Repositórios de dados para *E-Science*: *Open Data*, *Linked Data* e suas tecnologias

Divino Ignácio Ribeiro Junior

Doutor em Engenharia e Gestão do Conhecimento pela Universidade Federal de Santa Catarina (UFSC) – Florianópolis, SC - Brasil.
Professor da Universidade do Estado de Santa Catarina (UDESC) - Florianópolis, SC – Brasil.
http://lattes.cnpq.br/1599118875018269
E-mail: divino@udesc.br

Recebido em: 15/08/2014. Aprovado em: 23/1/2015. Publicado em: 07/08/2015.

Resumo
Um novo paradigma de atividade científica tem se caracterizado a partir dos constantes investimentos em computação de larga escala e repositórios de dados: a e-Science. Os repositórios de dados nesse contexto fazem uso de tecnologias da Web semântica, e são conceitualmente diferentes dos produtos tradicionais destinados à disseminação dos outputs da atividade científica tradicional. Este trabalho apresenta brevemente os princípios tecnológicos que definem tais repositórios de dados e ao mesmo tempo propõe um debate no âmbito da ciência da informação, no sentido de promover uma reflexão acerca do potencial de pesquisa e da participação dos pesquisadores dessa área. O artigo é realizado a partir de um levantamento bibliográfico sobre os temas envolvidos e conclui com considerações e questões identificadas na literatura, relevantes para promover a continuidade da discussão.

Palavras-chave: Open Data. Linked Data. E-Science. Web Semântica. RDF.

Data Repositories for E-Science: Open Data, Linked Data and its technologies

Abstract
A new paradigm of scientific activity has been characterized from the constant investments in large-scale computing and large-scale data repositories: e-Science. Data repositories that context make use of Semantic Web technologies, and are conceptually different from traditional products intended for dissemination of outputs of traditional scientific activity. This work briefly presents the technological principles that define such data repositories while proposing a debate in the context of the Information Science to promote a reflection about the research potential and the participation of researchers in this area. The article is realized from a literature survey on the issues involved and presents the end considerations and issues identified in the relevant literature to promote the continuity of the discussion.

Keywords: *Open Data. Linked Data. E-Science. Semantic Web. RDF.*

Repositorios de datos para la e-Ciencia: Open Data, Linked Data y sus tecnologías

Resumen
Un nuevo paradigma de la actividad científica se ha caracterizado por las inversiones constantes en computación de larga escala y repositorios de datos: la e-Ciencia. Repositorios de datos en este contexto hacen uso de las tecnologías de la web semántica, y son conceptualmente diferentes de los productos tradicionales para la difusión de los resultados de la actividad científica tradicional. Este artículo presenta brevemente los principios tecnológicos que definen a dichos repositorios de datos, mientras que propone un debate en la ciencia de la información, para promover una reflexión sobre el potencial de la investigación y la participación de los investigadores en este campo. El artículo se realiza a partir de una revisión de la literatura sobre los temas en cuestión y concluye con consideraciones y cuestiones identificadas en la literatura relevante para promover la continuidad de la discusión.

Palabras clave*: Open Data. Linked Data. E-Ciencia. Web Semántica. RDF.*

INTRODUÇÃO

Um fenômeno recorrentemente citado em toda a literatura relativa às temáticas sobre recuperação da informação, Web semântica, bibliotecas e repositórios digitais é o do enorme volume de informação e documentos – estruturados e não estruturados – que aumenta constantemente e provoca novas demandas de organização e gestão de informação e de documentos.

A geração das 'massas de dados ocultas'[1] produzidas diariamente nas organizações, no meio científico, pelas mídias, tendo por meio a infraestrutura as redes de telecomunicações e a Internet continua avançando deixando uma lacuna temporal entre a capacidade de resposta dos pesquisadores, organizações e tecnologias e os impactos decorrentes desse enorme volume de dados.

Todos os setores da sociedade, do governo e do setor produtivo são afetados, direta ou indiretamente, por esse fenômeno e suas consequências. A partir da necessidade de melhor gerenciar informações em qualquer formato, registradas em documentos estruturados ou dispersas em dados ou documentos com pouca estrutura, são realizados muitos investimentos em tecnologias e processos de gestão, a fim de minorar os impactos negativos dessa 'sobrecarga de informação'.

Nessa linha de pensamento, Sayão et al. (2009) afirmam que o século XXI consolida a cultura da disseminação da informação eletrônica por meio da Internet, do acesso às fontes de informação e dos canais de comunicação, como listas de discussão, blogs, *chats*, mídias sociais, entre outros. A cultura da disseminação da informação eletrônica movimenta a indústria de equipamentos de informática, telecomunicação e outros, que utilizam tecnologias com fins computacionais e para comunicação em rede; impulsiona o mercado de serviços de telefonia móvel, acesso à Internet, serviços de telefonia VoIP (Voice over IP), sistemas de IPTV – transmissão de sinal de TV através de canais de dados, entre outros setores.

Há fortes investimentos na 'criação de demanda' para que a população continue aderindo a novos serviços; por exemplo, as empresas de telecomunicação não vendem 'celulares' ou 'planos de telefonia e dados' – todo esforço de *marketing* é focado para vender 'conectividade', formar e alimentar a necessidade de estar sempre 'conectado' e 'acessível'. A partir daí instala-se um processo retroalimentado: a população desenvolve essa necessidade de consumo de serviços de telecomunicação, impulsionada pelas empresas que os fornecem. Outros mercados, como o de aplicativos para telefonia móvel, oferecem serviços de *software* em vários segmentos, criando novas demandas de consumo.

Além do aspecto econômico, há as dimensões sociais e culturais. Os impactos da cultura digital permeiam a formação de novas gerações, influenciando a maneira como pensam, escrevem e se comunicam.

Sob a ótica da gestão é aceitável que se observe a informação como insumo, artefato consumível para atender necessidades de produção. Nesse sentido, a área da ciência da informação tem se ocupado, também, em compreender como ocorre esse 'consumo', as relações entre os atores que apresentam demandas e produzem informação, as mudanças de comportamento nos diversos níveis – do indivíduo à sociedade.

Nesse contexto, a ciência da informação deve atender não somente demandas de investigação relacionadas à organização, recuperação e gestão da informação, a partir de problemas existentes. Em outras palavras, a concepção, pesquisa e inovação no campo dessa área devem antever soluções e modelos de serviços a partir de sua capacidade de prospectar ações de investigação baseadas no conhecimento da área e na capacidade de reflexão dos seus pesquisadores.

[1] Data Shadow, termo forjado pelo pesquisador Alan Westin da Universidade de Colúmbia nos anos 90, faz menção ao enorme volume de dados gerado pelas organizações a respeito de informações de cada indivíduo, redefinindo os conceitos de privacidade e liberdade.

Uma hipótese provável é a de que a área da ciência da informação, à medida que suas atividades de investigação (por meio da pesquisa e inovação) forem mais prospectivas do que reativas, será possível observar resultados e mudanças nos serviços de informação, na maneira como a sociedade define e pratica o consumo de tais serviços, gerando a transformação de paradigmas, nas relações sociais, nos processos de gestão organização, entre outros.

Essa lógica não é mera conjectura; um exemplo muito conhecido é o da disseminação das mídias sociais. Tais produtos foram agentes de mudança de ordem cultural e organizacional, operadas pela maneira como a informação é veiculada, pela redefinição da privacidade dos indivíduos, entre outras.

As mudanças também estão afetando o fazer científico, contexto em que se apresenta o conceito de E-Science, com práticas bem diferenciadas em toda a cadeia de produção do conhecimento científico.

Assim, o propósito deste artigo é apresentar um debate sobre a organização das informações como repositórios de dados abertos, seus padrões tecnológicos, motivação e implicações de ordem científica e da atuação de profissionais e serviços.

ECOSSISTEMAS DE INFORMAÇÃO

As organizações, pessoas, instituições de pesquisa estão 'cercadas' por dados em toda parte. Os dados estão nos documentos dispersos pelos computadores, *drives* virtuais, em bases de dados e repositórios, em bancos de dados estruturados, muitas vezes na ordem de milhões de registros com granularidade baixa.

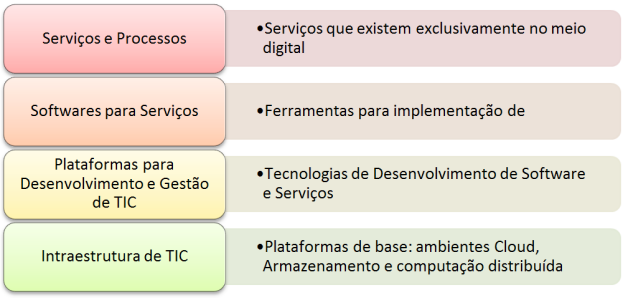
De acordo com Ericson (2010), as informações são elementos centrais em nossas vidas para que possamos direcionar e decidir nossas ações, numa teia complexa na qual cada indivíduo é ao mesmo tempo um produtor/consumidor de dados e informação.

Para se ter uma ideia, de acordo com Heath e Bizer (2011), existe todo um mercado de serviços no qual o principal insumo são repositórios de dados produzidos por outras empresas. As organizações abrem seus dados para empresas de terceiros oferecerem novos serviços a partir desses repositórios. Os autores citam o exemplo da *Amazon*, *Google e Yahoo!* que fornecem seus dados para um 'ecossistema de afiliados' que oferecem produtos e serviços a partir das informações produzidas por aquelas empresas. Um fornecedor de *software* para plataformas móveis pode, por exemplo, utilizar a plataforma de dados dos produtos *Amazon*[2] para obter receita a partir de propaganda ou intermediar venda de produtos.

Assim, determinado conjunto de tecnologias da informação e comunicação pode potencializar, no decorrer do tempo, a criação de novos modelos de negócio, de serviços, influenciar ou definir comportamentos sociais, possibilitar novos processos de gestão, entre outros.

Essa cadeia de produção pode ser mais bem compreendida a partir da ilustração da figura 1.

Figura 1 - Visão geral da cadeia de geração de serviços intensivos em TIC



Fonte: Elaborado pelo autor

A concepção, e até mesmo a existência, de produtos, serviços e processos é fortemente influenciada pelas tecnologias que baseiam sua infraestrutura. Esse pressuposto pode ser constatado pela observação das diferentes empresas no setor de TIC (tecnologias da informação e comunicação) e os produtos que elas ofertam.

[2] Saiba mais em https://associados.amazon.com.br/

No primeiro nível, 'Infraestrutrua de TIC', estão as empresas provedoras de serviços e produtos de equipamentos para computação, armazenamento de dados e telecomunicações. Podemos incluir nessa categoria as empresas que fornecem serviços no segmento do *Cloud Computing*, que possibilita a milhares de outras empresas terem infraestrutura computacional e de armazenamento de dados totalmente virtualizadas.

Os serviços de virtualização são pagos de maneira assemelhada a um contrato de locação; o cliente escolhe um plano de serviço de acordo com suas necessidades. Há também modelos de serviço em que o cliente contrata 'créditos' antecipadamente em determinado valor e ajusta, através de um painel de controle, os recursos consumidos pelo equipamento virtual (memória, espaço de armazenamento, numero de processadores, rede). Assim, por exemplo, não seria necessário contratar a configuração fixa de um computador virtual; basta ajustar um painel de controle para que sejam alocados automaticamente mais recursos computacionais dentro de limites configuráveis pelo cliente. Quanto mais recursos consumidos, mais 'créditos' são pagos.

A variedade de formatos de serviço transformou radicalmente o conceito de infraestrutura computacional. Pode-se afirmar com segurança que constituir infraestrutura para uma organização, independentemente de seu tamanho, não se resume mais à compra de equipamentos físicos e contratação de *links* de dados; atualmente qualquer pessoa pode contratar um ambiente virtual totalmente adaptado às suas necessidades de infraestrutura com confiabilidade maior (os níveis de SLA[3] são altos) e com menor custo direto[4].

No nível 'Plataformas para Desenvolvimento e Gestão de TIC' encontramos empresas do segmento de fornecimento e desenvolvimento de *software* voltados para gerenciamento de infraestrutura. São exemplos desses serviços os provedores de hospedagem de *sites*, *softwares* para criação e gestão de conteúdo *on-line* (CMS – *Content Management Systems*), empresas especializadas em serviços de gestão de sistemas de gerenciamento de banco de dados (SGBD), segurança e gerenciamento de dados. Geralmente os clientes dessas empresas têm perfil para gestão de TIC e atividades afins, no escopo de utilização de recursos de infraestrutura.

No nível 'Softwares para Serviços' encontramos as empresas dedicadas ao desenvolvimento de *softwares* e serviços voltados para atividades finalísticas das organizações.

Tais fornecedores consomem serviços de infraestrutura e gestão de infraestrutura citados anteriormente; assim, mantêm pessoal e *expertise* focados no seu negócio. Podemos citar como exemplo conhecido a empresa *Duracloud,* que oferece serviços para hospedagem de repositórios feitos com software *Dspace*. A empresa, por sua vez, utiliza intensivamente ambientes de virtualização e é exatamente por isso que seu negócio é economicamente viável.

Por fim, no último nível 'Serviços e Processos', enquadram-se de maneira genérica os serviços intensivos em TIC. Eles se sustentam fundamentalmente na estrutura anterior, seus modelos de negócio envolvem o consumo de serviços ofertados por empresas focadas nos níveis anteriores.

Um exemplo digno de citação são as bibliotecas e repositórios digitais. São baseados totalmente nos recursos apresentados anteriormente, e guardadas as proporções de escala, os modelos atuais desses serviços requerem um planejamento que envolva a utilização de recursos nos níveis anteriores. O projeto de um repositório digital ou biblioteca digital requer, claro, outras dimensões de concepção, porém elas não são suficientes para possibilitar o desenvolvimento de um projeto que seja sustentável.

[3] Service Level Agreement : nível de acordo de serviço. Trata-se das garantias contratuais ou termos de serviço que o fornecedor consegue garantir. Por exemplo, percentual de disponibilidade do serviço, nível de falhas toleradas, entre outros.

[4] Custo direto: são os investimentos aplicados com a contratação de serviços e/ou equipamentos e custos de pessoal para seu gerenciamento.

LINKED DATA E OPEN DATA

As noções de *Open Data* e *Linked Data* não são algo totalmente novo. Trata-se de conceitos relacionados à possibilidade de abrir – com ou sem restrições de acesso – dados de forma estruturada, nomeada, descentralizada, interconectada e compartilhável.

Heath e Bizer (2011) definem *Linked Data* do seguinte modo:

> O termo Linked Data se refere a um conjunto de melhores práticas para publicação e interligação de dados estruturados na Web. Estas melhores práticas foram introduzidas por Tim Berners-Lee no seu conceito de de arquitetura Web e tem sido conhecidas como Princípios do Linked Data. Estes princípios são os seguintes: 1. Usar URIs [Uniform Resource Identificator] como nomes para as coisas; 2. Usar HTTP URIs, então as pessoas possam identificar esses nomes. 3. Quando alguém usar uma URI deve-se identificar uma informação utilizável, especificada num padrão RDF; 4. Incluir links para outras URI, de modo que se possa descobrir outras coisa.s (tradução do autor)

É um paradigma de publicação de conteúdo diferente do que tradicionalmente se opera na Internet. Nos *sites* convencionas as informações e conteúdos estão armazenados em documentos baseados em HTML, em aplicativos para web e em dados que podem ser transformados, com apoio de algum aplicativo, em dados HTML legíveis por um navegador.

A vocação do HTML é estruturar conteúdo para pessoas. Mesmo com os avanços de outras linguagens de programação, que enriqueceram as possibilidades de desenvolvimento de conteúdo para Web, os dados e as informações são armazenados de maneira a serem tratadas e transformadas em conteúdo interpretável pelos navegadores web com foco na apresentação.

A ideia de *Linked Data* propõe a estruturação do conteúdo de acordo com uma arquitetura simples, de modo que seja facilmente compartilhável em escala global.

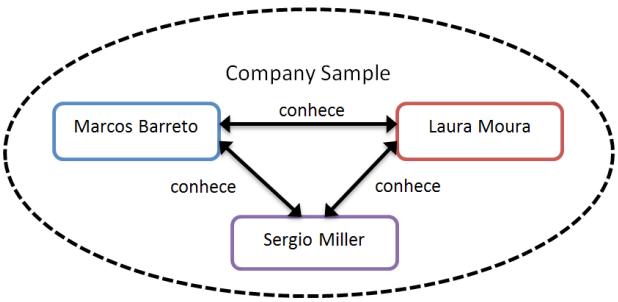
O primeiro princípio defende que o uso de URIs para identificar não apenas documentos e *sites* mas também os objetos e coisas (conceitos) do mundo real. Por exemplo, devemos nomear e categorizar as 'coisas' – doravante denominadas conceitos – como por exemplo, pessoas, lugares, categorias de objetos, entre outras coisas.

URI não é o mesmo que URL. Sayão (2007), fazendo menção à norma RFC 2396, define que URI é um formato padronizado (canônico) para nomear um recurso. Uma *Uniform Resource Locator* (URL) é um tipo de URI que identifica um recurso através da sua localização. Um tipo comum de URL são os endereços de páginas web que acessamos cotidianamente.

A nomeação de coisas do mundo real com URIs tem uma forma semelhante à que se observa nas URLs; nas últimas, um endereço http://www.empresa.com.br/pagina.html define incialmente o domínio chamado 'empresa' de natureza comercial registrado no Brasil, que contém um recurso chamado pagina.html. Nomear recursos no contexto de *Linked Data* segue os mesmos padrões da Web (Heath; Bizer, 2011), formando o que Berners-Lee et al. (2004) denominam HTTP URI.

Citando um exemplo de Heath e Bizer (2011), para referenciar três pessoas que se conhecem e trabalham na mesma empresa, podemos usar a seguinte estruturas (fig. 2):

Figura 2 - Exemplo de pessoas e relações representáveis através de URI



Fonte: adaptado de Heath e Bizer (2011)

As entidades e relações poderiam ser nomeadas da seguinte forma:

- Marcos Barreto:

http://www.semanticweb.org/divinoignacio/ontologies/2014/8/company-sample#marcos-barreto

- Laura Moura:

http://www.semanticweb.org/divinoignacio/ontologies/2014/8/company-sample#laura-moura

- Sergio Miller:

http://www.semanticweb.org/divinoignacio/ontologies/2014/8/company-sample#sergio-miller

- As ligações entre eles podem ser nomeadas usando o padrão FOAF (*Friend of a Friend) publicado por* Brickley e Miller (2014) na seguinte forma:

http://xmlns.com/foaf/0.1/knows

Cada uma dessas URI representa uma entidade ou relação do mundo real. Para implementar essas informações utilizamos o modelo de dados RDF, uma forma de estruturação de dados em linguagem XML destinada à codificação das declarações (as pessoas e suas relações), de tal maneira que os dados possam ser processados por computadores e utilizados para exibição.

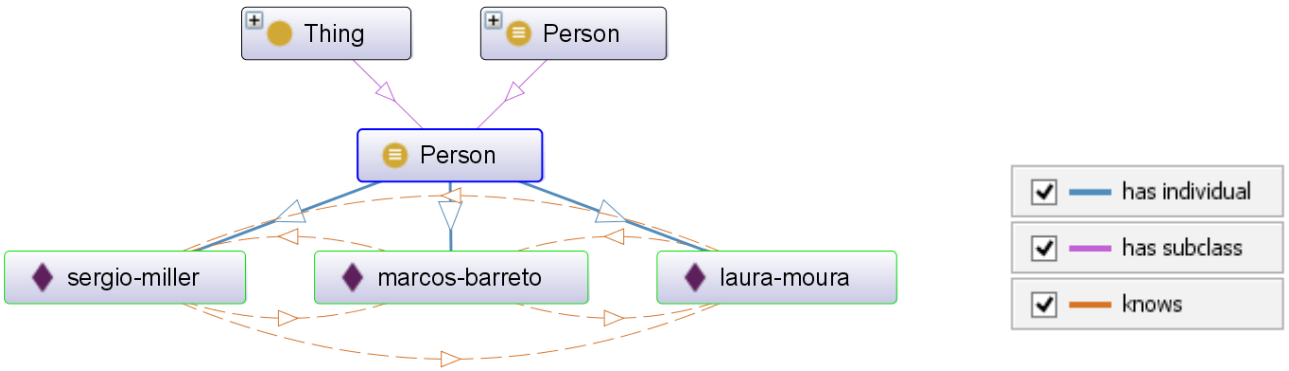
Isso significa que é possível implementar a semântica dos dados e de sua estrutura. Em outras palavras, o *software* que tiver condições de ler o arquivo RDF[5] com essas declarações poderá interpretar e utilizar a informação de que as pessoas Marcos Barreto, Laura Moura e Sérgio Miller se conhecem e trabalham na empresa Company Sample.

Tal implementação não seria possível utilizando-se apenas linguagem HTML e *links*. Os *links* apenas ligam recursos para exibição, mas os *softwares* que os processam não têm condições de extrair a semântica embarcada nos dados e ligá-los a outras fontes nas quais as mesmas entidades estão presentes. (Marcondes, 2012)

A padronização das URI na forma HTTP permite que os recursos sejam recuperáveis em qualquer parte da Internet. Quando elaboramos uma página da web e incorporamos a declaração URI para Laura Moura conhece Sergio Miller, qualquer outra página ou recurso da Web que citar uma dessas pessoas poderá identificar uma relação '*knows*' entre os dois. A implementação dessa semântica dos dados não é possível apenas com uso de HTML e hiperlinks.

A implementação em RDF dessas entidades e suas relações teria a da seguinte estrutura (fig. 3).

Figura 3 – Grafo de entidades e relações obtido a partir de uma estrutura RDF de dados ligados



Fonte: o autor

[5] Saiba mais em http://www.w3.org/RDF/

Figura 4 – Fragmento do arquivo em RDF que implementa as entidades e relações

Fonte: o autor

Esse exemplo de código mostra como as relações entre as entidades do mundo real aparecem codificadas, ilustrando a essência do conceito de *Linked Data*. Os dados são ligados e podem ser processados por um mecanismo de busca ou aplicativo com um *reasoner*, um mecanismo de inferência capaz de identificar as relações presentes nas declarações RDF e utilizá-las para realizar novas buscas e conexões com outros documentos, tanto no mesmo local quanto pela Internet.

Deve-se enfatizar a ideia de que *Linked Data* nos apresenta um novo paradigma de organização e disseminação de conteúdo sem, no entanto, reinventar a Internet e suas tecnologias. Isso significa que é possível embarcar esses dados em documentos da Web para que eles contenham a representação semântica que viabiliza a utilização de aplicações para Web semântica, através de sistemas de organização e busca que usem tais recursos.

Ilustrando uma comparação, citemos os repositórios digitais e as bases de dados bibliográficas. São produtos que ofertam acesso a documentos e metadados, com recursos para descrição dos documentos através de metadados e busca para recuperá-los. *Softwares* desse tipo têm evoluído bastante em termos tecnológicos com a agregação de técnicas e protocolos de compartilhamento e interoperabilidade, criação de padrões e preservação digital, armazenamento de documentos digitais em larga escala, técnicas e refinamentos para busca (busca por relevância, busca facetada, filtros, etc). Nos últimos 20 anos, vimos protocolos como o Z39.50[6], OAI-PMH[7], disseminação do padrão Dublin Core[8] em ambientes digitais, softwares como Dspace[9], Omeka[10] e Greenstone[11] se popularizarem como ferramentas para criação desses produtos. Nos últimos dez anos, popularizou-se bastante a criação de repositórios de documentos obtidos pelo compartilhamento de outras fontes, principalmente pela disseminação dos protocolos OAI-PMH, OAI-ORE e em menor escala, do protocolo SWORD[12].

Tais produtos têm na sua essência a mesma forma de concepção: um produto que organiza metadados previamente produzidos e os respectivos documentos digitais, associados a uma interface de busca e

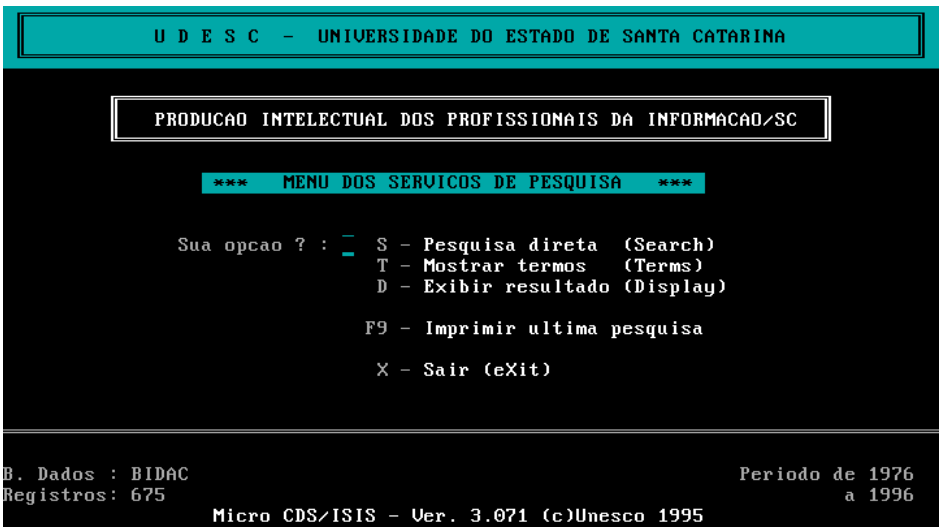
[6] http://www.niso.org/standards/resources/Z39.50_Resources
[7] http://www.openarchives.org/
[8] http://dublincore.org
[9] http://www.dspace.org/
[10] http://omeka.org/about/
[11] http://www.greenstone.org/
[12] http://swordapp.org/

navegação pelos registros bibliográficos. A figura 5 ilustra uma base de dados de registros bibliográficos desenvolvida com o software Micro CDS/ISIS há cerca de 20 anos, para ambiente MS DOS, atualmente recuperada e funcional em uma máquina virtual.

Ela oferece recursos de pesquisa por busca de termos e por navegação, e exibe listagens das referências bibliográficas.

O produto a seguir é uma base de dados desenvolvida com o software Dspace (fig. 6):

Figura 5 –Tela de entrada da base de dados BIDAC desenvolvida em ambiente DOS



Fonte: base de dados de autoria de Ohira et al. (1997), restaurada pelo autor[13]

Figura 6 – Página inicial do repositório experimental do LabTecGC/UDESC

Fonte: o autor

[13] A máquina virtual para Oracle Virtualbox com a base de dados BIDAC está disponível em http://www.labtecgc.udesc.br/bidac/basebidac.ova

Encontramos nesses dois produtos os mesmos conceitos de organização e recuperação de informação, mesmo que eles tenham suas tecnologias distantes em cerca de 20 anos.

Comparativamente, os repositórios de dados baseados em dados abertos e dados ligados (*Open Data* e *Linked Data*) possuem uma organização orientada aos princípios do *Linked Data* citados anterioremente.

Tais repositórios são organizados como de *Triple Stores*; o termo tripla faz menção à estrutura atômica de declaração usada em RDF: URI do recurso | URI da propriedade | URI do valor, como no exemplo da figura 4.

Alguns exemplos de repositórios de dados (*Triple Stores*):

- Databib: (http://databib.org/connect.php);
- DBpedia Data sets: (http://wiki.dbpedia.org/Datasets);
- Biblioteca Nacional da Espanha: (http://www.bne.es);
- Europeana: (http://labs.europeana.eu/api/linked-open-data/data-downloads/);
- Bigdata: (http://www.bigdata.com/).

No caso da Biblioteca Nacional da Espanha, os registros bibliográficos em formatos convencionais estão sendo convertidos para um *Triple Store* contendo as informações existentes e com o estabelecimento de novas ligações entre os conjuntos de dados, em padrões aderentes ao RDF e SKOS[14].

Diversas instituições estão abrindo seus dados na forma de repositórios de dados baseados na ideia de *Linked Data*, como uma maneira de facilitar o acesso às informações. Um mecanismo de busca semântica capaz de processar esses dados conseguirá identificar, mapear e interligar conceitos comuns existentes nesses repositórios, independentemente da necessidade de módulos de conversão ou exposição de metadados. Isso é bem diferente do que se conhece por interoperabilidade nos dias atuais; hoje dependemos de protocolos comuns para intercâmbio, padronização de metadados, módulos de conversão e de provedores de dados, ex.: protocolo OAI-PMH e padrão DC.

Mesmo que as bases de dados e repositórios ofereçam recursos para interoperar os metadados, ainda temos produtos de informação 'ilhados em escala', sempre com limitações impostas pelas diferenças dos padrões de metadados conhecidos e da disponibilidade de módulos de conversão ou intercâmbio.

E-SCIENCE

O que se entende por *E-Science*? Acredita-se que a ciência será cada vez mais realizada em grande escala por meio de colaborações em nível global, distribuídas pela Internet. Elas terão acesso a coleções de dados muito grande, computação em larga escala e de alta *performance*, que permitirão realizar experimentos de maneira escalar, incremental e compartilhada. Esse conceito apresentado por Taylor (2000), na época diretor geral do Conselho de Pesquisa Britânico, anunciava a iniciativa de implantação de uma infraestrutura computacional distribuída e de dados em larga escala que seria a base para um novo modo de fazer ciência.

Segundo Medeiros e Caregnato (2012):

> [...] é importante considerar que a e-Science altera fundamentalmente a maneira com que os cientistas realizam seu trabalho, as ferramentas que usam, os tipos de problemas que abordam e a natureza da documentação e da publicação que resulta da sua pesquisa.

De imediato se observa que as práticas da atividade científica nesse contexto diferem das formas tradicionais. O uso intensivo de poderosas infraestruturas computacionais (*grid computing*), compartilhamento e projetos conjuntos em nível global e a produção, armazenamento e compartilhamento de repositórios em larga escala são a principal característica do *modus operandi* dessas práticas.

[14] http://www.w3.org/2004/02/skos/

Nesse sentido, Appelbe e Bannon (2007) em seu trabalho denominam essas atividades *eResearch*, enfatizando ao afirmar que determinada comunidade científica promove suas atividades nesse contexto, faz necessário distinguir com clareza como e com que recursos elas o fazem, a fim de definir se realmente se trata do contexto do e-Science.

Esses autores destacam no quadro 1 algumas características que diferenciam o modo tradicional do e-Science:

QUADRO 1 – Paradigmas tradicionais versus e-Science

Característica	**e-Science (eResearch)**	**Pesquisa tradicional**
Participantes	Equipes de pesquisa distribuídas com diferentes habilidades	Pesquisador individual ou equipe de pesquisa local (pequena)
Dados	Gerados, armazenados e acessíveis de locais distribuídos (não centralizados)	Gerados, armazenados e acessíveis apenas no local da pesquisa
Computação e Instrumentação	Computação em larga escala ou sob demanda ou acesso a instrumentos computacionais compartilhados	Processamento de dados é executado nos computadores do próprio pesquisador
Rede	Suportada pela Internet e pela sua infraestrutura	Não é apoiada ou distribuída pela Internet
Disseminação da Pesquisa	Através de portais especializados	Através de publicações impressas/ eletrônicas e apresentações em conferências

Fonte: adaptado de Appelbe e Bannon (2007)

Os autores reiteram que não se trata apenas do contexto de pesquisa em computação ou engenharia; o e-Science pode se dar em qualquer área do conhecimento. Medeiros e Caregnato (2012, p.318) complementam a ideia do e-Science com o seguinte pensamento:

> Com a possibilidade de contar com uma infraestrutura capaz de subsidiar pesquisas através do compartilhamento e reuso de dados já coletados, como exposto, não cabe ao pesquisador despender um longo tempo no processo de coleta de dados se este procedimento já foi realizado e os dados estão disponíveis. Isto contribui também para a produção do conhecimento científico, uma vez que não há necessidade de retroceder à parte de um processo que já foi executada, bem como evita que recursos já concedidos por agências de fomento sejam duplicados.

Nesse sentido, as práticas no contexto do e-Science requerem investimentos de infraestrutura de telecomunicações, computação distribuída de larga escala e ambientes de armazenamento e repositórios de dados que sejam reusáveis e compartilháveis.

CONSIDERAÇÕES FINAIS

De maneira análoga à ideia de 'Ecossistema de informação' apresentada no início deste trabalho, a infraestrutura ou '*cyber structure*' como definem alguns autores, os repositórios de dados no contexto do e-Science dependem de um conjunto complexo de serviços e infraestrutura direcionados para a atividade científica.

Repositórios de dados estruturados segundo os princípios de *Linked Data* são, hoje em dia, uma alternativa mais consistente e viável do que o paradigma atual de produtos de organização e recuperação de informação e documentos (metadados + arquivos). Os *Triple Stores* oferecem um diferencial que esse paradigma atual não contempla: compartilhar dados segundo os princípios da Web semântica.

Os repositórios de dados requerem essa estruturação para que sejam, efetivamente, abertos e que possam ser acessados de maneira inteligente. Em outras palavras, não basta dispor tais dados ('*data dump*'); é necessária uma organização planejada, o que requer

novas competências e habilidades dos profissionais que lidam com esses repositórios.

Algumas questões que ainda merecem estudos científicos para que as práticas do e-Science se consolidem; pode-se citar as seguintes: Como os bibliotecários e os serviços de informação se capacitarão e atuarão no contexto do E-Science? Que habilidades precisamos para nos tornarmos '*Data Scientists*'? Como lidar com as restrições de direitos autorais, segurança, privacidade, interesses comerciais legítimos?

Há uma dimensão relevante: a das formas de comunicação científica no contexto do e-Science. Mons e Velterop (2009) nos trazem o conceito de 'Nanopublicações', uma ideia que faz menção à publicação de artefatos ou *outputs* das atividades do e-Science com uma granularidade baixa, através de portais especializados.

Os autores ilustram o processo da seguinte maneira: a) organização de conceitos; b) a partir dos conceitos, a organização de declarações (que serão registradas na forma RDF); c) a partir das declarações o registro de anotações (o mesmo processo conhecido como anotação semântica) relacionando contribuições dos pesquisadores com os conceitos e declarações registradas; d) das anotações para as nanopublicações.

As nanopublicações podem ser compartilhadas e encontradas por sistemas de busca semântica, pelo fato de estarem registradas como declarações e anotações semânticas em RDF. Elas devem ocorrer em todo o processo de pesquisa, em vez de acontecer ao final do projeto, como ocorre na atualidade.

Por fim, acredita-se que tais paradigmas redefinirão a maneira como se pensam os moldes canônicos da organização e recuperação da informação, especialmente no contexto da ciência da informação. Processos e princípios da catalogação, da comunicação científica e da gestão de periódicos ganharão novos matizes. É um caminho inevitável, e tal afirmação pode se verificar com os crescentes investimentos em estruturas de alto poder computacional e os investimentos das agências de fomento, inclusive no Brasil.

REFERÊNCIAS

APPELBE, Bill; BANNON, David. eResearch: paradigm shift or propaganda? *Journal of Research & Practice in Information Technology*. [s.l.], v.39, n.2, p.83-90, 2007. Disponível em:<http://search-ebscohost-com.ez46.periodicos.capes.gov.br/login.aspx?direct=true&db=iih&AN=24911628&lang=pt-br&site=ehost-live&authtype=ip,cookie,uid>. Acesso em: 13 ago. 2014.

BERNERS-LEE, Tim et al. *Architecture of the World Wide Web, volume one.* Disponível em: <http://www.w3.org/TR/webarch/>. Acesso em: 23 jun. 2014.

BRICKLEY, Dan; MILLER, Libby. *FOAF Vocabulary Specification 0.99.* Disponível em: <http://xmlns.com/foaf/spec/20140114.html>. Acesso em: 14 jan. 2014.

ERICSON, Jim. Net Expectations: what a web data service economy implies for businnes. *Information Management*. Nova York, v.1, 2010. Disponível em:<http://www.information-management.com/issues/20_1/net-expectations-10016922-1.html?zkPrintable=1&nopagination=1>. Acesso em: 07 jul. 2014.

HEATH, Tom; BIZER, Christian. *Linked Data*: evoluing the web into a global data space. Morgan e Claypool, 2011. 122 p., (Synthesis lectures on the semantic web).

MARCONDES, Carlos Henrique. *'Linked Data*': dados interligados e interoperabilidade entre arquivos, bibliotecas e museus na web. *Encontros Bibli (UFSC),* Florianópolis, v.17, n.34, p. 171-192, 2012.

MEDEIROS, Jackson da Silva; CAREGNATO, Sônia Elisa. Compartilhamento de dados e e-Science: explicando um novo conceito para comunicação científica. *Liinc em Revista*, v.8, n.2, p.311-322, set. 2012.

MONS, Barend; VELTEROP, Jan Nano-Publication in the e-science era. *Workshop on Semantic Web Applications in Scientific Discourse (SWASD 2009).* Washington: W3C, 2009.

OHIRA, Maria Lourdes Blatt et al. Produção científica em biblioteconomia no estado de Santa Catarina. *Transinformação*, v.9, n.3, p.68-87, set./dez., 1997.

SAYÃO, Luís Fernando. Padrões para bibliotecas digitais abertas e interoperáveis. *Encontros Bibli (UFSC),* Florianópolis, v.1, n. esp.1, 2007. Disponível em:<https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2007v12nesp1p18>. Acesso em: 10 mar. 2014.

SAYÃO, Luis et al. *Implantação e Gestão de Repositórios Institucionais*: políticas, memória, livre acesso e preservação. Salvador: EDUFBA, 2009. 365 p.

TAYLOR, John. *Defining e-Science.* Disponível em: <http://www.nesc.ac.uk/nesc/define.html>. Acesso em: 23 jul. 2014.