# Método para criação e recuperação de nanopublicações: uma aplicação no domínio de Análise de Redes de Colaboração Científica

**Romário Laltany G. da Silva**[1], **Andréa Sabedra Bordin**[1]

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

laltanyromario@gmail.com, andreabordin@unipampa.edu.br

***Abstract.*** *An excess of unstructured information is available in scientific repositories, causing problems in retrieving, extracting and reusing . To facilitate the retrieval of information, was proposed the use of nanopublication (NP), using semantic technologies for evidence important data, facilitating computational interpretation and reuse of statements.In this paper, we propose a method for the creation and recovery of NPs. The validation of the method is done in the domain of analysis of scientific collaboration networks, resulting in semantic artifacts capable of recovering information at a lower granularity level.*

***Resumo.*** *Um excesso de informações não estruturadas está disponível em repositórios científicos, ocasionando problemas para recuperá-las, extraí-las e reutilizá-las. Para facilitar a recuperação de informações, foi proposto o uso de nanopublicação (NP), que ao utilizar tecnologias semânticas, evidencia dados importantes, facilitando a interpretação computacional e o reúso de afirmações. Neste artigo, propõe-se um método para a criação e recuperação de NPs. A validação do método é feita no domínio de Análise de Redes de Colaboração Científica, resultando em artefatos semânticos capazes de recuperar informações em um nível de granularidade menor.*

## 1. Introdução

Ao longo dos anos um grande volume de publicações tem sido armazenado nos repositórios científicos digitais [Oliveira 2008]. Esse volume de dados ocasiona dificuldades na recuperação de informação. A busca por uma informação específica se torna difícil pela maneira como as publicações são armazenadas, comumente em formato não estruturado textual e sem o uso de técnicas de marcação e de destaque de seus conteúdos. Dessa forma, as informações se tornam ilegíveis para computadores dificultando a pesquisa, preservação, agregação de valor e o reúso das mesmas, evitando a extração máxima do seu potencial [Mons et al. 2011].

Sabendo que os computadores trabalham de maneira mais eficiente com dados estruturados, entende-se que o uso de artigos nesse formato não seja a melhor forma de disseminar conhecimento, já que na maioria das vezes o que realmente importa para os pesquisadores durante uma busca são pequenas unidades de informações, tais como os resultados obtidos em um estudo, denominadas de afirmações [Mons and Velterop 2009].

Nesse sentido, [Mons and Velterop 2009] criaram o conceito de nanopublicação (NP) com o objetivo de publicar resultados de pesquisas em formato de dados estruturados, mantendo informações sobre a proveniência dos mesmos. Na arquitetura miníma de uma NP o nível de asserção mantém informações relevantes obtidas em um estudo, no nível de proveniência é possível encontrar informações sobre a origem dos dados presentes na asserção, já no nível de proveniência da NP encontra-se informações sobre a origem da NP, tais como quem criou e quando ela foi criada. A implementação da arquitetura é feita através de tecnologias semânticas, tais como RDF, OWL, SPARQL, etc. De acordo com [Groth et al. 2010], o uso de NPs permite um melhoramento na recuperação e reúso de resultados presentes nos documentos, mantendo o contexto e os benefícios da abordagem tradicional, de forma que a atribuição, qualidade e procedência sejam relevantes.

Apesar dos benefícios entregues pela abordagem, a NP ainda é pouco utilizada fora do seu domínio de criação, Ciências da Vida.A NP foi projetada para ser extensível, podendo ser modificada conforme necessário e para diferentes domínios [Groth et al. 2013]. No entanto, observou-se que as *guidelines* propostas para a criação de NPs não parecem ser suficientes para efetivar sua operacionalização pois não oferecem um guia explicitando os passos necessários para a aplicação da abordagem. Além disso, maioria os artigos que trabalham com a abordagem focam em propósito distintos e não apresentam de maneira clara os passos necessário para a criação e recuperação de NPs.

Este artigo propõe um método para a criação e recuperação de NPs, servindo também como apoio às *guidelines*[1]. O método é validado através da sua aplicação no domínio de Análise de Redes de Colaboração Científica (ARCC). A escolha do domínio foi influenciada por sua estrutura, que é mais bem definida, visto que a maioria dos artigos apresentam seus resultados de acordo com as métricas de Análise de Redes Sociais. Como exemplo, a grande maioria dos artigos apresenta a densidade da rede, assim como os valores de centralidade de grau, etc. Verificou-se que as métricas utilizadas no estudo das redes trazem consigo resultados que podem ser interessantes para outros pesquisadores, que consequentemente poderiam utilizar desses dados em suas pesquisas.

Como contribuição principal, este trabalho apresenta um método para a criação e recuperação de NPs, além de uma série de artefatos que interessam ao domínio de ARCC, tais como: uma ontologia capaz de representar, recuperar e extrair conceitos do domínio de ARCC; uma ontologia para representar NPs de ARCC e uma aplicação semântica capaz de recuperar as NPs do domínio.

O artigo está organizado como segue: Fundamentação Teórica (Seção 2), Trabalhos Relacionados (Seção 3), Metodologia (Seção 4), Aplicação do Método (Seção 5) e Considerações Finais(Seção 5).

## 2. Fundamentação Teórica

A nanopublicação tem como ideia principal a subdivisão dos resultados científicos em pequenas partes denominadas de afirmações; a representação dessas afirmações em uma notação formal baseada em RDF; a adição de informações de proveniência neste nível atômico; o tratamento de cada uma destas pequenas entidades como uma publicação separada [Kuhn et al. 2013].

[1]http://www.nanopub.org/2013/WD-guidelines-20131215/

Um dos principais problemas da citação de dados científicos é a comprovação de que essas informações não foram alteradas ou corrompidas. Por isso, a arquitetura da NP deve ser capaz de identificar a origem e a autoria das afirmações estruturadas, assim como o autor da importação destas informações [Kuhn et al. 2016]. Isso permite a garantia da integridade dos dados e que os usuários possam avaliar a confiabilidade dos mesmos [Groth et al. 2010].

[Groth et al. 2013] propõe *guidelines* com três níveis básicos para a arquitetura de uma NP, onde: asserção (assertion) é uma unidade mínima de afirmação; proveniência (provenance) refere-se as informações sobre a origem da asserção, como a identificação dos autores, instituições, links para DOIs, etc; e informação da NP (publication info) que contém informações sobre a origem da NP, como o nome do autor da NP e a data de criação da mesma. Todos os níveis são expressos através de triplas RDF (sujeito-predicado-objeto), como exemplificado na Figura 1.

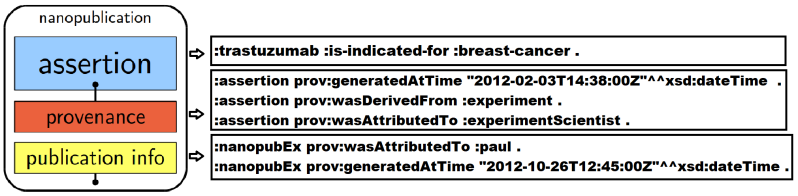


**Figura 1. Arquitetura mínima de uma NP. Fonte: nanopub.org**

Ainda de acordo com [Groth et al. 2013], independentemente do domínio no qual a NP será aplicada, é necessário que existam boas descrições do conhecimento para que se possa ter uma boa compreensão, reutilização e recuperação das informações desejadas. Por isso, a criação das NP deve ser guiadas por ontologias que descrevem e estabelecem relações entre os elementos e conceitos de um determinado domínio. Considera-se também que cada NP é única e tem sua própria IRI.

## 3. Trabalhos Relacionados

A nanopublicação é uma abordagem recente e pouco explorada. Através de um mapeamento sistemático, descobriu-se um total de 578 artigos publicados entre os anos de 2010 e 2019, nos repositórios ACM, IEEE Xplore, Scielo, Science Direct, Scopus e Google Acadêmico. A fim de resgatar somente os artigos que realmente trabalhavam com a NP, foram aplicados alguns critérios[2] de inclusão e exclusão, restando somente 57 publicações.

Analisou-se que 2015 foi o ano com mais publicações com o uso da abordagem. Como esperado, a maioria das publicações são do domínio de Ciências da Vida, com 47,4% das publicações. Os artigos selecionados focam em propósitos distintos, mas não definem os passos necessários para a criação e recuperação de NPs de qualquer domínio. Os artigos que exploraram com mais detalhes a criação de NPs são os apresentados a seguir.

No trabalho de [Beck et al. 2012] utilizou-se a abordagem de NP para disponibilizar um grande conjunto de dados linkados a partir das publicações do *Genome-Wide*

[2] http://tiny.cc/1gsclz

*Association Study*, permitindo que os pesquisadores que contribuíram com as publicações do GWAS possam ser adequadamente referenciados. O artigo deixa claro que a criação de NPs foi através das *guidelines* propostas pela *Open Phacts* e explicita as ontologias utilizadas na arquitetura das NPs criadas. Além disso, mostra o método adotado para a recuperação das NPs. No entanto, não há detalhamento da aplicação das guidelines de criação das NPs.

Em [Maiatsky et al. 2018] cria-se NPs para apresentar dados de fenômenos culturais presentes em documentos históricos, a fim de se obter uma melhora na representação do conhecimento e evidenciar a proveniência das informações. Apesar do artigo explicitar a extração dos dados realizada, pouco se detalha sobre os passos necessários para estruturar os dados em formato de NP. Além disso, foca na recuperação das NPs através do sistema VICOGLOSSIA.

Já em [Lipani et al. 2014] é abordada a criação de NPs do domínio de Recuperação de Informação (RI), a fim de evidenciar informações que auxiliem pesquisadores na reprodução de experiências de RI. Apesar de explicitar um fluxo de atividades para a extração de NP, pouco explica sobre a arquitetura e a recuperação das mesmas.

## 4. Método de Criação e Recuperação de Nanopublicações

O método proposto consiste em um conjunto de atividades dividas em 3 grupos, as quais estão representadas na Figura 2. Como pode ser visto, os grupos de atividades estão delimitados em um quadro pontilhado em cor cinza, as linhas pretas tracejadas representam o uso de um artefato em uma determinada atividade, a linha contínua expressa a colaboração de um especialista do domínio em uma determinada atividade.

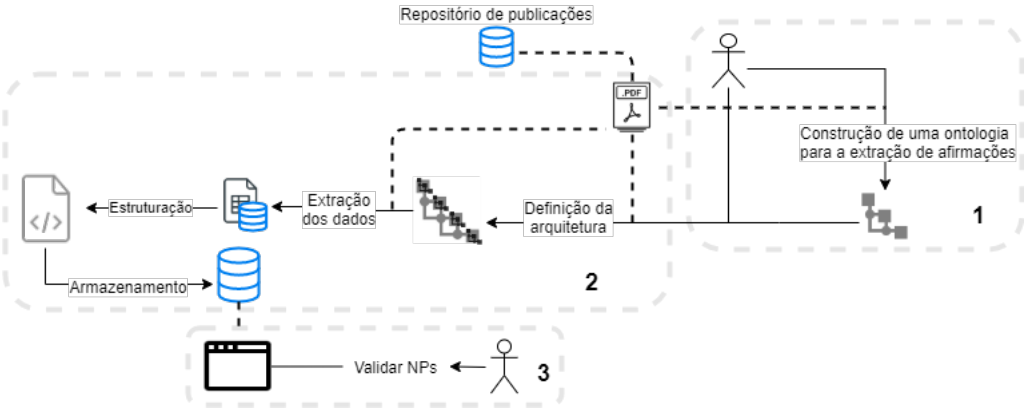


**Figura 2. Atividades para a criação de nanopublicações**

O primeiro grupo tem como propósito criar uma ontologia capaz de representar conceitos de um determinado domínio. Deve-se levar em consideração que os conceitos devem ser estabelecidos considerando a necessidade de recuperação das informações contidas nas publicações do domínio. Esta ontologia serve para guiar as atividades de extração das afirmações e estruturação das mesmas em triplas RDF. Para que se possa criar uma ontologia consistente aconselha-se seguir uma metodologia adequada ao cenário de criação da ontologia proposta. Como entrada para a atividade de construção da ontologia, tem-se um conjunto de publicações e o conhecimento de um especialista de determinado domínio.

O segundo grupo está relacionado à criação e armazenamento da NP. Primeiramente deve-se definir uma arquitetura para a NP, através da análise de documentos e conversas com especialistas no domínio. Essa definição consiste em estudar a necessidade de estender ou não a arquitetura miníma da NP e escolher as ontologias que irão compor os níveis da arquitetura definida. Entende-se que a extensão da arquitetura mínima deve ser feita quando houver a carência de informações importantes para dar contexto às afirmações presentes no nível de asserção.

As ontologias estabelecidas irão fornecer os conceitos que representarão as informações que serão mantidas em cada nível da arquitetura adotada, de acordo com suas respectivas finalidades. Como visto na Figura 1, a arquitetura de uma NP é composta minimamente por: asserção, proveniência da asserção e informações da NP. A ontologia resultante das atividades do primeiro grupo contém os conceitos necessários para evidenciar as afirmações presentes nos documentos do domínio e deve ser adotada para o nível de asserção. Para os níveis de proveniência, é necessário utilizar ontologias que representem a origem das asserções e da NP criada como, por exemplo, a PROV-O. Caso um novo nível tenha sido criado, é preciso fazer uso de uma ontologia que atenda o objetivo de sua criação.

Logo após, na atividade de extração dos dados, os dados das publicações devem ser extraídos de forma manual ou automática, a partir do conjunto de ontologias definidas na arquitetura e armazenados temporariamente. Após é necessário estruturá-los na arquitetura de NP estabelecida e, por fim, deve-se armazenar as NPs em um banco de triplas RDF.

Com o propósito de mostrar a viabilidade da NP, o terceiro e último grupo consiste em realizar a validação das NPs. A validação é realizada através da recuperação das NPs, por especialistas do domínio, e permite verificar se as informações contidas nas NPs estão corretas. A busca pelas NPs pode ser feita diretamente através de consultas SPARQL ou de alguma interface desenvolvida para este fim.

## 5. Aplicação do Método

A fim de validar o método proposto neste artigo, optou-se pelo domínio de Análise de Redes de Colaboração Científica. No método Análise de Redes Sociais (ARS), sistemas podem ser analisados sob a óptica de uma rede, composta por um conjunto de nós conectados através de arestas. Estes nós representam atores que correspondem a entidades, pessoas, empresas ou organizações, que podem ser analisados como unidades individuais ou sociais coletivas [Wasserman and Faust 1999]. O método de ARS costuma-se ser utilizado para estudar as relações da colaboração científica entre pesquisadores.

Para cumprir com o propósito do primeiro grupo de atividades do método proposto, utilizou-se a metodologia NeOn [Suárez-Figueroa et al. 2012], que abrange diferentes cenários de criação e reutilização de modelos ontológicos. Os documentos utilizados para dar apoio à criação da nova ontologia foram artigos científicos coletados do Google Acadêmico. A Figura 3 explicita a *SCNAS Ontology*[3], criada através das atividades propostas pela NeOn. A ontologia descreve conceitos como o Escopo da rede estudada (1); Métricas de ator, grupo e rede (2); Período temporal dos dados estudados (3); Localização geográfica da rede estudada (4).

[3]`https://github.com/Laltany/SCNAS-Ontology`

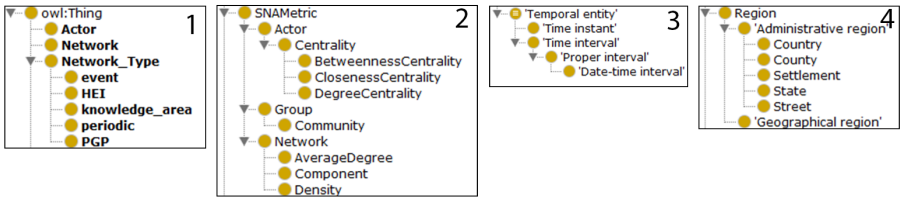


**Figura 3. Atual arquitetura da ontologia de extração**

No grupo 2, para a criação da arquitetura das NPs observou-se a necessidade de estender a arquitetura miníma com a criação de um quarto nível. Esse nível tem como objetivo descrever conceitos referentes a proveniência dos dados estudados nas análises, como o período temporal dos dados e a identificação do repositório que proveu os dados estudados, permitindo que os pesquisadores possam identificar a proveniência dos dados apresentados nos artigos do domínio.

A Figura 4 representa a arquitetura adotada para as NPs deste trabalho juntamente com as ontologias utilizadas em cada nível da arquitetura. Na camada de asserção usou-se a *SCNAS ontology*, anteriormente criada. No nível de proveniência das asserções optou-se por fazer uso da ontologia PROV-O[4] juntamente com a ontologia DataCite[5], possibilitando a identificação do título dos estudos, autores dos artigos e o DOI da publicação. Através da *DCMI Metadata* Terms[6] foi obtido o termo de anotação *publisher*. Para a proveniência da NP também utilizou-se a ontologia PROV-O, permitindo a identificação do autor e data de criação da NP. Por fim, para o nível de proveniência dos dados, foi utilizada a ontologia *Time* (contida na SCNAS ontology) e a classe *location* (contida na ontologia DataCite).

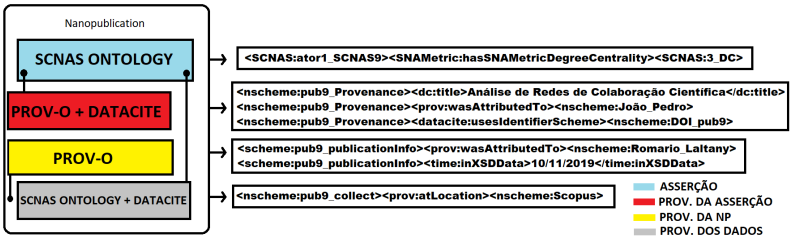


**Figura 4. Extensão da arquitetura miníma da NP.**

Ainda no grupo 2, após a definição da arquitetura com o seu conjunto de ontologias, foi iniciada a extração das afirmações a serem mantidas na estrutura de cada NP. Nessa atividade, utilizou-se como insumo os documentos coletados no grupo 1. Optou-se pela extração manual de uma quantidade relativamente pequena de dados, a fim de provar o conceito. Todas as afirmações foram registradas em planilhas e estruturadas automaticamente, com o uso da ferramenta Protégé.

Com os dados já estruturados, foi necessário armazená-los em um banco de triplas RDF. Neste estudo foi utilizado o Jena Fuseki que, através de uma interface gráfica,

[4] https://www.w3.org/TR/prov-o/

[5] https://sparontologies.github.io/datacite/current/datacite.html

[6] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/\#terms-publisher

permite a criação, monitoramento e administração do banco de triplas. Também permite a realização de consultas no banco através da linguagem SPARQL.

No último grupo de atividades foi implementada uma aplicação semântica capaz de recuperar as NPs contidas no banco de triplas. Nela é possível escolher o tipo de rede analisada nos estudos (área de conhecimento, periódico, evento, programa de pós-graduação e instituição de ensino ou pesquisa) e as métricas de análise de rede. Com isso, conseguiu-se responder a questões como:"Qual a Centralidade de Grau e de Proximidade dos atores de uma rede de colaboração científica de PPG de determinada área de conhecimento?", como apresentado na Figura 5.

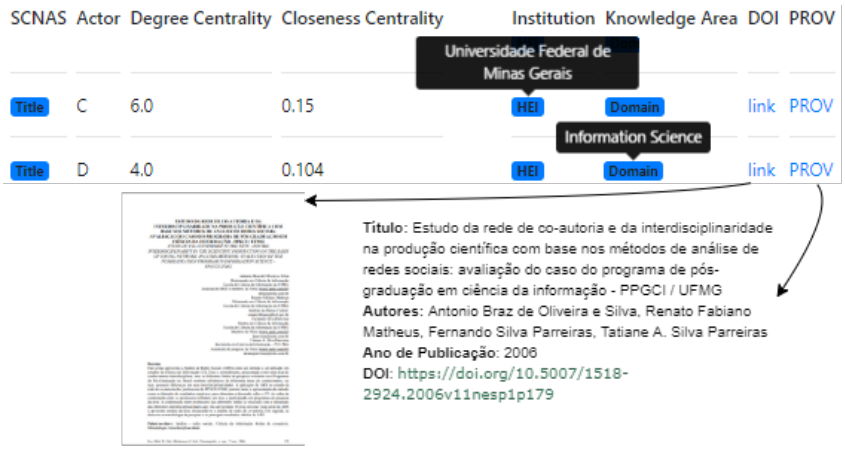


**Figura 5. Exemplo de consulta realizada na aplicação**

Uma especialista no domínio de ARCC realizou a validação das NPs recuperadas, através de consultas na aplicação que tiveram como base as questões de competência presentes na documentação da *SCNAS Ontology*[7], onde cada consulta deveria responder uma determinada questão e a resposta obtida deveria ser comparada com a resposta esperada. Todas as questões foram respondidas com sucesso e suas proveniências estavam corretas.

## 5.1. Considerações Finais

Neste artigo, apresentou-se um método para a criação e recuperação de NPs. Partindo de um cenário onde não há artefatos que especifiquem como é realizada a criação e recuperação de NPs, o método proposto pode auxiliar na implementação de NPs em diferentes domínios de conhecimento.

Com a aplicação do método ao domínio de ARCC, obteve-se um artefato ontológico que pode ser utilizado por pesquisadores deste domínio para diversos fins. A utilização de NPs do referido domínio trouxe melhorias na busca de resultados relacionados às métricas e outras informações presentes nas publicações. Com isso, comprovou-se que o uso da abordagem pode proporcionar melhorias para outros domínios além das Ciências da Vida, desde que a sua arquitetura seja adaptada e composta por ontologias que sejam capazes de descrever as informações do domínio ao qual deseja-se criar NPs, conforme consta no método.

Observa-se que o processo de extração das informações das publicações é um dos principais gargalos da abordagem, como afirmam [Do and Mobley 2015]. Visto isso,

[7] http://tiny.cc/a73ggz

assume-se como trabalho futuro a implementação de um extrator de dados para a criação de NPs do domínio de ARCC. A fim de evidenciar a real colaboração que a NP traz ao meio científico, pretende-se fazer uma comparação da abordagem com outras abordagens de recuperação de informação, apontando as vantagens e desvantagens de cada uma.

## Referências


Beck, T., Free, R. C., Thorisson, G. A., and Brookes, A. J. (2012). Semantically enabling a genome-wide association study database. *Journal of biomedical semantics*, 3(1):9.

Do, L. and Mobley, W. (2015). Single figure publications: Towards a novel alternative format for scholarly communication. *F1000Research*, 4:268–268.

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

Groth, P., Schultes, E., Thompson, M., Tatum, Z., and Dumontier, M. (2013). Nanopublication guidelines. `<http://www.nanopub.org/2013/WD-guidelines-20131215/>`.

Kuhn, T., Barbano, P. E., Nagy, M. L., and Krauthammer, M. (2013). Broadening the scope of nanopublications. In *Extended Semantic Web Conference*, pages 487–501. Springer.

Kuhn, T., Chichester, C., Krauthammer, M., Queralt-Rosinach, N., Verborgh, R., Giannakopoulos, G., Ngomo, A.-C. N., Viglianti, R., and Dumontier, M. (2016). Decentralized provenance-aware publishing with nanopublications. *PeerJ Computer Science*, 2:e78.

Lipani, A., Piroi, F., Andersson, L., and Hanbury, A. (2014). Extracting nanopublications from ir papers. In *Information Retrieval Facility Conference*, pages 53–62. Springer.

Maiatsky, M., Boyarsky, A., Boyarskaya, N., Velmezova, E., and Piotrowski, M. (2018). Vicoglossia: Annotatable and commentable library as a bridge between reader and scholar (a proof of concept study: Early soviet philological culture). *Umanistica Digitale*, (2):161–184.

Mons, B., van Haagen, H., Chichester, C., Hoen, P.-B. t., den Dunnen, J. T., van Ommen, G., van Mulligen, E., Singh, B., Hooft, R., Roos, M., Hammond, J., Kiesel, B., Giardine, B., Velterop, J., Groth, P., and Schultes, E. (2011). The value of data. *Nature Genetics*, 43(4):281–283.

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.

Oliveira, É. B. P. M. (2008). Periódicos científicos eletrônicos: definições e histórico. *Informação & Sociedade*, 18(2):69–77.

Suárez-Figueroa, M. C., Gómez-Pérez, A., and Fernández-López, M. (2012). The neon methodology for ontology engineering. In *Ontology engineering in a networked world*, pages 9–34. Springer.

Wasserman, S. and Faust, K. (1999). *Social network analysis—methods and applications, revised, reprinted edn*. Cambridge University Press, Cambridge.