

# Conceitos e tecnologias da *Web* semântica no contexto da colaboração acadêmico-científica: um estudo da plataforma Vivo

## *Concepts and technologies of the semantic Web for academic-scientific cooperation: A study within the Vivo platform*

José Eduardo SANTARÉM SEGUNDO<sup>1</sup>

Caio Saraiva CONEGLIAN<sup>2</sup>

Elaine Rosângela de Oliveira LUCAS<sup>3</sup>

### Resumo

A organização e disponibilização das informações são fundamentais para que os dados possam ser posteriormente recuperados. No contexto acadêmico, alunos, pesquisadores e gestores possuem diferentes necessidades quanto ao acesso às produções científicas. Tais necessidades são tratadas em diferentes ambientes da produção acadêmico-científica, disponível em bases de dados que normalmente não apresentam interoperabilidade entre si. Nesse sentido, as tecnologias e conceitos da *Web* Semântica apresentam uma nova perspectiva em relação à associação das informações contidas nessas bases de dados. Este artigo tem como objetivo apresentar os conceitos e tecnologias da *Web* Semântica que contribuem para a construção de redes semânticas colaborativas e para a organização das informações, identificando como plataformas semânticas se comportam nessa tarefa. Para isso, foi realizada uma prova de conceito na plataforma VIVO (Duraspace, Phoenix, Arizona, Estados Unidos da América), onde foram desenvolvidos testes e análises, utilizando-se um conjunto de publicações científicas de periódicos brasileiros da área da Ciência da Informação. Os testes mostraram uma grande aderência entre as tecnologias da *Web* Semântica e a proposta de organização executada em todas as interações entre o sistema e o usuário, permitindo inclusive a criação de outras relações, com o auxílio de linguagens de recuperação de dados estruturados. A partir dos resultados obtidos foi possível verificar que a *Web* Semântica pode proporcionar avanços em diversos níveis do ambiente acadêmico-científico, especialmente naqueles ligados aos temas: colaboração científica, estudos métricos e organização das informações.

**Palavras-chave:** Colaboração científica. Ontologia. Organização da Informação. Plataforma VIVO. *Web* Semântica.

### Abstract

*The organization and availability of information is fundamental to retrieve data at a later point in time. Within the academic context, students, researchers and managers have different needs regarding access to scientific productions. To address these needs, academic-*

<sup>1</sup> Universidade de São Paulo, Faculdade de Filosofia, Ciências e Letras, Departamento de Educação, Informação e Comunicação. Av. Bandeirantes, 3900, Monte Alegre, 14040-901, Ribeirão Preto, SP, Brasil. Correspondência para/Correspondence to: J.E. SANTARÉM SEGUNDO. E-mail: <santarem@usp.br>.

<sup>2</sup> Universidade Estadual Paulista Júlio de Mesquita Filho, Faculdade de Filosofia e Ciências, Departamento de Ciência da Informação, Marília, SP, Brasil.

<sup>3</sup> Universidade do Estado de Santa Catarina, Centro de Ciências Humanas e da Educação, Departamento de Biblioteconomia e Gestão da Informação, Florianópolis, SC, Brasil.

Apoio: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Processo nº 484899/2013-9) e à Fundação de Amparo à Pesquisa do Estado de São Paulo (Processo nº 2015/01517-2).

Recebido em 9/3/2016, reapresentado em 3/1/2017 e aprovado para publicação em 3/3/2017.

scientific production is available on databases in different environments but they are not normally interoperable with one another. In this sense, the technologies and concepts of the Semantic Web present a new perspective regarding the association of information contained in these databases. The aim of the article is to discuss the concepts and technologies of the Semantic Web that contribute to the development of collaborative semantic networks and information organization and identify how semantic platforms behave during this task. Therefore, a Proof of Concept was carried out on the VIVO (Duraspace, Phoenix, Arizona, United States of America) platform in which tests and analyses were conducted using a set of scientific publications from Brazilian journals in the field of Information Science. The tests showed great adherence to Semantic Web technologies and organization of all the interactions between the system and the user, allowing the creation of other relations with the aid of languages for structured data retrieval. With the results obtained it was found that the Semantic Web can provide advances in several levels of the academic-scientific environment, particularly those related to the following topics: scientific cooperation, metric studies, and information organization.

**Keywords:** Scientific cooperation. Ontology. Information organization. VIVO platform. Semantic Web.

## Introdução

Considerada a perspectiva de aumento da disponibilização de dados na *Web*, a Ciência da Informação, unida a outras ciências, como Ciência da Computação e Linguística, tem o desafio de encontrar meios eficientes para que os usuários consigam localizar e acessar as informações que sejam de seu interesse.

A *Web Semântica* posiciona-se como uma das propostas que visam promover a recuperação, processamento e mediação da informação em benefício dos usuários. Essa proposta tem como objetivo possibilitar que computadores 'compreendam' o sentido dos dados, de acordo com os domínios específicos onde estão inseridos, fornecendo ao usuário informações refinadas, com um agregado maior de conhecimento (BERNERS-LEE et al., 2001). Uma das premissas dos conceitos e tecnologias que envolvem a *Web Semântica* é dar estrutura e significado aos dados a serem publicados, de forma que o próprio computador processe e relacione as informações, levando em consideração o contexto.

A área acadêmico-científica é uma das que sofrem com a falta de organização e estrutura dos dados publicados na *Web*. Apesar da difusão do conceito de *Open Access* e de ferramentas e tecnologias que permitem a promoção do acesso ao conhecimento científico, entre elas o *Open Journals Systems* e o *DSPACE* (Duraspace, Phoenix, Arizona, Estados Unidos da América), existe uma grande dificuldade em integrar as diversas bases de documentos existentes. Embora o uso de repositórios tenha tornado possível aos pesquisadores o depósito, a disseminação e a recuperação de documentos, estes não são interligados

a informações de outras fontes - periódicos, eventos, currículos de autores, temas de pesquisa, entre outros -, o que seria de grande auxílio para a descoberta de dados das produções científicas, se fossem integrados e relacionados semanticamente.

A utilização de aplicações que utilizam conceitos e tecnologias da *Web Semântica* é uma das soluções para essa dificuldade. Entre elas está a plataforma VIVO (Duraspace, Phoenix, Arizona, Estados Unidos da América), que utiliza tecnologias como *Resource Description Framework* (RDF), *Web Ontology Language* (OWL) e *Uniform Resource Identifier* (URI) para descrever e relacionar recursos, possibilitando a interligação entre todos os dados armazenados dentro do sistema. Essa plataforma faz uso de diversos meios para definir uma estrutura de ontologias, com o intuito de representar as informações e, assim, poder estruturá-las semanticamente. Dessa forma, o objetivo do relato descrito é analisar como os conceitos e as tecnologias da *Web Semântica* contribuem no processo de construção de redes semânticas e na organização de informações, auxiliando os usuários a acessar informações de maior relevância.

## Web Semântica

A disponibilização das informações somente para a leitura humana tornou-se uma característica contrária à proposta inicial da *Web*, que, quando criada, explicitava a necessidade de que as máquinas também fossem capazes de ler e trabalhar com os dados.

Nesse contexto, Berners-Lee et al. (2001) propuseram o modelo de *Web Semântica*. Tal proposta definia como a *Web* poderia auxiliar o usuário a realizar tarefas cotidianas, sendo assessorado por mecanismos

computacionais chamados de agentes. Além disso, as informações deveriam estar todas interligadas e com fácil acesso, sendo todos os dados interoperáveis.

Posteriormente à sua proposta inicial, a *Web Semântica* foi se difundindo e agregando diversos conceitos, tecnologias e funcionalidades, com o intuito de tornar-se implementável e, assim, permitir a criação de ambientes que tenham características semânticas. Entre as tecnologias, destacam-se: *Resource Description Framework* (RDF), *eXtensible Markup Language* (XML), *SPARQL Protocol and RDF Query Language* [SPARQL], OWL e diversos outros conceitos descritos pelo *World Wide Web Consortium* (W3C), consórcio que administra a *Web*.

Nesse sentido, as tecnologias da *Web Semântica* têm um papel central para a aplicação dos ideais propostos inicialmente. Para atender à semântica formal de descrição da informação, foi desenvolvido o RDF, apresentando como base a estrutura de triplas, que relacionam duas informações por meio de uma propriedade. Em suma, as triplas RDF são compostas por sujeito, predicado e objeto, relacionando, assim, recurso (geralmente digital) a seu respectivo valor por meio de um predicado. Embora o RDF possa ser representado por diversas notações, a mais utilizada é o XML, justamente pela capacidade computacional dessa linguagem, sendo padrão de comunicação na maioria dos sistemas *Web*.

Outra tecnologia de importância ímpar para a *Web Semântica* é o SPARQL, protocolo responsável pela manipulação e recuperação de dados dentro de ambientes que utilizam as tecnologias supracitadas. O SPARQL se baseia na estrutura do RDF para fazer as buscas, relacionando, por meio das triplas RDF, as informações a serem procuradas com aquelas já definidas como ponto de partida para a busca.

Embora muitas dessas tecnologias sejam bastante difundidas e conhecidas, poucos ambientes digitais fazem uso efetivo e seguem o modelo proposto pela W3C para a *Web Semântica*. O desafio, então, está sendo a criação da *Web de Dados*, que tem como objetivo a interligação dos dados com todas as relações explícitas para os programas computacionais. A *Web de Dados* vem com o intuito de aperfeiçoar a *Web de Documentos*, que é basicamente um depósito de

documentos armazenados em *websites* e que podem ser recuperados por meio de serviços de buscas sintáticos (WORLD WIDE WEB CONSORTIUM, 2011).

Por meio do uso das tecnologias citadas, as ontologias caracterizam-se como elemento de organização dos dados semanticamente relacionados. Santarém Segundo (2015, p.226) afirma que: "Utilizar ontologias é uma das maneiras de se construir uma relação organizada entre termos dentro de um domínio, favorecendo a possibilidade de contextualizar os dados, tornando mais eficiente e facilitando o processo de interpretação dos dados pelas ferramentas de recuperação da informação". Santarém Segundo e Coneglian (2015, p.227) indicam que "[...] para o uso como tecnologia da *Web Semântica*, entende-se as ontologias como: artefatos computacionais que descrevem um domínio do conhecimento de forma estruturada, através de: classes, propriedades, relações, restrições, axiomas e instâncias".

Atualmente, a principal linguagem utilizada para a construção de ontologias é o OWL, capaz de implementar computacionalmente todas as características relatadas pelos autores. O OWL tem como base o RDF para a definição das relações, além de apresentar uma série de propriedades que inserem lógica e axiomas nas relações existentes. O uso de ontologias estruturadas em OWL torna possível uma descrição do contexto de um domínio com uma semântica formal bastante representativa.

A utilização das tecnologias da *Web Semântica* possibilita a construção de aplicações que tornam a tarefa das máquinas mais intuitiva e, como consequência, mais eficiente em auxiliar o ser humano em tarefas básicas. O uso das tecnologias da *Web Semântica* permite a realização de inferências a respeito dos dados, e assim as informações que são apresentadas aos usuários denotam maior relevância (WORLD WIDE WEB CONSORTIUM, 2011).

### **Colaboração Digital na Academia**

A partir da metade da segunda década do Século XXI, a evolução das tecnologias revolucionou a forma de comunicação entre as pessoas espalhadas

pelo globo. Por meio de ferramentas de mensagens instantâneas e de compartilhamento e colaboração de documentos em tempo real, as trocas de ideias e a cooperação entre pesquisadores passou a ser mais fácil, aumentando a capacidade de produzir pesquisa em colaboração.

O progresso das tecnologias proporcionou também um avanço no modo como documentos podem ser disponibilizados, armazenados e acessados digitalmente por meio da *Internet*. Nesse cenário, torna-se imprescindível que um documento esteja disponibilizado em meios digitais, fazendo com que o compartilhamento das informações seja algo cada vez mais comum e necessário (CATARINO; BAPTISTA, 2007).

Dentro do ambiente acadêmico, há uma necessidade crescente de tecnologias que permitam a expansão da colaboração digital, tornando-a mais acessível a um maior número de pesquisadores. Tal necessidade existe devido à relevância que a disponibilização das informações possui para a ciência, com o intuito de difundir mais efetivamente os conhecimentos produzidos. Como consequência dessa difusão, passam a ser reais o incentivo a trabalhos colaborativos e a diminuição do retrabalho em pesquisas em andamento ou concomitantes.

Nesse contexto, muitas propostas estão surgindo com o intuito de divulgar produções científicas abertas, para que um maior número de pessoas tenha acesso aos conhecimentos produzidos na academia. Uma dessas iniciativas são as publicações de acesso aberto (*open access*), que possibilitam a qualquer usuário o acesso irrestrito a artigos publicados em periódicos que seguem tais princípios, tendo como objetivo a disseminação ampla e irrestrita de resultados de pesquisa científica (COSTA, 2006).

Os repositórios, uma das iniciativas baseadas no modelo de acesso aberto, em sua maioria reúnem a produção científica temática ou institucional de uma organização, de forma a disponibilizá-la para as mais variadas comunidades. Esta tecnologia se mostra muito interessante para que o conhecimento produzido em uma instituição atravesse as fronteiras da universidade e do país, podendo influenciar uma gama grande de pesquisadores (LEITE, 2009).

O *Open Journals Systems* - principal ferramenta para publicação de revistas eletrônicas em formato aberto ao redor do mundo - apresenta funções distintas dos repositórios digitais. Entretanto, apresenta-se como figura importante e imprescindível no processo de compartilhamento e acesso facilitado à produção científica. No Brasil é possível identificar, por meio da Plataforma *Lattes*, uma plataforma pública de currículos, quais são os projetos, as áreas de interesse e pesquisa, a produção e o currículo de um pesquisador. No entanto, embora a Plataforma ofereça tais informações, ela ainda é restrita quando se pensa em relacionamentos ou interoperabilidade entre as produções e projetos realizados em colaboração - ou seja, qualquer estudo mais aprofundado que relacione informações de pesquisa e pesquisadores demanda trabalho extremo ou, em alguns casos, torna-se inviável. Apesar dos últimos avanços da Plataforma *Lattes*, principalmente na proposta de relacionar coautoria e implementar gráficos, isso ainda é muito pouco para atender às demandas da comunidade científica. Essas melhorias podem ser encaradas como uma versão rudimentar da semântica oferecida pela plataforma VIVO.

Os Estudos Métricos desenvolvidos no âmbito da Ciência da Informação acabam por se tornar bastante complexos do ponto de vista da coleta de dados. A maior dificuldade apontada pela literatura é a falta de interoperabilidade entre os sistemas que armazenam os dados científicos. Para a coleta e posterior análise, é comum o pesquisador realizar buscas manuais em diferentes bases de dados, exigindo diversas vezes que os documentos sejam verificados um a um.

Nesse cenário, o uso de plataformas que utilizam conceitos da *Web Semântica* com o objetivo de organizar as produções científicas - apresentando e explicitando os relacionamentos existentes - pode ser um grande auxílio para que os pesquisadores e a comunidade em geral acessem informações relevantes. Um exemplo de plataforma que atende a tais requisitos é a VIVO *Semantic Web*.

### **Plataforma VIVO: histórico e descrição**

A VIVO é uma plataforma *open-source* desenvolvida e implementada pela *Cornell University*, com o

intuito de permitir o acesso a informações de pesquisadores, pesquisas, fomentos e trabalhos acadêmicos. Basicamente, a plataforma VIVO é um ambiente informacional que permite a inserção de informações de uma instituição de ensino superior, atingindo todas as atividades, funções e pessoas que estão envolvidas com as atividades da academia. A plataforma agrupa e disponibiliza informações sobre pesquisadores e professores, publicações e eventos, disciplinas, cursos e currículos, dentre outras; assim, organiza e relaciona documentos e informações de todos os âmbitos dentro do ambiente acadêmico.

A plataforma começou em 2013, com um projeto que tinha a intenção de apresentar informações sobre pesquisas e pesquisadores, além de todos os relacionamentos existentes entre essas informações. Em 2009, houve um grande financiamento ao projeto, e diversas universidades se uniram para aprimorar a ferramenta. Apoiada nesses incentivos, a plataforma foi desenvolvida com instrumentos sofisticados de visualização, que possibilitam grande interação entre o programa e o usuário (BÖRNER *et al.*, 2012).

Uma das peculiaridades do projeto é que todas as informações disponibilizadas são públicas, permitindo a qualquer usuário acessar os dados, sem necessidade de autenticação no sistema. A utilização da plataforma possibilita que as relações semânticas entre os dados acadêmicos de uma instituição fiquem disponíveis e abertas aos usuários e programas computacionais, criando, assim, uma rede completamente relacionada e interligada, em que os dados e suas relações ficam sempre visíveis aos sistemas de busca.

### Funcionamento da Plataforma

Os dados armazenados na plataforma VIVO estão estruturados em RDF, permitindo, assim, que as informações fiquem todas interligadas, com os seus relacionamentos explícitos. Devido ao fato de a representação dos dados ser feita no formato RDF, a plataforma possibilita que todas as páginas dentro do sistema sejam visualizadas tanto em HTML quanto no próprio formato RDF.

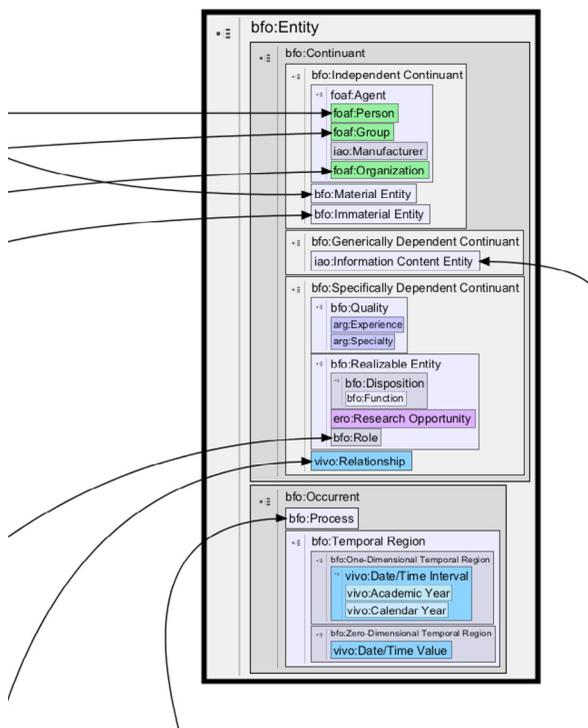
O uso de ontologias condiciona a inserção de novos dados na plataforma, que devem ser

associados a classes. Como consequência, todas as informações introduzidas no sistema passam a estar inter-relacionadas semanticamente. A VIVO faz uso de diversas ontologias para tratar as relações semânticas dos dados, além de utilizar uma ontologia própria, chamada de ontologia VIVO, desenhada especialmente para essa aplicação, relacionando diversas outras ontologias. Essa ontologia própria foi construída com o intuito de tratar as relações entre as diversas áreas que estão dentro do universo acadêmico, demonstrando todos os conceitos relacionados às informações de pesquisadores, organizações, eventos e diversas outras atividades.

A ontologia VIVO foi construída na linguagem OWL, o que permite um nível de detalhes e propriedades bastante elevado nas relações semânticas. Destaca-se que, em razão de a base de dados estar totalmente disponível em triplas RDF, é possível trabalhar com aplicações complexas, usando-se o protocolo SPARQL para identificar as relações existentes, criar novos relacionamentos e realizar inferências a respeito dos dados, entre outras ações semanticamente ricas.

O processo de criação da ontologia VIVO já foi pensado de forma que ela fosse independente e interoperável, ou seja, que pudesse trabalhar sobre outras aplicações. Esse aspecto permite que ela seja utilizada em sistemas com focos diferentes dos da plataforma VIVO, como sistemas comerciais e empresariais. Como mencionado, a ontologia VIVO é composta por diversas outras ontologias e vocabulários, realizando a integração com os dados contidos no domínio acadêmico. Existem quinze ontologias que compõem a ontologia VIVO: *Bibontology* (<http://purl.org/ontology/bibo/>); *Citation Counting and Context Characterization Ontology* (<http://purl.org/spar/c4o/>); *Citation Typing Ontology* (CiTO) (<http://purl.org/spar/cito/>); *Event Ontology* (<http://purl.org/NET/c4dm/event.owl#>); *FRBR-Aligned Bibliographic Ontology* (FaBio) (<http://purl.org/spar/fabio/>); *Friend of a Friend* (FOAF) (<http://xmlns.com/foaf/0.1/>); *Geopolitical Ontology* (<http://aims.fao.org/aos/geopolitical.owl#>); *OBO Foundry* (<http://purl.obolibrary.org/obo/>); *OCRe research* (<http://purl.org/net/OCRe/research.owl#>); *OCRe study design* ([http://purl.org/net/OCRe/study\\_design.owl#](http://purl.org/net/OCRe/study_design.owl#)); *SKOS (Simple Knowledge Organization System)* (<http://www.w3.org/2004/02/skos/core#>); *VCard* (<http://www.w3.org/2006/>

vcard/ns#); *Vitro Public Ontology* (<http://vitro.mannlib.cornell.edu/ns/vitro/public#>); *VIVO Core* (<http://vivoweb.org/ontology/core#>); e *VIVO Scientific Research Ontology* (<http://vivoweb.org/ontology/scientific-research#>). A partir da identificação dessas ontologias, é possível verificar os diversos domínios e áreas que são tratadas dentro da ontologia VIVO. Um fragmento dessa ontologia, com suas classes e relações, está esquematizado na Figura 1.



**Figura 1.** Fragmento das relações entre as diversas ontologias da plataforma VIVO.

Fonte: Adaptado de DURASPACE (2014).

Na Figura 1 é possível visualizar, por exemplo, o *Friend of a Friend* (FOAF) e o *Bibontology*, dentre outros vocabulários que se relacionam e trabalham em função da ontologia VIVO. É construída uma rede muito concisa, com relacionamentos fortes que, quando aplicados, estabelecem relações semânticas bastante significativas.

Outro aspecto interessante é que as diversas ontologias utilizadas tratam de áreas distintas umas das outras, pois toda informação incluída no sistema deverá ser representada por meio de uma classe da ontologia. Portanto, a estrutura ontológica deve abranger todos os tipos de dados que serão adicionados.

Analisando a Figura 1, pode-se verificar que a classe entidade (*Basic Formal* [bfo:entity]) é o centro da ontologia, originando a maior parte dos relacionamentos. Outro ponto de destaque são as especificações das classes mais genéricas da ontologia. Essa característica mostra o grau de especificidade que a ontologia e a própria plataforma VIVO apresentam durante a inclusão e descrição dos itens no sistema.

## Principais Módulos da Plataforma

A aplicação de diversos conceitos e tecnologias da *Web Semântica* possibilita à plataforma VIVO múltiplas ações, que atualmente vão além dos processos realizados pelos sistemas de informação tradicionais utilizados em diversas esferas organizacionais. A plataforma traz claramente as possibilidades que a *Web Semântica* oferece por meio do uso das suas tecnologias em todos os processos do sistema, com basicamente quatro grandes características: busca facetada, ligação semântica, reuso de dados e visualizações.

## Busca Facetada

A busca facetada utiliza-se de classificações para realizar uma pesquisa, ou seja, quando é realizada uma pesquisa na aplicação, o sistema classifica os resultados em vários *facets* ou várias categorias, apresentando-os aos usuários. Nesse contexto, 'classificar' significa ordenar e dispor em classes um determinado conjunto de informações, onde uma classe contém um número de elementos que possuem alguma característica em comum (TRISTÃO *et al.*, 2004).

Os motores de buscas tradicionais utilizam habitualmente buscas heurísticas ou critérios de relevância durante a apresentação dos resultados aos usuários. A busca facetada se contrapõe a esses mecanismos, pois possibilita ao usuário 'subir' ou 'descer' nas hierarquias das classificações, generalizando ou especificando o nível das informações apresentadas. Como a plataforma VIVO foi construída com conceitos da *Web Semântica*, existindo diversos níveis e tipos de relacionamentos, a busca facetada apresenta-se como a melhor solução para esse sistema, pois permite ao usuário uma navegação muito interativa com os

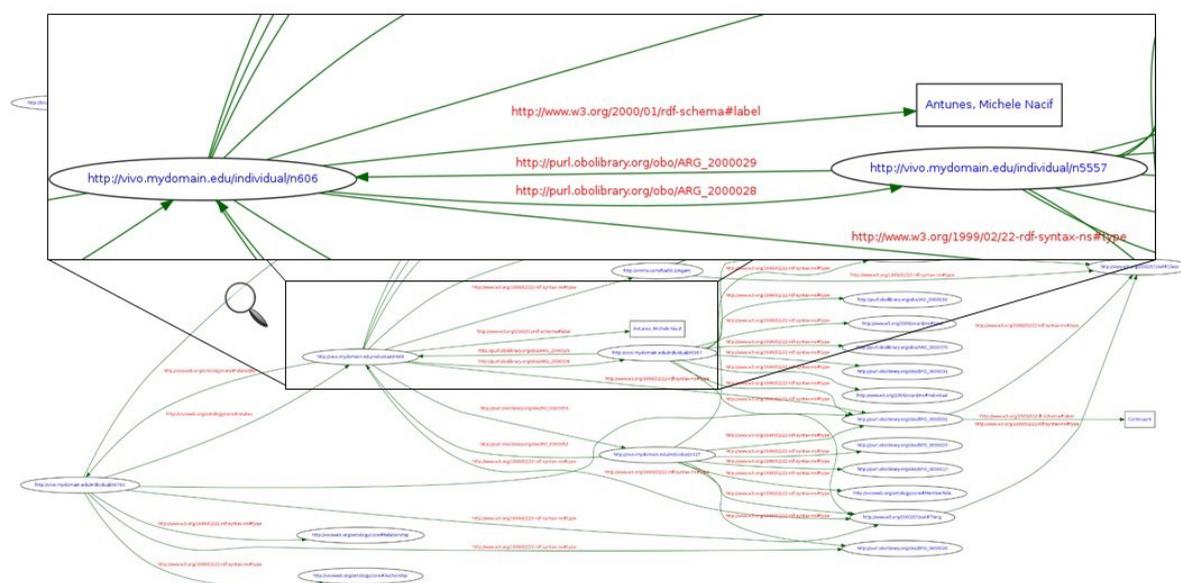
resultados apresentados, podendo verificar as relações existentes entre as informações pesquisadas (BÖRNER *et al.*, 2012).

## Ligação Semântica

Uma característica de extrema relevância dentro da plataforma VIVO é a ligação semântica que existe entre os dados contidos no sistema. Por esse motivo, todos os *links* dentro desse ambiente possuem um significado, e caso uma informação tenha um *link* com

outra, tal relacionamento está acessível e é bidirecional, de maneira que essa ligação semântica simples pode ser fundamental para a descoberta de informações. Assim, o acesso do usuário a uma determinada informação é algo simples e dinâmico, pois todos os *links* estão disponíveis (BÖRNER *et al.*, 2012).

Na Figura 2 visualiza-se como a página de um pesquisador é representada em RDF, por meio de grafos. A Figura mostra que todas as informações estão interligadas, pois se utiliza o RDF como método de armazenamento de informações.



**Figura 2.** Visualização do grafo RDF de um pesquisador.

Fonte: Elaborado pelos autores (2016).

## Reuso dos dados

Outro ponto de destaque na plataforma é a possibilidade de reuso dos dados. As propriedades que esses dados possuem, de serem abertos e estarem em formato RDF, os tornam ideais para que outras ferramentas possam consumir e utilizar essas informações.

A possibilidade de reutilização dos dados permite que sejam construídas aplicações que utilizem todos os dados disponíveis na plataforma, sendo possível, por exemplo, construir uma ferramenta que utiliza os dados da VIVO como base de dados,

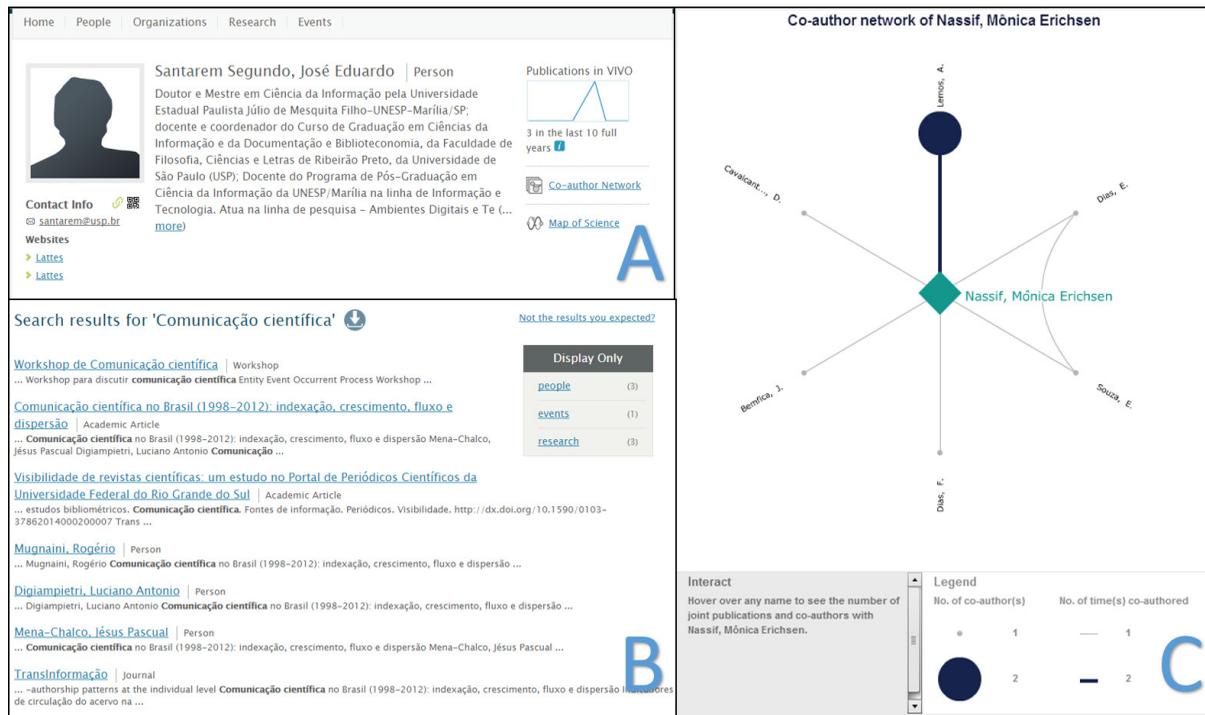
para a realização de estudos bibliométricos ou para verificação automatizada das publicações dos acadêmicos, sem a necessidade de que o pesquisador preencha planilhas para prestação de contas das suas publicações às instituições. Apesar de não seguir os mesmos princípios, a estratégia é a mesma da disponibilização de dados do conjunto de melhores práticas conhecido como *Linked Data*.

## Visualização

Por fim, a última característica considerada essencial na plataforma VIVO é o modo como a

visualização dos dados é apresentada. A plataforma possibilita visualizar redes de pessoas e de conceitos, além da geografia das publicações e dos autores, permitindo, assim, fazer diversas conexões que auxiliam o usuário no entendimento das informações.

A Figura 3 demonstra diversas interfaces visuais do sistema. Destaca-se o quadro com a letra A, que ilustra o perfil de um pesquisador, disponibilizando ao usuário informações gerais sobre ele, tais como currículo, vínculo institucional, mapa científico, rede de coautoria, dentre outras.



**Figura 3.** Visualização do perfil de um pesquisador. Fonte: Elaborado pelos autores (2016).

Outra opção de visualização apresentada pelo sistema são as redes de coautoria de um determinado autor. É possível verificar quais são os autores que dividiram a autoria de algum trabalho científico, com destaque para a visualização dos seus respectivos perfis. Na Figura 3, no quadro C, é possível verificar a rede de coautores de um determinado pesquisador: o losango verde representa o pesquisador central da rede, enquanto os círculos representam os pesquisadores coautores ao primeiro. As linhas que ligam os pesquisadores representam a quantidade de artigos escritos em coautoria: quanto mais espessa a linha, mais artigos em cooperação.

### Prova de Conceito: Metodologia, Testes e Resultados

Usar tecnologias e princípios da *Web Semântica* faz com que o compartilhamento e a recuperação de

informações ocorram de modo mais rápido e simples, acelerando e facilitando os processos para que os agentes computacionais possam ajudar as pessoas a encontrar as informações desejadas.

Para analisar mais profundamente a plataforma VIVO, foi realizada uma Prova de Conceito [*Proof of Concept*] - termo utilizado para denominar o modelo prático que pode provar o conceito teórico estabelecido por uma pesquisa -, a fim de verificar as funcionalidades e tecnologias envolvidas. Esse modelo foi realizado a partir da criação de um ambiente fechado com uma instalação da plataforma VIVO, contemplando três fases principais: a instalação da plataforma em um servidor local; a inserção de artigos científicos de diferentes revistas científicas brasileiras; e a realização de testes de funcionalidades, analisando como as ontologias e a semântica da plataforma se comportam. Estas fases são descritas a seguir.

Primeiramente, foi realizada a instalação e configuração de um servidor com a aplicação VIVO, seguindo o manual de instalação (DURASPACE, 2015). A aplicação ficou disponível apenas para testes no laboratório, estando totalmente indisponível para acessos externos. Após, foram inseridos no sistema 49 artigos publicados em revistas científicas brasileiras da área da Ciência da Informação, que possuíam relação temática ou autoral, dentre outras. Os artigos foram oriundos de dez revistas diferentes, contemplando 123 pesquisadores e 44 vínculos institucionais. Foram adicionados o resumo, as palavras-chave, os autores, o *Digital Object Identifier* (quando disponível), a data de publicação e o título do periódico de todos os artigos publicados. Em seguida, adicionaram-se também as informações relativas aos autores e suas instituições. Dessa forma, verificaram-se as relações entre instituições, publicações, autores e coautores, dentre outras.

Por fim, após a inserção de dados, observou-se que a plataforma VIVO relacionou os materiais acrescidos, através do uso das ontologias de que dispõe. Notou-se que dois artigos que têm o mesmo autor foram semanticamente interligados por um RDF, que explicitou a relações existentes. Como as páginas do sistema são geradas a partir de RDF, todas as informações mostradas ao usuário estão interligadas a outros dados, intensificando as relações semânticas e tornando-as acessíveis para os usuários e, principalmente, para os agentes computacionais. Devido a essas características de descrição utilizadas - ontologias, RDF e relações semânticas -, a plataforma possibilita a utilização do SPARQL, que propicia a descoberta de novas relações. Após as fases descritas, que se concentraram na tarefa de alimentar a plataforma de forma adequada, foram iniciados alguns testes, divididos em duas etapas.

A primeira etapa de testes ocorreu por meio da verificação das relações semânticas produzidas pelo sistema a partir dos dados inseridos. Foi analisado o processo de busca de informações pelos usuários, bem como o inter-relacionamento dos dados e os recursos que podem auxiliar no processo de recuperação da informação, tornando-o mais eficiente.

Na segunda etapa, com o intuito de analisar mais profundamente a semântica dos dados, foram

executadas consultas utilizando o protocolo SPARQL. Foi averiguado se o modo como a ontologia da plataforma VIVO foi construída permite que a recuperação das informações dos RDF seja executada com a semântica clara e, ainda, se os dados recuperados apresentam alta expressividade. Dessa forma, verificou-se a materialização dos conceitos e tecnologias da *Web Semântica*, com o uso de diversas ontologias trabalhando em conjunto, de maneira funcional. Com a análise e identificação do modelo, verificou-se que ele pode ser expandido para vários outros domínios.

Para a verificação, definiu-se que deveriam ser extraídas da base de dados as seguintes relações: quantidade de artigos por pesquisadores; quantidade de artigos por instituições; quantidade de pesquisadores por instituições; relações de coautoria entre as instituições. Como dito, para extrair relações, foram construídas consultas SPARQL, que permitem a execução das tarefas descritas. Para tal, a plataforma VIVO disponibiliza uma interface SPARQL *EndPoint* que, como relatado anteriormente, permite a realização de consultas para recuperação dos dados utilizando essa linguagem. É importante destacar que o uso dessa interface não é uma tarefa trivial; entretanto, o experimento teve a intenção de identificar as possibilidades de acesso. Toda ferramenta que dispõe desse tipo de tecnologia para acesso aos seus dados por meio do protocolo SPARQL está aberta a receber consultas de outros tipos de interfaces de usuários.

### Resultados da Prova de Conceito

Os testes realizados apresentaram resultados interessantes no que diz respeito à semântica dos dados e à própria interação que a plataforma dispõe ao usuário.

Nessa perspectiva, é importante identificar o retorno do sistema quando o usuário realiza uma pesquisa dentro da interface de busca. A pesquisa com o termo 'Competência Informacional' na plataforma VIVO fez com que o sistema apresentasse todos os conteúdos que continham a expressão buscada, partindo de uma busca sintática e, posteriormente, apresentando também informações relacionadas semanticamente aos documentos que continham a

expressão. Na busca do sistema foram recuperados diversos tipos de dados que se relacionavam aos termos descritos pelo usuário, recuperando, por exemplo, autores que escreveram sobre o assunto, eventos que tiveram publicações naquele enfoque, revistas que publicaram artigos sobre o tema, dentre diversas outras relações que a plataforma é capaz de realizar. Essa característica consegue explorar o uso da semântica no processo de recuperação dos dados, pois expande o modo como são realizadas a recuperação e a apresentação dos resultados aos usuários.

Em outro exemplo, caso seja realizada uma busca pelo nome de um pesquisador, serão mostrados ao usuário todos os dados relacionados àquela pessoa, como perfil, publicações, atividade editorial, dentre outros. Um exemplo pode ser visualizado na Figura 3, quadro B, onde são relacionados diversos elementos a partir de uma busca com a expressão 'Comunicação Científica'. Nesse caso, foram relacionados pesquisadores, revistas, eventos e publicações, demonstrando que a plataforma aprimora os níveis de relacionamentos.

No primeiro teste descrito, que observou a recuperação dos dados a partir de uma busca realizada pelo usuário, a plataforma VIVO apresentou

resultados satisfatórios, pois conseguiu relacionar semanticamente os dados, criando vínculos bem ricos e aptos a auxiliar efetivamente o usuário no processo de recuperação da informação. É possível, então, ter acesso a uma quantidade de informações com sustentação semântica, que auxiliem o usuário a encontrar dados que atendam às suas necessidades informacionais.

Posteriormente, o segundo teste utilizou o SPARQL para manipular os dados, criando diversos relacionamentos entre eles. Foi possível, a partir desse teste, tratar de questões de interesse dos estudos métricos da informação. Nesse contexto, a coleta e as análises matemáticas, estatísticas e relacionais, em sua maioria, são executadas por um pesquisador, manualmente ou com o auxílio de sistemas independentes e isolados, sendo que em diversos casos a verificação das informações, como autores, palavras-chave e instituições, ocorre individualmente em cada documento. Contudo, a disponibilização dos dados em RDF e a utilização de vocabulários e ontologias em sua descrição, possibilitam a realização de consultas SPARQL, que permitem inferir e contabilizar as informações desejadas, fornecendo, de forma automatizada, informações essenciais dentro de pesquisas com relações diversas.

```

SELECT ?nome_organizacao1 ?nome_organizacao2 (COUNT(?artigo) AS ?qtd_artigo)
WHERE
{
    {
        SELECT DISTINCT ?nome_organizacao1 ?nome_organizacao2 ?artigo
        WHERE
        {
            ?artigo rdf:type bibo:AcademicArticle;
                vivo:relatedBy ?ligArtigoAutor.
            ?ligArtigoAutor vivo:relates ?autor.

            ?autor rdf:type foaf:Person.

            ?autor obo:RO_0000053 ?ligOrganizacao.
            ?ligOrganizacao vivo:roleContributesTo ?organizacao.
            ?organizacao rdfs:label ?nome_organizacao1.
        }
        {
            ?artigo rdf:type bibo:AcademicArticle;
                vivo:relatedBy ?ligArtigoAutor1.
            ?ligArtigoAutor1 vivo:relates ?autor2.
            ?autor2 rdf:type foaf:Person.

            ?autor2 obo:RO_0000053 ?ligOrganizacao2.
            ?ligOrganizacao2 vivo:roleContributesTo ?organizacao2.
            ?organizacao2 rdfs:label ?nome_organizacao2.
        }
        FILTER (?nome_organizacao1 != ?nome_organizacao2)
    }
}
GROUP BY ?nome_organizacao1 ?nome_organizacao2
ORDER BY DESC(?qtd_artigo)

```

**Figura 4.** Exemplo de consulta SPARQL.

Fonte: Elaborado pelos autores (2016).

Para a investigação, foram construídas, em SPARQL, diversas consultas de relações bibliográficas, tratadas no final da subseção que apresentou a prova de conceito. Na Figura 4 é possível identificar uma das consultas SPARQL realizadas, cujo resultado apresenta as relações de coautoria entre instituições. Nessa Figura é possível verificar como uma consulta SPARQL é construída, identificando as relações existentes e os recursos que essa linguagem oferece.

A Figura 5 apresenta os resultados de diversas consultas realizadas. Vale destacar que os dados obtidos nessa consulta, conforme ressaltado

anteriormente, representam um conjunto simulado, em que foram inseridas as informações referentes a um grupo de artigos publicados nas principais revistas da Ciência da Informação no Brasil. Os resultados obtidos não podem ser utilizados para fins de comparação, pois não foi o objetivo desta pesquisa conferir a produção das organizações ou dos pesquisadores, mas apenas demonstrar a capacidade de construção de relações e compilação de informações dessa plataforma, como consequência do embasamento nos conceitos e nas tecnologias da Web Semântica.

nome_organizacao	qtd_artigo
"Universidade Federal de Minas Gerais"	7
"Universidade Estadual Paulista"	4
"Universidade Federal da Paraíba"	4
"Universidade Federal de Santa Catarina"	4
"Universidade de São Paulo"	4
"Universidade Federal do Estado do Rio de Janeiro"	3
"Fundação Oswaldo Cruz"	2
"Universidad Nacional Autónoma de México"	2
"Universidad de Pinar del Río"	2
"Universidade Federal Fluminense"	2
"Centro de Información y Gestión Tecnológica de Villa Clara"	1
"Centro de Investigación e Innovación en Tecnologías de la Información y Comunicación"	1
"Centro de Investigación y de Estudios Avanzados"	1
"Companhia de Tecnologia da Informação de Cachoeiro de Itapemirim"	1

A

nome_autor	qtd_artigo
"Nassif, Mônica Erichsen"	6
"Santarem Segundo, José Eduardo"	3
"Vidotti, Silvana Aparecida Borsetti Gregorio"	3
"Carrillo-Calvet, Humberto Andrés"	2
"Casarín, Helen Castro Silva"	2
"Díaz-Pérez, Maidelyn"	2
"Giráldez-Reyes, Raudel"	2
"Guimarães, Maria Cristina Soares"	2

B

nome_organizacao	qtd_pesquisador
"Universidade Federal de Santa Catarina"	11
"Universidad de la Republica"	9
"Universidade Federal da Paraíba"	8
"Centro de Información y Gestión Tecnológica de Villa Clara"	5
"Fundação Oswaldo Cruz"	5
"Centro de Investigación y de Estudios Avanzados"	4
"Universidad de Pinar del Río"	4
"Universidade Federal de Minas Gerais"	4
"Universidade Federal de São Carlos"	4
"Universidade Federal do Estado do Rio de Janeiro"	4
"Universidad de Lleida"	3

C

nome_organizacao1	nome_organizacao2	qtd_artigo
"Universidade Estadual Paulista"	"Universidade de São Paulo"	3
"Universidade de São Paulo"	"Universidade Estadual Paulista"	3
"Universidad Nacional Autónoma de México"	"Universidad de Pinar del Río"	2
"Universidad de Pinar del Río"	"Universidad Nacional Autónoma de México"	2
"Centro de Información y Gestión Tecnol"	"Hospital Docente Gineco-obstétrico "Ma"	1
"Centro de Investigación e Innovación en"	"Universidad Nacional Autónoma de México"	1
"Centro de Investigación e Innovación en"	"Universidad de Pinar del Río"	1
"Companhia de Tecnologia da Informação d"	"Universidade Cândido Mendes"	1
"Câmara dos Deputados"	"Universidade Federal de Santa Catarina"	1

D

Figura 5. Quantidade de artigos por organização.

Fonte: Elaborado pelos autores (2016).

Na Figura 5, o quadro A apresenta os resultados da primeira consulta, que relacionam as instituições à quantidade de artigos. É possível visualizar o modo de apresentação dos resultados que a consulta SPARQL retornou, por meio de tabela. Entretanto, todos eles podem ser apresentados também em formato RDF e XML, dentre outros, permitindo que outras ferramentas ou sistemas consumam os dados já com esse formato analítico de representação. No quadro B, apresentam-se os resultados de uma consulta com duas variáveis relacionadas: nomes dos autores e quantidade de artigos publicados por eles. Já o quadro C apresenta a relação existente entre as instituições e os pesquisadores, sendo possível também identificar a quantidade destes por instituição. Por fim, o quadro D apresenta uma relação entre três variáveis, sendo informadas as coautorias entre instituições e verificada a quantidade de artigos que duas instituições publicaram em conjunto.

A partir da Figura 5 verificou-se que a utilização de uma plataforma construída a partir das tecnologias e conceitos da *Web Semântica* permite disponibilizar uma base de dados muito rica, com alta gama de possibilidades. O uso do SPARQL mostrou-se muito eficiente nesse ambiente, pois todos os dados estão estruturados, permitindo que as relações criadas auxiliem significativamente no tratamento das questões relacionadas aos estudos métricos da informação.

Nessa prova de conceito, considera-se também a necessidade de tratamento dos dados para que possam ser apresentados como resultado final de uma análise em relatórios científicos. A proposta aqui apresentada sistematiza possibilidades, sem preocupação com o modo de apresentação dos dados, passo posterior a esta pesquisa.

## Conclusão

As tecnologias e conceitos da *Web Semântica* têm se aproximado de ferramentas e *frameworks*, favorecendo a recuperação semântica da informação pelos usuários de sistemas digitais. Entretanto, os *softwares* de repositórios existentes não permitem a integração de ontologias, com exceção de

customizações específicas que demandam muito investimento.

A *Web Semântica* expande essa visão por meio de tecnologias como RDF, OWL e SPARQL, que, quando aplicadas de forma concisa, transformam o poder de relação dos dados disponíveis. O uso de tais tecnologias integradas a repositórios aumenta o campo de visão e de relacionamento de conceitos, oferecendo aos usuários resultados mais ricos, do ponto de vista de relações semânticas.

A plataforma VIVO apresenta-se como um ambiente inovador e que, alicerçado pelas tecnologias da *Web Semântica*, propõe um novo paradigma estrutural para o armazenamento de documentos. As relações construídas internamente, baseadas em URI, e em tecnologias da *Web Semântica*, permitem que os dados armazenados estabeleçam ligações baseadas em ontologias, que se estruturam em vocabulários internacionalmente conhecidos, como *Bibliographic Ontology* (Bibo), SKOS e FOAF, dentre outros. As ontologias são a base para que o conjunto de documentos depositados na plataforma VIVO assumam uma condição de relação em estruturas baseadas em classes e propriedades. É possível, a partir dessa estrutura, realizar inferências e descobrir novas relações, expandindo a visão sobre as informações que estão armazenadas dentro do sistema.

Outro ponto são os módulos que foram construídos pensando justamente na utilização dos diversos conceitos e tecnologias da *Web Semântica*. Um desses módulos é a busca facetada, que se utiliza de um conceito mais intuitivo do que a realização de algoritmos heurísticos para a realização de buscas. Na busca facetada, o usuário irá interagir com as diversas relações que estão no sistema de forma ampliada.

A visualização das telas do sistema também traz um conceito diferente do encontrado em páginas *Web* tradicionais, pois todas as informações mostradas ao usuário estão representadas também em formato RDF, cujo conteúdo possui um significado claro, permitindo que os agentes computacionais se relacionem com as informações e, como consequência, proporcionem maior interação entre usuário e as interfaces do sistema. Apresentar um SPARQL *Endpoint* para a

execução de consultas torna a ferramenta aberta a pesquisas variadas e a modelos de recuperação amplos e repletos de recursos para selecionar o conjunto de dados mais adequado.

É imperativo dizer que a *Web Semântica* materializada, como se apresentou nesta pesquisa, é fundamental como auxílio ao usuário em tarefas complexas, facilitando a compreensão do sentido das informações pelos agentes computacionais. Portanto, plataformas baseadas na *Web Semântica* podem auxiliar significativamente na utilização da *Web* pelos usuários, pois, quando o significado dos dados está explícito para as máquinas, elas podem contribuir para as tarefas

cotidianas. A plataforma VIVO mostra-se um bom exemplo do uso da *Web Semântica*, como apresentado em aplicação prática, visto que realizar relações entre pesquisas e pesquisadores é sempre uma tarefa árdua. Assim, a VIVO consegue atender satisfatoriamente à navegação baseada na proposta da *Web Semântica*, o que aponta as possibilidades reais de uso dessa plataforma na questão da colaboração científica, podendo ser ampliado para outras áreas do conhecimento.

### Colaboradores

Todos os autores contribuíram para a concepção e desenho do estudo, análise de dados e redação final.

## Referências

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The semantic web. *Scientific American*, v. 284, n. 5, p. 28-37, 2001.

BÖRNER, K. *et al.* VIVO: A semantic approach to scholarly networking and discovery. *Synthesis Lectures on the Semantic Web: Theory and Technology*, v. 7, n. 1, p. 1-178, 2012.

CATARINO, M. E.; BAPTISTA, A. A. Folksonomia: um novo conceito para a organização dos recursos digitais na Web. *DataGramaZero*, v. 8, n. 3, 2007. Disponível em: <[http://www.dgz.org.br/jun07/Art\\_04.htm](http://www.dgz.org.br/jun07/Art_04.htm)>. Acesso em: 1 fev. 2016.

COSTA, S. M. S. Filosofia aberta, modelos de negócios e agências de fomento: elementos essenciais a uma discussão sobre o acesso aberto à informação científica. *Ciência da Informação*, v. 35, n. 2, p. 39-50, 2006. <http://dx.doi.org/10.1590/S0100-19652006000200005>

DURASPACE. *VIVO-ISF Ontology v1.6 Overviews: Classes*. Beaverton: DuraSpace, 2014. Available from: <<https://wiki.duraspace.org/display/VTDA/VIVO-ISF+Ontology+v1.6+Overview%3A+Classes>>. Cited: Jan. 23, 2017.

DURASPACE. *VIVO Technical Documentation*. Beaverton: DuraSpace, 2015. Available from: <<https://wiki.duraspace.org/display/VIVO/VIVO+Technical+Documentation>>. Cited: Jan. 18, 2017.

LEITE, F. C. L. *Como gerenciar e ampliar a visibilidade da informação científica brasileira*: repositórios institucionais de

acesso aberto. Brasília: IBICT, 2009. Disponível em: <<http://livroaberto.ibict.br/handle/1/775>>. Acesso em: 18 jan. 2017.

SANTAREM SEGUNDO, J. E. *Web Semântica, dados ligados e dados abertos: uma visão dos desafios do Brasil frente às iniciativas internacionais. Tendências da Pesquisa Brasileira em Ciência da Informação*, v. 8, n. 2, p. 219-239, 2015. Disponível em: <<http://inseer.ibict.br/ancib/index.php/tpbci/article/view/207>>. Acesso em: 18 jan. 2017.

SANTAREM SEGUNDO, J. E.; CONEGLIAN, C. S. Tecnologias da *Web Semântica* aplicadas a organização do conhecimento: padrão SKOS para construção e uso de vocabulários controlados descentralizados. In: GUIMARÃES, J. A. C.; DODEBEL, V. (Org.) *Organização do conhecimento e diversidade cultural*. Marília: ISKO-Brasil, 2015. (Estudos Avançados em Organização do Conhecimento, 3). Disponível em: <<http://isko-brasil.org.br/wp-content/uploads/2015/09/Organiza%C3%A7%C3%A3o-do-Conhecimento-e-Diversidade-Cultural-ISKO-BRASIL-2015.pdf>>. Acesso em: 18 jan. 2017.

TRISTÃO, A. M. D. *et al.* Sistema de classificação facetada e tesouros: instrumentos para organização do conhecimento. *Ciência da informação*, v. 33, n. 2, p. 161-171, 2004. <http://dx.doi.org/10.1590/S0100-19652004000200017>

WORLD WIDE WEB CONSORTIUM. *Web Semântica*. São Paulo: W3C, 2011. Disponível em: <<http://www.w3c.br/Padroes/webSemantica>> Acesso em: 18 jan. 2017.