Ontologias para Nanopublicações do Domínio de Análise de Redes de Colaboração Científica

Romário Laltany G. da Silva[1], Andréa Sabedra Bordin[2], Alexandre Leopoldo Gonçalves[2], Alessandro Botelho Bovo[3]

[1]Universidade Federal do Pampa
Alegrete, RS - Brasil

[2]Universidade Federal de Santa Catarina
Araranguá, SC - Brasil

[3]Universidade Tecnológica Federal do Paraná
Londrina, PR - Brasil

laltanyromario@gmail.com, {andrea.bordin, a.l.goncalves}@ufsc.br

alessandrobovo@utfpr.edu.br

Abstract. *Nanopublication is a model of granular representation of scientific information that emerges as an alternative to the traditional publication model. Its implementation with semantic technologies demands the use of general and domain-specific vocabularies. This article presents the detailed process of reuse and development of ontologies for the levels of the nanopublications architecture of the Scientific Collaboration Network Analysis domain. The Neon methodology was used to explain the creation of the ontology of the assertion layer. As a result, a set of ontological resources was obtained to support the extraction of information, the structuring and the recovery of nanopublications in this domain.*

Resumo. *A nanopublicação é um modelo de representação granular de informações científicas que surge como uma alternativa ao modelo de publicação tradicional. Sua implementação com tecnologias semânticas demanda a utilização de vocabulários gerais e específicos de domínio. Este artigo apresenta o processo detalhado de reutilização e desenvolvimento das ontologias para os níveis da arquitetura de nanopublicações do domínio de Análise de Rede de Colaboração Científica. A metodologia Neon foi utilizada para explicitar a criação da ontologia do nível de asserção. Como resultado se obteve um conjunto de recursos ontológicos para dar apoio a extração das informações, a estruturação e a recuperação de nanopublicações deste domínio.*

1. Introdução

Nanopublicações (NP) são pequenas unidades de informação, às quais são adicionadas informações de proveniência [Mons and Velterop 2009]. A abordagem, implementada com tecnologias de Web Semântica como RDF e Ontologias, surgiu como uma alternativa de representação de informação científica que permite uma recuperação granular da informação, a exploração do conhecimento científico através da inteligência de máquina e serve como ponto de entrada para bancos de dados científicos [Kuhn et al. 2018].

A arquitetura mínima de uma NP é composta pelos níveis de **Asserção** (*assertion*), que representa uma unidade mínima de informação, como por exemplo, um efeito observado de um medicamento em uma doença; **Proveniência** (*provenance*) que representa as informações sobre a origem da asserção, como a identificação dos autores da publicação, link para DOIs, etc; e **Informação da NP** (*publication info*) que mantém informações sobre criação da própria NP [Groth et al. 2010]. Para cada nível, deve existir uma ontologia que represente os conceitos dos dados estruturados na NP.

Um passo importante para o uso de NP é encontrar ou desenvolver vocabulários apropriados para descrever os conceitos do domínio e da proveniência dos dados envolvidos na NP. Nos trabalhos encontrados, percebeu-se que o processo de escolha ou criação das ontologias necessárias aos níveis da arquitetura da NP raramente é abordado em detalhes, ou seja, os trabalhos partem da pré-existência das ontologias. A representação dos conceitos necessários para representação de NP é uma atividade crucial para extração, armazenamento e recuperação dessas NP. Dessa forma, deve ser realizada através de um processo bem estabelecido, que garanta a correta expressividade dos conceitos.

Este trabalho tem como objetivo apresentar o processo de escolha e desenvolvimento das ontologias dos níveis da arquitetura de NP do domínio de Análise de Redes de Colaboração Científica (ARCC). O referido processo faz parte do método apresentado no trabalho de [da Silva and Bordin 2020] e neste artigo é apresentado de forma detalhada.

Como principais contribuições deste trabalho destacam-se: a) o processo detalhado de construção da ontologia da camada de asserção utilizando a metodologia NeOn; b) o artefato ontológico resultante deste processo, nomeado *Scientific Collaboration Network Analysis (SCNA Ontology)*; c) as escolhas das ontologias das demais camadas da arquitetura da NP.

O artigo está organizado da seguinte forma: Na Seção 2 é apresentado o processo de desenvolvimento das ontologias; na Seção 3 é apresentado, através de um exemplo, como os conceitos fornecem semântica a uma NP e, na Seção 4, as considerações finais.

2. Metodologia

No trabalho de Silva e Bordin [da Silva and Bordin 2020] foi proposta a extensão da arquitetura mínima da NP, com a criação de um quarto nível denominado Proveniência dos Dados. Diferentemente do nível de Proveniência, que mantém informações sobre o estudo a partir do qual as asserções foram geradas, o nível de Proveniência dos Dados mantém informações sobre a proveniência dos dados utilizados na criação das redes estudadas. Nesta seção é apresentado o processo de escolha e construção das ontologias necessárias para representar cada nível da arquitetura de NP no domínio proposto.

2.1. Ontologia do Nível de Asserção

Para o nível de asserção, criou-se uma ontologia capaz de representar os conceitos referentes aos resultados presentes nos estudos de Análises de Redes de Colaboração Científica (ARCC). Não se optou pela construção de uma ontologia a partir do zero, visto que uma pesquisa preliminar revelou a existência dos conceitos necessários em ontologias já existentes. Por isso, utilizou-se a metodologia NeOn de Suarez [Suarez 2010] para a construção de ontologias em rede a partir da reutilização de ontologias. A Figura 1 representa o processo adotado para a construção desta ontologia de domínio.

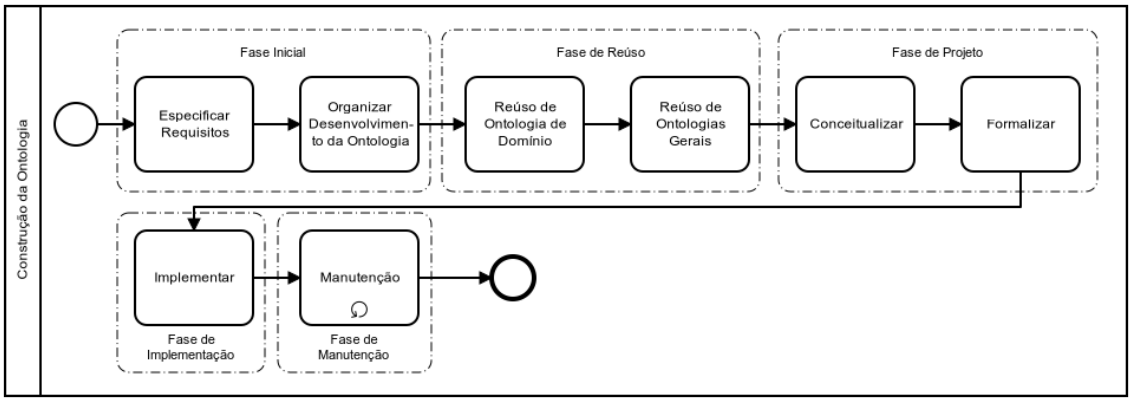

Figura 1. Atividades de criação ontologia. Adaptado de: [Suarez 2010].

2.1.1. Fase Inicial

Na atividade de especificação dos requisitos da ontologia foram criadas Questões de Competências (QCs) a partir de um conjunto de estudos de ARCC. As 52 QCs[1] foram organizadas e agrupadas em uma planilha e, em seguida, foram validadas por uma especialista do domínio para evitar possíveis conflitos ou contradições. Esta validação foi realizada à luz dos seguintes critérios: correção, consistência, compreensão, exclusão de ambiguidade e concisão. Um excerto das questões é mostrado na Tabela 1.

Por fim, realizou-se a criação de um Pré-Glossário[2] a partir da extração de termos presentes na lista de QCs, dos termos das respostas esperadas e das entidades nomeadas (como pode ser observado nos termos em negrito da Tabela 1). Com os resultados obtidos, criou-se o Documento de Especificações de Requisitos da Ontologia (DERO).

Tabela 1. Excerto das QCs

QC	Entrada	Resposta Esperadas
Qual a **densidade** e **grau médio** das **redes** de colaboração científica de um determinado **escopo**?	**Escopo**: **Área de Conhecimento**	**Densidade**: **0,006** **Grau Médio**: **2,45**
Qual o **número de atores** e **relações** das **redes** de um determinado **escopo** em uma determinada **localidade**?	**Escopo**: **IES** **Localidade**: **Portugal**	**Nº de atores**: **553** **Nº de relações**: **2419**
Quais as **centralidades de grau**, **intermediação** e **proximidade** dos **atores** das **redes** de determinado **escopo**?	**Escopo**: **PGP**	**L**: **2.0**, **0.074**, **13.08** **B**: **9.0**, **0.387**, **9.67** **V**: **21.0**, **0.542**, **60.0** **D**: **4.0**, **0.104**, **15.0**

Na atividade de organização do desenvolvimento, fez-se a escolha do modelo de ciclo de vida e o cenário de construção da ontologia. Optou-se pelo modelo cascata, diante da estabilidade e o conjunto finito de métricas presentes nos estudos do domínio de ARCC. Já o cenário escolhido foi o de Reúso de Recursos Ontológicos. Baseado nessas escolhas, o processo decisório da NeOn indicou o modelo cascata de cinco fases como o mais adequado para este projeto. A Figura 1 já contextualiza as atividades dentro deste modelo de cinco fases.

[1]QCs versão final: `https://tinyurl.com/pbjj82p4`
[2]Pré-Glossário: `https://tinyurl.com/vnmupkrp`

2.1.2. Fase de Reúso

Ao analisar os termos presentes no pré-glossário do DERO, percebeu-se a possibilidade de reúso de ontologias de domínio para os representar os termos relacionados à métricas, tais como: Centralidade de Grau, de Intermediação, de Proximidade; Grau Médio, dentre outras. Notou-se também a necessidade do uso de ontologias gerais para representar termos referentes à localização geográfica da rede e períodos temporais dos dados.

No que tange à reutilização de ontologias de domínio, a busca foi realizada em repositórios ontológicos como Swoogle e AberOWL, e também em repositórios científicos. No final desta atividade, identificou-se 3 possíveis recursos candidatas ao reúso: SemSNA, proposto por [Ereteo et al. 2009]), SNAMetric, proposto por [Bordin et al. 2015]) e ontoSELF-v2, proposto por [Yu 2008]). As ontologias passaram por uma avaliação, considerando a capacidade de resposta às questões de competência criadas na fase inicial. Constatou-se que a SNAMetric conseguiu responder todas as QCs a partir de seus conceitos, diferentemente das outras candidatas que responderam em partes ou não tiveram a capacidade de resposta.

Em relação às ontologias gerais ou comuns, realizou-se um processo de busca que resultou em um conjunto de ontologias. Através de um estudo comparativo, escolheu-se a ontologia *Time-OWL*, para representar os conceitos sobre o domínio de tempo, e a ontologia *Region* proposta no escopo do projeto NAZOU[3], para representar conceitos referentes a regiões geográfica básicas.

Para manter apenas os conceitos pertinentes para a nova ontologia, na *Time-OWL* reusou-se somente o módulo *Temporal Entity*, permitindo a representação de um ponto na linha de tempo e um intervalo temporal. Da ontologia *Region* manteve-se apenas o módulo *Administrative Region*, a fim de representar posições geográfica como país, estado e cidade; e a classe *Geographical Region*, que estabelece o continente de um país.

2.1.3. Fase de Projeto

Esta fase engloba as atividades de Conceituação e Formalização, que guiam a criação de um modelo formal a partir de um modelo conceitual. Para a criação de um modelo conceitual de estudos de ARCC, utilizou-se a metodologia Methontology proposta por Lopez et al. [Fernández-López et al. 1997], recomendada pela NeOn.

A partir do pré-glossário criou-se um glossário de termos do domínio[4] mais importantes para a ontologia. Os termos tipificados como "conceito" no glossário foram utilizados na construção da taxonomia dos termos.

A próxima tarefa foi a criação de um dicionário de conceitos[5] que especifica as instâncias, atributos, propriedades e relações binárias dos conceitos da ontologia desenvolvida. Todas as relações binárias entre as classes da ontologia[6], os atributos[7] e as constantes[8] e instâncias relevantes[9] foram descritas em documentos *online*.

[3]Projeto NAZOU: `http://nazou.fiit.stuba.sk`
[4]Glossário de termos: `https://tinyurl.com/y5ezsnxa`
[5]Dicionário de conceitos: `https://tinyurl.com/8scfh9pk`
[6]Documentação das relações binárias: `https://tinyurl.com/hyrf64b2`
[7]Documentação dos atributos: `https://tinyurl.com/b2tx4kch`
[8]Documentação das constantes: `https://tinyurl.com/kcrt32m6`
[9]Tabela de instâncias: `http://tiny.cc/ix3ggz`

A Figura 2 explicita o diagrama conceitual da ontologia nomeada *Scientific Collaboration Network Analysis (SCNA) Ontology*, onde cada cor representa a ontologia de origem do conceito. Como pode ser observado, cada rede presente em uma estudo do domínio de ARCC (*SCNA*) deve ter uma relação (*hasEscope*) com um escopo (*Scope*), por exemplo, uma rede de colaboração do evento BreSci deve ter *Event* como escopo.

Para representar o relacionamento entre uma rede (*Network*) e seus atores (*Actor*), utiliza-se o predicado *hasActor*. Um ator pode se relacionar com qualquer métrica de ator através de um predicado exclusivo para cada métrica. Por exemplo, *Actor* e *Closeness Centrality* se relacionam através do predicado *SNAMetricClosenessCentrality*.

Uma rede (*Network*) pode se relacionar com uma região geográfica (*Region*) através do predicado *hasRegion*, indicando a localização da rede, assim como pode ter uma relação com um período de tempo (*Temporal Entity*) a partir do predicado *hasTime*. Por exemplo: uma rede (*Network*) de colaboração científica da área de Ciência da Computação no Brasil tem escopo (*Knowledge Area*) Ciência da Computação cujos dados referem-se ao país (*Country*) Brasil .

A ontologia permite representar relações cruzadas entre os tipos de escopo, por exemplo: uma rede (*network*) de colaboração do programa de pós-graduação em engenharia de conhecimento da UFSC tem escopo (*PostGraduate*), o qual apresenta uma relação com *High Education Institution* a partir do predicado *hasHEI*.

Na formalização, o modelo conceitual foi utilizado para a criação de um modelo formal. Os conceitos encontrados foram transformados em classes, que foram estruturadas em taxonomias já definidas. Os atributos de instâncias foram definidos como propriedades de dados, assim como as relações binárias em propriedades de objetos.

2.1.4. Fase de Implementação e Manutenção

A ontologia[10] foi implementada em OWL e sua avaliação aconteceu a partir de atividades de verificação e validação. A verificação foi realizada durante todo o processo de desenvolvimento, visando garantir o produto esperado em cada fase. Para a validação, transcreveu-se todas as QCs presentes no DERO para a linguagem de consulta SPARQL. As saídas das consultas foram avaliadas de acordo com os critérios propostos pela metodologia, chegando-se a conclusão que todas responderam integralmente às questões.

Na fase de manutenção, identificou-se a necessidade de remover o termo SCNAS, cujo propósito era representar um estudo de análise de rede, e alterar as classes *Scope* e *Journal*, anteriormente nomeadas como *Type* e *Periodic*. Também houve alteração nos nome de algumas propriedades de objeto, a fim de padronizá-las.

2.2. Ontologia do Nível de Proveniência

Este nível tem como objetivo representar dados relacionados a proveniência dos resultados presentes no nível de asserção. Utilizou-se a ontologia PROV-O[11] para representar os autores da publicação a partir da qual a asserção se originou (classe *Person*), a data de publicação (propriedade *generatedAtTime*) e o DOI da publicação (propriedade *hadPrimarySource*). PROV-O é um padrão recomendado pelo W3C que fornece classes, relações

[10]SCNA Ontology: `https://github.com/Laltany/SCNAS-Ontology`
[11]Ontologia PROV-O: `https://www.w3.org/TR/prov-o/`

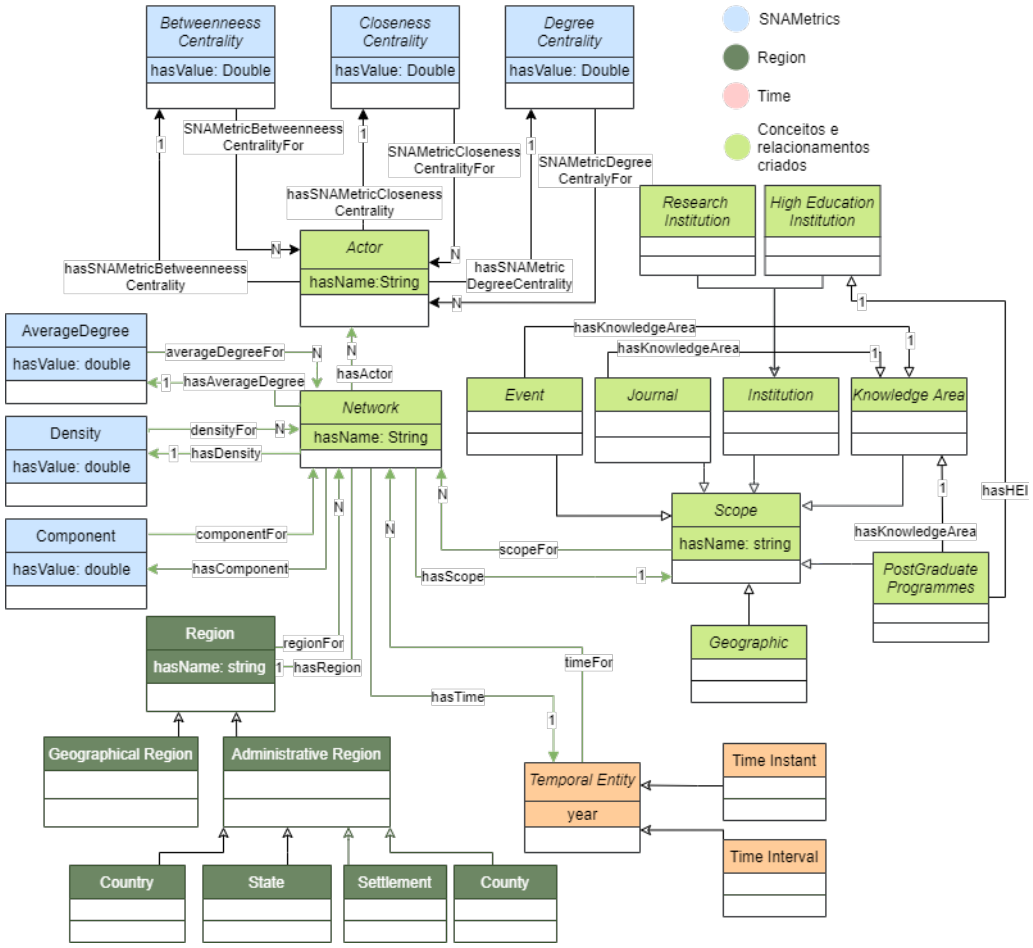

Figura 2. Conceitos e Relacionamentos da *SCNA Ontology*

e restrições para representar qualquer tipo de proveniência. Para representar o título e o periódico/conferência do artigo foram utilizados termos da DCMI Metadata Terms[12].

2.3. Ontologia do Nível de Informações da NP

Este nível tem como objetivo representar informações sobre a origem da NP criada. A ontologia PROV-O foi utilizada para representar o criador da NP (classe *Person*) e a data de criação (propriedade *generatedAtTime*).

2.4. Ontologia do Nível de Proveniência dos Dados

Este nível tem como propósito representar a proveniência dos dados utilizados na construção das redes analisadas nos estudos, tais como o período temporal dos dados coletados (ex.: de 2010 a 2015) e a identificação do repositório que proveu os dados (ex. Scopus). Utilizou-se a ontologia PROV-O que, a partir das propriedades *startedAtTime* e *endedAtTime*, representa o período dos dados coletados para a análise. Além disso, foi utilizada a propriedade *atLocation* para indicar o repositório.

[12]DCMI Metadata Terms: `https://www.dublincore.org/specifications/dublin-core/dcmi-terms/\#terms-publisher`

3. Nanopublicação de ARCC

Nanopublicações são implementadas através de grafos RDF nomeados utilizando a sintaxe TriG[13]. A Figura 3 explicita graficamente os cinco grafos de uma NP de ARCC, onde as elipses representam literais ou instâncias de um determinado tipo representado nas ontologias definidas nas seções anteriores.

Partindo do grafo *head* tem-se que uma NP relaciona-se com outros quatro grafos. O grafo *assertion* contém afirmações que indicam que uma rede tem escopo de área de conhecimento "Ciência da Computação", densidade "0.5" e grau médio "2.0". O grafo *provenance* indica que as triplas da asserção são provenientes do artigo cujo título é "Análise da Rede de Colaboração Científica dos Pesquisadores da Área de Ciência da Computação", de autoria de "Andréa", publicado em 04/05/2020 e DOI 11.1111/1.111111. O grafo *pubinfo* indica que o autor da NP é "Romário" e que ela foi criada em "04/05/2021". Por fim, o grafo *dataprovenance* indica que os dados de coautoria utilizados para gerar a rede compreendem um período que vai de "02-01-2009" a "30-12-2019" e foram coletados do repositório "Scopus".

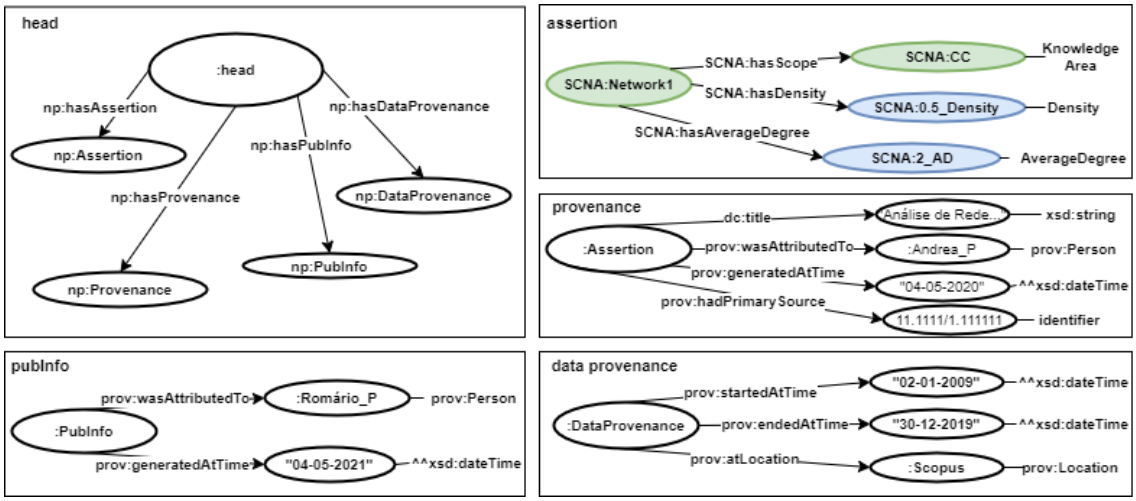

Figura 3. Visão gráfica de uma NP de ARCC

4. Considerações Finais

Neste artigo foi apresentado o processo de escolha e desenvolvimento de um conjunto de ontologias necessárias a cada nível da arquitetura de NP do domínio de Análise de Redes de Colaboração Científica. Além disso, foi demonstrado como os conceitos definidos nas ontologias são utilizados na criação de uma NP do referido domínio.

Entende-se que o detalhamento do processo apresentado neste trabalho se caracteriza como uma contribuição ao entendimento e disseminação da abordagem de NP como alternativa ao problema do volume e recuperação de informações científicas. De forma mais específica, a ontologia da camada de asserção (*SCNA Ontology*) caracteriza-se como um artefato maduro, que pode ser reutilizado em outros trabalhos deste domínio. Ela resultou de um processo executado com rigor metodológico, a partir de um grande número

[13]https://www.w3.org/TR/trig/

de questões de competências (52), que privilegiou tanto o reúso de outros recursos ontológicos, como a criação de conceitos novos.

É importante destacar que a proposta de uso de nanopublicações está fortemente atrelada à representação de conhecimento, o que caracteriza um gargalo na sua ampla e rápida adoção, uma vez que demanda a existência de vocabulários de diferentes domínios. Ainda assim, já existem aplicações em domínios diferentes do de Ciências da Vida (onde existem mais de 10 milhões de NP publicadas) e muitos estudos a apontam com uma promissora abordagem para recuperação e proveniência dos dados.

A abordagem de NP suscita diversas possibilidade de trabalhos futuros, como o desenvolvimento de extrator de dados de publicações para a criação de NPs; a comparação desta abordagem de recuperação de informação com abordagens tradicionais; a implementação de uma aplicação que permita a criação de NPs de forma colaborativa, onde os próprios autores e usuários interessados possam criar novas NPs, alimentando a base de conhecimento do domínio em questão, dentre outras.

Referências

Bordin, A. S., Souza, J. A. d., and Gonçalves, A. L. (2015). *Framework baseado em conhecimento para análise de rede de colaboração científica*. PhD thesis, Universidade Federal de Santa Catarina, Florianópolis.

da Silva, R. L. G. and Bordin, A. S. (2020). Método para criação e recuperação de nanopublicações uma aplicação no domínio de análise de redes de colaboração científica. In *Anais do XIV Brazilian e-Science Workshop*, pages 73–80. SBC.

Ereteo, G., Buffa, M., Gandon, F., and Corby, O. (2009). Analysis of a Real Online Social Network using Semantic Web Frameworks. In *ISWC 2009*, volume 5823/2009, pages 180–195, Washington, United States.

Fernández-López, M., Gómez-Pérez, A., and Juristo, N. (1997). Methontology: From ontological art towards ontological engineering. In *Proceedings of the Ontological Engineering AAAI-97 Spring Symposium Series*. American Asociation for Artificial Intelligence. Ontology Engineering Group OEG.

Groth, P., Gibson, A., and Velterop, J. (2010). The anatomy of a nanopublication. *Information Services & Use*, 30(1-2):51–56.

Kuhn, T., Chichester, C., Meroño-Peñuela, A., Dumontier, M., Malic, A., Poelen, J. H., Hurlbert, A. H., Ortiz, E. C., Furlong, L. I., Queralt-Rosinach, N., Banda, J. M., Willighagen, E., Ehrhart, F., Evelo, C., and Malas, T. B. (2018). Nanopublications: a growing resource of provenance-centric scientific linked data. In *14th IEEE International Conference on E-Science*, pages 83–92. IEEE Computer Society.

Mons, B. and Velterop, J. (2009). Nano-publication in the e-science era. In *Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009)*, pages 14–15.

Suarez, M. d. C. (2010). NeOn Methodology for Building Ontology Networks: Specification, Scheduling and Reuse. *Landscape and Urban Planning*, 102(4):268.

Yu, E. (2008). *Social Network Analysis Applied to Ontology 3D Visualization*. PhD thesis, Miami University, Miami.