesquisa encontra-se em estágio inicial, os elementos selecionados devem ainda ser avaliados para novos tipos de conjuntos de dados governamentais que venham a ser inseridos no projeto, considerando a diversidade de informações e estruturas dos dados existentes no domínio.

[2]https://open-metadata.org/

[3]https://datahubproject.io/

References

Alasem, A. (2009). An overview of e-government metadata standards and initiatives based on dublin core. *Electronic Journal of e-Government*, 7(1):pp1–10.

Alves, M. and Souza, M. I. F. (2007). Estudo de correspondência de elementos metadados: Dublin core e marc 21.

CEGE (2014). COMITÊ EXECUTIVO DO GOVERNO ELETRÔNICO. Padrão de metadados do governo eletrônico (e-pmg 1.1).

Core, D. (2024). Dublin core. *https://www.dublincore.org/*. Último acesso: 12 de julho de 2024.

da Silva, J. R., Ribeiro, C., and Lopes, J. C. (2019). Ranking dublin core descriptor lists from user interactions: a case study with dublin core terms using the dendro platform. *International Journal on Digital Libraries*, 20:185–204.

DCMI (2024). Dcmi metadata terms. *https://www.dublincore.org/specifications/dublin-core/dcmi-terms/section-1*. Último acesso: 12 de julho de 2024.

Koshman, S. (2010). Visualizing Metadata for Environmental Datasets. *International Conference on Dublin Core and Metadata Applications*, 2010.

Maali, F., Cyganiak, R., and Peristeras, V. (2010). Enabling interoperability of government data catalogues. In *Electronic Government: 9th IFIP WG 8.5 International Conference, EGOV 2010, Lausanne, Switzerland, August 29-September 2, 2010. Proceedings 9*, pages 339–350. Springer.

Weibel, S. L. (1995). Metadata the foundation of resource description. *Annual review of OCLC research*, pages 52–56.