



UNIVERSIDADE FEDERAL DE PERNAMBUCO
GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO
CENTRO DE INFORMÁTICA
2012.2

Criação e publicação de um dataset de dados interligados
das edições passadas do Simpósio Brasileiro de Banco de
Dados

PROPOSTA DE TRABALHO DE GRADUAÇÃO

Aluno: Mateus Gondim Romão Batista (mgrb@cin.ufpe.br)

Orientador: Bernadette Farias Lóscio (bfl@cin.ufpe.br)

Recife, 26 de Janeiro de 2013

1. Contextualização

Desde sua criação, a web passa por um processo contínuo de transformação. Nos seus primeiros anos, usuários podiam visualizar páginas web e navegar entre as mesmas através de hyperlinks, mas não tinham a possibilidade de contribuir com o conteúdo das respectivas páginas. Com a chegada da Web 2.0, usuários passaram a ter a experiência de interagir e colaborar entre si por meio de um diálogo em mídia social como criadores de conteúdo gerado por usuários em uma comunidade virtual, em contraste aos websites onde as pessoas eram limitadas à visualização passiva de conteúdo [5]. Aspectos como interatividade, participação e padronização ganharam destaque. Além das motivações diretamente relacionadas à experiência dos usuários, a quantidade de aplicações desenvolvidas para lidar de alguma forma com os dados que circulam nesse ambiente web cresceu consideravelmente. No entanto, há uma grande dificuldade de lidar com esses dados por uma série de motivos, como formato, mas principalmente pela ausência de semântica entre os dados. Dessa forma, a Web ainda não encontra-se em um estado interessante de legibilidade para aplicações.

Nesse contexto, surgiu o conceito de Web Semântica. Segundo [2], a Web Semântica é a extensão da *World Wide Web* que permite às pessoas compartilharem conteúdo além dos limites de aplicações e *websites*. Encorajando a inclusão de conteúdo semântico em páginas web, a Web Semântica visa converter a web atual dominada por documentos estruturados e semi-estruturados em uma “Web de dados” [4]. A implementação deste conceito possibilita que aplicações tirem mais proveito dos dados disponibilizados, ao dotá-los de mais semântica.

Para tornar a web semântica uma realidade, uma série de novas tecnologias foram adotadas, como *Resource Description Framework* (RDF), *Web Ontology Language* (OWL), *Extensible Markup Language* (XML), entre outros. Conectado com estes novos atores da Web, surgiu um conjunto de princípios e tecnologias chamado *Linked Data*. *Linked Data* refere-se a empregar o *Resource Description Framework* (RDF) e o *Hypertext Transfer Protocol* (HTTP) para publicar dados estruturados na Web e conectar dados de diferentes fontes de dados, efetivamente permitindo que dados em uma fonte sejam ligados a dados em outra fonte de dados [1]. Este conceito tem ganhado crescente destaque e adesão da comunidade acadêmica, e por isso, já existem grandes aplicações na web implementando o mesmo.

Como exemplo do uso de *Linked Data*, na World Wide Web Conference 2012, houve uma iniciativa com o objetivo de traduzir dados e metadados sobre papers aceitos, autores, comitês, programação e locais de vários formatos para RDF. Com isso, aplicações existentes ou futuras que usam a Web de dados poderiam integrar essa porção extra de conhecimento para aprimorar seu serviço, aperfeiçoar seu *mashup* e melhorar suas estatísticas [6].

Promovido anualmente pela Comissão Especial de Banco de Dados da Sociedade Brasileira de Computação (SBC), o Simpósio Brasileiro de Banco de Dados (SBBD) reúne

pesquisadores, estudantes e profissionais do Brasil e do exterior, que apresentam e discutem temas relacionados aos últimos avanços da área. Tradicionalmente, este evento tem reunido em torno de 500 participantes e é o maior evento na América Latina para a apresentação e discussão de resultados de pesquisa relacionados à área de banco de dados. Além de sessões técnicas e tutoriais, o simpósio também inclui tutoriais e palestras convidadas apresentadas por pesquisadores de renome da comunidade nacional e internacional [3].

2. Objetivos

Este trabalho tem como objetivo a criação e publicação de um dataset com dados relevantes sobre as últimas dez edições do Simpósio Brasileiro de Banco de Dados(SBBD), bem como o desenvolvimento de uma aplicação para a visualização dos respectivos dados. O conteúdo abrangerá desde dados sobre a estrutura e organização de cada evento, como comissões e palestras, até dados dos artigos apresentados, autores e suas respectivas afiliações.

O repositório será criado no formato RDF, e seu esquema será projetado para reaproveitar vocabulários relevantes com a adição de termos próprios para que o esquema contemple as particularidades do evento. Após essa fase, o povoamento será feito a partir dos dados presentes nos sites dos eventos passados e em seus respectivos anais, além de uma base relacional criada em 2010 para a celebração dos 25 anos de evento, cujos dados serão importados para o dataset criado. Na etapa seguinte, será criado um SPARQL endpoint para a publicação dos dados, disponibilizando a execução de consultas SPARQL sobre o repositório. Por fim, uma aplicação será desenvolvida para permitir a visualização de dados em formato de gráficos e tabelas.

Com este trabalho, o repositório de dados ligados terá um alto grau de flexibilidade e reusabilidade, podendo ser utilizado por demais aplicações que desejem manipular os dados históricos do evento. Além disso, com a aplicação para visualização dos dados do dataset, será possível ter um panorama histórico mais elaborado do evento.

3. Cronograma

O cronograma de atividades previsto para o desenvolvimento deste Trabalho de Graduação é apresentado na Tabela 1.

Atividade	Janeiro				Fevereiro				Março				Abril			
Formulação da proposta e análise da literatura	X	X	X	X												
Construção do esquema			X	X	X											
Povoamento do dataset						X	X									
Publicação do dataset em SPARQL Endpoint								X	X							
Desenvolvimento da aplicação e visualização de dados									X	X	X	X				
Preparação para apresentação e defesa													X	X		

Tabela 1: Cronograma de atividades

4. Possíveis Avaliadores

Os possíveis avaliadores desse trabalho de graduação serão:

- Ana Carolina Salgado
- Bernadette Farias Lóscio

5. Referências

- [1] BIZER, C., HEATH, T., IDEHEN, K., BERNERS-LEE, T. Abril, 2008. Linked data on the web (LDOW2008). In *Proceeding of the 17th international conference on World Wide Web* (pp. 1265-1266). ACM.
- [2] SEMANTIC WEB WIKI, 2013. Disponível em: <http://semanticweb.org/wiki/Main_Page>. Acesso em: 9 jan. 2013.
- [3] SIMPÓSIO BRASILEIRO DE BANCO DE DADOS. SBBD, 2012. Disponível em: <<http://sws2012.ime.usp.br/sbbd/index.php>>. Acesso em: 25 jan. 2013.
- [4] WIKIPEDIA, THE FREE ENCYCLOPEDIA. Semantic Web, 2013. Disponível em: <http://en.wikipedia.org/wiki/Semantic_Web>. Acesso em: 9 jan. 2013.
- [5] WIKIPEDIA, THE FREE ENCYCLOPEDIA. Web 2.0, 2013. Disponível em: <http://en.wikipedia.org/wiki/Web_2.0>. Acesso em: 17 jan. 2013.
- [6] WORLD WIDE WEB CONFERENCE 2012. Metadata Initiative, 2012. Disponível em: <<http://www2012.wwwconference.org/committee/metadata-initiative/>>. Acesso em: 25 jan. 2013.

6. Assinaturas

Mateus Gondim Romão Batista
Orientando

Bernadette Farias Lóscio
Orientador