XX ENANCIB
21 a 25 Outubro/2019 – Florianópolis
A Ciência da Informação e a era da Ciência de Dados

ISSN 2177-3688

GT-8 - Informação e Tecnologia

GRAFOS DE CONHECIMENTO PARA PREPARAÇÃO E REUTILIZAÇÃO DE DADOS CIENTÍFICOS

KNOWLEDGE GRAPHS FOR SCIENTIFIC DATA PREPARATION AND REUSE

Marcello P. Bax - Escola de Ciência da Informação, Universidade Federal de Minas Gerais
José E. A. Gonçalves - Escola de Ciência da Informação, Universidade Federal de Minas Gerais

Modalidade: Trabalho Completo

Resumo: A Ciência da Informação deve instrumentalizar os processos de gerenciamento de dados na *eScience*, viabilizando a maior integração e reuso de dados científicos. Este é um problema não resolvido, presente nos ciclos de gestão de dados científicos. A dificuldade de reuso leva à redundância de esforços custosos para os laboratórios e agências de fomento. É possível aumentar as taxas de reuso de dados entre pesquisas com a sua integração através de ontologias. Este artigo apresenta um método "ágil" de integração semântica de dados, que realiza a ingestão de conjuntos de dados (*datasets*), produzidos por diferentes estudos, na forma de grafos de conhecimento. O método utiliza uma ontologia de fundação estruturada como um "Dicionário Semântico de Dados", que, integrada a outras ontologias de domínio, gera um grafo de conhecimento. Este grafo facilita o reuso dos dados já que pode integrar conceitualmente dados oriundos de várias fontes em um único repositório. O grafo permite ao usuário navegar por facetas e escolher as dimensões de seu interesse. A organização facetada dos dados permite a seleção flexível e granular dos mesmos nos *datasets* integrados, fomentado o reuso e facilitando a tarefa de preparação de dados. O método apoia-se em uma adaptação da *design science*, com elementos do desenvolvimento ágil de sistemas, que permitem obter e avaliar resultados mais rapidamente (do que os métodos clássicos) e corrigir problemas precocemente. A hipótese de que a integração semântica é facilitada com o uso do método tem sido confirmada pelos resultados de sua aplicação em um projeto de integração de dados em epidemiologia.

Palavras-chave: Ontologias; Integração semântica de dados; Grafos de conhecimento; Web semântica

Abstract: Information Science must instrumentalize eScience data management processes, enabling greater integration and reuse of scientific data. This is an unresolved problem present in scientific data management cycles. The difficulty of reuse leads to the redundancy of costly efforts for laboratories and funding agencies. Data reuse rates can be increased by integrating them across ontologies. This paper presents an "agile" method of semantic data integration, which ingests datasets produced by different studies in the form of knowledge graphs. The method uses a foundation ontology structured as a "Semantic Data Dictionary", which, integrated with other domain ontologies, generates a knowledge graph. This graph facilitates data reuse as it can conceptually integrate data from multiple sources into a single repository. The graph allows the user to navigate through facets and choose the dimensions of interest. Faceted data organization enables flexible and granular selection of data in integrated datasets, fostering reuse and easing the task of data preparation. The method relies on an adaptation of design science, with elements of agile system development, which enable results to be obtained and evaluated faster (than classical methods) and to correct problems early. The hypothesis that semantic integration is facilitated using the method has been confirmed by the results of its application in an epidemiology data integration project.

Keywords: Ontologies; Semantic data integration; Knowledge graphs; Semantic Web.

1 INTRODUÇÃO

Pode-se afirmar que a ciência entrou em um novo modo de operação, tornando-se cada vez mais dependente de grandes volumes de dados, que as tecnologias tradicionais não estão preparadas para gerenciar (FOX; HENDLER, 2009). A *eScience* surge como uma forma de endereçar esses desafios (BOHLE, 2013). Ela combina tecnologia da informação e cibernética no apoio à investigação científica, incluindo coleta, preparação, organização e análise de dados científicos. Neste novo contexto do fazer científico, em que prepondera a chamada *data science*, metodologias fundamentadas em tecnologias semânticas permitem a modelagem do conhecimento científico com o uso de ontologias. Essa modelagem possibilita a anotação dos dados com metadados, facilitando a sua integração, sobretudo quando os dados são oriundos de fontes diversas. Facilitar o processo de preparação dos dados coletados pelas pesquisas e auxiliar na sua integração e reuso são contribuições necessárias para o avanço da ciência na era da *data science.*

O fluxo de trabalho na *data science* inicia-se com a coleta e preparação dos dados e finaliza-se com a sua análise na busca por novos conhecimentos. Um dos problemas não resolvidos dos ciclos de gestão dos dados científicos (fluxo que vai da preparação à análise) é como promover a reutilização dos dados gerados por estudos correlatos, integrando-os àqueles já disponíveis. A dificuldade de reuso dos dados por variados estudos científicos leva à redundância de esforços que são altamente custosos para os laboratórios e agências de fomento. Essa dificuldade é também um desafio quando se trata de reproduzir os resultados

dos estudos. Como dito acima, uma forma de aumentar as taxas de reuso dos dados entre as pesquisas seria promover a sua integração semântica com o uso de ontologias (HITZLER, 2016). A anotação baseada em ontologias apoia a descrição não ambígua dos dados, possibilita o seu reuso por outros estudos e viabiliza a reprodução dos resultados. Tecnologias semânticas, sobretudo o padrão RDF (*Resource Description Framework*), têm sido utilizadas para formalizar a descrição dos dados, anotando-os de forma sistemática com metadados (BRODARIC; GAHEGAN, 2010; WACHE et al., 2001).

O artigo apresenta uma abordagem que utiliza a ADSRM (*Agile Design Science Research Method*) proposta em (CONBOY; GLEASURE; CULLINA, 2015) e detalhada na Seção 5, para tornar incremental um método de integração semântica de dados atualmente em desenvolvimento no *Tetherless World Constellation* (TWC), já utilizado em diversos projetos. Esse método proposto em Rashid et (2017) e Pinheiro et al (2018), realiza a "ingestão" de dados tabulares de diferentes pesquisas em um grafo de conhecimento usando tecnologias semânticas (padronizadas pelo W3C) e faz uso de gabaritos (*templates*) de metadados para anotar os dados. Uma vez interpretados pelo algoritmo de anotação, esses *templates*, em conjunto com os dados, são integrados ao grafo[1] de conhecimento. O método adere aos princípios FAIR (WILKINSON et al., 2016) que permitem avaliar a qualidade dos conjuntos de dados publicados e dos métodos utilizados para sua disponibilização.

A abordagem proposta aqui atualiza o método do TWC, prescrevendo a sua aplicação organizada por ciclos incrementais de modelagem conceitual do problema e considera as especificidades encontradas no ciclo de gerenciamento de dados de estudos científicos. Como se verá mais adiante no texto, a proposição do método original apoia-se em uma ontologia de fundação que, de forma integrada a outras ontologias de domínio, por um processo de ingestão dos dados, gera um grafo de conhecimento que pode ser explorado de forma facetada. A gestão do grafo fomenta o reuso dos dados já que integra dados de várias fontes em um repositório único e permite que o pesquisador navegue por suas facetas e escolha as dimensões que caracterizam os dados integrados pelo grafo, que deseja explorar. A organização facetada dos dados permite a seleção flexível e granular dos mesmos, facilitando o seu entendimento, integração e reuso.

[1]Ou grafos RDF - *Resource Description Framework*, padrão de modelagem W3C (<https://www.w3.org/RDF/>) que permite organizar e adicionar semântica aos dados.

A motivação para incorporar a ADSRM ao método original, é a falta de agilidade deste último na realização da anotação dos dados e geração do grafo. Isso decorre da dificuldade do método original, que deve dispor de uma ontologia de domínio já madura ou finalizada para, somente então, realizar a ingestão dos dados e gerar o grafo. Pesquisadores que geram dados se beneficiarão de mecanismos mais ágeis e eficientes para permitir o armazenamento, a integração e a análise de dados de seus experimentos, podendo participar mais ativamente de todo processo. O aumento das taxas de reuso com a integração semântica de dados é reforçado com o método proposto, que permite essa integração de forma mais eficiente devido às características da metodologia ADSRM, as quais serão explicitadas na Seção 5.1.

As duas próximas seções introduzem os conceitos necessários para o entendimento da proposta e apresentam os principais trabalhos correlatos: abordagens e ferramentas baseadas em ontologias que visam facilitar a anotação semântica e integração de dados tabulares. A Seção 3 também detalha e discute o método de integração semântica tal como originalmente proposto pelo TWC. A Seção 4 descreve o *framework* utilizado para realizar a ingestão dos dados e persisti-los no formato de grafo de conhecimento. Em seguida a Seção 5 detalha o método incremental proposto. Para completar teoricamente a avaliação da abordagem proposta, o resultado do processo de anotação dos dados é avaliado na Seção 6. Finalmente a Seção 7 apresenta as conclusões da pesquisa, suas limitações e trabalhos futuros.

2 TECNOLOGIAS APLICADAS À *E-SCIENCE*

2.1 Ontologias e Grafos de Conhecimento

A preparação de conjuntos de dados - *datasets* - com vistas ao seu reuso por estudos correlatos se beneficia do uso de ontologias para anotar dados com metadados e promover a sua integração entre diferentes estudos. A integração de dados com a anotação semântica permite a interoperabilidade e reuso de dados entre estudos. Implementações de plataformas para gestão de dados científicos, ontologias e grafos de conhecimento têm sido utilizadas para instrumentalização da *eScience* (MCCUSKER et al.,2018; PINHEIRO et al., 2018b).

2.2 Tecnologias Semânticas

Ontologias fazem uso de URIs (*Uniform Resource Locators*), que são úteis para interligar diferentes ontologias, pois servem como identificadores. URIs identificam recursos próprios de uma ontologia ou referenciam recursos de outras ontologias (BERNERS-LEE, 2001). Um URI tem escopo global na web e é interpretado consistentemente entre contextos. Associar um URI a um recurso significa que ele pode ser referenciado, interligado ou recuperado por qualquer pessoa ou máquina que possua o URI (SHADBOLT; BERNERS-LEE; HALL, 2006).

A representação é padronizada, facilitando a interoperabilidade entre ontologias e permitindo uma forma consistente de acesso aos dados. A partir da linguagem padrão SPARQL[2] (W3C et al., 2013), por exemplo, é possível manipular os conceitos representados por uma ou mais ontologias (PÉREZ; ARENAS; GUTIERREZ, 2009).

2.3 Trabalhos correlatos

A criação de uma plataforma de publicação e reutilização de dados se beneficia cada vez mais das tecnologias semânticas padronizadas pelo W3C, mais especificamente do desenvolvimento de técnicas capazes de mapear dados tabulares para representações enriquecidas ou anotadas semanticamente por ontologias. Vários projetos como o Apache Any23[3], o Triplify (AUER et al., 2009), o Tabels, o Open Refine[4] e o Karma (GUPTA et al., 2012) foram motivados pela necessidade de facilitar a transformação de dados tabulares em estruturas de dados semanticamente vinculadas e representadas por grafos (*Linked Data*) (ERMILOV; AUER; STADLER, 2013; WAAL et al., 2014). Por serem genéricas, contudo, a maioria dessas abordagens e ferramentas abstraem o contexto em que são usadas, tornando a sua aplicação menos efetiva no escopo mais específico de dados oriundos de pesquisas científicas.

Uma proposta de solução correlata para tratar o problema, no contexto específico de dados científicos, é descrita em detalhes na Seção 3. Pretende-se, com o trabalho de pesquisa aqui relatado, estender esta solução, adotando uma estratégia incremental de enriqueci-

[2] Um acrônimo recursivo para: Protocolo e linguagem de consulta RDF SPARQL

[3] <https://any23.apache.org>

[4] <http://openrefine.org>

mento semântico que suporte o esforço de compartilhar dados no contexto de estudos científicos. As ideias iniciais que fundam a extensão proposta são descritas na Seção 5. A possibilidade de evoluir o grafo por alterações incrementais, no modelo e nas ontologias relacionadas, representa uma contribuição relevante da presente pesquisa.

3 INTEGRAÇÃO SEMÂNTICA DE DADOS CIENTÍFICOS

O processo de integração semântica de dados científicos, ao referenciar ontologias externas, se beneficia da semântica adicionada pela reutilização de termos formalmente definidos em cada uma dessas ontologias. Define-se Ingestão de Dados como o processo de obtenção e importação de dados para armazenamento em um repositório que integra semanticamente os dados ingeridos. Este processo é descrito a seguir.

3.1 Processo de Ingestão Semântica de Dados

3.1.1 Ontologias utilizadas

O processo de ingestão semântica de dados, tal como proposto em (RASHID et al., 2017) e (PINHEIRO et al., 2018b), envolve a utilização de um conjunto de ontologias relacionadas ao domínio do estudo, bem como de ontologias que fornecem a infraestrutura básica para o processo. Apresenta-se um exemplo de ingestão que utiliza as ontologias listadas na Tabela 1. Cada recurso de uma ontologia será identificado com um prefixo conforme mostra a tabela. Além destas ontologias, tem-se um protótipo rudimentar de ontologia de domínio, que no início do processo de ingestão é denominada “base” e identificada com o prefixo ’:’ (designando a ontologia local). Esta ontologia possui definições específicas dos conteúdos do estudo científico local em questão, cujos dados estão sendo anotados e integrados. Muitas vezes, no início do processo, definições precisas para estes conceitos ainda não existem ou ainda não foram encontradas em uma outra ontologia de referência. Estas definições imprecisas serão substituídas futuramente, à medida que iterações cíclicas de ingestão de dados acontecem. Além disso, é na ontologia “base” que o pesquisador faz a qualificação de certos dados para que sejam considerados como indicadores científicos do estudo em questão. Estes indicadores podem ser visualizados posteriormente, após a ingestão dos dados na plataforma, pela interface de busca facetada.

Tabela 1: Ontologias utilizadas no exemplo de ingestão semântica de dados

Ontologia	Prefixo	Referência
Semantic Science Ontology (SIO)	sio:	http://semanticscience.org/resource
Human-Aware Science Ontology (HAScO)	hasco:	http://hadatac.org/ont/hasco/

Fonte: Elaborada pelos autores.

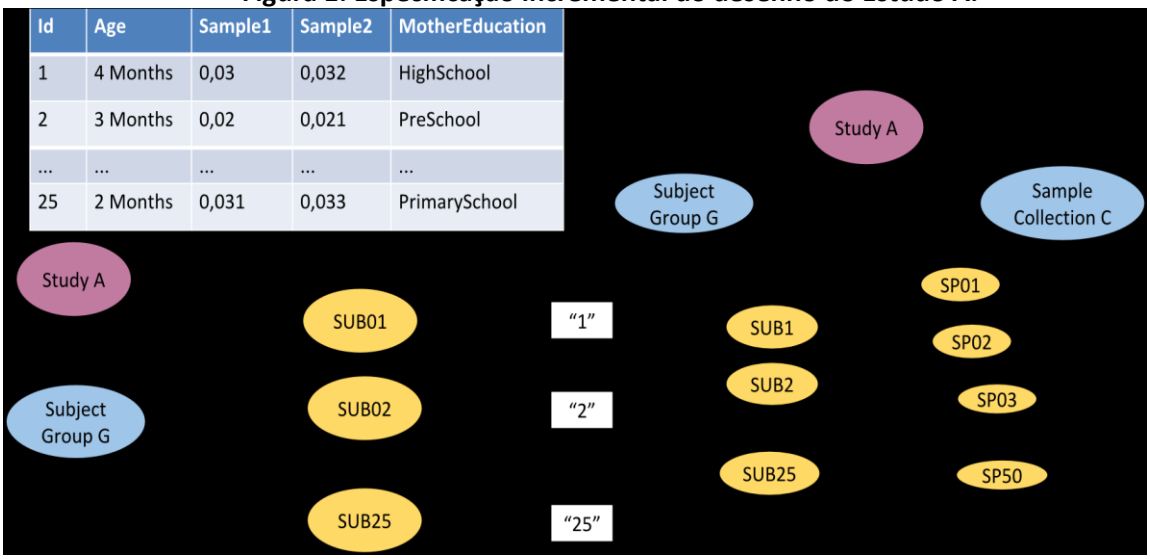
A primeira ontologia da Tabela 1, a *Semantic Science Ontology* (SIO), é a ontologia de fundamentação do método de anotação e define os conceitos e relações usados para descrever objetos, atributos, processos e a dimensão temporal dos estudos científicos (DUMONTIER et al., 2014). O uso da SIO, em conjunto com ontologias de domínio, permite que os cientistas usuários anotem as entidades e atributos que são objetos de estudo em seus domínios científicos mais específicos. A SIO fornece a estrutura integrada a partir da qual a segunda ontologia, a HAScO, está enraizada. A HAScO (PINHEIRO et al., 2018a) permite modelar estudos de domínios diversos para a anotação de dados e faz parte do *framework* que será apresentado na Seção 4.

O processo de ingestão pressupõe um conjunto de *templates* de metadados que são preenchidos pelo pesquisador para anotar os dados. Os principais *templates* utilizados no método definido por (PINHEIRO et al., 2018b) serão apresentados a seguir.

3.1.2 Organizando os objetos do estudo: Desenho Semântico do Estudo (SSD)

Realizar o Desenho Semântico do Estudo (SSD) é o primeiro passo do método. Os metadados que compõem o SSD formam o escopo inicial de descrição dos objetos relevantes para a pesquisa cujos dados estão sendo preparados: organizados em coleções, anotados e interligados. Dessa forma, inicia-se com a descrição do estudo em termos de seus principais objetos componentes. O SSD contém as coleções de objetos que o pesquisador analisa em sua pesquisa. Ou seja, usando o SSD o pesquisador descreve os objetos existentes no contexto do estudo, tais como são conhecidos à época em que o estudo foi concebido (parte inicial do design da pesquisa).

Figura 1: Especificação incremental do desenho do Estudo A.



Fonte: Elaborada pelos autores.

De acordo com o *dataset* da Figura 1(a), o pesquisador pretende observar 25 sujeitos humanos, identificados pelos Ids "01" a "25", 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. Nesse exemplo, os principais "objetos" cujos dados são coletados são crianças recém-nascidas. 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*[5]) 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 exemplo "Estudo A", da Figura 1(b) poderia 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.

O SSD é expresso em formato tabular (cf. Tabela 2). Nele, cada linha descreve uma coleção para os principais objetos que são caracterizados pelos dados a serem coletados no

[5] Os termos *hasCollection*, *isMemberOf*, *hasOriginalId* denotam conceitos definidos na Ontologia HAScO, criada para descrever estudos científicos (PINHEIRO et al., 2018a).

contexto da pesquisa. De acordo com o exemplo, o SSD define as coleções de participantes, de amostras de urina e dos meses de coleta das amostras de urina (mês 1 e mês 2).

Tabela 2: *Template* SSD para o Estudo A.

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

Fonte: Elaborada pelos autores.

Observe no SSD da Tabela 2, que a coleta de amostras de urina (STD-URINE) tem por escopo (*hasScope*) o grupo de participantes (STD-SUBJECTS) e o escopo temporal (*hasTimeScope*) definido pela coleção de meses (STD-MONTHS). 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 aos grafos RDF da Figura 1(b) e (c).

A próxima seção descreve o segundo contexto de geração de metadados da abordagem, que é o Dicionário Semântico de Dados (SDD). O SDD é responsável pela modelagem conceitual (ontológica) dos dados analisados no estudo e por guiar a criação do grafo RDF final. Nota-se que, no processo de criação do grafo RDF final que integra os dados do estudo, os objetos são criados no grafo em dois momentos.

a) Durante o processamento do SSD: realizado na preparação para ingestão dos dados, criando no grafo RDF as Coleções de objetos definidas pelo SSD;
b) Durante o processamento do SDD e da ingestão dos dados, que é realizada com o uso dos metadados definidos pelo SDD. O SDD é explicado na Seção 3.1.3.

3.1.3 Relacionando os objetos do estudo entre si: o Dicionário Semântico de Dados (SDD)

Após definir quais serão as coleções de dados do estudo (cujos dados estão sendo ingeridos e organizados no grafo RDF), estes dados devem ser instanciados no grafo como valo-

res 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 representadas no *dataset* da Figura 1(a) são os sujeitos principais do estudo, seus atributos aparecem como colunas dos arquivos de dados (*dataset* do estudo). Porém, no mesmo *dataset* podem aparecer também atributos de outros objetos implicitamente representados pelos dados, como, por exemplo, atributos das mães das crianças. 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*.

Estes objetos implícitos no *dataset* serão então explicitados e formalizados no grafo final gerado. É exatamente essa explicitação e formalização de objetos antes implícitos que favorece a integração de dados originários de outros *datasets*, gerados por estudos correlatos. Uma vez explicitados esses dados podem servir como "pontes de ligação" com os dados de outros estudos, permitindo a sua integração. Esses novos dados explicitados e formalizados podem ser tratados de forma automática por algoritmos diversos, já que são "compreensíveis" por máquinas. Alavanca-se assim o potencial para complementar os dados pesquisados no estudo corrente, pela integração semântica de outros dados, enriquecendo o conjunto total de dados relevantes para o entendimento dos fenômenos e facilitando a sua preparação para o teste de hipóteses correlatas e/ou complementares.

Tabela 3: 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	

Fonte: Elaborada pelos autores.

A Tabela 3 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 (em etapa anterior) para criar as coleções de objetos do estudo, associando um identificador para cada objeto da coleção criada. Assim, o SSD da Tabela 2 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. Outros atributos da mesma criança (no caso apenas Age *IsAttributeOf* de ??child) são também associados ao identificador, ou seja, ao objeto referente à criança com aquele *Id*. 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. Os objetos sem o atributo *hasco:originalID* são objetos implícitos, ou seja, objetos do estudo que serão adicionados ao grafo quando o arquivo de dados for ingerido e não no momento da interpretação do SSD. Assim, a partir da primeira linha (*Id*=1) do arquivo de dados mostrado na Figura 1(a) e da interpretação do SDD da Tabela 3, gera-se o seguinte grafo RDF:

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

As demais linhas, até a linha 25 do arquivo de dados, gerarão a continuidade do grafo acima, seguindo o 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 3), 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 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 arquivo de dados, cf.

Figura 1(a). A partir da segunda linha do SDD e da Figura 1(a), adiciona-se a tripla :SUB01 *:Age* "4 *Months*" ao grafo que está sendo construído.

Pela especificação do SDD (Tabela 3), vemos que o objeto ??child está relacionado ao objeto implícito ??mother (*inRelationTo*), sendo gerado o novo objeto :MSUB01 (identificador gerado internamente e atribuído ao objeto implícito) 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.

O processamento dos *templates* e arquivos de dados desta seção pode ser realizado com o *framework* que será assunto da próxima seção.

4 FRAMEWORK UTILIZADO PARA A GERAÇÃO DO GRAFO DE CONHECIMENTO

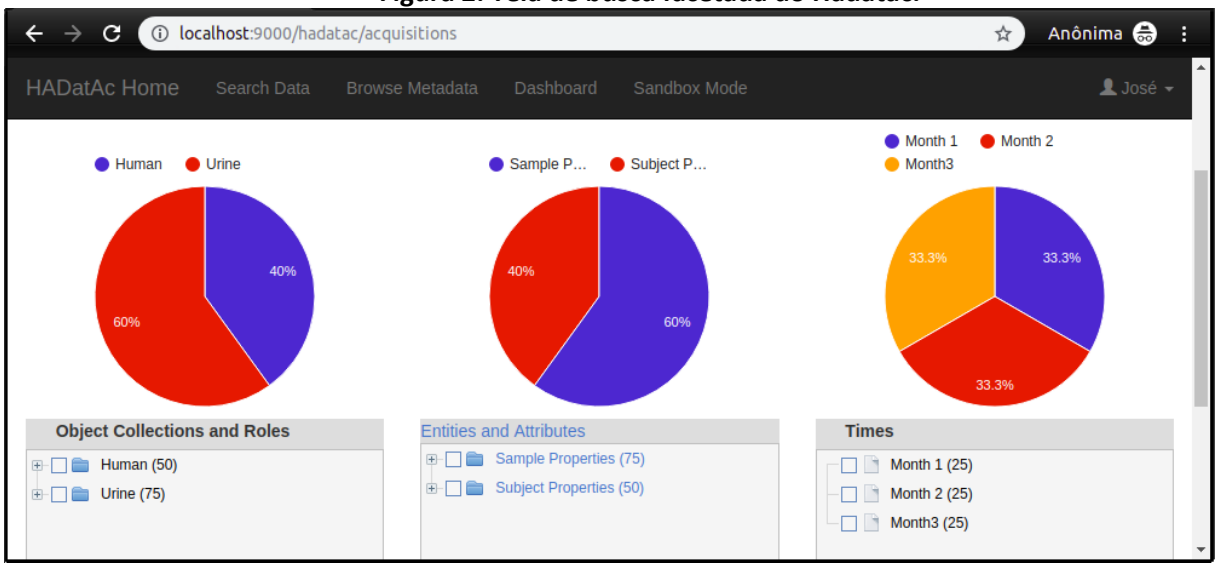
O método descrito na seção anterior foi implementado em um *framework* desenvolvido pelo TWC, o HADatAc (*Human-Aware Data Acquisition framework*)[6]. Ele é responsável pelo processamento da ingestão de dados, interpretando os *templates* apresentados na Seção 3 e armazenando-os em um grafo para serem utilizados posteriormente.

O método de ingestão de dados e geração do grafo de conhecimento apresentado inclui as seis etapas a seguir:

a) definição do *dataset*, contendo os dados a serem ingeridos, como um novo estudo no HADatAc;
b) definição e ingestão do SSD para gerar as coleções como instâncias RDF;
c) definição e ingestão do SDD que identifica como o conteúdo dos arquivos de dados será extraído, integrado e harmonizado;
d) ingestão e processamento dos arquivos de dados carregados;
e) acesso ao conteúdo do grafo integrado no HADatAc, como mostrado na Figura 2.

[6] Veja em: <http://hadatac.org>

Figura 2: Tela de busca facetada do Hadatac.



Fonte: Elaborada pelos autores.

Na tela de busca facetada exibida na Figura 2, o pesquisador pode, por exemplo, fazer o download de conjuntos de dados (*datasets* que são gerados a partir de uma determinada seleção ou filtro). Essa organização facetada permite uma seleção flexível e granular dos *datasets*.

4.1 Arquitetura do Repositório HADatAc

O grafo gerado é armazenado em dois bancos de dados: uma base de dados baseada em triplas, o Blazegraph [7] e um banco de dados NoSQL, o Apache Solr [8]. Esta divisão foi concebida por razões técnicas de escalabilidade, para permitir a separação entre dados e metadados do estudo científico. Vale lembrar que o tamanho dos arquivos de dados é bem superior ao tamanho dos arquivos de metadados. Não entraremos em detalhes aqui, mas a separação dessas duas partes permite obter escalabilidade explorando o que há de mais eficiente na tecnologia de cada tipo de banco de dados.

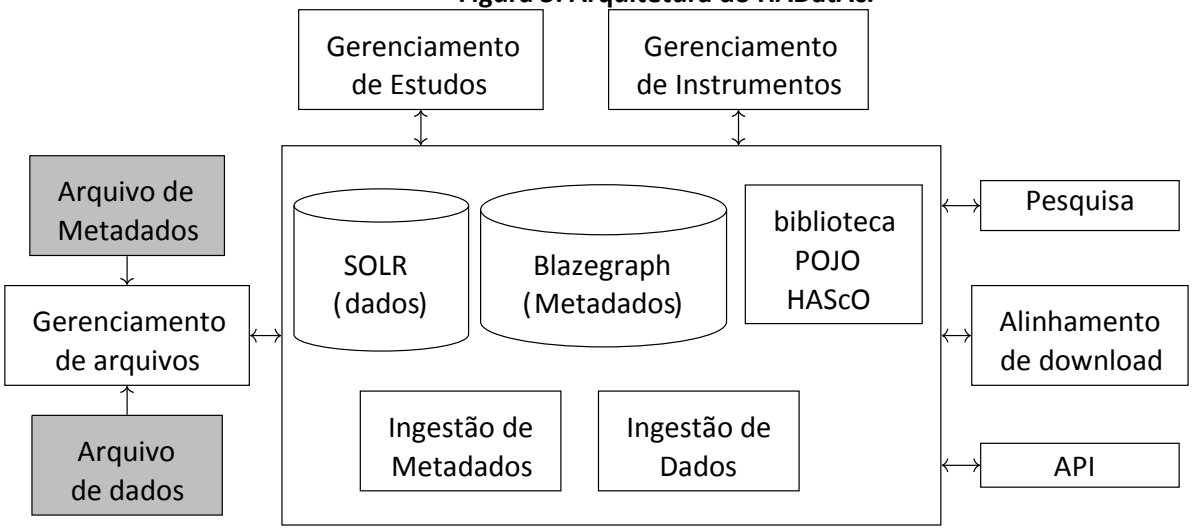
A arquitetura do HADatAc está esquematizada na Figura 3, O quadro principal, no centro da figura, representa o núcleo do sistema que é conectado a seis subsistemas satélites. O subsistema API é composto por uma coleção de serviços RESTful com acesso ao conteúdo do HADatAc. O componente principal tem os elementos necessários para suportar os subsistemas satélites, uma API Java codificando os conceitos da ontologia HAScO como Classes POJO e os

[7] <https://www.blazegraph.com/>

[8] <https://lucene.apache.org/solr/>

subsistemas responsáveis por extrair, anotar e armazenar conteúdo do estudo em arquivos de dados e metadados no SOLR e no Blazegraph. As classes HAScO POJO são utilizadas para construir e manter o grafo de conhecimento do HADatAc. O conteúdo é adicionado ao HADatAc através da análise de arquivos carregados através do Subsistema de Gerenciamento de Arquivos ou *on-line* através da interação do usuário com o Subsistema de Gerenciamento de Estudos e o Subsistema de Gerenciamento de Instrumentos. O conteúdo é apresentado diretamente aos usuários por meio do subsistema de pesquisa e baixado por meio do Subsistema de Alinhamento de Objeto e do Subsistema da API.

Figura 3: Arquitetura do HADatAc.



Fonte: Adaptação de (PINHEIRO et al., 2018b).

O método de ingestão de dados deve envolver a participação dos pesquisadores interessados no processo, que pode ser facilitada se a construção do grafo for realizada de forma que estes possam ter resultados mais rapidamente. Assim, propõe-se que sejam realizadas iterações curtas, com o envolvimento precoce dos interessados, em cada estágio do processo. O detalhamento deste método é o assunto da próxima seção.

5 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 originalmente por esta pesquisa prevê que a construção dos arquivos de metadados (SSD e SDD) utilize uma metodologia ágil. A complexidade e a morosidade no desenvolvimento de ontologias justificam a solução proposta, pela qual os resultados podem ser apresentados mais prontamente (BOURY-BRISSET, 2003).

No método proposto por nossa investigação e descrito a partir de agora, o pesquisador cria a primeira versão dos dados anotados, utilizando um modelo conceitual temporário, simples e idiossincrático (a ontologia "base") para descrever os objetos pesquisados, seus atributos e relações. Este modelo apenas organiza as colunas do *dataset* alvo da ingestão, sem mapeá-las para termos de ontologias pré-existentes. Logo, não é necessário investir esforço adicional concebendo a ontologia de domínio final 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). Em futuras iterações os metadados são revisados incorporando mudanças no modelo inicial.

As alterações nos metadados e na ontologia de domínio implicam na recriação do grafo a cada ciclo ou versão. Como o grafo é recriado, novos atributos podem ser adicionados com facilidade evitando-se possíveis inconsistências que ocorreriam se o grafo anterior não fosse removido. É importante salientar que, na integração de um estudo à base de dados existente, o procedimento de recriação do grafo se restringe ao subgrafo específico deste estudo. Isso é possível porque os grafos gerados são armazenados como *named graphs* (CARROLL et al., 2005), podendo ser manipulados de forma isolada.

O fluxo de trabalho proposto em nossa pesquisa permite iterações organizadas por ciclos com a participação do pesquisador e do ontologista durante todo o processo. Assim, à medida em que avança a sua compreensão sobre os fenômenos/objetos investigados, o pesquisador vai ajustando os SSDs e SDDs para incorporarem a versão em evolução da ontologia de domínio (derivada da ontologia "Base").

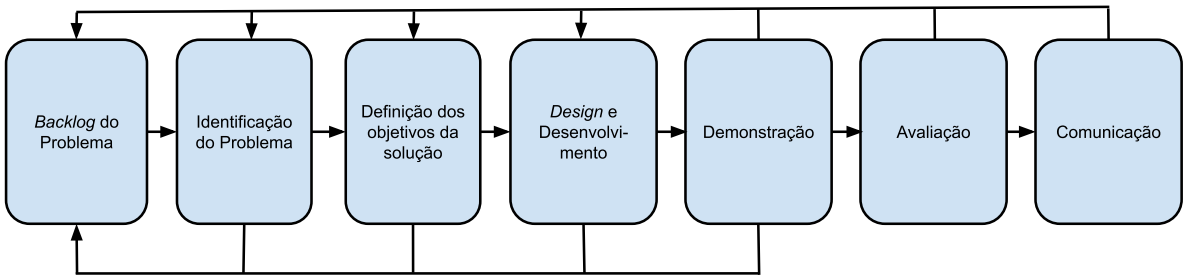
Esse fluxo incremental e assistido por um ontologista se justifica porque, mesmo que o pesquisador possua uma compreensão dos objetos do domínio pesquisado, ele não necessariamente domina eventuais formalizações de vocabulários pré-existentes estabelecidos por ontologias do domínio. Portanto, propõe-se 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 expectativa é de que se obtenha um maior reuso a cada ciclo.

5.1 Uso da ADSR como metodologia

O método de integração incremental se baseia na versão adaptada do *framework DSRM - Design Science Research Model* (PEFFERS et al., 2007), incorporando princípios da metodologia ágil conforme descrito em (CONBOY; GLEASURE; CULLINA, 2015). O modelo estendido, o ADSRM, é esboçado no diagrama apresentado na Figura 4.

Os artefatos são criados a partir de iterações, em que a cada ciclo um conjunto de tarefas é realizado. Uma iteração ou *Sprint* consiste em uma caixa de tempo fixa repetível durante a qual um artefato de maior valor possível é produzido.

Figura 4: *Agile Design Science Research Model* ADSRM



Fonte: Adaptação de (CONBOY; GLEASURE; CULLINA, 2015)

No método que propomos, uma *Sprint* está prevista para durar duas semanas. As sete etapas da Figura 4, que compõem uma *Sprint* ADSRM, adaptadas à nossa proposta de integração semântica incremental são:

a) ***Backlog* do Problema**: contém todas as tarefas que foram identificadas, no início do problema ou durante a execução das demais etapas. Como mostra a Figura 4, muitas etapas do processo podem gerar uma tarefa para o *Backlog* do Problema.

b) **Identificação do Problema**: na primeira *Sprint* do método de integração semântica proposto, esta etapa consiste na definição do dicionário de dados e das relações entre as colunas de dados, que irão fornecer subsídios para a criação dos *templates* para a ingestão de dados. Nas demais *Sprints,* o problema será tratado a partir das tarefas do *Backlog*, as quais serão produzidas nas etapas da *Sprint* anterior. A etapa de identificação do problema pode gerar tarefas para o *Backlog*.

c) **Definição dos objetivos da solução**: os objetivos definem as características da solução que abordará o problema definido. O método não se concentra apenas em objetivos específicos, mas também quebra os requisitos em histórias de usuário mais detalhadas. Para a integração incremental pode-se determinar, nesse momento, desde as características do grafo desejado até opções particulares da própria integração. Os objetivos definidos realimentam o *Backlog.*

d) ***Design* e desenvolvimento**: a partir do que foi definido nas etapas anteriores, deve-se projetar e implementar os artefatos que atenderão à solução (*templates* para a ingestão de dados ou mesmo alterações no *framewok*) e realizar a ingestão de dados. Mais uma vez, durante esta etapa, podem ser geradas tarefas para o *Backlog*.

e) **Demonstração**: com os *templates* criados e os dados ingeridos, pode-se apresentar aos envolvidos o resultado da *Sprint*. A implementação precoce e frequente da ingestão de dados permite identificar rapidamente possíveis problemas, os quais podem gerar novas tarefas para o *Backlog*. Nesta etapa é possível testar a estabilidade da solução frente ao problema. Um ponto importante da demonstração precoce da solução é que o pesquisador pode não conhecer ainda os recursos do HADatAc e poderá visualizar os dados ingeridos na ferramenta mais cedo, podendo sugerir futuras alterações.

f) **Avaliação**: o objetivo desta etapa é verificar se a solução resolve o problema definido. Isso corresponde aproximadamente ao componente de testes teóricos da pesquisa descritiva ou explanatória tradicional (CRESWELL, 2013), embora a ênfase esteja na utilidade do design e não na sua veracidade. Esta etapa pode fornecer subsídios para as seguintes etapas da *Sprint*: identificação do problema, definição do objetivo e design, além de poder gerar tarefas para o *Backlog*. No método de integração semântica proposto, isto equivale não somente a avaliar a ingestão de dados, mas também a integração destes com outros estudos já existentes no repositório do *framework*.

g) **Comunicação**: Representa o estágio final de um projeto, no qual as descobertas são compartilhadas com públicos relevantes por meio de publicação, tanto acadêmica quanto profissional. O *feedback* do público durante a comunicação possibilita o detalhamento dos objetivos da solução, do *design* e do desenvolvimento para tratamento na próxima *Sprint*. Componente central da a prática ágil (HUMMEL; ROSENKRANZ; HOLTEN, 2013), a comunicação é um processo contínuo e paralelo, enfatizando a diversidade de métodos e interações de comunicação.

O desafio de introduzir agilidade sem comprometer o rigor processual foi explorado e abordado na prática, em indústrias regulamentadas com altas demandas de transparência, documentação e responsabilidade durante o desenvolvimento (FITZGERALD et al., 2013). No ADSRM, uma *Hardening Sprint* é adicionada após um certo número de *Sprints*, com o objetivo de aumentar o rigor que pode estar faltando durante *Sprints* regulares. A frequência destas *Sprints* é baseada no contexto e dependente do rigor ou da falta percebida. Uma *Hardening Sprint* é diferenciada pelos seguintes mecanismos-chave:

a) Congelar o problema. Embora a capacidade de lidar com a mudança proporcionada pela agilidade esteja presente nos ambientes de *design* complexos atuais, um nível de rigor adicional pode ser aplicado ao utilizar-se uma fase em que não é permitida turbulência, dinamismo ou improvisação.

b) Congelar o processo. A preocupação de um ponto de vista rigoroso e regulador é que, para ter certeza do rigor, há momentos em que o processo deve ser valorizado sobre as pessoas. Novamente, uma única *Sprint* que requer uma adesão cuidadosa ao procedimento, verificações de conformidade e ausência de improvisação é valiosa para assegurar e manter o rigor.

c) Adicionar artefatos ao processo. A terceira forma de aumentar o rigor nesta *Sprint* extra é adicionar artefatos ao processo. Um exemplo seria adicionar alguma validação extra ou aprimorar as validações já realizadas.

Do exposto, pode-se observar que as características da ADSRM permitem maior agilidade mantendo o rigor do processo de ingestão de dados. A seguir, a Seção 6, apresenta uma avaliação preliminar do método proposto, de acordo com características importantes relacionadas aos dados ingeridos no HADatAc e com as facilidades de integração de dados que ele proporciona. Validações adicionais ainda são necessárias.

6 AVALIAÇÃO PRELIMINAR DO MÉTODO PROPOSTO

A *eScience* contemporânea requer que os dados gerados pelas pesquisas científicas e estudos sejam "Encontráveis" (*Findable*), Acessíveis, Interoperáveis e Reutilizáveis (FAIR). Os princípios FAIR (WILKINSON et al., 2016) permitem avaliar a qualidade dos conjuntos de dados publicados e dos métodos utilizados para sua disponibilização. Eles orientam a implementação

de algumas das "melhores práticas" em gerenciamento de dados no ciclo de gestão de dados científicos. Seguimos os princípios do FAIR para os dados gerados em nossa abordagem de anotação e, abaixo, avaliamos preliminarmente essa abordagem em relação aos mesmos.

Uma constatação deste trabalho é que o processo de criação do grafo RDF proposto é encontrável e acessível, na medida em que faz uso de ferramentas de código aberto e existe em repositório disponível ao público[9] além de ser ali documentado[10]. Mapear conjuntos de dados diversos para uma única estrutura conceitual atende à meta de interoperabilidade da Web Semântica, pois cada *datapoint*[11] presente no *dataset* é enriquecido por metadados que usam vocabulários e ontologias. Usando ontologias para anotar termos, *datasets* heterogêneos semanticamente podem ser alinhados, já que as conceituações podem ser comparadas a qualquer outro conjunto de dados que também tenha sido mapeado com termos ontologicamente equivalentes. Isso permite que conceitos que não tenham a mesma modelagem conceitual possam ser consultados de forma semelhante, permitindo o agrupamento de objetos similares a partir de conjuntos de dados distintos. Clusterizações adicionais podem ainda ser agregadas, alavancando a semântica dos conceitos mapeados.

A validação do método por experimentos empíricos está sendo realizada com a sua aplicação a alguns *datasets*, confirmando a hipótese de que o mesmo facilita o processo de integração semântica. Os resultados desses experimentos serão objeto de futuras publicações.

7 CONCLUSÃO

Apesar da diversidade de tipos de estudos científicos, nossa hipótese é que a consideração daqueles aspectos comuns, presentes na maioria deles, poderá tornar o método de anotação proposto mais adequado, para integrar essa natureza específica de dado, do que os trabalhos correlatos citados na Seção 2.3. Além disso, como visto na Seção 3.1, os metadados no contexto do *design* do estudo (SSD) e aqueles no contexto do Dicionário Semântico de

[9] https://tetherless-world.github.io/sdd/

[10] https:/tetherless-world.github.io/sdd/documentation

[11] Um datapoint é uma unidade discreta de informação. Em um sentido geral, qualquer fato único registrado é um datapoint. Em um contexto estatístico ou analítico, um datapoint é geralmente derivado de uma medição ou observação e pode ser representado numericamente.

Dados (SDD) devem advir de ontologias consolidadas. Porém, durante esta pesquisa constatou-se que a busca pela reutilização de termos de ontologias já estabelecidas é demorada e complexa, devendo sempre que possível ser realizada com o auxílio de um ontologista. Nesse sentido, a incrementalidade do enriquecimento semântico, peculiar ao nosso método, é vantajosa, pois os dados podem ser anotados, interpretados e comunicados mais prontamente. A principal contribuição do nosso método é o uso da ADSRM, que formaliza as etapas de cada iteração. Além disso, ao adicionar uma *Hardening Sprint,* o método permite realizar de forma incremental e ágil a integração semântica dos dados, sem perder de vista o rigor necessário à pesquisa científica.

Finalmente, além dos princípios FAIR discutidos na Seção 6, cabe ressaltar outros três requisitos importantes que nortearam até o momento a pesquisa do método de anotação descrito neste artigo: (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 também permite definir formas de melhor avaliá-lo no futuro.

REFERÊNCIAS

AUER, S. et al. Triplify: light-weight linked data publication from relational databases. In: **ACM**. Proceedings of the 18th international conference on World wide web. [S.l.], 2009. p. 621–630.

BERNERS-LEE T; HENDLER J; LASSILA O. The semantic web. **Scientific American**, [S.l.], v. 17, n. 5, p.28-37, maio 2001.

BOHLE, S. What is E-science and How Should it be Managed?. **Scilogs**: Nature: Spektrum der Wissenschaft, [S.l.], 9 maio 2013. Disponível em: http://www. scilogs. com/scientific_and_medicallibraries/what-is-e-science-and-how-should-it-be-managed

BOURY-BRISSET, A. C. Ontology-based approach for information fusion. **Proceedings of the sixth international conference on information fusion**, [S.l.: s.n.], v. 1, p. 522–529, 2003.

BRODARIC, B.; GAHEGAN, M. Ontology use for semantic e-science. **Semantic Web**, [S.l.], v. 1, n. 1, 2, p. 149-153, 2010.

CARROLL, J. J. *et al.* **Web Semantics**: Science, Services and Agents on the World Wide Web, Elsevier, v. 3, n. 4, p. 247–267, 2005.

CONBOY, K.; GLEASURE, R.; CULLINA, E. Agile design science research. **International Conference on Design Science Research in Information Systems,** [S.l.], p. 168–180, 2015.

CRESWELL, J. W. **Research design**: Qualitative, quantitative, and mixed methods approaches. [S.l.]: Sage publications, 2013.

DUMONTIER, M. et al. The semanticscience integrated ontology (sio) for biomedical research and knowledge discovery. **Journal of biomedical semantics**, [S.l.], v. 5, n. 1, p. 14, 2014.

ERMILOV, I.; AUER, S.; STADLER, C. Csv2rdf: User-driven csv to rdf mass conversion framework. **Proceedings of the ISEM.** [S.l.: s.n.], v. 13, p. 04–06, 2013.

FITZGERALD, B. *et al.* Scaling agile methods to regulated environments: An industry case study. *In:* IEEE PRESS. **Proceedings of the 2013 International Conference on Software Engineering**. [S.l.], 2013. p. 863–872.

FOX, P.; HENDLER, J. A. Semantic escience: encoding meaning in next-generation digitally enhanced science. **The Fourth Paradigm**, [S.l.], v. 2, 2009.

GUPTA, S. *et al*. Karma: A system for mapping structured sources into the semantic web. **Extended Semantic Web Conference**, [S.l.], p. 430–434, 2012.

HITZLER, P. *et al*. Ontology Design Patterns for Data Integration: The GeoLink Experience. Ontology Engineering with Ontology Design Patterns. **Foundations and Applications**, [S.l.], v. 25, p. 267, 2016. DOI: 10.3233/978-1-61499-676-7-267

HUMMEL, M.; ROSENKRANZ, C.; HOLTEN, R. The role of communication in agile systems development. **Business & Information Systems Engineering**, Springer, v. 5, n. 5, p. 343–355, 2013.

MCCUSKER, J.; RASHID, S.M.; AGU. N.; BENNETT, K.P.; MCGUINNESS, D.L. **Developing Scientific Knowledge Graphs Using Whyis**. Rensselaer Polytechnic Institute, Troy, p. 52-58, 2018.

PEFFERS, K. et al. A design science research methodology for information systems research. **Journal of management information systems, Taylor & Francis**, [S.l.], v. 24, n. 3, p. 45–77, 2007.

PÉREZ, J.; ARENAS, M.; GUTIERREZ, C. Semantics and complexity of sparql. **ACM Transactions on Database Systems** [S.l.], v. 34, n. 3, p. 16, 2009.

PINHEIRO, P. *et al.* Annotating diverse scientific data with hasco. **Proceedings of the Seminar on Ontology Research in Brazil**. [S.l.: s.n.], 2018.

PINHEIRO, P. et al. Hadatac: A framework for scientific data integration using ontologies. **Proceedings of the ISWC**, [S.l.: s.n.], 2018.

RASHID, S. M. et al. **The semantic data dictionary approach to data annotation & integration**. SemSci@ ISWC. [S.l.: s.n.], 2017. p. 47–54.

SHADBOLT, N.; BERNERS-LEE, T.; HALL, W. The semantic web revisited. **IEEE intelligent systems,** [S.l.], v. 21, n. 3, p. 96–101, 2006.

W3C et al. SPARQL 1.1 Overview. [S.l.]: World Wide Web Consortium, 2013. <https://www.w3.org/TR/sparql11-overview/>.

WAAL, S. van der *et al.* Lifting open data portals to the data web. **Linked Open Data–Creating Knowledge Out of Interlinked Data**, Springer, 2014. p. 175–195.

WACHE, H.; VÖGELE, T.; VISSER, U.; STUCKENSCHMIDT, H.; SCHUSTER, G.; NEUMANN, H.; HÜBNER, S. Ontology-based Integration of Information - A Survey of Existing Approaches, **In: Proceedings of IJCAI-01 Workshop**: Ontologies and Information Sharing, Seattle, WA, 2001, p. 108-117.

WILKINSON, M. D. *et al*. The fair guiding principles for scientific data management and stewardship. **Scientific data**, Nature Publishing Group, v. 3, 2016.