Método de Integração Semântica Incremental de Dados Científicos Baseado em Ontologias

Marcello P. Bax[1], **José E. A. Gonçalves**[1]

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

bax@ufmg.br, jeugenio@ufmg.br

Abstract. *Data integration can be accomplished by using a common model with the global or local conversion of different repositories to this common model (Global as view and Local as view). Semantic integration, however, requires the use of ontologies. We propose a method for semantically integrating scientific data using metadata standards annotated by ontologies. The contribution comes from the fact that the proposed method is inspired by agile principles. Its use allows the incremental evolution of the ontology and an agile evaluation of the integration results.*

Resumo. *A integração de dados pode ser realizada pela utilização de um modelo comum com a conversão, global ou local, de diferentes repositórios para este modelo (Global as view e Local as view). A integração semântica, contudo, exige a utilização de ontologias. Propõe-se um método de integração semântica de dados científicos usando padrões de metadados anotados por ontologias. A contribuição advém do fato de que o método proposto inspira-se nos princípios ágeis de desenvolvimento. Seu uso permite a evolução incremental da ontologia e uma avaliação ágil dos resultados da integração.*

1. Introdução

Pode-se afirmar que a ciência entrou em um novo modo de operação [Fox and Hendler 2009]. A *eScience* combina tecnologia da informação e cibernética no apoio à investigação científica, incluindo a coleta, preparação e análise de dados [Bohle 2013]. Metodologias fundamentadas em tecnologias semânticas permitem a modelagem do conhecimento científico pela anotação e integração de dados e informações. A anotação baseada em ontologias apoia a descrição não ambígua dos dados e possibilita o seu reuso por outros estudos; além de viabilizar a reprodução dos resultados. A próxima seção menciona ferramentas baseadas em ontologias para anotação semântica e integração de dados tabulares. Por serem genéricas, contudo, elas abstraem o contexto em que são usadas, tornando sua aplicação pouco efetiva no contexto de pesquisas (estudos) científicas.

Propõe-se um método de anotação semântica incremental que considera as especificidades encontradas no ciclo de gerenciamento de dados de estudos científicos e faz uso de templates de metadados para anotar os dados. Uma vez interpretados pelo algoritmo de anotação, esses templates, em conjunto com os dados, geram descrições dos dados no formato de grafos RDF (*Resource Description Framework*), padrão de modelagem do

W3C[1] que permite organizar e adicionar semântica aos dados. Sua estrutura forma um grafo orientado. O objetivo é utilizar princípios de metodologias ágeis para tornar incremental um método de integração semântica de dados atualmente em desenvolvimento no *Tetherless World Constellation* (TWC), já utilizado em diversos projetos.

2. Trabalhos correlatos

A criação de uma plataforma de publicação e reutilização de dados depende do desenvolvimento de técnicas capazes de mapear dados tabulares para representações enriquecidas ou anotadas semanticamente por ontologias. Segundo [Ermilov et al. 2013], vários projetos, como o Apache Any23, o Triplify, o Tabels e o Open Refine foram motivados pela necessidade de facilitar a transformação de dados tabulares em estruturas de dados semanticamente vinculadas (*Linked Data*) [van der Waal et al. 2014].

Uma forma de tratar o problema é apresentada em [Rashid et al. 2017] e [Pinheiro et al. 2018b]. Pretende-se estender esta solução, adotando uma estratégia incremental de enriquecimento semântico. Acredita-se que a possibilidade de evoluir o grafo por alterações incrementais no modelo e nas ontologias relacionadas represente um avanço importante da presente proposta.

3. Integração semântica de dados científicos baseada em ontologias

A seguir serão apresentados os principais templates utilizados para anotar os dados científicos. O método será exemplificado utilizando a ontologia HAScO[2] (prefixo 'hasco:') e a ontologia SIO[3] (prefixo 'sio:') e uma ontologia de domínio "base" com definições específicas da plataforma e termos não definidos nas demais (prefixo ':').

3.1. Desenho Semântico do Estudo (SSD)

O método proposto para anotar os dados de um estudo científico inicia-se com a descrição do mesmo em termos de seus objetos. Para esse fim, utiliza-se o template denominado "Desenho Semântico do Estudo" (SSD), que descreve as coleções de objetos que o pesquisador analisa em sua pesquisa. No *dataset* de exemplo da Figura 1(a), o pesquisador determinou que serão pesquisados 25 sujeitos humanos, identificados pelos Ids "01" a "25". A Figura 1(a) traz um conjunto simplificado de dados, coletados durante entrevistas e exames laboratoriais. Cada linha da tabela representa um registro de dados de um participante da pesquisa (criança recém-nascida). Tem-se *Id*, *Age*, *Sample1*, *Sample2* e *MotherEducation*. Parte do grafo RDF gerado pelo método para representar os objetos desse estudo pode ser visto na Figura 1(b). Tem-se o objeto `Study A`, composto (*hasco:hasCollection*) por uma coleção `SubjectGroup G`, vinte e cinco participantes são membros (*hasco:isMemberOf*) dessa coleção e cada participante tem um *Id* originalmente associado a ele (*hasco:hasOriginalId*).

O Estudo A, descrito na Figura 1(b) pode ser ainda mais especificado, de forma incremental (cf. Figura 1(c)), adicionando-se uma coleção de amostras de urina coletadas dos indivíduos (`Sample Collection C`), sendo duas amostras por indivíduo (*Sample1* e *Sample2*). Uma delas (*Sample1*, p.ex.) é coletada em um determinado mês e a outra (*Sample2*) um mês após o dia de coleta da primeira amostra.

[1] `https://www.w3.org/RDF/`
[2] `http://hadatac.org/ont/hasco/`
[3] `https://bioportal.bioontology.org/ontologies/SIO`

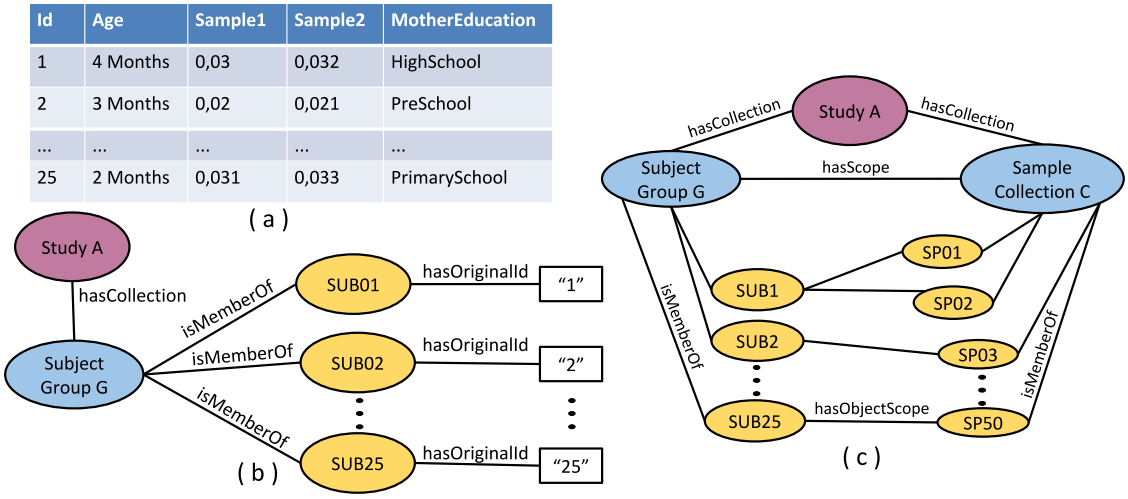

Figura 1. Os grafos (b) e (c) são a especificação incremental do desenho de um dado Estudo A, a partir do arquivo de dados (a).

O SSD é um template de metadados expressos em formato tabular (cf. Tabela 1). Cada linha descreve uma coleção de objetos. Dessa forma, conforme o exemplo do "Estudo A", o SSD deverá definir as coleções de participantes e de amostras de urina. Uma outra coleção representa os meses de coleta das amostras de urina (mês 1 e mês 2).

Tabela 1. Especificação do SSD para o Estudo A (STD).

id	type	isMemberOf	hasScope	hasTimeScope	cardinality
:STD	hasco:Study				
:STD-SUBJECTS	hasco:SubjectGroup	:STD			25
:STD-URINE	hasco:SampleCollection	:STD	:STD-SUBJECTS	:STD-MONTHS	1
:STD-MONTHS	hasco:TimeCollection	:STD			2

Observe no SSD da Tabela 1, que a coleta de amostras de urina (STD-URINE) tem por escopo (*hasScope*) o grupo de participantes (STD-SUBJECTS). Já a coleção de meses (STD-MONTHS) é definida no escopo temporal (*hasTimeScope*). A cardinalidade indica o número de elementos de uma coleção para um dado escopo. Assim, a cardinalidade 25 para STD-SUBJECTS estabelece que temos um grupo de 25 sujeitos. A cardinalidade 1 para a coleção STD-URINE indica que temos uma amostra para cada combinação STD-SUBJECTS e STD-MONTHS. Já a cardinalidade 2 para STD-MONTHS denota que cada sujeito terá amostras em 2 meses distintos. A interpretação do SSD dá origem ao grafo RDF da Figura 1(b) e (c).

3.2. Dicionário Semântico de Dados (SDD)

Após definir quais serão as coleções de dados de um estudo, estes devem ser instanciados como valores de atributos de objetos existentes nessas coleções. A especificação do "Dicionário Semântico de Dados" (SDD) permite definir os atributos e relações entre os objetos identificados de forma explícita ou implícita pelos dados dos estudos. Por exemplo, se crianças são os sujeitos principais de um estudo, seus atributos aparecem como colunas dos arquivos de dados tabulares (*dataset* do estudo). Porém, podem aparecer também atributos de outros objetos, como as mães das crianças, por exemplo. Portanto, se cada linha da tabela identifica uma criança, a cada vez que uma criança é instanciada

pela ingestão de uma linha de dados, um "objeto" mãe dessa criança também deverá ser criado no grafo RDF resultante. Diz-se que os objetos do tipo "mãe" estão implícitos no *dataset*.

Tabela 2. Dicionário Semântico de Dados (SDD)

Label	Attribute	IsAttributeOf	Entity	Role	inRelationTo	wasDerivedFrom
Id	hasco:originalID	??child				
Age	:Age	??child				
Sample1	:SolutionPH	??sample				
Sample2	:SolutionPH	??sample				
MotherEducation	:EducationLevel	??mother				
??child			sio:Human	:hasChild	??mother	
??sample			:Urine			??mother
??mother			sio:Human	:hasMother	??child	

A Tabela 2 apresenta o SDD que descreve o arquivo de dados da Figura 1(a), onde cada linha/registro representa uma criança. Quando o SDD é utilizado pelo algoritmo de anotação, o processamento do SSD já foi realizado para criar as coleções de objetos do estudo, associando um identificador para cada objeto da coleção criada. Assim, o SSD da Tabela 1 criou uma coleção de crianças (STD-SUBJECTS). O atributo *hasco:originalID* sinaliza ao processador que o *Id* (1a. coluna do arquivo de dados) deve ser associado ao identificador criado anteriormente no processamento do SSD. Os outros atributos são também atributos da mesma criança (no caso apenas `Age` *IsAttributeOf* de `??child`). Dito de outra forma, o *hasco:originalID* designa um objeto do estudo que já existe no grafo RDF no momento em que cada linha de um arquivo de dados é ingerida. Desde que o objeto não possua o atributo *hasco:originalID*, a expressão "??" representa um objeto implícito, ou seja, um objeto do estudo que será adicionado ao grafo no momento em que um arquivo de dados é ingerido, e não no momento da interpretação do SSD. Os objetos implícitos não aparecem explicitamente nas colunas dos arquivos de dados do estudo.

A partir da primeira linha (*Id*=1) do arquivo de dados mostrado na Figura 1(a) e da interpretação do SDD da Tabela 2, gera-se o seguinte grafo RDF:

```
:SUB01 rdf:type sio:Human;
   :Age "4 Months".
:SP01 rdf:type :UrineSample;
   :hasValue "0,03".
:SP02 rdf:type :UrineSample;
   :hasValue "0,032"
:SUB01 :hasMother :MSUB01.
:MSUB01 rdf:type sio:Human;
   :EducationLevel "HighSchool".
:MSUB01 :hasChild :SUB01;
:SP01 :wasDerivedFrom :MSUB01.
:SP02 :wasDerivedFrom :MSUB01.
```

As demais linhas, até a linha 25 do arquivo de dados, gerarão a continuidade do grafo RDF acima, seguindo o mesmo procedimento, descrito a seguir:

Considerando o *dataset* da Figura 1(a): a primeira e a segunda colunas da primeira linha do SDD especificam o *Id*=1 e o atributo *hasco:originalID*. Da primeira linha do SDD (na Tabela 2), deduz-se que o objeto referido pelo registro na Figura 1(a) é `:SUB01`

(criado no momento da interpretação do SSD, quando foi gerada uma coleção de sujeitos (`Subject Group G`) contendo 25 identificadores (`:SUB01` até `:SUB25`)). A coluna seguinte do SDD (*IsAttributeOf*) determina, portanto, que `??child` é representado por `:SUB01`; o que significa que o objeto em questão, ao interpretar aquele registro na Figura 1(a), é aquele identificado por `:SUB01`, que já existe no grafo (cf. Figura 1(b)). O `:SUB01` é o identificador da primeira criança do arquivos de dados, cf. Figura 1(a). A partir da segunda linha do SDD e também da Figura 1(a), adiciona-se a tripla `:SUB01` *:Age* `"4 Months"` ao grafo que está sendo construído.

Pela especificação do SDD (Tabela 2), vemos que o objeto `??child` está relacionado ao objeto implícito `??mother` (*inRelationTo*), sendo gerado o novo objeto `:MSUB01` (identificador gerado internamente e atribuído a objetos implícitos) e associando `??mother` com `:MSUB01`. As triplas `:MSUB01` *rdf:type* `sio:Human` e `:SUB01` *:hasMother* `:MSUB01` são derivadas da segunda e terceira colunas do SDD. A tripla `:MSUB01` *:EducationLevel* `"HighSchool"` vem da quinta linha do SDD. Seguindo o mesmo procedimento, os demais objetos `SUB02` até `SUB25` (da Figura 1(a)) são criados juntamente com seus objetos implícitos relacionados. Em seguida, o segundo registro do arquivo de dados será associado ao Id `:SUB02`; e usando a segunda linha do SDD (Tabela 2) e da Figura 1(a), adiciona-se a tripla `:SUB02` *:Age* `"3 Months"` ao grafo que está sendo construído.

3.3. Construindo o grafo de forma incremental

Em relação ao processo de ingestão original, tal como descrito até aqui e apresentado em [Pinheiro et al. 2018a], o método proposto por esta pesquisa prevê que a construção dos arquivos de metadados (SSD e SDD) seja realizada de forma incremental e apoiada por *software*. Numa primeira versão, o pesquisador deve utilizar-se de um modelo conceitual temporário, simples e idiossincrático (a ontologia "base") para descrever os objetos pesquisados, seus atributos e relações. Logo, não seria necessário investir esforço adicional concebendo a ontologia de domínio antes de obter uma primeira versão do grafo RDF. Esta primeira versão do grafo constitui-se num artefato que pode ser utilizado, o mais cedo possível, no ciclo da pesquisa. Desse modo, os dados pesquisados podem ser compreendidos e comunicados pelo pesquisador o quanto antes, em busca de avaliação e *feedbacks* (por parte de seus pares).

O fluxo de trabalho proposto pelo projeto de tese gerencia esse processo incremental, permitindo a iteração organizada por ciclos de desenvolvimento entre o pesquisador e o ontologista durante todo o processo. Assim, à medida em que avança a sua compreensão sobre os fenômenos/objetos investigados, o pesquisador vai ajustando incrementalmente seus SSDs e SDDs para incorporarem a versão em evolução da ontologia de domínio (derivada incrementalmente da ontologia "Base"). Isso se justifica porque mesmo que o pesquisador possua uma compreensão dos objetos do domínio pesquisado, ele não necessariamente domina eventuais formalizações de vocabulários pré-existentes estabelecidos por ontologias do domínio. Portanto, propõe-se que o reuso dos termos existentes seja feito aos poucos para não se tornar um empecilho ao avanço do trabalho de pesquisa. Tal reuso é importante pois um grafo RDF de qualidade deve reutilizar termos de outras ontologias já consolidadas e minimizar a referência aos termos da ontologia "base". Assim, a evolução incremental permitirá que se obtenha um maior reuso a cada ciclo.

4. Conclusão

Apesar da diversidade de tipos de estudos científicos, nossa hipótese é que a consideração de aspectos presentes em sua maioria poderá tornar nosso método de anotação semântica mais adequado do que os trabalhos correlatos citados na Seção 2 para integrar esse tipo de dado. Os metadados no contexto do design do estudo (SSD) e aqueles no contexto do dicionário semântico de dados (SDD) devem advir de ontologias consolidadas. Porém, essa busca por reutilização de termos é demorada e complexa, devendo ser realizada com o auxílio de um ontologista. Além disso, a incrementalidade do enriquecimento semântico traz vantagens, pois os dados podem ser anotados, interpretados e comunicados mais prontamente.

Dessa forma, três requisitos principais norteiam a pesquisa do processo de anotação proposto: (1) facilidade de uso - o método deve poder ser utilizado por pesquisadores que não são especialistas em Web Semântica; (2) contexto específico - diferentemente de outras abordagens genéricas o processo descrito é pensado especificamente para o contexto da integração de dados em estudos científicos; (3) incrementalidade - o grafo que organiza os dados integrados deve poder ser criado de forma incremental. O atendimento aos requisitos acima não somente justifica o método proposto nesta pesquisa, como permite definir formas de avaliá-lo.

Referências

Bohle, S. (2013). What is e-science and how should it be managed? nature.com. *Spektrum der Wissenschaft, http://www. scilogs. com/scientific_and_medical_libraries/what-is-e-scienceand-how-should-it-be-managed*.

Ermilov, I., Auer, S., and Stadler, C. (2013). Csv2rdf: User-driven csv to rdf mass conversion framework. In *Proceedings of the ISEM*, volume 13, pages 04–06.

Fox, P. and Hendler, J. A. (2009). Semantic escience: encoding meaning in next-generation digitally enhanced science. *The Fourth Paradigm*, 2.

Pinheiro, P., Bax, M., Santos, H., Rashid, S. M., Liang, Z., Liu, Y., McCusker, J. P., and McGuinness, D. L. (2018a). Annotating diverse scientific data with hasco. In *Proceedings of the Seminar on Ontology Research in Brazil*.

Pinheiro, P., Santos, H., Liang, Z., Liu, Y., Rashid, S. M., McGuinness, D. L., and Bax, M. P. (2018b). Hadatac: A framework for scientific data integration using ontologies. In *Proceedings of the ISWC*.

Rashid, S. M., Chastain, K., Stingone, J. A., McGuinness, D. L., and McCusker, J. (2017). The semantic data dictionary approach to data annotation & integration. In *SemSci@ ISWC*, pages 47–54.

van der Waal, S., Węcel, K., Ermilov, I., Janev, V., Milošević, U., and Wainwright, M. (2014). Lifting open data portals to the data web. In *Linked Open Data–Creating Knowledge Out of Interlinked Data*, pages 175–195. Springer.