Universidade Federal de Pernambuco

Graduação em Ciência da Computação

Centro de Informática

2012.2



Criação e publicação de um dataset de dados interligados sobre os docentes do CIn

**Proposta de Trabalho de Graduação**

**Aluno:** Wanderson de Lima Carneiro (wlc@cin.ufpe.br)

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

Recife, 23 de Janeiro de 2013

# Contextualização

A web, hoje, é uma ferramenta que faz parte da vida de milhares de pessoas em todo o mundo, sendo composta de uma imensa quantidade de dados que cresce bastante a cada dia. A maioria desses dados são criados apenas com fins de apresentação para o usuário como, por exemplo, as páginas web que visitamos em nosso dia a dia. Uma página web é composta basicamente de conteúdo em linguagem natural, tags html/css e linguagens de scripts, sendo esses últimos responsáveis pela forma como o conteúdo da página deve ser apresentado para o usuário. As tecnologias usadas na criação e formatação das páginas web atingem o objetivo de proporcionar conteúdo compreensível para o usuário, porém o mesmo é de difícil interpretação por parte dos computadores devido à falta de semântica contida nos dados, o que faz com que a imensa quantidade de dados disponível na web não possa ser usada em sua plenitude.

Para a resolução da falta de semântica dos dados na web, novas tecnologias e novos conceitos foram criados e um deles é o de web semântica. Web semântica ou web de dados não é uma substituição da web atual, mas sim a evolução da mesma, onde os dados possuem significado semântico, de forma que a informação é interpretável por um computador, permitindo que pessoas e softwares possam trabalhar em cooperação. Na web semântica, os dados podem ser acessados pela arquitetura web atual e devem apresentar relacionamentos um com os outros [1]. O propósito da web semântica pode ser obtido através do uso de tecnologias integradas(XML,RDF,OWL, SPARQL, entre outras...) e do uso de metadados [2] - dados sobre outros dados - os quais proverão informação semântica que permite aos computadores interpretar os conteúdos disponíveis na web, tornando assim a web interpretável.

Para que a web semântica seja uma realidade, é preciso que uma grande quantidade de dados esteja disponível na web em formatos padrões, acessíveis e gerenciáveis pelas ferramentas que fazem parte da web semântica. Além disso, é necessário que esses mesmos dados estejam interligados para que ocorra a criação da web de dados [4]. Da necessidade da criação de padrões para a publicação dos dados na web e do modo como eles devem se interligar, surgiu, dentro da web semântica, o conceito de Linked Data (dados ligados). Linked Data refere-se ao conjunto de melhores práticas para a publicação de dados estruturados na Web. Tim Berners-Lee definiu quatro princípios que caracterizam dados ligados os quais podem ser resumidos da seguinte forma: (i) o uso de URIs como nome para coisas; (ii) Usar URIs HTTP para que as pessoas possam encontrar esses nomes ; (iii) Quando alguém procura um URI, fornecer informações úteis usando os padrões (RDF, SPARQL); (iv) Incluir links para outras URIs, de modo a permitir que itens relacionados possam ser descobertos [3]. Linked Data faz uso de tecnologias web padrões (HTTP,URI), porém, ao invés de apenas servir páginas web às pessoas, ela permite a leitura dos dados conectados de forma automática por um computador, permitindo, desta forma, que dados de diferentes fontes de dados possam ser conectados e consultados [5].

# Objetivos

O objetivo deste trabalho consiste na criação e publicação de um dataset de dados interligados com as informações acerca dos docentes do CIn-UFPE principalmente os dados relativos às publicações realizadas pelos mesmos. A disponibilização de dados dos docentes na web será em formato aberto, padronizado no formato RDF e interligado com outros datasets relevantes, permitindo que o dataset possa ser acessado, analisado e processado por qualquer sistema.

Na criação do dataset serão, ao máximo, reutilizados termos de vocabulários conhecidos de forma a propiciar integração de dados, facilitar o acesso aos dados por softwares que desejem consumir os dados e evitar que ocorra mais pré-processamento dos dados. Após a criação do esquema do dataset, será feita a população do mesmo através de dados colhidos de fontes relevantes de informações dos docentes disponíveis na web, fazendo sempre, quando possível, uso de links de relacionamento e links de entidade com outros datasets importantes, tornando-o o mais conectado possível. Por último, será realizada a publicação na web e a disponibilização de um SPARQL endpoint, o que possibilitará a realização de consultas sobre os dados e a disponibilização dos mesmos para consumo.

# Cronograma

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

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Atividade | Dezembro | | | | Janeiro | | | | Fevereiro | | | | Março | | | | Abril | | | |
| Estudo da Literatura |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Especificação da solução |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Implementação da solução |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Elaboração do relatório |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Validação e testes da solução |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Preparação da apresentação |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Defesa |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

Tabela 1: **Cronograma de atividades**

# Possíveis Avaliadores

Os possíveis avaliadores desse trabalho de graduação serão:

* Ana Carolina Salgado
* Bernadette Farias Lóscio

# Assinaturas

Recife, 23 de Janeiro de 2013

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Wanderson de Lima Carneiro

**Orientando**

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Bernadette Farias Lóscio

**Orientador**

# Referências

[1] HERMAN, I. W3C Semantic Web Frequently Asked Questions. W3C, 2001. (atualização 2009). Disponível em: <http://www.w3.org/2001/sw/SW-FAQ>. Acesso em: 6 dez. 2012.

[2] WIKIPÉDIA, A ENCICLOPÉDIA LIVRE. Web semântica. , 2012. Disponível em: <[http://pt.wikipedia.org/wiki/Web\_semântica](http://pt.wikipedia.org/wiki/Criptografia_qu%C3%A2ntica)>. Acesso em: 6 dez. 2012.

[3] Berners-Lee, T. (2006, July). Linked Data - Design Issues.

Disponível em: <http://www.w3.org/DesignIssues/LinkedData.html>. Acesso em: 6 dez. 2012.

[4] W3C. Linked data. , 2012. Disponível em: < http://www.w3.org/standards/semanticweb/data>. Acesso em: 6 dez. 2012.

[5] WIKIPÉDIA, THE FREE ENCYCLOPEDIA. Linked data. , 2012. Disponível em: <http://en.wikipedia.org/wiki/Linked\_data>. Acesso em: 6 dez. 2012.