Arcabouço Comparativo de Ferramentas para Integração Semântica de Dados Tabulares

Marcello P. Bax[1], **Rafael Rocha**[1]

[1]Escola de Ciência da Informação – Universidade Federal de Minas Gerais (UFMG)
Belo Horizonte – MG – Brasil

{bax,rafael-rocha}@ufmg.br

Abstract. *Comma-Separated Values (CSV) is used due to its low computational cost, on the other hand it is necessary to integrate it with a high-level repository to aggregate value to the data. In addition, Linked Data (LD) brought a new approach to enrich data and generate knowledge. Institutions in the most diverse areas can extract more knowledge from the data when it is coated in a semantic format. Each semantic tool has different features that directly impact the integration of LD from CSV. Hence, it is a recurring problem to know if the tool has all the features required for an integration project. This work proposes an objective evaluation of the features present in the tools that perform the semantic integration of data in CSV. The classification uses the features created in the literature in a new structure with their grouping. The salient features in the classification form the comparative framework that uses a positive number line facilitating the evaluation of the tools. The classification and framework can be used in all areas where there is semantic integration of data. The result of this work is possible to evaluate the features implemented in the tools that generate semantic integration in an agile comparative analysis.*

Resumo. *O Comma-Separated Values (CSV) é utilizado pelo seu baixo custo computacional, em contrapartida é necessário integrá-lo a um repositório de alto nível para agregar valor aos dados. Neste sentido, o Linked Data (LD) trouxe uma nova abordagem para enriquecer dados e gerar conhecimento. Instituições das mais diversas áreas podem extrair mais conhecimento dos dados quando estes estiverem em um formato semântico. Cada ferramenta semântica apresenta diferentes recursos que impactam diretamente na integração do LD a partir do CSV. Neste contexto, é um problema recorrente saber se a ferramenta possui todos os recursos demandados de um projeto de integração. Este trabalho propõe uma avaliação objetiva dos recursos presentes nas ferramentas que realizam integração semântica de dados em CSV. A classificação construída utiliza os recursos previamente avaliados na literatura em uma nova estrutura com agrupamento dos mesmos. Os recursos salientados na classificação formam o arcabouço comparativo que utiliza uma reta numérica positiva facilitando a avaliação das ferramentas. A classificação e o arcabouço podem ser utilizados em todas as áreas que haja integração semântica de dados. O resultado deste trabalho é possível avaliar os recursos implementados nas ferramentas que geram integração semântica em uma análise comparativa ágil.*

1. Introdução

A utilização do formato *Comma-Separated Values* (CSV) é popular entre os cientistas devido à sua simplicidade, flexibilidade e ao seu baixo custo de processamento e armazenamento[1]. Os arquivos CSV armazenam dados separados por vírgulas onde as linhas representam os registros (indivíduos), as colunas constituem propriedades (ou atributos dos indivíduos) e os cabeçalhos das colunas representam metadados sobre os conteúdos das células de uma coluna [Shafranovich 2005]. Entretanto, dados tabulares em CSV não possuem semântica explícita, sendo útil convertê-los para *Linked Data* (LD) quando é importante que estes sejam interoperáveis. O padrão LD é composto de diversas estruturas e recursos que possibilitam a interoperabilidade e conhecimento compartilhado [Bizer et al. 2011]. Destaca-se o *Resource Description Framework* (RDF), estruturado em triplas que relacionam sujeito, predicado e objeto, pois cada um desses elementos pode ser identificado por uma *Uniform Resource Identifier* (URI) única na web. Para integrar semanticamente dados em CSV são necessárias ferramentas que automatizem a conversão para LD, na sequência, sua integração em um repositório de alto nível como, por exemplo, um *triplestore*, *data lake* ou *big data* [Adamou and D'Aquin 2020]. O processo de integração é contínuo, já que novas fontes de dados necessitam ser integradas frequentemente ao repositório. Cada ferramenta possui recursos que influenciam o LD resultante e consequentemente a integração. Vários trabalhos elaboraram metodologia para avaliar ferramentas e linguagens para conversão de dados tabulares para LD. Um arcabouço comparativo de recursos funcionais que utiliza uma escala conceitual para classificar as linguagens de mapeamento é apresentado em [Hert et al. 2011] e [Crotti Junior et al. 2017]. Outros dois arcabouços que analisam os fatores que influenciam a geração do LD aparecem em [Dimou et al. 2018] e [Rashid et al. 2020].

Este artigo apresenta pesquisa em andamento para categorizar e avaliar requisitos funcionais e não-funcionais de ferramentas de integração semântica em uma reta numérica positiva, em que as maiores notas representam características mais avançadas. A partir dessa classificação, pretende-se desenvolver um arcabouço comparativo atualizado que permita analisar um conjunto de abordagens presentes nas ferramentas, de forma simples e ágil, identificando aquelas que apresentam mais recursos (e recursos mais avançados), e em que classe a respectiva ferramenta é melhor avaliada. No artigo, a Seção 2 apresenta os trabalhos correlatos e a metodologia na comparação das abordagens. A Seção 3 discute o conceito de integração semântica e apresenta a classificação dos recursos. A Seção 4 discute os recursos presentes nas abordagens e analisa o arcabouço proposto. Finalmente, a Seção 5 apresenta as conclusões, limitações e trabalhos futuros.

2. Trabalhos Relacionados

[Hert et al. 2011] analisam recursos necessários para converter bancos de dados relacionais para RDF. Os recursos do arcabouço comparativo estado da arte são: *Logical Table to Class*, *M:N Relationships*, *Project Attributes*, *Select Conditions*, *User-defined Instance URIs*, *Literal to URI*, *Vocabulary Reuse*, *Transformation Functions*, *Datatypes*, *Named Graphs*, *Blank Nodes*, *Integrity Constraints*, *Static Metadata*, *One Table to n Classes* e

[1]Os repositórios do Portal Brasileiro de Dados Abertos (http://dados.gov.br/dataset), FiveThirty Eight (https://data.fivethirtyeight.com) e Kaggle (https://www.kaggle.com/datasets), juntos, fornecem milhares de datasets em CSV.

Write Support. A linguagem de mapeamento pode oferecer suporte completo ao recurso, suporte parcial ou sem suporte. Além dos recursos apresentados até aqui, dois recursos são considerados por [Crotti Junior et al. 2017]: *Filter* e *Reusability*.

Para [Dimou et al. 2018] os fatores que influenciam a conversão para LD são analisados em caráter técnico e não-técnico; avaliam se a ferramenta possui os fatores ou se os implementam parcialmente, ou ainda se não possui os fatores. Em [Rashid et al. 2020] os autores propõem agrupar as abordagens analisadas e definir critérios de avaliação aplicados em três grupos: dicionários de dados tradicionais, linguagens de mapeamento e ferramentas de integração de dados. Cada recurso das abordagens recebe um valor entre 0 (zero) e 1 (um).

A avaliação conceitual ou com números decimais dificulta a sumarização dos resultados. Nenhum trabalho é dedicado à análise de requisitos funcionais e não-funcionais na integração semântica de dados tabulares em CSV.

3. Classificação de Recursos para Integração Semântica

Conforme [Doan et al. 2004], a participação do usuário no processo de geração de LD é fundamental, pois o mapeamento dificilmente será completamente automatizável e autônomo. Assim os recursos não-funcionais que gerem maior aderência do usuário à ferramenta são tão importantes quanto recursos funcionais. Em [Bizer et al. 2011] os autores reforçam que as ferramentas carecem de interfaces cada vez mais intuitivas para se popularizarem, portanto, nessa seção é discutido os recursos salientes para elevação semântica do CSV e a integração dos dados.

Recursos para metaDados (RD) contém funcionalidades voltadas para gerenciar os dados dos *datasets*. Por exemplo, alguns dados do arquivo original não devem constar no LD gerado (dados sensíveis ou desatualizados), portanto deve ser possível filtrá-los (RD.I). Este recurso é descrito como: *Project Attributes* em [Hert et al. 2011, p. 27]; *Select* em [Crotti Junior et al. 2017, p. 407]; *Subset Selection* em [Rashid et al. 2020, p. 466]. Ao gerar o LD, alguns metadados podem ser criados para enriquecer o LD (RD.II). Este recurso é descrito como: *Static Metadata* em [Hert et al. 2011, p. 28]; *Additional Data* em [Crotti Junior et al. 2017, p. 407]. Histórico dos dados, sua proveniência e como foram gerados são informações importantes que a ferramenta deve poder adicionar durante a integração (RD.III). Este recurso é descrito como: *Static Metadata* em [Hert et al. 2011, p. 28]; *Additional Data* em [Crotti Junior et al. 2017, p. 407]; *Provenance* em [Rashid et al. 2020, p. 469]. Não é possível determinar o tipo de um dado em um arquivo CSV, logo a ferramenta deve permitir associar ao dado o seu tipo (RD.IV). Este recurso é descrito como: *Datatypes* em [Hert et al. 2011, p. 27]; *Data types* em [Crotti Junior et al. 2017, p. 409]; *Data type assignment* em [Rashid et al. 2020, p. 466]. *Blank Nodes* (ou BNodes) são uma forma disponível no LD de representar dados complexos cuja identificação de seus componentes não seja necessária, este recurso pode ou não ser suportado pela ferramenta (RD.V). É descrito como: *Blank Nodes* em [Hert et al. 2011, p. 27] e [Crotti Junior et al. 2017, p. 409]. O RD.VI diz respeito a técnica utilizada para gerar as URIs, estas podem ser geradas automaticamente pela ferramenta ou manualmente explicitando-se os valores através de *templates* definidos pelo usuário. A URI utilizada no LD pode ser gerada automaticamente (RD.VI.1) ou pode ser definida pelo usuário (RD.VI.2). Este recurso é descrito como: *User-defined Instance URIs* em [Hert et al. 2011, p. 27].

Recursos para Mapeamento (RM) estão relacionados à representação da conversão do dado tabular sem semântica explícita para LD. O RM.I define o tipo de mapeamento realizado. O mapeamento é *Direto* (RM.I.1) quando há uma replicação da estrutura tabular para o LD. O mapeamento *Aumentado* (RM.I.2) ocorre quando a ferramenta oferece algum enriquecimento semântico nos dados tabulares, por exemplo, a explicitação de algumas relações, tal como a relação de subsunção. O mapeamento é *Semântico* (RM.I.3) quando a ferramenta utiliza uma forma de mapear os dados para conceitos de uma ontologia. RM.I é discutido como: *Mapping* em [Hert et al. 2011, p. 29]; *Mapping-driven* e *Data-driven* em [Dimou et al. 2018, p. 3]; *Approaches* em [Rashid et al. 2020, p. 470]. O mapeamento pode ocorrer em dois momentos antes da importação dos dados tabulares, chamado de *Mapeamento pré* (RM.II). O *Mapeamento pós* (RM.II) ocorre quando o dado tabular já está carregado. Os itens RM.II e RM.III não são excludentes. É discutido a temporalidade da ação como *Generation's execution* em [Dimou et al. 2018, p. 2].

Em relação aos Recursos para Processamento (RP), esses podem ser *processamento em lote* (RP.I) que não exige interação do usuário; ou o *processamento online* (RP.II) que permite ciclos interativos com o usuário. Ambos os recursos são discutidos como *Trigger* em [Dimou et al. 2018, p. 3].

Já os Recursos para Serialização (RS) são classificados em *serialização do CSV* (RS.I) quando a ferramenta gera um arquivo de saída contendo os dados tabulares em seu estado original; *serialização do LD* (RS.II) quando gera um aquivo contendo os dados semânticos; *serialização do mapeamento* (RS.III) quando gera um arquivo com as informações do mapeamento realizado. A serialização é discutida como: *Reusability* em [Crotti Junior et al. 2017, p. 409]; *Materialization* em [Dimou et al. 2018, p. 3]; *Graph materialization* em [Rashid et al. 2020, p. 467].

A pontuação dos recursos segue a seguinte maneira, o RD.VI recebe 1 (um) ponto quando a ferramenta gerar a URI automaticamente e recebe 2 (dois) pontos quando permitir o usuário definir a URI, ao passo que o RM.I recebe 1 (um) ponto para o mapeamento direto, 2 (dois) pontos para o mapeamento aumentado e 3 (três) pontos para o mapeamento semântico. O restante dos recursos recebem 1 (um) ponto quando a ferramenta possui o respectivo item e 0 (zero) caso contrário. Os recursos não-funcionais segue os critérios de conformidade de usabilidade e funcionalidade da ISO/IEC 9126 [Jung et al. 2004], portanto são representados pelos recursos: RD.VI.2; RM.II; RM.III; RP.I; RP.II.

4. Avaliação das Ferramentas

A avaliação é realizada seguindo a literatura científica e os manuais produzidos pelos autores. Ademais, a ferramenta deve possuir: (i) interface gráfica para usuário final; (ii) manipular o arquivo CSV com recursos próprios, ou seja, sem *software* de terceiros; (iii) gerar LD apartir do CSV. Assim, cada ferramenta é instalada em ambiente conforme a documentação e a avaliação utiliza um arquivo CSV e uma modelagem semântica. A classificação foi aplicada ao LD-R[2] e ao Karma[3].

O LD-R precisa importar o arquivo CSV para iniciar o processo, portanto RM.III recebe 1 (um) ponto e RM.II recebe 0 (zero) ponto. O mapeamento é tipo *aumentando*,

[2]https://github.com/ali1k/ld-r

[3]https://github.com/usc-isi-i2/Web-Karma

ou seja, é replicada a estrutura tabular além de atribuir semântica. O RM.I recebe 2 (dois) pontos. O LD-R permite filtrar metadados, mas não permite criar metadados. O RD.I recebe 1 (um) ponto e RD.II recebe 0 (zero) ponto. A proveniência é gerada automaticamente pela ferramenta, RD.III recebe 1 (um) ponto. É possível tipar os dados literais, mas a ferramenta não gera BNodes de forma automática e não permite o usuário definir. O RD.IV recebe 1 (um) ponto e RD.V recebe 0 (zero) ponto. A URI é definida no mapeamento, portanto o RD.VI recebe 2 (dois) pontos. LD-R oferece somente o processamento *online*, logo RP.I recebe 0 (zero) ponto e RP.II recebe 1 (um) ponto. Não há serialização pela ferramenta, recebe 0 (zero) em todos os RS.

Karma utiliza mapeamento *semântico*, RM.I recebe 3 (três) pontos. O mapeamento pode ser carregado antes ou após a carga do arquivo CSV, assim RM.II e RM.III recebem 1 (um) ponto cada. No mapeamento do Karma é possível filtrar metadados, RD.I recebe 1 (um) ponto. A ferramenta permite criar metadados, RD.II recebe 1 (um) ponto. Karma permite que o usuário forneça os metadados para proveniência, RD.III recebe 1 (um) ponto. Os BNodes são gerados de forma automática pela ferramenta, RD.V recebe 1 (um) ponto. A URI no LD é gerado de forma automática, RD.VI recebe 1 (um) ponto. Karma oferece processamento em lote e *online*, logo RP.I e RP.II recebem 1 (um) ponto cada. A ferramenta permite serializar o arquivo CSV, o LD e o mapeamento realizado. Nos RS, cada item recebe 1 (um) ponto.

4.1. Resultados

Como pode-se observar, o Karma é a ferramenta que concentra a maior quantidade de recursos e optou por utilizar mapeamento semântico, além de gerar automaticamente a URI no LD. O LD-R não possui nenhum recurso para serialização, além disso adotou o mapeamento aumentado e o usuário define a construção da URI no LD. A Figura 1 representa graficamente o arcabouço comparativo.

	RD.I	RD.II	RD.III	RD.IV	RD.V	RD.VI	RD	RM.I	RM.II	RM.III	RM	RP.I	RP.II	RP	RS.I	RS.II	RS.III	RS	Total
LD-R	1	0	1	1	0	2	5	2	1	1	4	0	1	1	0	0	0	0	10
Karma	1	1	1	1	1	1	6	3	1	1	5	1	1	2	1	1	1	3	16

Figura 1. Arcabouço comparativo do Karma e LD-R

A Figura 2 enfatiza a concentração de recursos por classe nas ferramentas. Neste gráfico evidencia que o Karma possui mais recursos em todas as classes.

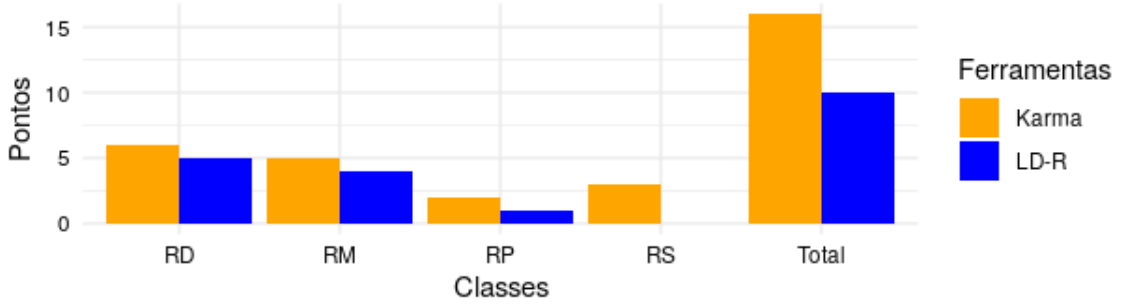

Figura 2. Arcabouço sumarizado do Karma e LD-R

5. Conclusão

Os dados em CSV são recorrentemente utilizados pela sua eficácia computacional. A integração semântica de dados agrega conhecimento aos dados tabulares. Diante dos trabalhos analisados, é possível perceber a complexidade envolvida ao gerar LD a partir de dados tabulares.

Na literatura as ferramentas semânticas e as linguagens de mapeamento foram analisadas em uma comparação em forma de arcabouço. A classificação proposta neste trabalho eleva o estado da arte ao agrupar os recursos por classe e utilizar uma reta numérica positiva. O arcabouço comparativo gera uma identificação ágil de quais recursos as ferramentas possuem, dessa forma fica evidente que o Karma é a ferramenta que fornece mais recursos para integração semântica de dados tabulares.

Portanto, este artigo não pretende esgotar o levantamento de recursos para realizar a integração semântica de dados tabulares. Os trabalhos futuros são: (i) incluir novos recursos; (ii) avaliar novas ferramentas.

Referências

Adamou, A. and D'Aquin, M. (2020). *Linked Data Principles for Data Lakes*, chapter 7, pages 145–169. John Wiley & Sons, Ltd.

Bizer, C., Heath, T., and Berners-Lee, T. (2011). Linked data: The story so far. In *Semantic services, interoperability and web applications: emerging concepts*, pages 205–227. IGI Global.

Crotti Junior, A., Debruyne, C., Brennan, R., and O'Sullivan, D. (2017). An evaluation of uplift mapping languages. *International Journal of Web Information Systems*, 13(4):405–424.

Dimou, A., Heyvaert, P., De Meester, B., and Verborgh, R. (2018). What factors influence the design of a linked data generation algorithm? In *LDOW2018 workshop, part of WWW2018, the International World Wide Web Conference*, pages 1–6.

Doan, A., Noy, N., and Halevy, A. (2004). Introduction to the special issue on semantic integration. *SIGMOD Record*, 33:11–13.

Hert, M., Reif, G., and Gall, H. C. (2011). A comparison of rdb-to-rdf mapping languages. In *Proceedings of the 7th International Conference on Semantic Systems*, pages 25–32. ACM.

Jung, H.-W., Kim, S.-G., and Chung, C.-S. (2004). Measuring software product quality: A survey of iso/iec 9126. *IEEE software*, 21(5):88–92.

Rashid, S. M., McCusker, J. P., Pinheiro, P., Bax, M. P., Santos, H., Stingone, J. A., Das, A. K., and McGuinness, D. L. (2020). The semantic data dictionary–an approach for describing and annotating data. *Data Intelligence*, pages 443–486.

Shafranovich, Y. (2005). Common format and mime type for comma-separated values (csv) files. RFC 4180, IETF.