EAComp
Encontro Acadêmico de Computação

VI Encontro Acadêmico de Computação - UFMA
Copyright (c) EAComp. ISSN: 2594-486X

Abrindo os dados públicos da Universidade Federal do Maranhão

Micael Lopes da Silva[1], **Sérgio Souza Costa**[1]
Walysson Carlos dos Santos Oliveira[1], **Thamyla Maria de Sousa Lima**[1]

[1]Curso de Engenharia da Computação – Universidade Federal do Maranhão (UFMA)
São Luís – MA – Brasil

{micael.lopes,walysson.oliveira,thamyla.lima}@aluno.ecp.ufma.br

Abstract. *The Federal University of Maranhão (UFMA) provides various information in its data portals. However, the data is only readable to humans, not to computers. Therefore, this project aimed at the integration and availability of some unstructured public data published in the scope of the Federal University of Maranhão, through the implementation of a RESTful API for the disclosure and use of data opened by the developer society.*

Resumo. *A Universidade Federal do Maranhão (UFMA) disponibiliza diversas informações em seus portais de dados. No entanto, os dados encontram-se apenas legíveis para humanos, e não para computadores. Logo, este projeto visou a integração e disponibilização de alguns dados públicos não estruturados publicados no âmbito da Universidade Federal do Maranhão, através da implementação de uma API RESTful para divulgação e utilização de dados abertos pela sociedade desenvolvedora.*

1. Introdução

A Universidade Federal do Maranhão, sendo uma fundação federal presente em diversos municípios do estado do Maranhão, é geradora de diversos tipos de dados e, portanto, deve disponibilizar tais informações públicas para todos. Sua estrutura organizacional é composta por diversos departamentos, unidades suplementares e núcleos que geram os mais diversos tipos de dados como produção acadêmica, informações de acervos da biblioteca, dados de diretores de departamentos, coordenadores de cursos, professores e discentes em geral, portais que disponibilizam periódicos locais, publicação de editais, calendário acadêmico, refeições, estre outras informações não sensíveis e que podem ser visualizadas e redistribuídas pois são de origem pública.

No entanto, até este presente trabalho, os dados da Universidade são exclusivamente disponibilizados em páginas HTML o que não permite uma análise automatizada fácil por computador. Portanto, neste artigo mostra-se o desenvolvimento de uma plataforma de distribuição de dados abertos em forma de API. Apresenta-se na seção 2, as definições para dados públicos e abertos; na seção 3, as técnicas de *scraping* empregadas; e na seção 4, a plataforma de disponibilização de dados estruturados.

2. Fundamentação

Segundo a Controladoria Geral da União do Brasil, os dados públicos são informações que não lesem leis de privacidade, integridade e segurança [Brasil 2013]. Em governos

democráticos é direito fundamental de todo cidadão o acesso livre e sem restrição desses dados de modo que eles mesmos ajam como fiscalizadores do governo. No Brasil, a Lei nº 12.527, de 18 de novembro de 2011 regulariza o direito de acesso à informações de órgãos públicos administrativos, autarquias, fundações e empresas estatais, e todas as instituições controladas direta ou indiretamente pelo governo ou toda e qualquer instituição privada sem fins lucrativos que receba verbas públicas [lei 2011].

A interoperabilidade dos sistemas de serviço Web permite o reúso de código a partir de sistemas distribuídos garantindo a comunicação entre plataformas de origem diferentes [Ferreira Filho 2009]. Nesse quesito, a arquitetura REST é um paradigma utilizado em sistemas hipermídia que difunde as interfaces, a escalabilidade e a implementação dos componentes [Fielding 2000]. [Leonard and Ruby 2007] cunhou o termo "Serviços RESTful" para todo serviço Web que segue os paradigmas que os próprios definem, como a integração do HTTP ao REST. Logo, serviços RESTful são, segundo [Ferreira Filho 2009] compostos por cinco concepções: recurso, representação, identificador uniforme, interface unificada e escopo de execução.

3. Metodologia

A publicação de dados abertos governamentais é um processo contínuo, pois sempre existirá a demanda para abertura de novos dados e a manutenção e melhorais dos já existentes. Deste modo desenvolveu-se, como primeira etapa, uma plataforma de dados estruturados em JSON na arquitetura RESTful onde os dados requisitados via requisições HTTP são retornados serializados em JSON.

3.1. Escolha e modelagem de dados

Em aspectos gerais, esta etapa consiste no levantamento, na modelagem e na autorização para a publicação dos dados. Contudo, este projeto tem como objetivo disponibilizar em formato aberto dados que já são públicos, portanto, foi feita uma análise dos dados já disponíveis nos portais da Universidade e foram selecionados as informações mais importantes para o conhecimento de uma instituição. A partir do portal SIGAA público[1], os seguintes conjuntos de dados foram escolhidos:

- **Discentes**: engloba informações de alunos ativos da universidade, contendo os atributos nome do aluno e matrícula.
- **Docentes**: engloba as informações disponibilizadas na página pessoal de cada docente no SIGAA, podendo conter os atributos nome, descrição pessoal, áreas de interesses e link do currículo *lattes*.
- **Cursos**: dados de cada curso de graduação da UFMA. Os atributos selecionados são área de conhecimento CNPQ, convênio acadêmico, coordenador do programa, descrição, modalidade do curso, título do profissional, endereço URL no SIGAA e o número de identificação do curso.
- **Turmas**: engloba dados de instância de componentes curriculares. Atributos selecionados são nível, semestre de ocorrência, número de identificação, código do componente curricular, nome do componente curricular e carga horária.
- **Unidades**: dados das unidades acadêmicas da UFMA, contendo os atributos nome, sigla, endereço URL no SIGAA, descrição da unidade, nome e código SIAPE do diretor.

[1] `https://sigaa.ufma.br/sigaa/public/home.jsf`

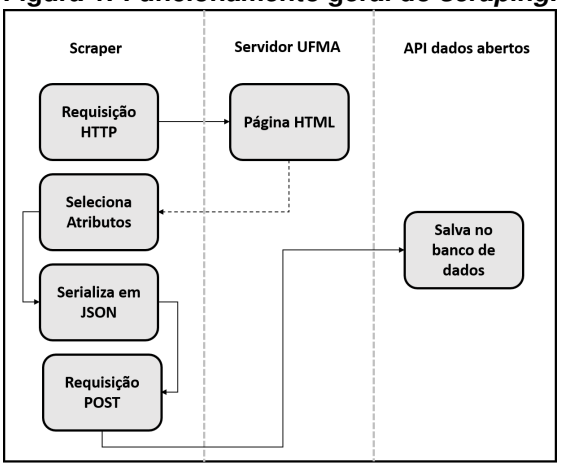
- **Subunidades**: este conjunto engloba dados das subunidades da UFMA, comumente departamentos e coordenações acadêmicas. Os atributos extraídos são nome, localidade, programa e número da subunidade.
- **Monografias**: engloba metadados das monografias de alunos, tais quais título da monografia, nome do aluno, nome do orientador, nome do curso, ano e data da defesa.

3.2. Definição de arquiteturas e formatos

REST foi a arquitetura escolhida para disponibilização dos dados, e JSON o seu formato de disponibilização. Os dados foram replicados e armazenados em um banco de dados não relacional, o MongoDB. Esse sistema de banco de dados é orientado a documentos e tem como vantagem não requerer a definição prévia de um esquema de dados.

Quanto a obtenção dos dados, desenvolveu-se técnicas de *web scraping* para primeiramente extraírem os atributos selecionados na seção 3.1 e serializá-los em JSON, formato escolhido como padrão para a representação das informações. Quanto a forma como as páginas HTML são obtidas dos servidores da UFMA, constatou-se que elas podem ser obtidas de duas formas principais: a partir de requisições GET e POST.

Figura 1. Funcionamento geral do *scraping*.



A figura 1 dá uma visão geral da estratégia para obtenção dos dados dos servidores da UFMA. Essencialmente, um algoritmo *scraper* realiza uma requisição HTTP GET ou POST (a depender da investigação anterior) que retornará uma página HTML, sendo que nesta o algoritmo procura pelas tags HTML através de seletores únicos (CSS ou XPath) para que assim seja feita a extração dos valores dos atributos. Após, os atributos e seus valores são serializados em JSON e armazenados temporariamente no computador local em um arquivo de mesmo formato. Ao final, uma requisição POST é realizada para a API de dados abertos, onde os dados salvos em JSON são enviados e armazenados no banco de dados não relacional (MongoDB).

3.3. Disponibilização de dados

A consulta das informações é realizada a partir de ações GET HTTP na página advindas de qualquer fonte cliente. A partir da URL raiz, `https://dados-abertos-ufma.herokuapp.com`, os conjuntos de dados podem ser consultados da forma `/api/v01/`

`nome-do-conjunto-dados/?params` onde `params` são os atributos buscados via requisição GET, cujos nomes estão apresentados na tabela 1.

Um exemplo de consulta é mostrado na figura 2, onde é feita uma requisição GET solicitando todos o dados de subunidades. A especificação de atributos para o refinamento da busca pode ser feita através de *query string* (*string* de consulta) [Masinter 1998], como exemplificado na figura 3, onde um ou mais atributos são passados na própria URL seguindo o esquema padronizado de *string* de consulta.

A tabela 1 mostra as URLs para recuperação dos conjuntos de dados extraídos e disponíveis em formato aberto, assim como a quantidade de recursos extraídos.

Figura 2. Esquematização do processo de consulta à API de dados abertos, enfatizando o papel do cliente e da API quanto as suas formas de requisição e resposta, respectivamente, para o conjunto de subunidades acadêmicas.

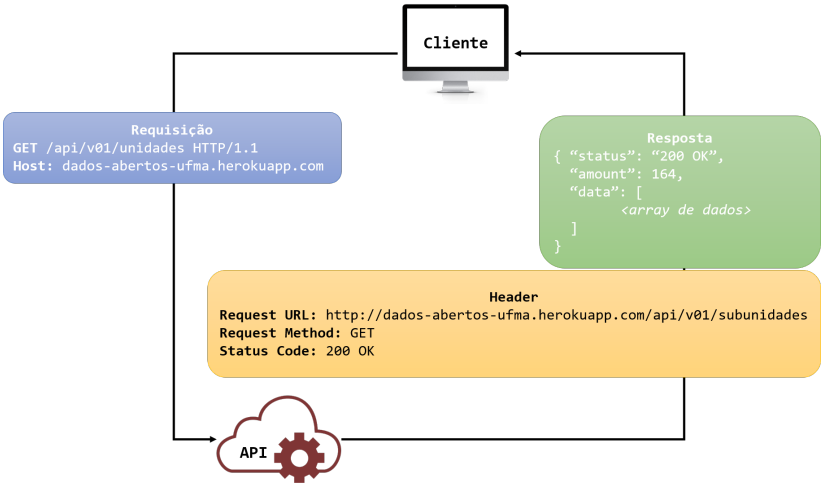
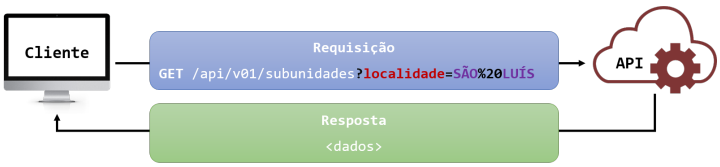


Figura 3. Em vermelho, atributo seletivo, e em roxo, *string* de consulta "São Luís".



4. Resultados

Neste trabalho, argumentou-se que a disponibilização dos dados da UFMA em formato aberto e estruturado, favorecerá o acesso a informações e no desenvolvimento de soluções e aplicativos. Por exemplo, a API desenvolvida já foi utilizada no projeto de [Oliveira and Costa 2016].

O sistema SIGRH da UFMA disponibiliza uma série de relatórios em formato PDF. Contudo, o agente reutilizador pode precisar de algum relatório específico que não foi previsto durante o desenvolvimento do portal. Em outros casos, as informações para gerar estes relatórios, gráficos ou tabelas já se encontram disponibilizados no portal, contudo, acessá-las iria requer uma cansativa busca no SIGAA público envolvendo a seleção

Tabela 1. Conjunto de dados por seus respectivos atributos, e quantidade total de documentos extraídos.

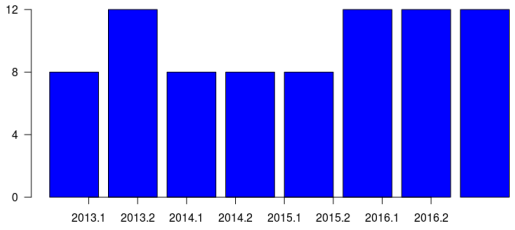
URL	Atributos	Total
/discentes	nome, matricula, curso_id	23.981
/docentes	siape, nome, descricao, subunidade, codigo_subunidade, areas_de_interesse, formacao_academica, url_sigaa, url_lattes	1.752
/turmas	nivel, semestre, codigo, nome, carga_horaria*, siape_docente, tid	52.157
/cursos	_id, area_cnpq, convenio_academico, coordenador*, descricao, modalidade*, nome*, titulo_profissional, url_sigaa	239
/subunidades	codigo, nome, localidade, programa	164
/unidades	nome, sigla, url_sigaa, descricao, diretor, siape_diretor	7
/monografias	ano, data, discente, orientador, titulo, curso_id	7.990

manual das informações procuradas, além da anotação de entrada por entrada em algum editor de planilha, para assim então gerar gráficos ou algum outro tipo de representação. Essa tarefa custaria um longo tempo de trabalho e seria repetida para cada nova busca. Por exemplo, as seguintes perguntas não seriam fáceis de serem respondidas sem o engajamento das informações da UFMA:

1. Qual foi a evolução da carga horária semanal de um dado docente?
2. Qual é a distribuição de carga horária entre os docentes de um dado curso?

Na API desenvolvida, foram extraídas os dados das turmas, que incluem a matrícula do docente e o semestre de oferta. Deste modo, pode-se escrever um algoritmo que recupera as turmas de um dado docente em um dado semestre. Sendo assim, foi escrito um algoritmo que retorna um gráfico de barras com a carga horária de um dado professor, escolhido aleatoriamente, em cada semestre. O gráfico é apresentado na figura 4. Pelo gráfico, observamos que nos últimos oito semestres este dado professor ficou com 8 horas semanais no primeiro semestre de 2013, em ambos os semestres de 2014, e no primeiro semestre de 2015. Nos semestres restantes, em sua carga horária houve um aumento de 12 horas por semana.

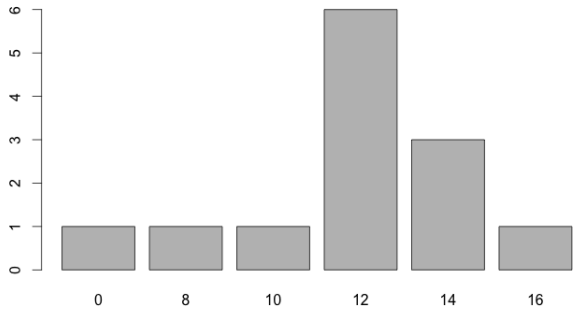
Figura 4. Carga horária semanal de um dado docente entre 2013.1 e 2016.2.



Para responder a última pergunta proposta neste capítulo, pode-se usar o conjunto de dados de docentes e de turmas. Dado todos os docentes de uma dada subunidade, é

possível verificar a quantidade de carga horária e gerar um gráfico que apresenta a quantidade de docentes com uma dada carga horária. Na figura 5 é apresentada a distribuição da carga horária entre os 13 docentes lotados na Coordenação de Engenharia da Computação no segundo semestre de 2016.

Figura 5. Distribuição da carga horária entre os 13 docentes lotados na Coordenação de Engenharia da Computação.



Na figura 5, observa-se que existe um docente sem disciplinas, devido a um afastamento. A maioria dos docentes estão com 12 horas. Porém, tem professores com 14 e 16 horas semanais.

5. Considerações finais

O portal de dados abertos proposto e implementado neste trabalho para a Universidade Federal do Maranhão possibilitou uma análise computacional rápida e eficiente sobre os dados antes fragmentados, e trouxe conhecimento a cerca do número de alunos ativos na instituição, assim como o efetivo de professores, cursos, turmas e departamentos que formam a universidade como um todo. A estruturação dessas informações e divulgação também é capaz de ampliar a transparência da instituição, pois os dados agora estão acessíveis para toda a sociedade, e permite que a criatividade e inovação por parte da comunidade possa vigorar de forma a trazer benefícios para todos.

Referências

(2011). Lei nº 12.527, de 18 de novembro de 2011.

Brasil (2013). Manual da lei de acesso à informação para estados e municípios.

Ferreira Filho, O. F. (2009). *Serviços semânticos: uma abordagem RESTful.* PhD thesis, Universidade de São Paulo.

Fielding, R. T. (2000). *Architectural styles and the design of network-based software architectures*. PhD thesis, University of California, Irvine.

Leonard, R. and Ruby, S. (2007). *RESTful web services*.

Masinter, L. (1998). The data url scheme.

Oliveira, W. and Costa, S. S. (2016). Fila virtual: Ocultando o tempo de espera em restaurantes universitários. *VI Jornada de Informática do Maranhão*.