BIG DATA - O QUE OS MÉTODOS TRADICIONAIS NÃO CONSEGUEM VER

O conceito de *distant reader* aplicado à pesquisa na Comunicação

Márcio Carneiro dos Santos[1]

Resumo: São descritas as linhas gerais de uma adaptação metodológica das técnicas de processamento de grandes volumes de dados da área de *big data* para iniciativas de pesquisa em Ciências Sociais. Inspirado no conceito de *distante reader* de Moreti (2007) a abordagem apresentada, que tem foco em situações de excesso de informação, tem o objetivo de identificar padrões de difícil apreensão em amostras reduzidas. Através de um recorte na área do jornalismo são apresentados dois exemplos práticos: o primeiro onde a metodologia foi usada para estudar mais de 90 mil registros de sites de notícia brasileiros e comprovar a característica da atualização constante normalmente atribuída ao processo de produção digital e o segundo onde, com poucas linhas de código, se gerou uma visualização da distribuição das revistas acadêmicas da subárea de Comunicação e Informação a partir de uma planilha da CAPES com mais de 120 mil linhas.

Palavras-chave: *big data; distant reader;* métodos digitais; jornalismo digital; *python*.

1. INTRODUÇÃO

O avanço da ciência não se faz apenas através do seu desenvolvimento interno, baseado na consolidação progressiva do conhecimento produzido ou herdado, mas também a partir do alinhamento dos recursos que utiliza, no caso seus métodos e técnicas, às transformações históricas que cada momento da humanidade oferece como cenário para estudo e apreensão. Na sociedade contemporânea reconstruída sobre fluxos de informação binária, fenômenos complexos, caracterizados pelo excesso de dados para análise, tornaram-se comuns entre os objetos da comunicação digital, área onde trabalhos recentes como Franciscato (2017), Machado (2016); Machado e Rohden (2016) mostram os estudos de caso e o olhar individualizado como escolhas majoritárias, em

[1] Prof. Adjunto do DCS/UFMA. Dr.em Tecnologias da Inteligência e Design Digital pela PUC-SP . Bolsista de Produtividade DT-2 do CNPq. Email: mcszen@gmail.com .

detrimento de abordagens que permitam a identificação de padrões mais gerais, baseadas em amostras significativamente mais amplas e, por isso, com maior potencial para gerar inferências, testar teorias existentes ou sugerir novas.

O presente texto apresenta um recorte (direcionado ao jornalismo) de um trabalho de caráter teórico e metodológico mais amplo, ainda em andamento, que problematiza a necessidade de incorporação de metodologias até então pouco comuns na Comunicação, para a modelagem de pesquisas onde a situação do excesso de dados é uma das variáveis a considerar; mesmo sabendo que tal direcionamento vai de encontro à agenda descritiva-interpretativa das Humanidades, principal pilar epistemológico de um campo que agora tem de olhar para objetos regidos pela lógica numérica.

Além da discussão proposta no texto, essa iniciativa contempla os princípios de um *framework* que considera três aspectos básicos: a) uma ontologia especializada dos objetos digitais; b) a configuração ou estrutura assumida pelos fluxos de informação/comunicação e c) a utilização de recursos computacionais intensivos, na linha do que definimos como métodos digitais (SANTOS, 2016), para identificação de padrões e tendências em grandes volumes de dados, utilizando o conceito de *distant reader* (MORETTI, 2007) como inspiração.

A discussão dessa temática recortada para o âmbito do jornalismo digital não se dá ao acaso. Foi o excesso de informação disponível no ambiente da internet, através dos portais de transparência e dos repositórios *online* ou ainda nos acervos que foram convertidos para formas binárias de arquivamento, que levou os jornalistas a enfrentarem a necessidade de adaptar métodos de coleta e análise capazes de lhes proporcionar uma compreensão mais ampla dos cenários em que estavam trabalhando. 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; termos associados a outro conceito, 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 disponível hoje.

As possíveis conexões da área de *big data* com a atividade jornalística já tem sido estudadas sobre diversos aspectos. Para citar apenas alguns, poderíamos listar Codding-

ton (2015) que analisa as diferenças entre reportagem assistida por computador, jornalismo de dados e jornalismo computacional; Lewis e Westlund (2015) que examinam as questões epistemológicas, técnicas, econômicas e éticas dessa relação e ainda Lima Jr. (2012), um dos precursores dessa discussão no Brasil, que já há alguns anos levantava a necessidade de atualização profissional e da aproximação com o campo das Ciências da Computação. Em termos mais gerais a aproximação das Ciências Sociais com a temática de *big data* também tem sido foco de muitos trabalhos como em Gonález-Bailón (2013), que argumenta que a partir desse tipo de abordagem novas questões podem ser feitas e antigas revisitadas; Mahrt e Scharkow (2012) que discutem as situações onde a análise de grandes quantidades de dados pode ser útil ou ainda Bruns (2013) que problematiza as dificuldades em conciliar procedimentos e terminologia oriundos das ciências ditas duras ao trabalho acadêmico das Humanidades e Ciências Sociais.

Em paralelo, a necessidade de traduzir enormes massas de dados em formas de mais fácil apreensão para os consumidores de informação também levou esses profissionais a utilizarem ferramentas de visualização e infografia (RODRIGUES, 2009; CORDEIRO, 2013), capazes de traduzir, em imagens mais simples, padrões, tendências e inferências, num tipo de lógica de síntese já comum em áreas como a Economia e a Estatística.

Se em termos de técnicas a extração de dados (*scraping*) e a visualização têm se transformado em novas habilidades fundamentais para o jornalista digital, mesmo conhecimentos mais distantes como o de programação tem recebido interesse crescente entre os profissionais. Um exemplo a ser citado foi o curso de "*Python for Data Journalists*" promovido entre junho e julho de 2017 pelo *Knight Center* da Universidade do Texas que na modalidade online e gratuita conseguiu mais de 2.700 interessados entre profissionais de cerca de 120 países. Mesmo sendo o curso em inglês, ministrado pelo jornalista Ben Welsh do *Los Angeles Times*, o Brasil foi o segundo país com maior número de inscritos (319), indicando que tal necessidade não tem sido sentida apenas em redações da América do Norte e Europa, mas também por grupos locais. A cidade de São Paulo, por exemplo, teve mais inscritos que Londres e Nova York.

name	total	percent	code
United States	1064	0.388747	USA
Brazil	319	0.116551	BRA
Spain	104	0.037998	ESP
United Kingdom	87	0.031787	GBR
Mexico	81	0.029594	MEX
India	70	0.025575	IND
Germany	63	0.023018	DEU
Canada	62	0.022653	CAN
Argentina	52	0.018999	ARG
Nigeria	39	0.014249	NGA
Australia	34	0.012422	AUS
Colombia	32	0.011692	COL
Ukraine	30	0.010961	UKR
Venezuela, Bolivarian Republic Of	28	0.010230	VEN
China	26	0.009499	CHN
Italy	26	0.009499	ITA
Netherlands	26	0.009499	NLD
France	23	0.008403	FRA
Chile	23	0.008403	CHL
Portugal	22	0.008038	PRT

name	total	percent
	513	0.187226
São Paulo	46	0.016788
London	41	0.014964
New York	41	0.014964
Madrid	33	0.012044
Los Angeles	31	0.011314
Rio de Janeiro	27	0.009854
Austin	27	0.009854
Washington	22	0.008029
San Francisco	22	0.008029
Chicago	20	0.007299
Washington, DC	18	0.006569
Lagos	18	0.006569
Buenos Aires	18	0.006569
Brooklyn	17	0.006204
Santiago	16	0.005839
Toronto	15	0.005474
Berlin	15	0.005474
Seattle	14	0.005109
Brasília	14	0.005109

Tabela 1 – Participação de alunos por país e número de alunos nas principais cidades que tiveram inscritos no curso de Python para Jornalistas do Knight Center. Fonte: Welsh (2017)

Podemos perguntar então por que tais habilidades e práticas não tem se refletido na atividade acadêmica da área através da incorporação de metodologias e técnicas de coleta e análise de dados condizentes com o cenário contemporâneo de excesso de informações disponível e por que quando isso acontece dá-se de forma periférica, basicamente em abordagens meramente descritivas, que não operam efetivamente com a lógica do digital e com as particularidades dos seus objetos.

Partimos das seguintes premissas:

a) De que a formação não apenas dos profissionais e também dos pesquisadores da área naturalmente valoriza metodologias oriundas das Humanidades. O que de forma alguma é um problema, mas potencialmente um limitador em função de determinados objetos de estudo do ambiente digital.

b) De que essa formação não tem se atualizado no sentido de gerar e incorporar métodos e práticas mais alinhados com as características inerentes desses objetos ou, nos termos de Santos (2016), de sua ontologia especializada, incluin-

do ai suas interfaces com questões de ordem tecnológica e econômica que impactam muitos dos fenômenos que nos propomos a estudar.

c) De que essa formação segue uma tendência reducionista onde o aprofundamento da análise individualizada deixa pouca margem para a apreensão das interconexões e interfaces desses fenômenos em termos coletivos, interna e externamente, dificultando a exploração de inferências, a avaliação e a reformulação teóricas; transformando o campo num enorme conjunto de estudos de caso que não permitem a visualização dos efeitos sistêmicos e coletivos das transformações por que passa o ecossistema de meios da atualidade como apontam Franciscato (2017), Machado (2016) e Machado e Rohden (2016).

Diante dessas premissas o presente texto pretende colaborar alinhando alguns tópicos teóricos e metodológicos que poderão ser utilizados como um *framework* de orientação inicial para iniciativas de pesquisa que se arrisquem a explorar caminhos pouco comuns até então e que, seguindo em direção contrária ao fluxo principal dos trabalhos produzidos, enfatizam as visões de conjunto e a interconexão entre as partes; a utilização de conceitos focados nos aspectos da materialidade dos processos comunicacionais estudados e, por fim, a utilização de recursos computacionais mais robustos para execução de parte dessas atividades.

2. O CONCEITO DE *DISTANT READER*

A utilização de métodos quantitativos nas Ciências Sociais e Humanas não é nova, entretanto a disponibilidade de recursos computacionais intensivos e o surgimento de projetos de pesquisa interdisciplinares onde essas áreas puderam interagir e atuar em conjunto com pesquisadores com outro tipo de formação, notadamente da Ciência da Computação, do Design e da Estatística, ofereceram um cenário onde novas questões de pesquisa puderam ser feitas e uma leitura diferente dos dados disponíveis passou a ser possível.

Uma dessas iniciativas descrita por Moretti (2007), surgiu de um inesperado encontro entre os estudos de literatura e a utilização de ferramentas computacionais para analisar a produção inteira de movimentos literários antes profundamente estudados,

mas, sempre sob o olhar do especialista em determinado autor ou escola, a partir da análise de obras selecionadas consideradas mais representativas daquela produção. Um olhar próximo que se atém aos detalhes de um determinado texto ou, no máximo, a um conjunto pequeno de textos tidos como semelhantes. Utilizando ferramentas de visualização como gráficos, mapas e árvores, o autor conseguiu traduzir um grande conjunto de informações sobre esses objetos em visualizações que representavam descobertas nunca antes observadas. Partindo de amostras bem maiores do que seria a produção de determinado movimento ou escola literária, Moretti conseguiu um tipo de apreensão semelhante ao gerado pelo movimento de zoom out das câmeras de filmagem que, abrindo o ângulo de visão, consegue definir um quadro com mais informações e, principalmente, onde as partes que o compõem podem ser percebidas através das inter-relações que estabelecem com as outras partes num "tipo de abordagem onde a distância não é um obstáculo e sim uma forma específica de conhecimento com poucos elementos mas uma apurada percepção do seu conjunto de interconexões" (MORETTI, 2007, pág. 2) .

Tal abordagem descrita por alguns como uma herética mistura entre métodos quantitativos, geografia e teoria evolucionária gerou gráficos como os da figura abaixo onde estudos voltados a épocas anteriores identificaram as formas hegemônicas da literatura britânica entre 1760 a 1850, a distribuição dos gêneros novelísticos da produção inglesa entre 1740 e 1900 ou, mais recentemente um mapa com a penetração da comédia americana no cinema em diversos países do mundo, identificando, por exemplo, as dificuldades desse gênero nos países asiáticos.

Há que se observar que as amostras utilizadas para os referidos estudos são compostas de milhares de obras, o que seria um problema para a avaliação desse conjunto por um único leitor ou mesmo um grupo de pesquisadores. Para poder analisa-las individualmente tal equipe teria que passar toda a sua vida lendo o material, provavelmente sem conseguir terminar seu trabalho e muito menos identificar tais padrões mais gerais que só foram possíveis de apreensão através do processamento do grande volume de informações e sua posterior organização utilizando formas de visualização em gráficos e mapas como formas simplificadoras.

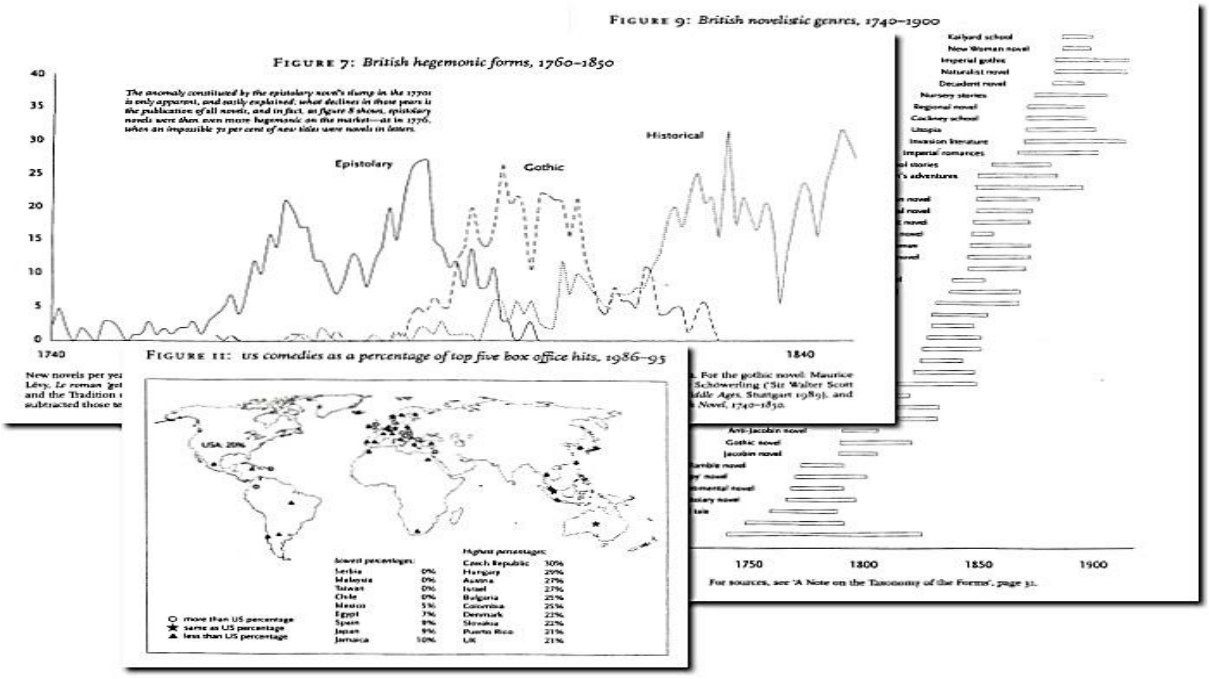

Figura 1 – Exemplos de gráficos e mapas gerados a partir do processamento de grande volume de informações sobre formas e gêneros narrativos. Fonte: Moretti (2007).

O conceito de *distant reader* proposto pelo autor, baseado na busca de padrões e tendências mais gerais a partir da estruturação de volumosos conjuntos de dados pode ser replicado em outros campos além da literatura, entre eles o do jornalismo. Para efetivar tal empreitada algumas questões básicas devem ser listadas em termos metodológicos:

. Como determinar o contexto onde se insere tal objeto e mapear suas interconexões ainda que de forma preliminar?

. Como abordar grandes repositórios digitais ou arquivos com diferentes formatos de estruturação e deles coletar dados empíricos para análise?

. Como operar sobre esses dados utilizando ferramentas computacionais que exijam baixo conhecimento específico e tem aplicabilidade geral?

3. ANÁLISE DE CONTEXTO

A apreensão de quadros semelhantes aos desenvolvidos pela experiência do *distant reader* relatada acima deve considerar os diferentes aspectos relacionados a deter-

minado objeto de estudo e implica normalmente em algum tipo de visão sistêmica que orienta a percepção de um conjunto representado pelas inter-relações que ele (o objeto) estabelece com seu entorno.

Para o campo da Comunicação uma das formas viáveis de avaliar a produção de sentido no ambiente digital é estabelecer que forças atuam nesse processo a partir de três vetores básicos : tecnológico, econômico e cultural.

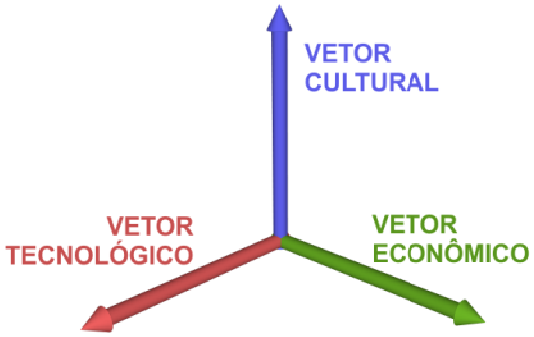

Figura 2: Representação do modelo de análise de transformações a partir de três vetores fundamentais.
Fonte: do autor.

Poderíamos citar como exemplo o estudo de determinado conjunto de mensagens numa plataforma de mídia social. Analisar apenas os textos (vetor cultural), num esforço interpretativo bastante comum nos estudos da Comunicação poderia levar a conclusões incompletas ou até erradas já que não se consideraria aspectos tecnológicos e econômicos relacionados ao ambiente ou contexto onde essa produção se desenvolve. Avaliar apenas as intenções do autor sem considerar que naquele ambiente um software com regras específicas impacta a quem será exposta a mensagem (vetor tecnológico) e que tal procedimento está alinhado aos interesses da plataforma baseada num modelo de negócios que vive da venda de inteligência de mercado e se alimenta a partir do crescimento da própria rede para manter a confiança dos seus acionistas (vetor econômico) pode comprometer seriamente qualquer esforço de análise, principalmente se tal iniciativa se basear em amostras pequenas e de baixa representatividade.

A contextualização a partir da busca da influência desses aspectos sobre determinado fenômeno estudado oferece um quadro mais claro das forças envolvidas, podendo tal relação inclusive ser formalizada e traduzida em termos quantitativos para modela-

gem de problemas como a definição de um índice de impacto de publicações em redes sociais.

Um desdobramento mais detalhado sobre esse tipo de análise contextual está no trabalho de Van Dick (2013) para estudar plataformas digitais de mídias sociais. Em situações assim a autora afirma que é preciso considerar diversos aspectos em dois níveis, que ela chama de micro e macro. No primeiro seria preciso avaliar questões como as características da própria tecnologia; o tipo de conteúdo que permite criar; bem como os usos e apropriações que dela advém. No segundo a análise incorporaria questões como a propriedade, ou seja, quem é o dono do aplicativo e que interesses representa; a governança, traduzida por suas regras de utilização e, por fim, os modelos de negócio que a sustentariam ou permitiriam ao dono obter retorno financeiro a partir do crescimento do processo de adoção.

Para discutirmos a relação da abordagem de *big data* com as pesquisas onde há grande quantidade de informação disponível aprofundaremos a análise do vetor tecnológico que impacta a estruturação dos dados, as possibilidades de coleta e principalmente a própria essência dos objetos de interesse.

3.1 Vetor Tecnológico e Métodos Digitais

Rogers (2013) afirma que mesmo portando métodos tradicionais para o emprego em pesquisas ligadas ao digital, podemos, em algumas situações, estar utilizando um ferramental inadequado.

> 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 digi-

> tais, como uma prática de pesquisa, contribuem para o desenvolvimento de uma metodologia do próprio meio (ROGERS, 2013, E-book).[2]

A proposta de Rogers vai ao encontro do percurso que ora propomos partindo de uma visão do mundo contemporâneo onde o digital apresenta uma centralidade crescente, composto por entes com características específicas e, por isso, demandando também uma adequação ou extensão metodológica capaz de colaborar com pesquisas cujos objetos de alguma forma têm essa característica.

Desse modo, definimos métodos digitais como o conjunto de ferramentas, processos e abordagens de pesquisa que consideram a ontologia dos objetos de estrutura binária e as estruturas de redes por onde circulam, utilizando-se de recursos computacionais intensivos para coleta e análise de dados.

Tais soluções oferecem uma espécie de escala de utilização como representada no gráfico abaixo:

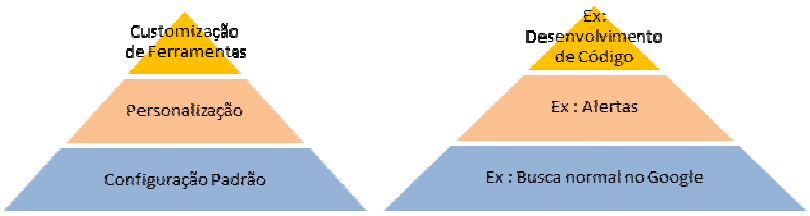

Figura 3: Representação da escala de utilização dos métodos digitais. Fonte: Santos (2016)

Tal escala vai da utilização de ferramentas e técnicas já existentes em sua configuração padrão num nível inicial; com ajustes a fim de personaliza-las para atender nossas necessidades específicas, num nível médio; ou ainda, num nível mais alto, através da criação de soluções baseadas em programação e desenvolvimento de código.

Nas pirâmides acima exemplificamos a escala numa situação de coleta de dados que utiliza a busca do Google, inicialmente com sua interface normal, depois a partir de uma solução com maior poder de personalização como os alertas[3] e por fim através de um código específico para coletar e armazenar esses dados.

[2] Tradução do autor.
[3] https://www.google.com/alerts

Em termos gerais, a abordagem que propomos resume-se às seguintes etapas:

Etapa 1 - Identificar a estrutura que contém os dados que precisamos. Algumas possibilidades apresentam-se com mais frequência:

a) Bases de Dados que permitem consultas amigáveis via preenchimento de formulários ou procedimentos simples. Exemplo: portais de transparência governamentais onde é possível requisitar dados sobre determinado tema e período.
b) APIs[4] que exigem requisições estruturadas no formato que estabelecem, ou seja, respeitando sua sintaxe própria. Exemplo: APIs do *Twitter* e do *Facebook* que precisam ou de uma aplicação específica para solicitar conteúdo, como os aplicativos que as acessam em nossos celulares, ou de um código customizado que consiga estabelecer tal diálogo e coletar as informações que a API entrega a partir de cada tipo de requisição.
c) Conteúdo disponível em páginas de internet que podem ser extraídos diretamente via técnicas de *scraping* (raspagem de dados). Como textos de matérias em portais jornalísticos ou tabelas e informações gerais publicadas, tais como previsão do tempo, cotação do dólar e resultados de competições esportivas.
d) Informações protegidas em ambientes fechados, acessadas apenas por usuários cadastrados e que contam com mecanismos de proteção como encriptação de dados e outros. Tais ambientes eventualmente podem ser acessados por técnicas de *hacking* que estão além do escopo deste texto.

Etapa 2 – Formatar a consulta ou requisição de dados alinhada ao tipo de repositório onde eles se encontram de acordo com as opções acima.

[4] 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.

Etapa 3 – Analisar os dados coletados a partir do processamento possível partindo do que foi efetivamente conseguido.

Podemos combinar então a escala de utilização da figura 3 com as diversas formas de estruturação de dados mais comuns numa tabela. A partir do cruzamento das quatro formas mais comuns de repositórios estruturados em bases de dados *online*, listamos algumas possibilidades de aplicação dos métodos digitais em seus três níveis.

Estrutura dos Dados	**Nível Inicial: Ferramentas Padrão**	**Nível Médio: Ferramentas com personalização**	**Nível Alto: Desenvolvimento de Código**
a) Bases de Dados e Repositórios Acessíveis	Solicitação de dados através da própria interface da base de dados, recebendo o resultado no formato padrão de entrega. Ex: Acesso à base SIDRA do IBGE e download do arquivo no formado do Excel ou em CSV.	Utilização de filtros e recursos de análise e visualização oferecidos pela plataforma, alterando a forma de entrega do resultado de acordo com as opções oferecidas. Ex: Uso das funções avançadas da SIDRA e geração de gráfico.	Código para automatizar o acesso ao banco de dados fazendo requisições sucessivas, customizadas, coletando e salvando os registros em outro tipo de estrutura ou formato de dados. Ex: Python com módulos Splinter ou Selenium
b) Servidores com acesso via API específica	Acesso via aplicação oficial da plataforma ou através de sua página web padrão. Ex: Uso do app do Facebook no celular ou acesso à pagina www.facebook.com .	Acesso através de aplicativos de terceiros que também acessam o servidor da plataforma mas oferecem funcionalidades adicionais. Ex: Node XL .	Código para acessar diretamente a API da plataforma coletando todas as informações disponibilizadas por ela e também fazendo requisições sucessivas capazes de coletar volumes maiores de dados.
c) Conteúdo em páginas web com padrão HTML[5]	Busca do Google, acesso manual e eventual coleta via CTRL+C e CTRL+V	Utilização de ferramentas específicas para scrapping. Ex: Portia	Desenvolvimento de código para coleta e análise. Ex: Python com módulo Beautiful Soup
d) Dados protegidos mediante acesso logado	Acesso via solicitação de cadastro e *login* normal.	Ferramentas de hacking geral tipo brute force ou engenharia social.	Desenvolvimento de códigos de invasão tipo worm ou trojan.

Tabela 2 – Matriz de possibilidades de coleta via métodos digitais em função da forma e local dos dados e dos níveis de aplicação. Fonte: do autor.

[5] *HiperText Markup Language*

4. *DISTANT READER* – EXERCÍCIOS DE APLICAÇÃO

Como exemplo de aplicação do caminho inspirado no conceito de *distant reader*, descrevemos a seguir um esforço anterior de pesquisa já documentado onde tal abordagem foi utilizada e possibilitou a identificação e comprovação empírica da velocidade com que as grandes redações jornalísticas brasileiras implementaram uma das principais características atribuídas ao jornalismo no ambiente digital : a atualização constante.

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 e ROHUMAA, 2011; 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, assumindo o papel de um *distant reader* . Como exemplo de aplicação acessamos o projeto da internet *WayBackMachine* – WBM (Fig. 4), também conhecido como *Internet Archive,* que se constitui de uma biblioteca digital de sites de internet com mais de 430 bilhões de páginas arquivadas.

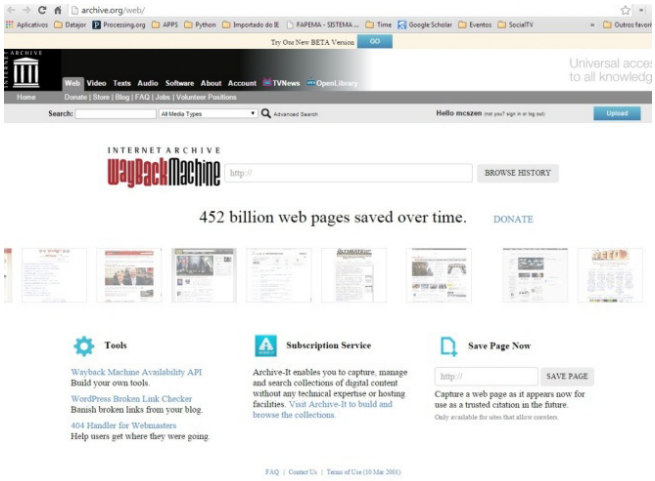

Figura 4: Tela principal do site Internet Archive. Fonte: Internet Archive (2014).

O objetivo desse experimento foi verificar com dados empíricos a característica da atualização constante atribuída ao jornalismo digital por muitos autores, numa afirmação normalmente baseada na potencial facilidade de publicar conteúdo possibilitada pelos atuais sistemas de CMS (*Content Management Systems*), sem contudo aferir isso

com dados. Aplicando a abordagem descrita acima nossa metodologia baseou-se na coleta e análise de todos os registros de alteração dos principais sites de conteúdo jornalístico brasileiro catalogados pelo *Internet Archive,* num recorte temporal desde o início desses registros para cada um dos sites até o ano de 2014.

Nos termos do caminho proposto, decidimos por uma solução de customização de ferramenta, num nível elevado de aplicação dos métodos digitais (ver figura 3), diante de um conjunto de dados estruturado na modalidade c) da etapa 1. Para isso desenvolvemos um código utilizando a linguagem de programação Python, capaz de analisar todo o conjunto de versões da página principal dos principais portais de conteúdo informativo do Brasil. Para seleção dos sites jornalísticos do nosso estudo utilizamos a classificação da plataforma Alexa[6] 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 pertenciam 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 .

O gráfico abaixo (Fig. 5) traz a plotagem da série temporal de atualizações extraídas do registro da WBM a partir dos dados do site www.ig.com.br em que foram contabilizados mais de 19 mil versões diferentes.

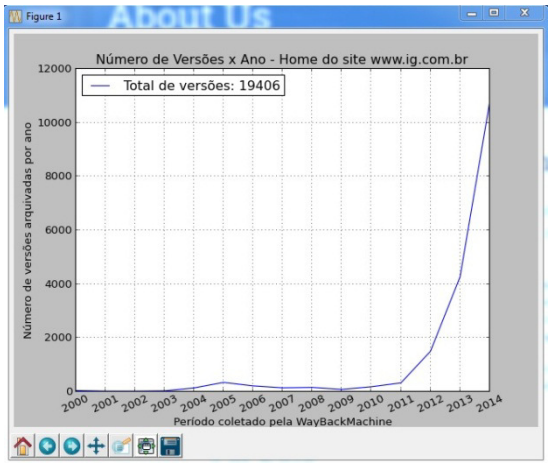

Figura 5 - Gráfico plotado com as atualizações registradas entre os anos de 2000 e 2014 do site www.ig.com.br . Fonte: do autor.

[6] www.alexa.com

As visualizações abaixo (Fig. 6) 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.

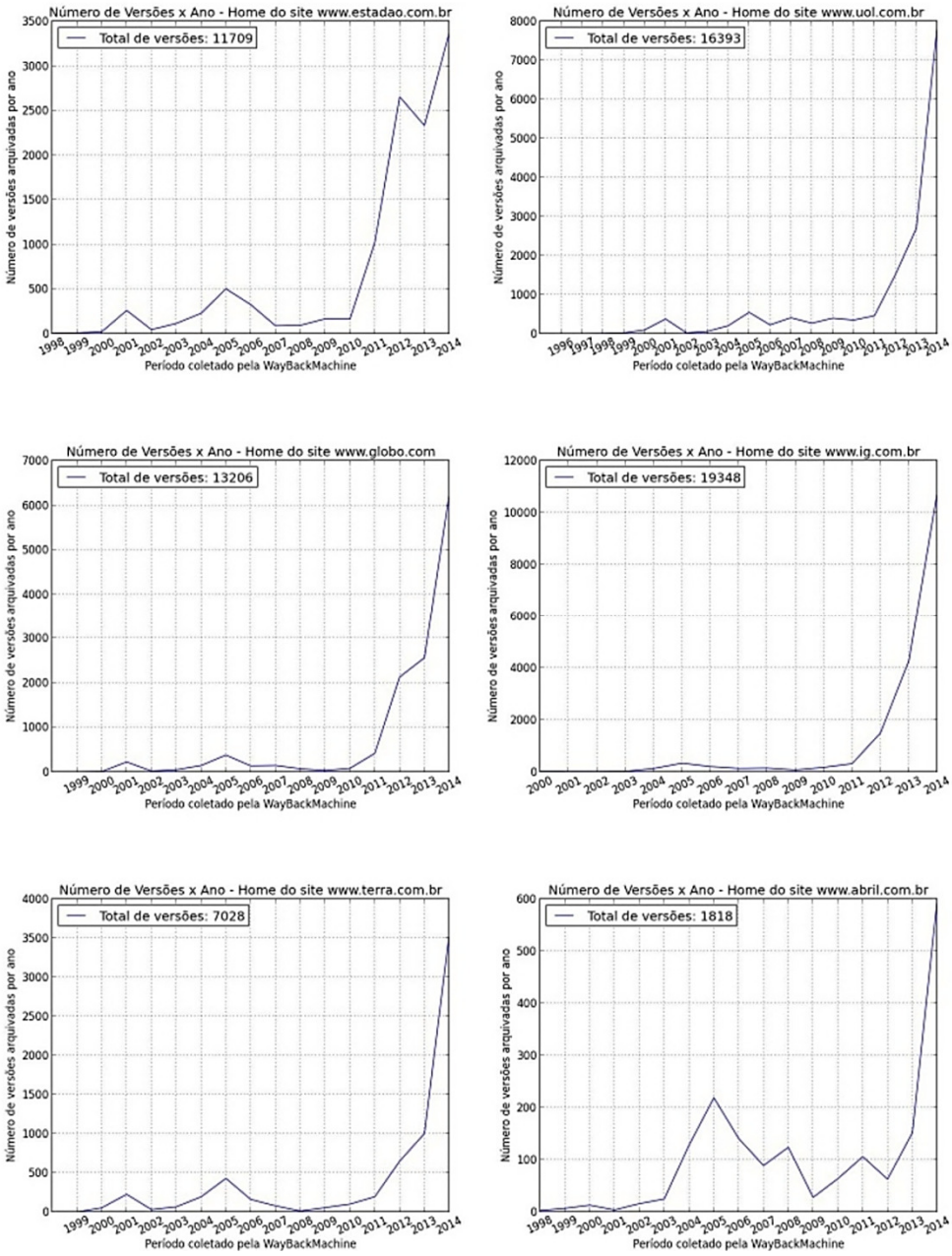

Figura 6 - Gráficos mostrando o crescimento dos números de atualizações a partir dos anos 2010 e 2011 nos principais sites jornalísticos brasileiros. Fonte: do autor.

Outro exemplo do potencial de operação e síntese gráfica de um código Python personalizado pode ser observado a partir de uma planilha com mais de 128 mil linhas contendo a classificação Qualis Capes de todas as revistas acadêmicas registradas em todas as 48 subáreas de conhecimento.

Com apenas 6 linhas de código é possível carregar a planilha , selecionar apenas as revistas de Comunicação e Informação e gerar um gráfico representando a quantidade de publicações por cada faixa de classificação ou estrato mostrando que existem mais de 400 revistas B5 nessa subárea e menos de 100 no estrato mais qualificado A1.

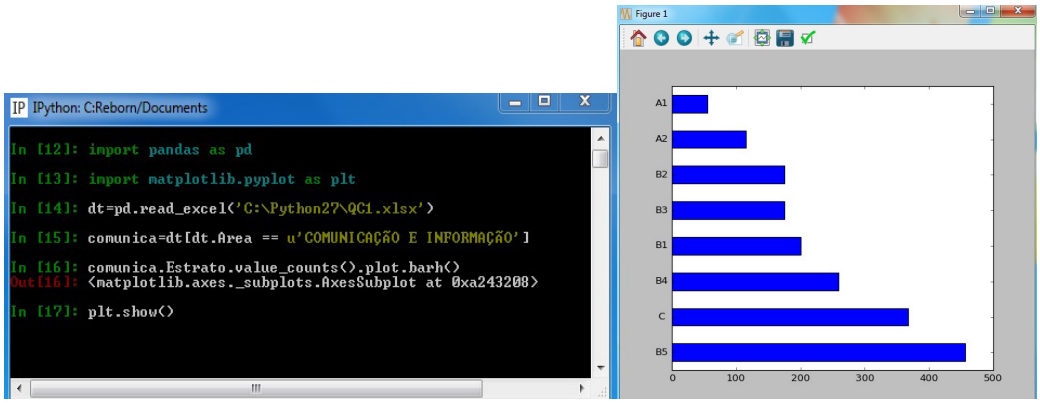

Figura 7 - Exemplo de código customizado em Python para processar e gerar uma visualização a partir de uma planilha com mais de 120 mil linhas. Fonte: do autor.

5. CONSIDERAÇÕES FINAIS

Dentro do conceito de *distant reader*, que nada mais é do que uma implementação das técnicas de *big data*, foram analisados no caso dos sites jornalísticos brasileiros mais de 90 mil registros ou versões das páginas iniciais o que seria difícil ou praticamente impossível de ser feito mesmo por uma equipe relativamente grande dedicada à tarefa, principalmente pelos custos que envolveriam uma empreitada assim.

A transformação dos dados em visualizações simples tornou perceptível a aceleração entre as versões, ou seja, a efetivação de uma política de atualização mais ágil do conteúdo nas páginas principais, fato que também, utilizando a metodologia que propomos, poderia ainda ser investigado mais profundamente através de uma análise de

contexto para identificar o impacto dos vetores econômico, tecnológico e cultural nesses eventos.

No exemplo da planilha com as revistas acadêmicas da CAPES foram processados mais de 120 mil registros com um pequeno código onde conseguirmos facilmente observar a divisão das revistas pelos estrados de classificação.

Nos dois exemplos seguimos as etapas de identificar a estrutura onde os dados estão inseridos e a partir dai escolher dentro dos níveis de complexidade dos métodos digitais que listamos a variante mais adequada para coletar e tratar as massas de dados objetos dos estudos.

Tais procedimentos tem aplicabilidade em grande número de casos onde o excesso de dados disponíveis pode intimidar o pesquisador e induzi-lo eventualmente ao erro de observar amostras com baixa relevância e pequeno potencial de apreensão, em cenários onde a complexidade das inter-relações pode não ser percebida usando os métodos tradicionais.

É óbvio que o caminho ora proposto de forma alguma invalida ou contraria as formas tradicionais de investigação nas Ciências Sociais e Humanidades constituindo-se apenas em uma extensão possível e de utilidade verificável em casos específicos. No jornalismo, a aproximação com tais métodos começou dentro da atividade profissional, principalmente nos trabalhos ligados ao jornalismo investigativo e ao chamado jornalismo guiado por dados. A atenção para esses procedimentos por parte da academia nos parece relevante considerando o grande conjunto de questões de pesquisa e objetos inseridos em contextos de excesso de informação e com características digitais que não podem ser esquecidas.

A aplicação desse caminho de pesquisa além de trazer resultados de difícil alcance por outros meios ainda permite potencialmente enfrentar novas questões de investigação ou ainda submeter questões já abordadas a análises diferentes, caminho recomendável em toda investigação científica.

REFERÊNCIAS

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

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

_____.; ROHUMAA, Liisa. **The online journalism handbook**: skills to survive and thrive in the digital age. Essex: Pearson Education, 2011.

BRUNS, Axel. Faster than the speed of print: reconciling big data social media analysis and academic scholarship. In: **First Monday – Peer Reviewed Journal On The Internet.** Volume 18, Nº 10, 2013. Disponível em http://firstmonday.org/article/view/4879/3756 . Acessado em 23/07/2017 .

CODDINGTON, Mark. Clarifying Journalism's Quantitative Turn in: **Digital Journalism**, 3:3, 331-348, DOI: 10.1080/21670811.2014.976400 , 2015. Disponível em http://dx.doi.org/10.1080/21670811.2014.976400 . Acessado em 21/07/2017.

CORDEIRO, William. **Infografia interativa na redação:** o exemplo do Diário do Nordeste. Mossoró, RN: Sarau das Letras, 2013.

FRANCISCATO, Carlos. A Inovação Metodológica Como Problema Na Pesquisa Em Jornalismo Digital. In: **Contemporânea – Revista de Comunicação e Cultura.** Vol. 15, nº 1, 2017, págs. 25-46. Disponível em https://portalseer.ufba.br/index.php/contemporaneaposcom/ . Acessado em 23/07/2017 .

GONÁLEZ-BAILÓN, Sandra. **Social Science in the era of Big Data**. Penn Libraries, 2013. Disponível em https://pdfs.semanticscholar.org/6e78/b1133713cb17aabbc3bf421a6e51bc538eca.pdf . Acessado em 23/07/2017;

INTERNET ARCHIVE. Digital library of millions of free books, movies, software, music, websites, and more. 2014. Disnível em: <https://archive.org/index.php>. Acesso em: 7 maio 2015.

LEWIS, Seth; WESTLUND, Oscar. Big Data and Journalism **–** Epistemology, expertise, economics and ethics**.** In: **Digital Journalism**. Vol. 3, 2015, pags. 447-466.

LIMA JR. , Walter. Big data, jornalismo computacional e data jornalismo**:** estrutura, pensamento e prática profissional na Web de dados. In: **Estudos em Comunicação** nº 12, págs. 207 a 222. Covilhã : UBI, 2012. Disponível em http://www.ec.ubi.pt/ec/12/pdf/EC12-2012Dez-11.pdf . Acessado em 21/07/2017.

MACHADO, Elias. As limitações metodológicas nas pesquisas em Jornalismo: um estudo dos trabalhos apresentados no GT de Jornalismo da Associação Nacional de Pós Graduação em Comunicação. In: ENCONTRO NACIONAL DE PESQUISADORES EM JORNALISMO, 10. 2012. **Anais...** Curitiba, 2012. Disponível em <http://soac.unb.br/index.php/ENPJor/XENPJOR/paper/view/2146>. Acesso em: 26 jan. 2015.

_____.; ROHDEN, J. Metodologias de Pesquisa Aplicadas ao Jornalismo: Um estudo dos trabalhos apresentados na SBPJor (2003-2007). **Brazilian Journalism Research.** v. 12, n. 1, p. 228-245. 2016.

MAHRT, Merja; SCHARKOW, Michael. The value of Big Data in Digital Media Research. In: **Journal of Broadcasting & Eletronic Media.** Vo. 57, 2013. Disponível em: http://www.tandfonline.com/doi/abs/10.1080/08838151.2012.761700 . Acessado em 23/07/2017.

MANOVICH, L. **The language of new media.** Massachusetts: Mit Press. 2001.

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

RODRIGUES, Adriana Alves. **Infografia Interativa em Base de Dados no Jornalismo Digital**. 130f. Dissertação (Mestrado em Comunicação) – Universidade Federal da Bahia, Salvador. 2009.

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

WELSH, BEM . **Análise MOOC Knight Center Python for Journalists** . 2017. Disponível em https://github.com/california-civic-data-coalition/python-calaccess-notebooks/blob/master/project-management/mooc-students.ipynb . Acessado em 19/07/2017.

SANTOS, MARCIO. **Comunicação Digital e Jornalismo de Inserção –** Como big data, inteligência artificial, realidade aumentada e internet das coisas estão mudando a produção de conteúdo informativo. Disponível em: https://drive.google.com/file/d/0BwblN2uXiXNjQnNMOFFUQjc2enM/view . Acessado em 21/07/2017.

VAN DICK, José. **The culture of connectivity: a critical history of social media**. [E-book]. New York: Oxford Press, 2013.