Coleta de dados para agregação de repositórios digitais: Entidades vinculadas à Secretaria Especial de Cultura do Brasil

Data collection for aggregation of digital repositories: Entities linked to the Special Secretariat of Culture of Brazil

Luis Felipe Rosa de Oliveira
Universidade de Brasília, Brasil.
E-mail: rosa.luis@aluno.unb.br

Dalton Lopes Martins
Universidade de Brasília, Brasil.
E-mail: daltonmartins@unb.br

RESUMO

A proposta deste artigo é explicitar o uso de técnicas de ciência de dados na ciência da informação, através do relato da coleta de dados dos repositórios digitais das entidades vinculadas à Secretaria Especial de Cultura do Brasil, processo que faz parte de um projeto de pesquisa fomentado pela Fundação de Amparo à Pesquisa do Estado de São Paulo, e executado na Universidade de Brasília pelo Laboratório de Inteligência de Redes. A metodologia quantitativa descritiva aplicada, faz alusão ao processo de extração, transformação e carga de dados utilizado na constituição de data warehouses, em que foram desenvolvidos scripts em Python para coleta dos dados. Como resultados são apontados quantos scripts foram necessários, bem como quantos arquivos de armazenamento foram gerados, além da descrição dos dados coletados, denotando uma maior utilização da técnica de raspagem de dados, dificultando o processo de coleta. Assim, o artigo aponta como a atual realidade das instituições culturais brasileiras analisadas está longe de possibilitar a agregação de seus repositórios digitais, mas ao mesmo tempo aponta como estratégias de ciência de dados permitem ao profissional da ciência

Como citar: Oliveira, L. F. R.; & Martins, D. L. (2022). Coleta de dados para agregação de repositórios digitais: Entidades vinculadas à Secretaria Especial de Cultura do Brasil. *Advanced Notes in Information Science, volume 2* (pp. 171-181). Tallinn, Estonia: ColNes Publishing. DOI: 10.47909/anis.978-9916-9760-3-6.106.

da informação, superar as barreiras técnicas existentes, e promover a análise e o reuso de dados.
Palavras-chave: Ciência de dados; Ciência da Informação; Repositórios digitais; Instituições culturais; Reuso.

ABSTRACT

This article aims to clarify the use of data science techniques in information science by reporting data collection from the digital repositories of entities linked to the Special Secretariat of Culture of Brazil. This process is part of a research project promoted by the Foundation of Research Support of the State of São Paulo and carried out at the University of Brasília by the Network Intelligence Laboratory. The quantitative descriptive methodology applied alludes to extracting, transforming, and data load used in the constitution of data warehouses, in which Python scripts were developed to collect the data. The results indicate how many scripts were needed and how many storage files were generated, in addition to the description of the data collected, denoting a greater use of the web scraping technique, making the collection process more difficult. Thus, the article points out how the current reality of the analyzed Brazilian cultural institutions is far from enabling the aggregation of their digital repositories. At the same time, it points out how data science strategies allow information science professionals to overcome existing technical barriers and promote data analysis and reuse.
Keywords: *Data Science; Information science; Digital repositories; Cultural institutions; Reuse.*

INTRODUÇÃO

A digitalização de objetos culturais, e consequente criação de repositórios digitais, é um fator determinante na presença das instituições culturais na internet, que não só dá visibilidade às entidades culturais, como também denota uma grande oportunidade para a produção de pesquisas. Na ciência da informação, por exemplo, os estudos sobre repositórios digitais podem ser encontrados na linha de pesquisa de organização da informação (Bräscher & Monteiro, 2010). Ainda no contexto da ciência da informação alguns estudos utilizam técnicas de ciências de dados, como análise e mineração de dados para promover o reuso dos dados de repositórios digitais (Virkus & Garoufallou, 2019).

E é nessa linha de estudos, que envolvem o uso de técnicas de ciência de dados no contexto da ciência da informação, que o estudo apresentado neste artigo se situa, mais especificamente no relato de uma das etapas do projeto de pesquisa: "Interoperabilidade entre os repositórios digitais do patrimônio cultural brasileiro: da web semântica e dados abertos ligados às ferramentas de busca e recuperação da informação", fomentado

pela Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), e executado na Universidade de Brasília pelo Laboratório de Inteligência de Redes.

O projeto tem como objetivo principal problematizar o desenvolvimento de um serviço de agregação dos repositórios digitais das entidades brasileiras vinculadas à secretaria especial de cultura (antigo Ministério da Cultura). São sete as instituições culturais no escopo deste projeto: Biblioteca Nacional (BN), Fundação Nacional de Artes (Funarte), Fundação Cultural Palmares (FCP), Agência Nacional do Cinema (ANCINE), Instituto do Patrimônio Histórico e Artístico Nacional (IPHAN), Instituto Brasileiro de Museus (IBRAM), e Fundação Casa de Rui Barbosa (FCRB).

Dentro do processo de desenvolvimento do projeto existem 5 etapas: análise da informação dos repositórios digitais; coleta amostral dos dados; análise de modelos conceituais semânticos para interoperabilidade entre os repositórios; coleta dos dados e implementação da base de dados semântica; e desenvolvimento e customização do repositório digital para busca integrada.[1] Neste estudo será abordada a etapa de coleta de dados e implementação da base de dados semântica, mais especificamente o processo de análise e desenvolvimento dos scripts de coleta de dados dos repositórios digitais das entidades culturais no escopo do projeto. Essa etapa de coleta dos dados está diretamente subordinada aos resultados da análise da informação dos repositórios digitais,[2] em que foram definidos os critérios de diagnóstico de repositórios nos sites de cada uma das entidades culturais, além da análise das prováveis técnicas de coleta disponíveis para cada repositório identificado.

Como resultados dessa etapa de análise, em suma, foram denotadas dificuldades de identificar documentação sobre os repositórios, além da prospecção de que a maioria dos repositórios encontrados não apresentaram uma solução efetiva de interoperabilidade para permitir a coleta dos dados de forma estruturada, por exemplo, sistemas com OAI-PMH, ou API habilitados.

Sem informações sobre a documentação utilizada nos repositórios institucionais, além da dificuldade de encontrar repositórios com protocolos de comunicação dos dados, um problema importante é identificado: como estudar um possível serviço de agregação desses repositórios digitais culturais, sem a possibilidade de interoperabilidade entre sistemas através de protocolos de comunicação? Como resposta a esta

[1] Todos os relatórios das etapas já concluídas estão disponíveis em - https://tinyurl.com/relatoriosInteropFAPESP

[2] Relatório da etapa de análise da informação dos repositórios digitais - https://tinyurl.com/realtorio1InteropDiagFAPESP

problemática, que vai de encontro com a realidade atual dos repositórios digitais culturais brasileiros, entende-se que o encontro da ciência de dados com a ciência da informação, é um dos caminhos possíveis. Utilizar técnicas de extração e transformação do contexto da ciência de dados, para entender melhor como estão organizados os dados dos repositórios digitais na perspectiva de um serviço de agregação, se demonstrou uma solução importante para superar a barreira da recuperação de dados nesses repositórios.

Foi aplicada uma metodologia de extração, coleta e carga (ETL) de dados semelhante ao processo utilizado em *Data Warehouses* (DW). Um DW é um sistema de armazenamento de dados agregados de uma ou mais fontes de dados, para posterior análise e tratamento dos dados. Os dados armazenados em um DW passam por um processo de ETL, utilizando ferramentas específicas ou programação, em que os dados das fontes são extraídos, transformados (opcionalmente) e carregados no DW (Ferreira, Miranda, Abelha & Machado, 2010). No caso do estudo aqui relatado, não foi o objetivo propor um sistema de *DataWarehouse*, mas a aplicação das demais etapas de coleta e agregação dos dados foi semelhante, devido à natureza inconsistente dos dados e metadados dos repositórios digitais identificados, por isso a metodologia e os resultados da etapa de coleta aqui descritos estão sob este escopo conceitual.

Vale ainda ressaltar os conceitos das três técnicas identificadas para extração dos dados dos repositórios digitais: através do protocolo de comunicação OAI-PMH *(Open Archives Initiative Protocol for Metadata Harvesting),* através de API (*Application Program Interface)* ou através de raspagem de dados. O OAI-PMH é um protocolo de interoperabilidade entre sistemas baseado na coleta de metadados. Para sua implementação, é necessária uma infraestrutura que envolve os provedores de dados (sistemas de repositório digital) e dos provedores de serviços (sistemas de coleta dos dados dos provedores) (Lagoze, Sompel, Nelson & Warner, 2002).

Nesta pesquisa os dados dos repositórios digitais encontrados com a opção de coleta através do OAI-PMH,[3] foram acessados pela URL *(Uniform Resource Locator)* de acesso através de requisições, e retornados no formato XML. Já a API, é uma interface de programação de aplicações, e neste caso específico as aplicações REST *(Representational State Transfer),* que dentre outras funções, permite a consulta aos dados através do protocolo HTTP *(Hypertext Transfer Protocol),* utilizando verbos de consulta em requisições ao ponto de acesso da API (Pautasso, Zimmermann & Leymann, 2008). Por fim, a raspagem de dados, "é uma técnica para

[3] Uma URL de acesso ao OAI-PMH normalmente em o formato http://<sitedorepositório>/OAI-script.

extrair dados de sites da web e salvá-los em um sistema de arquivos ou banco de dados para posterior recuperação ou análise" (Zhao, 2017). Quanto à coleta dos dados dos repositórios digitais encontrados, a raspagem de dados foi utilizada em todas as situações em que não foi possível realizar a coleta através do OAI-PMH ou de API REST.

OBJETIVOS

O objetivo principal deste artigo é descrever as técnicas de coleta utilizadas para acessar e coletar os dados dos repositórios digitais mapeados a partir dos sites das entidades vinculadas à Secretaria Especial de Cultura. E como objetivos específicos, apontar quantos e quais scripts foram necessários para extrair os dados dos repositórios digitais identificados; descrever os resultados obtidos da extração dos dados, denotando a quantidade de registros coletada e a variabilidade dos metadados identificados; indicar os problemas encontrados no processo de coleta dos dados.

PROCEDIMENTOS METODOLÓGICOS

Esta pesquisa, de cunho quantitativo descritivo, ocorreu através de um processo metodológico que faz alusão a uma atividade comum na ciência de dados, denominado ETL *(Extract, Transform, Load)*. Todo o processo de extração, transformação e coleta dos dados foi executado entre dezembro de 2020 e junho de 2020, realizado utilizando scripts na linguagem de programação Python. Os scripts desenvolvidos estão acessíveis através do GitHub.[4] Vale ressaltar o papel fundamental da biblioteca Pandas,[5] principalmente do *Dataframe,* que é uma estrutura tabular de dados de duas dimensões, cujas funções possibilitam armazenar os dados, permitindo a aplicação de transformações nos dados e exportação para outros formatos de dados.

Na etapa de extração dos dados, foi confirmado para cada fonte de dados qual a técnica mais adequada de coleta de acordo com as características do sistema de exibição dos objetos digitais do repositório. Por exemplo, se a fonte de dados exibia o repositório no formato de páginas web (http://portal.iphan.gov.br/pa/videos), caso corriqueiro no caso do IPHAN, a técnica mais indicada foi a raspagem de dados, já que nenhuma outra opção foi identificada. No caso de repositórios exibidos em sistemas de repositórios digitais (DSpace, Tainacan, Sophia), foi identificado se estes possuíam algum protocolo de comunicação, como OAI-PMH

[4] Scripts desenvolvidos para a coleta de dados - https://github.com/tainacan/data_science/tree/master/FAPESP/scripts_extracao

[5] Pandas - https://pandas.pydata.org/

(http://rubi.casaruibarbosa.gov.br/), ou uma API (https://mhn.acervos.museus.gov.br/reserva-tecnica/), para coleta automática dos dados, caso não encontrada essa opção (http://acervo.bndigital.bn.br/sophia/index.html) também foi aplicada a técnica de raspagem de dados. Dessa maneira a etapa de extração dos dados foi executada por uma de três formas encontradas de coleta dos dados, através do OAI-PMH, através de API, ou através da raspagem de dados.

Para a raspagem de dados, foram utilizados dois tipos de scripts diferentes. Para páginas estáticas, foi utilizada a biblioteca *Beautiful Soup,*[6] que permite interpretar a estrutura de árvore do HTML, e acessar os valores de componentes de texto ou tabelas. Já para repositórios digitais em que não foi encontrada outra forma de coleta dos dados, foi necessário aplicar a biblioteca *Selenium,*[7] que permite automatizar o navegador, e acessar os dados através de facetas, caixas de busca, e atravessar mais facilmente *iframes* em *JavaScript*. Para a coleta dos repositórios digitais que dispunham do protocolo OAI-PMH, foi utilizada a biblioteca *Sickle*[8] para fazer conexão com o ponto de acesso, e permitir realizar a coleta de acordo com os verbos de requisição.[9] Como os dados foram retornados no formato XML, foi utilizada a biblioteca *ElementTree,*[10] para interpretar os dados.

Por fim, para a coleta através de API, foi utilizada a biblioteca *requests,*[11] para fazer a requisição a partir do ponto de acesso, e assim obter os dados em JSON. No caso de redes sociais, como o *Flickr* da FCP, foi utilizada a biblioteca *flickr_api,*[12] que permitiu coletar os dados através de métodos específicos da rede social de fotografias. A etapa de transformação não contemplou técnicas de limpeza ou normalização dos dados, já que foi previsto no projeto manter o valor original dos dados. Assim, a transformação se limitou ao formato dos dados. No caso dos dados obtidos através do OAI-PMH, que geralmente são originalmente no formato XML, no caso das APIs que são geralmente no formato JSON, e no caso de dados obtidos pela raspagem de dados, que normalmente são apresentados através do HTML. De todos esses formatos, os dados foram transformados em dicionários e/ou *Dataframes* no contexto dos *scripts* em Python.

[6] Beautiful Soup - https://www.crummy.com/software/BeautifulSoup/bs4/doc/

[7] Selenium - https://selenium-python.readthedocs.io/index.html

[8] Sickle - https://sickle.readthedocs.io/en/latest/#

[9] Verbos de requisição OAI-PMH - http://www.openarchives.org/OAI/openarchivesprotocol.html#ProtocolMessages

[10] ElementTree - https://docs.python.org/3/library/xml.etree.elementtree.html

[11] Requests - https://docs.python-requests.org/en/master/

[12] Flickrapi- https://stuvel.eu/software/flickrapi/

Por fim, na etapa de carga, utilizando a as possibilidades de exportação da biblioteca Pandas,[13] todos os dados foram carregados para múltiplos arquivos no formato tabular CSV, cada arquivo representando os resultados de coleta de uma fonte de informação. Não necessariamente os dados têm que ser mapeados para o CSV, eles poderiam facilmente ser carregados em um banco de dados relacional como MySQL por exemplo. Por conveniência foi escolhido o formato CSV que posteriormente poderia ser transformado e carregado em outra opção de armazenamento com certa facilidade.

RESULTADOS

Os resultados do processo de coleta de dados dos acervos serão apresentados em três instâncias: quanto aos scripts de extração desenvolvidos; quanto aos arquivos de dados gerados, registros coletados e metadados obtidos; e quanto aos problemas encontrados no processo de coleta.

Scripts de extração

Foram desenvolvidos ao todo 23 scripts de extração de dados de 41 repositórios digitais encontrados para coleta. A maioria (86,95%) dos scripts aplica a técnica de raspagem de dados, somente três deles utilizam técnicas diferentes: os scripts de extração de dados via API para os repositórios disponíveis através do *Tainacan,* e pelo *Flickr;* e o script de coleta através do OAI-PMH para o repositório RUBI.

Como a raspagem de dados é uma técnica que prevê a coleta dos dados através do HTML das páginas web dos repositórios, foi preciso adaptar um script específico para cada situação, por exemplo, no IPHAN, os objetos digitais exibidos através das páginas de "regiões", "galerias", "vídeos", etc., demandaram um script específico para cada tipo de exibição. Salvo a exceção dos repositórios disponíveis através do SophiA Biblioteca, utilizados pela Funarte e pela Biblioteca Nacional, em que foi possível compartilhar o mesmo script para as duas situações. Já para o script de coleta via API dos repositórios publicados através do Tainacan, foi possível coletar dados de acervos do IBRAM e Funarte por exemplo, sem alteração direta no código, somente modificando os parâmetros de acesso à API.

[13] Funções de leitura/escrita de Dataframes no Pandas - https://pandas.pydata.org/pandas-docs/stable/user_guide/io.html

Dados coletados

Como resultado dos scripts de coleta, foram gerados 41 arquivos CSV para armazenamento dos dados de cada repositório. Foram coletados ao todo 396.557 registros, sendo a Biblioteca Nacional e entidade com mais registros coletados (185.427), e a Fundação Cultural Palmares aquela em que menos foram encontrados registros (4.191). Não foram coletados dados para a ANCINE, pois na etapa de diagnóstico não foram identificados repositórios digitais para a instituição. Vale ressaltar que não foi coletada a totalidade de registros identificados para todos os repositórios digitais. No caso dos repositórios da Biblioteca Nacional, em que foi prevista a existência de mais de 2 milhões de registros, como os dados foram coletados através de raspagem de dados, devido a onerosidade (lentidão e consumo de hardware) do processo de coleta, foi definida uma data de encerramento da coleta em junho de 2021. Além dos registros coletados, foi calculada a frequência de metadados encontrados, que revelou 561 metadados diferentes para estes 41 repositórios digitais coletados, apontando uma alta variabilidade de elementos descritivos entre os acervos.

PROBLEMAS ENCONTRADOS

Os problemas encontrados no processo de desenvolvimento dos scripts estão relacionados com as técnicas de coleta utilizadas, a raspagem de dados e a coleta por API apresentaram certas barreiras que dificultaram o processo de extração. A raspagem de dados foi a técnica de coleta que mais apresentou problemas no desenvolvimento dos scripts. A principal barreira foi a variabilidade de conteúdo, cada repositório apresentou um contexto de exibição diferente de seus dados através de páginas em HTML, isso gerou lentidão no processo de desenvolvimento dos scripts, já que foi o principal método de extração utilizado.

Outra questão importante foi a execução dos scripts de raspagem de dados, como o processo de coleta leva em conta a consulta às páginas da web e interpretação da estrutura HTML, a extração dos dados através desta técnica sofreu com os problemas de disponibilidade da rede, ora por parte do servidor do repositório, ora por parte do computador local onde o script estava sendo executado. Já quanto aos scripts de coleta através de API, a maior barreira foi a interpretação dos métodos de acesso aos dados disponíveis na API. Cada sistema utiliza um conjunto diferente de métodos para acesso às funcionalidades da API, e isso exige um período de aprendizagem para entender como coletar os dados de maneira mais efetiva.

A extração de dados através do OAI-PMH foi a única técnica que não apresentou maiores problemas, pois a forma de acesso através dos verbos de requisição é simples, e direcionada especificamente para coleta dos registros do repositório.

CONSIDERAÇÕES FINAIS

Como foi possível identificar nos resultados apresentados, a realidade de agregação dos repositórios digitais das entidades culturais brasileiras ainda está bastante distante de uma possibilidade prática. Desde os problemas da falta de documentação explícita sobre os acervos, identificados no projeto de pesquisa, até os problemas de efetivação da coleta e armazenamento dos dados apontados neste estudo, indicam as dificuldades de se promover pesquisas e outras técnicas de reuso dos dados de maneira agregada sobre a cultura brasileira.

Mesmo assim, com essa discrepância na disponibilização de dados dos acervos culturais públicos, o relato aqui apresentado, aponta como técnicas oriundas do contexto da ciência de dados podem auxiliar a explicitar, de maneira prática, as barreiras e possibilidades de reuso dos dados dos repositórios. O conjunto de técnicas utilizadas para extração de dados não é novidade em aplicações de recuperação de dados, a raspagem de páginas web, por exemplo, é amplamente utilizada em situações em que não existe outra maneira de se obter os dados na internet. Porém, neste trabalho, a forma e o contexto em que essas técnicas foram utilizadas reforçam sua importância. Sem a aplicação dessas formas de extração de dados, não seria possível executar as demais etapas do projeto de maneira ampla, ao se tratar dos metadados e dados obtidos.

Com os dados coletados, foi possível levantar os metadados utilizados para descrever os objetos nos repositórios digitais das instituições culturais. De posse desses dados, foi possível realizar uma análise conceitual de quais metadados são compartilhados entres os repositórios e assim promover insumos para a proposta de um padrão de metadados para a agregação conceitual dos dados. Elencando, por exemplo, a afirmação de Virkus e Garaoufallou (2020), que apontam a curadoria dos metadados através de técnicas de ciência de dados, como uma responsabilidade dos cientistas da informação. Dessa forma, como apontamento final deste artigo, reitera-se o potencial da aplicação de técnicas de ciência de dados em estudos da ciência da informação, tanto na análise de dados de pesquisa, quanto em formas de reuso dos dados.

CONFLITOS DE INTERESSE

Os autores declaram não haver conflito de interesses.

DECLARAÇÃO DE CONTRIBUIÇÃO

Curadoria de Dados, Investigação, Análise Formal, Software, Recursos, Redação- rascunho original: Oliveira, L.F.R

Conceituação, Administração do Projeto, Supervisão, Validação, Redação - revisão e edição: Martins, D.L.

DECLARAÇÃO DE CONSENTIMENTO DE DADOS

Os dados coletados/extraídos, gerados durante o desenvolvimento deste estudo foi depositado no Tainacan e está acessível em: https://culturabr.tainacan.org/agregador-fapesp. Os scripts de coleta/extração de dados, gerados durante o desenvolvimento deste estudo foi depositado no Tainacan e está acessível em: https://github.com/tainacan/data_science/tree/master/FAPESP

FINANCIAMENTO

Esta pesquisa foi apoiada por: Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), e Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES).

REFERÊNCIAS

Bräscher, M., & Monteiro, F. de S. (2010). Organização da informação em repositórios digitais. *Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, 15*(29).

Ferreira, J., Miranda, M., Abelha, A., & Machado, J. (2010, September). O processo etl em sistemas data warehouse. In *INForum* (pp. 757-765).

Lagoze, C., Van de Sompel, H., Nelson, M., & Warner, S. (2002). *Open archives initiative-protocol for metadata harvesting-v. 2.0*. Recuperado de http://www.openarchives.org/OAI/2.0/openarchives-protocol.htm

Pautasso, C., Zimmermann, O., & Leymann, F. (2008). Restful web services vs. "Big" web services: making the right architectural decision. In *Proceedings of the 17th international conference on World Wide Web,* 805-814. doi: https://doi.org/10.1145/1367497.1367606

Virkus, S., & Garoufallou, E. (2019). Data science from a library and information science perspective. *Data Technologies and Applications,* 422-441. doi: https://doi.org/10.1108/DTA-05-2019-0076

Virkus, S., & Garoufallou, E. (2020). Data science and its relationship to library and information science: a content analysis. *Data Technologies and Applications,* 643-663. doi: https://doi.org/10.1108/DTA-07-2020-0167

Zhao, B. (2017). Web scraping. *Encyclopedia of big data,* 1-3. Recuperado de https://www.researchgate.net/profile/Bo-Zhao-3/publication/317177787_Web_Scraping/links/5c293f85a6fdccfc7073192f/Web-Scraping.pdf