Métodos digitais e a memória acessada por APIs: desenvolvimento de ferramenta para extração de dados de portais jornalísticos a partir da WayBack Machine

Marcio Carneiro dos Santos

ISSN nº 2447-4266 Vol. 1, nº 2, Setembro-dezembro. 2015

DOI: http://dx.doi.org/10.20873/uft.2447-4266.2015v1n2p23

Métodos digitais e a memória acessada por APIs: desenvolvimento de ferramenta para extração de dados de portais jornalísticos a partir da *WayBack Machine*

Digital methods and the memory accessed by APIs: Development tool for extracting data from journalistic portals with the WayBack Machine

Métodos digitales y memoria visitada por APIs:Herramienta de desarrollo para extraer datos de los portales periodísticos por la Wayback Machine

Marcio Carneiro dos Santos[1, 2]

RESUMO

Explora-se a possibilidade de automação da coleta de dados em sites, a partir da aplicação de código construído em linguagem de programação Python, utilizando a sintaxe específica do HTML (*HiperText Markup Language*) para localizar e extrair elementos de interesse como links, texto e imagens. A coleta automatizada de dados, também conhecida como raspagem (*scraping*) é um recurso cada vez mais comum no jornalismo. A partir do acesso ao repositório digital do site www.web.archive.org, também conhecido como *WayBackMachine*, desenvolvemos a prova de conceito de um algoritmo capaz de recuperar, listar e oferecer ferramentas básicas de análise sobre dados coletados a partir das diversas versões de portais jornalísticos ao longo do tempo.

PALAVRAS-CHAVE: Raspagem de dados. Python. Jornalismo Digital. HTML. Memória.

[1] Professor Adjunto da área de Jornalismo em Redes Digitais do Departamento de Comunicação Social da Universidade Federal do Maranhão. Mestre em Comunicação e Doutor em Tecnologias da Inteligência e Design Digital pela PUC-SP. E-mail: mcszen@gmail.com.

[2] Endereço de contato do autor (por correio): Universidade Federal do Maranhão. Centro de Ciências Sociais – CCSo. Coordenação/Departamento de Comunicação Social. Avenida dos Portugueses, s/n. Campus Universitário do Bacanga. São Luís – MA. Brasil CEP: 65.085-580.

ISSN nº 2447-4266 Vol. 1, nº 2, Setembro-dezembro. 2015

DOI: http://dx.doi.org/10.20873/uft.2447-4266.2015v1n2p23

ABSTRACT

We explore the possibility of automation of data collection from web pages, using the application of customized code built in Python programming language, with specific HTML syntax (Hypertext Markup Language) to locate and extract elements of interest as links, text and images. The automated data collection, also known as scraping is an increasingly common feature in journalism. From the access to the digital repository site www.web.archive.org, also known as WayBackMachine, we develop a proof of concept of an algorithm able to recover, list and offer basic tools of analysis of data collected from the various versions of newspaper portals in time series.

KEYWORDS: Scraping. Python. Digital Journalism. HTML. Memory.

RESUMEN

Se explora la posibilidad de automatización de los sitios de recolección de datos, desde el código de aplicación construida en lenguaje de programación Python, utilizando la sintaxis específica de HTML (Hypertext Markup Language) para localizar y extraer elementos de interés, tales como enlaces, texto e imágenes. La colección de datos automatizada, también conocido como el raspado es una característica cada vez más común en el periodismo. Desde el acceso a la www.web.archive.org, sitio de repositorio digital, también conocida como WayBackMachine, desarrollamos una prueba de concepto de un algoritmo para recuperar, listar y ofrecer herramientas básicas de análisis de los datos recogidos de las diferentes versiones de portales de periódicos en el tiempo.

PALABRAS CLAVE: Raspar datos. Python. Periodismo digital. HTML. Memoria.

Recebido em: 09.09.2015. Aceito em: 01.12.2015. Publicado em 08.12.2015.

ISSN nº 2447-4266 Vol. 1, nº 2, Setembro-dezembro. 2015

DOI: http://dx.doi.org/10.20873/uft.2447-4266.2015v1n2p23

Introdução

Muitos estudos do ciberjornalismo dependem da coleta de dados a partir dos sites e portais objetos de pesquisa. A aplicação de métodos, ferramentas e processos que considerem a ontologia dos objetos digitais, descritos de forma numérica, e as estruturas de rede por onde circulam, utilizando-se de recursos computacionais para sua aplicação, pode em tais casos agregar efetividade e expansão das estratégias de amostragem entre outros benefícios.

Tal abordagem tem sua fundamentação teórica no trabalho de Manovich (2001) a partir da discussão que faz sobre as características dos objetos digitais, especificamente na que denomina de transcodificação.

Para Manovich (2001), os objetos digitais apresentam cinco traços ou características que podem ou não estar presentes simultaneamente em sua existência, a saber: descrição numérica, modularidade, automação, variabilidade e transcodificação.

A descrição numérica indica, como já citamos, que os objetos digitais constituem-se no final das contas de sequencias de números, podendo, por isso, sofrer muitas das transformações que se aplicam a essa categoria, entre elas a possibilidade de replicação idêntica, desde que a nova sequencia mantenha a estrutura e a ordem original da primeira.

A modularidade nos termos de Manovich (2001) descreve os objetos digitais como compostos de partes que podem ser arranjadas de diversas formas, sem que cada um desses módulos perca sua identidade original. Ao visitarmos a página de um site na internet não estamos vendo a imagem de um único elemento completo, mas sim o resultado da construção feita pelo *browser*[3] a partir de diversas partículas de informação; os pequenos arquivos enviados pelo servidor onde o site está

[3] *Browser* é uma categoria de software que age como um cliente de internet solicitando conteúdo aos servidores da rede e organizando os elementos recebidos nas páginas que visitamos em nossa navegação pela *web*.

ISSN nº 2447-4266 Vol. 1, nº 2, Setembro-dezembro. 2015

DOI: http://dx.doi.org/10.20873/uft.2447-4266.2015v1n2p23

hospedado. Esses são agrupados e estruturados pela ordem descrita no código da programação HTML (*HiperText Markup Language*) que define onde e de que jeito cada texto, foto, título, vídeo, ou o que mais a página possua, vão estar.

A partir dessas duas primeiras características, as duas seguintes estabelecem-se como consequências. Se posso aplicar operações ou transformações matemáticas sobre esses objetos e recombiná-los em diversas configurações, porque são compostos de forma modular, posso também programar essas ações e automatizar parte delas, para que sejam realizadas de forma transparente, sem que o usuário sequer perceba o que está acontecendo. A automação permite que, ao apertar a tecla *ENTER* do computador, uma grande quantidade de linhas de código de programação seja executada e algo novo aconteça na tela, sem a necessidade de sermos programadores ou entendermos que processos estão por trás dessa ação.

Para Manovich (2001) as diversas possibilidades de combinação entre esses elementos faz com que eles também reajam de forma diferente a partir de contextos ou situações distintas. A ideia de interatividade seria para o autor uma forma de expressão da variabilidade dos objetos digitais, adaptáveis, programáveis e recombináveis oferecendo aos usuários novas formas de contato e fruição. A não linearidade das narrativas construídas a partir de hiperlinks ou a imersão que um game oferece são bons exemplos do que o autor entende como variabilidade.

Por fim, através do que ele denomina de transcodificação, cada objeto digital é constituído de duas camadas ou *layers*, uma utilizada para carregar o sentido a ser interpretado e processado pelos humanos, a camada da representação ou cultural, que nos oferece o material para que possamos lidar com tal objeto. Entretanto, pela transcodificação, existe ainda uma segunda camada (FIG. 1), que também descreve ou traz informações sobre esse objeto só que para o processamento maquínico, automatizado, o *layer* dos dados estruturados que os computadores entendem e que é usado para fazer esse objeto trafegar pelas redes digitais.

ISSN nº 2447-4266 Vol. 1, nº 2, Setembro-dezembro. 2015

DOI: http://dx.doi.org/10.20873/uft.2447-4266.2015v1n2p23

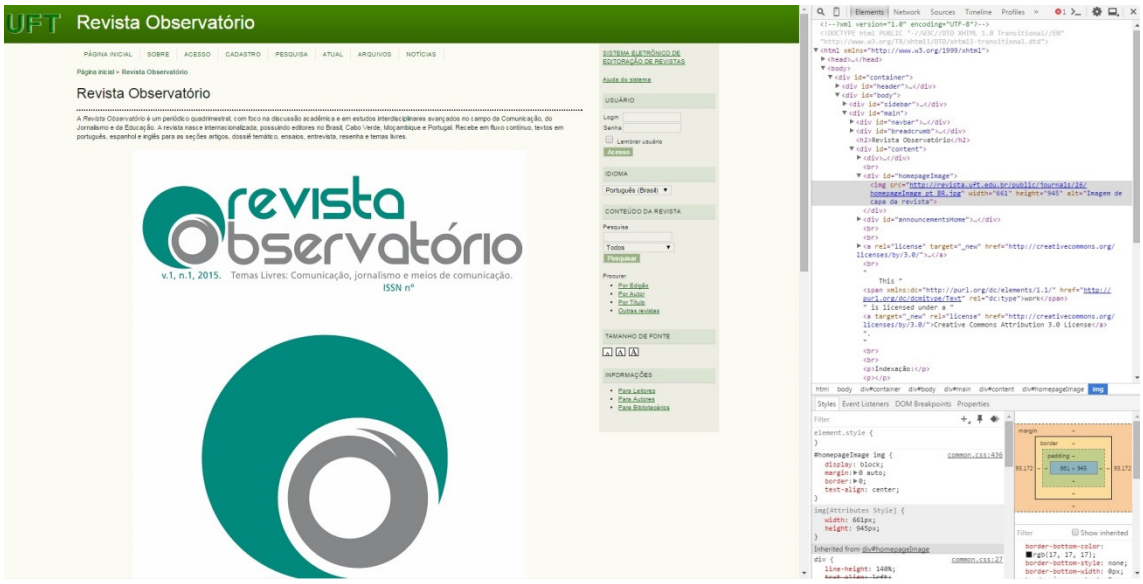

FIGURA 1- Tela do site da revista Observatório com a parte do código HTML explicitada demonstrando os dois *layers* da transcodificação.
FONTE: do autor.

A ideia de métodos do meio (ROGERS, 2013), ou seja, métodos que exploram a lógica interna inerente aos objetos digitais, ou nos termos que estamos propondo, que consideram sua ontologia específica, permitem novas abordagens e formas mais eficientes de enfrentar dificuldades implícitas em algumas temáticas contemporâneas.

> Por exemplo, varredura e extração de dados, inteligência coletiva e classificações baseadas em redes sociais, ainda que de diferentes gêneros e espécies, são todas técnicas baseadas na internet para coleta e organização de dados. Page Rank e algoritmos similares são meios de ordenação e classificação. Nuvens de palavras e outras formas comuns de visualização explicitam relevância e ressonância. Como poderíamos aprender com eles e outros métodos online para reaplica-los? O propósito não seria tanto contribuir para o refinamento e construção de um motor de buscas melhor, uma tarefa que deve ser deixada para a Ciência da Computação e áreas afins. Ao invés disso o propósito seria utiliza-los e entender como eles tratam *hiperlinks, hits, likes, tags, datestamps* e outros objetos nativamente digitais. Pensando nesses mecanismos e nos objetos com os quais eles conseguem lidar, os métodos digitais, como uma prática de pesquisa, contribuem para o desenvolvimento de uma metodologia do próprio meio (ROGERS, 2013).[4]

[4] Tradução do autor.

ISSN nº 2447-4266 Vol. 1, nº 2, Setembro-dezembro. 2015

DOI: http://dx.doi.org/10.20873/uft.2447-4266.2015v1n2p23

A necessidade de iniciativas nessa linha pode ser justificada também por algumas condições verificáveis relacionadas à produção de informação a partir das redes: volume, variedade, velocidade. Não à toa esses termos estão associados a outro conceito contemporâneo, o de *big data*, que de forma simplificada poderia ser definido como o conjunto de métodos, ferramentas e processos destinados a lidar com a verdadeira enxurrada informacional com a qual nos deparamos hoje; tema que Gleick (2013) descreve numa perspectiva histórica e técnica.

São situações assim que exigem a incorporação de métodos que considerem as características inerentes aos objetos digitais, entre elas a transcodificação nos termos de Manovich. Como veremos a seguir, uma alternativa viável para casos onde os dados são gerados e armazenados em plataformas na internet, como o *Twitter* ou a *WayBackMachine* (que utilizaremos nesse trabalho), é o contato direto com os servidores que as sustentam ou, em termos técnicos, a utilização da sua API (*Application Programming Interface*)[5] para realizar consultas e extração de informação a partir do *layer* da máquina.

Explorando a memória digital

Pensar nos sites da internet como representantes contemporâneos dos arquivos que antes apenas podíamos encontrar nas bibliotecas ou locais de memória tradicional é um fato que deve ser considerado como caminho possível para os pesquisadores das Ciências Sociais incluindo os da Comunicação e do Jornalismo.

Muitas pesquisas partem da necessidade de coletar dados sobre objetos que hoje tem suas versões digitais à disposição do acesso via internet. Apesar da

[5] Uma API – *Application Programming Interface* (Interface de Programação de Aplicações) é o conjunto de rotinas, padrões e instruções de programação que permite que os desenvolvedores criem aplicações que possam acessar e interagir com determinado serviço na internet, inclusive extraindo dados dele.

ISSN nº 2447-4266 Vol. 1, nº 2, Setembro-dezembro. 2015

DOI: http://dx.doi.org/10.20873/uft.2447-4266.2015v1n2p23

aparente facilidade para acessar sites é preciso considerar três problemas que se apresentam. Em primeiro lugar a constatação de que a memória digital, apesar de extensa e em constante crescimento, não é eterna e pode ser apagada, a qualquer hora, por decisão do administrador que gerencia o servidor de *web* onde está hospedada. A segunda diz respeito justamente ao fato de que mesmo tendo acesso a esse site, talvez não estejamos coletando toda a informação disponível, olhando apenas para a camada cultural ou da representação e, por isso, tendo uma visão parcial de um todo maior. Por fim a própria coleta pode tornar-se difícil considerando a quantidade de informação disponível e as frequentes mudanças às quais os sites, principalmente os jornalísticos, estão sujeitos.

Nesse cenário a possibilidade de automatização parcial ou completa da fase de coleta de dados em pesquisas da nossa área pode tornar-se um caminho oportuno e que poderá impactar principalmente as decisões sobre as estratégias de amostragem, oferecendo uma relação otimizada entre universo pesquisado e quantidade de elementos considerados na análise (BONACICH; LU, 2012).

A coleta automatizada de dados, também conhecida como raspagem (*scraping*) ou mineração é um recurso cada vez mais comum no jornalismo digital e investigativo (BRADSHAW, 2014) podendo, no caso do trabalho acadêmico, ser utilizada tanto para a execução de rotinas repetitivas, permitindo ao pesquisador mais tempo para as tarefas de maior complexidade, como para identificar padrões e tendências em grandes volumes de informação que, em algumas situações, podem passar despercebidos no processo exclusivamente manual, como em Moretti (2007).

Nosso experimento inicial acessa o projeto da internet *WayBackMachine* - WBM (FIG. 2) também conhecido como *Internet Archive,* que constitui-se de uma biblioteca digital de sites de internet com mais de 430 bilhões de páginas arquivadas. A iniciativa da WBM, que oficialmente não tem fins lucrativos, deu início aos trabalhos em 1996 tendo, a partir de 1999, incluído novos formatos em seu acervo

ISSN nº 2447-4266 Vol. 1, nº 2, Setembro-dezembro. 2015

DOI: http://dx.doi.org/10.20873/uft.2447-4266.2015v1n2p23

tais como vídeos, arquivos de som e de texto, software e outros se constituindo numa base de dados útil para certas pesquisas.

Para acessar esse repositório, desenvolvemos a prova de conceito de um código capaz de recuperar, listar e oferecer ferramentas básicas de análise sobre dados coletados a partir das diversas versões de portais jornalísticos ao longo do tempo.

Utilizando o conteúdo arquivado das séries disponibilizadas é possível avaliar métricas como o número de versões ou atualizações anuais, palavras mais frequentes ao longo do tempo, alterações na organização de conteúdo e design entre outras.

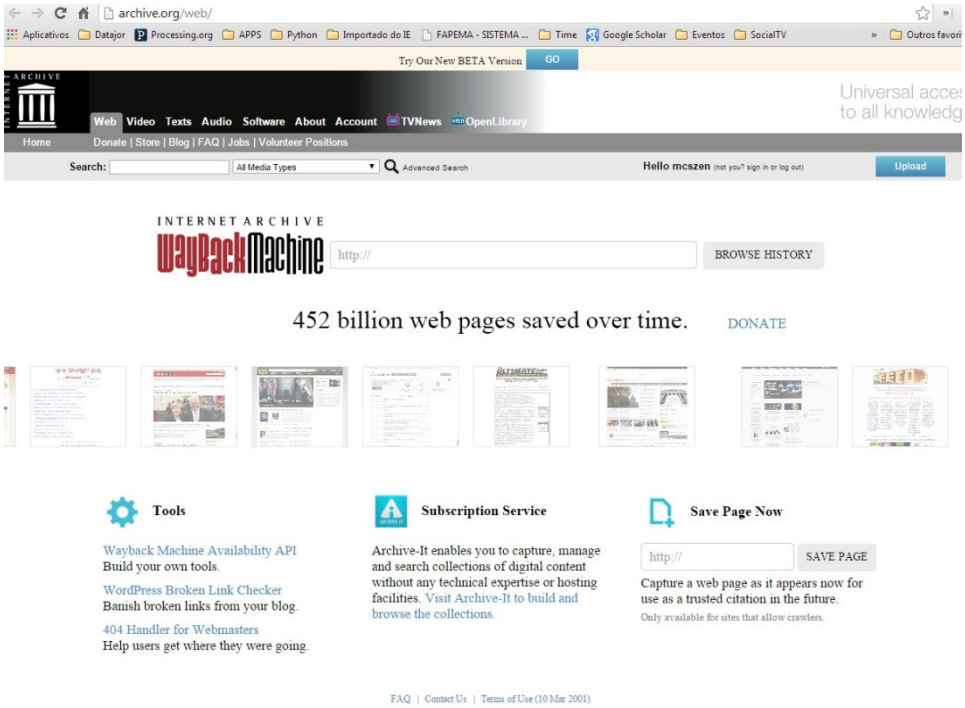

FIGURA 2- Tela da Home do site *Internet Archive* .
FONTE: Internet Archive (2014)

Waybackmachine e a memória dos sites jornalísticos

A WBM, na sua área de sites, a partir da indicação de determinado domínio, permite visualizar todas as versões arquivadas do mesmo, incluindo sua página inicial

ISSN nº 2447-4266 Vol. 1, nº 2, Setembro-dezembro. 2015

DOI: http://dx.doi.org/10.20873/uft.2447-4266.2015v1n2p23

(*home page*) e links principais, numa interface que mostra em formato de *timeline* (Fig. 3) e calendários as datas onde uma nova versão daquele site foi arquivada. Na imagem abaixo é possível ver o resultado de um teste feito a partir do endereço da Universidade Federal do Maranhão (www.ufma.br) que indica o número de versões disponíveis, as datas da primeira e da última versão em destaque e todos as outras marcadas nos calendários mensais com pontos azuis que a WBM chama de "*spots*". A plataforma mantém uma API que responde a consultas com uma sintaxe própria.

Segundo dados do projeto as coletas são feitas diariamente de forma a documentar novas versões que ao serem registradas podem ser acessadas pelos usuários a qualquer tempo através das ferramentas oferecidas.

FIGURA 3- Tela do Internet Archive com a marcação das versões arquivadas (399 entre 1997 e 2014) do site da UFMA em suas respectivas datas no ano de 2007 que podem ser extraídas via código.
FONTE: Internet Archive (2014)

O objetivo desse experimento constituiu-se no desenvolvimento de um código escrito utilizando a linguagem de programação Python, capaz de realizar as seguintes funções:

a) A partir da indicação de um endereço na internet pelo usuário, buscar no repositório da WBM informações básicas sobre o número de versões, datas

ISSN nº 2447-4266 Vol. 1, nº 2, Setembro-dezembro. 2015

DOI: http://dx.doi.org/10.20873/uft.2447-4266.2015v1n2p23

da primeira e última coletas e links para as páginas de todas os registros documentados.

b) Extrair o número de atualizações por ano de forma a identificar padrões relativos à intensidade da atualização do conteúdo do mesmo. Como focamos nosso estudo em sites dos grandes portais jornalísticos é possível associar essa métrica à característica da atualização constante, que nos livros teóricos é comumente descrita como traço distintivo do jornalismo digital, apesar de poucos tratarem o assunto com dados empíricos.

c) Extrair elementos de interesse para pesquisa como links e textos das versões coletadas. Tal material presta-se portanto a estudos onde, por exemplo, as transformações associadas a temáticas específicas são alvo de investigação.

d) Gerar visualizações das métricas avaliadas como demonstraremos abaixo em relação às atualizações por ano.

Uma das vantagens da linguagem Python é a grande quantidade de módulos disponíveis para a execução das mais diversas funções, fato que facilita muito o programador inexperiente ou oriundo de outras áreas de conhecimento. Neste experimento além das funções internas básicas da linguagem utilizamos os módulos Mathplotlib, Numpy e NLTK (BIRD; EDWARD; KLEIN, 2009) como ferramentas para gerar as visualizações e analisar as métricas relacionadas aos textos extraídos.

O fato de utilizarmos em nosso experimento os sites de caráter jornalístico não impede a utilização da ferramenta em outros cenários de pesquisa onde a série histórica de versões de sites tenha algum interesse.

Para atingir nosso objetivo trabalhamos com a metodologia descrita a seguir. Inicialmente fizemos um estudo da própria plataforma avaliando a estruturação do código HTML que a suporta e identificando os padrões de resposta da API para as requisições das versões de um endereço específico.

ISSN nº 2447-4266 Vol. 1, nº 2, Setembro-dezembro. 2015

DOI: http://dx.doi.org/10.20873/uft.2447-4266.2015v1n2p23

A partir do conhecimento de como a WBM trabalha internamente, de início implementamos no algoritmo as funções de consulta, registro de informações básicas, listagem dos endereços da páginas arquivas, estruturação da quantidade de versões por ano e geração de gráfico com a evolução das atualizações ao longo do tempo.

O que nosso algoritmo permite é fazer uma consulta idêntica à que é feita diretamente no site da WBM, entretanto, permitindo que de forma automática todos os endereços das páginas registradas sejam listados para posterior acesso e análise.

Inicialmente o código recupera as informações básicas oferecidas pela plataforma que são o número de versões registradas e as datas do primeiro e do último registro (FIG. 4) que serão utilizados também como parâmetros para a coleta de todas as outras atualizações arquivadas.

Depois dessa etapa o programa vai processar e salvar numa lista e em um arquivo de texto todos os endereços das páginas (URLs) onde estão as versões registradas na plataforma. O exemplo abaixo (FIG. 5) contém todas as versões arquivadas do site da UFMA. Essa lista posteriormente pode ser lida por outra função do software que vai extrair de cada uma os links e textos associados, constituindo assim um corpus empírico bem mais amplo para o pesquisador que terá ainda a possiblidade de aplicar outras ferramentas específicas em sua análise.

ISSN nº 2447-4266 Vol. 1, nº 2, Setembro-dezembro. 2015

DOI: http://dx.doi.org/10.20873/uft.2447-4266.2015v1n2p23

```
Python 2.7.3 (default, Apr 10 2012, 23:31:26) [MSC v.1500 32 bit (Intel)] on win
32
Type "copyright", "credits" or "license()" for more information.
>>> ================================ RESTART ================================
>>>
Digite o site que deseja:www.ig.com.br

Você digitou: www.ig.com.br

Se digitou errado aperte a tecla ESC, dê OK e depois F5 para rodar de novo.

Versões registradas= 19406
Início do Registro= Sat Apr 08 20:28:24 UTC 2000
Último Registro= Sun Feb 22 19:44:09 UTC 2015

Processando ano de 2000
Processando ano de 2000
Processando ano de 2000
Processando ano de 2000
Processando ano de 2000
Processando ano de 2000
Processando ano de 2000
Processando ano de 2000
Processando ano de 2000
Processando ano de 2000
Processando ano de 2000
Processando ano de 2000
Processando ano de 2000
Processando ano de 2000
```

FIGURA 4 – Print da tela do código com as informações iniciais básicas (item a da lista de objetivos) do site jornalístico www.ig.com.br .
FONTE: Elaborado pelo autor.

```
siteufmabr399 vers - Bloco de notas
Arquivo  Editar  Formatar  Exibir  Ajuda
Links para versões do site ufma.br
https://web.archive.org/web/19970601184857/http://ufma.br/
https://web.archive.org/web/19971023004327/http://ufma.br/
https://web.archive.org/web/19980110155537/http://ufma.br/
https://web.archive.org/web/19980124223026/http://ufma.br/
https://web.archive.org/web/19980213083048/http://ufma.br/
https://web.archive.org/web/19981206044342/http://ufma.br/
https://web.archive.org/web/19990125104123/http://ufma.br/
https://web.archive.org/web/19990223191249/http://ufma.br/
https://web.archive.org/web/19990427042721/http://ufma.br/
https://web.archive.org/web/20000510141013/http://ufma.br/
https://web.archive.org/web/20000512002401/http://ufma.br/
https://web.archive.org/web/20000520085250/http://ufma.br/
https://web.archive.org/web/20000609135500/http://ufma.br/
https://web.archive.org/web/20000817190520/http://ufma.br/
https://web.archive.org/web/20001024165243/http://ufma.br/
https://web.archive.org/web/20001109170300/http://ufma.br/
https://web.archive.org/web/20001203143500/http://ufma.br/
https://web.archive.org/web/20001206220500/http://ufma.br/
https://web.archive.org/web/20010202202800/http://ufma.br/
https://web.archive.org/web/20010208123658/http://ufma.br/
https://web.archive.org/web/20010301221839/http://ufma.br/
https://web.archive.org/web/20010302022029/http://ufma.br/
https://web.archive.org/web/20010330015858/http://ufma.br/
https://web.archive.org/web/20010405021416/http://ufma.br/
https://web.archive.org/web/20010429153813/http://ufma.br/
https://web.archive.org/web/20010429160028/http://ufma.br/
https://web.archive.org/web/20010507183545/http://ufma.br/
https://web.archive.org/web/20010925032612/http://ufma.br/
https://web.archive.org/web/20020122051318/http://ufma.br/
https://web.archive.org/web/20020402124440/http://ufma.br/
https://web.archive.org/web/20020526154405/http://ufma.br/
https://web.archive.org/web/20020531030548/http://ufma.br/
https://web.archive.org/web/20020721084530/http://ufma.br/
https://web.archive.org/web/20020928154419/http://ufma.br/
https://web.archive.org/web/20020930092207/http://ufma.br/
https://web.archive.org/web/20021120041705/http://ufma.br/
https://web.archive.org/web/20021128093629/http://ufma.br/
https://web.archive.org/web/20030203141728/http://ufma.br/
https://web.archive.org/web/20030215011707/http://ufma.br/
https://web.archive.org/web/20030329120458/http://ufma.br/
https://web.archive.org/web/20030501055043/http://ufma.br/
https://web.archive.org/web/20030528031219/http://ufma.br/
https://web.archive.org/web/20030620171808/http://ufma.br/
https://web.archive.org/web/20030621042319/http://ufma.br/
https://web.archive.org/web/20030804051232/http://ufma.br/
https://web.archive.org/web/20030927125649/http://ufma.br/
https://web.archive.org/web/20031008104833/http://ufma.br/
```

FIGURA 5 – Print do arquivo com a lista de links extraídos automaticamente que levam às páginas arquivadas do site www.ufma.br (item a da lista de objetivos).
FONTE: Elaborado pelo autor.

ISSN nº 2447-4266 Vol. 1, nº 2, Setembro-dezembro. 2015

DOI: http://dx.doi.org/10.20873/uft.2447-4266.2015v1n2p23

Depois de processar todas as versões coletadas, o código as conta e classifica por ano a fim de que seja possível identificar o número de atualizações por cada período (FIG. 6). Tal métrica nos permitirá identificar a velocidade com que os sites estudados tem se modificado ao longo do tempo, um fator que, no caso dos sites jornalísticos pode ser associado à característica da atualização constante, frequentemente atribuída ao jornalismo de internet.

É importante ressaltar que o número de versões identificadas pela plataforma WBM não representa o universo total de mudanças. Segundo dados da própria WBM, os resultados são conseguidos através de um *crawler*[6] próprio e de dados do portal Alexa que também varre a internet diariamente. De qualquer forma, pela quantidade de registros, é possível perceber que a amostra oferecida pela WBM é bastante significativa e, considerando que usa a mesma metodologia para a coleta dos diferentes sites que arquiva, tal amostra pode ser utilizada em estudos comparativos de métricas específicas, como nesse estudo.

[6] *Crawler*s, também conhecidos como robôs, são programas que varrem a internet registrando endereços de páginas e arquivando-os. Motores de busca como Google, plataformas de análise como Alexa (www.alexa.com) e bibliotecas digitais como a WBM usam algoritmos assim para executar suas funções.

ISSN nº 2447-4266 Vol. 1, nº 2, Setembro-dezembro. 2015

DOI: http://dx.doi.org/10.20873/uft.2447-4266.2015v1n2p23

```
Processando ano de 2015
Processando ano de 2015
Processando ano de 2015
Processando ano de 2015
Processando ano de 2015
Processando ano de 2015
Processando ano de 2015
Processando ano de 2015
Processando ano de 2015
Processando ano de 2015
Processando ano de 2015
Processando ano de 2015
Processando ano de 2015
Processando ano de 2015
Processando ano de 2015
Processando ano de 2015
Processando ano de 2015
Processando ano de 2015
Processando ano de 2015
Processando ano de 2015
Processando ano de 2015
Processando ano de 2015
Processando ano de 2015
Processando ano de 2015
Processando ano de 2015
Processando ano de 2015
Itens encontrados:19406
Itens informados  no site:19406

Lista final com 19406 itens.
['2002', '2003', '2000', '2001', '2006', '2007', '2004', '2005', '2015', '2014',
 '2008', '2009', '2011', '2010', '2013', '2012']
[16, 26, 39, 17, 211, 139, 135, 340, 1210, 10764, 154, 81, 322, 179, 4272, 1501]
[(2002, 16), (2003, 26), (2000, 39), (2001, 17), (2006, 211), (2007, 139), (2004
 135), (2005, 340), (2015, 1210), (2014, 10764), (2008, 154), (2009, 81), (2011
 322), (2010, 179), (2013, 4272), (2012, 1501)]
[2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2
013, 2014, 2015]
[39, 17, 16, 26, 135, 340, 211, 139, 154, 81, 179, 322, 1501, 4272, 10764, 1210]
[2000, 2015, 16, 10764]
```

FIGURA 6 – Print com destaque para a parte do código que conta por ano o número de atualizações registradas (item b da lista de objetivos).
FONTE: Elaborado pelo autor.

Com o número de versões contabilizadas é possível então gerar uma primeira visualização que representa a série temporal de atualizações extraídas do registro da WBM. O gráfico abaixo (FIG.7) traz essa métrica plotada a partir dos dados do site www.ig.com.br.

Para efeito desse estudo, apesar de coletados, os dados relativos ao ano de 2015 foram excluídos dos gráficos já que se referem a apenas alguns meses, configurando uma unidade temporal diferente do restante.

ISSN nº 2447-4266 Vol. 1, nº 2, Setembro-dezembro. 2015

DOI: http://dx.doi.org/10.20873/uft.2447-4266.2015v1n2p23

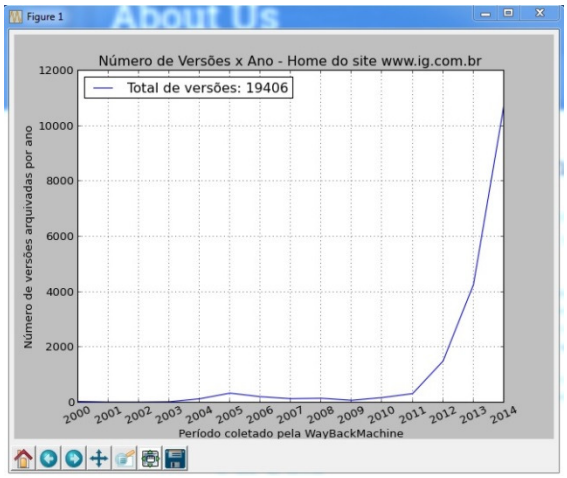

FIGURA 7 – Gráfico plotado com as atualizações registradas entre os anos de 2000 e 2014 do site www.ig.com.br (item d da lista de objetivos).
FONTE: Elaborado pelo autor.

Para seleção dos sites jornalísticos do nosso estudo utilizamos a classificação da plataforma Alexa[7] que, entre outras ferramentas, ranqueia sites e portais da internet em função do número de acessos. Entre os 50 sites com os maiores números no Brasil, selecionamos os que pertencem à categoria jornalismo. Por esse critério foram escolhidos os sites estadão.com.br; uol.com.br; globo.com; ig.com.br; terra.com.br e abril.com.br .

As visualizações abaixo (FIG. 8) foram conseguidas seguindo as etapas já descritas e demonstram como a característica da atualização constante passou a ter uma relevância entre os anos de 2010 (estadão) e 2011 (uol, globo, ig e terra) impactando de forma maior ou menor, de acordo com cada caso, a quantidade de atualizações registradas. Apenas o site abril.com.br parece ter aumentado o número de atualizações tardiamente com um incremente significativo apenas em 2013. Tal fato talvez se justifique pela periodicidade semanal e não diária da produção jornalística original gerada pelos veículos administrados pela empresa que, em 2013,

[7] www.alexa.com

ISSN nº 2447-4266 Vol. 1, nº 2, Setembro-dezembro. 2015

DOI: http://dx.doi.org/10.20873/uft.2447-4266.2015v1n2p23

passaria a ter uma integração mais forte à internet como canal de distribuição desse conteúdo.

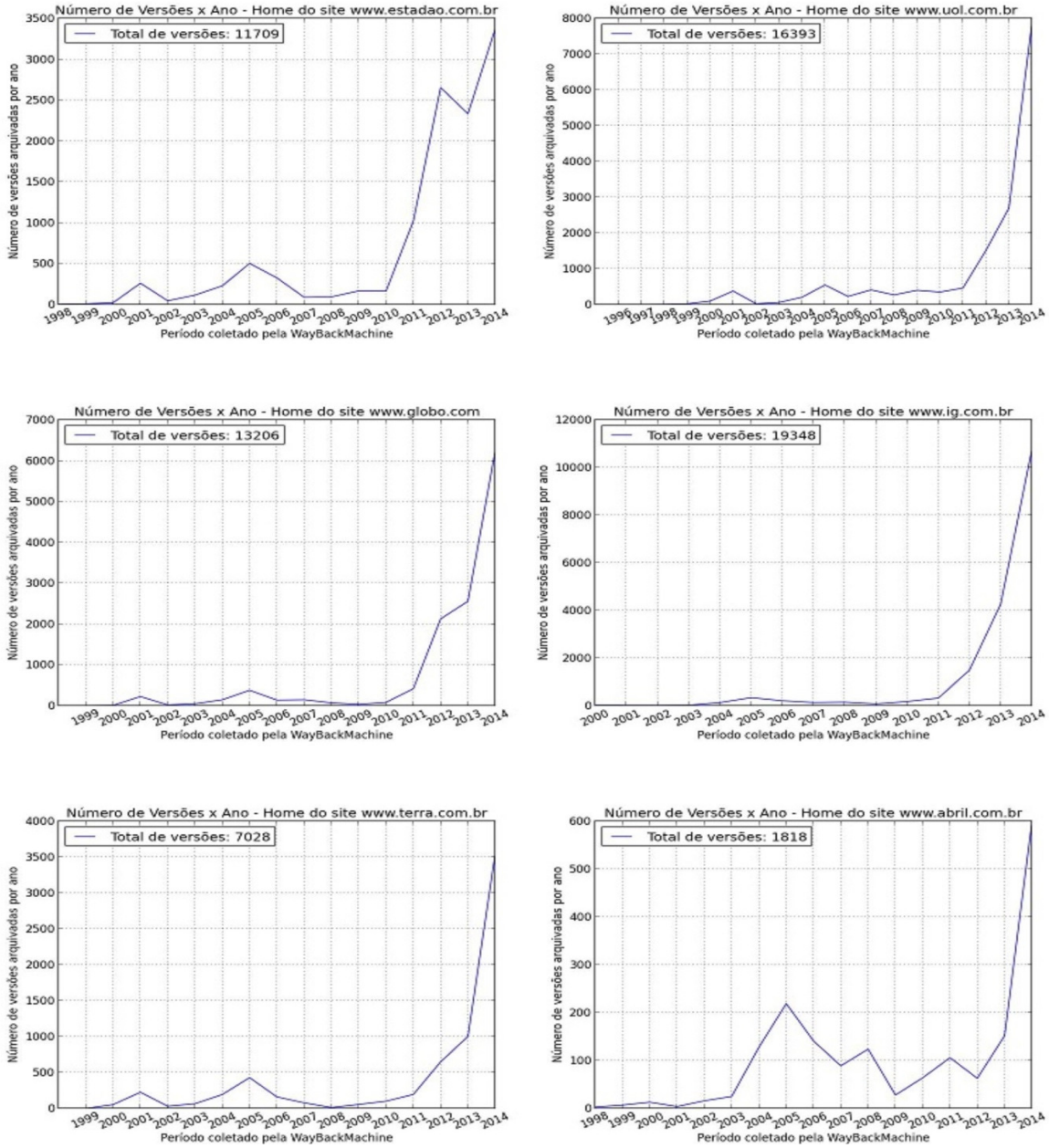

FIGURA 8 – Gráficos mostrando o crescimento dos números de atualizações a partir dos anos 2010 e 2011 nos principais sites jornalísticos brasileiros.

FONTE: Elaborado pelo autor.

ISSN nº 2447-4266 Vol. 1, nº 2, Setembro-dezembro. 2015

DOI: http://dx.doi.org/10.20873/uft.2447-4266.2015v1n2p23

Verificando o site abril.com em suas versões anteriores observamos também que durante um bom período de tempo a página inicial apenas era usada para divulgar as diversas publicações semanais da editora e não para divulgação direta de notícias, procedimento que só foi implementado nos últimos anos e ainda de forma parcial. Tal situação explica as diferenças encontradas nos gráficos acima e nos permite também explorar outro aspecto dos arquivos que é a sua estrutura gráfica ou visual.

Uma função ainda em fase de teste permite que também salvemos *prints*, ou seja, visualizações das versões arquivadas (FIG. 9), facilitando a compreensão das mudanças estéticas ou funcionais que os administradores do site foram definindo ao longo da série histórica analisada.

FIGURA 9 – Recorte de *print* salvo a partir do site www.abril.com.br demonstrando que, nesse caso a utilização da *home* é mais utilizada para divulgação das revistas do que das notícias.
FONTE: Elaborado pelo autor.

Por fim, a partir dos endereços que contém as páginas arquivadas é possível coletar os textos utilizados nos links da página principal que indicam os temas de interesse e, no caso de sites jornalísticos, em grande parte, as chamadas para as matérias que foram publicadas.

ISSN nº 2447-4266 Vol. 1, nº 2, Setembro-dezembro. 2015

DOI: http://dx.doi.org/10.20873/uft.2447-4266.2015v1n2p23

Apenas como teste utilizamos a ferramenta no site do LABCOM (www.labcomufma.com) que tem poucas versões arquivadas na WB para extrair os testos dos links e verificar a frequência de utilização de cada termo (FIG. 10).

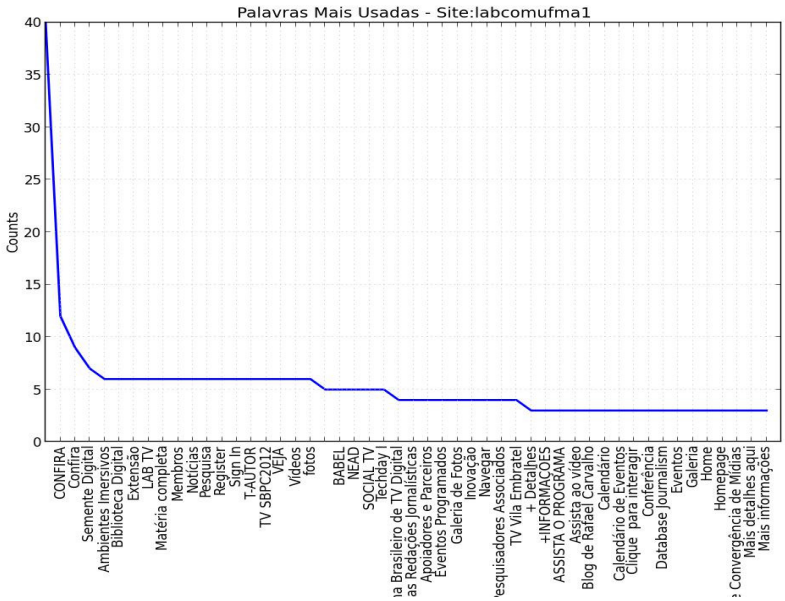

FIGURA 10 – Gráfico que mostra as 50 palavras ou expressões mais usadas nas versões arquivadas do site www.labcomufma.com .
FONTE: Elaborado pelo autor.

Pelo gráfico é possível identificar que o projeto Semente Digital, que trabalha a preservação do patrimônio histórico da cidade de São Luís utilizando tecnologia, teve mais atenção nas publicações do site, perdendo apenas para a palavra "confira" muito utilizada para indicar links e chamadas de matérias.

Considerações finais

A vertente aplicada do presente trabalho é um recorte de uma iniciativa mais ampla voltada ao desenvolvimento de métodos específicos e de uma epistemologia especializada para os estudos da Comunicação Digital. A automatização de processos repetitivos e a análise de grandes volumes de dados têm demonstrado um potencial

ISSN nº 2447-4266 Vol. 1, nº 2, Setembro-dezembro. 2015

DOI: http://dx.doi.org/10.20873/uft.2447-4266.2015v1n2p23

de oportunidades em termos de pesquisa na área de Comunicação e o acesso à memória digital como no experimento aqui apresentado é um exemplo desse caminho.

A escala de aplicação de tais ferramentas, que implica num gradiente de possibilidades de utilização, não obriga nenhum pesquisador a aprender a programar, mas aponta para um caminho onde a formação de equipes multidisciplinares e a compreensão técnica das características dos meios de comunicação, principalmente a internet, pode trazer fundamental diferença nos horizontes a serem vislumbrados.

Tal fato se reflete principalmente na estratégia de amostragem permitida que, com o software e a coleta automatizada passa a oferecer mais abrangência e, consequentemente, potencial de inferência maior.

No atual estádio de desenvolvimento, o código já consegue cumprir os objetivos básicos inicialmente propostos oferecendo um caminho simplificado para a extração dos endereços de todas as versões arquivadas na WBM e posterior utilização dos mesmos para análise da frequência de mudanças ao longo do tempo, arquivamento de imagens das páginas principais e coleta e análise das palavras e expressões mais utilizadas na série histórica em estudo.

Este e outras soluções de código, tais como as também desenvolvidas em Santos (2013 e 2014), que constituem a parte aplicada da proposta dos métodos digitais em pesquisas da área de Comunicação serão em breve oferecidas à comunidade científica através de um site específico ainda em construção que utilizará o domínio www.labcomdados.com.br .

Referências

BIRD, Steven; LOPER, Edward; KLEIN, Ewan. **Natural Language Processing with Python:** analyzing text with the Natural Language Toolkit. New York: O'Reilly Media Inc., 2009.

ISSN nº 2447-4266 Vol. 1, nº 2, Setembro-dezembro. 2015

DOI: http://dx.doi.org/10.20873/uft.2447-4266.2015v1n2p23

BONACICH, Phillip; LU, Phillip. **Introduction to mathematical sociology**. New Jersey: Princeton University Press, 2012.

BRADSHAW, Paul. **Scraping for Journalists.** Leanpub, 2014, [E-book].

GLEICK, James. **A Informação.** Uma história, uma teoria, uma enxurrada. São Paulo, Companhia das Letras, 2013.

MANOVICH, Lev. **The Language of New Media.** Cambrige: Mit Press, 2001.

MORETTI, Franco. **Graphs, maps, trees.** Abstract models for literary history. New York, Verso, 2007.

ROGERS, Richard. **Digital Methods.** Cambridge: Mit Press, 2013. E-book.

SANTOS, Márcio. Conversando com uma API: um estudo exploratório sobre TV social a partir da relação entre o twitter e a programação da televisão. **Revista Geminis**, ano 4 n. 1, p. 89-107, São Carlos. 2013. Disponível em: <www.revistageminis.ufscar.br/index.php/geminis/article/view/129/101>. Acesso em: 20 abr. 2013.

SANTOS, Márcio. Textos gerados por software. Surge um novo gênero jornalístico. **Anais XXXVII Congresso Brasileiro de Ciências da Comunicação.** Foz do Iguaçu, 2014. Disponível em: <http://www.labcomufma.com/biblioteca-digital>. Acesso em 26 jan. 2014.

Acesse esse e outros artigos da **Revista Observatório** em: