INTEGRACuBe: Exploração de dados analíticos em RDF

Jones O. Avelino[1,2], Kelli F. Cordeiro[1,2], Maria C. Cavalcanti[1]

[1]Instituto Militar de Engenharia (IME)
Praça General Tibúrcio 80, Praia Vermelha – 22.290-270 – Rio de Janeiro

[2]Centro de Análise de Sistemas Navais
Ed. 23 do AMRJ – R. da Ponte, s/n, Centro – 20.091-000 – Rio de Janeiro

{jones.avelino,kelli}@marinha.mil.br, yoko@ime.eb.br

Abstract. *The growth of web-available datasets that use the RDF standard enables data analysis that involves multiple dimensions. According to the W3C, one of the resources for analyzing multidimensional data is the use of the RDF Data Cube vocabulary. However, there is still a lack of support tools for applying this vocabulary in datasets. In this sense, this article proposes INTEGRACuBe, an environment that uses a meta-scheme and semi-automated mechanisms to support the mapping of data resources to the RDF Data Cube metamodel. As a result, an exploration of analytical data in RDF will be possible. Additionally, a case study is presented in the Software Development Management scenario.*

Resumo. *O crescimento de conjuntos de dados disponíveis na Web que utilizam o padrão RDF propicia análises de dados que envolvem múltiplas dimensões. Segundo a W3C, um dos recursos para analisar dados multidimensionais é a utilização do vocabulário RDF Data Cube. Contudo ainda há uma carência de instrumentos de apoio para aplicação deste vocabulário em conjuntos de dados. Nesse sentido, este artigo propõe o INTEGRACuBe, um ambiente que utiliza um metaesquema e mecanismos semiautomatizados para apoiar o mapeamento de recursos de dados ao metamodelo RDF Data Cube. Como resultado, será possível a exploração de dados analíticos em RDF. Adicionalmente, um estudo de caso é apresentado no cenário de Gerência de Desenvolvimento de Software.*

1. Introdução

Nos últimos anos, o crescimento de dados na Web, principalmente aqueles publicados no padrão RDF (*Resource Description Framework*), propiciou análises de dados a partir de interligações entre diferentes conjuntos de dados, permitindo que organizações tenham suas decisões sustentadas em dados [Tadesse et al. 2019]. Por um lado, alguns trabalhos apresentam abordagens de interligações entre conjuntos de dados através, por exemplo, do grau de similaridade entre seus recursos com base na sintaxe [Avelino et al. 2020]. Por outro lado, há trabalhos que exploram aspectos semânticos e suas interligações, geralmente estabelecidas por meio de vocabulários e/ou ontologias [Figueiredo et al. 2020, Silveira and Cavalcanti 2020]. No geral, esses trabalhos buscam agregar conhecimento, permitindo a recuperação de dados a partir de consultas SPARQL.

Embora consultas SPARQL consigam recuperar dados, há limitações quando se tratam de dados agregados. Neste sentido, a W3C adotou o vocabulário RDF Data Cube[1]

[1]https://www.w3.org/TR/vocab-data-cube/

como mecanismo para explorar conjuntos de dados baseado na análise multidimensional [Escobar et al. 2020]. O RDF Data Cube foi implementado por meio de um metamodelo capaz de lidar com dados agregados a partir do mapeamento de recursos dos conjuntos de dados [Cyganiak et al. 2014]. No entanto, o mapeamento não é trivial e existe uma dificuldade de estabelecer quais recursos serão mapeados, o papel de cada um deles e suas relações [Mountantonakis et al. 2019].

Neste trabalho, propomos um ambiente denominado INTEGRACuBe (**INTE**rlevel data inte**GRA**tion **C**u**B**e) cujo objetivo é apoiar o mapeamento de recursos de dados de um conjunto de dados ao metamodelo RDF Data Cube. Para tal, implementamos a ferramenta INTEGRACuBeTool que se baseia na Modelagem Dimensional (MD), combinado com um conjunto de passos, aqui representados por *Steps* implementados no Pentaho Data Integration (PDI)[2]. As contribuições deste trabalho são: (i) um ambiente formado por um metamodelo e uma ferramenta; (ii) uma ferramenta que apoia o mapeamento do metamodelo RDF Data Cube; e (iii) um estudo de caso real que demonstra a viabilidade e utilidade.

2. Ambiente analítico para apoio a tomada de decisão utilizando grafos RDF

Em ambientes que a tomada de decisão é um fator decisivo para os negócios, há necessidade de analisar dados históricos. Um meio viável é a análise multidimensional que dá suporte à tomada de decisão. Nesses ambientes, as análises são baseadas em fatos e utilizam operações OLAP (*Online Analytical Processing*) associadas a um *Data Warehouse* (DW). Um DW é composto por processos ETL (Extração, Transformação e Carga) e utiliza tabelas dimensão e fatos para representar os dados. As tabelas dimensão estabelecem atributos comuns entre fontes de dados e agrupam dados, descrevendo "quem, o quê, onde, quando, como e por quê". Já as tabelas fato agregam valores de medições comuns ao negócio. Ambas as tabelas são representadas em um modelo de dados, denominado Modelo Dimensional (MD). O MD é uma técnica para representar dados analíticos, caracterizado pela simplicidade e alto desempenho [Kimball and Ross 2013].

Considerando as necessidades de análises de dados, em RDF é possível realizar consultas utilizando a linguagem SPARQL para explorar dados. Apesar de a linguagem ser padronizada, as análises são limitadas a dimensões não agregadas, reduzindo a abrangência dos resultados. Em alguns casos é necessário analisar dados multidimensionais e agregados [Escobar et al. 2020]. Por isso, a W3C adotou o RDF Data Cube, que é composto de um vocabulário para troca e compartilhamento de dados estatísticos e metadados entre organizações [Cyganiak et al. 2014]. O metamodelo descreve classes e propriedades. Ele é constituído de dois componentes principais: (i) *ComponentSpecification* que agrega classes do tipo *ComponentProperty* e permite definir dimensões, medidas e atributos; e (ii) *DataStructureDefinition* que corresponde aos conjuntos de dados vinculados às dimensões, às métricas e aos dados, composto por *observations*, *slices* e *datasets*.

3. Trabalhos Relacionados

Há trabalhos, como em [Escobar et al. 2020] e [Etcheverry and Vaisman 2017], que apresentam abordagens que adotam o metamodelo RDF Data Cube. Em [Escobar et al. 2020],

[2]https://www.hitachivantara.com/en-us/products/data-management-analytics/pentaho-platform.html

os autores propuseram uma abordagem que descreve o processo de melhoria e enriquecimento da qualidade dos dados publicados baseado em dados multidimensionais de modo que usuários consigam interagir, evitando a complexidade do SPARQL. Já o trabalho de [Etcheverry and Vaisman 2017] está centrado na exploração de dados governamentais abertos, utilizando uma extensão do RDF Data Cube por meio do QB4OLAP. Apesar de ambos os trabalhos fazerem o mapeamento dos recursos de dados com o vocabulário RDF Data Cube, os mesmos não oferecem um suporte semiautomatizado como o INTEGRACuBe. Assim, este trabalho diferente dos trabalhos apresentados, propõe uma ferramenta semiautomatizada que utiliza recursos de um metaesquema, oferecendo mecanismos que apoiam o mapeamento dos recursos de dados ao metamodelo do RDF Data Cube, permitindo aos usuários a exploração de dados analíticos em RDF, detalhado na Seção 4.

4. INTEGRACuBe

O INTEGRACuBe é um ambiente composto pelo metaesquema da abordagem INTEGRA (Figura 1) e a ferramenta INTEGRACubeTool (Figura 2). O metaesquema é um grafo RDF composto de quatro recursos (*table, tuple, hasAttribute, hasTuple*) capazes de serem associados aos recursos de dados das fontes de dados, permitindo a identificação do objeto e seu nível de representação (esquema ou instância) [Avelino et al. 2020].

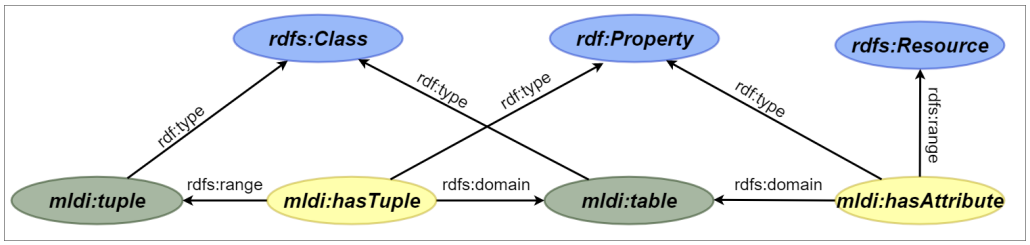

Figura 1. Metaesquema do INTEGRA (Adaptado) [Avelino et al. 2020]

O INTEGRACuBeTool[3] tem como objetivo aplicar os recursos do vocabulário RDF Data Cube a um conjunto de dados armazenado no *Triplestore*, através de mecanismos semiautomatizados, transformando o grafo tradicional em um grafo multidimensional. Como ilustrado na Figura 2, INTEGRACuBeTool foi implementado por meio de um *Job*, composto por seis *Steps* que permitem identificar os recursos dimensão e fato candidatos do MD, utilizando os recursos, *table* e *tuple*, do metaesquema do INTEGRA [Avelino et al. 2020]. Em seguida, ele realiza o mapeamento das classes do metamodelo RDF Data Cube, até a geração do cubo de dados e sua persistência no *Triplestore*.

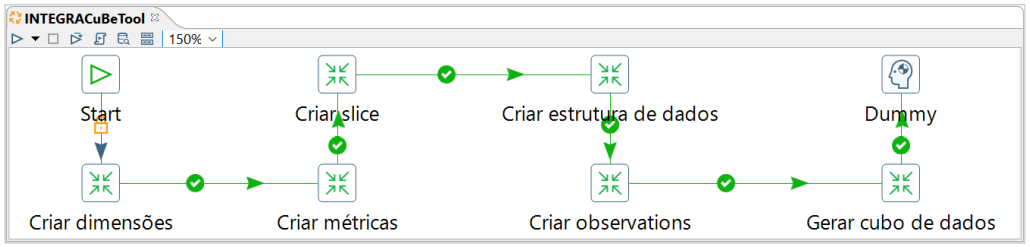

Figura 2. Visão Geral dos *Steps* de INTEGRACuBeTool.

[3]https://github.com/jonesavelino/INTEGRACUBE

A implementação foi norteada pela estrutura do RDF Data Cube. Com isso, os *Steps* Criar dimensões, Criar métricas e Criar slice foram baseados nas classes do *ComponentSpecification*, a partir dos recursos vinculados à classe *table* do metaesquema do INTEGRA. O objetivo é identificar os recursos dimensão e fato do conjunto de dados e, por fim, mapeá-los através das classes *DimensionProperty, MeasureProperty e SliceKey*. Já os *Steps* Criar estrutura de dados, Criar observations e Gerar cubo de dados foram baseados nas classes do *DataStructureDefinition*. O objetivo é criar o cubo de dados a partir dos recursos vinculados à classe *tuple* que representam as instâncias. Esses dados são mapeados através das classes *DataSet e Observations*. Na Seção 5, um experimento será apresentado para elucidar a ferramenta, explorando sua utilidade no estudo de caso.

INTEGRACuBeTool foi desenvolvido na plataforma do PDI através do *framework* ETL4LOD+[4] para operações em grafos RDF [Cordeiro et al. 2011]. O resultado da execução dos *Steps* do PDI, ilustrado na Figura 2, é a geração de um cubo de dados armazenado no Openlink Virtuoso (*https://virtuoso.openlinksw.com*), que está ilustrado em modo gráfico, na Figura 4, por meio de uma consulta via GraphDB *(https://www.ontotext.com/products/graphdb/)*. Além disso, o cubo pode ser exportado no formato *NTriples*, permitindo a exploração por consultas SPARQL ou através da ferramenta OpenCube Toolkit (*https://github.com/opencube-toolkit*), abordado na Seção 5.

5. Experimento e Exploração de Cubo de Dados

O INTEGRACuBe foi aplicado no experimento a partir de uma amostra do estudo de caso real de um projeto na área de Gerência de Projetos de Desenvolvimento de *Software* baseado no trabalho de [Avelino et al. 2020]. O objetivo é avaliar a alocação de pessoal da equipe. Para tal, é disponível um conjunto de dados, representado em um grafo RDF, com 7.716 triplas do *software* de tarefas Redmine[5]. Nesse cenário, um exemplo de análise é a quantidade de tarefas executadas por papel, podendo envolver outras dimensões.

Na Figura 3, é ilustrado o MD com as entidades do negócio. Nele, há quatro tabelas dimensão, em que *dimTask* descreve as tarefas, *dimPerson* a equipe, *dimStatusTask* os estados e *dimTime* as séries temporais. Além disso, a tabela fato *factProjectAction* é responsável por agregar valores e métricas quantitativas das tarefas.

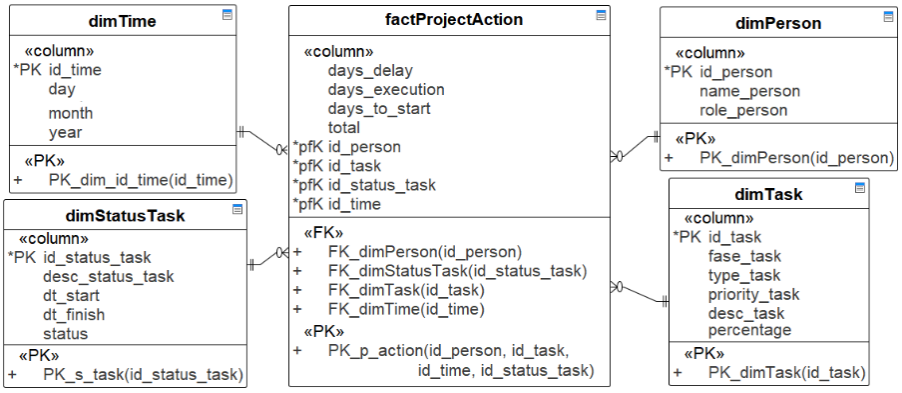

Figura 3. Modelo dimensional com as entidades do estudo de caso.

[4]https://ufrj.gitbook.io/etl4lod/

[5]https://www.redmine.org/

Na Figura 4, dois grafos RDF do experimento são ilustrados. O primeiro grafo representa a equipe, sem a adoção do metamodelo RDF Data Cube. Nele, são destacados *person_task*, as classes do metaesquema do INTEGRA, *person_task_papel* e literais. Já o segundo, após a execução do INTEGRACuBeTool, é gerado o grafo multidimensional que representa o cubo de dados *project_action*. Nele, são destacados os recursos baseado nas classes (*ComponentSpecification* e *DataStructureDefinition*) e os *Steps* envolvidos.

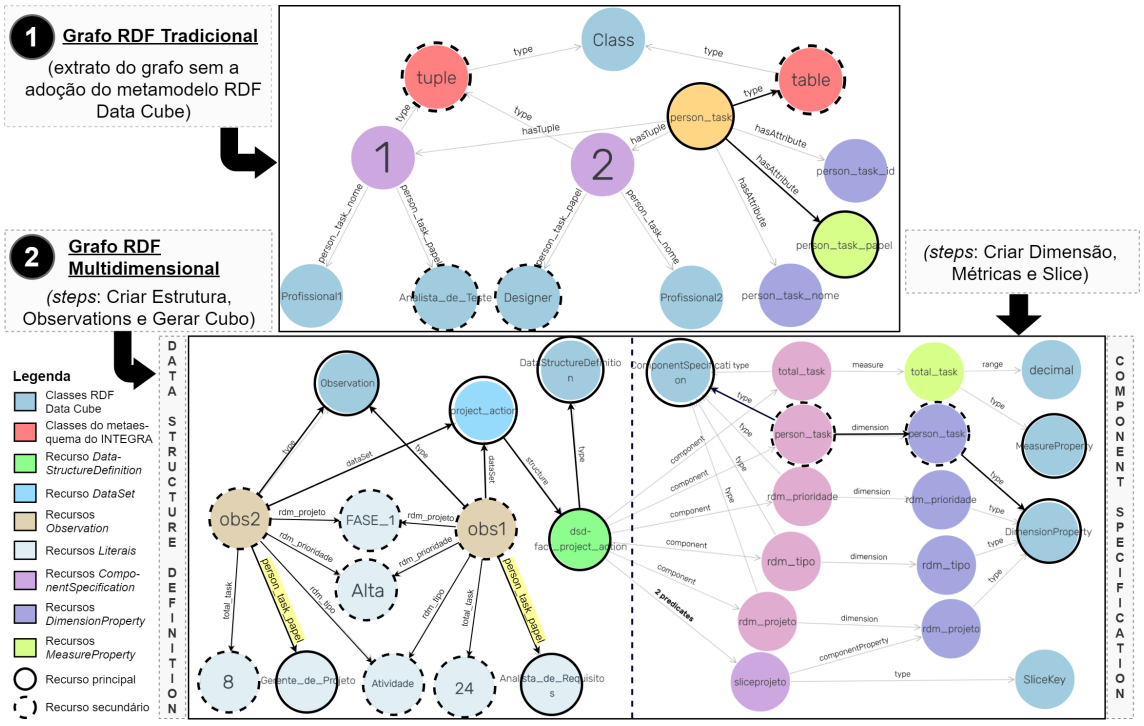

Figura 4. Grafos RDF tradicional e multidimensional.

Como abordado na Seção 4, dada a flexibilidade do INTEGRACuBe, o cubo de dados pode ser explorado de duas formas: (i) consultas SPARQL; e (ii) ferramentas que adotam o paradigma OLAP com o vocabulário RDF Data Cube. Na Figura 5, são ilustradas as duas formas de exploração do cubo de dados *project_action*. Na primeira, são destacadas as propriedades, *qb:dimension* e *qb:measure*, da classe *ComponentSpecification*, assim como as propriedades, *qb:structure* e *qb:dataset*, da classe *DataStructureDefinition*. Já na segunda, a ferramenta OpenCube Toolkit retorna os dados analíticos via interface, permitindo o usuário avaliar a alocação de tarefas por múltiplas dimensões.

6. Conclusão

Este artigo apresentou o INTEGRACuBe, um ambiente que apoia o mapeamento dos recursos de dados do grafo RDF tradicional utilizando o vocabulário RDF Data Cube, transformando-o em um grafo multidimensional. Para isso, foi implementada uma ferramenta que fornece mecanismos semiautomatizados baseado nos recursos *table* e *tuple* do metaesquema da abordagem INTEGRA. Nesse sentido, INTEGRACuBeTool é uma ferramenta flexível que permite adicionar novos recursos de dados por meio da execução de um *Job* com seis *Steps* de modo a agregar desde novas dimensões e fatos até a criação do cubo de dados. Ao INTEGRACuBeTool foi submetido um conjunto de dados, baseado

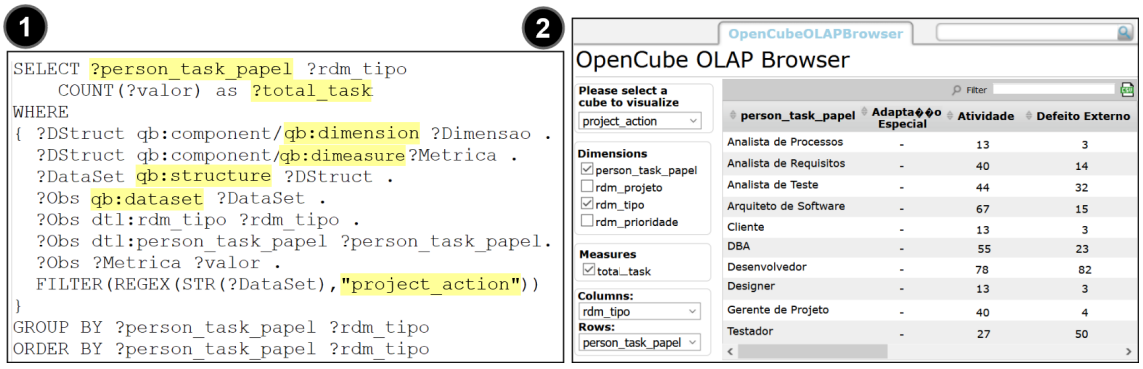

Figura 5. Formas de exploração do cubo de dados *project_action*.

em um estudo de caso sobre um cenário real. Os resultados obtidos evidenciaram tanto a utilidade quanto a viabilidade da proposta. Trabalhos futuros incluem: (i) implementar no INTEGRACuBeTool o RDF *Knowledge Graph* a partir de conjuntos de dados interligados, que propicia a exploração do conhecimento em grafos semânticos; e (ii) estender o metaesquema do INTEGRA para contemplar outras propriedades de integração de dados como os atributos de proveniência dos dados.

Referências

Avelino, J. O., Cordeiro, K. F., and Cavalcanti, M. C. (2020). An RDF based approach for integrating data at different levels of abstraction. In *Proceedings of the Brazilian Symposium on Multimedia and the Web*, WebMedia '20, page 81–88.

Cordeiro, Kelli, F., Pereira, et al. (2011). An approach for managing and semantically enriching the publication of linked open governmental data. In *SBBD*, pages 82–95.

Cyganiak, R., Reynolds, et al. (2014). The RDF data cube vocabulary. *World Wide Web Consortium (W3C), 16th Jan*, 2:014.

Escobar, P. et al. (2020). Adding value to linked open data using a multidimensional model approach based on the RDF data cube vocabulary. *Computer Standards Interfaces*.

Etcheverry, L. and Vaisman, A. A. (2017). Efficient analytical queries on semantic web data cubes. *Journal on Data Semantics*, 6(4):199–219.

Figueiredo, G., Cordeiro, K. F., and Campos, M. L. M. (2020). LigADOS: Interlinking datasets in open data portal platforms on the semantic web. *Metadata and Semantic Research*, 1355:73 – 84.

Kimball, R. and Ross, M. (2013). *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. Wiley, Indianapolis, IN, 3 edition.

Mountantonakis et al. (2019). Large-scale semantic integration of linked data: A survey. *ACM Comput. Surv.*, 52(5).

Silveira, R. and Cavalcanti, M. (2020). Método para rotular ligações semânticas na web de dados. In *Anais do XXXV Simpósio Brasileiro de Bancos de Dados*, pages 49–60.

Tadesse, S. et al. (2019). ARDI: Automatic generation of RDFS models from heterogeneous data sources. In *2019 IEEE 23rd International Enterprise Distributed Object Computing Conference (EDOC)*, pages 190–196.