XIII Encontro Nacional de Pesquisa em Ciência da Informação - XIII ENANCIB 2012

GT 8: Informação e Tecnologia

LINKED DATA COMO FORMA DE AGREGAR VALOR ÀS INFORMAÇÕES CLÍNICAS

Modalidade de apresentação: Comunicação oral

FERNANDO HADAD ZAIDAN – UFMG

MARCELLO PEIXOTO BAX – UFMG

fhzaidan@gmail.com

Resumo: *Linked Open Data* (LOD) - dados abertos vinculados - tem sido assunto constante nos principais congressos e *journals* de *web* semântica por todo o mundo. Diversos estudos comprovam que o consumo destes dados é potencialmente importante para melhorar a qualidade dos dados dos sistemas de informações nas mais diversas áreas. A informação clínica é uma destas áreas, e a iniciativa LOD tem se esforçado para padronizar a sua publicação, tornando a interligação do conjunto de dados (*datasets*) mais eficiente. Contudo, agregar valor aos dados internos dos sistemas de informação, em especial os clínicos (sistemas de informação clínicos - CIS), é desafiador. Com base nos autores que propuseram a *web* semântica e os que acompanham a evolução do LOD, foi feita a revisão de literatura dos principais conceitos. Uma pesquisa documental foi realizada obtendo uma bibliografia consistente. A partir dos trabalhos correlatos propostos e do estado da arte, apresentou-se de que forma pode existir a agregação de valor aos dados dos CIS.

Palavras-Chave: Dados Abertos Vinculados; *Web* Semântica; Informação Clínica; Sistemas de Informação.

Abstract: Linked Open Data (LOD) has been a constant subject in the main Semantic Web conferences and journals around the world. Several studies confirm that the consumption of this data is potentially important for improving the data quality of information systems in several areas. The clinical information is just one of them, and the LOD initiative has endeavored to standardize the publication, making a more efficient interconnection of the datasets. However, adding value to the internal data of information systems, particularly clinicians (clinical information systems - CIS), is challenging. Based on the authors who proposed the semantic web and the once that accompany the progress of the LOD, was performed a literature review of key concepts. From the related work proposed and the state of the art, it was presented how can there be to add value to the data of the CIS.

Keywords: Linked Open Data; Semantic Web; Clinical Informatics; System Information.

1. INTRODUÇÃO

O objetivo do presente artigo é apresentar o *Linked Open Data*, sua evolução e nuvem de dados abertos vinculados. A fim de constatar se é possível agregar valor aos dados e informações clínicos, será feito uma breve análise sobre a publicação e consumo destes dados abertos, e de que forma eles poderão ser interligados aos dados fechados dos sistemas de informação clínicos.

Os sistemas de informação organizacionais têm evoluído e tendem a sofisticação, com um grau de inteligência elevado. Mas, isto não é tudo o que se necessita para agregar valor aos dados. A obtenção de dados ainda é realizada de maneira precária, desconectada e às vezes manual, comprometendo resultados e a integração. Os bancos de dados relacionais, muito eficientes em diversos contextos, não conseguem fornecer a capacidade, por si só, de uma operabilidade integrada em uma *web* distribuída.

A publicação e o consumo de dados na *web*, entre os sistemas interconectados, com flexibilidade, é bastante facilitado com as tecnologias semânticas. Em uma *web* mais sofisticada, com consistência e integridade dos dados, necessita-se do apoio, no nível dos dados. Logo, não teremos uma página apontando para outra, mas um dado apontando para outro dado, usando referências globais (URIs - *Uniform Resource Identifiers*[1]).

Valores serão agregados aos sistemas de informação, na medida em que, em um novo cenário de interoperabilidade e de dados abertos vinculados (*Linked Open Data*), torna-se possível um modelo de dados onde a informação sobre uma única entidade esteja distribuída na *web*, sendo acessada por inúmeras organizações (HEATH; BIZER, 2011; LINKED DATA, 2012). Este modelo de dados não estará nas aplicações, mas fará parte da infraestrutura da *web* (ALLEMANG; HENDLER, 2011).

Diferentemente do início da *web* semântica, onde eram escassos os dados estruturados disponíveis, hoje são encontrados inúmeros domínios com uma grande quantidade de dados. Dentre eles, os dados da Wikipédia, cujo banco de dados no *Linked Data* é o DBPedia, assim como os dados de governos, saúde, educacionais, acadêmicos, entretenimento, etc. Não são poucas as iniciativas para um efetivo crescimento no formato RDF (*Resource Description Framework)* – modelo de dados da *web* semântica (LINKED DATA, 2012).

[1] O conceito de URIs (*Uniform Resource Identifiers*), assim como o de RDF (*Resource Description Framework*), serão fundamentados no item abaixo denominado: *Web* semântica e seus conceitos.

Diante do que foi exposto, a seguinte questão de pesquisa emerge para ser investigada: de que forma a utilização dos dados abertos (*Linked Data*) pode agregar valor aos dados dos sistemas de informação clínicos?

Este artigo esta dividido em cinco partes, onde a primeira o tema foi contextualizado e apresentado os objetivos e a questão de investigação. Abaixo se encontra a metodologia. Na terceira seção apresentam-se três trabalhos correlatos. Na quarta parte do artigo os conceitos são elucidados no referencial teórico, sendo apresentadas diversas citações de autores seminais. No intuito de alcançar os objetivos do estudo, na quinta parte é esclarecida a forma de agregar valor aos dados internos dos sistemas de informação clínicos. O que se segue são as considerações finais e a lista das referências bibliográficas.

2. PROCEDIMENTOS METODOLÓGICOS

Esta pesquisa tem o carácter exploratório, pois visa a maior familiaridade do pesquisador com o problema apresentado e, consequentemente, torná-lo mais explícito para a evolução e continuidade dos estudos.

A técnica a ser utilizada será a pesquisa documental. Após a leitura de livros sobre o assunto, cujos autores são seminais deste tema, foram selecionados os últimos trabalhos nos principais congressos e *journals* da área. Desta forma, obteve uma referência bibliográfica consistente, discriminada ao final deste artigo.

Por fim, dentre as referências selecionadas, escolheu-se três delas que constituirão os trabalhos relacionados. As pesquisas realizadas nestes estudos englobam os dados abertos vinculados na informática em saúde e biomedicina.

3. TRABALHOS RELACIONADOS

3.1 Estudos de Matthias Samwald e outros (2011) na Universidade de Viena, Áustria - sobre o LODD

Em um trabalho preliminar (*Preliminary Communication*), publicado em 2011, os autores reafirmam a enorme quantidade de informações sobre drogas disponíveis na web. O esforço do *Linked Open Data* na área médica (Saúde e Ciências da Vida) tem disponibilizado dados nesta área específica. Existem recomendações para a publicação destes dados, tornando-os conectados, como será visto em tópico específico deste artigo.

Os autores apresentam uma evolução histórica do LODD[2] (*Linked Open Drug Data*), e examinam a crescente importância dos dados abertos para o compartilhamento de dados na área farmacêutica. Discutem o TripleMap[3], um aplicativo voltado pra web, cujo foco é fornecer interface capaz de lidar com dados RDF. Estes estudos contribuem na medida em que apresentam esforços que podem agregar valor às informações clínicas.

3.2 Proposta de um *Workbench* para o *Linked Data* - Haase, Schimidt e Schwarte (2011) e Fluid Operations (2012)

A iniciativa do *Linked Open Data* é louvável, contudo existem dificuldades e barreiras para adentrar neste mundo. Haase, Schimidt e Schwarte (2011), bem como Fluid Operations (2012) propõem um conjunto integrado de ferramentas (*workbench*) para facilitar a utilização dos dados abertos vinculados.

Fontes de dados locais podem ser integradas através de camadas específicas, assim como utilizar uma interface onde os usuários podem se servir (*self-service)* de todos os recursos. Também é disponibilizado um grande conjunto do *widgets*[4] para a completa interação com dados RDF. Todos os detalhes ficam transparentes ou ocultos para os usuários. Dados locais e fontes virtualmente integradas podem ser consultados de forma integrada.

3.3 Processamento de consulta em um framework baseado em mediador para integração de dados no padrão de Linked Data - João Carlos Pinheiro (2011) - Tese de Mestrado e Doutorado em Ciência da Computação

O domínio específico deste estudo é a área médica, utilizando os *datasets* do LOD: Diseasome, DrugBank, DailyMed, Sider e DBPedia. A área de pesquisa é a integração de dados no padrão do *Linked Data*. Apesenta-se um cenário de necessidade de integração a partir de múltiplas fontes de dados públicas (PINHEIRO, 2011).

É apresentado um *framework*[5] baseado em mediador para a integração de dados no padrão *Linked Data*, acessíveis via SPARQL[6] *endpoint*, que é um serviço para implementação

[2] http://www.w3.org/wiki/HCLSIG/LODD
[3] http://triplemap.com/
[4] Componente de GUI (interface gráfica do usuário) que pode ser: menus, janelas, botões, ícones, etc. Podem ser também pequenos aplicativos da área de trabalho em um computador.
[5] *Framework* é uma estrutura (ou arcabouço) conceitual no intuito de facilitar a agregação.
[6] SPARQL (*Simple Protocol and RDF Query Language*) é a linguagem de consulta aos dados em RDF.

de consultas no modelo de dados RDF. O esquema de mediação é representado por uma ontologia de domínio, que fornece um vocabulário compartilhado (PINHEIRO, 2011).

O objetivo da tese é propor um método para o processamento distribuído de consultas SPARQL. O autor consegue alcançar o objetivo na medida em que fornece uma "solução não intrusiva e de fácil utilização para processamento de consultas em um ambiente mediado no contexto de dados publicados no padrão de Linked Data" (PINHEIRO, 2011, p. 143).

4. FUNDAMENTAÇÃO DOS CONCEITOS

4.1 Informação Clínica e os Sistemas de Informação Clínicos

A Informação Clínica é aquela originada dos procedimentos relacionados ao tratamento da saúde de um indivíduo. São resultantes dos exames de laboratórios, procedimentos, entrevistas, internação hospitalar, pronto atendimento, etc.

Alguns autores inserem os Sistemas de Informação Clínicos (CIS) como um subsistema dos Sistemas de Informação de Saúde de uma Comunidade (CHIS) - *Community Health Information Systems*, os quais realizam a gestão direta dos pacientes (VELDE; DEGOULET, 2003). Contudo, sabe-se que se trata apenas de convenções. De fato, a necessidade de sistemas de informação clínica tornou-se óbvia.

A evolução da Informação Clínica se deu a partir dos anos 1960, quando os sistemas ainda eram denominados Sistemas de Informação de Hospital (HIS), abarcando informações médicas e administrativas. Os CHIS vieram em seguida e, a partir dos anos 1990, visavam à redução de custos, ajuda às instituições e comunidades médicas nas atividades diárias de tomada de decisão, bem como integração dos recursos e melhora da gestão do paciente. Os sistemas tiveram uma evolução natural, denominando, assim, CIS.

4.2 Os Sistemas de Informação (SI) e os Dados Internos (fechados)

Os dados e informações dos sistemas clínicos, que são processados dentro das instituições médicas, hospitais, clínicas, etc., sem interoperabilidade, são considerados internos ou fechados. Isto acontece com todos os dados e informações dos SI no contexto interno das organizações.

Sistemas de Informação consistem num conjunto de partes que estão constantemente interagindo e se integrando, sempre com o propósito de atingirem objetivos e alcançar

resultados. Nenhum sistema sozinho pode fornecer todas as informações que uma empresa necessita. Os sistemas formam um todo unificado (LAUDON; LAUDON, 2011).

Para dar o aporte necessário aos dados dos SI necessita-se de um banco de dados (BD). O BD é um conjunto de dados inter-relacionados. O problema de interligar BD de sistemas de informações distintos, de tecnologias heterogêneas, sempre foi um desafio e motivo de pesquisas.

O conceito de interoperabilidade emerge da necessidade dos dados, primeiramente produzidos de forma independente e acarretando em heterogeneidade, terem uma estrutura uniforme e integrada, permitindo o compartilhamento.

4.3 Interoperabilidade

O W3C (2012) esclarece que a heterogeneidade semântica é um empecilho para alcançar a interoperabilidade, impedindo que os sistemas se comuniquem ou troquem informações. Para que dois ou mais sistemas troquem informações e utilizem as informações trocadas, precisa ocorrer a interoperabilidade.

Dentre diversos conceitos, o W3C (2012) define a interoperabilidade de maneira ampla, como a capacidade de dois ou mais sistemas de interagir e trocar dados e informações, de acordo com métodos definidos, objetivando resultados esperados. Este consórcio vem se dedicando no desenvolvimento de padrões para avançar rumo a excelência em termos de interoperabilidade.

A partir do surgimento da web semântica a interoperabilidade na web está em processo de melhoria, pelo fato da possibilidade de expressividade para dados.

4.4 Web semântica e seus conceitos

É notório que a *web* semântica é o resultado da aplicação de tecnologias de representação de conhecimento[7] em sistemas distribuídos em geral, com a finalidade de preencher o hiato de comunicação existente entre o ser humano e a máquina.

No clássico artigo em 2001, "*The semantic web*", a *web* semântica é descrita como extensão da *web* atual[8], com o objetivo de desenvolver meios para que as máquinas possam

[7] Área da inteligência artificial, cuja investigação se destina a representar o conhecimento em símbolos para facilitar a inferência e a partir destes elementos criarem novos elementos do conhecimento. A lógica é a base para a maioria dos formalismos necessários (HALPIN; LAVERENKO, 2009).
[8] A *World Wide Web* (www) foi criada por Berners-Lee em 1989.

servir aos humanos de maneira mais eficiente. Entretanto, é necessária a construção de instrumentos, no intuito de fornecer sentido lógico e semântico aos computadores (BERNERS-LEE, 2001).

Neste contexto, as ontologias cumprem um importante papel, pois é a conceituação formal de um domínio, com compromisso no compartilhamento semântico. São instâncias (também denominados nós) representadas por relações que fazem sentido, formando mecanismos de controles terminológicos. Havendo uma ontologia existe um consenso.

A proposição da arquitetura da *web* semântica foi baseada na linguagem de marcação estendida XML (*eXtensible Markup Language*). Esta linguagem surgiu em 1998 com o objetivo inicial de estruturar dados de forma aberta, criando uma infraestrutura única para diversas linguagens. Promove, também, um padrão de integração para a troca de documentos eletrônicos (W3C, 2012).

Tim Berners-Lee partiu do princípio que todo recurso *web*[9] necessita de uma URI (*Uniform Resource Identifiers*) única. URIs são a base da *web* semântica, assim como são fundamentais para toda a *web*, pois foram criadas para serem os nomes na *web* e todo recurso da *web* possui uma URI única e pode ser definido por ela[10] (BERNERS-LEE *et al.*, 2006).

Berners-Lee apresentou uma linguagem declarativa que se tornou um padrão, chamada de RDF (*Resource Description Framework)*. A escrita é em XML e foi recomendada pelo W3C em 2004. O modelo de dados RDF fornece uma semântica simplificada com boa representação para o tratamento de metadados[11]. O RDFs é o esquema (*Squema*) para declaração de classes e tipos em RDF.

Entretanto, pelo fato da RDF não fornecer subsídios necessários para a expressividade exigida de uma ontologia para a *web* semântica, criou-se a OWL (*Web Ontology Language)*. OWL permite a descrição dos aspectos semânticos dos termos utilizados e seus respectivos relacionamentos, favorecendo uma representação mais abrangente da RDF, contemplando a interoperabilidade.

O SPARQL (*Simple Protocol and RDF Query Language*) foi recomendado em 2008 pelo W3C. É uma linguagem de consulta RDF para expressar *queries* em diversas fontes de dados armazenados nativamente em RDF e também é um protocolo. O SPARQL *endpoint* é um serviço para implementação do SPARQL. Com isto, serão permitidas consultas a uma

[9] Os recursos são os conteúdos publicados na *web*.
[10] URI é uma *string* (cadeia) de caracteres. Alguns exemplos: para acesso a página *web* do Google (http://www.google.com) e URI do recurso Torre Eiffel, na enciclopédia colaborativa Wikipédia (http://en.wikipedia.org/ wiki/Eiffel_Tower).
[11] Os metadados são destinados à definição dos dados.

base de dados RDF utilizando a linguagem SPARQL. Máquinas e seres humanos utilizam estes serviços. Já o SPARQL 1.1 permite fazer consultas federadas[12] que buscam informações em múltiplas fontes de dados.

4.5 *Linked Data* e a sua fundamentação

O *Linked Data* descreve um conjunto de práticas para publicar e conectar dados estruturados na *web*. Os conjuntos de dados existentes da *web* têm representação através das triplas RDF, utilizando os *links* para os conjuntos de dados (*datasets*) participantes. Este projeto tem os mesmos princípios básicos da *web*, proposto por Tim Berners Lee: ser simples, possuir um *design* modular e contemplar a descentralização (BERNERS-LEE, 2001).

O W3C, desde o surgimento do *Linked Data* em 2007, faz o suporte devido aos dados abertos, e com isto impulsiona a produção de dados na web. Contudo, já são numerosos os *datasets* do *Linked Data* e o vocabulário heterogêneo dos dados e sua fragmentação natural no ambiente *web*, faz com que o consumo e reutilização tornem-se difícil. Mecanismos cada vez mais eficientes são criados a fim de permitir a utilização por qualquer interessado (BIZER; HEATH; BERNERS-LEE, 2009).

Berners-Lee *et al.* (2006) citam que os *Linked Data* são:

- Abertos e não proprietários: podem ser acessados por meio de uma ilimitada variedade de aplicações;
- Modulares: não necessita de planejamento prévio para combinar com outros dados*;*
- Escaláveis: uma vez que já exista dados no *Linked Data*, a adição de mais dados é feita de maneira fácil.

Os autores complementam que a *web* de documentos foi a inspiração para a web de dados, que envolve a padronização da semântica por trás dos dados. As ontologias, por sua vez, por se tratarem de consenso, fornecem a alternativa de integração dos dados de determinado domínio. Para que as aplicações acessem o *Linked Data*, precisam fazer através de consultas em um SPARQL *endpoint*.

4.6 *Dados abertos* e o governo brasileiro

[12] É baseado no conceito de distribuição do processamento de consultas para múltiplas fontes de dados autônomas. No âmbito de banco de dados heterogêneos, este conceito tem sido estudado há algum tempo. "Consultas federadas sobre SPARQL *endpoints* permitem que os dados possam ser integrados, sendo potencialmente benéficas em fontes de dados heterogêneos cujos componentes individuais possam usar *wrappers* SPARQL para publicar dados" (PINHEIRO, 2011).

Segundo o INDA (2012), governos de diversos países, cujas referências principais são o americano e o britânico, já estão em processo de definição de políticas e vêm desenvolvendo plataformas tecnológicas automatizadas através da internet, para promover a disseminação das informações públicas, de forma que a reutilização seja possível pela sociedade. “Na medida em que mais dados governamentais estejam disponíveis de forma aberta, espera-se que o próprio governo passe gradualmente a utilizar esses dados abertos como plataforma ágil de integração entre sistemas de informação” (INDA, 2012, p. 4). Esta citação foi retirada do “Guia de Abertura de Dados”, elaborado com o apoio da Infraestrutura Nacional de Dados Abertos (INDA). Os assuntos contemplados neste guia vão desde os objetivos para a abertura de dados, conceitos, melhores práticas, assim como os aspectos gerenciais e técnicos.

Outro documento importante é o “Manual dos dados abertos: governo” (W3C MANUAL DOS DADOS ABERTOS, 2011), elaborado para governos que desejam abrir dados, mas pode ser utilizado por quaisquer pessoas que queiram saber mais sobre os aspectos técnicos, sociais e políticos dos dados abertos. O W3C, juntamente com o INDA, vêm auxiliando o governo brasileiro neste esforço de transparência e abertura dos dados públicos, no intuito de aparecerem os primeiros casos de uso com o *Linked Data*.

4.7 Diagrama de nuvem do *Linked Open Data* (LOD)

O diagrama de nuvem do LOD mostra os conjuntos de *datasets* que foram publicados em determinadas datas. A FIG. 1 representa como foi a primeira publicação em Maio/2007 com apenas 12 *datasets.* Poucos, quando comparados com a FIG. 2, que é a representação de Setembro/2011 – 295 *datasets.* Uma evolução significativa ocorreu neste período.

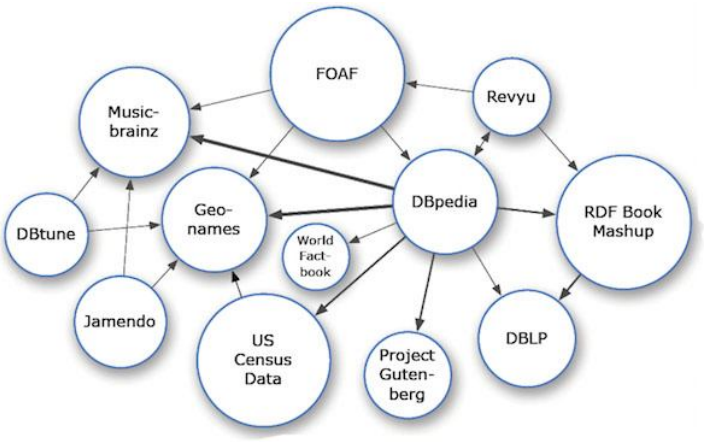

Figura 1*:* LOD *Cloud Diagram* – Maio/2007.
Fonte: Linked Data, 2012.

Cabe apresentar um grupo de datastes denominado LODD (*Linked Open Drug Data*), que os autores Samwald *et al*. (2011), correlatos a este estudo, indicam como específicos de "*Health Care"* e "*Life Science"*. Este esforço trabalha com um conjunto de tecnologias e convenções dentro do *Linked Data*, facilitando a conexão dos dados neste domínio específico. Alguns dos *datasets* do projeto LODD podem ser vistos na nuvem abaixo (FIG. 2), em uma localização específica. Estão todos abaixo do *dataset* DBPEDIA, que está ao centro.

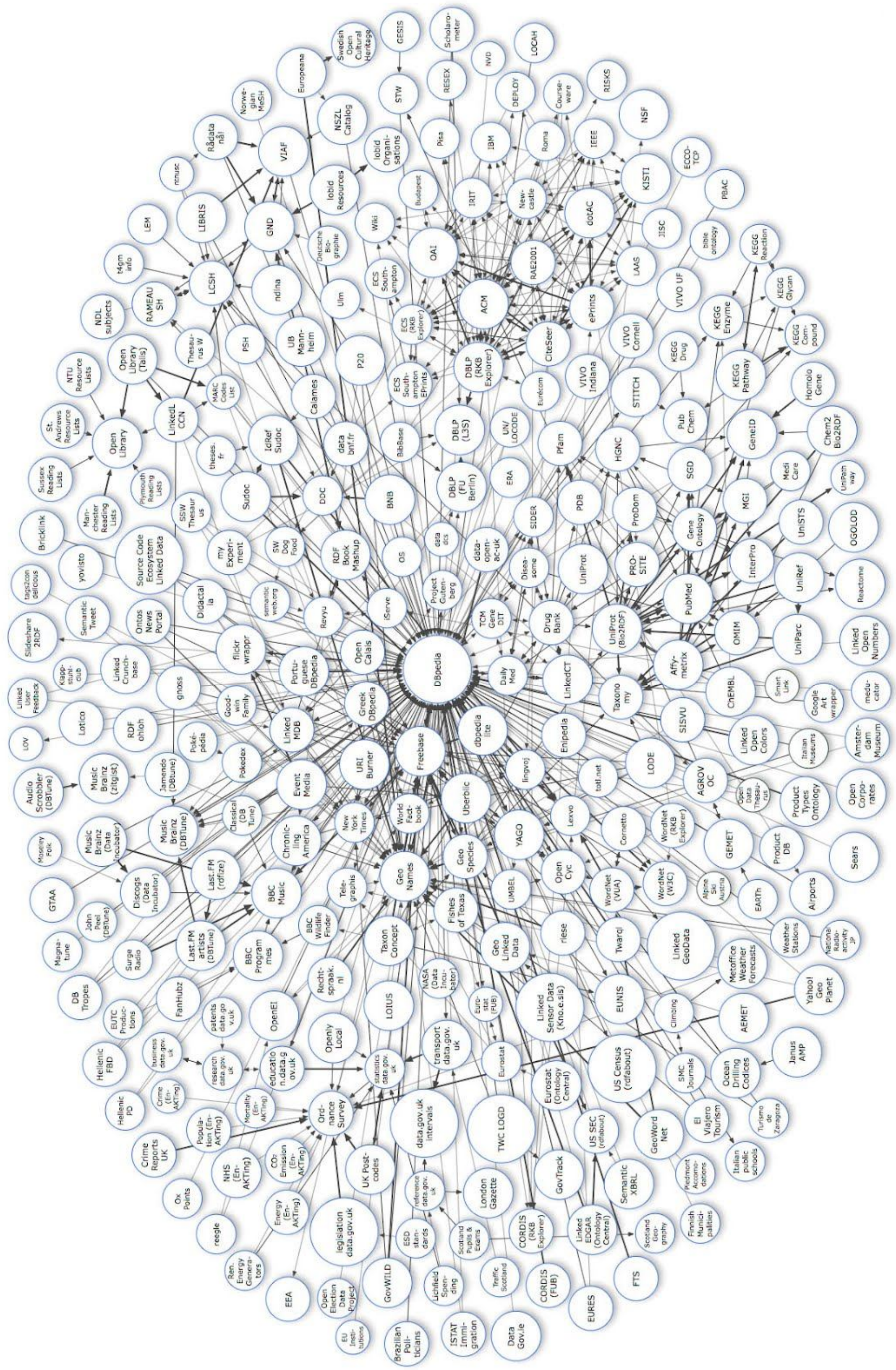

Figura 2*:* LOD *Cloud Diagram* – Setembro/2011.
Fonte: BIZER; JENTZSCH; CYGANIAK (2011).

O diagrama de nuvem é mantido por Richard Cyganiak, do DERI[13] - *Digital Enterprise Research Institute*, da Universidade Nacional da Irlanda, Galway, e Jentzsch Anja, da Universidade Freie[14], Berlin.

A imagem da nuvem mostra os conjuntos de dados que são publicados e como estão interligados com outros conjuntos de dados. O tamanho dos círculos corresponde ao número de triplas RDF que possuem. As setas bidirecionais indicam que os *links* estão em ambos os conjuntos de dados. Contudo, a seta em uma única direção (por exemplo, de A para B), indica que o *dataset* A contém triplas RDF que usam identificadores em B (LINKED DATA, 2012).

De Maio/2007 a Setembro/2011 houve onze novas "nuvens" publicadas. Abaixo o quadro com a lista das datas de publicação e quantidade de *datasets*.

Data Publicação	Quantidade *Datasets*
01/05/2007	12
08/10/2007	25
07/11/2007	28
28/02/2008	32
31/03/2008	34
18/09/2008	45
05/03/2009	89
27/03/2009	93
14/07/2009	95
22/09/2010	203
19/09/2011	295

Quadro 1: Publicação dos *datasets*.
Fonte: LINKED DATA, 2012. Adaptado pelo autor.

Chama-se a atenção para o crescimento acentuado de Julho/2009 para Setembro/2010.

4.8 Como publicar e consumir os *Linked Data*

Heath e Bizer (2011) esclarecem que basicamente são dois tipos de aplicações que realizam o consumo do *Linked Data*:

- Aplicações genéricas: fazem o uso dos dados do *Linked Data* em qualquer domínio;
- Aplicações de domínio específico: são as que focam em um domínio específico.

[13] http://www.deri.ie/
[14] http://www.fu-berlin.de/en

Existem algumas maneiras para armazenar dados no padrão necessário do *Linked Data*, a saber:

- Utilizando APIs (*Application Programming Interface*) para as triplas RDFs nativas. Exemplos: Sesame[15] e Jena[16];
- Fornecendo *wrappers* (tradutores) para banco de dados relacionais. Neste caso, um *paper* de Bizer e Cyganiak (2006) faz a proposição desta funcionalidade. O Virtuoso[17] é um exemplo.
- Fazendo a triplificação (disponibilização de dados em triplas RDFs) de fontes de dados relacionais. Os exemplos são o Jena SDB[18] e Jena TDB[19].

Como comentado anteriormente, o consumo (utilização) dos dados do LOD necessita, incialmente, de conceitos como URIs e SPQRQL. O acesso pode ser feito por meio do conhecimento prévio do vocabulário das fontes de dados e da sintaxe das consultas SPARQL submetidas nos SPARQL *endpoints*. Esta abordagem é denominada tradicional, como explicam Hartig e Langegger (2010), e os dados são materializados em um *data warehouse*[20]. Para *datasets* muito grandes, a materialização tende a ser demorada e utilizar grande espaço de armazenamento. A vantagem fica por conta da velocidade das respostas, pois não há necessidade de comunicação em rede.

A outra abordagem, denominada federação de consultas, prevê "um mediador transparente que decompõe a consultas em subconsultas e encaminha as subconsultas a múltiplos serviços de consulta distribuídos" (PINHEIRO, 2011, p. 17-18). O acesso aos dados é realizado por meio de um vocabulário padrão especificado na ontologia de domínio[21]. Depois de obtido os resultados, os dados são integrados e a resposta final é entregue ao usuário. Existem algumas vantagens nesta abordagem, por exemplo, não necessita tempo ou espaço adicional para materialização de dados e obtêm os dados totalmente atualizados, pois a extração é no momento que a consulta é requisitada (PINHEIRO, 2011). Trabalhos vêm sendo desenvolvidos no âmbito de otimizar estas consultas.

---

[15] http://www.openrdf.org/
[16] http://jena.sourceforge.net/
[17] http://virtuoso.openlinksw.com/
[18] http://openjena.org/SDB/
[19] http://openjena.org/TDB/
[20] Conceito que significa um local de grande capacidade, onde podem ser armazenados dados para consultas posteriores.
[21] Contém termos essenciais de um domínio específico do conhecimento, diferentemente das ontologias fundacionais (GUARINO, 1995).

Para que os dados sejam publicados no *Linked Data*, um conjunto de princípios deverá ser seguido[22]. Alguns deles são:

- Dados no formato RDF;
- Conter pelo menos 1.000 triplas;
- Deve ser conectado por *links* RDF para um conjunto de dados que já está no diagrama (pelo menos 50 *links*);
- Ser acessado por SPARQL *endpoint*, dentre outros.

É importante contextualizar o conceito de objetos desreferenciados, ou seja, que não se conhece, a princípio, a URI. Neste caso, quando se obtêm informações, o resultado é uma descrição RDF do recurso identificado. Uma série de benefícios é encontrada, como a criação de *links* RDF (ou URI-*links*) entre dados de diferentes fontes de dados (ALLEMANG; HENDLER, 2011).

Vale uma ressalva sobre o termo dados abertos. Segundo o *Linked Data* (2012), nem todos os publicadores dos dados fizeram uma licença explícita dos seus dados. É sempre conveniente para quem for consumir os dados consultar se o site de quem publicou tem a licença com os termos e condições de uso.

4.9 A interligação dos dados internos com *Linked Data*

Para alcançar o objetivo deste estudo, necessita-se apresentar algumas soluções que possibilitam a interligação dos dados internos com o *Linked Data*.

Haase, Schimidt e Schwarte (2011) propõem um *workbench* no intuito de diminuir a barreira de entrada para o mundo do *Linked Data*. O apoio à descoberta e exploração de fontes de dados facilita a integração e processamento de dados abertos vinculados. Fontes de dados remotas podem ser virtualmente integradas através de uma camada de federação, e o desenvolvimento de uma interfase de usuário *self-service* também torna mais fácil. O uso é baseado em um *wiki* semântico[23], combinados com um grande conjunto de *widgets* para interação com os dados. Finalmente, vale ressaltar que esta abordagem de integração fica transparente para o usuário. Não existe a preocupação com aspectos da distribuição física dos dados ou protocolos de acesso, pois detalhes da integração ficam ocultos em tempo de

[22] Os princípios constam no link <http://www4.wiwiss.fu-berlin.de/bizer/pub/linkeddatatutorial/>. Acesso em: 12 dez. 2011.
[23] Conceito que implementa tecnologias da *web* semântica nas ferramentas wikis.

execução. Dados locais e fontes virtualmente integradas podem ser consultados de forma integrada.

A empresa Fluid Operations[24] oferece o "*Information Workbench – for a world where all data is Linked*": uma plataforma *web* aberta, para soluções de *Linked Data* para organizações. Dados de diferentes fontes podem ser integrados e conectados, utilizando uma camada de dados do *Linked Data* em cima do conteúdo, facilitando o acesso semântico e a busca inteligente. A Fluid Operations tem seus princípios baseados em Haase, Schimidt e Schwarte (2011).

A ligação de recursos internos e externos (dados internos e externos) são feitas por meio da URI (HEATH; BIZER, 2011). Os autores exemplificam o uso de chaves primárias em banco de dados internos, como ISBN de livros, para ligar com dados do *Linked Data*.

Em uma importante abordagem de ligação de dados internos, Passant *et al*. (2010) propõem interfaces de navegação e *mashups*[25] semânticos. A novidade desta aplicação está na reutilização de dados RDF do GeoNames[26] para fornecer um *mashup* semântico da combinação de fonte dados externos e internos. Os dados internos são combinados com dados do GeoNames pelas coordenadas de localização, permitindo uma navegação avançada dos recursos. A representação visual é outro ponto alto desta aplicação.

Os autores acreditam que *mashups* semânticos podem ser parte significativa do futuro das aplicações Enterprise 2.0[27]. O uso do *Linked Data* é fundamental para este sucesso, na medida em que as empresas se beneficiam de informações públicas a custo zero. Passant *et al*. (2010) complementam a importância desta ligação inclusive com dados legados das organizações.

Um interessante conceito emerge: *Linked Data Enterprise* (WOOD, 2010). As empresas de dados vinculados são organizações em que, o ato de lidar com a informação (criar, armazenar, disseminar), está intimamente associado com o ato de compartilhar, ou seja, o compartilhamento de dados é tão importante como produzi-lo. Neste tipo de empresas, indivíduos ou grupos continuam a produzir e consumir informação de modo específica para suas necessidades, contudo com vistas no compartilhamento, reduzindo as barreiras de troca de informação.

---

[24] http://www.fluidops.com.
[25] Sites ou aplicações webs que usam informações de mais de uma fonte no intuito de criar novos serviços.
[26] *Dataset* do LOD com informações sobre mais de 6 milhões de lugares e características geográficas.
[27] Este termo indica uma linha de evolução das organizações voltadas para o conhecimento, que utilizam ferramentas da *web* 2.0 (cooperação e colaboração) e da *web* 3.0 (semântica).

A princípio, pode parecer que trabalhadores do conhecimento não dispõem de tempo ou incentivo para fazer parte deste esforço. Mas, quando é percebido que seus dados podem agregar valor a outros dados, o caminho inverso é verdadeiro e recíproco, pois um minuto de esforço na partilha de informação, resulta em várias horas economizadas. O contexto do *Linked Open Data* proporciona uma ampla gama de situações adequadas para operacionalizar esta agregação de valor aos dados (WOOD, 2010).

5. AGREGANDO VALOR AOS DADOS INTERNOS DOS SISTEMAS DE INFORMAÇÃO CLÍNICOS

Existe uma grande diversidade no conceito de agregar valor, inclusive colocando o cliente ao centro, quando estabelece a sua satisfação com um produto, um serviço ou um sistema de informação. Com o foco no produto significa incorporar tecnologia ao mesmo, tornando-o cada vez mais sofisticado. Contudo, agregar valor aos dados dos sistemas de informações significa a possibilidade de não tê-los em ambientes isolados, mas interoperáveis, transformando-os em informações passíveis de análises estratégicas para uma tomada de decisão precisa (LAUDON; LAUDON, 2011).

Velde e Degoulet (2003) explicam que, em geral, focaliza-se na quantidade de dados no contexto da informação clínica, muitas vezes não proporcionando nenhum tipo de análise, e sem capacidade de rastreabilidade ou historicidade dos dados. O mais importante, de acordo com estes autores, é que seja viável a análise pelos sistemas de informação.

É impossível imaginar a possibilidade de agregar valor aos dados de informação clínica sem especificar e consumir alguns *datasets* do LOD voltados para esta área. O número de *datasets* da "*Life Science*", segundo Bizer, Jentzsch e Cyganiak (2011) é de 41 (no total de 295)[28], com mais de três bilhões de triplas. Acima deste número, encontram-se apenas 87 *datasets* de publicações e 49 *datasets* de dados governamentais. A lista abaixo relaciona alguns *datasets* da "Ciência da Vida":

- DailyMed: publicado pela Biblioteca Nacional de Medicina, fornece informações de qualidade sobre drogas comercializadas;
- Diseasome: rede pública de mais de 4.300 doenças e Genes ligados a distúrbios;
- DrugBank: repositório de quase 5.000 moléculas e informações detalhadas sobre drogas;

[28] Publicação da nuvem de *datasets* em 19 set. 2011.

- Gene Ontology: ou GO, é uma iniciativa importante de bioinformática para unificar a representação dos atributos dos Genes e dos atributos do produto dos Genes de todas as espécies;
- InterPro: Banco de dados de famílias de proteínas, com a iniciativa de possuir as mais novas proteínas;
- SIDER: contém informações sobre drogas comercializadas e seus efeitos colaterais. As informações são extraídas de documentos públicos e de bulas.
- STITCH: contém informações sobre produtos químicos e proteínas, bem como suas interações e *links*;
- TaxonConcept: as espécies são conhecidas por muitos nomes diferentes. Esta base de conhecimento tem URIs para conceitos das espécies.
- Dentre outras (LINKED DATA, 2012).

Pinheiro (2011) apresenta na Figura 3 um esquema mínimo de fontes de dados da área médica com respectivas ligações para a integração no padrão *Linked Data* de um domínio específico (domínio médico), que incluem informações sobre doenças (*Diseasome*), drogas (*DrugBank*), bulas de drogas (*DailyMed*), medicamentos e efeito diversos (*Sider*), assim como do DBPedia, que interliga praticamente todos os domínios do LOD, como uma fonte de dados de temas variados.

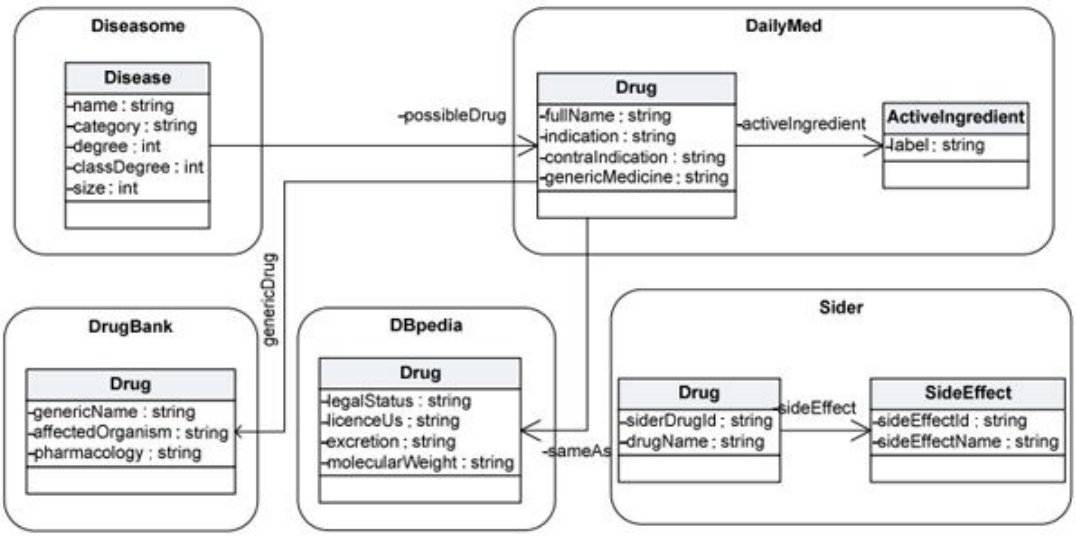

Figura 3*:* Interligação de fonte de dados do domínio médico no padrão de *Linked Data*.
Fonte: PINHEIRO (2011, p. 20).

Nos tópicos abaixo, serão apresentados algumas tecnologias utilizadas com o *Linked Data*, com vistas a agregar valor aos dados internos.

5.1 Os extratores do *Linked Data*

A *web* dos dados nos dá a ideia concreta de que mais e mais dados estão interligados. Isto é uma realidade que confirmam Rizzo e Troncy (2011), quando complementam que, para ir em direção a este mundo, com um foco estruturado dos dados, existe a necessidade de proporcionar anotações mais estruturadas aos documentos, utilizando vocabulários comuns ou ontologias. Textos semiestruturados, como os científicos, médicos ou artigos de notícias, tem uma maior possibilidade de serem semanticamente anotados. Entidades extratoras desempenham um papel fundamental para a extração de informações estruturadas, identificando características (entidades) e ligando-as a outros recursos da *web* por meio de inferências.

Os autores desenvolvem um trabalho que agregam valor aos dados, onde avaliam os extratores mais populares do *Linked Data*, como DBPedia Spotlight, Extractiv, OpenCalais, Alche-MyAPI e Zemanta. O resultado da pesquisa é um *framework* com avaliação realizada pelo homem, atribuindo um valor a detecção da entidade, tipo de entidade e desambiguação de entidade.

Heath e Bizer (2011), também falam sobre os extratores para o *Linked Data*. Reforçam que onde a entrada de dados é textual e há recursos naturais de linguagem, por exemplo, uma série de notícias ou relatórios de negócios, é possível passar estes documentos através de extratores de entidades para *Linked Data*. Publicar estas anotações, junto dos documentos, melhora a tarefa de recuperação da informação e permite aplicativos usarem as fontes do *Linked Data* referenciadas, como um pano de fundo para mostrar as informações sobre as páginas, além de oferecer a navegação facetada.

5.2 LDIF - *Linked Data Integration Framework*

Schultz *et al.* (2011), na Universidade Freie - Berlin - desenvolveram um *framework* para a construção de aplicações de dados no *Linked Data*. O LDIF traduz dados vinculados heterogêneos da *web*, em uma representação mais limpa, para o uso local, mantendo a procedência dos dados. Fornece uma linguagem de mapeamento expressiva para traduzir os dados de vocabulários diferentes, com vistas ao uso local. Inclui também um componente de resolução de identidade, que descobre URIs baseado nos dados de entrada.

O estudo de caso destes autores foi, justamente, o domínio da Ciência da Vida do LOD, onde foi escolhido dois *datasets*:

- Kegg Gened: uma coleção de catálogos de Genes gerada a partir de recursos públicos disponíveis;
- UniProt: um conjunto de dados contendo sequencias de proteínas, genes e funções.

Os autores vão estender o LDIF, contemplando: a distribuição dos processos em *cluster* de máquinas, permitindo uma escala de grande quantidade de dados; *crawlers* para Linked Data e SPARQL *endpoint*; adicionando uma avaliação da qualidade dos dados.

5.3 Aplicativos *web*

Samwald *et al*. (2011) apresentam o TripleMap, um aplicativo para *web* que fornece uma interface rica, dinâmica e integrada, para conjunto de dados RDF. Pode ser feita a escolha de qual área se deseja trabalhar, por exemplo, o LODD, compondo doenças, drogas, ensaios de pesquisa, etc. As entidades podem ser arrastadas para dentro da tela, possibilitando a navegação e visualizando as associações.

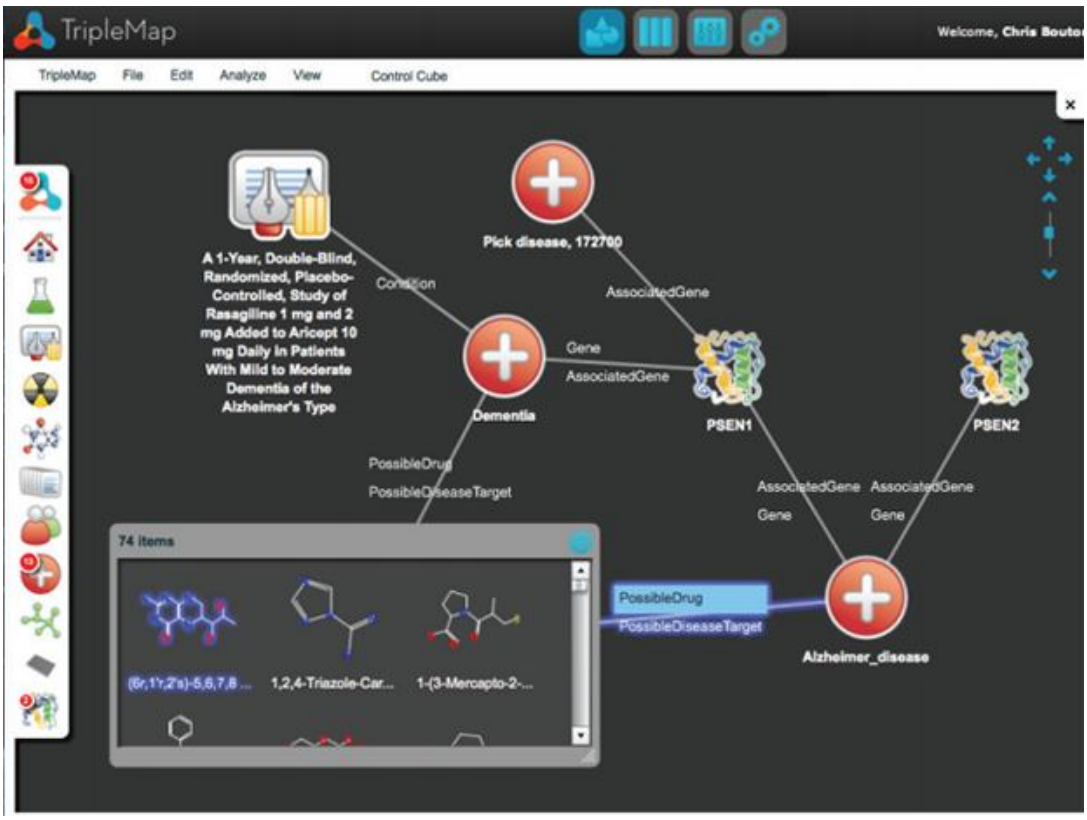

Figura 4*:* Tela do TripleMap com datasets do LODD.
Fonte: SAMWALD *et al.* (2011).

6. CONSIDERAÇÕES FINAIS

Neste artigo procurou-se apresentar a evolução do *Linked Open Data* (LOD) e de que forma existe a agregação de valor aos dados dos sistemas de informação clínicos (CIS).

Quanto aos trabalhos correlatos, todos utilizaram os dados abertos das áreas da Saúde e das Ciências da Vida. Nestes trabalhos foram propostos *frameworks* e interfaces para os usuários lidarem, de maneira fácil e transparente, com os dados do LOD.

Na revisão de literatura, para elucidar os principais conceitos, utilizaram-se autores que participaram da proposição da *web* semântica e do LOD, como Tim Berners-Lee, Christian Bizer, James Hendler, Richard Cyganiak, dentre outros. Além deles, os sites que encarregam da proposição, validação e acompanhamento do LOD, como o www.w3c.org e www.linkeddata.org. Introduziu-se ao estudo de publicação e consumo dos dados abertos.

Foi mostrado como é possível agregar valor aos dados internos dos sistemas clínicos, mesmo quando se tem disponível, para ser consumido, mais de três bilhões de triplas RDF, somente nos 41 *datasets* do LOD – Saúde e Ciências da Vida (dados da última nuvem do LOD, que foi publicada antes do fechamento deste artigo, em Setembro/2011).

Outras formas apresentadas neste estudo, com vistas à proposição de agregação de valor aos dados clínicos, foram os extratores de *Linked Data*, o LDIF (*Linked Data Integration Framework*) e o TripleMap, mais especificamente com os dados abertos vinculados.

Este artigo alcança seu objetivo, mas é apenas o início de um estudo em um vasto campo de pesquisa que é o *Linked Open Data*, utilizando os *datasets* da Saúde e da Ciência da Vida. Adentrar os estudos nas ontologias de domínio da área da informática em saúde, RDF, OWL, SPARQL, assim como nas interligações dos *datasets* do LOD, é um desafio e uma necessidade.

Estudos embrionários nestas áreas estão espalhados pelas principais universidades do mundo, principalmente o DERI e Universidade de Freie, citadas neste artigo, cujas referências e *links* apresentados contemplaram o estado da arte.

REFERÊNCIAS BIBLIOGRÁFICAS

ALLEMANG, D. HENDLER, J. **Semantic web for the working ontologist**: effective modeling in RDFS as OWL. 2. ed. USA – MA: Elsevier Inc., 2011.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The Semantic Web. **Scientific**

American, (5), 2001.

BERNERS-LEE, T.; *et al*. Tabulator: exploring and analyzing linked data on the semantic web. Proceedings of the **3rd International Semantic Web User Interaction**, 2006.

BIZER, C.; CYGANIAK, R. Publishing Relational Databases on the Web as SPARQL-Endpoints. **15th International World Wide Web Conference**, 2006.

BIZER, C.; HEATH, T.; BERNERS-LEE, T. Linked Data: The Story So Far. **International Journal on Semantic Web and Information Systems**, 5(3), 1-22, 2009.

BIZER, C.; JENTZSCH, A.; CYGANIAK, R. State of the LOD Cloud. 2011. Disponível em <http://www4.wiwiss.fu-berlin.de/lodcloud/state/>. Acesso em 01 jul. 2012.

CAO, L.; ZHANG, C.; LIU. J. Ontology-based integration of business intelligence. **Web Intelligence and Agent Systems**, IOS Press, 2006.

FLUID OPERATIONS. For a world where all data is Linked - Self-service Platform for Linked Data Solutions in the Enterprise. Disponível em: <http://www.fluidops.com/wp-content/uploads/downloads/2010/10/fluidOps_Information_Workbench_Brochure.pdf>. Acesso em: 01 jul. 2012.

GUARINO, N. Formal Ontology, Conceptual Analysis and Knowledge Representation. **International Journal of Human-Computer Studies**, v. 43, n. 5/6, 1995.

HAASE, P.; SCHMIDT, M.; SCHWARTE, A. The Information Workbench as a Self-Service Platform for Linked Data Applications. **Second International Workshop on Consuming Linked Data** (COLD2011). 2011.

HALPIN, H.; LAVERENKO, V. Relevance feedback between hypertext and semantic search. In: **Proceedings of Semantic Search Workshop at the World Wide Web Conference**, 2009.

HARTIG, O.; BIZER, C.; FREYTAG, J.-C. Executing SPARQL queries over the Web of Linked Data. Proceedings of the **8th International Semantic Web Conference**, Springer-Verlag, 2009.

HARTIG, O.; LANGEGGER. A Database Perspective on Consuming Linked Data on the Web. **Datenbank-Spektrum**, 14(2):1–10, 2010.

HEATH, T.; BIZER, C. **Linked Data**: Envolving the web into a global data space. USA – CA: Morgan & Claypool Publishers, 2011.

INDA - Guia de Abertura de Dados da Infraestrutura Nacional de Dados Abertos (INDA): documento versão final. 2012. 2. ed. Disponível em: <https://www.consultas.governoeletronico.gov.br/ConsultasPublicas/download.do?acao=arquivoDocumentoFinal&tipo=pdf&id=93>. Acesso em: 23 ago. 2012.

LAUDON, K. C.; LAUDON J. P. **Sistemas de informação gerenciais:** administrando a empresa digital. 7. ed. São Paulo: Prentice Hall, 2011.

LINKED DATA. Disponível em: <www.linkeddata.org>. Acesso em: 01 jul. 2012.

W3C MANUAL DOS DADOS ABERTOS: governo. 2011. Disponível em: < http://www.w3c.br/pub/Materiais/PublicacoesW3C/Manual_Dados_Abertos_WEB.pdf>. Acesso em: 23 ago. 2012.

NEIRA, R. *et al.* Como incorporar conhecimento aos sistemas de registro eletrônico em saúde. **X CBIS - Congresso Brasileiro de Informática na Saúde.** 2008.

PASSANT, A.; *et al.* **Enhancing Enterprise 2.0 Ecosystems Using Semantic Web and Linked Data Technologies**: The SemSLATES Approach. In: WOOD, D. (Org.) Linking Enterprise Data. New York: Springer, 2010.

PINHEIRO, J. C. **Processamento de consulta em um framework baseado em mediador para integração de dados no padrão de Linked Data**. 2011. Tese (Mestrado e Doutorado em Ciência da Computação), Universidade Federal do Ceará, 2011.

RIZZO, G.; TRONCY, R. NERD: A framework for evaluating named entity recognition tools in the Web of data. **10th International Semantic Web Conference - ISWC'11** - Workshop on Web Scale Knowledge Extraction. Bonn, Germany, Oct., 2011.

SAMWALD, M. *et al.* Linked open drug data for pharmaceutical research and development. **Journal of Cheminformatics**, 3:19, 2011.

SCHULTZ, A;. *et al.* **LDIF -Linked Data Integration Framework. Second International Workshop on Consuming Linked Data** (COLD2011), 2011.
TAGGART, C. The economics of open data & the big society. 2011. Disponível em: <http://countculture.wordpress.com/2011/10/13/the-economics-of-open-data-the-big-society/>. Acesso em: 10 jul. 2012.

VELDE, R. V. de; Degoulet, P. **Clinical Information Systems**: a Component-Based Approach. New York: Springer-Verlag New York, Inc., 2003.

W3C. Disponível em: <www.w3c.org>. Acesso em: 01 jul. 2012.

W3C MANUAL DOS DADOS ABERTOS: governo. 2011. Disponível em: < http://www.w3c.br/pub/Materiais/PublicacoesW3C/Manual_Dados_Abertos_WEB.pdf>. Acesso em: 23 ago. 2012.

WOOD, D. (Ed.) **Linking Enterprise Data**. New York: Springer, 2010.

WYLOT, M.; *et al.* dipLODocus[RDF] Short and Long-Tail RDF Analytics for Massive Webs of Data. **10th International Semantic Web Conference** (ISWC2011), 2011.