Universidade Federal de Juiz de Fora

Programa de Pós-graduação em Modelagem Computacional

Edson Bruno Novais

e-ScienceNet: uma Rede Ponto a Ponto Semântica para aplicações em e-Science.

Juiz de Fora

2012

e-ScienceNet: uma Rede Ponto a Ponto Semântica para aplicações em e-Science.

Edson Bruno Novais

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Modelagem Computacional da Universidade Federal de Juiz de Fora, como parte dos requisitos necessários à obtenção do grau de Mestre em Ciências em Modelagem Computacional.

Orientadora: Regina Maria Maciel Braga Villela

Juiz de Fora

Agosto de 2012

Edson Bruno Novais

e-ScienceNet: uma Rede Ponto a Ponto Semântica para aplicações em e-Science.

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Modelagem Computacional da Universidade Federal de Juiz de Fora, como parte dos requisitos necessários à obtenção do grau de Mestre em Ciências em Modelagem Computacional.

Aprovada em 10 de Agosto de 2012.

BANCA EXAMINADORA

_____

Profa. Dra. Regina Maria Maciel Braga Villela - Orientadora

Universidade Federal de Juiz de Fora

_____

Prof. Dr. Alcione de Paiva Oliveira

Universidade Federal de Viçosa

_____

Profa. Dra. Fernanda Cláudia Alves Campos

Universidade Federal de Juiz de Fora

Dedico à família e amigos.

AGRADECIMENTOS

Aos professores e colaboradores do Programa de Pós-graduação em Modelagem Computacional; As agências de fomento CAPES e CNPq pelo financiamento do trabalho; A Universidade Federal de Juiz de Fora e a todos que, direta ou indiretamente, contribuíram para a realização deste trabalho.

"Scientia potentia est, sed parva; quia scientia egregia rara est, nec proinde apparens nisi paucissimis, et in paucis rebus. Scientiae enim ea natura est, ut esse intelligi non possit, nisi ab illis qui sunt scientia praediti." - Thomas Hobbes

RESUMO

Atualmente estão acontecendo significativas mudanças na natureza do processo de pesquisa científica. Em particular, tem-se uma maior colaboração entre grandes grupos de pesquisadores, o que leva a um aumento no uso de técnicas de processamento de informação, e uma maior necessidade de compartilhar resultados e observações entre os participantes do processo. Utilizando tecnologias como Redes Ponto a Ponto e Web Semântica, conseguimos criar um único ponto de acesso a bases de conhecimento dispersas e aplicações científicas distribuídas, onde os cientistas possam trabalhar com informações heterogêneas e criar comunidades científicas de acordo com suas especialidades e interesses. Este trabalho apresenta a e-ScienceNet, uma arquitetura de suporte ao armazenamento, compartilhamento e execução de experimentos científicos em uma Rede Ponto a Ponto Semântica científica. Para tal, desenvolvemos um protótipo que realiza a interação entre os diferentes nós da rede nas distintas comunidades científicas semânticas.

Palavras-chave: e-Science, Redes Ponto a Ponto, Sistemas Distribuídos e Web Semântica.

ABSTRACT

Currently, there are taking place significant changes in the nature of scientific research. In particular, there is a greater collaboration among large groups of researchers, which leads to an increase in the use of information processing techniques, and a greater need to share results and observations among the participants in the process. Using technologies such as Peer to Peer Networks and the Semantic Web, we can create a single point of access to distributed knowledge bases and scientific applications, where scientists can work with heterogeneous information and create scientific communities according to their specialties and interests. This work presents the e-ScienceNet as an architecture that support the storage, share and execution of scientific experiments in a scientific Semantic Peer to Peer Network. To this end, we developed a prototype that performs the interaction between the different network nodes in different semantic scientific communities.

Keywords: e-Science, Distributed computing, Peer-to-Peer Networks and Semantic Web.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

1 INTRODUÇÃO

Atualmente estão acontecendo significativas mudanças na natureza do processo de pesquisa científica. Em particular, tem-se uma maior colaboração entre grandes grupos de pesquisadores, o que leva a um aumento no uso de técnicas de processamento de informação, e uma maior necessidade de compartilhar resultados e observações entre os participantes do processo [1].

Utilizando tecnologias como Redes Ponto a Ponto [2] e Web Semântica [3], conseguimos criar um único ponto de acesso a bases de conhecimento dispersas, onde os cientistas possam trabalhar com informações heterogêneas e criar comunidades científicas de acordo com suas especialidades e interesses.

1.1 Motivação

A partir das últimas décadas do século XX, com o progresso contínuo da informática, o processo de concepção, execução e análise de experimentos científicos conduzidos por pesquisadores de diversas áreas tem buscado o apoio de ferramentas computacionais de forma crescente. Percebe-se também que o emprego de tais recursos está relacionado ao aumento na complexidade dos experimentos desenvolvidos que, em alguns casos, podem tornar-se inviáveis sem algum tipo de suporte computacional adequado [4]. Entretanto, cabe ponderar que os benefícios oferecidos por essas ferramentas computacionais implicam em novos e complexos desafios no cenário da pesquisa científica [5].

Com os recentes avanços, a ciência moderna assiste a um crescimento exponencial em complexidade e escopo. A necessidade de maior colaboração entre os cientistas em diferentes instituições, considerando diferentes especialidades, e através de diferentes disciplinas científicas faz parte deste cenário de crescimento [6].

Considerando esta colaboração entre cientistas, o volume de dados resultante das pesquisas tende a crescer cada vez mais. Neste cenário, as Redes Ponto a Ponto se tornam um bom recurso a ser utilizado para armazenamento, processamento e distribuição desses dados. Dado o grande volume de dados, a tarefa de pesquisar informações relevantes se torna difícil e custosa, alocando tempo em tarefas que podem ser automatizadas computacionalmente. Outro fator importante é não suprir localmente, na maioria das vezes, a necessidade requerida de processamento por parte de simulações científicas. Neste contexto, as Redes Ponto a Ponto se tornam relevantes, pois permitem a distribuição de processamento entre os seus usuários.

Por fim, é válido mencionar que o tema desenvolvido no presente trabalho encontra respaldo a partir das diretrizes formuladas no documento Grandes Desafios [7] elaborado sob os auspícios da Sociedade Brasileira de Computação (SBC), particularmente em dois dos cinco tópicos propostos: (i) gestão da informação em grandes volumes de dados distribuídos; (ii) Modelagem Computacional de sistemas complexos artificiais, naturais e sócio-culturais e da interação homem-natureza.

1.2 Justificativa

O uso de recursos computacionais para a realização de pesquisas científicas não é trivial, pois, em geral os cientistas de outras áreas que não a área de computação, têm dificuldade para lidar com as tecnologias computacionais disponíveis em baixo nível de abstração [8]. Além da dificuldade de utilização das ferramentas computacionais por cientistas, existe também a dificuldade para se encontrar artefatos de software que executem determinada tarefa necessária ao experimento que está sendo realizado [8].

Neste contexto, e-Science [9] pode ser utilizada para caracterizar o importante papel das tecnologias computacionais na pesquisa, colaboração, compartilhamento de dados e documentos, e uso de recursos para automatizar a execução e análise de dados de experimentos científicos [10].

Considerando este cenário, tecnologias como Workflows Científicos [11], Ontologias [12] e Serviços Web Semânticos [1] podem ser utilizadas para a composição de uma

infraestrutura de apoio a e-Science, sendo todas essas tecnologias possíveis de trabalharem em um ambiente distribuído de uma Rede Ponto a Ponto.

A maioria das pesquisas recentes endereçando interações na Web Semântica [13, 14, 15, 16, 17, 18] estão na maioria das vezes fragmentadas e abordam uma visão global do problema. Além disso, a maioria das Redes Ponto a Ponto desenvolvidas atualmente são utilizadas somente para o compartilhamento de aplicativos, músicas e vídeos, desperdiçando um potencial de trabalho em Sistemas Distribuídos [19].

Este trabalho propõe e detalha como é realizada a junção entre as tecnologias de Redes Ponto a Ponto e Web Semântica, criando uma Rede Ponto a Ponto Semântica. Dessa forma, é importante ressaltar a utilização dos conceitos da Web Semântica para criar uma descrição semântica dos dados disponíveis na Rede Ponto a Ponto utilizando Ontologias, proporcionando uma descrição semanticamente rica dos dados, facilitando o acesso dos cientistas a informações públicas disponíveis por outros cientistas e correlacionadas aos seus interesses e especialidades.

1.3 Objetivos

Esta dissertação tem como objetivo principal projetar e desenvolver uma infraestrutura de apoio a e-Science em uma Rede Ponto a Ponto Semântica científica. A infraestrutura de apoio a e-Science será responsável por disponibilizar dados científicos em uma rede de colaboração científica, facilitando o acesso de cientistas a bases de dados geograficamente dispersas através de um único ponto de acesso.

Alguns requisitos para a criação de uma infraestrutura de apoio a e-Science são detalhados no capitulo de pressupostos teóricos. Desses requisitos, podemos selecionar como objetivo da primeira versão da infraestrutura desenvolvida: (i) armazenamento e compartilhamento de dados, onde o cientista deve ser capaz de armazenar grandes volumes de dados de forma distribuída; (ii) os dados devem ser compartilhados de acordo com configurações dos cientistas, indicando políticas de acesso as comunidades científicas; (iii) os dados compartilhados devem ser disponibilizados de forma transparente; (iv) notificações devem ser enviadas na rede através de eventos; (v) a infraestrutura deve suportar a expansão

da quantidade de serviços e usuários na rede sem causar desordem ou perda de desempenho; e (vi) se basear em componentes de software que se adequem as novas necessidades dos cientistas.

1.4 Metodologia

A metodologia de pesquisa incluiu a revisão bibliográfica, a proposta de uma arquitetura de apoio a e-Science com a utilização de Redes Ponto a Ponto e Web Semântica no contexto de e-Science e a sua avaliação através de um cenário de uso.

Na revisão bibliográfica, foram apresentados os conceitos que permitem a criação de uma infraestrutura para apoio a e-Science através da junção das Redes Ponto a Ponto e Web Semântica. Após apresentação dos conceitos, foi proposto a e-ScienceNet e um cenário de uso que mostra a sua utilização através de uma implementação.

1.5 Estrutura da Dissertação

Este trabalho está organizado em cinco capítulos. O primeiro é o capítulo de introdução, ao qual apresentamos a motivação para realização deste trabalho, os objetivos gerais que pretendemos alcançar juntamente com as justificativas, além da estrutura da dissertação apresentada nesta seção.

No capítulo dois são apresentados os principais conceitos e tecnologias utilizados durante o desenvolvimento deste trabalho, sendo e-Science, Redes Ponto a Ponto e Web Semântica os conceitos mais relevantes.

O capítulo três detalha o planejamento e desenvolvimento da arquitetura para uma infraestrutura de apoio a e-Science, fazendo a junção dos diversos conceitos apresentados no capítulo dois.

No capítulo quatro, é apresentado um protótipo da arquitetura, com alguns exemplos de utilização considerando usuários (cientistas) distribuídos na rede.

Para finalizar, no capítulo cinco apresentamos as considerações finais do trabalho.

2 PRESSUPOSTOS TEÓRICOS

Este capítulo introduz os principais conceitos utilizados ao longo deste trabalho. Na primeira seção é apresentado o conceito de e-Science, área a qual este trabalho está fortemente relacionado. Na segunda seção tem-se o conceito de Redes Ponto a Ponto, arquitetura de rede que foi adaptada neste trabalho para suporte a e-Science. A terceira seção detalha conceitos relacionados a Web Semântica, aparato tecnológico que auxiliou na adaptação de uma Rede Ponto a Ponto para apoio a e-Science. Na seção quatro detalha-se a relação entre os conceitos apresentados nas seções anteriores e finalmente na seção cinco os trabalhos relacionados são apresentados.

2.1 e-Science

Podemos chamar de e-Science conceitos relacionados à integração da computação nas pesquisas científicas em diversas áreas, sendo este apoio computacional de grande importância para aumentar a eficiência das pesquisas [8]. Atualmente os sistemas computacionais têm se tornado importantes para a pesquisa científica, suportando todos os aspectos relacionados ao seu ciclo de vida. A comunidade científica tem utilizado os termos e-Science e e-Research para englobar o importante papel das tecnologias computacionais na pesquisa, colaboração, compartilhamento de dados e documentos, e uso de recursos para automatizar a execução e análise de dados de experimentos científicos [10].

O termo e-Science foi introduzido no Reino Unido por Taylor, encapsulando as tecnologias necessárias ao suporte à pesquisa colaborativa e multidisciplinar que emergiu em vários campos da ciência [9]. Taylor reconheceu a importância do uso de ferramentas computacionais na pesquisa científica colaborativa, multidisciplinar e com grande volume de dados, e usou o termo e-Science para englobar as ferramentas e tecnologias necessárias ao suporte a esse tipo de pesquisa [10].

O termo e-Science pode ser utilizado, portanto, para descrever o desenvolvimento de uma infraestruturas de serviços de software capazes de prover acesso a facilidades remotas, recursos computacionais distribuídos, armazenamento de informações em bancos de dados dedicados, disseminação e compartilhamento de dados, resultados e conhecimento [11].

Esta utilização de recursos computacionais no desenvolvimento da pesquisa beneficia o trabalho das comunidades científicas, facilitando o compartilhamento de dados e serviços computacionais, além de contribuir para a construção de uma infraestrutura de dados e de uma comunidade científica distribuída [12]. Na literatura existem diversos trabalhos se aprofundando em e-Science, como, [14, 20, 21, 22].

O poder computacional vem sendo largamente explorado por cientistas, como, por exemplo, para a criação de simulações sofisticadas relacionadas ao clima e terremotos, para a extração de resultados de dados científicos oriundos da astronomia ou física, e para a criação de simulações e análise de dados relacionados à biologia [23]. No entanto, aplicações como estas ainda carecem de uma infraestrutura que facilite a utilização de todo um aparato computacional de suporte. Muitas vezes os cientistas que utilizam este tipo de aplicação não conseguem usufruir de maneira eficiente do poder computacional a eles disponibilizado. Assim, a área de e-Science está em busca de aprimorar e facilitar o uso deste aparato computacional, através de uma infraestrutura que permita projetar, reusar, anotar, validar, compartilhar e documentar artefatos gerados pela pesquisa científica [11]. Além disso, toda informação deve ser gerenciada de maneira eficiente por diversos processos, como, por exemplo, armazenamento, recuperação e integração [20].

Considerando o contexto deste trabalho, uma infraestrutura globalmente distribuída para e-Science utilizando tecnologias como Redes Ponto a Ponto e Web Semântica levanta diversos requisitos que devem ser considerados no seu planejamento. Abaixo seguem alguns destes requisitos relevantes [20, 21].

- Armazenamento: o cientista deve ser capaz de armazenar e processar grandes volumes de dados de forma eficiente, independente da sua localização geográfica;
- Gerenciamento de propriedade: o cientista deve ser capaz de manter propriedade sobre os seus dados e serviços, disponibilizando seus recursos somente depois que outros cientistas aceitarem os termos de uso do recurso solicitado;

- Transparência: o cientista deve ser capaz de descobrir, acessar e processar dados de forma transparente, independente de onde estes dados estejam localizados;
- Comunidades: o cientista deve ser capaz de criar, manter e desfazer comunidades, sejam elas restritas ou não. Para tal, é necessário criar regras indicando as permissões dos cientistas aos recursos providos pela comunidade;
- Segurança: o cientista deve ser capaz de compartilhar seus dados de forma segura. A utilização de técnicas de criptografia e autenticação é crucial para manter a privacidade dos dados transmitidos;
- Mobilidade: o cientista deve ser capaz de acessar os dados disponíveis de qualquer dispositivo, incluindo computadores pessoais e dispositivos móveis;
- Fluxo de trabalho: suportar o processo de automação, descrevendo os processos científicos de forma clara e computacionalmente processável, criando assim um fluxo de trabalho totalmente automatizado;
- Proveniência: informações suficientes devem ser armazenadas durante a execução, provendo evidências da concisão dos dados gerados, possibilitando a reutilização dos resultados obtidos e viabilizando a reprodução dos experimentos científicos;
- Notificações: os cientistas devem receber notificações de novos dados e serviços disponibilizados de acordo com seus interesses;
- Suporte a decisão: prover informações e sugestões relevantes para os cientistas de acordo com suas necessidades;
- Expansão: suportar o crescimento da infraestrutura. A quantidade de cientistas, dados e serviços não deve afetar ou limitar o seu uso;
- Componentes: basear o desenvolvimento em Componentes de Software, se adaptando as novas necessidades dos cientistas com o tempo.

Levando em consideração os requisitos apresentados para uma infraestrutura globalmente distribuída para e-Science, planejou-se a e-ScienceNet de forma que cada requisito apresentado seja tratado da forma mais adequada possível.

2.2 Redes Ponto a Ponto

Uma Rede Ponto a Ponto é uma arquitetura de rede completamente descentralizada que utiliza recursos distribuídos de vários computadores. Chamados “nós de rede”, os computadores distribuídos trabalham para realização de uma tarefa em comum, geralmente o compartilhamento de arquivos na internet [19].

As Redes Ponto a Ponto são em sua natureza um Sistema Distribuído, sem nenhuma organização hierárquica ou controle centralizado sobre os dados, onde um nó na rede pode atuar tanto como cliente ou como servidor [2]. A filosofia de projeto de sistemas de Rede Ponto a Ponto, similar a apresentada pela e-Science, é prover aos seus usuários flexibilidade de cooperação com os diversos usuários na rede [24].

Nos últimos anos ocorreu um aumento substancial no número de computadores pessoais. Cada novo computador pessoal disponível na internet, é um possível cliente ou servidor em uma Rede Ponto a Ponto. Devido a esse crescimento, a tecnologia de Redes Ponto a Ponto vem recebendo grande atenção da comunidade científica [25], que aproveita o poder de processamento e armazenamento distribuído em seus experimentos científicos.

Fatores que influenciam no ganho de popularidade das Redes Ponto a Ponto são a sua escalabilidade, tolerância a falhas, auto-organização e privacidade, propondo assim um paradigma para criação de aplicativos para recuperação de dados a um custo baixo [26], o qual é de grande importância para pesquisas científicas. Apesar das características de uma Rede Ponto a Ponto serem um auxilio importante para a comunidade científica, atualmente elas são utilizadas mais para a troca de arquivos na internet, especialmente arquivos de música e vídeos [27].

Considerando as vantagens de Sistemas Distribuídos em relação a Sistemas Centralizados, é importante ressaltar que os Sistemas Centralizados são simples de programar e gerenciar, entretanto são gargalos em potencial, uma vez que o servidor central tem capacidade limitada e pode não suportar o aumento da demanda. Por outro lado, os sistemas descentralizados são escaláveis e robustos, mas isso demanda certa complexidade de implementação, principalmente nas questões de tolerância a falhas e descoberta de novos

recursos. Muitos Sistemas Distribuídos combinam características das duas arquiteturas, parte do sistema utilizando o modelo tradicional cliente/servidor e outra parte Redes Ponto a Ponto.

A topologia de uma Rede Ponto a Ponto descreve como é o esquema das conexões na rede e como as informações são trafegadas entre os seus nós. Uma Rede Ponto a Ponto se divide basicamente entre as redes estruturadas[1] e não estruturadas[2]. Em [2] existe uma discussão extensiva sobre esses dois tipos de Redes Ponto a Ponto. Outra abordagem atualmente sendo estudada é a construção de camadas não estruturadas sobre uma camada estruturada. Chamadas de híbridas, esse tipo de topologia consegue aumentar a flexibilidade das conexões e a eficiência do acesso à rede [28].

Como exemplo de redes estruturadas, temos CAN [29], Tapestry [30], Chord [31], Pastry [32], Kademlia [33] e Viceroy [34]. Como não estruturada temos o Gnutella [35].

Estruturas híbridas são implementadas notavelmente em Sistemas Distribuídos Colaborativos. A principal preocupação em muitos desses sistemas é como se conectar, para o qual muitas vezes um esquema tradicional cliente/servidor é adotado. Uma vez que o cliente se conecte ao sistema, é utilizado um esquema totalmente descentralizado para colaboração. Um exemplo dessa abordagem pode ser vista em clientes BitTorrent [32].

Apesar do drástico crescimento das Redes Ponto a Ponto nos últimos anos, as buscas por dados nas redes atuais são muito ineficientes e não escalam muito bem. A ineficiência aparece, grande parte, devido à criação de Redes Sobrepostas aleatórias, conceito detalhado nos parágrafos seguintes, onde as buscas são repassadas de um nó para outro de forma cega [32].

Na maioria das Redes Ponto a Ponto atuais, a propagação de requisições para os outros nós da rede é feita de forma aleatória. Ao enviar buscas de dados de forma aleatória para os outros nós, temos uma diminuição na eficiência da comunicação e uma perda na acurácia das pesquisas.

Criar grupos com os nós em uma Rede Ponto a Ponto é uma tarefa conhecida como Redes Sobrepostas, ou Redes Overlay. Esses grupos tem conhecimento um dos outros, formando assim um grafo com os diversos grupos e seus respectivos nós.

[1] Uma rede estruturada garante através de um protocolo que um nó possa rotear sua busca para outro nó na rede sem que exista muita comunicação.
[2] Uma rede não estruturada as conexões são estabelecidas de maneira aleatória, dessa forma, quando é realizada uma busca, temos uma maior comunicação e uma menor precisão.

Se utilizarmos os domínios de pesquisas e interesses de cada um dos cientistas conectados na rede, podemos evitar que as informações sejam propagadas pela rede de forma aleatória, sendo repassadas as requisições somente aos nós que tenham similaridade semântica com os domínios de pesquisas e interesse daquele pesquisador.

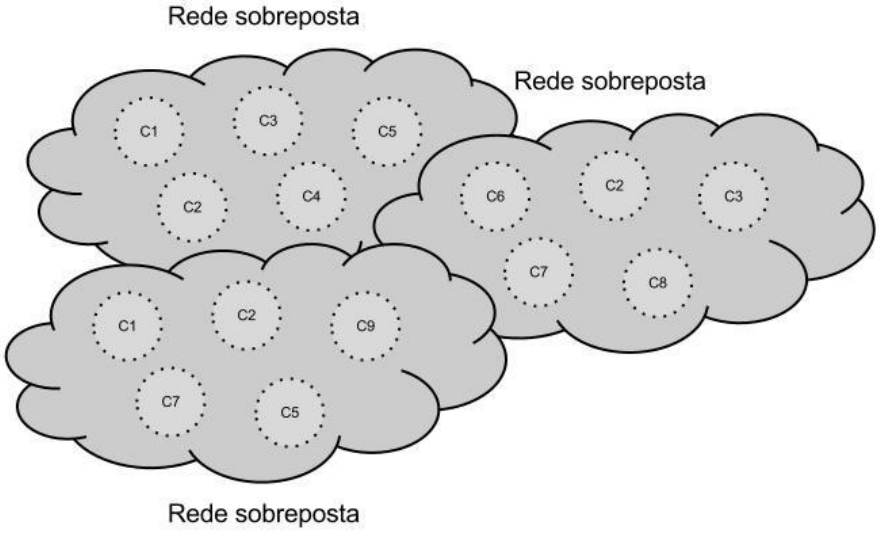

Figura 2.1: Redes Sobrepostas

Como visto na Figura 2.1, um mesmo nó (C1 … C9) pode estar contido em quantas redes sobrepostas forem necessárias, limitando-se pelo seu poder computacional e interesses na rede.

Em [36] é proposto um esquema de Redes Sobrepostas que criam seus grupos de nós de acordo com a similaridade semântica dos seus conteúdos disponíveis. Também temos em [37] um modelo que propõe a criação de Redes Sobrepostas de acordo com a especialidade do nó. Com o uso de Rede Sobrepostas as buscas são roteadas para o grupo apropriado, aumentando a chances que a pesquisa tenha sucesso com mais rapidez, diminuindo assim o processamento de nós que não tem conteúdo similar.

Com dados reais de usuários, os testes realizados em [36] mostram que o uso de Redes Sobrepostas semânticas utilizou somente 10%-20% do poder computacional que redes aleatórias utilizariam na troca de mensagens. Com 92 mensagens, o sistema com Redes Sobrepostas semânticas proposto em [36] foi capaz de achar 20% dos resultados, enquanto

que um sistema similar que usa Redes Sobrepostas sem semântica precisou de 285 mensagens para chegar à mesma quantidade de resultados. Exemplos dessa nova classe de Redes Ponto a Ponto podem ser visualizados em [38, 39, 40, 41, 42].

Outro conceito importante são as Tabelas de Espalhamento Distribuídas (DHT), que são considerados como uma classe de Sistemas Distribuídos descentralizados, provendo um serviço de pesquisa onde qualquer nó participante pode eficientemente recuperar o valor associado a uma dada chave na tabela. A responsabilidade de manter o mapeamento de chaves para valores é distribuída entre os nós, de forma que mudanças no conjunto de participantes causem um mínimo de desordem. Isso faz com que as DHTs escalem a um número extremamente grande de nós e gerenciem chegadas, saídas e falhas contínuas [43].

2.3 Web Semântica

A Web Semântica é definida como uma extensão da web existente, onde a informação disponibilizada tem um significado bem definido [3]. A maioria da informação na web atual é destinada para humanos. Os princípios estruturais são fracos, diferentes tipos de informações coexistem e a maioria dessas informações é representada como texto puro [24].

Uma das tecnologias mais importantes que possibilitam a Web Semântica são as Ontologias [3, 12, 44]. São as Ontologias que fazem o papel principal, provendo semântica das informações, podendo então ser processadas por máquinas entre sistemas heterogêneos [1].

Segundo [45], o termo Ontologia tem origem no grego ontos, que significa ser e logos, ou seja, palavra. A origem é a palavra aristotélica “categoria”, que pode ser usada para classificar e caracterizar alguma coisa [8].

Em Ciência da Computação, uma Ontologia é uma conceitualização compartilhada que define uma especificação formal e explícita dos termos de um domínio e das relações entre eles [46, 47, 48]. Uma Ontologia consiste tipicamente de uma descrição hierárquica de importantes conceitos em um domínio, juntamente com as descrições e propriedades de cada um desses conceitos [1].

As Ontologias foram primeiramente desenvolvidas pela comunidade de Inteligência Artificial para facilitar o compartilhamento e reuso da informação [1]. Alguns benefícios de sua utilização são o reuso, o compartilhamento e portabilidade de conhecimento, a manutenibilidade, a documentação e a confiabilidade [49].

O estudo e o uso de Ontologias na área de software ganharam novo impulso com a Web Semântica introduzida por [3]. Neste contexto, uma Ontologia consiste de uma taxonomia e um conjunto de regras de inferência, que permitem capturar o conhecimento que não está explícito na taxonomia [46, 47, 48].

A captura de conhecimento não explícito é dada através da utilização de máquinas de inferência. Uma máquina de inferência é um pedaço de software capaz de inferir consequências lógicas através de fatos e axiomas adicionados na ontologia. É importante ressaltar que uma máquina de inferência não realiza alterações na ontologia. A máquina de inferência utiliza uma ontologia previamente criada e como saída gera outra ontologia com novos conhecimentos inferidos.

O desenvolvimento de Ontologias é geralmente um processo cooperativo que envolve diferentes entidades em diferentes locais [1]. O objetivo é construir uma Ontologia compartilhada com o conhecimento comum de um domínio entre as pessoas, organizações e sistemas computacionais. Nos dias atuais, as Ontologias estão cada vez mais utilizadas para possibilitar o acesso e processamentos de dados de forma semântica [1].

Considerando o contexto da Web Semântica, a linguagem Ontology Web Language (OWL) [50] é uma linguagem utilizada para implementação de Ontologias. OWL é uma revisão da linguagem DAML+OIL[3] incorporando lições aprendidas durante o design e aplicações da DAML+OIL. OWL têm três versões que representam a sua expressividade semântica [1], que são: OWL Lite, OWL DL e OWL Full. Uma lista exaustiva de linguagens de Ontologias é apresentada em [51].

A linguagem OWL foi projetada para ser utilizada por aplicações que necessitam processar o conteúdo da informação em vez de apenas apresentar informações para os seres humanos [8]. A OWL visa atender às necessidades de uma linguagem de Ontologia para a web, superando algumas limitações das linguagens que a precederam, como, por exemplo, o RDF.

[3] DAML+OIL é uma linguagem de marcação semântica para recursos web.

O objetivo da OWL é prover uma linguagem de Ontologia que possa ser usada para descrever, de um modo natural, classes e relacionamentos entre classes em documentos e aplicações web [8].

Como é baseada em lógica descritiva, OWL possibilita o uso de mecanismos de inferência, os quais permitem explicitar conhecimentos que estão implícitos em uma base de conhecimento. Dessa forma, um documento OWL não deve ser considerado apenas sob o ponto de vista de sua sintaxe, mas também de sua semântica [8].

O editor Protégé-OWL [52] permite a construção de Ontologias para a Web Semântica, em especial utilizando a linguagem OWL especificada pelo W3C [8].

Quando utilizado no contexto de Redes Ponto a Ponto, as Ontologias dos diversos nós necessitam de ser mapeadas entre si. O Mapeamento de Ontologias consiste em realizar o mapeamento entre duas ou mais Ontologias utilizando regras para definir a similaridade entre seus conceitos e propriedades. Alguns autores utilizam o termo alinhamento entre Ontologias para caracterizar o processo de pesquisa de similaridade dos conceitos e propriedades entre duas ou mais Ontologias distintas.

Um alinhamento calcula as relações semânticas entre as duas Ontologias de entrada utilizando as entidades das Ontologias e um conhecimento K relevante ao processo de mapeamento [25].

O problema de mapear duas ou mais Ontologias de forma efetiva aparece em diversas aplicações [53, 54, 55]. Estudos realizados em [56] mostram que o mapeamento manual entre Ontologias é muito mais preciso que os mapeamentos automáticos realizados por softwares. Em contra partida, são muito mais caros que os automáticos [56].

O mapeamento manual consiste de um modelo metodológico para identificar mapeamentos que possam ser feitos de forma semiautomática, onde o sistema propõe ou critica mapeamentos existentes e o usuário prove retorno do método utilizado. Mapeamentos automáticos que tentam achar mapeamentos entre Ontologias sem a intervenção do usuário pagam o preço de um possível mapeamento incorreto [24]. Os métodos mais conhecidos para mapeamento de Ontologias são Naive Ontology Mapping [57] e Quick Ontology Mapping [58].

O mapeamento entre as Ontologias na e-ScienceNet é um requisito necessário para estabelecer a interoperabilidade entre as diferentes Ontologias utilizadas pelos diferentes cientistas da e-ScienceNet. Atualmente existem alguns métodos propondo o mapeamento de Ontologias em Redes Ponto a Ponto [59, 60, 61].

Entretanto, em Redes Ponto a Ponto frequentemente encontramos situações onde grandes estruturas ontológicas devem ser mapeadas em apenas alguns segundos ou menos, para que possam ser utilizadas [53]. Quando os métodos de mapeamento atuais foram aplicados em alguns casos de Redes Ponto a Ponto, como, Bibster [62] e Xarop [63], os resultados mostram que nenhum dos métodos foi aceitável para a tarefa aplicada em Redes Ponto a Ponto, já que todos eles negligenciaram eficiência [53].

Em uma Rede Ponto a Ponto, quanto maior o número de nós conectados, a eficiência dos mapeamentos decresce de forma exponencial [54]. Esse é um problema recorrente na maioria dos algoritmos de mapeamento atuais. Isso se deve ao fato de que, quanto maior o número de Ontologias distribuídas, mais demorado será o mapeamento entre cada um dos seus conceitos e propriedades [54]. Existem métodos como o mapeamento rápido de Ontologias [53] que decresce a qualidade geral do mapeamento em Ontologias para aumentar a eficiência do seu processamento. No contexto da e-ScienceNet, utilizamos uma solução inicial para o problema que realiza o mapeamento completo entre duas ou mais Ontologias. No entanto, estudos mais detalhados devem ser realizados em trabalhos futuros de forma a tentar solucionar o problema de maneira mais eficaz.

2.4 e-Science, Redes Ponto a Ponto e Web Semântica

A e-Science proporciona uma visão promissora de como as tecnologias da informação podem ajudar a aperfeiçoar o processo de pesquisa científica. Entretanto, atualmente existe uma grande lacuna entre os esforços empregados e a visão da e-Science, a qual tem um grau de facilidade de uso e automação muito mais elevados, e que propõe maneiras de colaboração e computação em escala global [1].

O modelo tradicional de cliente/servidor utilizado pela web atual pode trazer diversos problemas quando aplicado a e-Science. O grande volume de dados proporcionado pelos

experimentos científicos e sua necessidade cada vez maior de poder de processamento, podem causar uma grande perda de desempenho em servidores centralizados, que, com somente um ponto de acesso, pode deixar cientistas sem acesso aos serviços por um tempo indeterminado durante o experimento científico.

Os problemas de processamento e espaço físico para armazenamento de dados em servidores podem ser solucionados com a utilização de técnicas de clusterização de diversos servidores. No entanto, além de não ser uma solução simples de ser implementada, a clusterização também é uma solução bem custosa, que muita das vezes inviabiliza sua implementação no meio acadêmico.

Desta forma, a descentralização vem se mostrando uma solução plausível para os problemas apresentados, mas trazendo os seus próprios desafios [20] [21]. As Redes Ponto a Ponto por sua vez, utilizam recursos de diversos computadores distribuídos geograficamente e apresentam uma boa alternativa para se trabalhar com arquiteturas distribuídas totalmente descentralizadas.

As Redes Ponto a Ponto tem se mostrado eficientes para o compartilhamento de dados disponíveis atualmente. Considerando que a informação científica em sua natureza é distribuída e de livre acesso, assim como os dados, o uso de uma Rede Ponto a Ponto pode ser uma das soluções a serem analisadas.

Ao utilizarmos as Redes Ponto a Ponto em conjunto com e-Science, temos um único ponto de acesso para que os cientistas acessem as bases de conhecimento dispersas entre os cientistas conectados a rede. Além disso, temos a possibilidade das aplicações científicas existentes se conectarem e utilizarem as funcionalidades disponíveis na rede, e o seu baixo custo de utilização e manutenção quando comparado a um sistema distribuído desse porte.

Recentemente as Redes Ponto a Ponto também têm sido utilizadas com sucesso para a interconexão entre grandes bases de dados científicas heterogêneas distribuídas, possibilitando a troca e pesquisa em estruturas de dados complexas [27]. Um problema dessa abordagem é que a maioria dessas estruturas são muito complexas para o entendimento de cientistas de outros domínios que não a computação, além de não possuir uma descrição semântica dos dados, dificultando portanto seu entendimento.

Considerando a e-ScienceNet, adicionamos um nível acima da pesquisa semântica dos dados em Redes Ponto a Ponto, criando comunidades semânticas para agrupar cientistas com mesmo conteúdo, especialidade e interesses em comunidades similares.

Mas as Redes Ponto a Ponto tendem a ter um volume de comunicação elevado quando o fluxo de trabalho científico aumenta. Considerando que a pesquisa científica se concentra em pequenas comunidades com algumas conexões entre si, podemos então limitar a comunicação entre os nós da rede para somente aqueles que trabalham em tal comunidade científica, diminuindo assim a comunicação entre os nós e aumentando a eficiência das pesquisas na rede [27].

Simulações de experimentos realizados mostraram que a seleção de nós baseados em sua especialidade para montar um grupo, aumentou o desempenho de Redes Ponto a Ponto, tanto para a precisão das pesquisas quanto para a quantidade de mensagens transmitida [37]. Considerando este contexto, a Web Semântica pode auxiliar na gerência das especialidades dos cientistas, facilitando a especificação de grupos de pesquisa formados por nós na rede. Utilizando Ontologias para descrever os grupos e suas especialidades, podemos utilizar técnicas de inferência, descobrindo relacionamentos relevantes entre grupos de pesquisa dispersos na rede. A vantagem dessa abordagem é que o processamento e resultados de um dado experimento não será distribuído para um número aleatório de nós, mas sim para aqueles nós que possam ter dados relevantes para a especialidade daquela pesquisa.

Considerando que o conhecimento em pesquisa científica pode ser representado por Ontologias de forma hierárquica, podemos utilizar essa hierarquia e montar um grafo de conceitos semânticos, onde cada nó representa um conceito que foi estruturado de forma semântica utilizando técnicas de inferência para descobrir novos nós e conexões entre eles. Podemos utilizar essa estrutura semântica como base da indexação dos nós em uma Rede Ponto a Ponto, redirecionando as pesquisas para somente comunidades no grafo que contenham dados mais relevantes para aquela pesquisa realizada.

A Web Semântica visa resolver o problema da complexidade da informação, provendo suporte a avançados meios de representar e processar essas informações. Já as Redes Ponto a Ponto, visão solucionar a complexidade do sistema, possibilitando um meio flexível e decentralizado para armazenar e processar essas informações [24]. Fazendo a junção dessas duas tecnologias para resolver alguns dos problemas disponibilizados pela e-Science,

podemos conseguir uma forma eficiente, barata e redundante para ser utilizada pelos cientistas no meio acadêmico.

2.5 Trabalhos Relacionados

Na literatura são encontrados trabalhos relacionados a e-ScienceNet que tratam sobre os principais tópicos abordados por essa dissertação, sendo eles e-Science, Redes Ponto a Ponto e Web Semântica.

O principal trabalho relacionado é [36], onde é proposto a utilização de Redes Sobrepostas Semânticas. Com a utilização de Redes Sobrepostas Semânticas, podemos organizar os nós da Rede Ponto a Ponto de acordo com suas especialidades e interesses, assim como proposto em [37]. A diferença para a e-ScienceNet é a utilização de Ontologias para criação dessas Redes Sobrepostas de acordo com a especialidade e interesse do nó, aumentando a precisão das pesquisas, diminuindo a comunicação entre nós e proporcionando a descoberta de novas comunidades semânticas de acordo com as inferências realizadas sobre essas Ontologias, assim como detalhado no capítulo 3, seção 3.4.

Considerando as Redes Sobrepostas Semânticas criadas através dos termos das Ontologias, podemos agrupar os cientistas nessas Redes Sobrepostas Semânticas e obter uma infraestrutura para apoio a e-Science similar a proposta em [20].

Além disso, a e-ScienceNet proporciona uma modularização das suas funcionalidades com a utilização de Componentes de Software, similar a [64], onde temos uma Rede Ponto a Ponto que utiliza o padrão OSGi[4] para pesquisa e instalação de componentes. A diferença para a e-ScienceNet é, além da pesquisa e instalação de componentes, a utilização desses componentes com a sua lógica de trabalho para propiciar as funcionalidades à rede, se diferenciando da maioria de Redes Ponto a Ponto existentes, como, [29, 30, 31, 32, 33, 34, 35] que são comumente utilizadas para compartilhamento de arquivos. Mais informações sobre os Componentes de Software na e-ScienceNet são detalhadas no capítulo 3, seção 3.2.

[4] OSGi é um conjunto de especificações que define um sistema dinâmico de componentes para Plataforma Java.

Dentre as lógicas de componentes que podem ser implementados na e-ScienceNet podemos citar [65, 66, 67, 68, 69, 70, 71].

Em [65, 66] temos uma repositório para Serviços Web Semânticos com a utilização de uma infraestrutura de Rede Ponto a Ponto. Toda a pesquisa de Serviços Web Semânticos é realizada de forma distribuída, utilizando Serviços Web Semânticos conhecidos pelos diversos nós na rede. Na e-ScienceNet um componente de compartilhamento de Serviços Web Semânticos pode utilizar OWL-S [72] para pesquisa semântica desses serviços com auxilio do Gerente de Dados, com a possibilidade de utilização das Ontologias e mapeamentos disponibilizados na rede através do Gerente de Semântica.

Outros exemplos similares a [65] e [66] são [67] e [68], onde [67] é utilizado para compartilhamento de dados bibliográficos, com a vantagem da sua implementação na e-ScienceNet ser o agrupamento através das Redes Sobrepostas Semânticas. Em [68] é proposto uma banco de dados distribuído para dados biológicos, onde na e-ScienceNet toda essa informação poderia ser utilizada por outros componentes através da comunicação entre componentes, assim como detalhado na seção de 3.2 do capítulo 3. Além disso, o banco de dados proposto em [68] teria uma grande redundância dos dados disponibilizados na rede com um custo muito baixo de implementação e manutenção.

Na e-ScienceNet temos também o compartilhamento e mapeamento de Ontologias pelos nós na rede, similar a [69] onde é proposto o compartilhamento de Ontologias, com a diferença de na e-ScienceNet além de serem disponibilizadas as Ontologias, temos também o mapeamento entre essas Ontologias.

O grande diferencial da e-ScienceNet em relação a [70] é o uso de estruturas semânticas para auxiliar na futura composição de componentes em um Workflow Científico. Com a possibilidade de comunicação entre os Componentes de Software na e-ScienceNet, a composição de Workflows pode usufruir de outros componentes através da comunicação entre componentes, como, por exemplo, se comunicar com um Componente para Descoberta e Publicação de Serviços Web Semânticos durante a composição. Além disso, a composição pode ser realizada em tempo de execução, ou seja, quando solicitado um resultado, é realizada a composição se baseando no objetivo final, e a partir disso podemos compor quais Serviços Web precisamos utilizar para chegar ao objetivo final e quais serviços estão disponíveis na rede naquele momento.

Alguns trabalhos relacionados podem ser visualizados e analisados em [71] com mais detalhes. Grande parte desses trabalhos podem ser implementados na e-ScienceNet como Componentes de Software.

3 E-SCIENCENET

Este capítulo detalha uma arquitetura para apoio a e-Science chamada de e-ScienceNet. Esta arquitetura engloba diversas tecnologias com foco na criação de um ponto único e de fácil acesso para que cientistas dispersos geograficamente possam gerar, analisar, compartilhar e discutir seus trabalhos de uma maneira mais produtiva. A arquitetura e-ScienceNet tenta abranger os principais requisitos apresentados na seção 2.1 do capítulo 2.

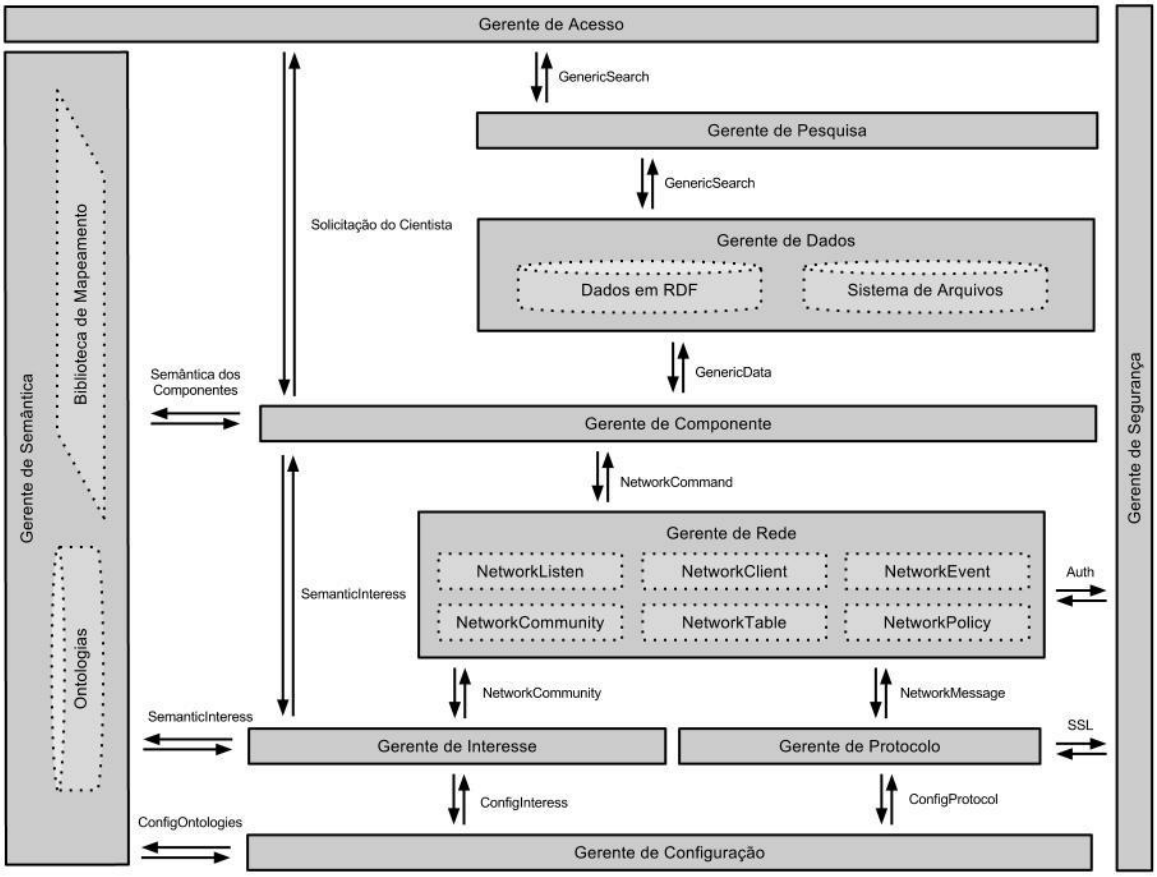

Figura 3.1: arquitetura da e-ScienceNet

A arquitetura é dividida em camadas, as quais chamamos de gerentes, sendo esses gerentes responsáveis por prover as funcionalidades requeridas para a solução dos requisitos apresentados na primeira seção do capítulo dois. Na Figura 3.1, são apresentados os gerentes da e-ScienceNet em uma visão de alto nível (abstrata).

Desta forma, este capítulo foca em apresentar os requisitos, a estrutura, questões relacionadas à implementação do protótipo e desafios associados a desenvolver uma arquitetura que dê suporte a e-Science. Na primeira seção, apresentamos o Gerente de Acesso, responsável pela interação com os cientistas utilizando a e-ScienceNet. Na segunda seção, detalhamos o Gerente de Componentes, responsável pelas funcionalidades disponibilizadas na rede pelos componentes (aplicações e dados) criados pelos cientistas. Na terceira seção, o Gerente de Rede é descrito, sendo responsável pela comunicação entre os cientistas geograficamente dispersos, sendo considerado o componente principal da e-ScienceNet. Na quarta seção, apresentamos o Gerente de Semântica, responsável por prover as funcionalidades relacionadas à Web Semântica. Na quinta seção, temos o Gerente de Interesse, ao qual com auxilio do Gerente de Semântica, permite a criação das comunidades semânticas a serem conectadas pelo Gerente de Rede. Na sexta seção, temos o Gerente de Pesquisa responsável por receber uma requisição do cientista e repassar para a rede. Na sétima seção, apresentamos o Gerente de Dados, responsável por acessar fisicamente os dados disponíveis na rede. Na oitava seção, apresentamos o Gerente de Protocolo, camada que auxilia o Gerente de Rede a entender o que os outros cientistas estão transmitindo na rede. Na nona e última seção, temos o Gerente de Segurança, responsável por prover a segurança de acesso aos dados disponibilizados pelos cientistas na e-ScienceNet.

3.1 Gerente de Acesso

Para facilitar o processo de pesquisa científica, a e-ScienceNet foi desenvolvida segundo o modelo de uma Interface de Programação de Aplicativos (API). Utilizando esta abordagem, independente do dispositivo, qualquer aplicativo pode ter acesso às funcionalidades desenvolvidas pela e-ScienceNet.

A API viabiliza o acesso de aplicações distintas a base de conhecimento distribuída da e-ScienceNet. Ao viabilizar este acesso, melhora-se a busca pelos dados disponíveis por parte dos cientistas, independente da sua preferência de aplicativo e interesse científico.

Ao iniciar um processo de pesquisa científica, um cientista necessita utilizar serviços e aplicativos existentes. Atualmente muitos cientistas utilizam a web para realizarem diversas tarefas relacionadas à pesquisa científica, mas tem dificuldades de lidar com características essencialmente técnicas, como, por exemplo, a forma de distribuição de dados e serviços pela web.

É importante que os cientistas não gastem tempo de pesquisa se preocupando com problemas não relacionados ao seu trabalho científico. Utilizando a web como uma base de conhecimento distribuída, ou seja, sem o uso da e-ScienceNet, um dado cientista tem que manter o seu próprio banco de dados com a descrição de aplicativos, bases de conhecimento, entre outros, sendo de sua inteira responsabilidade o carregamento dos aplicativos, as possibilidades de composição e os mecanismos para acesso a esses dados.

Utilizando a e-ScienceNet, o acesso a estes aplicativos fica facilitado, uma vez que os detalhes técnicos de acesso ficam transparentes para o cientista, independente da localização física, tipo de informação solicitada e dispositivo utilizado. A vantagem da e-ScienceNet ser disponibilizada em especificações de uma API, não é somente para facilitar o acesso dos cientistas à base de conhecimento globalmente distribuída, mas também facilitar o provimento de bases de conhecimento distribuídas e algoritmos relacionados a aplicativos científicos específicos.

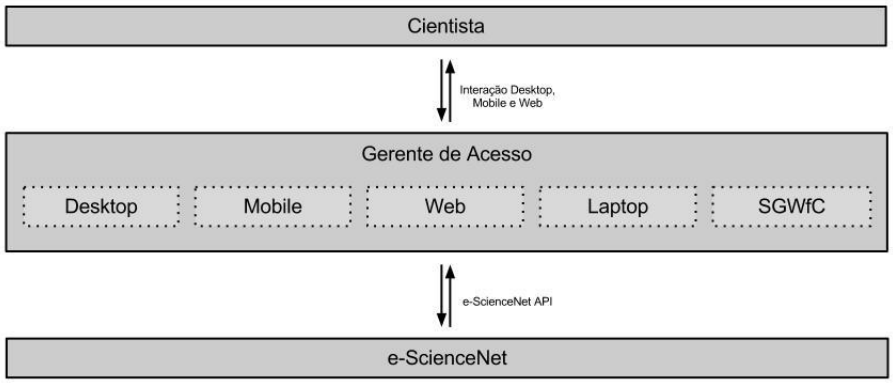

Figura 3.2: Gerente de Acesso

Como podemos ver na Figura 3.2, o Gerente de Acesso não faz parte da API da e-ScienceNet, mas é parte fundamental do seu funcionamento.

Alguns exemplos de utilização da e-ScienceNet a partir de diferentes contextos são apresentados a seguir.

a) Acesso através de um computador desktop.

Inicialmente podemos ter uma interface para ser utilizada em computadores pessoais. Esta interface gráfica genérica disponibiliza acesso as funcionalidades da e-ScienceNet, podendo, no entanto, ser adaptada às necessidades específicas de cada cientista ou domínio de pesquisa.

```
public class MainForm extends JFrame implements WindowListener {
    public MainForm() {
        this.loadConfg();
        this.loadComponents();
        this.loadEvents();

        eScienceNet.load();

        this.pack();
        this.setLocationRelativeTo(null);
    }

    @Override
    public void windowClosing(WindowEvent aEvent) {
        if (eScienceNet.getState() == eScienceNetState.RUNNING) {
            eScienceNet.unload();
        }

        this.setVisible(false);
        this.dispose();
    }
```

Figura 3.3: execução da e-ScienceNet através de um projeto desktop

Na Figura 3.3, temos um exemplo de um Gerente de Acesso para clientes desktop. Essa versão inicial do cliente desktop não proporciona acesso a todas as funcionalidades desenvolvidas na e-ScienceNet. O código mostrado na Figura 3.3 mostra a criação de um formulário em um ambiente desktop. Durante a sua criação, o formulário inicia a e-ScienceNet através do comando eScienceNet.load(). Já em windowClosing(...), método executado quando o cientista fechar o formulário, é verificado o estado atual da e-ScienceNet. Caso esteja em execução, chamamos o comando eScienceNet.unload() para descarregar a e-ScienceNet.

b) Acesso através da API por outros aplicativos.

Em alguns casos não é necessário que o acesso à rede seja disponibilizado de forma tão genérica. Dependendo da necessidade do cientista, do dispositivo ou aplicativo a ser utilizado, certos detalhes não são importantes na hora da utilização da e-ScienceNet. Podemos citar como exemplo alguns Sistemas Gerenciadores de Workflows Científicos [73], como, Kepler [74] e Taverna [75]. Esses aplicativos podem acessar os dados disponíveis na base de conhecimento da e-ScienceNet através da API, enriquecendo e facilitando o processo de criação de Workflows Científicos pelos cientistas.

c) Acesso através de um servidor dedicado.

Alguns centros de pesquisas podem não disponibilizar acesso irrestrito para os cientistas da sua rede local à internet. Para solucionar tal problema, alguns centros de pesquisa utilizam Firewall[5] para limitar o acesso a sua rede interna. Considerando que um centro de pesquisa queira ter acesso a base de conhecimento provida pela e-ScienceNet, mas sem interferir na segurança interna da sua rede local, devemos possibilitar a sua execução em um servidor dedicado.

[5] Dispositivo de uma rede de computadores que tem por objetivo aplicar uma política de segurança a um determinado ponto da rede.

```
public static void main(String[] args) {
    eScienceNet.load();

    Runtime.getRuntime().addShutdownHook(new Thread() {
        @Override
        public void run() {
            if (eScienceNet.getState() == eScienceNetState.RUNNING) {
                eScienceNet.unload();
            }
        }
    });
}
```

Figura 3.4: execução da e-ScienceNet em seu pacote para disponibilização

Como mostrado na Figura 3.4, a própria e-ScienceNet em seu pacote para disponibilização implementa uma execução simples. Essa execução é ideal para servidores dedicados, onde a e-ScienceNet é carregada e fica na escuta de conexões até que a máquina virtual Java continue em execução.

d) Acesso através de dispositivos móveis.

Enquanto super computadores auxiliam na resolução de problemas complexos, os dispositivos móveis facilitam muito o acesso de cientistas às informações disponibilizadas na internet. Pela sua facilidade de uso e locomoção, os dispositivos móveis podem aumentar o poder de colaboração entre cientistas, mostrando seu potencial inexplorado ao facilitar o processo científico. Na versão 1.0 da e-ScienceNet a interface para dispositivos móveis ainda não foi disponibilizada.

3.2 Gerente de Componente

Na e-ScienceNet os componentes são responsáveis por disponibilizar os serviços de rede aos cientistas, além de permitir o acoplamento de novos aplicativos na rede científica. Os componentes utilizam o conceito de plug-in[6], onde os componentes utilizados só precisam ser carregados pela e-ScienceNet para que o serviço seja disponibilizado na rede.

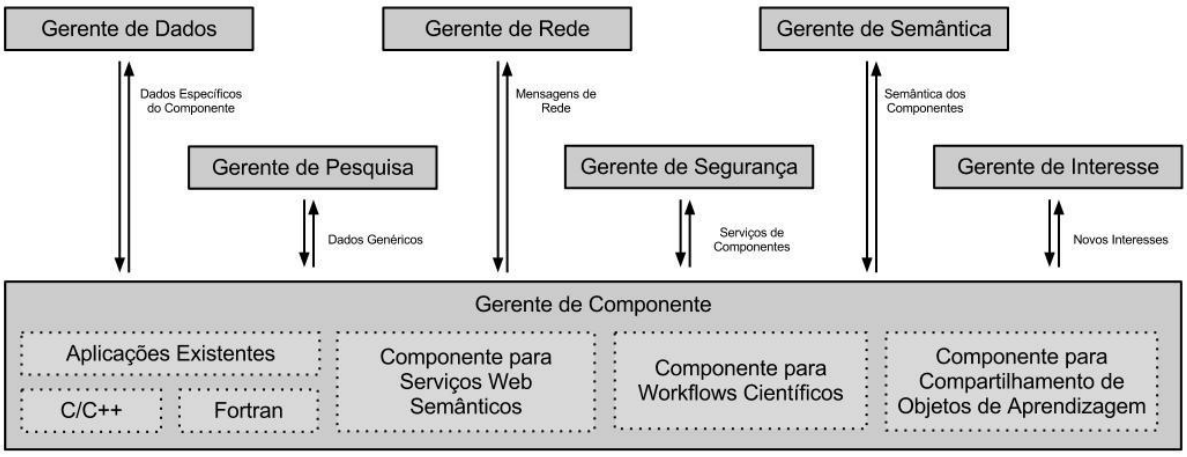

Figura 3.5: Gerente de Componente

Como podemos ver na Figura 3.5, o Gerente de Componente é responsável por prover acesso as principais funcionalidades da rede aos componentes. Na parte superior da Figura 3.5 temos alguns exemplos de solicitações que os componentes podem fazer aos outros gerentes através do Gerente de Componentes. Na parte inferior da Figura 3.5 mostramos alguns exemplos de componentes e como o Gerente de Componente tenta abstrair a chamada a aplicações existentes[7].

A adaptabilidade da e-ScienceNet depende diretamente dos seus componentes. As necessidades dos cientistas e as novas tecnologias que auxiliam no processo de pesquisa

[6] Plug-in são componentes utilizados para adicionar funções a softwares existentes, provendo funcionalidades especiais ou específicas, sendo geralmente leves e utilizados somente sob demanda.

[7] Os componentes utilizados como exemplo na Figura 3.5 ainda não estão efetivamente implementados na e-ScienceNet. Já a chamada a aplicações existentes realizada pelo Gerente de Componente é apresentado mais a diante nesta seção.

científica vêm sofrendo diversas mudanças ao longo dos anos. Considerando o contexto da Web Semântica e sua importância para a operacionalização das buscas e uso de aplicativos e dados disponibilizados na rede, todos os componentes devem ter uma descrição semântica indicando os serviços disponíveis. Utilizando esta descrição semântica dos componentes, podemos realizar pesquisas e instalações de novos componentes sempre que o cientista necessitar de novas funcionalidades que já foram criadas e distribuídas na e-ScienceNet.

A escalabilidade de funcionalidades da e-ScienceNet é dependente dos componentes disponibilizados. Um cientista pode instalar em seu dispositivo quantos componentes desejar, aproveitando ao máximo os recursos disponibilizados na e-ScienceNet.

Atualmente na e-ScienceNet a pesquisa de componentes existentes é feita de maneira simplificada. O objetivo é que em versões futuras este processo seja aprimorado. Cada componente tem um ou mais termos ontológicos para descrever suas funcionalidades. Utilizando como exemplo um componente para alinhamento distribuído de sequências e a Ontologia Sequence Aligning Ontology[8], podemos dizer que o componente de alinhamento distribuído de sequências é descrito pela classe "sequence_aligning" da referida Ontologia. Quando um cientista realizar uma pesquisa por algum componente que é descrito pela classe "sequence_aligning", o Gerente de Rede ao qual apresentamos na próxima seção, faz uma busca entre os nós aos quais esse cientista está conectado e verifica se em alguma dessas conexões existe algum nó com um componente descrito por essa classe. Caso seja retornado algum componente é apresentado ao cientista para que ele instale (i.e. seja disponibilizado o link para este cientista) ou não aquele componente.

A instalação de novos componentes na e-ScienceNet é um trabalho que pode ser aprimorado em versões futuras, onde podemos, por exemplo, expandir a pesquisa de novos componentes para comunidades semânticas descritas na mesma Ontologia, ou então, retornar todos os componentes que tem sua descrição semântica englobando aquela classe.

Além disso, alguns aspectos de segurança que fogem do escopo deste trabalho devem ser levados em consideração, sendo um dos principais, a verificação se o componente é seguro ou não de ser executado.

[8] A Sequence Aligment Ontology é uma modificação da MyGridOntology, considerando somente aspectos relacionados ao alinhamento de sequencias. Disponível em http://gabriellacastro.com.br/SequenceAligningOntology.owl.

A chamada à aplicações existentes atualmente implementado na e-ScienceNet se da dá seguinte forma: nas configurações da e-ScienceNet é informado o caminho da aplicação existente a ser executada e os seus parâmetros. Quando o Gerente de Rede recebe uma mensagem indicando que uma dada aplicação deve ser executada, o Gerente de Componente faz a chamada a essa aplicação existente passando como parâmetro os valores recebidos na mensagem.

```
public abstract class Component {
    public abstract void load();
    public abstract void unload();
    public abstract String getName();
    public abstract String[] getSemanticDescription();
    public abstract String[] getDependency();
    public abstract NetworkCommand[] getCommands();
}
```

Figura 3.6: implementação de um componente

A implementação de um componente se dá através da extensão da classe mostrada na Figura 3.6. O método load() e unload() são executados respectivamente durante o carregamento e descarregamento da rede. O método getName() retorna o nome do componente. O método getSemanticDescription() retorna uma lista de descrições semânticas para aquele componente. Já o método getDependency() retorna uma lista de classes que devem estar carregadas antes que esse componente seja carregado. No método getCommands() deve ser implementado a lista com os comandos (serviços) disponíveis na rede através deste componente. É importante ressaltar que por questões de organização as classes de um componente deve ser empacotado dentro de um arquivo JAR[9].

```
<SemanticWebService.getServices>
        <filter>SequenceAligningOntology#sequence_aligning</filter>
</SemanticWebService.getServices>
```

Figura 3.7: comando de invocação de um componente

[9] Java Archive é um formato de arquivo tipicamente utilizado para agrupar classes Java em somente um arquivo.

Na Figura 3.7 tem-se um exemplo de como é realizado a invocação de um método entre os nós da rede. Os comandos são strings no formato "Component.Comando", onde o método responsável por aquele comando é invocado recebendo os mesmos parâmetros passados ao comando. Na Figura 3.7 tem-se a invocação do comando getServices do componente SemanticWebService, sendo um filtro passado como parâmetro, onde somente Serviços Web Semânticos com essa descrição serão retornados.

```
public NetworkCommand getServicesCommand(NetworkCommand aCommand) {
    // Cria um comando de retorno sem nenhum parâmetro.
    NetworkCommand lReturnCommand = new NetworkCommand();

    // Cria um filtro com o valor do campo "filter" passado como parâmetro ao comando.
    ServiceFilter lFilter = new ServiceFilter(aCommand.getParam("filter"));
    // Pega a lista de serviços disponíveis baseado no filtro.
    ServiceList lServiceList = this.getServices(lFilter);

    // Passa por todos os serviços na lista de Serviços Web disponíveis localmente.
    for (Service lService : lServiceList) {
        // Adiciona como parâmetro de retorno ao comando o nome e a URI do Serviço Web.
        lReturnCommand.addParam(new NetworkParam("service", lService.getName() + ":" + lService.getURI()));
    }

    // Retorna o comando que será enviado como resposta ao outro nó.
    return lReturnCommand;
}
```

Figura 3.8: invocação de um componente

A Figura 3.8 mostra o método executado pelo Gerente de Rede quando recebe um comando para aquele componente. O Gerente de Rede é detalhado na próxima seção. Na Figura 3.8 o método getServicesCommand() recebe um comando como parâmetro e retorna outro comando em resposta. Por questões de didática e de simples implementação, utilizamos somente um filtro para os Serviços Web Semânticos, mas como será visto no Gerente de Dados, filtros mais complexos podem ser implementados em versões futuras da e-ScienceNet.

3.3 Gerente de Rede

O Gerente de Rede é responsável por prover a abstração lógica de rede da e-ScienceNet. Ao abstrair a forma de acesso física da rede, conseguimos manter a liberdade para que todos os cientistas acessem a e-ScienceNet através de qualquer dispositivo disponível.

A e-ScienceNet se baseia na afirmação que os cientistas podem utilizar qualquer sistema operacional com qualquer componente com capacidade de conexão de rede. Para tal, utilizamos o modelo de camadas TCP/IP [77], sendo que a e-ScienceNet trabalha na camada de aplicação do modelo utilizado.

É no Gerente de Rede que são implementadas as funções de escuta de conexões, conexões com outros cientistas, conexão às comunidades semânticas, envio de mensagens entre os componentes e outras funções que serão discutidas ao longo dessa seção.

Ao utilizar a camada de aplicação do modelo de camadas TCP/IP, a topologia de rede da e-ScienceNet apresentada a seguir descreve a lógica do fluxo de dados na rede, e não como essa comunicação é realizada através da estrutura física de rede utilizada.

A topologia de rede da e-ScienceNet se caracteriza como uma topologia de Rede Ponto a Ponto híbrida, onde são utilizadas técnicas de Redes Ponto a Ponto centralizadas e descentralizadas, conforme detalhado na segunda seção do capítulo de pressupostos teóricos. Com o uso de uma arquitetura de Rede Ponto a Ponto híbrida, conseguimos adequar às diversas necessidades dos cientistas conectados a e-ScienceNet, diminuindo os custos de infraestrutura e aumentando a flexibilidade e expansibilidade dos cientistas conectados na rede, conforme será detalhado a seguir. Na e-ScienceNet utilizamos também o conceito de redes sobrepostas apresentado na segunda seção do capítulo de pressupostos teóricos.

As comunidades semânticas são grupos de cientistas que compartilham das mesmas especialidades e interesses mantidos em um mesmo grupo de conexões de rede. Neste contexto, os cientistas na e-ScienceNet são caracterizados por um conjunto de elementos ontológicos que descrevem semanticamente suas especialidades e interesses. Neste trabalho, com auxilio da Web Semântica, utilizam-se Ontologias para descrever semanticamente o conteúdo de cada comunidade semântica de cientistas. Os grupos de conexões de rede aqui chamados de comunidades semânticas são termos, relacionamentos e restrições extraídos das Ontologias. Um grupo pode ser caracterizado por um ou mais elementos (termos, relacionamentos e/ou restrições) de diferentes Ontologias, que, através de mapeamentos e possíveis inferências, são descobertas suas similaridades semânticas. Uma das grandes vantagens no uso das comunidades semânticas para indexação dos cientistas na rede, é que os cientistas na e-ScienceNet serão responsáveis por gerenciar somente algumas conexões com outros cientistas que compartilham os mesmos dados e serviços. Dessa forma, conseguimos

diminuir a quantidade de recursos necessários para utilização da e-ScienceNet, mas aumentando a precisão das pesquisas realizadas pelos cientistas.

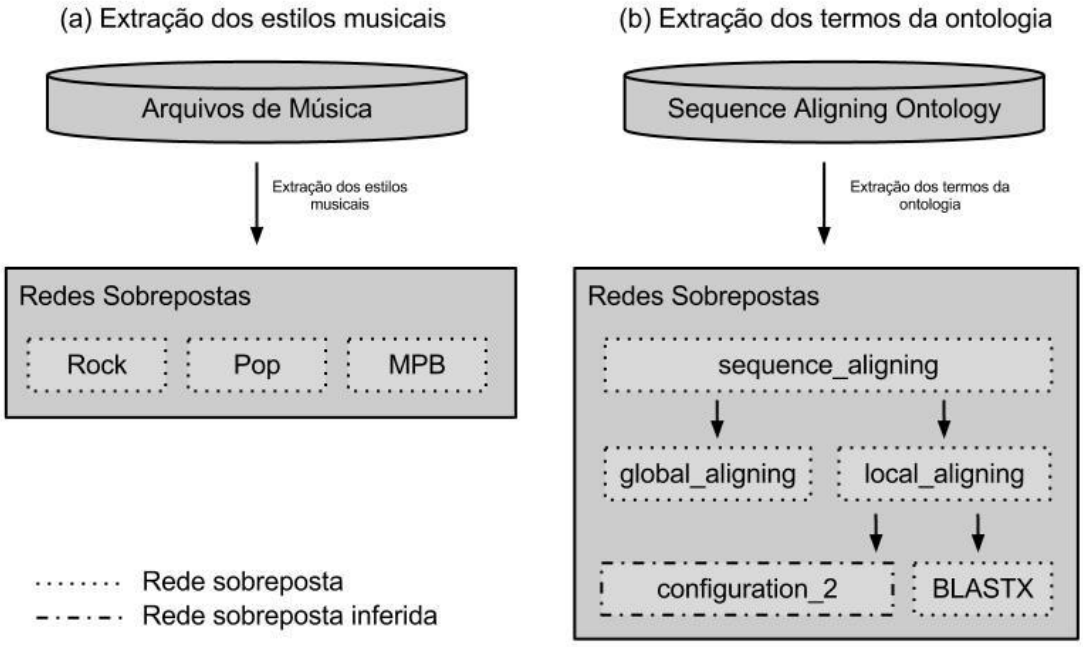

Figura 3.9: Redes Sobrepostas criadas através de estilos musicais e Ontologias

Na e-ScienceNet temos uma Rede Ponto a Ponto estruturada similar a [36], que utiliza uma estrutura de árvores para organizar os seus nós na rede. A grande diferença entre [36] e a e-ScienceNet é o uso de Web Semântica e seus relacionamentos semânticos entre os termos, montando um grafo mais refinado com as Redes Sobrepostas. Como visto na Figura 3.9 (a), em [36] as Redes Sobrepostas são criadas com os estilos dos arquivos de música utilizados. Já na Figura 3.9 (b), temos a criação das comunidades semânticas através da extração de termos da ontologia Sequence Aligning Ontology. É interessante notar que as Redes Sobrepostas criadas através de Ontologias possuem um relacionamento semântico entre as redes, possibilitando a expansão das pesquisas para Redes Sobrepostas distintas, considerando as ligações semânticas entre os termos da Ontologia, o que não é possível em [36]. Além disso, temos a possibilidade de Redes Sobrepostas inferidas a partir de regras criadas na Ontologia, assim como ocorrido com a Rede Sobreposta “configuration_2” mostrado na Figura 3.9.

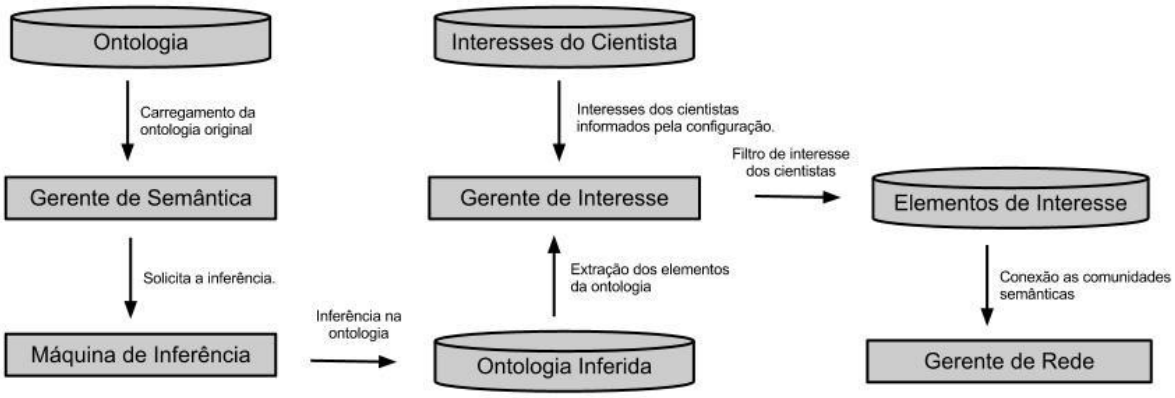

Figura 3.10: processo de extração de elementos para comunidades semânticas

Com o uso de uma rede estruturada para as Redes Sobrepostas, conseguimos gerenciar de forma eficiente os nós dispersos pela e-ScienceNet de acordo com suas especialidades e interesses, sendo essas especialidades e interesses especificadosa partir dos elementos extraídos das Ontologias. Desta forma, as buscas na e-ScienceNet ficam mais direcionadas e, portanto, mais precisas, otimizando o trabalho dos pesquisadores. Como pode ser visto na Figura 3.10, a extração dos termos das Ontologias é um processo realizado pelo Gerente de Semântica, filtrado pelo Gerente de Interesse e repassado ao Gerente de Rede para que sejam realizadas as devidas conexões. O Gerente de Semântica e o Gerente de Interesse são apresentados em mais detalhes na seção quatro e cinco deste capítulo. Primeiramente para que uma busca seja realizada na e-ScienceNet, o Gerente de Semântica verifica as Ontologias existentes na rede e os mapeamentos existentes. Assim que o Gerente de Interesse recebe os dados das Ontologias existentes e seus mapeamentos do Gerente de Semântica, é realizado um filtro dos interesses dos cientistas através de uma comparação simples entre os interesses armazenados nas configurações para aquele cientista e o campo rdf:about. Após receber do Gerente de Interesse os elementos da Ontologia relacionados aos interesses do cientista e o detalhamento dos mesmos, tais como, rdf:about, rdfs:subClassOf e as restrições, o Gerente de Rede realiza uma pesquisa na rede através dos super nós conhecidos, e tenta assim descobrir se existem comunidades semânticas representando aqueles elementos que foram selecionados como de interesse do cientista pelo Gerente de Interesse. Caso necessário, realiza também inferências nas Ontologias relacionadas, descobrindo assim novos relacionamentos e/ou restrições além do explicito na Ontologia com o objetivo de descobrir novas comunidades

semânticas com conteúdo similar, justificando o repasse das buscas para essas comunidades semânticas mapeadas e/ou inferidas.

Uma das grandes vantagens no uso das comunidades semânticas para indexação dos cientistas na rede, é que os cientistas na e-ScienceNet serão responsáveis por gerenciar somente algumas conexões com outros cientistas que compartilham os mesmos dados e serviços. Dessa forma, conseguimos diminuir a quantidade de recursos necessários para utilização da e-ScienceNet, mas aumentando a precisão das pesquisas realizadas pelos cientistas.

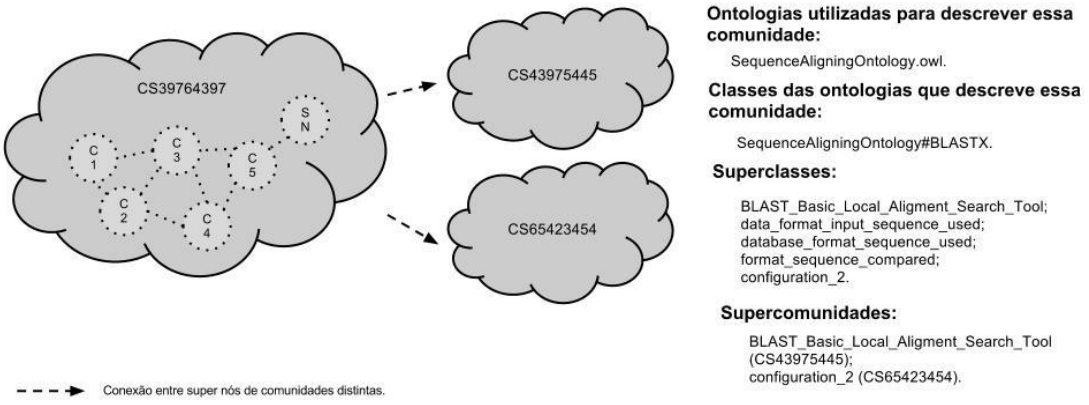

Figura 3.11: comunidades semânticas

A Figura 3.11 apresenta as comunidades semânticas (CS) representando uma única classe extraída da Ontologia Sequence Alignment Ontology. Cada comunidade semântica tem o seu próprio grafo com os cientistas (C) interessados naquela comunidade semântica e um ou mais super nós (SN) responsáveis pela gerência daquela comunidade semântica.

No exemplo apresentado na Figura 3.11, as comunidade semânticas CS43975445 e CS65423454 têm relacionamentos com a comunidade semântica CS39764397. Estes relacionamentos entre as comunidades semânticas são criados a partir de mapeamentos e possíveis inferências entre as Ontologias utilizadas pelos cientistas.

Atualmente na e-ScienceNet são realizadas conexões entre alguns dos nós em uma comunidade semântica de forma aleatória. Um problema visível dessa abordagem é o crescimento de nós em uma comunidade semântica, onde, quanto maior a quantidade de nós

em uma comunidade semântica, maior a propagação e transferência de mensagens naquela comunidade. Em versões futuras, podemos realizar um balanceamento da quantidade de conexões entre os nós, limitando a um número de conexões para cada nó, ou de acordo com configurações informadas pelos cientistas com um número máximo de conexões.

Para manter as comunidades semânticas em uma Rede Ponto a Ponto de forma estruturada, precisamos manter uma lista com as comunidades semânticas existentes na e-ScienceNet e os respectivos cientistas daquela comunidade semântica. Quando aplicado em maior escala, a quantidade de comunidades semânticas abrangendo os diversos interesses e especialidades dos cientistas pode crescer muito rápido, interferindo no desempenho da e-ScienceNet e limitando o acesso de dispositivos com menor capacidade de armazenamento e processamento. Para resolver este problema na e-ScienceNet, utilizamos super nós e Tabelas de Espalhamento Distribuídas.

3.3.1 Super Nó

Os super nós são nós com maior poder de processamento e armazenamento conectados a e-ScienceNet [78]. Atualmente existem super nós dedicados em Redes Ponto a Ponto, onde seu único objetivo na rede é manter os índices de arquivos disponíveis para os outros nós. Uma arquitetura de rede utilizando super nós cria uma centralização em sistemas decentralizados, formando uma rede de super nós estruturada [27].

Na e-ScienceNet, os super nós são responsáveis por gerenciar as comunidades semânticas. Como já dito, todas as comunidades semânticas têm um ou mais super nós responsáveis por gerenciar os cientistas conectados nas respectivas comunidades semânticas. Um super nó pode gerenciar diversas comunidade semânticas ao mesmo tempo, dependendo somente da sua disponibilidade computacional.

Uma comunidade semântica com mais de um super nó gerenciando os seus nós, proporciona uma redundância de super nós caso aconteça algo inesperado com outro super nó daquela comunidade, mantendo assim a comunidade semântica em funcionamento enquanto o super nó com problema não volte ao seu funcionamento normal.

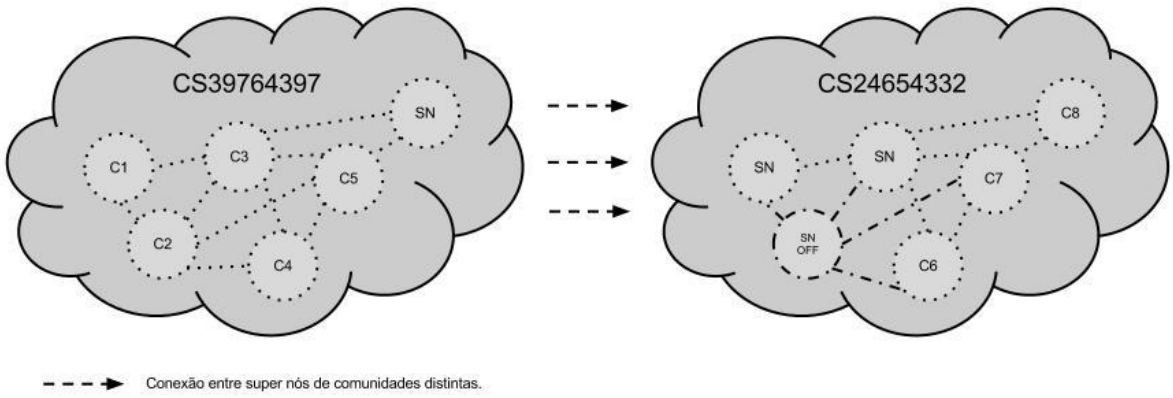

Figura 3.12: super nós e comunidades semânticas

Na Figura 3.12 mostramos duas comunidade semânticas interligadas entre si pelos seus respectivos super nós. A comunidade semântica a esquerda possui somente um super só, já a comunidade semântica a direita possui três super nós, sendo um deles marcado como *offline*.

As arquiteturas de super nós geralmente são baseadas em protocolos de roteamento de duas fases, onde o roteamento das mensagens primeiramente é enviado para os super nós das comunidades e então distribuídas para os outros nós da rede [78]. Em [79, 80] podemos ver que a pesquisa em super nós vem recebendo grande atenção nos últimos anos.

A escolha de super nós de uma comunidade semântica pode utilizar diversas características para avaliar qual o nó atual daquela comunidade semântica se comporta melhor como super nó. Dentre essas características podemos citar componentes físicos do dispositivo utilizado pelo cientista, como, o poder de processamento e capacidade de armazenamento, a distância física entre a maioria dos nós da comunidade semântica e o tempo de conexão daquele nó a e-ScienceNet. O ideal é a configuração da e-ScienceNet para utilizar qualquer uma destas características. Atualmente a e-ScienceNet implementa uma seleção de super nó aleatória. São gerados números aleatórios pelos nós conectados a comunidade semântica, o nó que gerar o menor número é selecionado para ser o super nó atual daquela comunidade. A renovação de super nó é feita somente quando o super nó atual se desconecta da comunidade semântica.

Uma abordagem mais desafiadora que poderá ser futuramente implementada na e-ScienceNet é a possibilidade de um dado nó em determinada comunidade científica se candidatar a super nó daquela comunidade. Todos os nós nas comunidades semânticas revezam o gerenciamento de suas comunidades com outros nós conectados por um determinado espaço de tempo.

Dispositivo	**Cientista**	**08:00 AM**	**09:00 AM**	**10:00 AM**
Desktop	Nó 001	Super Nó	-	-
Desktop	Nó 002	-	Super Nó	-
Móvel	Nó 003	-	-	-
Desktop	Nó 004	Super Nó	Super Nó	Super Nó

Tabela 3.1: revezamento de super nós em uma comunidade semântica.

Na Tabela 3.1, podemos ver que a cada faixa de tempo negociada entre os nós da comunidade semântica, temos diferentes nós gerenciando aquela comunidade semântica. Nesta política de revezamento, é importante ressaltar, considerando ainda a Tabela 3.1, que um dispositivo móvel que não possui recursos computacionais para gerenciar uma comunidade semântica não entrou na lista de super nós daquela comunidade semântica em momento nenhum.

Para se conectar na e-ScienceNet o nó precisa conhecer pelo menos outro nó de interesse que já esteja conectado na rede. Quando descobrimos esse nó de interesse, seja ele um super nó ou não, podemos solicitar as informações das comunidades semânticas atuais na rede ou outros nós que possuam essa informação.

Para tal, podemos nos basear em algumas técnicas implementadas nas Redes Ponto a Ponto atuais. Por exemplo, podemos utilizar a centralização proporcionada pelo BitTorrent [81] que tem um servidor central chamado de tracker responsável por gerenciar os nós com aquele arquivo.

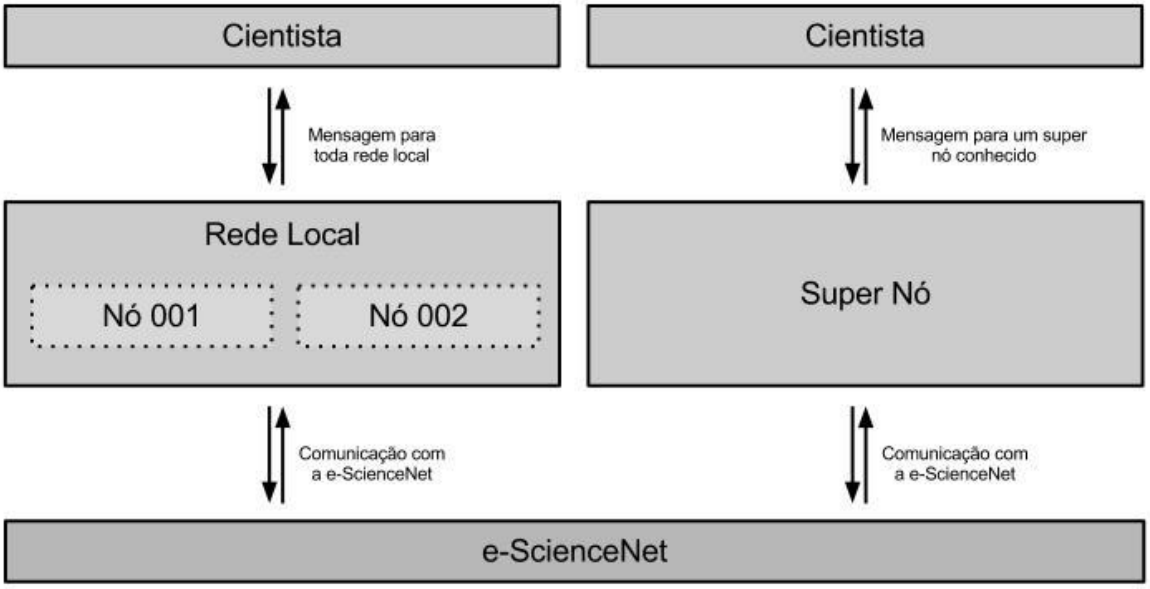

Figura 3.13: descoberta de super nós locais e centrais

Assim como mostrado na Figura 3.13, na e-ScienceNet podemos nos conectar a um servidor central, seja ele um super nó com endereço fixo ou um servidor web através do protocolo HTTP. Outra abordagem utilizada em redes locais é o envio de mensagens para toda a rede em broadcast[10]. Em um centro de pesquisa, por exemplo, podemos descobrir a localização dos super nós da e-ScienceNet através dos cientistas que já estão conectados na rede. Primeiramente é enviado um broadcast para toda a rede local, caso algum cliente (nó) da rede local responda que tem conhecimento de super nós da e-ScienceNet, é iniciado o processo de conexão a esse super nó informado.

3.3.2 Tabelas de Espalhamento Distribuídas

Para resolver o problema de que um super nó não precise armazenar todas as comunidades semânticas disponíveis na e-ScienceNet e todos os nós das suas comunidades semânticas, a e-ScienceNet utiliza Tabelas de Espalhamento Distribuídas (DHT).

[10] Processo pelo qual se transmite ou difunde determinada informação, tendo como principal característica que a mesma informação está sendo enviada para muitos receptores ao mesmo tempo.

As DHTs são uma classe de Sistemas Distribuídos completamente descentralizados, provendo um serviço de pesquisa similar às Tabelas de Espalhamento. A diferença é a distribuição das chaves e valores da tabela de espalhamento pelos nós na e-ScienceNet. A forma como as DHTs são projetadas causam um mínimo de desordem entre os nós da e-ScienceNet. Isso faz com que as DHTs escalem a um número considerável de nós, gerenciando as entradas, saídas e falhas contínuas dos cientistas.

Para armazenar as comunidades semânticas utilizamos uma DHT pelos super nós da e-ScienceNet. A DHT da e-ScienceNet utiliza Secure Hash Algorithm (SHA)[11] como técnica de espalhamento a partir da descrição semântica dos elementos nas Ontologias relacionadas, para criar a chave da tabela. Os valores da DHT para comunidades semânticas são os super nós conhecidos que estão gerenciando aquela comunidade semântica.

Nome	Descrição Semântica	Chave	Valor
BLAST_Basic_Local_Alignment_Search_Tool	http://gabriellacastro.com.br/SequenceAligningOntology#BLAST_Basic_Local_Alignment_Search_Tool	a14a94c70e3905ea94e3202c663ae0707f54119c	Super Nó 001, Super Nó 004
DNA_sequence	http://gabriellacastro.com.br/SequenceAligningOntology#DNA_sequence	440fa43087c8efac7f9bfdbb78ab8b1b9cdcce2a	Super Nó 002
BLASTX	http://gabriellacastro.com.br/SequenceAligningOntology#BLASTX	fbcfd463acfb449b018c6874df0fa2a2c931da1b	Super Nó 001, Super Nó 003
BLOSUM	http://gabriellacastro.com.br/SequenceAligningOntology#BLOSUM	8387865a7de8c3145fe73161d79d3719a79a0f92	Super Nó 004

Tabela 3.2: tabelas de espalhamento distribuídas em comunidades semânticas.

A Tabela 3.2 apresenta quatro comunidades semânticas com suas informações replicadas entre quatro super nós distintos. As chaves geradas utilizam o endereço na coluna Descrição Semântica.

[11] Na criptografia, SHA é uma função hash projetada pela agência de segurança dos Estados Unidos.

Como os super nós mantêm uma tabela com os outros super nós e as suas respectivas comunidades semânticas, um cientista que tiver interesse em se conectar a outra comunidade semântica deve fazer uma requisição a algum super nó conhecido, e então esse super nó se encarregara de procurar aquela comunidade semântica na sua tabela de comunidades semânticas.

Após selecionar uma comunidade semântica de acordo com a descrição semântica de seu interesse, um nó utiliza a coluna valor para se comunicar com os super nós responsáveis por aquela comunidade semântica. Caso tenha mais de um super nó gerenciando aquela comunidade científica, é selecionado um super nó aleatório e realizada a conexão para descoberta dos nós conectados naquela comunidade semântica.

Para armazenar os nós em comunidades semânticas, a e-ScienceNet utiliza outra DHT mais específica para aquela comunidade semântica replicada entre todos os nós daquela comunidade semântica. Nessa tabela temos uma chave que indica o serviço disponível por um componente daquela comunidade semântica. Os valores dessa tabela são relacionados aos cientistas que disponibilizam aquele componente naquela comunidade semântica.

Nome	**Tipo**	**Chave**	**Valor**
DNA_sequence	Dados	1ef6502c84f95bc6202a26bbb5d217ca5a4f5c98	Nó 001, Nó 002, Nó 003
RNA_sequence	Dados	daf4a03db66448212d9edf947356f319bd9b7ceb	Nó 001, Nó 004, Nó 005
sequence_aligning	Serviço	379b4fd19ad0df90254574e678a546882428b164	Nó 001, Nó 003
BLOSUM	Dados	02a7ebcb6237c146570d4cbe8e7d58297f3ab6a6	Nó 001, Nó 006

Tabela 3.3: tabelas de espalhamento distribuídas em componentes e nós[12].

[12] Por questões didáticas foram utilizados conceitos semanticamente similares em comunidades semânticas distintas, mas que poderiam facilmente se enquadrar em uma única comunidade semântica devido a sua semelhança.

É importante notar que, a coluna Valor da Tabela 3.3 deve conter valores que possibilitam a conexão a aquele nó, como o endereço de rede IPv4 ou IPv6, aporta de conexão TCP e o protocolo de aplicação utilizado pelo Gerente de Protocolo, ao qual é detalhado na seção oito deste capítulo. Além disso, tem-se uma nítida diferença entre as DHTs comumente utilizadas que armazenam referências a arquivos disponíveis na rede. A e-ScienceNet não armazena referências aos arquivos disponíveis, mas sim as comunidades semânticas disponíveis e os respectivos serviços e nós disponíveis nas comunidades semânticas.

3.3.3 Políticas de Conexões de Nós

Os super nós responsáveis por gerenciar as comunidades semânticas podem aplicar políticas as estas comunidades semânticas. Com a utilização de políticas para se conectar as comunidades semânticas, conseguimos limitar o acesso às informações disponíveis em uma comunidade semântica para somente os nós com acesso autenticado e seguro a e-ScienceNet.

Por exemplo, um centro de pesquisa, pode criar uma estrutura de compartilhamento entre os nós do seu laboratório utilizando a e-ScienceNet. Ao utilizar uma Ontologia personalizada para descrever as comunidades semânticas de interesse do laboratório, criando um super nó em sua rede local para gerenciar as suas comunidades semânticas, pode-se aplicar políticas para que outros nós se conectem a essa comunidade semântica, limitando o acesso aos dados e aumentando a segurança do compartilhamento das suas pesquisas.

As políticas de acesso ainda não estão efetivamente implementadas na e-ScienceNet, mas dentre as possíveis políticas para as comunidades semânticas podemos citar a autenticação de nós para acesso a comunidade semântica e o local físico de onde o nó está se conectando na e-ScienceNet.

As politicas podem ser estendidas de acordo com as necessidades dos nós e carregadas na e-ScienceNet modularmente pelo super nó. Por padrão, uma comunidade semântica não aplica nenhuma das políticas apresentadas, assim todos os nós de uma comunidade semântica têm a liberdade de compartilhar todos os seus dados e serviços com outros nós na mesma comunidade semântica.

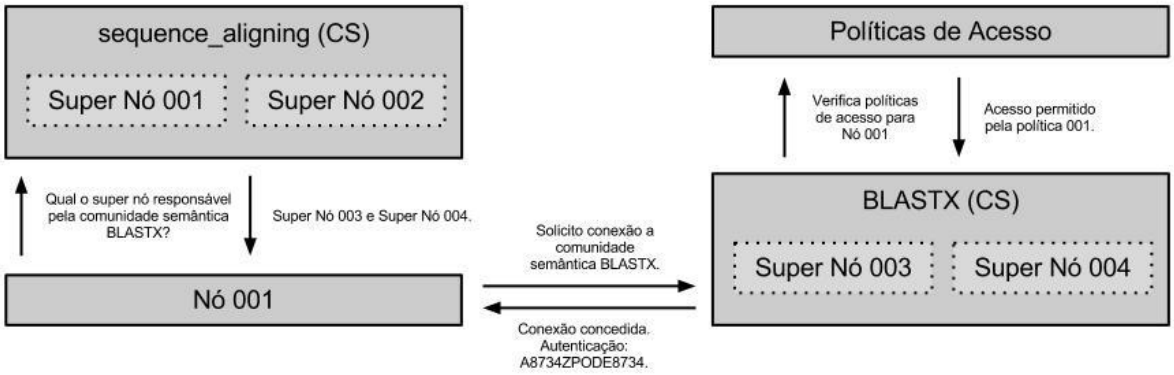

Figura 3.14: solicitação de conexão e as políticas de acesso

Na Figura 3.14 primeiramente é realizada a pesquisa de quais super nós são responsáveis pela comunidade semântica “BLASTX”. Com essa informação, o Nó 001 realiza a solicitação de conexão a comunidade semântica “BLASTX”. Ao receber essa solicitação, o super nó responsável pela comunidade semântica verifica se existe alguma política que libere o acesso do Nó 001 aquela comunidade semântica. Caso positivo, são repassadas ao Nó 001 as informações dos outros nós conectados aquela comunidade semântica e uma chave gerada aleatoriamente para que a conexão seja aceita nos outros nós.

Atualmente na e-ScienceNet um nó aceita conexões de qualquer local de rede, independente se o código para autenticação está correto ou não. Além disso, é verificada as políticas de acesso quando um super nó recebe uma conexão solicitando informações para conexão a uma comunidade semântica, mas nenhuma dessas políticas de segurança tem código retornando se é possível ou não sua conexão.

O Gerente de Rede gerencia diversas conexões e desconexões autônomas pelos nós na e-ScienceNet. Para manter sempre atualizada as informações das DHTs e otimizações propostas pela e-ScienceNet, utilizamos o conceito de eventos.

Os eventos são mensagens transmitidas entre os nós da e-ScienceNet indicando uma ação. Com o uso de eventos conseguimos manter as informações de índice da e-ScienceNet sempre atualizada. Dividimos os eventos em estruturados e não estruturados para caracterizar a transmissão dos eventos.

Os eventos estruturados são transmitidos entre os super nós na e-ScienceNet. Na Tabela 3.4 apresentamos alguns eventos estruturados.

Evento	Descrição
Adicionar Comunidade	Adiciona uma nova comunidade semântica à lista de comunidades semânticas disponíveis nos super nós da e-ScienceNet.
Adicionar Super Nó	Adiciona um novo super nó como parte da gerência de uma comunidade semântica.
Adicionar Serviço	Adiciona um novo serviço proposto por um componente relacionado à determinada comunidade semântica.
Adicionar Nó	Adiciona um nó aos provedores de determinado serviço em uma comunidade semântica.
Remover Comunidade	Remove uma comunidade semântica da lista de comunidades semânticas disponíveis nos super nós da e-ScienceNet.
Remover Super Nó	Remove um super nó como parte da gerência de uma comunidade semântica.
Remover Serviço	Remove um serviço proposto por um componente relacionado à determinada comunidade semântica.
Remover Nó	Remove um nó aos provedores de determinado serviço em uma comunidade semântica.
Adicionar Mapeamento	Adiciona um mapeamento a lista de mapeamentos realizados em determinada comunidade semântica.
Remover Mapeamento	Remove um mapeamento da lista de mapeamentos realizados em determinada comunidade semântica.
Mesclar Comunidade	Mescla duas ou mais comunidades semânticas em uma só, realizando a junção dos serviços e nós nas comunidades.

Tabela 3.4: lista de eventos estruturados da e-ScienceNet.

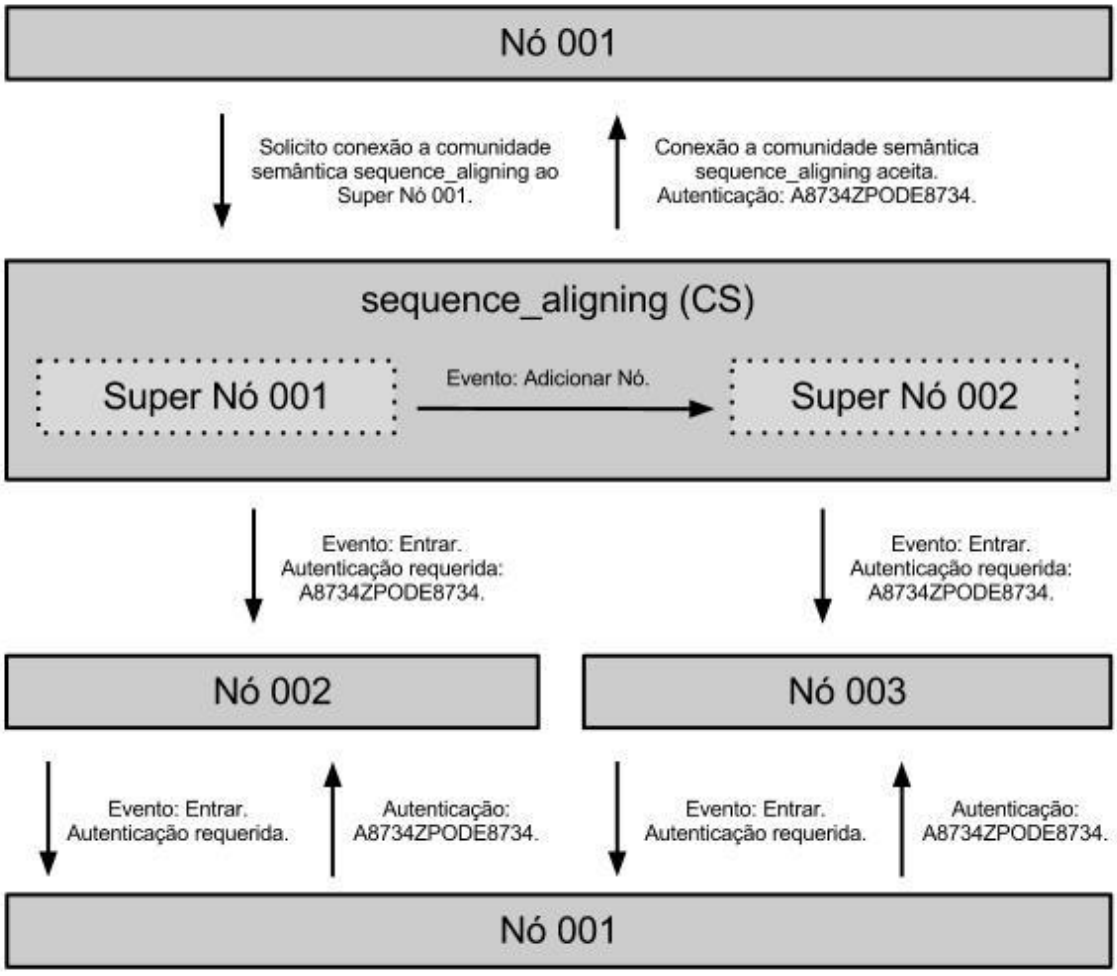

Figura 3.15: fluxo de eventos estruturados e não estruturados

Na Figura 3.15, mostramos a transmissão dos eventos ocorridos quando um cientista solicita uma conexão a uma comunidade semântica. É realizada uma solicitação de conexão ao super nó daquela comunidade semântica. O super nó que recebeu e aceitou a conexão daquele novo nó envia um evento estruturado para os outros super nós daquela comunidade semântica. Como essa é uma comunidade restrita, é enviada para os nós uma autenticação no formato de evento não estruturado para que os nós aceitem a conexão daquele nó somente com a autenticação informada.

Os eventos não estruturados são transmitidos entre qualquer nó na e-ScienceNet. Na Tabela 3.5 apresentamos alguns eventos não estruturados.

Evento	Descrição
Entrar	Responde a uma solicitação de entrada na comunidade semântica.
Sair	Fecha a conexão a uma comunidade semântica.
Sucesso	Evento genérico indicando que algo ocorreu com sucesso. Podemos citar como exemplo a aprovação de um nó em determinada comunidade semântica ou a aprovação de mais um super nó como gerente de determinada comunidade semântica.
Falha	Evento genérico indicando que algo falhou durante a sua execução. Podemos citar como exemplo a negação de um nó em determinada comunidade semântica ou a negação de mais um super nó como gerente de determinada comunidade semântica.

Tabela 3.5: lista de eventos não estruturados da e-ScienceNet.

Os eventos não estruturados podem ser enviados por componentes da e-ScienceNet. Os componentes da e-ScienceNet têm toda a liberdade de criar e enviar os seus próprios eventos não estruturados para outros nós na e-ScienceNet. Com o envio de eventos pelos componentes conseguimos, por exemplo, criar um componente que notifique outros nós na e-ScienceNet quando tiver novas informações disponíveis para determinada comunidade semântica. Além disso, eventos não estruturados são a primeira etapa para a possibilidade de criação de um Workflow Científico estruturado no contexto da e-ScienceNet.

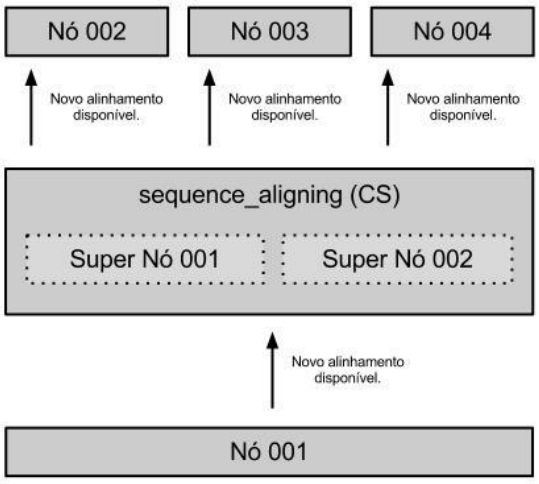

Figura 3.16: fluxo de eventos enviados por componentes na rede

Na Figura 3.16 temos uma demonstração de como funcionaria os eventos disparados na rede por componente. Considerando a Figura 3.16 temos o Nó 001 que acabou de realizar um alinhamento de uma nova sequência pelo componente de alinhamento de sequências. Para repassar as informações aos outros nós da rede indicando que o Nó 001 tem novos dados de alinhamento disponíveis, o componente envia um evento para os outros nós da rede indicando o ocorrido.

É importante ressaltar que esses eventos tem um código único de identificação que redireciona o evento para o componente correto, sendo assim, cada componente quando criar um novo evento deve criar uma código único igual a todos os clientes. Ao utilizar uma função como o SHA se baseando no nome completo da classe do componente, conseguimos criar uma identificação única para cada evento de acordo com o componente.

3.4 Gerente de Semântica

O Gerente de Semântica é responsável por prover as características semânticas da e-ScienceNet. Podemos dizer que o Gerente de Semântica é uma interface que abstrai o acesso dos outros gerentes da e-ScienceNet às APIs e bibliotecas relacionadas a Web Semântica.

O Gerente de Semântica também é responsável por gerenciar as Ontologias utilizadas pelos cientistas, além do compartilhamento e mapeamento dessas Ontologias.

No contexto da e-ScienceNet, os nós participantes estão geograficamente dispersos, apresentando interesses e especialidades distintas. Dessa forma, devemos manter a liberdade destes cientistas com relação aos seus interesses e especialidades durante o processo de criação e manutenção das comunidades semânticas. Utilizando Ontologias para descrever as comunidades semânticas, conseguimos obter uma estruturação das comunidades e fácil acesso ao conhecimento descritivo provido por cada uma dessas Ontologias [24].

A escolha de quais Ontologias serão utilizadas para descrever os interesses e especialidades dos cientistas é muito importante, já que interfere diretamente nas comunidades semânticas e dados disponíveis na e-ScienceNet para aquele cientista. Além

disso, estudos mostram [36] que a seleção aleatória e não criteriosa de uma hierarquia de classificação pode reduzir, ou até mesmo limitar os benefícios do uso de redes sobrepostas semânticas em Redes Ponto a Ponto [36].

Um benefício adicional de Ontologias comparado com hierarquias de classificação simples é que com o uso de Ontologias conseguimos capturar conhecimento sobre comunidades semânticas que não estão explícitos na Ontologia através de regras de inferência. Com o poder de expressividade semântica das Ontologias, conseguimos representar um domínio de pesquisa científica em um sistema axiomático, inferindo novas comunidades semânticas baseadas em suas propriedades.

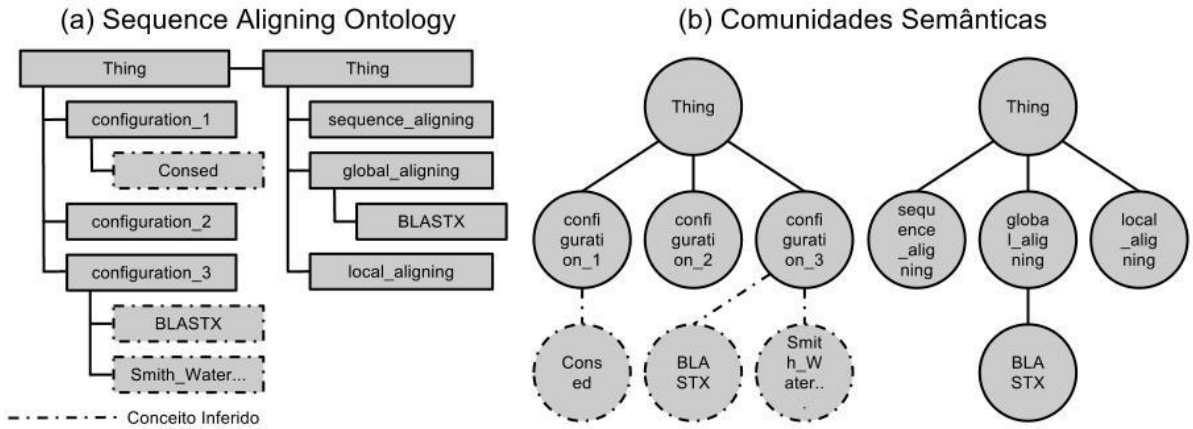

Figura 3.17: comunidades semânticas e Ontologias

Na Figura 3.17 mostramos duas hierarquias extraídas da Ontologia Sequence Aligning Ontology. Podemos ver que a comunidade “BLASTX” está ligada em duas comunidades semânticas distintas. Isso se dá pelo fato, que através da realização de inferências, foi descoberta a relação semântica da comunidade “BLASTX” com a comunidade “configuration_3”. Sendo assim, ao realizar a expansão de uma pesquisa, ou seja, enviar uma pesquisa para nós em outras comunidades semânticas, aumentamos a chance dos resultados retornados serem semanticamente similares. A expansão das pesquisas é detalhada mais a frente no Gerente de Pesquisa.

O uso de Ontologias para descrever as comunidades semânticas têm benefícios, mas também traz desafios quando utilizadas em Sistemas Distribuídos. Um dos principais desafios

é o armazenamento das Ontologias de forma que todos os cientistas na e-ScienceNet tenham suas comunidades semanticamente interoperáveis com outras comunidades. Para solucionar o problema de distribuição das Ontologias pelos cientistas na e-ScienceNet, mostramos o uso de Ontologias centrais e Ontologias locais como base da proposta de uso de Ontologias compartilhadas na e-ScienceNet.

3.4.1 Ontologias compartilhadas

Uma Ontologia central leva em consideração os esforços realizados principalmente na década de 90, onde eram propostos padrões para definirem Ontologias de topo ou estabelecer repositórios públicos para favorecer o reuso do conhecimento [82]. No entanto, a utilização de Ontologias centrais limita a liberdade de cientistas de descreverem as suas comunidades semânticas e insere o problema de centralização de um recurso. Se a Ontologia central fica indisponível, não é possível criar novas comunidades semânticas, sendo o uso da e-ScienceNet limitado às comunidades pré-existentes.

Uma Ontologia local é armazenada localmente pelos nós na e-ScienceNet. Com a utilização de uma Ontologia local para cada nó, podemos resolver o problema de centralização, e ainda conseguimos manter a liberdade dos cientistas descreverem as suas comunidades semânticas de forma adequada segundo seus interesses e especialidades.

O problema de utilizar uma Ontologia local é a diversidade semântica entre os nós conectados na e-ScienceNet, dificultando que as informações das comunidades semânticas sejam distribuídas entre ambientes que devem ser semanticamente interoperáveis [54]. Outro problema do uso de Ontologias locais é a sincronização de alterações realizadas em uma Ontologia por qualquer nó, onde cada alteração na Ontologia deve ser propagada para todos os outros nós na e-ScienceNet.

Uma Ontologia compartilhada consiste no uso de Ontologias locais com técnicas de mapeamento para realizar a junção das Ontologias locais disponibilizadas pelos nós, criando assim uma Ontologia compartilhada entre os nós da comunidade semântica. Dessa forma, mantemos as comunidades semânticas interoperáveis entre os nós da e-ScienceNet em diferentes comunidades semânticas.

Em um ambiente distribuído é complexo manter uma Ontologia compartilhada pela dificuldade de negociar uma Ontologia que resolva as necessidades de ambas as partes envolvidas, além de ser uma tarefa custosa manter uma Ontologia compartilhada em um ambiente extremamente dinâmico [83]. No entanto, considerando o contexto de redes de pesquisa científica, onde os pesquisadores conectados compartilham informações em áreas de aplicação específicas, esta complexidade pode ser reduzida, uma vez que podemos contar com a existência de Ontologias largamente aceitas pelas comunidades científicas e portanto mais facilmente compartilháveis pelos cientistas. Dessa forma, fazemos o uso desta técnica. Como a e-ScienceNet utiliza as Ontologias locais dos nós para criar uma Ontologia que seja compartilhada na rede, temos uma Ontologia compartilhada que atenda as necessidades de ambas as partes envolvidas. Para transformar as Ontologias locais em Ontologias que possam ser compartilhadas entre os nós, a e-ScienceNet utiliza técnicas de mapeamento entre as Ontologias locais de cada nó para descobrir a similaridade entre os conceitos definidos nas Ontologias distintas.

Na e-ScienceNet um cientista pode ter interesses em diversas Ontologias distintas. Atualmente a e-ScienceNet considera as Ontologias de interesse de um dado cientista e realiza um mapeamento entre essas Ontologias, criando assim a Ontologia que será utilizada por aquele nó na rede.

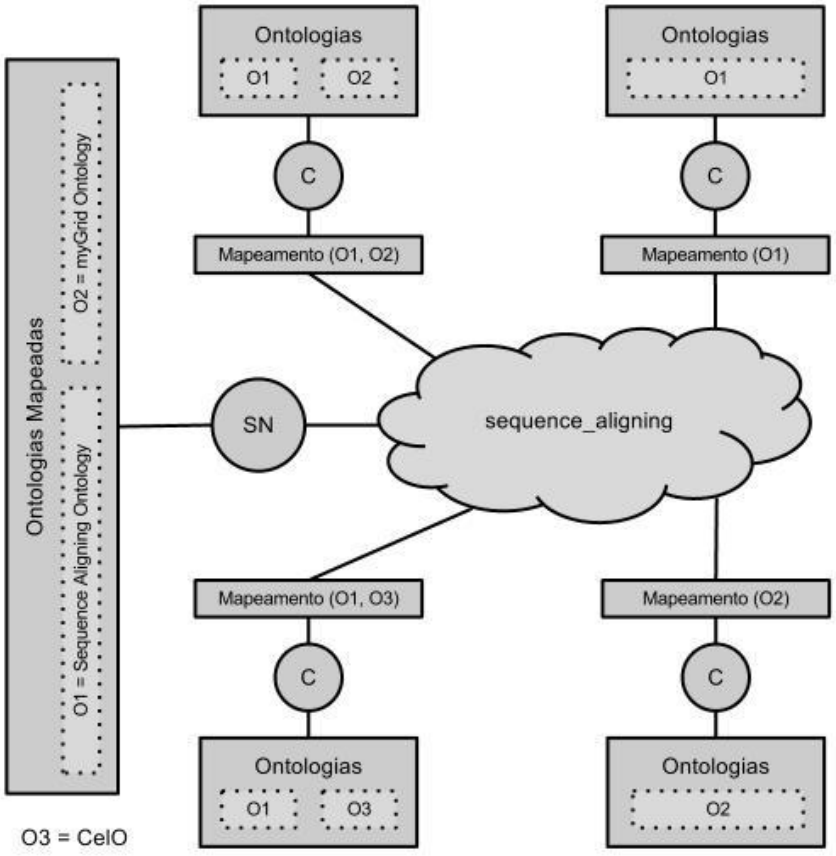

Figura 3.18: Ontologias locais e Ontologias compartilhadas na e-ScienceNet

Como mostrado na Figura 3.18, temos um exemplo de quatro cientistas em uma mesma comunidade semântica. Cada um desses cientistas tem as suas Ontologias locais, sendo O1 a Ontologia Sequence Aligning Ontology, a O2 representada por MyGridOntology[13] e O3 como CelO[14]. O super nó da comunidade semântica é o detentor de uma Ontologia compartilhada com o mapeamento entre todas as Ontologias de todos os cientistas naquela comunidade semântica.

Essa abordagem apresenta alguns problemas, como, o mapeamento entre Ontologias semanticamente não correlatas, onde uma Ontologia mapeada pode ficar extensa caso existam muitos nós com Ontologias diferentes em determinada comunidade semântica, e

[13] A MyGridOntology descreve o domínio de pesquisa da Bioinformática e a dimensão ao qual um serviço possa ser caracterizado pela perspectiva do cientista. Disponível em http://www.mygrid.org.uk/tools/service-management/mygrid-ontology/.

[14] A Ontologia CelO captura a estrutura de um modelo celular e as propriedades de componentes funcionais. É utilizada para descrever, pesquisar e compor modelos CellML, melhorando o reuso e composição de componentes existentes e permitindo a validação de novos modelos.

consequentemente, um aumento no processamento e comunicação entre os nós. Na Figura 3.18, o problema de mapeamento por Ontologias não correlatas é apresentado com a Ontologia O3, ao qual é utilizada por um cientista que está naquela comunidade, mas não tem relação semântica com a comunidade, sendo mapeada desnecessariamente por aquela comunidade semântica.

Em versões futuras da e-ScienceNet, a melhor abordagem seria manter uma matriz indicando se uma Ontologia tem relação semântica com outras Ontologias utilizadas pelos nós. Cada linha e coluna dessa matriz representaria uma Ontologia, sendo os valores dos elementos a similaridade entre as duas Ontologias. No entanto, considerando os domínios de interesse dos cientista, que no geral são correlatos, a não correlação semântica entre as Ontologias é uma exceção a regra geral na e-ScienceNet.

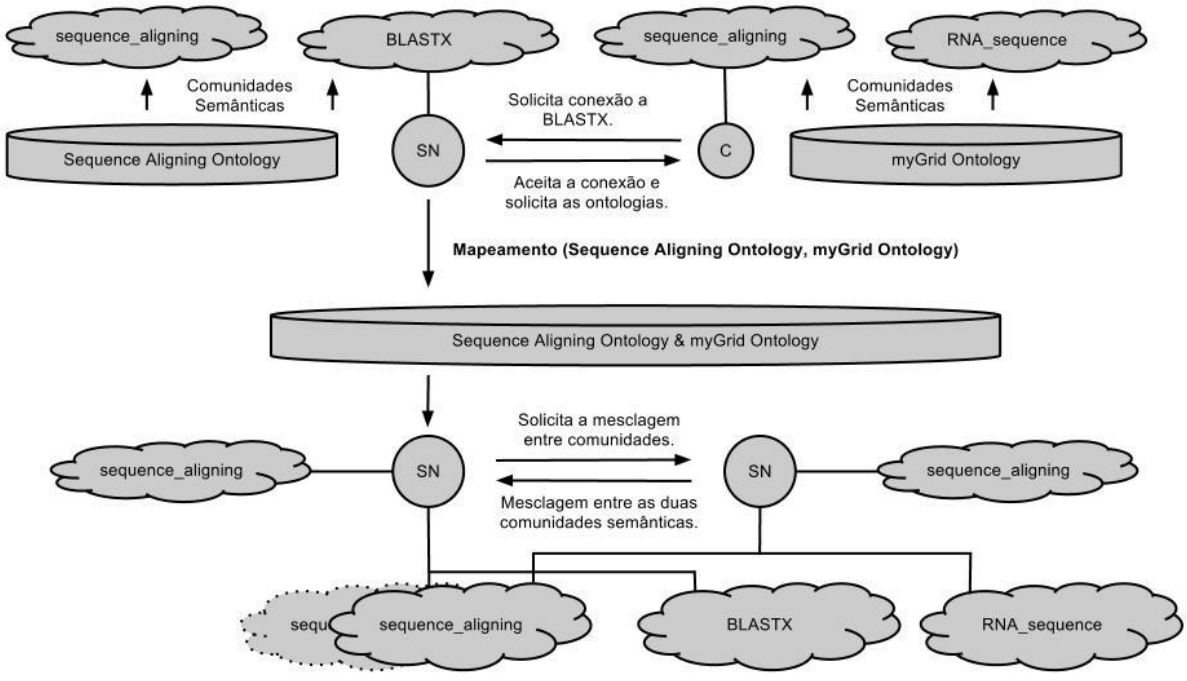

Figura 3.19: comunidades semânticas e Ontologias mapeadas

Além da descoberta de novas comunidades semânticas com o uso de Ontologias, o mapeamento entre as Ontologias proporciona a descoberta de comunidades semânticas similares definidas por outras Ontologias. Na Figura 3.19, podemos ver que existem quatro comunidades semânticas criadas a partir de duas Ontologias distintas. Quando um nó (C)

solicitar a conexão a uma comunidade semântica descrita por outra Ontologia através do seu super nó (SN), as Ontologias relacionadas ao nó são enviadas ao super nó para que seja realizado o mapeamento entre as Ontologias. Após realizar o mapeamento e descobrir a similaridade entre os elementos das Ontologias, o super nó verifica se existe alguma comunidade semântica criada de elementos similares. Caso aquela não seja uma comunidade semântica privada, é realizada uma operação de merge entre as duas comunidades semânticas distintas, possibilitando que os nós em ambas as comunidades compartilhem suas informações.

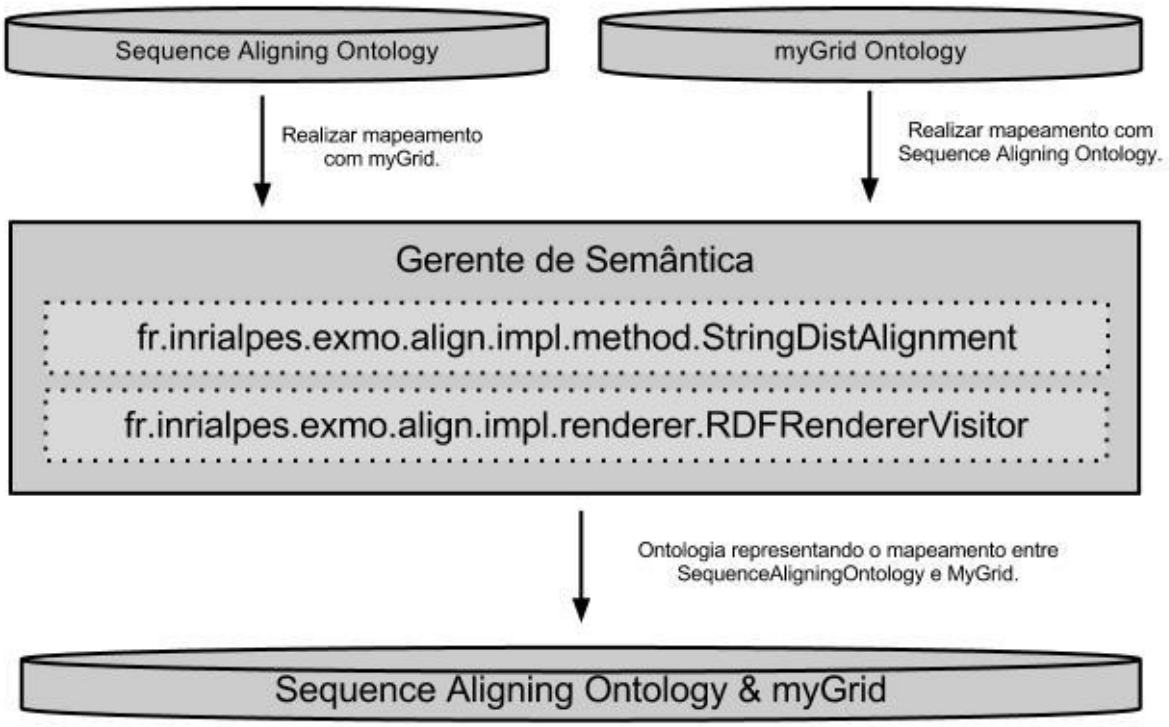

Figura 3.20: Alignment API na e-ScienceNet

Assim como mostrado na Figura 3.20, todo o mapeamento na e-ScienceNet é realizado através do método de StringDistAlignment da Alignment API [84]. Este método utiliza a distância de Levenshtein para medir a similaridades entre os termos, onde é considerado o mínimo de edições necessárias para transformar uma string em outra. Em versões futuras este método será substituído, sendo que foi adotado na primeira versão com o intuito de simplificar a abordagem.

Além do compartilhamento de dados na e-ScienceNet, podemos compartilhar os mapeamentos existentes entre as Ontologias através do super nó. Na e-ScienceNet existem diversos mapeamentos de Ontologias com suas similaridades entre os conceitos já calculados. Eventualmente, depois que um mapeamento foi calculado, podemos salvá-lo localmente e compartilhar com outros cientistas na rede [85]. Todos os cientistas conectados na e-ScienceNet podem disponibilizar os diversos mapeamentos calculados e armazenados com os outros cientistas conectados na e-ScienceNet.

Com o compartilhamento dos diversos mapeamentos existentes, podemos utilizar técnicas para avaliar a qualidade geral daqueles mapeamentos. O trabalho [86] utiliza técnicas de "fofoca na rede" para tentar descobrir erros e avaliar a qualidade dos mapeamentos que estão sendo compartilhados em uma Rede Ponto a Ponto.

O reuso e compartilhamento de mapeamentos existentes em conjunto com as comunidades semânticas aumenta a eficácia das pesquisas por esses mapeamentos. Ao carregar os interesses de um cientista em determinado domínio de pesquisa, e fazer a pesquisa nas comunidades semânticas daquele domínio, a chance de que a Ontologia do cientista já esteja mapeada entre as Ontologias disponíveis na comunidade semântica é muito maior que se fosse realizada uma pesquisa de forma aleatória entre os cientistas conectados na e-ScienceNet.

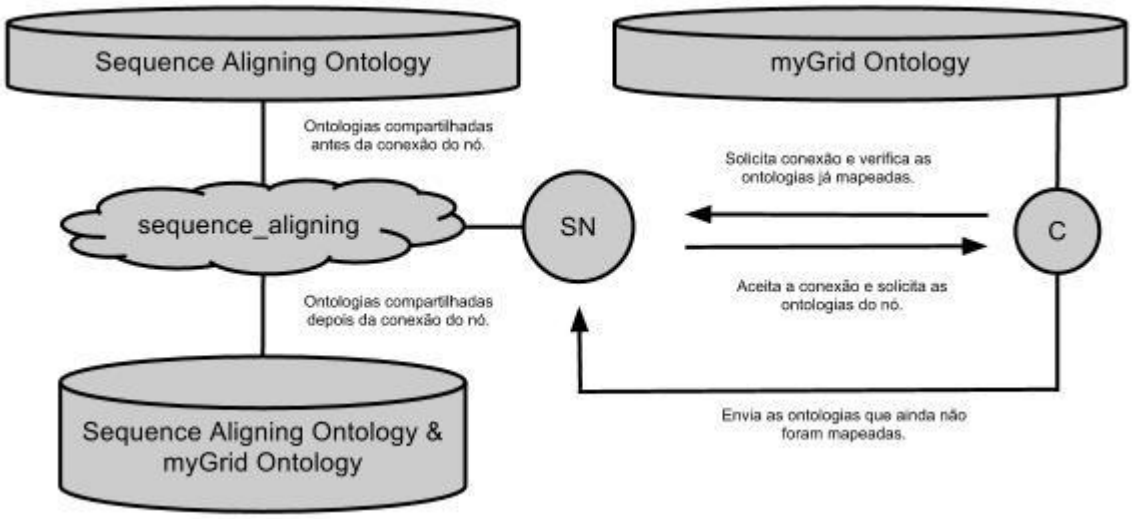

Figura 3.21: compartilhamento de mapeamentos em comunidades semânticas

Na Figura 3.21, mostramos a conexão de um nó a uma comunidade semântica. Primeiramente aquela comunidade semântica tem somente uma Ontologia descrevendo a comunidade. Depois que um nó se conecta, tendo uma de suas Ontologias não mapeada na comunidade semântica, o nó envia ao super nó da comunidade para que seja realizado o mapeamento. Após realizado o mapeamento, as duas Ontologias serão compartilhadas entre os nós daquela comunidade semântica.

3.4.2 Otimização de compartilhamento e mapeamento de Ontologias

Algumas técnicas para mapeamento em Redes Ponto a Ponto podem ser aplicadas em versões futuras da e-ScienceNet, otimizando o mapeamento e compartilhamento de Ontologias. Dentre essas técnicas podemos citar o poder de processamento distribuído de outros nós para ajudar durante o processo de mapeamento das Ontologias e o mapeamento incompleto dessas Ontologias.

O mapeamento incompleto de Ontologias consiste em realizar o mapeamento somente de partes da Ontologia que mais interessar ao cientista.

Os cientistas na e-ScienceNet podem utilizar Ontologias descrevendo vários conceitos, sendo possível que muitos desses conceitos não tenham influência direta nas comunidades semânticas e nas pesquisas à base de conhecimento da e-ScienceNet. Utilizando mapeamento incompleto de Ontologias, conseguimos filtrar partes das Ontologias e obter somente os conceitos utilizados direta ou indiretamente pelo cientista.

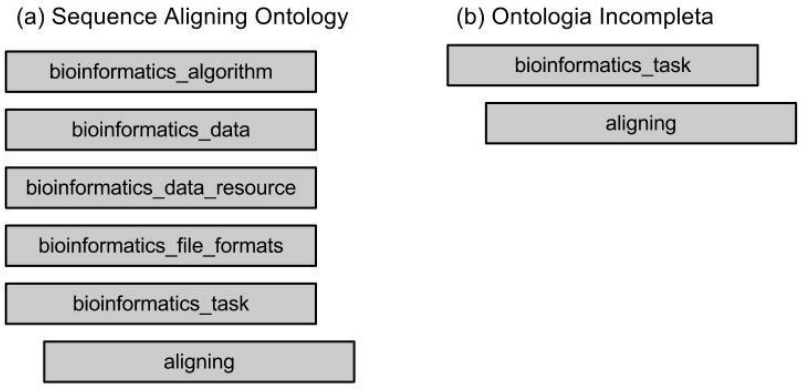

Figura 3.22: mapeamentos incompletos por interesse e especialidade

Considerando a Figura 3.22 (a), mostramos alguns termos na Sequence Aligning Ontology. Dentre esses termos, consideramos que o cientista tenha interesse somente em “aligning”. Dessa forma, quando forem realizado os mapeamentos para aquele cientista, seria realizado somente o mapeamento dos termos relacionados a “aligning”, assim como mostrado na Figura 3.22 (b).

Utilizando mapeamento incompleto de Ontologia podemos criar um mapeamento contínuo, onde cada parte interessada realiza o mapeamento de forma incremental a partir dos anteriores já realizados [85]. Dessa forma, quando um cientista em uma comunidade semântica realizar o mapeamento incompleto de uma Ontologia, outros cientistas com interesses similares definidos na mesma Ontologia, podem continuar a mapear alguns conceitos que lhes interessem.

O mapeamento incompleto de Ontologia pode ser utilizado pelo próprio cientista que iniciou o mapeamento incompleto, mesmo que não tenha interesse total nos conceitos definidos por aquela Ontologia. Depois de realizar pesquisas e se conectar nas comunidades semânticas de seu interesse, o dispositivo utilizado pelo cientista pode continuar o mapeamento incompleto da Ontologia de forma contínua.

Muitos cientistas conectados a e-ScienceNet podem ter o seu dispositivo ocioso em determinados momentos. A e-ScienceNet pode aproveitar o poder de processamento computacional dos dispositivos utilizados pelos cientistas para auxiliar no processo de mapeamento das Ontologias incompletas utilizadas pela e-ScienceNet.

Outra técnica a ser utilizada em versões futuras da e-ScienceNet é o mapeamento distribuído de Ontologias, que consiste em dividir uma Ontologia em partes menores distintas, similar ao mapeamento incompleto. Sendo assim, todas as partes interessadas em uma comunidade semântica podem disponibilizar seu poder de processamento computacional para auxiliar o processo de mapeamento total da Ontologia referida.

3.5 Gerente de Interesse

O Gerente de Interesse é responsável por selecionar as comunidades semânticas a serem conectadas pelo Gerente de Rede. Com isso o Gerente de Interesse fica responsável por analisar as pesquisas dos cientistas, a semântica dos dados que deseja compartilhar e os interesses e especialidades informadas diretamente pelo cientista.

Para avaliar e propor as comunidades semânticas de acordo com os interesses, componentes e dados a serem disponibilizados pelo cientista na e-ScienceNet, o Gerente de Interesse utiliza informações para avaliar a adequabilidade de determinada comunidade semântica. Dentre essas informações, podemos citar os interesses informados diretamente pelo cientista, os serviços dos componentes disponíveis em uma comunidade similar aos disponíveis localmente pelo cientista e a análise das pesquisas realizadas pelo cientista na e-ScienceNet. Atualmente a e-ScienceNet repassa ao Gerente de Rede somente as comunidades explicitamente informadas pelos cientistas.

Para adicionar um interesse, o cientista seleciona uma Ontologia já previamente carregada na rede, e a partir dessa Ontologia o cientista pode selecionar os elementos ontológicos de interesse. Assim que o cientista seleciona os elementos daquela Ontologia, esses elementos são armazenados como interesse daquele cientista nas configurações.

Com o uso de Ontologias podemos inferir novos conhecimentos e propor ao Gerente de Rede outras comunidades semânticas que possam conter informações relevantes de acordo com o interesse do cientista. Por exemplo, em uma Ontologia podemos realizar pesquisas em comunidades semânticas utilizando as relações semânticas entre os elementos da Ontologia.

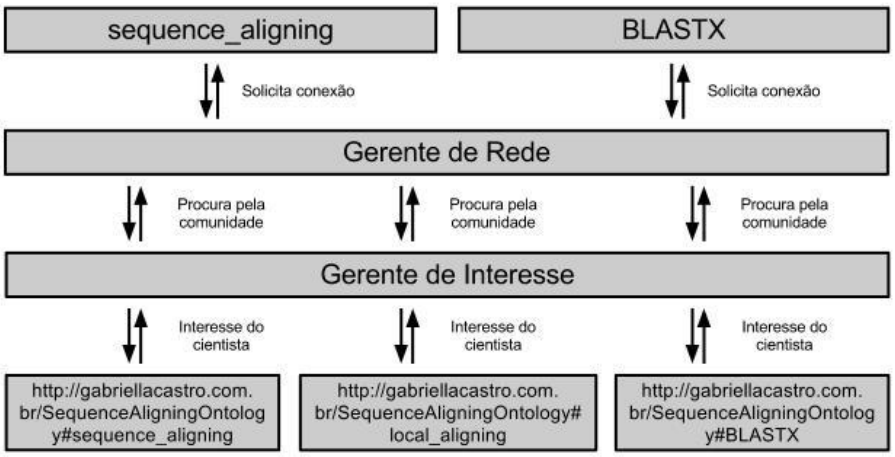

Figura 3.23: Gerente de Interesse

Na Figura 3.23 temos os interesses do cientista carregados a partir das configurações. Depois de serem carregados os interesses e realizado o filtro nas Ontologias, é repassado ao Gerente de Rede para que sejam solicitadas as conexões as comunidades semânticas. Pela Figura 3.23 podemos ver que existem três interesses, mas somente duas comunidades semânticas. Isso acontece devido ao fato de o interesse “local_aligning” ser subclasse de “sequence_aligning”, propondo nesse caso em específico a conexão a uma comunidade semântica superior, já que de acordo com a Ontologia os dois termos tem uma relação semântica de pai e filho. Se fosse realizada a criação de mais uma comunidade semântica somente para “local_aligning”, teríamos inicialmente somente um nó nessa comunidade, não sendo interessante nesse momento realizar essa operação de criação de comunidade semântica.

Como a e-ScienceNet é baseada em componentes de softwares, as informações indicando as melhores comunidades semânticas podem ser desenvolvidas e carregadas modularmente na rede. Os componentes podem prover informações ao Gerente de Interesse que permitam uma seleção mais acurada de quais são as comunidades semânticas mais adequadas de acordo com suas informações específicas.

Por exemplo, um componente que analise o perfil de cientistas em bases de dados cientificas, pode obter informações para auxiliar o Gerente de Interesse a selecionar quais são as comunidades semânticas mais adequadas. Outro exemplo possível neste contexto seria o

uso de um banco de dados com publicações, como, Association for Computing Machinery[15] ou IEEE Xplore [16] que poderiam ser analisados para achar publicações dos cientistas e conectá-los nas comunidades semânticas referentes aos seus trabalhos. Outra possibilidade é um componente poder analisar localmente os arquivos disponíveis no dispositivo do cientista e propor ao Gerente de Interesse determinadas comunidades semânticas onde o cientista tem conteúdo a disponibilizar. Alguns tipos de arquivo como textos e artigos podem ser analisados de forma a obter sua relevância semântica. A classificação dos documentos pode ser feita de forma manual ou automática. Exemplos de classificadores automáticos são os de correspondência textual [87], Redes Bayesiana [88] e algoritmos de clusterização [73].

As pesquisas realizadas pelos cientistas podem nos fornecer informações sobre qual assunto o cientista está trabalhando no momento. Dessa forma conseguimos que nas próximas conexões do cientista aquelas comunidades semânticas se conectem automaticamente. Em [89], por exemplo, é realizada uma analise das ultimas pesquisas dos usuários para criar um perfil de uso. No entanto, estas funcionalidades ainda não estão disponíveis na versão atual da e-ScienceNet.

3.6 Gerente de Pesquisa

O Gerente de Pesquisa é responsável por analisar e realizar o roteamento de pesquisas para as comunidades semânticas relacionadas àquela solicitação.

A maioria das Redes Ponto a Ponto atuais utilizam pesquisas por palavra chave. Na e-ScienceNet utilizamos recursos da Web Semântica (Ontologias) para criar as comunidades semânticas, sendo assim, dispomos da descrição semântica das comunidades que permitem realizar pesquisas semânticas na rede. Este é o principal diferencial da e-ScienceNet em relação a propostas similares descritas na literatura. Com o uso da semântica, as pesquisas são mais focadas e direcionadas para as comunidades semânticas relevantes no contexto da referida pesquisa.

[15] http://www.acm.org/
[16] http://ieeexplore.ieee.org/

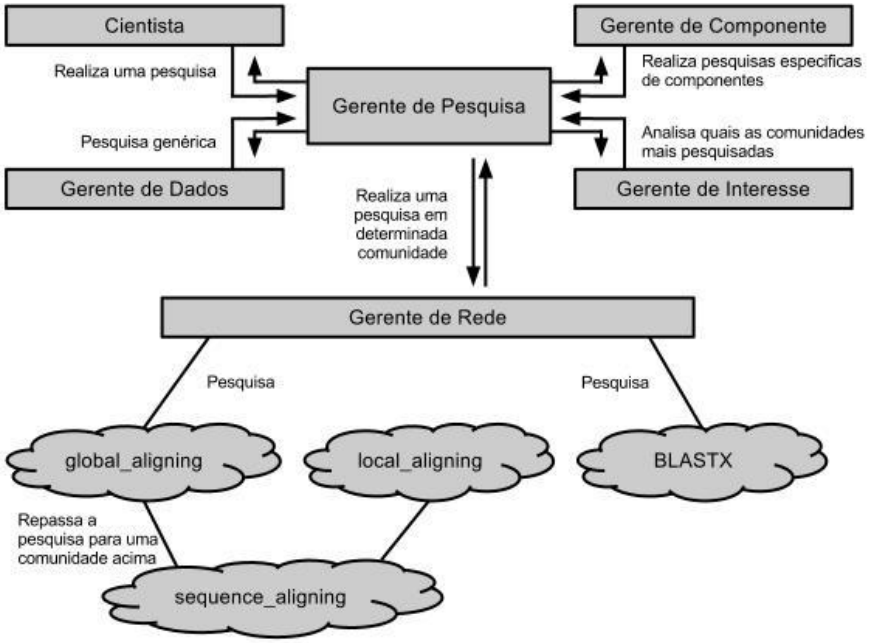

Figura 3.24: Gerente de Pesquisa

Na Figura 3.24 mostramos a interação entre o Gerente de Pesquisa com o Gerente de Rede. O cientista está conectado em duas comunidades semânticas, a "global_aligning" e a "BLASTX". Considerando as relações semânticas especificadas na Ontologia Sequence Aligning Ontology, a comunidade "global_aligning" tem relação semântica com a comunidade "sequence_aligning", através de uma relação de subClassOf. Neste exemplo, a relação ontológica já estava explicitada na Ontologia, bastando um caminhamento na relação semântica da ontologia, justificando assim o repasse da pesquisa para outras comunidade semânticas. No entanto relações mais complexas podem ser utilizadas, como relações não explícitas na ontologia, ou no mapeamento entre ontologias, onde as restrições nas relações (propriedades) e/ou o uso de máquinas de inferência e regras em SWRL [90] podem auxiliar.

Assim, considerando ainda o exemplo da Figura 3.24, quando um cientista realizar uma pesquisa, o Gerente de Pesquisa solicita ao Gerente de Rede que repasse a solicitação a todos os nós na comunidade semântica "global_aligning". Dessa forma, se a pesquisa estiver configurada para ser repassada as comunidades semânticas relacionadas, temos uma expansão da pesquisa, enviando para mais nós em outras comunidades com possíveis conteúdos semanticamente relacionados, como é o caso da "global_aligning" com "sequence_aligning" mostrado na Figura 3.25.

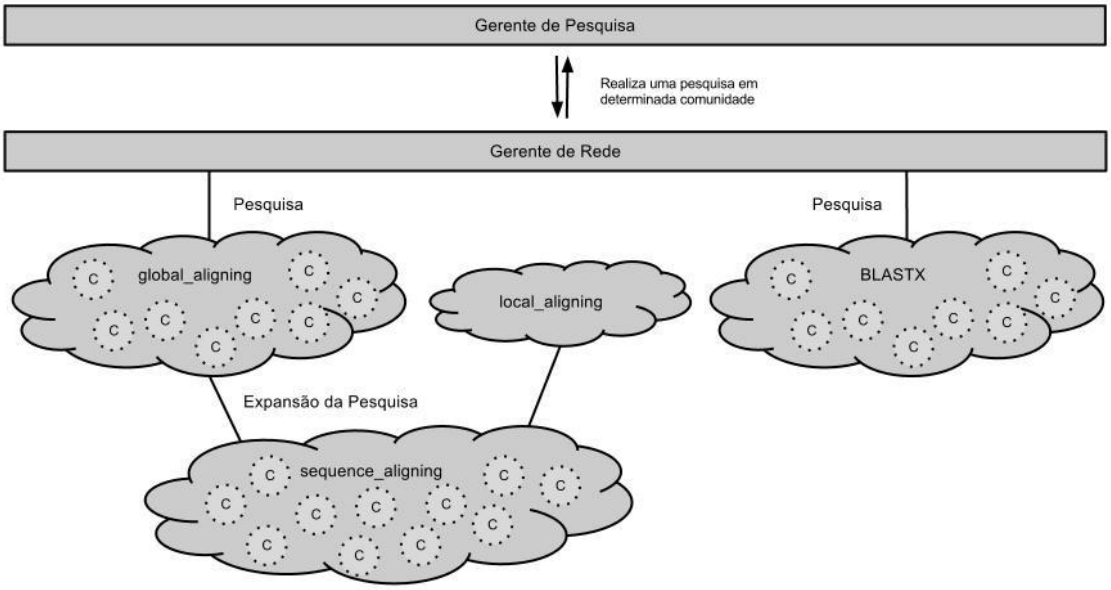

Figura 3.25: propagação de pesquisas

Outra possibilidade de pesquisa semântica é a possibilidade de o usuário realizar uma pesquisa por texto livre. Ao analisar o texto de busca disponibilizado pelo cientista, realiza-se uma filtragem por palavras chaves e uma associação destas palavras com os termos ontológicos, sendo possível assim selecionar quais comunidades semânticas melhor descrevem sua pesquisa. Após descobrir quais comunidades semânticas são relacionadas aquela pesquisa, o Gerente de Pesquisa solicita ao Gerente de Rede que se conecte as comunidades semânticas que ainda não estão conectadas. Enquanto é realizada a conexão a essas comunidades semânticas a pesquisa do cientista é repassada para as comunidades semânticas que já estão conectadas.

Com a pesquisa semântica, temos também a diferenciação dos resultados de acordo com o contexto do cientista. Tipicamente, dois cientistas diferentes com o mesmo texto de pesquisa vão receber os mesmos resultados. Com uma pesquisa semântica, são avaliados os interesses daquele cientista e são retornados resultados somente naquele contexto.

Na Figura 3.24 temos também a interação com o Gerente de Interesse. As pesquisas realizadas são analisadas pelo Gerente de Interesse para descobrir quais comunidades semânticas são mais relevantes, ou seja, onde a pesquisa retornou mais informações

relevantes de acordo com a pesquisa do cientista. Dessa forma, em versões futuras da e-ScienceNet conseguimos manter o cientista conectado somente nas comunidades que tem valor semântico para suas pesquisas, economizando assim recursos do dispositivo utilizado. Se os nós de uma comunidade semântica estão retornando poucos dados para as pesquisas realizadas, talvez a alocação dos seus recursos a aquela comunidade semântica não esteja sendo tão bem aproveitada.

A interação do Gerente de Pesquisa com o Gerente de Componente se da através de pesquisas específicas a certos tipos de informações. Essa interação é detalhada na próxima seção. Quando os nós recebem a pesquisa de outros nós na comunidade semântica, é realizada a pesquisa em seu Gerente de Dados local, retornando as informações que tem relação semântica com a pesquisa realizada.

Geralmente a propagação das pesquisas é de acordo com o escopo e os parâmetros daquela pesquisa. Na e-ScienceNet a propagação das pesquisas é gerenciada através do caminho da pesquisa. O caminho da pesquisa indica por quais nós determinada pesquisa já passou. Dessa forma, se um nó que se encontra no caminho da pesquisa estiver na lista de nós conhecidos, essa pesquisa não deve ser repassada a aquele nó. Além do caminho da pesquisa, temos também a quantidade máxima de saltos de uma pesquisa, onde é configurado um valor inteiro, e a cada nó da rede esse valor é decrementado até chegar à zero, onde a pesquisa para de ser transmitida.

Depois que a pesquisa é realizada e as informações relevantes são retornadas ao Gerente de Pesquisa, o mesmo realiza uma filtragem dos resultados removendo informações duplicadas que foram retornadas por mais de um nó. Além disso, é criado um filtro com a quantidade de dados retornados, diminuindo a comunicação de dados na rede.

Nas próximas vezes que uma pesquisa similar é realizada, primeiramente é verificado um cache local com os resultados das pesquisas anteriores, mostrando quase que instantaneamente resultados obtidos anteriormente enquanto a pesquisa na rede é realizada.

3.7 Gerente de Dados

O Gerente de Dados é responsável por acessar fisicamente os dados/informações de componentes disponíveis nos nós da rede. Utilizando um Gerente de Dados conseguimos abstrair o acesso a esses dados, podendo disponibilizar informações na rede independente da sua fonte. Na e-ScienceNet existem duas formas de disponibilizar os dados disponíveis aos nós. A primeira é de forma genérica, já a segunda utiliza uma lógica específica do componente em questão.

A pesquisa genérica define o seu próprio modelo de dados genérico, não se preocupando com a lógica do componente utilizado. A lógica de trabalho de cada componente varia muito de acordo com o objetivo de cada um, sendo complexo estabelecer um modelo de dados funcional as necessidades dos diversos componentes distintos. O modelo genérico utiliza como formato de arquivo RDF e permite que a pesquisa considere a descrição semântica das informações.

Para realizar a pesquisa do tipo genérica, a e-ScienceNet utiliza SPARQL [91]. SPARQL é um padrão já estabelecido em diversas ferramentas e proposto pelo W3C.

```
<?xml version='1.0'?>

<rdf
        xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
        xmlns:generic="http://localhost:8080/generic.rdf">

        <rdf:Description rdf:ID="name">
                <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
                <rdfs:label>Resource Name</rdfs:label>
        </rdf:Description>

        <rdf:Description rdf:ID="description">
                <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
                <rdfs:label>Resource Description</rdfs:label>
        </rdf:Description>

        <rdf:Description rdf:ID="type">
                <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
                <rdfs:label>Resource Type</rdfs:label>
        </rdf:Description>

        <rdf:Description rdf:ID="component">
                <rdf:type rdf:resource="http://www.w3.org/1999/02/22-rdf-syntax-ns#Property"/>
                <rdfs:label>Component Name</rdfs:label>
        </rdf:Description>
</rdf>
```

Figura 3.26: RDF para pesquisas genéricas

Na Figura 3.26 temos o RDF utilizado para representar os dados que podem ser utilizados em uma pesquisa genérica. Nessa versão da e-ScienceNet, o Gerente de Dados, assim como mostrado no RDF da Figura 3.26, realiza uma pesquisa genérica considerando o nome, o tipo e o componente dos dados. O que cada um desses campos representa é apresentado logo abaixo com exemplos reais na Figura 3.27.

```
<?xml version="1.0" encoding="UTF-8"?>

<rdf:RDF
        xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:generic="http://localhost:8080/generic.rdf#">

        <rdf:Description>
                <generic:name>1AIM</generic:name>
                <generic:description>CRUZAIN INHIBITED BY BENZOYL-TYROSINE-ALANINE-FLUOROMETHYLKETONE</generic:description>
                <generic:type>text</generic:type>
                <generic:component>Protein Data Bank</generic:component>
        </rdf:Description>

        <rdf:Description>
                <generic:name>2FPB</generic:name>
                <generic:description>Structure of Strictosidine Synthase, the Biosynthetic Entry to the Monoterpenoid Indole Alkaloid Fa
                <generic:type>text</generic:type>
                <generic:component>Protein Data Bank</generic:component>
        </rdf:Description>

        <rdf:Description>
                <generic:name>2VOA</generic:name>
                <generic:description>STRUCTURE OF AN AP ENDONUCLEASE FROM ARCHAEOGLOBUS FULGIDUS</generic:description>
                <generic:type>text</generic:type>
                <generic:component>Protein Data Bank</generic:component>
        </rdf:Description>

        <rdf:Description>
                <generic:name>3TMS</generic:name>
                <generic:description>PLASTIC ADAPTATION TOWARD MUTATIONS IN PROTEINS: STRUCTURAL COMPARISON OF THYMIDYLATE SYNTHASES</ge
                <generic:type>text</generic:type>
                <generic:component>Protein Data Bank</generic:component>
        </rdf:Description>
```

Figura 3.27: dados genéricos de um componente

Na Figura 3.27 temos um arquivo simples mostrando como dados de um componente são disponibilizados de forma genérica através da e-ScienceNet. O campo generic:name indica o um nome para os dados. O generic:description é uma descrição textual que possa enriquecer as informações disponibilizadas sobre aquele dado. Já no campo generic:type, temos o tipo do dado, geralmente relacionado ao campo generic:component que indica qual componente é responsável por aquele dado.

(a) Pesquisa por nome

```
PREFIX generic:<http://localhost:8080/generic.rdf#>

SELECT ?name ?description ?type ?component
FROM <generic.xml>
WHERE {
        ?n generic:name "3TMS" .
        ?n generic:name ?name .
        ?n generic:description ?description .
        ?n generic:type ?type .
        ?n generic:component ?component .
}
```

(b) Pesquisa por componente

```
PREFIX generic:<http://localhost:8080/generic.rdf#>

SELECT ?name ?description ?type ?component
FROM <generic.xml>
WHERE {
        ?n generic:component "Protein Data Bank" .
        ?n generic:name ?name .
        ?n generic:description ?description .
        ?n generic:type ?type .
        ?n generic:component ?component .
}
```

(c) Código para geração de pesquisa

```
    // Fecha o arquivo em RDF.
    lGenericModelStream.close();
}

// Verifica qual o tipo da pesquisa.
switch (aType) {
    case GenericSearchType.COMPONENT: {
        // Filtra por componente com esse nome.
        lFilter = "?n generic:component \"" + aQuery + "\" .'; break;
    }
    case GenericSearchType.TYPE: {
        // Filtra por tipo com esse nome.
        lFilter = "?n generic:type \"' + aQuery + '\" ."; break;
    }
    default: {
        // Filtra por esse nome.
        lFilter = "?n generic:name \"' + aQuery + '\" ."; break;
    }
}

// Monta a pesquisa;
lQueryString = "PREFIX generic:<http://localhost:8080/generic.rdf#>\n" +
        'SELECT ?name ?description ?type ?component\n" +
        'WHERE {\n" +
            lFilter +
            "?n generic:name ?name .\n" +
            "?n generic:description ?description .\n" +
            "?n generic:type ?type .\n" +
            "?n generic:component ?component .\n" +
        '}';

// Criar um objeto para realizar a pesquisa.
Query lQuery = QueryFactory.create(lQueryString);
```

Figura 3.28: simples pesquisas utilizadas pelo Gerente de Dados

Na Figura 3.28 (a) é mostrado um exemplo de pesquisa em SPARQL, ao qual seleciona dados genéricos com nome Halostylodinium, nome esse fornecido por uma pesquisa de um nó. Já na Figura 3.28 (b) temos uma pesquisa que retorna somente dados genéricos do componente de Objetos de Aprendizagem, que será utilizado para demonstrar a pesquisa no próximo capítulo. O código responsável por gerar essas pesquisas pode ser visualizado na Figura 3.28 (c).

Quando levado em consideração a arquitetura baseadas em componente, temos um problema para disponibilização dos diversos dados utilizados por esses componentes. Cada componente deve poder escolher o seu modelo de dados de acordo com as suas necessidades. Através da extensão da classe NetworkSearch, todos os componentes (que se registrarem) recebem as solicitações e realizam a pesquisa localmente de acordo com a suas necessidades.

A pesquisa genérica é um meio de facilitar a disponibilização de dados pelos componentes na rede, sendo a implementação de cada pesquisa mais rica, consequentemente mais complexa, de responsabilidade de cada componente.

Esses componentes podem disponibilizar arquivos de dados, tais como, textos científicos, vídeos, imagens e até mesmo serviços na rede, inviabilizando assim um modelo único de dados utilizável por todos os componentes.

É aconselhável que todos os componentes disponibilizem seus dados de forma genérica através do Gerente de Dados, mas esquemas criados e aceitos na literatura, tais como, metadados para pesquisa de Objetos de Aprendizagem e OWL-S para pesquisa de Serviços Web Semânticos também podem ser utilizados.

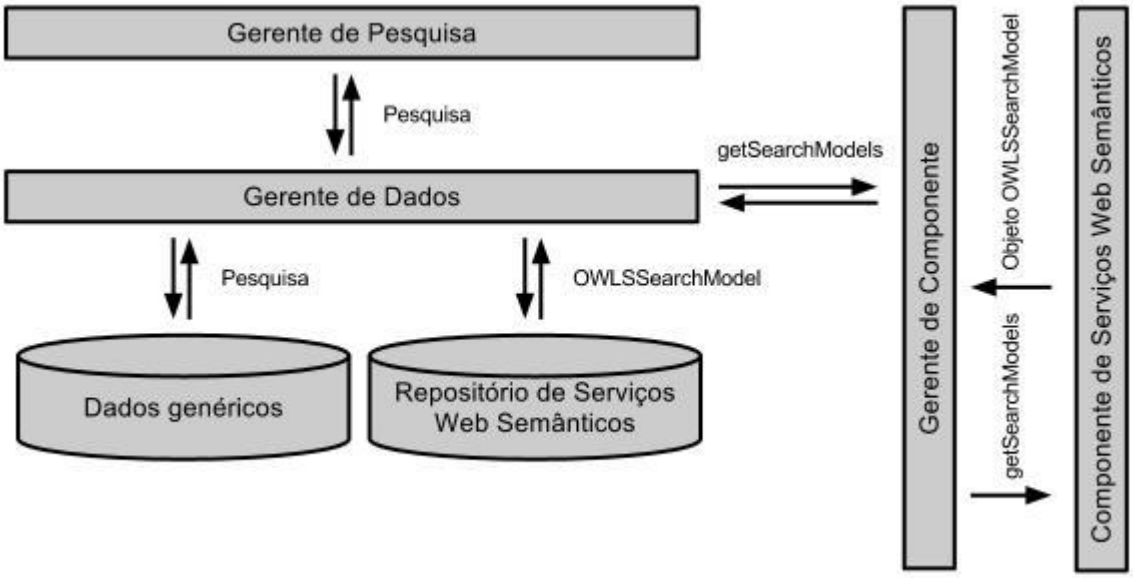

Figura 3.29: modelo de dados específicos para componentes

Conforme apresentado na Figura 3.29, cada componente pode descrever os dados de acordo com o seu esquema de dados. Para isso, o Gerente de Dados solicita ao Gerente de Componente que selecione entre seus componentes os objetos responsáveis por realizar as pesquisas específicas. Dessa forma, quando o Gerente de Pesquisa solicitar uma pesquisa ao Gerente de Dados, esses objetos também recebem a pesquisa. Na Figura 3.29 temos como exemplo o componente para disponibilização de Serviços Web Semânticos, que, além de disponibilizar utilizando o modelo de dados genérico, pode realizar uma pesquisa mais rica semanticamente sobre os seus dados utilizando OWL-S.

3.8 Gerente de Protocolo

Definimos o Gerente de Protocolo para abstrair a troca de mensagens entre os cientistas na e-ScienceNet.

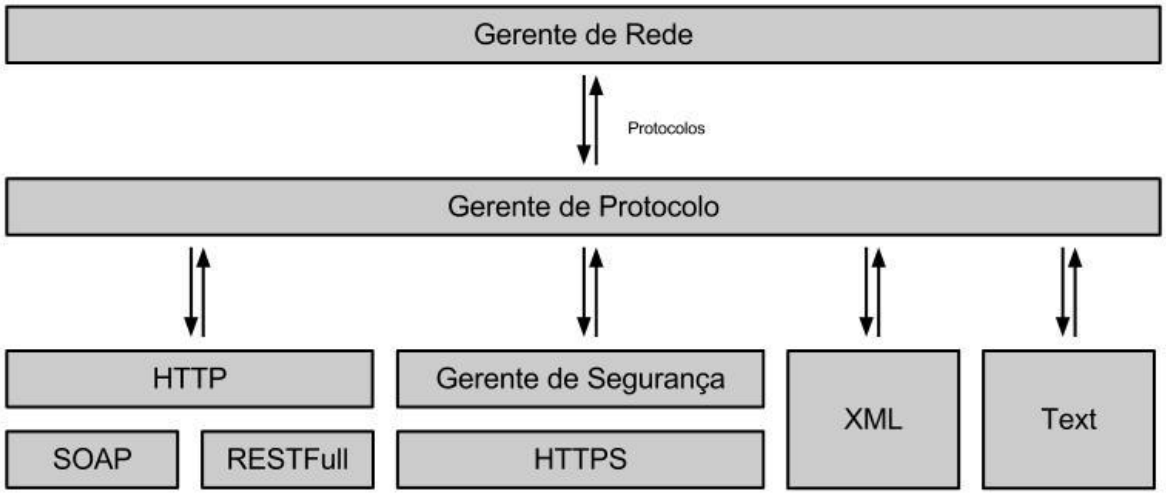

Figura 3.30: Gerente de Protocolo

Na Figura 3.30 temos a interação entre o Gerente de Rede e o Gerente de Protocolo, sendo mostrado alguns possíveis protocolos a serem utilizados na e-ScienceNet.

Muitos centros de pesquisas tem seu acesso a redes externas limitadas, impossibilitando que cientistas acessem a e-ScienceNet através de protocolos nativos de Redes Ponto a Ponto. Com a utilização de uma camada para o Gerente de Protocolo, a troca de mensagens da e-ScienceNet pode ser realizada através de diversos protocolos, como, por exemplo, o protocolo HTTP, que na maioria dos Firewalls existentes tem a sua comunicação permitida.

O Gerente de Protocolo também é responsável por proporcionar informação ao Gerente de Segurança sobre a criptografia da comunicação entre os cientistas. Podemos dizer, por exemplo, que a conexão entre dois cientistas está sendo realizada de forma segura

utilizando o protocolo Secure Socket Layer (SSL)[17] ou de forma textual simples utilizando protocolos otimizados para diminuir a quantidade de informação transferida em aplicações de alto desempenho cientifico.

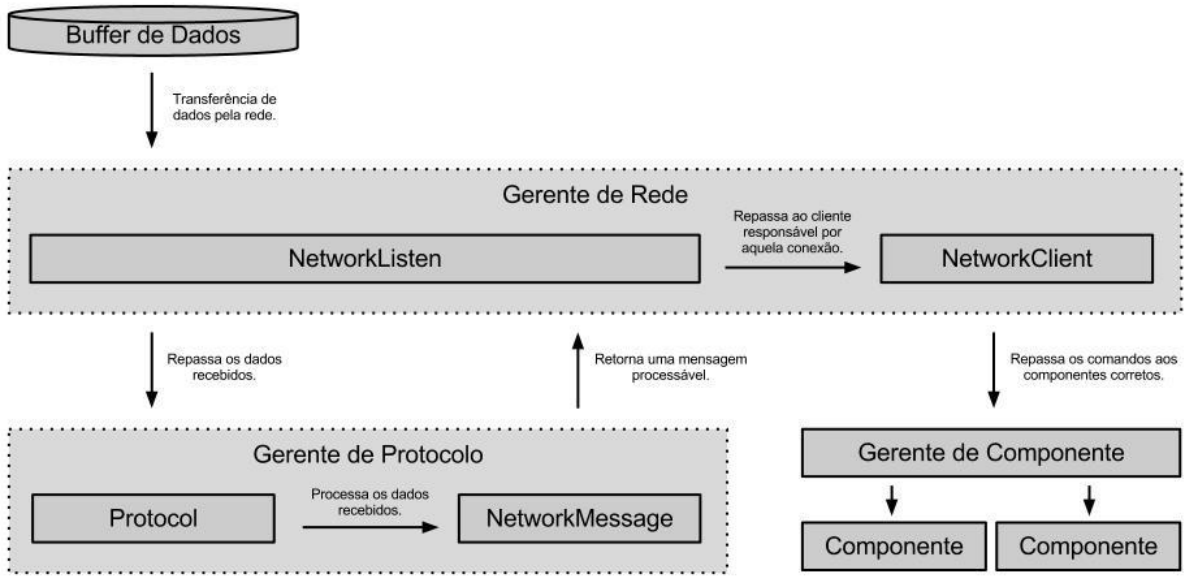

Figura 3.31: fluxo de dados pelo Gerente de Protocolo

Na Figura 3.31 mostramos o fluxo de dados pelo Gerente de Protocolo. Primeiramente temos os dados não lapidados vindo da rede. Ao receber esses dados por um "NetworkListen", objeto responsável pela escuta na rede, é criado um buffer com os dados para que eles sejam processados pelo protocolo responsável por aquele "NetworkListen". Todo "NetworkListen" tem um objeto "Protocol" associado, dessa forma a e-ScienceNet consegue escutar em diversas portas através do "NetworkListen", cada uma com seu protocolo específico. Após a mensagem ser processada pelo "Protocol" associado, é retornado ao "NetworkListen" um objeto "NetworkMessage", contendo diversos "NetworkCommand" a serem executados pelos seus respectivos componentes.

[17] SSL é um protocolo de comunicação segura que utiliza criptografia para transferência de informações..

3.9 Gerente de Segurança

Atualmente a segurança é uma grande preocupação na interligação de sistemas distribuídos. Qualquer sistema distribuído envolve todos os quatro aspectos de segurança: confidencialidade, integridade, disponibilidade e autenticidade. Segurança em ambientes distribuídos é um problema complexo que requer interação dos diversos recursos administrados de forma autônoma de um jeito que não cause impacto na usabilidade do sistema e que não apresente problemas de segurança individuais ou no ambiente como um todo [20].

Um dos requisitos apresentados para uma infraestrutura para aplicações em e-Science é a segurança das informações disponibilizada pelos cientistas. Os cientistas devem ser capazes de compartilhar seus dados de forma segura com auxilio de autenticação, criptografia e privacidade dos dados disponibilizados [20].

Para solucionar os problemas apresentados pelo requisito segurança, a e-ScienceNet prevê o uso de um Gerente de Segurança, capaz de prover funcionalidades de segurança da informação ao seu nível mais básico, não envolvendo a lógica de nenhum componente da e-ScienceNet.

Primeiramente, todos os cientistas na e-ScienceNet devem ter a liberdade de se identificar. Algumas funcionalidades devem estar disponíveis somente para cientistas autenticados. Os cientistas podem disponibilizar para o Gerente de Segurança uma assinatura digital confirmando a sua identidade na e-ScienceNet.

Outra forma de autenticação é a utilização de componentes de autenticação, onde um nó da rede pode se comportar como um servidor de autenticação. Podemos citar como exemplo, um centro de pesquisa onde os cientistas que trabalham nos laboratórios são autenticados remotamente pelo servidor de autenticação desse centro de pesquisa. Dessa forma, estamos confiando que aquele cientista é quem ele diz ser através do centro de pesquisa com o servidor de autenticação, que por sua vez, deve possuir uma assinatura digital confirmando que é o centro de pesquisa informado.

Limitação e privacidade dos dados disponibilizados devem ser levadas em consideração na hora de disponibilizar qualquer informação na e-ScienceNet. Na hora de

disponibilizar qualquer informação na e-ScienceNet o cientista deve selecionar os outros cientistas que podem ter acesso a aqueles dados. Para que um cientista possa liberar certos dados para outro cientista em específico, ambos os cientistas devem utilizar a funcionalidade descrita como assinatura digital.

A comunicação física entre os cientistas pode ser feita utilizando protocolos criptografados, como, por exemplo, o SSL. Para indicar se uma comunicação física está sendo criptografa ou não, o Gerente de Segurança precisa do auxilio do Gerente de Protocolo para dizer se aquela comunicação é ou não realizada de forma segura.

Os componentes disponibilizados na e-ScienceNet podem aumentar a sua credibilidade através de certificados digitais. Com o uso de certificados digitais os componentes disponibilizados podem assegurar as suas funcionalidades disponíveis na e-ScienceNet.

As Redes Ponto a Ponto por sua natureza já envolvem um dos aspectos da segurança, a disponibilidade. Utilizando técnicas de autenticação através de assinaturas digitais, conseguimos disponibilizar a autenticidade e integridade da informação fornecida por determinado cientista. Com auxilio do Gerente de Protocolo e com as políticas de comunidades semânticas conseguimos garantir a confidencialidade dos dados.

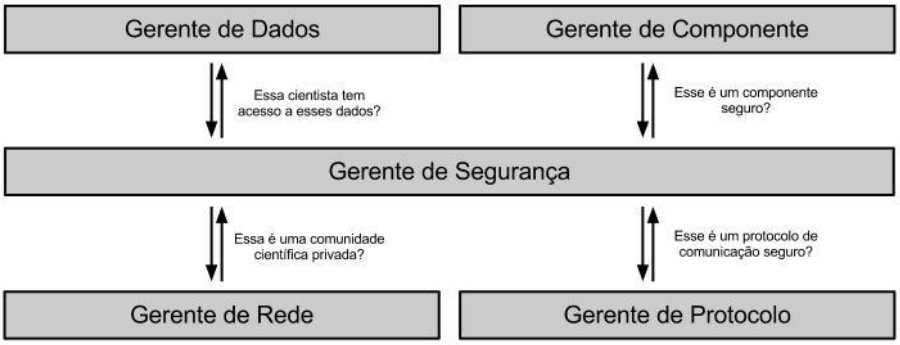

Figura 3.32: Gerente de Segurança

A Figura 3.32 mostra as possíveis interações entre o Gerente de Segurança e os outros gerentes da e-ScienceNet. No Gerente de Dados garantimos se determinado cientista tem acesso aos dados disponibilizados. No Gerente de Componente garantimos que aquele é um componente seguro a ser disponibilizado na rede. O Gerente de Protocolo indica se a

comunicação entre dois nós é realizada de forma segura. E no Gerente de Rede, garantimos que o acesso a determinadas comunidades semânticas é restrito.

Diversos componentes podem ser criados para aumentar a segurança da e-ScienceNet, tais como: uma lista negra de componentes criados e disponibilizados na e-ScienceNet que não são confiáveis e uma análise das respostas de pesquisas na e-ScienceNet para avaliar o conteúdo disponibilizado pelos cientistas.

O componente responsável por disponibilizar uma lista negra com informações de outros componentes que não confiáveis pode evitar que cientistas instalem em seus dispositivos componentes que não fazem o que se propõe. Já a análise das respostas de pesquisas pode nos dizer a validade dos arquivos disponibilizados pelos cientistas, levando em consideração o uso por outros cientistas conectados a e-ScienceNet.

Apesar de ainda não implementado na e-ScienceNet, o Gerente de Segurança é um importante passo a ser dado para uma arquitetura que suporte a e-Science.

3.10 e-ScienceNet e os requisitos para uma infraestrutura de apoio a e-Science

Considerando os requisitos apresentados no capítulo 2, seção 2.1, tem-se o armazenamento, gerenciamento de propriedade, transparência, comunidades, segurança, mobilidade, fluxo de trabalho, proveniência, notificações, suporte a decisão, expansão e componentes. De acordo com os objetivos, somente alguns dos requisitos foram implementados na primeira versão da e-ScienceNet.

O primeiro objetivo trata do armazenamento de dados, ao qual é implementado pelo Gerente de Dados, onde podemos acessar e disponibilizar de forma distribuída dados adicionados a rede. A pesquisa por esses dados é implementada pelo Gerente de Pesquisa, onde realizamos pesquisas distribuídas pelas comunidades semânticas.

O segundo objetivo é implementado pelo Gerente de Configurações, onde as preferências dos cientistas são armazenadas para futuras execuções. Também temos a implementação das políticas de acesso, funcionalidade essa implementada pelo Gerente de Rede em conjunto com o Gerente de Segurança.

O terceiro objeto trata da transparência de disponibilização dos dados. O acesso a e-ScienceNet é de responsabilidade do Gerente de Acesso, onde o cientista acessa de forma transparente os dados distribuídos pela rede como se fossem dados locais ao seu ambiente de trabalho.

O quarto objetivo refere-se a as notificações, disponíveis através da implementação de eventos pelo Gerente de Rede.

O quinto objetivo, também implementado pelo Gerente de Rede, faz uso de Tabelas de Espalhamento Distribuídas, onde conseguimos manter uma expansão de serviços e usuários na rede.

O sexto e último objetivo, é implementado pelo Gerente de Componente, onde as funcionalidades disponíveis na rede são carregadas através de Componentes de Software.

Alguns requisitos que vão além dos objetivos iniciais da e-ScienceNet também foram implementados. Temos as comunidades semânticas implementadas pelo Gerente de Rede e Gerente de Semântica, e a mobilidade implementada pelo Gerente de Acesso.

Alguns requisitos ainda não implementados na primeira versão da e-ScienceNet, mas tendo a arquitetura projetada de acordo que a inclusão dos requisitos cause um mínimo de alterações na estrutura da e-ScienceNet. Podemos citar o gerenciamento de propriedade, segurança, fluxo de trabalho, proveniência e suporte a decisão como requisitos ainda não implementados.

4 PROTÓTIPO: IMPLEMENTAÇÃO E CENÁRIO DE USO

A implementação do protótipo da e-ScienceNet baseia-se no emprego de um conjunto de tecnologias oriundas da Ciência da Computação onde a ênfase reside no emprego de software livre e de código aberto.

Pode-se destacar o emprego das seguintes tecnologias no desenvolvimento do protótipo da e-ScienceNet:

- Linguagem de Programação Java[18];
- Biblioteca para manipulação de Ontologias Apache Jena[19];
- Biblioteca para mapeamento de Ontologias Alignment API[20];
- Ambiente de desenvolvimento integrado Eclipse[21].

Neste capítulo, apresentamos o primeiro protótipo da e-ScienceNet. Este protótipo foi desenvolvido utilizando Java, e teve como principal objetivo o teste das ideias e funcionalidades apresentadas no capítulo anterior.

Para ilustrar o carregamento de componentes e as interações na rede, foram utilizadas as Ontologias Sequence Aligning Ontology, CelO e MyGridOntology. Além disso, mostramos o acesso aos dados do Protein Data Bank[22] (PDB) de maneira distribuída pela rede como cenário de uso. O PDB é um banco de dados em 3D de proteínas e ácidos nucléicos. Os dados armazenados no PDB tem um formato padrão e geralmente são obtidos através de difração de raios X ou ressonância magnética nuclear. Esses dados são armazenados e distribuídos em domínio publico, sendo seu acesso disponível a qualquer cientista através da Web. Um importante uso das informações contida no PDB é a criação de novas drogas, onde a estrutura 3D da proteína é utilizada para realizar a predição de cavidades e ligamentos.

[18] Java disponível em http://www.java.com
[19] Apache Jena disponível em http://jena.apache.org
[20] Alignment API disponível em http://alignapi.gforge.inria.fr
[21] Eclipse disponível em http://www.eclipse.org
[22] PDB disponível em http://www.rcsb.org

A Figura 4.1 apresenta uma interface gráfica genérica da e-ScienceNet que disponibiliza as funcionalidades propostas no capítulo três.

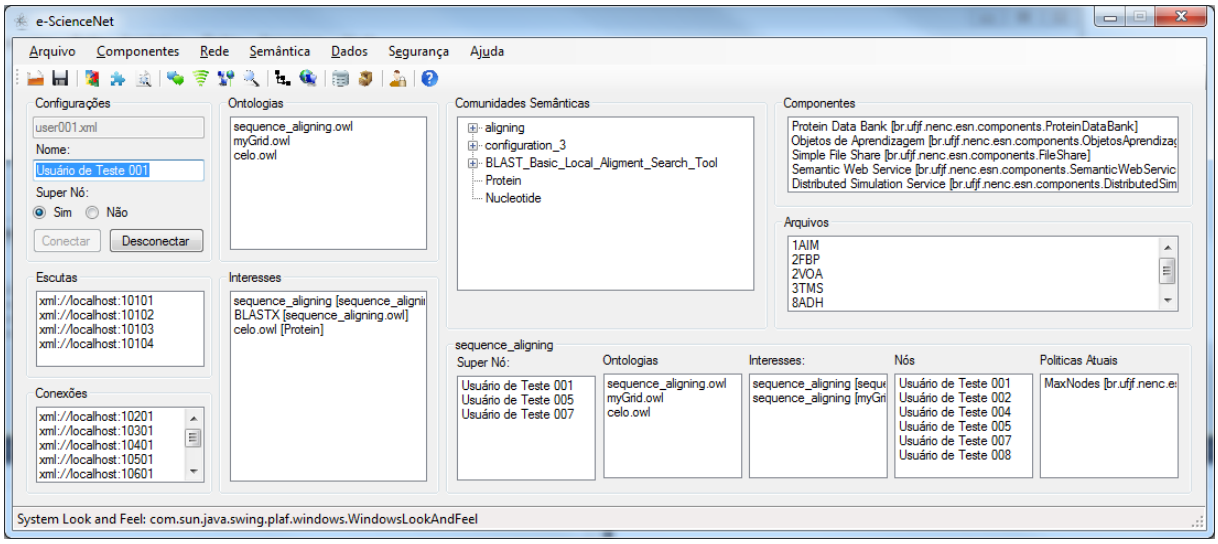

Figura 4.1: interface gráfica do protótipo da e-ScienceNet

A Figura 4.1 apresenta algumas informações relevantes para o funcionamento da eScienceNet, como, as conexões realizadas, Ontologias responsáveis por descrever os interesses do cientista e as comunidades semânticas conectadas. No decorrer deste capítulo explicaremos o significado de cada um desses campos apresentados na Figura 4.1 e qual sua relação com a arquitetura apresentada no capítulo anterior.

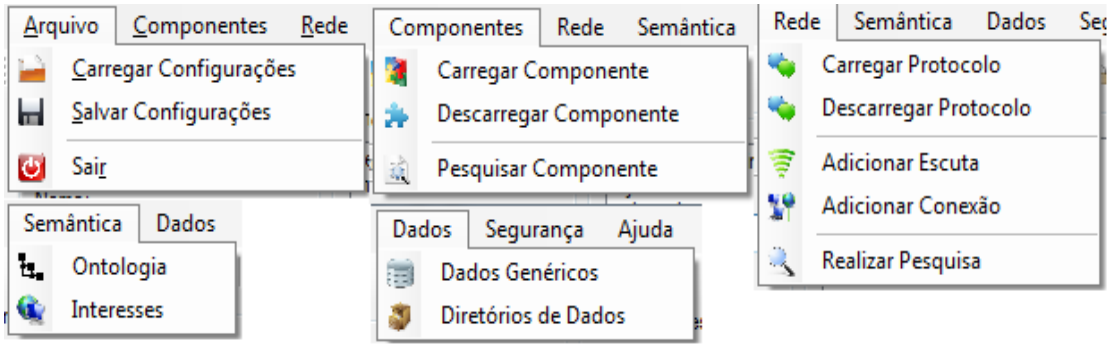

Figura 4.2: menus do protótipo da e-ScienceNet

Para acessar a edição das configurações em uso pela e-ScienceNet mostrados na Figura 4.2, é necessário que o cientista acesse as respectivas telas através dos menus mostrado na Figura 4.2.

Inicialmente temos o menu “Arquivo”, onde é possível carregar ou adicionar através de um arquivo XML as configurações em uso pela e-ScienceNet. Já na opção “Salvar Configurações”, cria-se um arquivo XML com as configurações atuais de execução da eScienceNet, para que em futuras execuções não seja necessário informar todos os dados novamente. O menu “Componentes” engloba as funções do Gerente de Componente, onde é possível carregar e descarregar componentes em tempo de execução, assim como através do menu “Pesquisar Componente”, responsável por realizar pesquisas na e-ScienceNet por mais componentes que realizem determinadas tarefas. No menu “Rede” tem-se concentrado as funcionalidades relacionadas ao Gerente de Protocolo, ao Gerente de Rede e ao Gerente de Pesquisa. Em “Semântica” estão disponíveis as funcionalidades relacionadas ao Gerente de Semântica e ao Gerente de Interesses. O menu “Dados” refere-se as funcionalidades do Gerente de Dados e no menu “Segurança”, nessa primeira versão do protótipo, tem-se as “Políticas de Acesso” as comunidades semânticas, as quais são responsabilidade do Gerente de Rede em conjunto com o Gerente de Segurança.

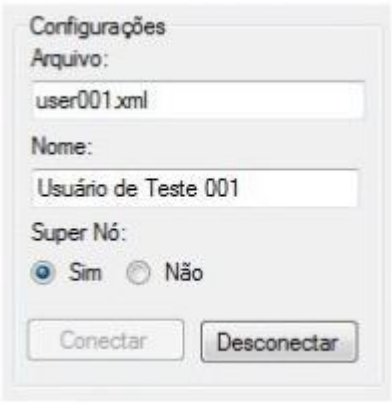

Figura 4.3: configurações básicas da e-ScienceNet

Na Figura 4.3 temos as configurações básicas iniciais de execução da e-ScienceNet. No campo “Arquivo” é mostrado o nome do arquivo inicial onde as configurações foram carregadas, arquivo esse que será utilizado para salvar as configurações ao finalizar a

execução do protótipo. O campo “Nome” indica o nome do nó de rede que será disponibilizado para os outros nós. Já em “Super Nó”, informamos se aquele nó pode ou não funcionar como super nó na rede. Essas configurações podem ser digitadas diretamente nos campos caso seja a primeira execução, ou então carregadas através da opção “Carregar Configurações” disponível no menu “Arquivo”.

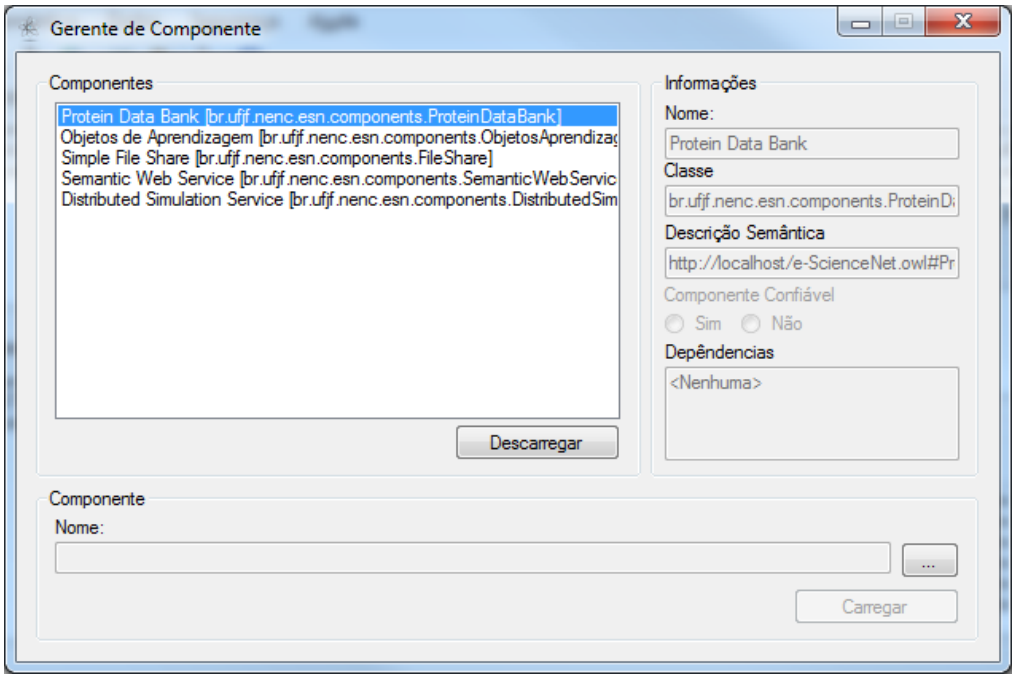

Figura 4.4: componentes disponíveis no protótipo

Na Figura 4.4, mostramos alguns componentes carregados na e-ScienceNet para testar a capacidade de carregamento de componentes em tempo de execução. Vamos destacar o componente de Pesquisa de Proteínas, que será utilizado como exemplo neste capítulo para apresentação das funcionalidades da e-ScienceNet. Com especial ênfase, este componente foi utilizado para apresentar as funcionalidades do Gerente de Dados.

Em “Informações” na Figura 4.4 temos algumas informações básicas sobre aquele componente, como a Descrição Semântica e as Dependências com outros componentes. Mais informações sobre o que cada um desses campos representa podem ser encontradas na seção Gerente de Componente do capítulo e-ScienceNet.

Para realizar o carregamento de um novo componente, clica-se no botão "...", onde podemos selecionar o caminho completo do arquivo com o código do componente. Esse componente é uma classe Java empacotada em um arquivo compactado do tipo JAR. Os componentes podem ser carregados durante a inicialização da rede através do arquivo de configuração, ou diretamente pelo cientista através do menu "Carregar Componente". Depois de carregado, o componente aparece na lista de componentes em execução, sendo possível a sua descarga a qualquer momento através do menu "Descarregar Componente".

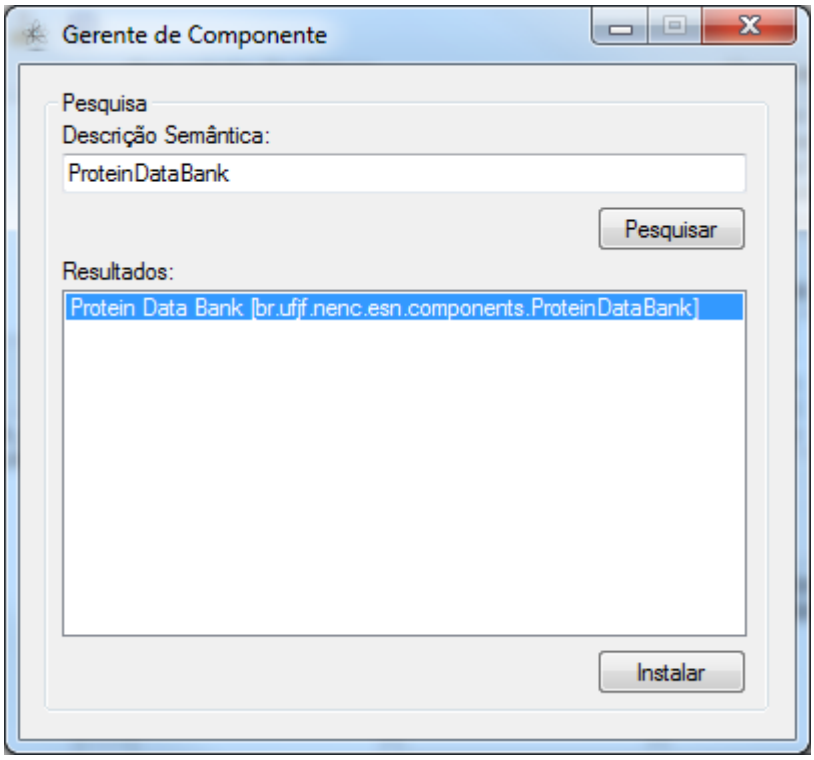

Figura 4.5: pesquisa de componentes na e-ScienceNet

A Figura 4.5 mostra a tela para pesquisa de componentes na e-ScienceNet. Como mostrado na seção do Gerente de Componente no capítulo anterior, para realizar a pesquisa de um componente existente na rede é necessário informar a descrição semântica daquele componente. Quando informamos a descrição semântica, ou seja, um ou mais termos ontológicos relacionados a Ontologias que descrevem aquele nó, a e-ScienceNet pesquisa nos nós em que está conectada e verifica se algum componente corresponde com a descrição semântica informada. A verificação é feita através de uma pesquisa considerando os termos

informados pelo cientista na descrição semântica do componente através de uma comparação de textual.

Figura 4.6: escutas de rede e suas conexões

A partir do menu “Rede” é possível acessar algumas funcionalidades inerentes ao Gerente de Rede, sendo possível monitorar as conexões existentes. Na Figura 4.6 em “Escutas” temos os endereços locais de rede ao qual estamos recebendo conexões. Já em “Conexões”, mostramos as conexões atuais a outros nós na rede. Para começar a escutar em um novo endereço ou porta, assim como realizar uma conexão direta a outro nó, temos os campos “Endereço”, “Porta” e “Protocolo”.

O campo “Endereço” é informado no formato IP ou pelo nome do dispositivo na rede. A “Porta” vai de 1025 até 65535. Já no campo “Protocolo”, é disponibilizado uma lista com os protocolos conhecidos pelo Gerente de Protocolo. Quando um nó recebe uma conexão, é necessário que se tenha informações de qual protocolo está sendo utilizado naquela porta para que a conexão seja estabelecida com sucesso.

Todos os nós apresentados na Figura 4.6 e utilizados ao longo das demais telas do protótipo estão na mesma rede local. Sendo assim, além da forma de conexão direta a esses nós, um simples broadcast na rede local pode disponibilizar todas as informações necessárias de conexão a um novo cliente que deseja se conectar a e-ScienceNet..

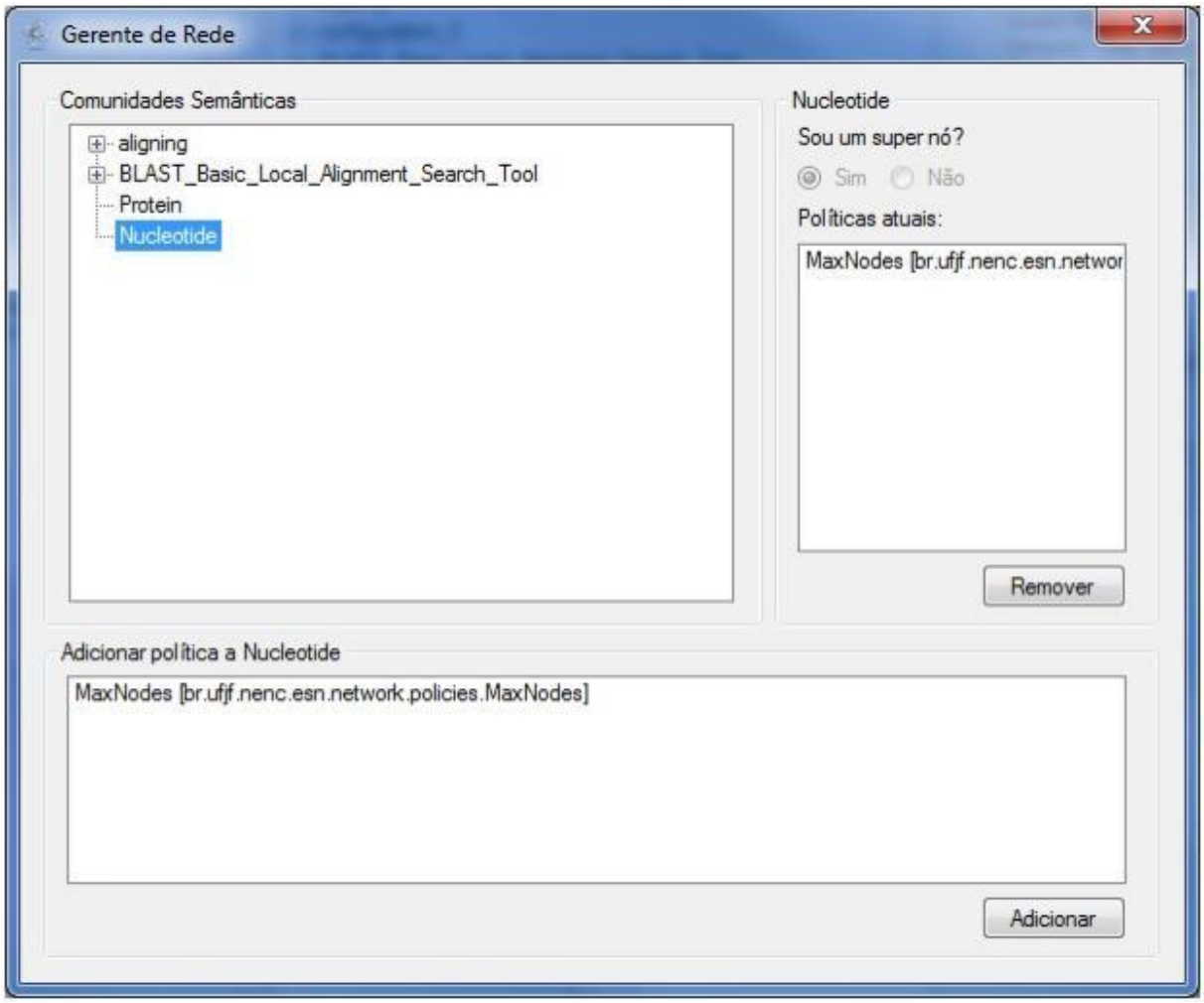

Figura 4.7: políticas de acesso às comunidades semânticas

Na Figura 4.7 temos uma tela para adição de políticas de acesso as comunidades semânticas. No campo “Comunidades Semânticas” são mostradas as comunidades semânticas conhecidas por esse nó. Caso o nó não estiver configurado para trabalhar como super nó, as comunidades semânticas não são apresentadas.

Após selecionar a comunidade semântica, ao lado direito são exibidas informações sobre a comunidade semântica selecionada. No campo “Sou um super nó” temos a indicação se podemos ou não adicionar ou remover políticas de acesso aquela comunidade semântica. Caso o nó não tenha permissão de realizar modificações naquela comunidade, ou seja, não é um super nó daquela comunidade semântica, o botões de “Adicionar” e “Remover” ficam desativados para aquele cientista. Em “Políticas Atuais” temos a lista de políticas de acesso que estão atualmente adicionadas naquela comunidade semântica. Em “Adicionar política a Nucleotide” temos a lista de políticas de acesso disponíveis para serem adicionadas a comunidade semântica “Nucleotide”.

A política de acesso mostrada na Figura 4.7 cria um limite de conexões a comunidade semântica “Nucleotide”. A primeira versão da política de acesso “MaxNodes” limita a somente 7 conexões a comunidade semântica a qual ela for aplicada.

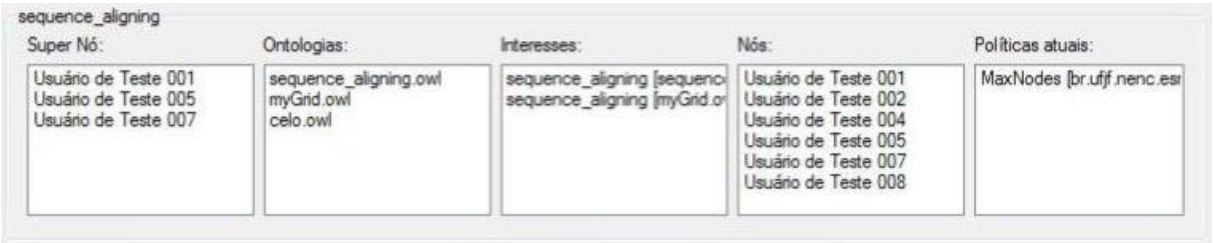

Figura 4.8: informações sobre a comunidade semântica

Na Figura 4.1 quando selecionamos uma comunidade semântica, como no exemplo a “sequence_aligning”, são exibidas informações específicas sobre aquela comunidade semântica. Na Figura 4.8 temos algumas informações relevantes sobre a comunidade semântica “sequence_aligning”, como, “Super Nó” indicando quais são os super nós atuais daquela comunidade semântica, “Ontologias” que apresentam quais Ontologias estão mapeadas nos super nós daquela comunidade semântica, “Interesses” que mostra quais são os interesses (termos ontológicos) que descrevem semanticamente aquela comunidade, “Nós” que apresenta quais são os nós que estão conectados aquela comunidade e “Políticas atuais” que mostra quais são as políticas de acesso aquela comunidade semântica que estão vigentes.

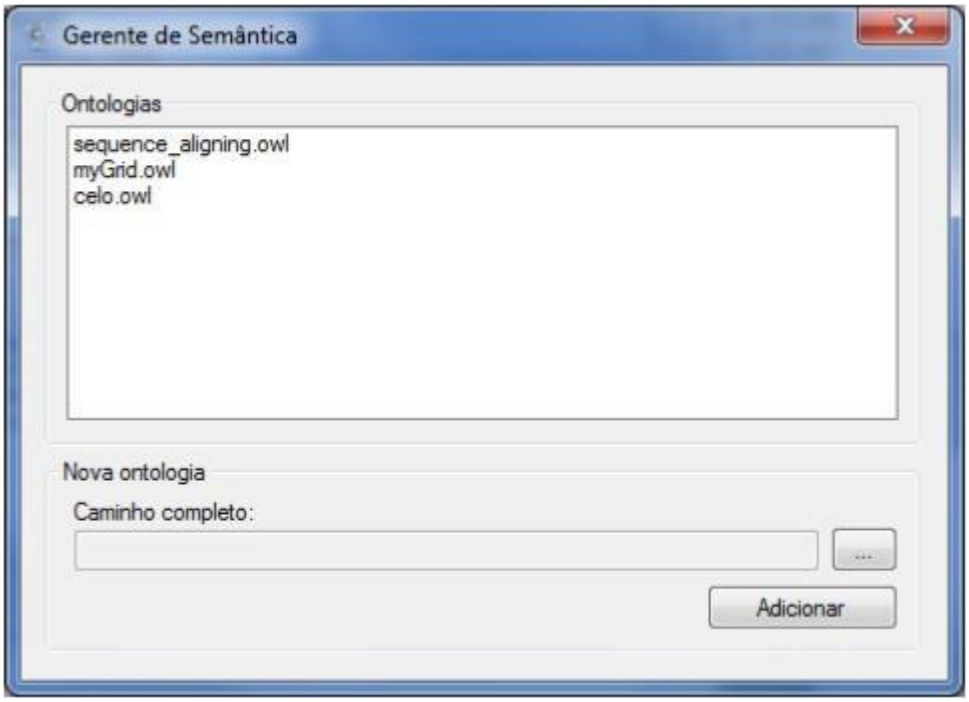

Figura 4.9: adição de Ontologias na e-ScienceNet

A Figura 4.9 mostra a tela para carregamento de Ontologias na e-ScienceNet. Para que um cientista indique os seus interesses, primeiramente é necessário realizar a adição da Ontologia a e-ScienceNet. Para adicionar uma Ontologia basta clicar no botão "..." e clicar no botão "Adicionar". Assim que adicionada a Ontologia, é realizado o processo de mapeamento com as outras Ontologias locais que aquele cientista está utilizando, assim como mostrado no Gerente de Semântica. As Ontologias utilizadas na Figura 4.9 são respectivamente Sequence Aligning Ontology, MyGridOntology e CelO.

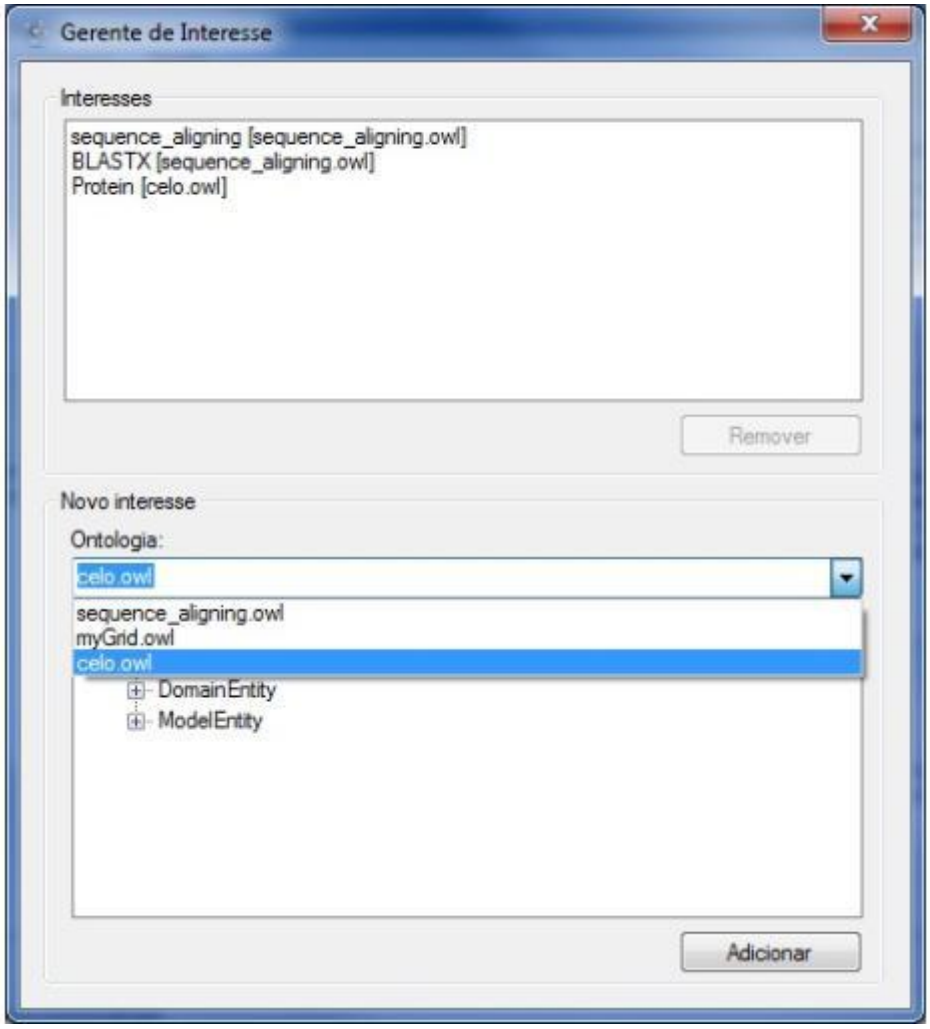

Figura 4.10: adição de interesses na e-ScienceNet

Para adicionar um interesse à e-ScienceNet, basta selecionar a Ontologia desejada que serão carregadas os termos daquela Ontologia e exibidos para que o cientista realize a seleção dos seus interesses relacionados aquela Ontologia. Um cientista pode ter interesse em Ontologias distintas. Assim como mostrado na lista de "Interesses" da Figura 4.10, temos três interesses, sendo "sequence_aligning" e "BLASTX" da Ontologia Sequence Aligning Ontology e "Protein" da Ontologia CelO.

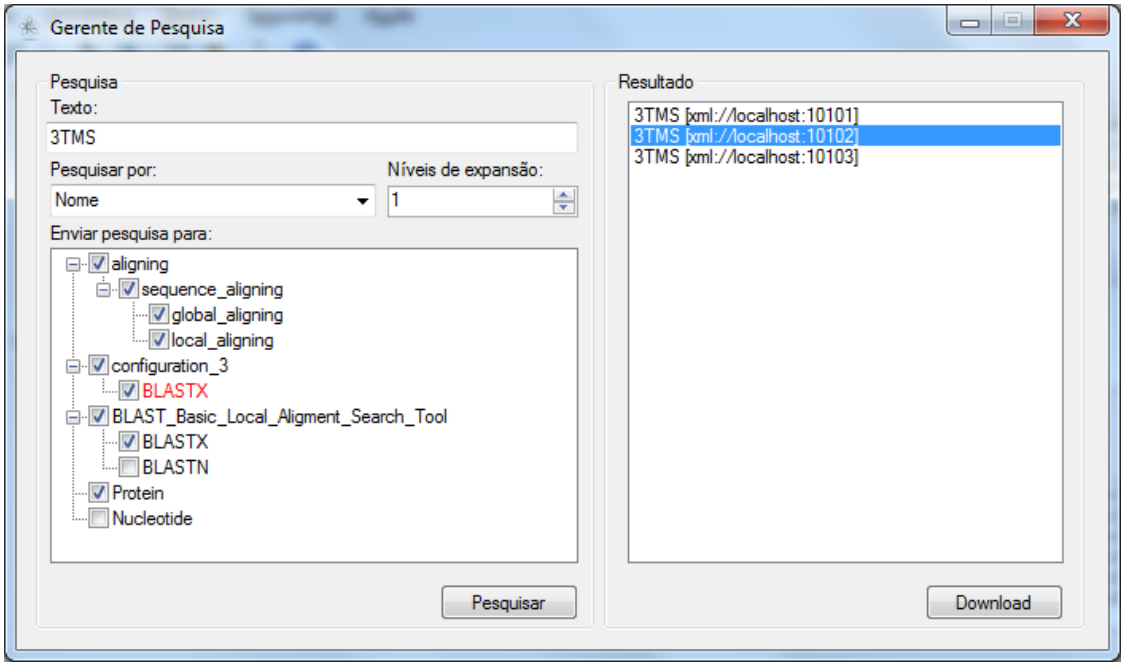

Figura 4.11: pesquisa na e-ScienceNet

A Figura 4.11 apresenta uma pesquisa sendo realizada na e-ScienceNet. No campo "Texto" indica-se um ou mais termos para pesquisa. Em "Pesquisar por" indica-se onde a pesquisa será realizada. Assim como mostrado na seção do Gerente de Dados no capítulo anterior, a pesquisa pode ser realizada considerando o nome, o tipo e o componente. Em "Níveis de expansão" tem-se o repasse das pesquisas para as comunidades semânticas relacionadas. Considerando a Figura 4.11, o cientista está devidamente conectado somente nas comunidades semânticas "sequence_aligning", "BLASTX" e "Protein". Caso a expansão estivesse configurada para 0, ou seja, sem expansão, a pesquisa seria enviada para somente as comunidades que estão conectadas. Já que a pesquisa está configurada para expansão de até um nível, a pesquisa será expandida para outras comunidades semânticas relacionadas semanticamente com estas, sendo elas "aligning", "configuration_3" e "BLAST_Basic_Local_Aligment_Search_Tool" considerando relacionamentos de hiperonimos e "global_aligning" e "local_aligning" considerando relacionamentos hipônimos.

É importante notar que uma das comunidades "BLASTX" foi inferida. Isso se da pelo fato de essa comunidade ter sido inferida na Ontologia, indicando que a comunidade semântica "configuration_3" pode ter conteúdo semanticamente similar a "BLASTX". O campo "Forçar pesquisa na comunidade" possibilita que a pesquisa seja enviada para outras

comunidades semânticas, mesmo que essa não esteja na lista de interesses informados pelo cientista.

Os resultados mostrados na Figura 4.11 são as informações retornadas por aquela pesquisa nas comunidades semânticas informadas. O termo da pesquisa quando pesquisado pelo campo nome retornou que existem dados com esse termo em três nós distintos na rede, onde todos eles estão no localhost em portas distintas, mas com o mesmo protocolo.

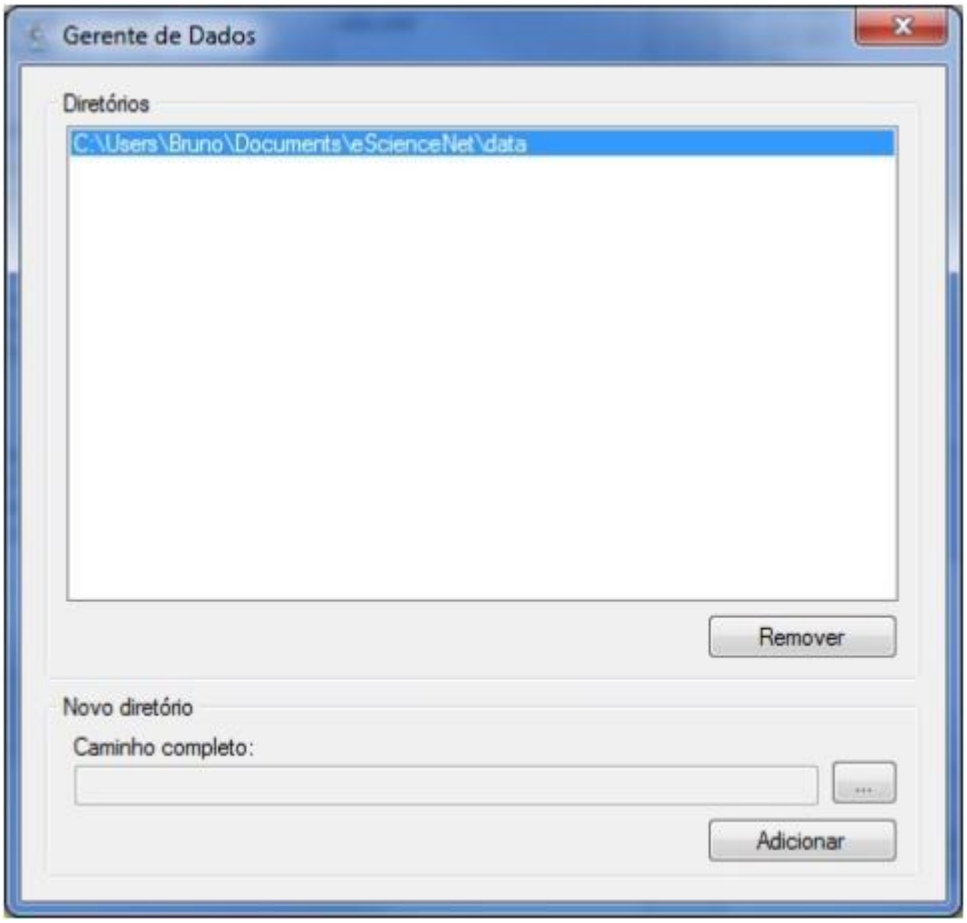

Figura 4.12: diretórios com dados de pesquisa

No Gerente de Dados precisamos informar os diretórios onde os dados estão armazenados. O RDF com os dados apresentados na Figura 3.27 armazena somente um índice semântico para realização das pesquisas. Os dados reais são armazenados nos diretórios especificados pela tela da Figura 4.12. Quando é requisitado um dado nos outros nós, o Gerente de Dados concatena o nome do dado no RDF com os diretórios nesta lista e verifica

sua existência em disco, caso mais de um arquivo seja encontrado, é retornado o do primeiro diretório.

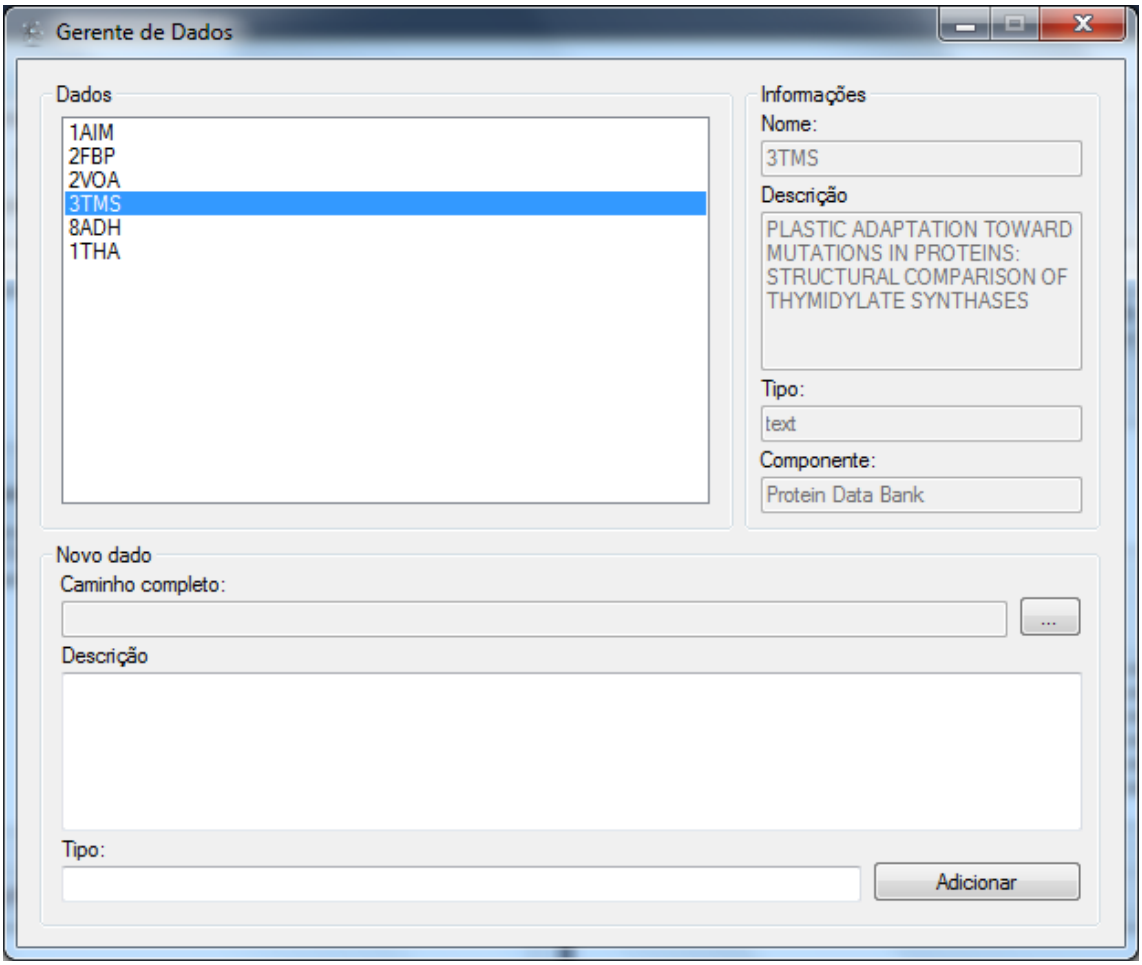

Figura 4.13: dados genéricos disponíveis no Gerente de Dados

Na Figura 4.13 temos os dados cadastrados no Gerente de Dados. Ao selecionar um dado item na lista, o lado direito mostra as informações cadastradas no RDF sobre aquele arquivo. Para adicionar um novo arquivo, basta selecionar o seu caminho completo através do botão "...", criar uma descrição e informar seu tipo. Quando a adição é feita pelo cientista, o campo "Componente" fica vazio, indicando que aquela informação foi adicionada pelo cientista e não por um componente. Quando o arquivo é adicionado, os seus dados são copiados para o primeiro diretório informado na Figura 4.12.

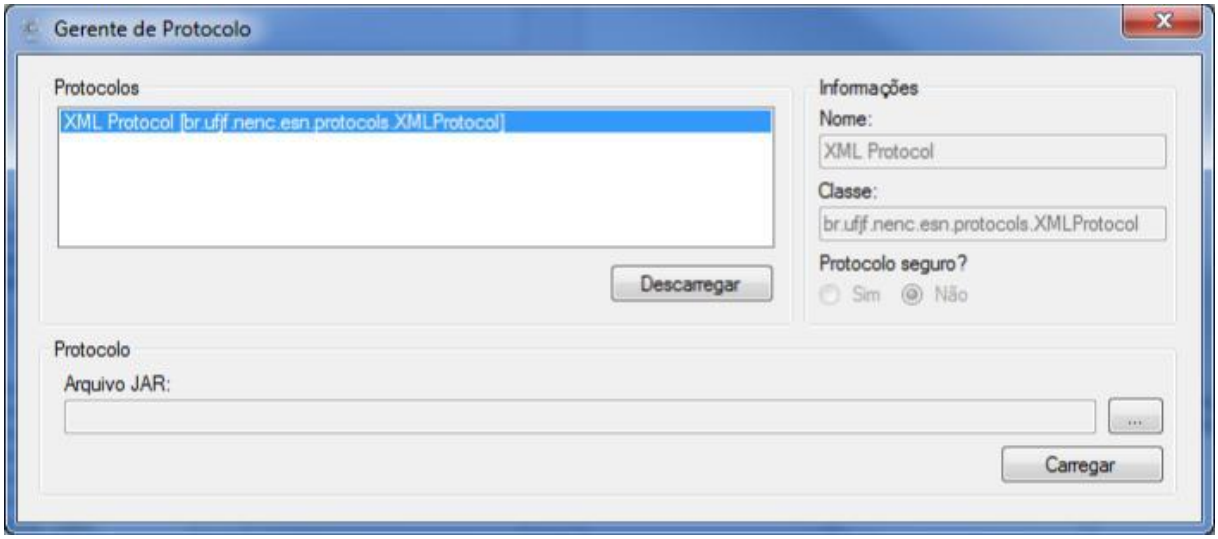

Figura 4.14: gerenciamento de protocolos na e-ScienceNet

A Figura 4.14 mostra como acontece a adição e remoção de protocolos. Para poder criar escutas ou se conectar a nós que utilizem determinados protocolos, ambos os nós precisam ter esse protocolo instalado. A instalação de um protocolo é similar a instalação de um componente. Basta selecionar um arquivo JAR com o código do protocolo que ele fica disponível para ser utilizado na rede. O campo Protocolo seguro indica se aquele é um protocolo de comunicação seguro, informações essas relevantes ao Gerente de Segurança.

5 CONSIDERAÇÕES FINAIS

Os recursos computacionais estão se tornando cada vez mais importantes no ciclo de vida da pesquisa científica. A quantidade de dados científicos gerados computacionalmente vem crescendo cada vez mais, onde o poder computacional viabiliza a execução de diversos experimentos, que, quando não automatizados, ficariam praticamente impossíveis de serem executados. Além disso, temos o grande crescimento do volume de dados no contexto científico, onde fica cada vez mais desafiador o armazenamento, processamento e pesquisa desses dados.

Projetou-se e desenvolveu-se a arquitetura da e-ScienceNet, com objetivo de apoiar as pesquisas científicas em um contexto distribuído. Para tal, consideramos o uso de Redes Ponto a Ponto e Web Semântica. As Redes Ponto a Ponto proporcionam um baixo custo, mas ao mesmo tempo disponibiliza um grande potencial para compartilhamento de dados distribuído. A Web Semântica proporciona uma rica descrição semântica dos dados disponibilizados pela Rede Ponto a Ponto.

Com a junção das Redes Ponto a Ponto com Web Semântica, conseguimos criar comunidades semânticas e agrupar cientistas na mesma comunidade de acordo com seus interesses e especialidades. Além disso, com o uso de Ontologias e Máquinas de Inferência, conseguimos inferir relações entre as diferentes comunidades semânticas, podendo expandir ou não a pesquisa para outros nós com conteúdo semanticamente similar.

O desenvolvimento de experimentos científicos computacionais é um processo complexo. Conseguimos diminuir a complexidade de criação dos experimentos científicos computacionais ao basear a e-ScienceNet em Componentes de Software, criando uma estrutura mais rica em funcionalidades disponíveis aos cientistas e desenvolvedores de componentes. Existem diferentes Componentes de Software que podem ser utilizados em conjunto para a realização de um experimento.

Considerando os objetivos iniciais da e-ScienceNet, temos os distintos gerentes na arquitetura responsáveis por determinados objetivos. Temos também os requisitos para uma

infraestrutura de suporte a e-Science, assim como visto na seção 3.10 do capítulo 3. Os objetivos abrangem um ou mais requisitos para uma infraestrutura de apoio a e-Science, sendo alguns requisitos não desenvolvidos na primeira versão da e-ScienceNet.

O protótipo da e-ScienceNet apresentado têm algumas limitações plausíveis de trabalhos futuros, além dos diversos componentes que podem ser criados com novas funcionalidades disponíveis a rede.

Uma limitação que podemos citar é o não desenvolvimento total de todos os gerentes da e-ScienceNet, como, por exemplo, o Gerente de Segurança. O Gerente de Segurança teve seus objetivos apresentados, e seu espaço projetado na arquitetura inicial da e-ScienceNet, mas suas funcionalidades não foram desenvolvidas nesta primeira versão.

Outra limitação importante que podemos citar é o mapeamento incompleto de Ontologias, implementado pelo Gerente de Semântica. Com a implementação do mapeamento incompleto de Ontologias, conseguimos diminuir o tamanho de armazenamento e a eficiência de processamento dos mapeamentos, sendo esse um importante fator no tempo total das conexões e pesquisas em comunidades semânticas.

REFERÊNCIAS

[1] MEDJAHED, B.; BOUGUETTAYA, A. “Service Composition for the Semantic Web”, *Springer*, 191 p, ISBN 978-1-4419-8464-7, 2011.

[2] LUA E. K.; et al. “A Survey and Comparison of Peer-to-Peer Overlay Network Schemes”, *IEEE Communications Surveys and Tutorials*, v. 7, p. 72-93, 2005.

[3] BERNERS-LEE T.; HENDLER J.; LASSILA O. “The Semantic Web”, *Scientific American*, v. 284, p. 34-43, 2001.

[4] MATTOSO, M.; et al. “Gerenciando experimentos científicos em larga escala”, *XXVIII Seminário Integrado de Software e Hardware, Sociedade Brasileira de Computação*, p. 121-135, 2008.

[5] VALENTE W. A. G. “SciProv: uma Arquitetura para a Busca Semântica em Metadados de Proveniência no Contexto de e-Science”, *Programa de Pós-graduação em Modelagem Computacional*, Dissertação de Mestrado. Universidade Federal de Juiz de Fora, Juiz de Fora, 2010.

[6] HENDLER, J. “Science and the Semantic Web”, *Science*, v. 299, n. 5606, p. 520-521, 2003.

[7] CARVALHO A. “Grandes Desafios da Pesquisa em Computação no Brasil: 2006-2016”, *Seminário Grandes Desafios da Pesquisa em Computação no Brasil*, São Paulo, 2006.

[8] SILVA, L. A. M. “Composer-Science: um framework para a composição de Workflows Científicos”, *Programa de Pós-graduação em Modelagem Computacional*, Dissertação de Mestrado. Universidade Federal de Juiz de Fora, Juiz de Fora, Julho de 2010.

[9] HINE, C. M. “New infrastructures for knowledge production: understanding e-science”, *Information Science Publishing*, v. 1, p. 306, 2006.

[10] PARASTATIDIS, S. “A Platform for All That We Know: Creating a Knowledge-Driven Research Infrastructure”, *The Fourth Paradigm: Data Intensive Scientific Discovery*, p. 165-172, 2009.

[11] DIGIAMPIETRI, L. A. “Gerenciamento de workflows científicos em bioinformática”, *Instituto de Computação*, Tese de Doutorado. Universidade Estadual de Campinas, Campinas, 2007.

[12] FENSEL D. “Ontologies: A Silver Bullet for Knowledge Management and Electronic Commerce”, *Springer*, Verlag, ISBN 3540003029, 2003.

[13] ADAM, N. R.; YESHA, Y. “Strategic Directions in Electronic Commerce and Digital Libraries: Towards a Digital Agora”, *ACM Computing Surveys*, v. 28, p. 818-835, 1996.

[14] BUSSLER, C. “B2B Protocol Standards and their Role in Semantic B2B Integration Engines”, *IEEE Data Engineering Bulletin*, v. 24, p. 3-11, 2001.

[15] DOGAC, A. “A Survey of the Current State-of-the-Art in Electronic Commerce and Research Issues in Enabling Technologies”, *Proceddings of the Euro-Med Net Conference*, Electronic Commerce Track, p. 50-53, 1998.

[16] DOGAC, A.; CINGIL, I. “A Survey and Comparison of Business-to-Business E-Commerce Frameworks”, *ACM SIGecom Exchanges*, v. 2 p. 16-27, 2001.

[17] LARSEN, G. “Component-based Enterprise Frameworks”, *Communications of the ACM*, v. 43, p. 24-26, 2000.

[18] SHIM, S. S. Y.; et al. “Business-to-Business E-Commerce Frameworks”, *IEEE Computer*, v. 33, p. 40-47, 2000.

[19] TANENBAUM, A. S.; STEEN, M. V. “Distributed Systems: Concepts and Design 5th Edição”, *Addison-Wesley*, ISBN 978-0132143011, 2011.

[20] ROURE, D.; JENNINGS, N.; SHADBOLT N. “The Semantic Grid: A Future e-Science Infrastructure”, *Concurrency and Computation: Practice and Experience*, v. 15, n. 11, 2003.

[21] ROURE, D.; JENNINGS N.; SHADBOLT N. “Research Agenda for the Semantic Grid: A Future e-Science Infrastructure”, *Grid Computing - Making the Global Infrastructure a Reality*, p. 437-470, ISBN 0470853190, 2001.

[22] LAUSCHNER, T. “Ambientes de e-Science e a Evolução dos Padrões de Arquitetura de Computação em Grade”, Rio de Janeiro, 2005.

[23] GIL, Y.; et al. “A Semantic Framework for Automatic Generation of Computational Workflows Using Distributed Data and Component Catalogs”, *Journal of Experimental and Theoretical Artificial Intelligence*, 2010.

[24] STUCKENSCHMIDT, H.; et al. “Peer-to-Peer and Semantic Web”, *Semantic Web and Peer to peer*, p. 1-17, ISBN 978-3-540-28346-1, 2006.

[25] PHAM, T. V.; LAU, L. M. S.; DEW, P. M. “The integration of Grid and Peer-to-peer to Support Scientific Collaboration”, *Proceedings of GGF11 Semantic Grid Applications Workshop*, Hawaii, p. 71-77, 2004.

[26] TANG, C.; XU, Z.; DWARKADAS S. “Peer-to-peer information retrieval using self-organizing semantic overlay networks”, *Proceedings of the 2003 conference on Applications*, technologies, architectures, and protocols for computer communications, p. 175-186, ISBN 1-58113-735-4, 2003.

[27] LOSER, A.; et al. “Efficient data store discovery in a scientific P2P network”, *Proceedings of the WS on Semantic Web Technologies for Searching and Retrieving Scientific Data*, CEUR WS 83, 2003.

[28] SIEBES, R. “pnear: combining content clustering and distributed hash tables”, *In The Second International Workshop on to-Peer Knowledge Management (P2PKM05)*, 2005.

[29] RATNASAMY S.; et al. “A scalable content addressable network”, *Processings of the ACM SIGCOMM*, p. 161-172, 2001.

[30] ZHAO, B. Y.; et al. “Tapestry: A resilient global-scale overlay for service deployment”, *IEEE Journal on Selected Areas in Communications*, v. 22, n. 1, p. 41-53, 2004.

[31] STOICA, I.; et al. “Chord: A scalable peer-to-peer lookup protocol for internet applications”, *IEEE/ACM Transactions on Networking*, v. 11, n. 1, p. 17-32, 2003.

[32] ROWSTRON, A.; DRUSCHEL, P. “Pastry: Scalable, distributed object location and routing for large-scale peer-to-peer systems”, *Proceedings of the Middleware*, 2001.

[33] MAYMOUKOV, P.; MAZIREZ, D. “Kademlia: A peer-to-peer information system based on the xor metric”, *Processings of the IPTPS*, Cambridge, MA, USA, p. 53-65, 2002.

[34] MALKHI, D.; NAOR, M.; RATAJCZAK, D. “Viceroy: a scalable and dynamic emulation of the butterfly”, *Processings of the ACM PODC’02*, Monterey, CA, USA, p. 183-192, 2002.

[35] RIPEANU, M. “Peer-to-peer architecture case study: Gnutella network”, *Peer-to-Peer Computing. First International Conference on Computing*, p 99-100, 2001.

[36] CRESPO, A.; GARCIA-MOLINA, H. “Semantic Overlay Networks for P2P Systems”, *Lecture Notes in Computer Science*, v. 3601, p. 1-13, ISBN 10.1007/11574781, 2005.

[37] SIEBES, R.; HAASE, P.; HARMLEN, F. “Expertise-Based Peer Selection”, *Semantic Web and Peer to peer*, p. 125-142, ISBN 978-3-540-28346-1, 2006.

[38] ABERER, K.; CUDR-MAUROUX, P.; HAUSWIRTH, M. “The chatty web: Emergent semantics through gossipeing”, *Proceedings of the Twelfth International World Wide Web Conference (WWW2003)*, Budapest, Hungary, 2003.

[39] BERNSTEIN, P. A.; et al. “Data management for peer-to-peer computing: A vision”, *Proceedings of the Fifth International Workshop on the Web and Databases, Madison*, Wisconsin, 2002.

[40] HALEVY, A. Y.; et al. “Piazza: Data management infrastructure for semantic web applications”, *Proceedings of the Twelfth International World Wide Web Conference (WWW2003)*, Budapest, Hungary, 2003.

[41] ALEXANDER, L.; et al. “Information Integration in o Schema-Based Peer-To-Peer Networks”, *Proceedings of the 15th International Conference of Advanced Information Systems Engieering (CAiSE 03)*, Klagenfurt, 2003.

[42] NEJDL, W.; et al. “Super-Peer-Based Routing and Clustering Strategies for RDF-Based Peer-To-Peer Networks”, *Proceedings of the Twelfth International World Wide Web Conference (WWW2003)*, Budapest, Hungary, 2003.

[43] NAOR, G., S., M.; WIEDER U. “Know thy Neighbor's Neighbor: the Power of Lookahead in Randomized P2P Networks”, *STOC*, p. 54-63, 2004.

[44] MCILLRAITH, S. A.; SON, T. C.; ZENG, H. “Semantic Web Services”, *IEEE Intelligent Systems*, v. 16, p. 46-53, 2001.

[45] ALMEIDA, M. B.; BAX, M. P. “Taxonomia para projetos de integração de fontes de dados baseados em ontologias”, *V Encontro Nacional de Pesquisa em Ciência da Informação*. 2003.

[46] GRUBER, T. “A Translation Approach to Portable Ontology Specification”, *Proceedings of Japanese Knowledge Acquisition Workshop (JKAW92)*, 1992.

[47] GRUBER, T. R. “Toward Principles for the Design of Ontologies Used for Knowledge Sharing”, *International Journal Human-Computer Studies*, v. 43, n. 5-6, p. 907-928, 1995.

[48] BORST, W. N. “Construction of Engineering Ontologies”. Disponível em http://www.ub.utwente.nl/webdocs/inf/1/t0000004.pdf, acessado em 10 de Outubro de 2010.

[49] LAMBRIX, P.; et al. “Biological Ontologies”, *Semantic Web: Revolutionizing Knowledge Discovery in the Life Sciences*, p. 85-99, 2007.

[50] HORROCKS, I. “OWL: A Description Logic Based Ontology Language”, *International Conference on Logic Programming*, p. 1-4, 2005.

[51] GOMEZ-PEREZ, A.; CORCHO, O. “Ontology Languages for the Semantic Web”, *IEEE Intelligent Systems*, v. 17, p. 54-60, 2002.

[52] PROTÉGÉ. 2010. “The Protégé Ontology Editor and Knowledge Acquisition System”, Disponível em http://protege.stanford.edu, acessado em 15 de Junho de 2010.

[53] EHRIG, M.; et al. “The SWAP data and metadata model for semantics based peer-to-peer systems”, *First German Conference on Multiagent Technologies*, Erfurt, Germany, 2003.

[54] HOTHO, A.; STAAB, S.; STUMME, G. “Ontologies improve text document clustering”, *Proceedings of the International Conference on Data Mining*, IEEE Press, 2003.

[55] EHRIG, M.; STAAB, S. “Satisficing Ontology Mapping. Semantic Web and Peer to peer”, p. 217-233 ISBN 978-3-540-28346-1, 2006.

[56] JASON, J.; JUNG, J. “Reusing ontology mappings for query routing in semantic peer-to-peer environment”, *Information Sciences*, v. 180, n. 17, p. 3248-3257, 2010.

[57] SHVAIKO, P.; EUZENAT J. “A survey of schema-based matching approaches”, *Journal on Data Semantics*, p. 146-171, 2005.

[58] SHVAIKO, P.; et al. “OpenKnowledge Deliverable 3.1.: Dynamic Ontology Matching: a Survey”, *Technical Report*, DIT-06-046, 2006.

[59] BARTINI, C.; LENZERINI, M.; NAVATHE, S. B. “A comparative analysis of methodologies for database schema integration”, *ACM Computing Surveys*, v. 18, p. 323-364, 1986.

[60] MITRA, P.; WIEDERHOLD, G.; MARTIN, K. “A graph-oriented model for articulation of ontology interdependencies”, *Lecture Notes in Computer Science*, 2000.

[61] HAASE, P.; et al. “Bibster - A Semantics-Based Bibliographic Peer-to-Peer System”, *Semantic Web and Peer to peer*, p. 349-363, ISBN 978-3-540-28346-1, 2006.

[62] BONIFACIO, M. “A Peer-to-Peer Solution for Distributed Knowledge Management”, *Semantic Web and Peer to peer*, p. 323-334, ISBN 978-3-540-28346-1, 2006.

[63] FRÉNOT S.; ROYON, Y. "Component Deployment Using a Peer-to-Peer Overlay", *Lecture Notes in Computer Science*, v. 3798, p. 33-36, 2005.

[64] HAASE, P.; et al. "Bibster - A Semantics-Based Bibliographic Peer-to-Peer System", *The Second Workshop on Semantics in Peer-to-Peer and Grid Computing*, 2004.

[65] ARUMUGAM, M.; SHETH, A.; ARPINAR, B. "Towards Peer-to-Peer Semantic Web: A Distributed Environmentfor Sharing Semantic Knowledge on the Web", *Technical Report*, 2001.

[66] YANG, Y.; et al. "Peer-to-Peer Based Grid Workflow Runtime Environment of SwinDeW-G", *Proceedings of the Third IEEE International Conference on e-Science and Grid Computing*, p. 51-58, 2007.

[67] YANG, X.; WANG, L.; JIE, W. "Guide to e-Science: Next Generation Scientific Research and Discovery", *Springer*, ISBN 978-0857294388, 2011.

[68] WORLD WIDE WEB CONSORTIUM . "OWL-S: Semantic Markup for Web Services", 22 de Novembro de 2004.

[69] AALST, W.; HEE, K. "Workflow Management: Models, Methods, and Systems", *The MIT Press*, ISBN 0-262-72046-9, 2002.

[70] ALTINTAS, I.; et al. "Kepler: an extensible system for design and execution of scientific workflows". *Scientific and Statistical Database Management*, p. 423-424, 2004.

[71] OINN, T.; et al. "Taverna: lessons in creating a workflow environment for the life sciences". *Concurrency and Computation: Practice and Experience*, v. 18, n. 10, p. 1067-1100, 2006.

[72] URBAN, S. D.; et al. "Interconnection of Distributed Components: An Overview of Current Middleware Solutions", *Journal of Computer and Information Sciences and Engineering*, v. 1, p. 23–31, 2001.

[73] RFC 675. "Specification of Internet Transmission Control Protocol", Dezembro de 1974. Disponível em http://tools.ietf.org/html/rfc675 acessado 10 de Janeiro de 2012.

[74] CHIRITA, P.; et al. "Designing Semantic Publish/Subscribe Networks Using Super-Peers", *Semantic Web and Peer to peer*, p. 159-179, 978-3-540-28346-1, 2006.

[75] MONTRESOR, A. "A Robust Protocol for Building Superpeer Overlay Topologies", *Proceedings of the 4th International Conference on Peer-to-Peer Computing*, Zurich, Switzerland, 2004.

[76] WACHE, H.; et al. “Ontology-based integration of information - a survey of existing approaches”, *Proceedings of the workshop on Ontologies and Information Sharing at IJCAI*, p. 108-117, 2001.

[77] DO, H.; MELNIK, S.; RAHM, E. “Comparison of schema matching evaluations”, *Proceedings of the second int. workshop on Web Databases (German Informatics Society)*, 2002.

[78] MCCALLUM, A.; NIGAM, K.; LYLE, H. U. “Efficient clustering of high-dimensional data sets with application to reference matching”, *Knowledge Discovery and Data Mining*, p. 169-178, 2000.

[79] HE, B.; CHANG, K. C. C. “Automatic complex schema matching across web query interfaces: A correlation mining approach”, *ACM Transactions on Database Systems*, v. 31, p. 1-45, 2006.

[80] KALFGLOU, Y.; SCHORLEMMER, M. “IF-Map: An ontology-mapping method based on information-flow theory”, *Journal on Data Semantics I*, p. 98–127, 2003.

[81] KANG, J.; NAUGHTON, J. F. “On schema matching with opaque column names and data values”, *Proceedings of SIGMOD*, p. 205-216, 2003.

[82] MADHAVAN, J.; BERNSTEIN, P.; RAHM, E. “Generic schema matching with Cupid”, *Proceedings of VLDB*, p. 49-58, 2001.

[83] MELNIK, S.; GARCIA-MOLINA, H.; RAHM, E. “Similarity flooding: A versatile graph matching algorithm”, *Proceedings of ICDE*, p. 117-128, 2002

[84] NOY, N.; MUSEUN, M. “Anchor-PROMPT: using non-local context for semantic matching”, *Proceedings of the workshop on Ontologies and Information Sharing at IJCAI*, p. 63-70, 2001.

[85] DOAN, A.; DOMINGOS, P.; HALEVY, A. “Learning to match the schemas of data sources: A multistrategy approach”, *VLDB Journal*, v. 50, p. 279-301, 2003.

[86] ZHONG J.; YING, H. "A Semantic Web based Peer-to-Peer Service Registry Network", *SKG '05 Proceedings of the First International Conference on Semantics, Knowledge and Grid*, p. 122, 2005.

[87] VERMA, K.; et al. "METEOR-S WSDI: A Scalable P2P Infrastructure of Registries for Semantic Publication and Discovery of Web Services", *Journal of Information Technology and Management*, v. 6, n. 1, p. 17-39, 2005.

[88] ALVAREZ, D.; SMUKLER, A.; VAISMAN, A. A. "Peer-To-Peer Databases for e-Science: a Biodiversity Case Study", *Brazilian Symposium on Databases - SBBD*, p. 220-234, 2005.

[89] NTARMOS, N.; TRIANTAFILLOU, P. "AESOP: Altruism-Endowed Self Organizing Peers", Proceedings of the 2nd International Workshop on Databases, Information Systems and Peer-to-Peer Computing, 2004.

[90] BOUQUET, P.; SERAFINI, L.; ZANOBINI, S. "Semantic Coordination of Heterogeneous Classifications Schemas", *Semantic Web and Peer to peer*, p. 185-200, ISBN 978-3-540-28346-1, 2006.

[91] ABERER, K.; CUDRÉ-MAUROUX, P.; HAUSWIRTH, M. "Semantic Gossiping: Fostering Semantic Interoperability in Peer Data Management Systems", *Semantic Web and Peer to peer*, p. 259-275, ISBN 978-3-540-28346-1, 2006.