Utilização de *Bots* para Obtenção Automática de Dados Públicos usando as Técnicas de *Web Crawling* e *Web Scraping*

Igor Martins Galdino[1], **Erica de Lima Gallindo**[1], **Mário W. L. Moreira**[1]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Ceará (IFCE)
Aracati, CE – Brazil

igormgaldino@gmail.com, {erica.gallindo, mario.wedney}@ifce.edu.br

Abstract. *The Escola Virtual.Gov (EV.G) receives resources from partner institutions to provide a range of courses that are required by them. In order to promote active transparency, and following the Lei de Acesso à Informação, accountability for the application of these resources needs to be available to ordinary people. From this, EV.G manages the application of resources through its system. In this way, the system is fed manually. In this situation, given the need for EV.G, this work proposes to simplify the process of updating the Portal in Numbers, automating the manual feeding activities performed by EV.G today and publishing the information obtained in the portal's data source.*

Resumo. *A Escola Virtual.Gov (EV.G), recebe recursos de instituições parceiras para disponibilizar a oferta de cursos que são demandados por elas. A fins de promover uma transparência ativa e em conformidade com a Lei de Acesso à Informação, a prestação de contas da aplicação desses recursos precisa estar disponível ao cidadão comum. A partir disto, a EV.G gerencia a aplicação dos recursos através de um sistema próprio. Desse modo, a alimentação do sistema é feita de forma manual. Nessa situação, dada a necessidade da EV.G, este trabalho propõe simplificar o processo de atualização do Portal em Números, automatizando as atividades de alimentação manual hoje realizadas pela EV.G e publicando as informações obtidas na fonte de dados do portal.*

1. Introdução

A Lei nº 12.527, de 18 de novembro de 2011, conhecida como Lei de Acesso à Informação (LAI), regulamenta o direito constitucional de acesso às informações públicas. É obrigação dos governos federal, estaduais e municipais garantir o direito de acesso à informação de modo transparente, ágil, claro e com linguagem de fácil compreensão. Para permitir que os cidadãos pudessem solicitar as informações regulamentadas pela LAI, a Controladoria Geral da União (CGU)[1] criou o Sistema Eletrônico de Informação ao Cidadão (e-SIC). Usando o e-SIC, qualquer cidadão pode solicitar informações aos órgãos federais que, por sua vez, terão a obrigação de fornecê-las, excetuando-se casos específicos previstos na LAI. Em 2012, ano de início de funcionamento do e-SIC, foram registrados 55.212 pedidos de acesso à informação em todos os órgãos da administração federal[2]. Em 2019, este número já era 200.725, demonstrando um aumento significativo do interesse da população pelas ações governamentais.

[1]A CGU é um órgão do Governo Federal responsável pela defesa do patrimônio público e pelo incremento da transparência da gestão, através de atividades de controle interno, auditoria pública, correição, prevenção e combate à corrupção e ouvidoria.

[2]Disponível em: https://esic.cgu.gov.br/sistema/Relatorios/Anual/RelatorioAnualPedidos.aspx

Para lidar com a alta demanda de solicitações, os órgãos da administração pública estão aderindo a iniciativas de transparência ativa[3], *i.e.*, iniciativas de divulgação de informações de interesse geral, independentemente de terem sido solicitadas por algum cidadão. O Governo Federal, especificamente os órgãos de controle da administração pública, vêm incentivando a criação de *sites* para o compartilhamento dessas informações ao cidadão, viabilizando mais um meio que permita o combate à corrupção [Santos 2018]. O Portal da Transparência do Governo Federal, lançado em 2004, é o principal exemplo de aplicação do conceito da transparência ativa em *sites* governamentais.

Em 2018, o Governo Federal lançou o novo Portal da Transparência, proporcionando buscas mais intuitivas, painéis gráficos e agregação de redes sociais. Embora permitisse que o cidadão comum obtivesse dados de seu interesse, por ter sido construído como uma ferramenta de propósito geral e com visualizações de dados padronizadas, o portal não atende às necessidades específicas de todos os órgãos da administração pública, que precisam construir outras visualizações dos mesmos dados que estão disponibilizados no referido ambiente. A Escola Nacional de Administração Pública (Enap), é um exemplo de entidade da administração pública com requisitos próprios para disponibilização de informação sobre a execução de recursos orçamentários e financeiros que ela gerencia.

A Enap coordena a Escola Virtual.Gov (EV.G), que oferta cursos a distância para capacitação profissional de servidores públicos de todo país. Diversos órgãos da administração pública utilizam a EV.G como principal mecanismo de capacitação de seus servidores. Com frequência, os cursos ofertados pela EV.G são demandados por instituições parceiras, confeccionados e posteriormente disponibilizados em seu portal. No ano de 2018 foram realizadas cerca de 442.719 inscrições em cursos ofertados na EV.G, tendo este número aumentado para 945.545 em 2019[4], demonstrando o crescente interesse pelo tipo de capacitação ofertada.

Para viabilizar a confecção e o desenvolvimento dos cursos, as instituições parceiras descentralizam recursos à Enap (EV.G), por meio de um Termo de Execução Descentralizada (TED)[5], para execução de um plano de trabalho previamente acordado, permitindo que a instituição parceira acompanhe a aplicação dos recursos.

A Enap (EV.G) gerencia vários TEDs simultâneos e para isso possui uma solução de *software* para armazenar as informações dos principais documentos associados à execução financeira. Estas informações são inseridas manualmente no sistema, através de um servidor da Enap (EV.G). Como a frequência de lançamentos de novos dados é muito alta, o tratamento manual não garante que os registros estejam sempre atualizados. Logo, faz-se necessário um mecanismo para atualização automática dos dados armazenados no sistema da Enap (EV.G), com base nas fontes de dados oficiais existentes.

Neste contexto, o presente trabalho tem como objetivo propor uma solução para automatizar a atualização dos dados de execução das despesa pública, necessários ao monitoramento realizado pela Enap (EV.G), usando para isso informações do Portal da Transparência. Os dados necessários ao monitoramento das ações da Enap (EV.G) serão extraídos automaticamente do referido portal, inseridos nos sistemas de registros existen-

[3]Disponível em: https://www12.senado.leg.br/perguntas-frequentes/perguntas-frequentes/canais-de-atendimento/transparencia-1/o-que-e-transparencia-ativa

[4]Disponível em: https://emnumeros.escolavirtual.gov.br/indicadores/

[5]Cada TED é acompanhado por um conjunto de informações, tais como: plano de trabalho, definição do objeto, propósitos a serem alcançados, etapas e recursos envolvidos, permitindo o acompanhamento da unidade que descentralizou o recurso.

tes na Enap (EV.G), e publicados por meio de *dashboards* disponíveis no *site* da entidade. Isso viabilizará a transparência ativa incentivada pelos órgãos de controle, o controle social e uma visão global da aplicação dos recursos que circulam na Enap (EV.G).

2. Fundamentação Teórica

Nesta seção são conceituados os *bots* que utilizam técnicas para obtenção das informações na *web*. Em seguida, a modalidades de aplicação de recursos federais e os tipos de descentralização dos recursos são detalhados. Finalmente, a execução financeira no âmbito da Enap (EV.G) e o processo existente para proporcionar a transparência ativa no contexto dessa instituição são detalhados.

2.1. Bots da web

O *bot* da *web*, também chamado de *spider*, é um algoritmo usado para analisar e extrair informações dos *websites* de forma sistemática e automatizada [Omari et al. 2016]. Estes *bots* capturam informações das páginas, cadastram os *links* identificados, para que possam ser posteriormente utilizados na localização de novas páginas, além do mais, podem obter os dados contidos nesses sites. Estes *bots*, fazem rastreamento e raspagem de dados, utilizando duas técnicas, a saber: *web crawling* e *web scraping*. Estas técnicas podem ser usadas de forma simultânea ou como duas tarefas distintas [Khalil and Fakir 2017].

2.2. Técnicas para extração de dados na web

Web crawling é o processo responsável por efetuar as buscas das páginas *web* e indexá-las. Esta técnica captura informações dos *websites* e cadastra os links encontrados, de forma que possam ser localizados futuramente [D'Haen et al. 2016]. O processo é iniciado a partir de uma *seed* (semente), que consiste em um ponto de partida para encontrar novos endereços a serem visitados. À medida que o *bot* visita estes endereços, os *hyperlinks* são identificados e adicionados à lista de endereços a serem visitados [Khalil and Fakir 2017], proporcionando a localização de outras páginas, mantendo assim o banco de dados atualizado.

O *web scraping* é o processo de extração utilizado para coletar dados relevantes de *sites* de forma automática, convertendo as informações desestruturadas em estruturadas, para serem posteriormente analisadas [Zhao 2017], em um procedimento conhecido como raspagem de dados.

O *bot* que utiliza a técnica *web scraping* é programado para efetuar requisições a um servidor *web* a partir de uma lista predefinida de URLs ou retornado pela técnica *web crawling*. Após a solicitação são extraídos os dados necessários. Os dados obtidas são copiados e podem ser exportados em arquivos nos formatos JSON[6], CSV[7], entre outros. Normalmente, este processo simula uma navegação humana na utilização de um *site*, porém, o *bot* consegue efetuar mais requisições do que um ser humano.

2.3. Orçamento Público

A Lei Orçamentária Anual (LOA) é a lei que anualmente estabelece as despesas e as receitas que serão realizadas pelo governo no próximo ano. Por meio da LOA é definido o recurso financeiro total que o Governo Federal espera arrecadar em receitas

[6]JSON (*JavaScript* Object Notation - Notação de Objetos *JavaScript*) é um formato leve de troca de dados. É de fácil compreensão para leigos.

[7]CSV (*Comma separated value* - valores separados por vírgulas) é um tipo de arquivo de texto, fundamental para transferência de informações entre aplicativos diferentes.

e fixado o valor máximo de despesas que podem ser efetuadas com dinheiro público [Arruda and Araújo 2017]. Este orçamento ajuda na transparência das contas públicas permitindo que todo cidadão fiscalize e acompanhe a aplicação dos recursos.

2.4. Modalidades de aplicação de recursos federais

Uma modalidade de aplicação classifica a despesa e indica a aplicação do recurso diretamente pelas entidades de mesmo nível de governo ou indiretamente por meio de transferências para outros órgãos[8].

A descentralização de créditos é a forma utilizada na Administração Pública Federal para se transferir o poder do crédito orçamentário de uma unidade gestora (UG) para outra do mesmo órgão ou de órgão distinto. Desta forma, a descentralização pode ser classificada como **externa** quando a movimentação dos créditos orçamentários é feita entre órgãos diferentes e **interna** quando a movimentação de créditos ocorre entre UGs de um mesmo órgão.

Conforme mencionado anteriormente, o contexto deste trabalho é na transparência ativa da execução financeira de recursos no âmbito da Enap (EV.G), que recebe recursos de órgãos para disponibilização de cursos e estabelece parcerias com outros órgãos para a realização de ações de pesquisas e desenvolvimento, sendo ambos tipos de órgãos chamados de instituições parceiras.

O repasse de recursos entre as instituições parceiras e a Enap é formalizado por meio de um TED, definido no Decreto nº 8.180, de 30 de dezembro de 2013, como "o instrumento por meio do qual é ajustada a descentralização de crédito entre órgãos e/ou entidades integrantes, com objetivo executar ações de interesse da unidade orçamentária descentralizadora". Um TED contém uma descrição abrangente do objeto a ser executado a partir dos recursos descentralizados, das metas a serem alcançadas, das etapas e dos recursos envolvidos.

Após a assinatura do TED, a Enap é autorizada então a descentralizar o crédito previsto para a instituição parceira, sendo esta operação de descentralização formalizada por meio de um documento denominado **Nota de Crédito ou Nota de Movimentação de Crédito** (NC)[9].

2.5. A Escola Virtual.Gov

A EV.G é uma iniciativa coordenada pela Enap para centralizar a oferta de cursos à distância de capacitação profissional a servidores, empregados públicos de todo o país, bem como cidadãos e estrangeiros interessados nos cursos. A EV.G possui um sistema de gestão acadêmica próprio, catálogo de cursos, base de dados de cursos, alunos e capacitações, serviços de atendimento ao usuário de primeiro nível, entre outros. Os cursos ofertados no âmbito da EV.G são oriundos das instituições associadas.

A Enap (EV.G) pode desempenhar dois papéis distintos, a depender do mecanismo de descentralização de recursos usados. Quando a Enap (EV.G) recebe recursos de outros órgão para a realização de cursos em seu ambiente, ela está exercendo o papel de **entidade**

[8]Manual do SIAFI - Classificações orçamentárias está disponível em: https://conteudo.tesouro.gov.br/manuais/index.php?option=com_content&view=article&id=1567:020332-classificacoes-orcamentarias&catid=749&Itemid=376

[9]A NC é o documento utilizado para registrar eventos vinculados a movimentação interna e externa de créditos para execução da despesa pública, no qual se descreve os tipos de despesas que podem ser realizadas com aquele recurso e seus respectivos valores máximos.

descentralizada. Por outro lado, quando a Enap (EV.G) repassa recursos para que outro órgão desenvolva pesquisas para ela, o papel exercido é de **unidade descentralizadora**.

O processo de transferências e aplicação de recursos é feito por meio de **NC**. Para que possa disponibilizar a transparência ativa na aplicação os recursos, ela precisa ter, sob sua gestão, todos os demais documentos de formalização da despesa emitidos pelas entidades parceiras. Inevitavelmente, tem-se um grande volume destes documentos financeiros, tornando-se necessário existir uma gestão automatizada de todos eles, tenham eles sido emitidos pela própria Enap (EV.G) ou pelas entidades parceiras.

3. Proposta

Este trabalho tem como objetivo simplificar o processo de atualização do Portal em Números, automatizando as atividades de alimentação manual hoje realizadas pela Enap (EV.G) e publicando as informações que foram obtidas na fonte de dados do portal. Com a inserção da proposta apresentada neste estudo, o processo será alterado essencialmente na parte de obtenção e registro dos dados no sistema próprio da Enap (EV.G), que será automatizado na solução proposta. Antes do início do trabalho, foi solicitada uma autorização para a realização desta pesquisa junto à Enap, sendo devidamente aprovada através de um termo de autorização.

A solução proposta atua principalmente em duas partes distintas: na extração de dados do Portal da Transparência e em sua inserção no sistema de armazenamento da EV.G. Nesta proposta, o rastreio dos documentos é realizado por meio de *bots* da *web* que buscam as informações de interesse, oriundas do Portal da Transparência, extraindo e estruturando os conteúdos obtidos em seguida. Os documentos obtidos, por sua vez, são armazenados na nuvem e publicados no Portal em Números.

O processo de extração de dados, apresentado na Figura 1, é realizado através de *bots*, utilizando as técnicas de *web crawling* e de *web scraping*, que fazem a rastreabilidade e a obtenção dos dados, respectivamente.

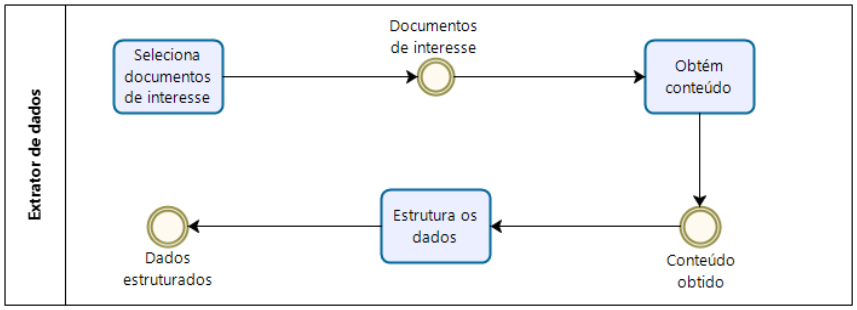

Figura 1. Processo de extração de dados.

Os *bots* fazem a busca dos registros no Portal da Transparência e selecionam quais são os documentos de interesse. Posteriormente, eles realizam a varredura dos dados, obtendo os conteúdos necessários. Para que o servidor não seja sobrecarregado, os *bots* foram desenvolvidos incluindo um tempo de espera entre uma requisição e outra. Após a obtenção dos dados, resultantes de vários arquivos, são estruturados em apenas um único arquivo CSV[10].

Com a obtenção dos dados estruturados, realizado pelo processo anterior, segue-se com o armazenamento dessas informações. O funcionário da Enap (EV.G) é o usuário

[10]Os dados estruturados em arquivo CSV foram solicitados pela própria Enap (EV.G).

responsável pelo processo de inserção de dados ilustrado na Figura 2. Em resumo, este processo se inicia com a aquisição dos dados por meio da utilização de uma aplicação *web* proposta. Os dados coletados por meio da aplacação são posteriormente avaliados pelo usuário público-alvo da aplicação (funcionário da Enap). Após os dados serem validados, o usuário obtém os arquivos que são importados no sistema de gerenciamento da Enap (EV.G), que é um *Redmine* customizado para àquela instituição.

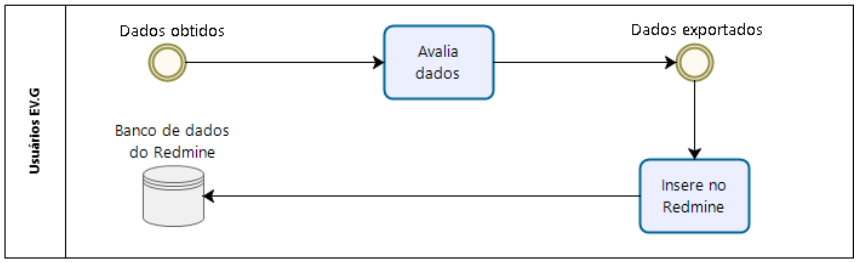

Figura 2. Processo de armazenamento dos dados.

A Enap (EV.G) utiliza o *Redmine* como sistema para o gerenciamento dos documentos financeiros. Atualmente, os documentos financeiros são cadastrados um a um na ferramenta pelo funcionário responsável após a consolidação dos dados por meio de um processo manual de busca e agrupamento de informações. Esta forma de trabalho é lenta e suscetível a erro humano na inserção individualizada de cada item. Por meio do *bot* desenvolvido neste trabalho, este procedimento pode vir a ser rapidamente realizado, dependendo muito mais da velocidade de acesso à *internet* do que da agilidade do usuário que está responsável pela ação.

Com os documentos armazenados no banco de dados do *Redmine*, a próxima etapa é a publicação dos dados em uma ambiente de armazenamento na nuvem. Este passo consiste em uma aplicação, na qual é executada diariamente com a finalidade de buscar os documentos que estão salvos no *Redmine* e armazená-los no *Google Drive* para carga de dados no Portal em Números.

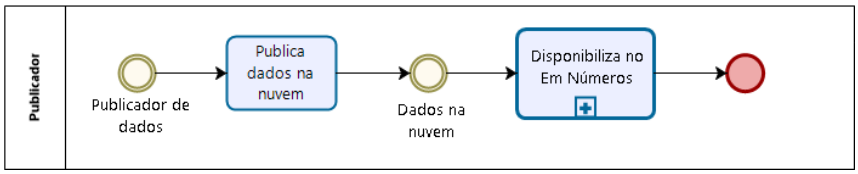

Figura 3. Processo de publicação dos dados.

O processo de publicação, demonstrado na Figura 3, se inicia com uma aplicação que obtém os documentos necessários oriundos do *Redmine*. Em seguida, essa aplicação passa pelo procedimento de conexão com o serviço do *Google Cloud Plataform* (GCP) para conectar ao *Google Drive*. Após ser autenticada, a aplicação armazena os dados na nuvem. Posteriormente, a base de dados no *Google Drive* é utilizado como fonte de dados do Portal em Números em um processo de atualização diária e sem intervenção humana.

4. Resultados

A partir da utilização da solução proposta, a Enap (EV.G) não precisará mais buscar um a um, em um processo manual e lento, os documentos das despesas disponíveis no Portal da Transparência. Então, a partir da aplicação da solução proposta neste trabalho, os

documentos necessários foram obtidos de forma simplificada e ágil do Portal da Transparência, de forma que a base de dados armazenada no banco do *Redmine* esteja sempre consistente e atualizada em relação à execução dos recursos públicos envolvidos.

Com o desenvolvimento da aplicação web, desenvolvida especificamente para este fim, foi possível obter-se dois resultados, diferenciados pelo tipo do documento. A busca das informações de um pagamento individual referente a vários favorecidos, no qual, é realizado pela Ordem Bancária (OB), e os dados de vários pagamentos de um determinado empenho.

Download

Nº do documento	Data	Descrição	Fase	Tipo de documento	Valor do documento	Observação do documento	Favorecido Final	CPF / CNPJ	Valor do favorecido
2018OB802002	04/10/2018	Ordem Bancária (OB)	Pagamento	OBB PARA MESMO BANCO/AGENCIA	R$ 146.600,00	PAGAMENTO DE FOLHA 9/2018-172, AUXILIO PARA PESQUISA, SEI 23106.114365/2018-75, GEPRO_TED_ENAP_ESCOLAVIRTUAL_2016.	[illegible]	***.056.471-**	950,00
2018OB802002	04/10/2018	Ordem Bancária (OB)	Pagamento	OBB PARA MESMO BANCO/AGENCIA	R$ 146.600,00	PAGAMENTO DE FOLHA 9/2018-172, AUXILIO PARA PESQUISA, SEI 23106.114365/2018-75, GEPRO_TED_ENAP_ESCOLAVIRTUAL_2016.	[illegible]	***.550.311-**	950,00
2018OB802002	04/10/2018	Ordem Bancária (OB)	Pagamento	OBB PARA MESMO BANCO/AGENCIA	R$ 146.600,00	PAGAMENTO DE FOLHA 9/2018-172, AUXILIO PARA PESQUISA, SEI 23106.114365/2018-75, GEPRO_TED_ENAP_ESCOLAVIRTUAL_2016.	[illegible]	***.453.161-**	1.000,00
2018OB802002	04/10/2018	Ordem Bancária (OB)	Pagamento	OBB PARA MESMO BANCO/AGENCIA	R$ 146.600,00	PAGAMENTO DE FOLHA 9/2018-172, AUXILIO PARA PESQUISA, SEI 23106.114365/2018-75, GEPRO_TED_ENAP_ESCOLAVIRTUAL_2016.	[illegible]	***.607.661-**	1.000,00

Figura 4. Aplicação com tabela de informações da OB obtidos pelo *bot*

A consulta realizada por meio do número da OB teve como resultado as informações desse documento em específico e os dados de todos os favorecidos finais desse pagamento. Como é ilustrado na Figura 4, as informações foram adquiridas pelo bot, passaram pelo procedimento de estruturação e foram exibidas em formato de tabela, na interface web. Com essa tabela o usuário analisa as informações, verifica e as valida. Com isso, os dados são disponibilizados para download em um arquivo CSV.

A busca pela NE retorna uma quantidade maior de dados do que a busca pela OB. Ao se buscar por uma OB, somente são recuperados os dados de um único objeto, enquanto que ao se buscar uma NE, podem existir diversas ordens bancárias vinculadas a ela. Para os casos analisados neste trabalho, o maior número de OBs associados a uma mesma NE foi 35 (trinta e cinco). Devido ao tempo necessário para a recuperação dos dados, este processo de busca precisou ser dividido em duas partes, uma denominada **busca simples** e outra denominada **busca completa**.

Busca completa

Numero do empenho	2017NE000310
Data do empenho	29/12/2017
Descrição do empenho	Nota de Empenho (NE)
Fase do empenho	Empenho
Especie	ORIGINAL
Unidade orçamentária	47101 - MINIST. DO PLANEJAMENTO, DESENVOLV. E GESTAO
Número	2018OB802406
Data	18/12/2018
Descrição	Ordem Bancária (OB)

(i)

Download

Numero do empenho	2017NE000310
Data do empenho	29/12/2017
Descrição do empenho	Nota de Empenho (NE)
Fase do empenho	Empenho
Especie	ORIGINAL
Unidade orçamentária	47101 - MINIST. DO PLANEJAMENTO, DESENVOLV. E GESTAO
Número	2018OB802465
Data	20/12/2018
Descrição	Ordem Bancária (OB)

(ii)

Figura 5. Resultados de buscas usando Notas de Empenho.

A Figura 5 (i) ilustra os resultados obtidos por meio da **busca simples** que resulta nas informações da NE com os dados da primeira OB a ela vinculada. A busca simples retorna as informações de uma NE, agrupadas com a primeira OB associada a ela. Na Figura 5 (ii) é possível visualizar as informações que são recuperadas por meio de uma **busca completa**, que retorna as informações da NE com os dados de todas as OBs a ela vinculadas. Desta forma, o usuário poderá realizar o *download* de todas essas informações em um arquivo CSV.

5. Considerações Finais

Com a aplicação da proposta, a Enap (EV.G) não precisa mais buscar manualmente os dados com informações sobre a aplicação dos recursos públicos executados pelas suas entidades parceiras, minimizando o tempo gasto nas execução manual deste processo, que é naturalmente lento. A aplicação, objeto deste trabalho, foi construída para realizar a requisição desses dados, utilizando *scripts* que sincronizam diariamente os dados contidos no sistema de armazenamento da Enap (EV.G) para o *Google Drive*. Os dados disponibilizados no armazenamento em nuvem do *Google Drive*, por sua vez, são utilizados como fonte para a geração dos *dashboards* que são publicados no Portal em Números.

Com esta proposta de solução, foi possível obter-se os dados de interesse da Enap (EV.G) diretamente do Portal da Transparência, sem a necessidade de uma intermediação manual. Por meio de uma *interface web*, o usuário consegue solicitar os documentos de interesse em um dado momento.

Como trabalho futuro, pode-se avaliar o *redesign* da solução para usar a API do Portal da Transparência, realizando uma comparação do desempenho das consultas nas duas abordagens. Espera-se também, colocar esse serviço em produção no âmbito da Enap para coletar *feedback* dos usuários, a partir de sua utilização.

Referências

Arruda, D. G. and Araújo, I. P. (2017). *Contabilidade pública*. Editora Saraiva, São Paulo, SP.

D'Haen, J., Van den Poel, D., Thorleuchter, D., et al. (2016). Integrating expert knowledge and multilingual web crawling data in a lead qualification system. *Decision Support Systems*, 82:69–78.

Khalil, S. and Fakir, M. (2017). RCrawler: An R package for parallel web crawling and scraping. *SoftwareX*, 6:98–106.

Omari, A., Shoham, S., and Yahav, E. (2016). Cross-supervised synthesis of web-crawlers. In *Proceedings of the 38th International Conference on Software Engineering*, pages 368–379, New York, NY, USA. ACM.

Santos, M. G. (2018). Portal da transparência da cidade de Bananeiras: uma análise segundo parâmetros da lei de acesso à informação e requisitos de usabilidade. Master's thesis, Universidade Estadual da Paraíba (UEPB), João Pessoa, PB.

Zhao, B. (2017). Web scraping. In Schintler, L. and McNeely, C., editors, *Encyclopedia of Big Data*. Springer, Cham.