LINKED DATA WORKFLOW PROJECT ONTOLOGY: UMA ONTOLOGIA DE DOMÍNIO PARA PUBLICAÇÃO E PRESERVAÇÃO DE DADOS CONECTADOS

LINKED DATA WORKFLOW PROJECT ONTOLOGY: A DOMAIN ONTOLOGY FOR LINKED DATA PUBLISHING AND DIGITAL PRESERVATION

Sandro Rautenberg
srautenberg@unicentro.br
Universidade Estadual do Centro-Oeste

Edgard Marx
marx@informatik.uni-leipzig.de
Universität Leipzig

Ivan Ermilov
ivan.s.ermilov@gmail.com
Universität Leipzig

Sören Auer
auer@cs.uni-bonn.de
Universität Bonn

Resumo: No domínio da *Web* Semântica, a criação, a produção e a manutenção de bases de dados conectados não são atividades triviais. Esforços e recursos consideráveis são consumidos durante a execução de *workflows* para tais fins. É desejável que estes esforços sejam planejados, baseando-se em processos bem estabelecidos e conduzidos de forma sistemática ao longo do tempo, a fim de garantir a preservação de dados na *Web* de Dados. Neste contexto, este artigo apresenta a *Linked Data Workflow Project Ontology*, uma ontologia para modelar os aspectos de planejamento e de execução para a produção e a manutenção de bases de dados conectados na *Web* de Dados. Caracterizando-se como uma pesquisa aplicada, metodologicamente, o desenvolvimento dessa ontologia fundamentou-se em um conjunto de melhores práticas da Engenharia de Ontologias. Como resultado imediato, salienta-se que a *Linked Data Workflow Project Ontology* já é utilizada na preservação do histórico do Índice Qualis na *Web* de Dados como dados abertos conectados. Diante disso, conclui-se que a ontologia desenvolvida permite: a) a descrição de *workflows* como planos para produção de bases de dados conectados; b) o apoio à automatização da execução desses *workflows* em um ambiente computacional controlado; c) o reuso de planos, permitindo a reprodutibilidade e a repetibilidade de resultados e de bases de dados conectados; e d) a documentação dos planos e de suas execuções, promovendo a proveniência da produção bases de dados conectados.

Palavras-chave: Ontologia. Web Semântica. Processos de Gestão. Preservação Digital.

Abstract: The creation, production, and reproduction of linked data datasets are not trivial activities. Substantial efforts and resources are consumed during a workflow execution. It is suitable that these efforts should be planned, based on a development process, and performed in a systematic way. This

paper presents the Linked Data Workflow Project Ontology, a lightweight ontology for modeling the method, plan, and execution aspects for linked data dataset production and maintenance. Being characterized as applied research, methodologically, the development of this ontology was conducted based on a set of best practices Ontology Engineering. As an immediate result, the Linked Data Workflow Project Ontology is already used for preserving the historical data of the Qualis Index in Web of Data as linked open data. Facing these efforts, it is concluded that the developed ontology allows: a) the description of workflows as plans for producing linked data datasets; b) the support to the automatization of those workflows; c) the reuse of plans over time, as consequence, facilitating the reproducibility e repeatability of linked data datasets; and d) the documentation of plans and executions of the workflows, promoting the provenance on producing linked data datasets.

Keywords: Ontology. Semantic Web. Management Process. Digital Preservation.

1 INTRODUÇÃO

Historicamente, os anos noventa foram marcados por uma revolução digital, sendo a Internet a principal plataforma tecnológica dessa mudança. Diante essa revolução, Studer *et. al.* (2004) descreveram uma nova perspectiva para gestão de dados, informação e conhecimento. Esta perspectiva é alinhada a duas premissas: a) o uso de ontologias como paradigma para representação de elementos de conhecimento; e b) o emprego da *web* como plataforma de compartilhamento desses elementos.

No mesmo período, Tim Bernes-Lee cunhou o conceito da *Web* Semântica. Em sua essência, a *Web* Semântica estava promovendo uma nova forma de utilização da *web*, antes baseada em documentos e seus *links*, para uma *web* de compartilhamento de dados e seus relacionamentos (BERNES-LEE; HENDLER e LASSILA, 2001). Com isso, foram estabelecidos os dados conectados (*linked data*) e suas boas práticas para estruturação, publicação e conexão de dados em escala global, formando a *Web* de Dados.

Alguns anos depois, baseando-se em princípios dos Processos Ágeis da Engenharia de Software e na Engenharia do Conhecimento, Auer (2007) propôs a *Agile Knowledge Engineering*. Um constructo voltado a produzir procedimentos metodológicos e tecnologias para desenvolvimento de aplicações da *Web* Semântica, em especial, aplicações com dados conectados. Como perspectivas, foram traçadas: a) as ontologias e os vocabulários são utilizados para representar os dados na *Web* de Dados; b) o *Linked Data Lifecycle* (AUER, 2014) é o ciclo de vida de dados utilizado para o desenvolvimento e compartilhamento de bases de dados conectados na *web*; e c) o *Linked Data Stack* (van NUFFELEN et. al., 2014) é um conjunto de ferramentas tecnológicas empregadas para o desenvolvimento de aplicações com dados conectados na *Web* de Dados. Nota-

damente, os procedimentos metodológicos e o desenvolvimento de tecnologias para dados conectados estão em fase de amadurecimento e de adoção (AUER, 2014; van NUFFELEN, 2014). Neste sentido, neste artigo objetiva-se avançar na discussão de um dos problemas comumente enfrentados na *Web* Semântica, em especial, pela comunidade ligada à pesquisa com dados conectados: a proveniência e, consequente, reprodutibilidade e repetibilidade de bases de dados conectados na *Web* de Dados.

Em particular, foca-se na representação do conhecimento que circunscreve o planejamento e a execução de *workflows* (fluxos de trabalho) para manutenção de bases de dados conectados. Tal representação é modelada na forma de uma ontologia. Sendo assim, é apresentada a *Linked Data Workflow Project ontology* (LDWPO). Resumidamente, a LDWPO é uma ontologia que descreve e organiza os conceitos *Plan* e *Execution* para a operacionalização da produção e da preservação de bases de dados conectados na *Web* de Dados. A ontologia é estendida a partir de outras ontologias e vocabulários encontrados na literatura: *Publishing Workflow ontology* - PWO (GANGEMI et. al., 2014), *the Open Provenance Model Vocabulary* - OPMV (MOREAU et. al., 2011) e PROV *ontology* - PROV-O (LEBO et. al., 2015). Diante essa extensão, a LDWPO suporta: a) a descrição de *workflows* como planos para produção de bases de dados conectados; b) a automatização da execução desses *workflows* para produção bases de dados conectados em um ambiente computacional controlado; c) o reuso de planos, permitindo a reprodutibilidade e a repetibilidade das bases de dados conectados ao longo do tempo; e d) a documentação dos planos e das execuções, promovendo a proveniência e preservação de bases de dados conectados na *Web* de Dados. Ressalta-se que a LDWPO já é utilizada no domínio da Cientometria em um importante estudo de caso denominado QualisBrasil. Tal estudo tem por objetivo preservar o histórico do Índice Qualis na *Web* de Dados de forma automatizada (RAUTENBERG e BURDA, 2016), como também será descrito neste artigo.

Para discutir o potencial da LDWPO, além desta seção introdutória, este artigo compreende: i) os trabalhos correlatos (demais ontologias e vocabulários) encontrados na literatura e que são utilizados como fontes de conhecimento no desenvolvimento da LDWPO; ii) o procedimento metodológico adotado para construir a ontologia proposta; iii) uma breve apresentação dos principais componentes da LDWPO; iv) o QualisBrasil como o estudo de caso principal utilizado para verificar a ontologia; e v) as considerações finais.

2 TRABALHOS CORRELATOS

No contexto do desenvolvimento e evolução da LDWPO, considera-se como a motivação principal o atendimento aos requisitos de planejamento, proveniência, reprodutibilidade, repetibilidade e documentação da produção de bases de dados conectados na *Web* de Dados. Nesta seção, são apresentadas as ontologias e vocabulários aderentes a essa motivação. Também é apresentada uma discussão das limitações encontradas nestes trabalhos frente às motivações traçadas.

2.1 ONTOLOGIAS E VOCABULÁRIOS PARA *WORKFLOW*

De acordo com os preceitos da Engenharia de Ontologias (GÓMEZ-PÉREZ; CORCHO e FERNÁNDEZ-LÓPEZ, 2004), no desenvolvimento de uma ontologia é relevante buscar por outras ontologias e outros vocabulários caracterizados como correlatos para promover a reutilização de conceitos. Neste sentido, de acordo os requisitos da LDWPO, são citados: a) PWO; b) PROV-O; e c) OPMV. Além destes, são citadas as ontologias integradas aos ambientes de *Workflow* Científico Taverna (TAVERNA, 2015) e Keppler (KEPLER, 2015).

2.1.1 PWO (GANGEMI et. al., 2014)

A PWO é uma ontologia para descrever *workflows* associados à publicação de um documento na *web*. Utilizando os componentes principais desta ontologia, é possível: a) definir o passo inicial para um dado *workflow*; b) relacionar os passos antecessor e sucessor de um determinado passo, estruturando um *workflow*; e c) definir as entradas e saídas para cada passo. Perante este trabalho, esta ontologia inspira o núcleo de operação da LDWPO.

2.1.2 OPMV (MOREAU *et. al.*, 2011)

O OPMV é um vocabulário recomendado para modelar a proveniência de dados, permitindo a publicação e o compartilhamento de dados entre sistemas. No OPMV: a) um processo é controlado por um agente; b) um processo utiliza artefatos em determinado tempo/ação; c) um artefato é gerado por um processo; d) um artefato pode ser derivado de outro artefato; e e) para executar um *workflow*, um processo dispara o processo subsequente. Assim, como característica principal, a OPMV descreve como se dá a gestão de artefatos de dados ao longo do tempo. Entretanto, o OPMV não define o conceito de *workflow*, explicitamente.

2.1.3 PROV-O (LEBO et. al., 2015)

A ontologia PROV-O é uma recomendação do *World Wide Web Consortium* (W3C) para representar e compartilhar informações sobre proveniência e reprodutibilidade geradas por diferentes sistemas e contextos. Com os conceitos principais, na PROV-O: a) uma atividade é associada a um agente; b) uma entidade é atribuída a um agente; c) uma atividade utiliza uma ou mais entidades por determinado tempo; d) uma entidade pode ser derivada de outra entidade; e e) para manter um *workflow*, uma atividade é sucedida por outra atividade. Assim como a OPMV, na PROV-O, o conceito de *workflow* não é explicitamente descrito.

2.1.4 *Scientific Workflow ontologies*: ontologias Scufl2 e de Keppler

Em outro domínio, a comunidade científica cunhou o conceito *Workflow* Científico como "o processo automatizado que combina dados e processos em um conjunto estruturado de passos para implementar soluções computacionais à problemas científicos" [tradução dos autores] (ALTINTAS; BARNEY e JAEGER-FRANK, 2006). Para facilitar a gestão de *workflows*, foram desenvolvidos os Sistemas de Gerenciamento de *Workflows* Científicos, tais como o Kepler (LUDÄSCHER et. al., 2006) e o Apache Taverna (HULL et. al., 2006). Estes sistemas de gerenciamento empregam ontologias para modelar os *workflows*. Estas ontologias são denominadas Kepler e Scufl2 *ontologies*, respectivamente.

Listagem 1: EXEMPLOS DE CLASSES EXTRAÍDAS DAS KEPLER *ONTOLOGIES*

```
01  [...]
02   <owl:Class rdf:ID ="Workflow">
03    <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
04      Workflow
05    </rdfs:label>
06   </owl:Class>
07   [...]
08   <owl:Class rdf:ID="WorkflowOutput">
09     <rdfs:label rdf:datatype="http://www.w3.org/2001/XMLSchema#string">
10       Workflow Output
11     </rdfs:label>
12     <rdfs:subClassOf>
13       <owl:Class rdf:about="#DataOutput"/>
14     </rdfs:subClassOf >
15   </ owl : Class >
16   [...]
```

Fonte: (KEPLER, 2015).

As Kepler *ontologies* são parte integrante do Kepler *framework* e podem ser encontradas

na instalação computacional da referida ferramenta. Estas ontologias não incluem a descrição de seus conceitos em linguagem natural (*rdfs:comment*), conforme representado na Listagem 1. A descrição dos conceitos em linguagem natural é uma boa prática da Engenharia de Ontologias, a qual facilita o reuso de recursos ontologicamente estabelecidos. A ausência de tais descrições limitou a adoção das Kepler *ontologies* no desenvolvimento da LDWPO.

Listagem 2: EXEMPLO DE CLASSES EXTRAÍDAS DA SCUFLE2 *ONTOLOGY*

```
01  […]
02    <owl:Class rdf:about="http://ns.taverna.org.uk/2010/scufl2#Workflow">
03      <rdfs:label xml:lang="en">Workflow</rdfs:label>
04      <rdfs:subClassOf rdf:resource="http://ns.taverna.org.uk/2010/scufl2#Named"/>
06      <rdfs:subClassOf>
07        <owl:Restriction>
08          <owl:onProperty rdf:resource="http://ns.taverna.org.uk/2010/scufl2#name"/>
09          <owl:someValuesFrom rdf:resource="&xsd;string"/>
10        </owl:Restriction>
11      </rdfs:subClassOf>
12      <rdfs:subClassOf>
13        <owl:Restriction>
14          <owl:onProperty rdf:resource="http://ns.taverna.org.uk/2010/scufl2#workflowIdentifier"/>
15          <owl:someValuesFrom rdf:resource="&owl;Thing"/>
16        </owl:Restriction>
17      </rdfs:subClassOf>
18      <owl:hasKey rdf:parseType="Collection">
19        <rdf:Description rdf:about="http://ns.taverna.org.uk/2010/scufl2#workflowIdentifier"/>
20      </owl:hasKey>
21    </owl:Class>
22  [...]
```

Fonte: (SCUFLE2, 2016).

A falta de descrição de conceitos também é percebida na Scufl2 *ontology*, conforme exemplificado na Listagem 2. Para muitos dos elementos dessa ontologia não há uma descrição, o que dificulta o entendimento da estrutura desta ontologia, uma vez que a nomenclatura de seus componentes é pouco usual.

2.2 LIMITAÇÕES DOS TRABALHOS CORRELATOS

Ao averiguar algumas das ontologias e vocabulários disponíveis, como ponto comum, observa-se a capacidade de descrever a execução de *workflows*. Ou seja, a seu modo, a PWO, a PROV-O e o OPMV comportam as descrições de recursos, agentes e passos usados na execução de *workflows*. Porém, na perspectiva de gestão de recursos, visando à reutilização de um esque-

ma em várias execuções, tais artefatos não possibilitam a descrição de *workflow* ao nível do planejamento (somente ao nível de execução). Em outras palavras, nas ontologias e vocabulários investigados, a cada execução, um novo *workflow* deve ser instanciado e não reutilizado. Salienta-se que a descrição de esquemas (ou planos) é requisito essencial em ambientes de gerenciamento de *workflows*, facilitando a automatização, a reprodução e a preservação de resultados ao longo do tempo.

Considerando as limitações apontadas e o contexto da preservação de bases de dados conectados, este artigo estende o vocabulário OPMV e as ontologias PWO e PROV-O, propondo a LDWPO. Infelizmente, uma análise semelhante não é passível de ser realizada para as Kepler e Scufle2 *ontologies*, uma vez que descrições de seus elementos não estão disponíveis em sua integralidade. Diante disso, a seguir discute-se o procedimento metodológico adotado no desenvolvimento da LDWPO.

3 PROCEDIMENTO METODOLÓGICO

O procedimento metodológico do desenvolvimento da LDWPO se baseia nas práticas de alguns métodos para a construção de ontologias. Combina-se os artefatos metodológicos oriundos da *On-to-Knowledge* (SURE; STUDER, 2002), da METHONTOLOGY (GÓMEZ-PÉREZ; CORCHO e FERNÁNDEZ-LÓPEZ, 2004) e do guia *Ontology Development 101* (NOY; MCGUINNESS, 2008), conforme segue:

- ***On-to-Knowledge*** - contribui na especificação dos requisitos da ontologia, por meio do emprego de questões de competência como modo simples e direto para confirmar o escopo e o propósito de uma ontologia. Tal fato permite identificar previamente, conceitos, propriedades, relações e instâncias de uma ontologia.
- **METHONTOLOGY** - por meio de uma rica gama de artefatos, contribui na documentação e na verificação de ontologias. Esta metodologia serviu de guia para a elaboração do relatório técnico[1] da LDWPO.
- ***Ontology Development 101*** - contribui com uma visão clara de como se dá um processo iterativo para o desenvolvimento de ontologias.

[1] Relatório técnico da Linked Data Workflow Project Ontology disponível em: https://github.com/AKSW/ldwpo/blob/master/misc/technicalReport/LDWPO_technicalReport.pdf.

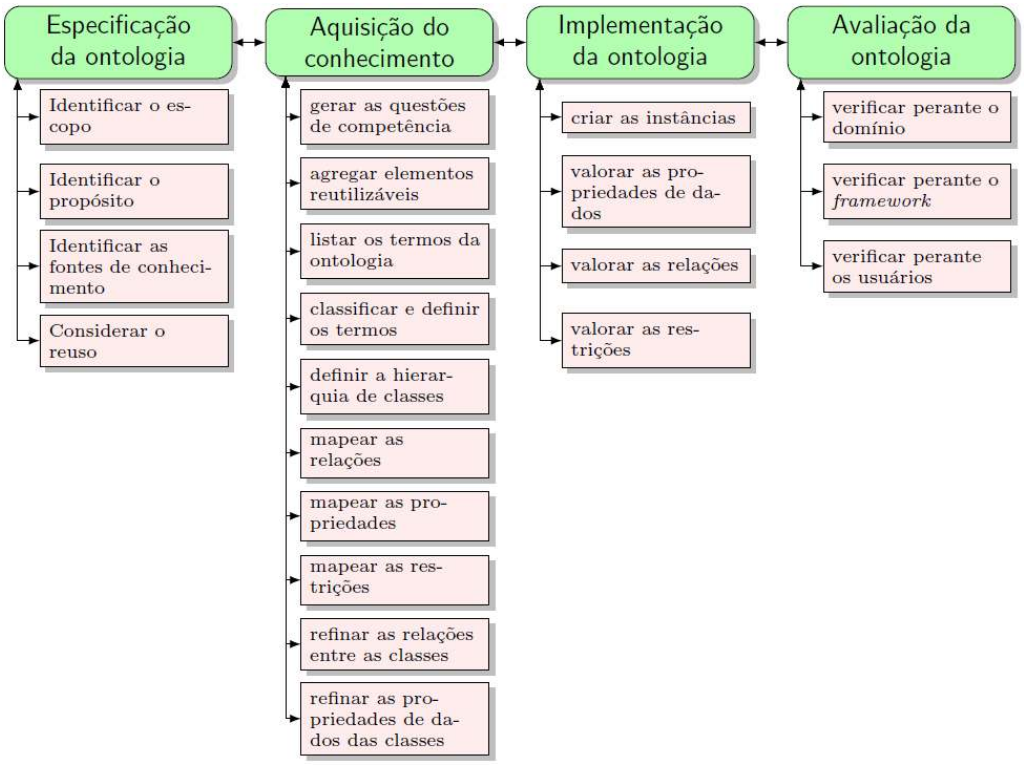

Figura 1: PROCEDIMENTO METODOLÓGICO ADOTADO
Fonte: elaborada pelos autores.

Representado na Figura 1, o procedimento metodológico baseia-se em quatro grandes atividades, com suas respectivas tarefas:

1. **Especificação da ontologia** - é uma atividade também presente no guia *Ontology Development 101*, *On-to-Knowledge* e METHONTOLOGY. Nesta atividade tende-se a discernir a respeito dos custos do desenvolvimento da ontologia. Pretende-se: a) identificar o escopo; b) identificar o propósito; c) identificar as fontes de conhecimento; e d) considerar o reuso com os elementos das fontes de conhecimento.
2. **Aquisição do conhecimento** - é uma atividade que também compreende as tarefas de conceitualização e de formalização da ontologia. Representa o ponto de maior interação do ontologista com os especialistas de domínio. Desta interação, se abstrai a maioria dos elementos de conhecimento da ontologia. Interativamente, consideram-se as tarefas de: a) gerar as questões de competência; b) agregar os elementos reutilizáveis; c) listar os termos da ontologia; d) classificar e definir em linguagem natural os termos da ontologia; e) definir a hierarquia de classes; f) mapear as relações de cada classe; g) mapear as propriedades de dados de cada classe; h) mapear as restrições de cada classe; i) refinar as relações entre as classes, atrelando algumas características (funcional, inversa funcional, reflexiva, irreflexiva, simétrica, assimétrica e transitiva); e j) refinar as propriedades de dados das classes, definindo qual o tipo de da-

dos comportado (*string*, número, data ou lógico) e se a propriedade tem a característica funcional.

3. **Implementação** - é uma atividade de menor interação com especialistas de domínio, sendo reservada às tarefas de: a) criar as instâncias de cada classe; b) valorar as propriedades de dados de cada instância; c) valorar as relações de cada instância, conectando uma instância para com outras instâncias da ontologia; e d) valorar as restrições das classes, definindo as restrições presentes no domínio quanto aos valores possíveis de suas propriedades de dados e de suas relações.
4. **Verificação** - trata-se de uma atividade que prevê maior interação com especialistas de domínio e com os usuários da ontologia para averiguar a ontologia, sendo as tarefas: a) verificar a ontologia perante as fontes de conhecimento; b) verificar a ontologia perante um *frame* de referência gerado a partir do escopo, do propósito e das questões de competência; e c) verificar a ontologia perante a visão do usuário, considerando a usabilidade e a utilidade da ontologia.

4 DESENVOLVENDO A LDWPO

O escopo da LDWPO está circunscrito aos projetos para preservação de bases de dados conectados na *Web* de Dados, descrevendo os conceitos sobre planos e execução dos planos para manutenção do referido tipo de base de dados. Neste sentido, a LDWPO tem como propósitos:

1. A descrição de *workflows* como planos para produção de bases de dados conectados. Neste sentido, um plano é caracterizado como um conjunto de passos. Cada passo corresponde ao uso de uma ferramenta sobre um conjunto de dados de entrada para produzir um conjunto de dados de saída. Disto, o encadeamento lógico de passos formula um fluxo de trabalho, ou *workflow*.
2. A automatização da execução dos *workflows* para produção bases de dados conectados. A automatização é alcançada num ambiente controlado, por executar as ferramentas sobre base de dados de entrada, produzindo as bases de dados de saída.
3. O reuso de planos. Isso permite a reprodutibilidade e a repetibilidade das bases de dados conectados ao longo do tempo.
4. A documentação dos planos e das execuções de *workflows*. Tal requisito vem ao encontro das atividades de manutenção e evolução dos dados, promovendo a proveniência e preservação de bases de dados conectados na *Web* de Dados.

Diante desses propósitos, como fontes de conhecimento, foram identificadas as ontologias aderentes ao conceito *workflow* e vocabulários úteis na estruturação de elementos de conheci-

mento. Outras fontes e vocabulários disponíveis na *web* também foram identificadas. Disto, são enumeradas como fontes de conhecimento: a) Dublin Core (DUBLIN CORE, 2015); b) *Description of a Project vocabulary* - DOAP (DOAP, 2015); c) *Friend-of-a-Friend* - FOAF (FOAF, 2015); d) OPMV; e) PROV-O; f) PWO; e g) o sítio do Instituto *Agile Knowledge and Semantic Web* - AKSW (2015).

De acordo com os procedimentos metodológicos, a próxima atividade é a aquisição do conhecimento. Assim, o objetivo primordial perseguido foi indagar os especialistas de domínio (engenheiros de dados) na perspectiva que estes formulem questões de competência pertinentes ao entendimento dos requisitos funcionais traçados. A Tabela 1 enumera as questões percebidas durante as entrevistas com os especialistas, organizando-as segundo o *framework* de Zachman (SOWA; ZACHMAN, 1992).

Tabela 1: QUESTÕES DE COMPETÊNCIA LEVANTADAS JUNTO AOS ESPECIALISTAS DE DOMÍNIO (ORIGINALMENTE EM LÍNGUA INGLESA)

Dimensões	**Questões de competência da ontologia**
Qual	01. Qual é o nome deste projeto? 02. Qual é o sítio deste projeto? 03. Quais são os nomes das pessoas que contribuem no projeto? 04. Qual é o nome da pessoa que criou este projeto? 05. Quais são os objetivos deste projeto? 06. Quais são as ferramentas empregadas nesta atividade? 07. Qual é a configuração de ferramenta neste passo? 08. Quais são as bases de dados de entrada do projeto? 09. Quais são as bases de dados deste projeto? 10. Quais são as atividades deste projeto? 11. Quais são as bases de dados de saída deste estágio? 12. Quais são as melhores práticas aplicadas neste projeto? 13. Qual é o processo aplicado neste projeto? 14. Quais são as atividades cobertas por este processo? 15. Quais são as tarefas nesta atividade? 16. Quais são as ações realizadas nesta base de dados neste passo? 17. Quais são os passos realizados nesta atividade? 18. Qual é o passo anterior a este passo? 19. Qual é o plano para este passo? 20. Quais ações são realizadas neste passo? 21. Qual é o próximo passo a este passo? 22. Quais são os passos realizados para esta atividade? 23. Qual é o formato desta base de dados? 24. Qual é a licença de uso desta base de dados? 25. Qual é o valor padrão deste parâmetro?
Onde	26. Onde esta base de dados de entrada está armazenada? 27. Onde esta base de dados de saída está armazenada? 28. Onde esta ferramenta é localizada? 29. Onde está localizado o arquivo de configuração da ferramenta?
Por que	*Não foram feitas questões de competência para esta dimensão*

Dimensões	Questões de competência da ontologia
Quando	30. Quando esta base de dados foi atualizada pela última vez? 31. Quando este passo foi executado? 32. Quando este *workflow* iniciou? 33. Quando este *workflow* finalizou?
Quem	34. Quem são as pessoas que contribuem neste projeto? 35. Quem é o responsável por este passo?
Como	36. Como esta base de dados de entrada é armazenada? 37. Como esta base de dados de saída é armazenada?

Fonte: elaborada pelos dos autores a partir das entrevistas com especialistas de domínio.

Ressalta-se que a evolução da LDWPO é mantida em contribuição com especialistas internacionais. Por isso, originalmente, os termos da ontologia e suas definições são encontrados em língua inglesa, no relatório técnico disponibilizado a partir do sítio do projeto da ontologia[2].

A partir das questões de competência e demais fontes de conhecimento, implementou-se a LDWPO, a qual é ilustrada na Figura 2. Ressalta-se que os elementos os elementos principais (*Project, Workflow, Step, WorkflowExectuion* e *StepExecution*) são especializados para o domínio *Linked Data Workflow*, utilizando-se o prefixo "LDW".

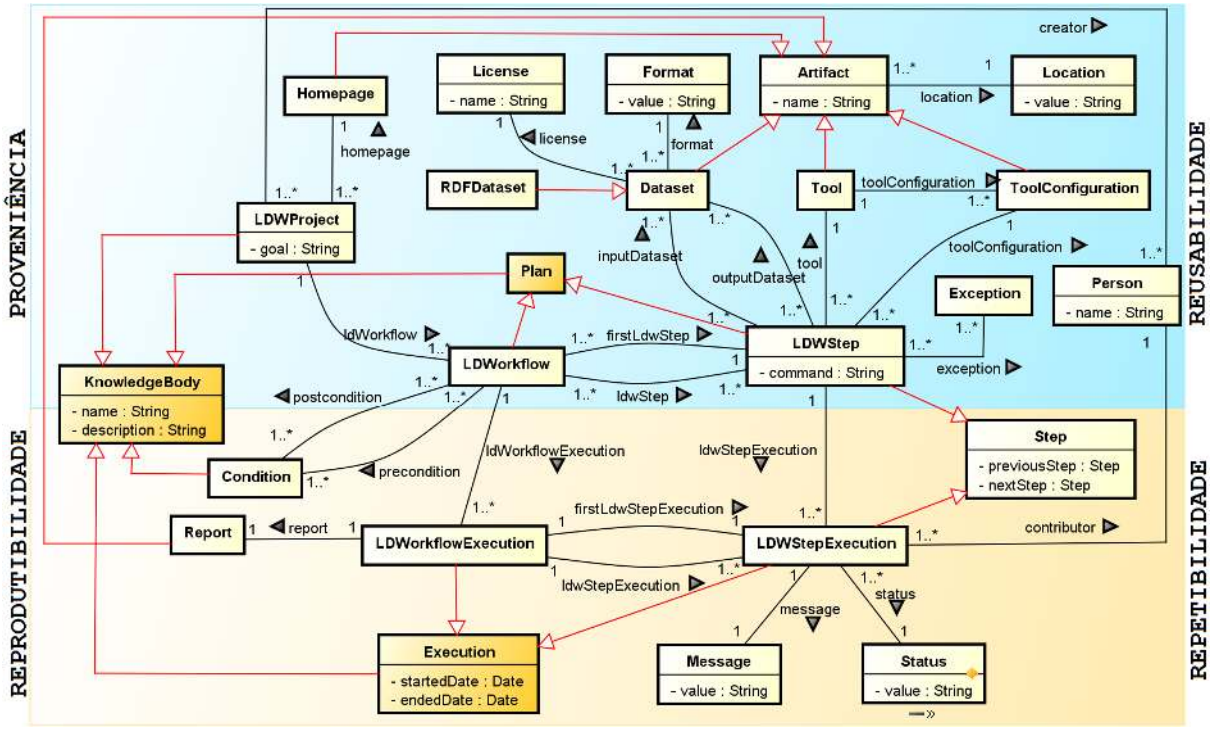

Figura 2: REPRESENTAÇÃO DA LDWPO
Fonte: elaborada pelos autores.

[2] Disponível em: http://aksw.org/Projects/LDWPO.html.

A seguir são apresentadas as principais classes da ontologia e suas correlações para com as fontes de conhecimento investigadas:

- ***LDWProject* (Projeto de base de dados conectados)** - classe que representa um empreendimento para criação e manutenção de uma base de dados conectados na *Web* de Dados. Tem correlação com a classe *foaf:Project*.
- ***Plan* (Plano)** - superclasse das classes *LDWorkflow* e *LDWStep*. Organiza semanticamente a dimensão de planejamento dos *workflows* de um *LDWProject*.
- ***LDWorkflow* (*Workflow* de base de dados conectados)** - subclasse de *Plan* utilizada para envolver os passos reservados para a produção/manutenção de uma base de dados conectados em um *LDWProject*. Em suma, esta classe organiza uma lista de *LDWSteps*. Tem correlação com *pwo:Workflow*.
- ***LDWStep* (Passo em um *workflow* de base de dados conectados)** - subclasse de *Plan* and *Step* que representa uma unidade atômica de um *LDWorkflow*. A uma instância de *LDWStep* são relacionados: uma base de dados de entrada (relação *inputDataset*); uma base de dados de saída (relação *outputDataset*); e a ferramenta (relação *tool*) que processa a base de dados de entrada, gerando a base de dados de saída, de acordo com um arquivo de configuração (relação *toolConfiguration*). Tem conotação às classes: *pwo:Step*, *opmv:Process* e *prov:Activity*.
- ***Execution* (Execução)** - superclasse das classes *LDWorkflowExecution* e *LDWStepExecution*. Organiza semanticamente a dimensão de execução dos *workflows* de um *LDWProject*. Com essa classe define um ponto particular no tempo (propriedades *startedDate* e *endedDate*) em que algum evento relacionado a um *LDWProject* ocorre. Permite fazer apontamentos de repetibilidade e reprodutibilidade de resultados.
- ***LDWorkflowExecution* (Execução de um *workflow* de base de dados conectados)** - subclasse de *Execution* relacionada a uma execução particular de um *LDWorkflow*, a qual produz fisicamente uma versão de base de dados conectados. Internamente, ela organiza uma lista de *LDWStepExecutions*.
- ***LDWStepExecution* (Execução de um passo em um *workflow* de base de dados conectados)** - subclasse de *Execution* e *Step* que representa uma unidade de processamento de um *LDWorkflowExecution*. Está intimamente relacionado a um *LDWStep* previamente planejado.

Diante essas definições, o ponto de partida na LDWPO é o conceito *LDWProject*, uma descrição para criar ou preservar uma base de dados conectados. Dentre suas propriedades, um

LDWProject é associado a um *LDWorkflow*. Resumidamente, um *LDWorkflow* organiza um plano necessário para produzir uma base de dados conectados, encapsulando um conjunto de *LDWSteps*. *LDWStep* é um conceito que representa uma unidade de processamento, descrevendo um procedimento a ser realizado sobre um conjunto de dados, utilizando uma ferramenta computacional, a fim de produzir um conjunto de dados de saída. Adicionalmente, um *LDWStep* pode ser automaticamente executado em um ambiente computacional controlado. Essa automatização é melhor detalhada na próxima seção, a qual descreve os esforços empreendidos na preservação do histórico do Índice Qualis na *Web* de Dados.

5 QUALISBRASIL - VERIFICANDO A LDWPO COM UM ESTUDO DE CASO

QualisBrasil é um *LDWProject* tem como objetivo preservar uma base de dados do histórico do índice Qualis de acordo com os princípios de Dados Abertos Conectados. Neste projeto, a base de dados é mantida anualmente, com dados provenientes de um sistema legado. Tal empreendimento visa suprimir algumas limitações enfrentadas por pesquisadores do campo da Ciência da Informação: a) os dados históricos do índice Qualis não estão disponíveis em outro sítio na *web*, tornando dificultoso o estudo de séries temporais; b) nos anos iniciais, os dados eram abertos, porém, sob formato fechado (em documento PDF ou XLS), consumindo recursos consideráveis no pré-processamento de dados; c) somente as últimas versões do índice estão disponíveis para *download* (em formato XLS); e d) o índice Qualis não é relacionado a outras bases de dados conectados, o que pode ser um desafio na *Web* de Dados.

Tabela 2: CAPTURA E PRÉ-PROCESSAMENTO DO ÍNDICE QUALIS

Ano de coleta	Período referência	Origem	Formato	Tuplas validadas
2007	2005-2007	WebQualis	XLS	35.020
2009	2008-2010	WebQualis	PDF	54.233
2013	2011-2013	WebQualis	PDF	107.429
2015	2014	Internet[3]	XLS	108.622
2016	2015	Sucupira	XLS	44.463

Fonte: elaborada pelos autores.

Ressalta-se que o índice Qualis foi coletado ao longo dos últimos dez anos, principalmente, a partir do Sistema WebQualis (WEBQUALIS, 2013) e a Plataforma Sucupira (SUCUPIRA, 2016),

[3]NIEVINSKI, F. G. [ciência aberta] Planilha Qualis (em anexo). [mensagem eletrônica]. Disponível em: https://lists.okfn.org/pipermail/cienciaaberta/2014-October/000559.html. Acesso em: 2 de Fevereiro de 2015.

sendo armazenado em um sistema legado pessoal. A Tabela 2 resume a coleta e o pré-processamento dos dados, associando: um período de referência para construção de histogramas, a fonte de dados, o formato e as tuplas validadas. Os referidos dados foram convertidos ao modelo RDF, conforme a Figura 3, utilizando os seguintes vocabulários e ontologias:

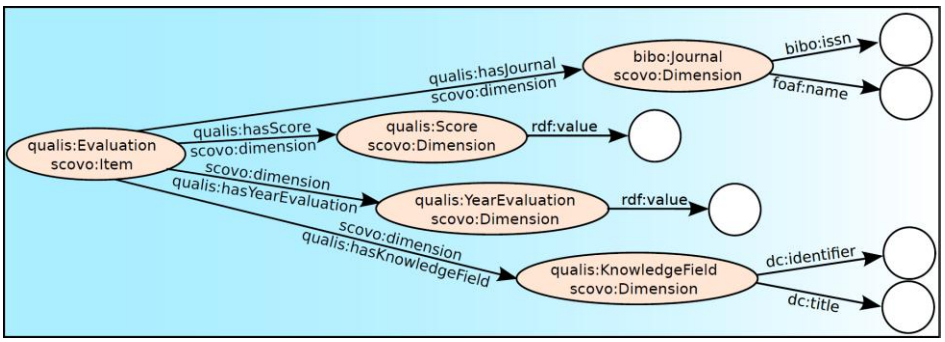

Figura 3: MODELO RDF PARA O ÍNDICE QUALIS NA *WEB* DE DADOS
Fonte: elaborada pelos autores.

1. ***The Statistical COre VOcabulary*[4] (scovo)** - é um vocabulário simples para representar dados estatísticos na *web*. Neste trabalho, é usado para organizar o Índice Qualis na forma multidimensional.
2. **Dublin Core[5] (dc)** – é um vocabulário amplamente utilizado para descrever recursos. É utilizado para melhor representar as áreas de conhecimento no grafo QualisBrasil (elementos *dc:identifier* e *dc:title*).
3. ***Bibliographic Ontology Specification*[6] (bibo)** – é uma ontologia que modela os conceitos e as propriedades para descrever referências bibliográficas. Seus elementos são usados para representar os *journals*.
4. ***Friend-of-a-Friend*[7] (foaf)** – é um vocabulário utilizado para relacionar entidades a informações na *web*, como por exemplo, o nome de um *journal*.

Para preservar e compartilhar os dados do Índice Qualis na *Web* de Dados, com a LDWPO, foi formalizado um *LDWProject* denominado QualisBrasil. De acordo a Figura 4, este *LDWProject* é baseado em um *LDWorkflow* composto por cinco *LDWSteps*. Salienta-se que tais *LDWSteps* mantém as informações sobre a proveniência *LDWProject*, podendo ser executados automaticamente da seguinte forma:

1. Recupera-se os dados do Índice Qualis do um sistema legado, salvando-os em um arquivo

[4] Disponível em: http://vocab.deri.ie/scovo
[5] Disponível em: http://dublincore.org/documents/dcmi-terms/
[6] Disponível em: https://github.com/structureddynamics/Bibliographic-Ontology-BIBO/blob/master/bibo.owl
[7] Disponível em: http://xmlns.com/foaf/spec/

com formato CSV;

2. Converte-se os dados do arquivo CSV em uma base de dados conectados, de acordo o modelo RDF da Figura 3, usando a ferramenta de conversão Sparqlify[8];
3. Armazena-se o conjunto de dados resultante em um grafo denominado http://lod.unicentro.br/QualisBrasil/ em um *endpoint* na *Web* de Dados disponível em http://lodkem.led.ufsc.br:8890/sparql. Tal *endpoint* é baseado no servidor de dados universal Open Link Virtuoso[9];
4. Utiliza-se a ferramenta LIMES[10] para interligar o Índice Qualis com os recursos do grafo DBpedia[11], os quais representam os periódicos científicos, considerando o ISSN dos periódicos; e
5. Armazena-se os *links* adquiridos no passo anterior ao grafo do Índice Qualis.

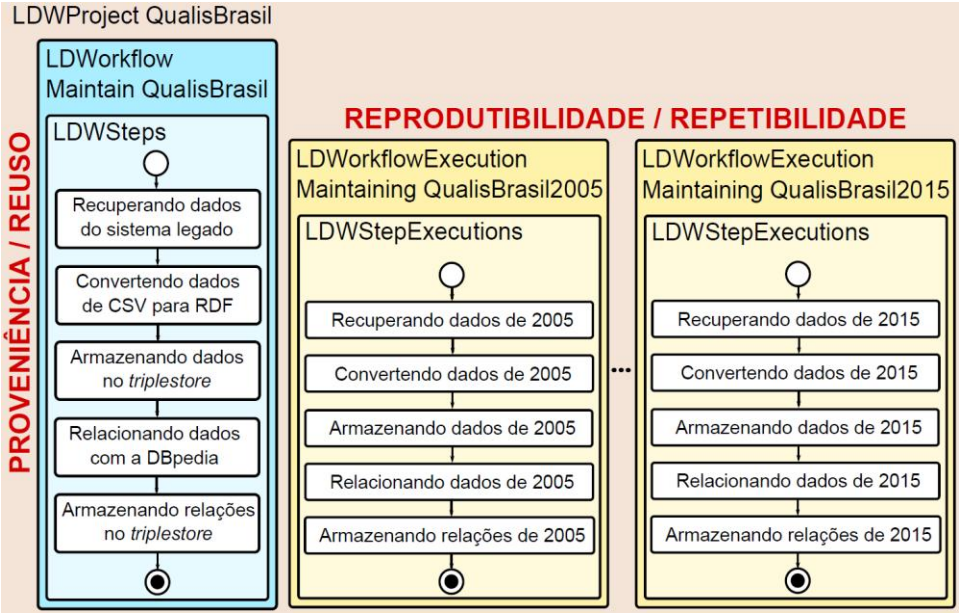

Figura 4: *LDWPROJECT* QUALISBRASIL – SEU *WORKFLOW* E EXECUÇÕES.
Fonte: elaborada pelos autores.

Tabela 3: SUMARIZAÇÃO DO PROCESSAMENTO ANO A ANO

EXECUÇÃO	ANO DE REFERÊNCIA	# AVALIAÇÕES CRIADAS	# AVALIAÇÕES PRESERVADAS
1°	2005	35.020	35.020
2°	2006	35.020	70.040
3°	2007	35.020	105.060
4°	2008	54.233	159.293

[8] Disponível a partir de: http://aksw.org/Projects/Sparqlify.html
[9] Disponível em: http://virtuoso.openlinksw.com/
[10] Disponível em: http://aksw.org/Projects/LIMES.html
[11] Um esforço conjunto para extrair informação estruturada da Wikipedia. Disponível em: http://dbpedia.org.

5°	2009	54.233	213.526
6°	2010	54.233	267.759
7°	2011	107.429	375.188
8°	2012	107.429	482.617
9°	2013	107.429	590.046
10°	2014	108.622	698.668
11°	2015	44.463	743.131

Fonte: elaborada pelos dos autores.

Ainda conforme a Figura 4, o *LDWorkflow* formalizado é reutilizado como um plano de ação para os 11 *LDWorkflowExecutions*, um para cada ano da série histórica, de 2005 a 2015. Para executar o plano de ação do QualisBrasil *LDWProject*, utilizou-se a *LODFlowEngine*[12], uma ferramenta que interpreta a base de conhecimento da LDWPO e executa automaticamente o plano de ação formalizado em um ambiente computacional controlado. No caso do QualisBrasil *LDWProject*, com as execuções, são preservadas 743.131 avaliações Qualis na *Web* de Dados. Como resultado, a base de dados conectados do Índice Qualis está à disposição para ser livremente utilizada em demais estudos bibliométricos ou cientométricos. A Tabela 3 sumariza o processamento e a preservação do Índice Qualis na Web de Dados nas 11 execuções.

Ao executar os *LDWorkflowExecutions* do QualisBrasil *LDWProject*, foi possível comprovar os requisitos traçados para a ontologia LDWPO. Neste sentido, com o estudo de caso, percebe-se o potencial da ontologia desenvolvida para: descrever *workflows* como planos para produção de bases de dados conectados; suportar a automatização da execução dos *workflows* na produção bases de dados conectados; reutilizar os planos ao longo do tempo, permitindo a reprodutibilidade das bases de dados conectados; e documentar os planos e as execuções de *workflows*, promovendo a proveniência e preservação de bases de dados conectados.

6 CONSIDERAÇÕES FINAIS

Neste artigo, é apresentada a *Linked Data Workflow Project ontology* (LDWPO). A ontologia desenvolvida é baseada na premissa que a LDWPO modela o planejamento e a execução de *workflows* para preservação de bases de dados conectados na *Web* de Dados.

Para suportar esta visão, a LDWPO representa os elementos de conhecimento sobre: projetos, pessoas, *workflows*, passos e artefatos. Além disso, a ontologia também reutiliza elementos das fontes de conhecimento: a) Dublin Core; b) DOAP; c) FOAF; d) OPMV; e) PROV-O; e f) PWO.

[12] Ferramenta computacional disponível em: https://github.com/AKSW/LODFlow/tree/master/tools/LODFlowEngine

Ao empregar a LDWPO na preservação de uma base de dados conectados no campo da Cientometria, é demonstrada sua aplicabilidade sob a ótica da gestão de projetos. Em particular, a ontologia suporta um projeto e seu respectivo plano para preservar o histórico do Índice Qualis na forma de dados abertos conectados na *Web* de Dados. O plano é executado de forma automática em um ambiente controlado, munido de tecnologias do *Linked Data Stack*. Disto, como resultado adicional, ressalta-se que a manutenção do índice Qualis na *Web* de Dados subsidia o compartilhamento do referido índice a demais estudos da Ciência da Informação.

Considera-se a LDWPO como um passo importante na área da *Web* Semântica, ao mediar a proveniência e a reprodutibilidade de bases de dados conectados na *Web* de Dados. Como trabalhos futuros, pretende-se evoluir a ontologia desenvolvida, assim como adotá-la em outros estudos de caso. Ademais, almeja-se implementar uma ferramenta computacional como interface à LDWPO, incorporando a ontologia ao *Linked Data Stack* para promover o suporte integrado na gestão de outras bases de dados conectados.

AGRADECIMENTOS

Os autores agradecem à Fundação Araucária pelo suporte financeiro (Projeto n° 601/2014 - Modelo para Compartilhamento de Informações sobre Pesquisas baseado em *Linked Open Data* para Estudos Cientométricos) e a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES - Processo número: 18228/12-7).

REFERÊNCIAS

AKSW. **Agile Knowledge and Semantic Web**. Disponível em: http://aksw.org/About.html. Acesso em: 17 de novembro de 2015 21:00

ALTINTAS, I; BARNEY, O; JAEGER-FRANK, E. Provenance collection support in the kepler scientific workflow system. **Lecture Notes in Computer Science**, v. 4145, p. 118–132, 2006.

AUER, S. **Towards Agile Knowledge Engineering: Methodology, Concepts and Applications**. Tese, University of Leipzig – Leipziger Informatik-Verbund (LIV), Leipzig, 2007.

AUER, S. Introduction to lod2. In AUER, S.; BRYL, V.; TRAMP, C (ed). **Linked Open Data – Creating Knowledge Out of Interlinked Data**. Springer-Verlag, 2014.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web. **Scientific American**, v. 284, n. 5, p. 34–43, 2001.

BOURQUE, P.; FAIRLEY, R. E. **Guide to the Software Engineering Body of Knowledge (SWEBOK)**. IEEE, 2014.

DUBLIN CORE. **Dublin Core Metadata Element - Version 1.1**. Disponível em: http://dublincore.org/documents/dces/. Acesso em: 17 de novembro de 2015 21:00.

DOAP. **Description of a Project vocabulary**. Disponível em: https://github.com/edumbill/doap/blob/master/schema/doap.rdf. Acesso em: 17 de novembro de 2015 21:00.

FOAF. **FOAF Vocabulary Specification 0.99**. Disponível em: <http://xmlns.com/foaf/spec/>. Acesso em: 17 de novembro de 2015 21:00.

GANGEMI, A.; PERONI, S.; SHOTTON, D.; VITALI, F. A pattern-based ontology for describing publishing workflows. In **Proceedings of the 5th Workshop on Ontology and Semantic Web Patterns** (WOP2014), 2014, Riva del Garda, Italy. Anais... Riva del Garda: CEUR-WS.org, 2014. p. 2–13.

GÓMEZ-PÉREZ, A; CORCHO, O.; FENÁNDEZ-LÓPEZ, M. **Ontological Engineering: with examples from the areas of knowledge management, e-commerce and the semantic web**. Heidelberg: Springer, 2004.

HULL, D.; WOLSTENCROFT, K.; STEVENS, R.; GOBLE, C.; POCOCK, M. R.; LI, P.; OINN, T. Taverna: a tool for building and running workflows of services. **Nucleic Acids Res**., v. 34, p. 729–732, 2006.

KEPLER. **The Kepler Project**. https://kepler-project.org/. Acesso em: 25 de Agosto de 2015 10:00.

LEBO, T.; SAHOO, S.; MCGUINNESS, D.; BELHAJJAME, K.; CHENEY, J.; CORSAR, D.; GARIJO, D.; SOILAND-REYES, S.; ZEDNIK, S.; ZHAO, J. **PROV-O: The prov ontology**. Disponível em: http://www.w3.org/TR/prov-o/. Acesso em: 13.01.2015.

LUDÄSCHER, B.; ALTINTAS, I.; BERKLEY, C.; HIGGINS, D.; JAEGER, E.; JONES, M.; LEE, E. A.; TAO, J.; ZHAO, Y. Scientific workflow management and the kepler system. **Concurrency and Computation: Practice and Experience**, v. 18, i. 10, p. 1039–1065, 2006.

MOREAU, L.; CLIFFORD, B.; FREIRE, J.; FUTRELLE, J.; GIL, Y.; GROTH, P.; KWASNIKOWSKA, N.; MILES, S.; MISSIER, P.; MYERS, J.; PLALE, B.; SIMMHAN, Y.; STEPHAN, E.; van den BUSSCHE, J. The open provenance model core specification (v1.1). **Future Generation Computer Systems**, v. 27, n. 6, p. 743–756, 2011.

NOY, N. F.; MCGUINNESS, D. L. **Ontology Development 101: A Guide to Creating Your First Ontology**. Disponível em: <http://wwwksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness.pdf>. Acesso em: 02 abril 2008 17:00.

van NUFFELEN, B.; JANEV, V.; MARTIN, M.; MIJOVIC, V.; TRAMP, S. Supporting the linked data life cycle using an integrated tool stack. in: AUER, S.; BRYL, V.; TRAMP, S. (eds.). **Linked Open Data – Creating Knowledge Out of Interlinked Data**. Springer-Verlag, 2014.

RAUTENBERG, S.; BURDA, A. LINKED OPEN DATA PARA CIENTOMETRIA: Compartilhando e Mantendo o índice Qualis na Web de Dados In: ENCONTRO BRASILEIRO DE BIBLIOMETRIA E CIENTOMETRIA, 5., 2016, São Paulo. **Anais**... São Paulo: USP, 2016. p. A34

SCUFLE2. **incubator-taverna-language/scufl2.rdf** Disponível em: https://github.com/apache/incubator-taverna-language/blob/master/taverna-scufl2-schemas/src/main/resources/org/apache/taverna/scufl2/rdf/scufl2.rdf. Acesso em: 10 de Julho de 2016 18:00.

SOWA, J.; ZACHMAN, J. Extending and formalizing the framework for information systems architecture. **IBM Systems Journal**, v. 31, p. 590-616, 1992.

STUDER, R.; DECKER, S.; FENSEL, D.; STEFFEN, S. Situation and perspective of knowledge engineering. In CUENA, J.; DEMAZEAU, Y.; SERRANO, A. G.; TREUR, J. (eds). **Knowledge Engineering and Agent Technology**. IOS Press, Oxford, 2004.

SUCUPIRA. **Plataforma Sucupira**. Disponível em: https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/veiculoPublicacaoQualis/listaConsultaGeralPeriodicos.jsf. Acesso em: 03 de Abril de 2016 21:00.

SURE, Y.; STUDER, R. On-To-Knowledge methodology. In DAVIES, J.; FENSEL, D.; van HARMELEN, F. (eds). **On-To-Knowledge: Semantic Web enabled Knowledge Management**. J. Wiley & Sons, 2002.

TAVERNA. **Taverna - open source and domain independent Workflow Management System**. Disponível em: http://www.taverna.org.uk/. Acesso em: 25 de Agosto de 2015 10:00.

WEBQUALIS. **Sistema WebQualis - Portal Capes**. Disponível em: http://qualis.capes.gov.br/webqualis/principal.seam. Acesso em: 25 de Agosto de 2013 10:00.

WFMC. **The workflow reference model**. Technical report, The Workflow Management Coalition, 1995.