UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

CENTRO DE ARTES

DEPARTAMENTO DE COMUNICAÇÃO SOCIAL

JOHANNA INÁCIA HONORATO

JORNALISMO E VISUALIZAÇÕES DE DADOS: METODOLOGIA, QUESTÕES E DESAFIOS

VITÓRIA

2016

JOHANNA INÁCIA HONORATO

JORNALISMO E VISUALIZAÇÕES DE DADOS: METODOLOGIA, QUESTÕES E DESAFIOS

Trabalho de Conclusão de Curso apresentado ao Departamento de Comunicação Social do Centro de Artes da Universidade Federal do Espírito Santo, como requisito parcial para obtenção do grau de Bacharel em Jornalismo.

Orientador: Prof. Dr. Fábio Gomes Goveia

VITÓRIA

2016

BANCA EXAMINADORA

_____

Prof. Dr. Fábio Gomes Goveia

Orientador

_____

Profª. Drª Ruth de Cássia dos Reis

_____

Prof. Dr. Rafael Paes Henriques

Aprovado em __ de _____ de 2016

AGRADECIMENTOS

Primeiramente, Fora Temer.

Quando pensei que 2015 tinha sido avassalador, veio 2016 e resolveu ser o ano do encerramento de um ciclo. A graduação vai deixar saudades e sinto um aperto no peito quando vejo o inevitável adeus que darei ao portal do aluno, à minha matrícula 2012101103, ao "status" de graduanda em uma universidade federal, às bolsas PAD e afins.

Posso dizer que definitivamente entrei verde no curso. 16 anos de vida e de repente uma montanha de oportunidades e responsabilidades surgiram. Como qualquer calouro que se preze, a vontade era de fazer tudo o que fosse possível: ir em todos os congressos, todas as palestras, ganhar todos os certificados, pegar todas as optativas, ler todos os livros das disciplinas (obrigatórios e extras), entrar em todos os grupos de estudos. E como todo calouro que se preze perceber, logo no segundo período, que tudo o que você quer é não ter aula 7h da manhã, que o cardápio do dia do RU não seja frango ou linguiça, e algumas horas a mais de sono no CA.

Logo no primeiro período me foi ofertada a possibilidade de ingressar em um projeto de pesquisa. Passei na prova dos Jovens Talentos e caí de paraquédas no laboratório que iria se tornar minha segunda casa, minha segunda família. Entrei na linha de pesquisa em imagem e com o passar do tempo fui me descobrindo interessada nas discussões de cibercultura, big data e visualização, realizando minhas duas Iniciações Científicas nessa mesma área e culminando nesse Trabalho de Conclusão de Curso.

Em primeiro lugar, gostaria de agradecer à Deus, que nunca me falhou quando eu mais precisava. Me deu conforto nos momentos de desespero, me deu forças quando eu achava que não ia conseguir, me deu paz quando eu finalmente atingia meus objetivos. Obrigada.

Em segundo lugar, à minha família. Se eu estou onde estou é porque devo absolutamente tudo a eles. Eles compreenderam os caminhos que escolhi trilhar e sei que, por onde eu for, eles estarão sempre junto de mim. Obrigada.

Em terceiro lugar, ao meu orientador Fábio Goveia. Desde o primeiro período como sua orientanda, agora no final do curso não podia ser diferente. Esse mês de julho foi

repleto de orientações surpresas, leituras complementares, mensagens incessantes de revisão de texto e prazos de uma semana. Obrigada por aceitar me orientar e confiar na minha capacidade enquanto graduanda e pesquisadora do laboratório.

Agradeço ao Laboratório de Estudos Sobre Imagem e Cibercultura, por ter viabilizado toda a minha trajetória de pesquisa dentro da universidade e me dado respaldo técno-científico para produzir, publicar e apresentar meus trabalhos nos congressos e eventos de comunicação. Aos seus integrantes que, além de colegas de trabalho, se tornaram amigos que levarei para a vida toda: Tasso, o qual permanece sendo meu melhor amigo, aguentando meus surtos e paranoias, trazendo chocolate quando eu estava na bad; Veronica, minha companheira de imagem que foi capturada pelo time da modelagem, mas nem tanto assim; Lia, mentora dos artigos, inspiração de vida acadêmica, agradeço por botar pressão em manter meu Lattes atualizado; Willian e Andrei, por criarem o ImageCloud e o Leticia, possibilitando assim a maioria das pesquisas, e me aturarem perguntando mil coisas sobre os scripts ou pedindo ajuda porque deu ruim; Milena, por ajudar uma graduanda desesperada e insegura com a escrita acadêmica; Marianne, pelos conselhos de vida maravilhosos que me ajudaram a me tornar um ser superior; Prof. Fábio Malini, por ter idealizado o laboratório e incentivado as pesquisas de iniciação científica, monografias e dissertações; Prof. Patrick Ciarelli, por desenvolver o Crawler e sempre ter uma solução pra qualquer problema informático que tínhamos; e Profª. Adriana Ilha, por administrar esse povo todo e garantir o pão nosso de cada dia.

Por fim, agradeço às coisas simples da vida universitária: as idas à Cantina do Onofre, o sagrado dia da bolsa cair na conta, o bom cardápio do dia do RU, os pufes do Labic, a saída de sexta pra pracinha de Jardim da Penha, as estadias com os amigos nos hosteis desse Brasil, os bolos dos aniversários comemorados, os pdfs compartilhados na internet. São esses pequenos prazeres que fizeram da graduação um momento único e eterno.

O dilúvio informacional jamais cessará. A arca não repousará no topo do monte Ararat. O segundo dilúvio não terá fim. Não terá nenhum fundo sólido sob o oceano das informações. Devemos aceitá-lo como nosssa nova condição. Temos que ensinar nossos filhos a nadar, a flutuar, talvez a navegar.

Pierre Lévy

RESUMO

O presente trabalho ilustra a importância das visualizações de dados imagéticos coletados de sites de redes sociais, referentes a um determinado acontecimento ou fenômeno, juntamente ao trabalho dos profisisionais do jornalismo. Para isso, foi realizado um estudo de caso com imagens publicadas no Instagram que possuíam hashtags vinculadas à crise de epidemia mundial de Zika, Chikungunya e Dengue.

A divisão do trabalho se deu em três capítulos: no primeiro, foi montado o arcabouço teórico com leitura de bibliografia a respeito do surgimento da cibercultura, passando pelo fenômeno contemporâneo do Big Data e culminando nas formas de se visualizar os dados vindos dos sites de redes sociais. No segundo, a metodologia de pesquisa em imagem do Laboratório de Estudos sobre Imagem e Cibercultura é colocada de forma a mostrar a evolução, desdes os primeiros datasets até o panorama atual de análise, com os scripts criados e pesquisas realizadas.

Por fim, o terceiro capítulo traz o estudo de caso, com sua metodologia de coleta e dois tipos de visualizações: ImageCloud e plotagem em mapa. A partir delas foi possível realizar uma análise dos tipos imagéticos mais frequentes ao tema, bem como uma análise regional baseada na contextualização de cada local delimitado pela pesquisa.

Palavras-chave: cibercultura; Big data; data visualization; Instagram; Zika

LISTA DE IMAGENS

SUMÁRIO

INTRODUÇÃO

Este trabalho de conclusão de curso é resultado de todo um caminhar de pesquisa realizado durante a graduação enquanto pesquisadora do Laboratório de Estudos sobre Imagem e Cibercultura da Ufes, mais especificamente na linha de que estuda as imagens publicadas e compartilhadas no meio digital. Por meio do projeto "Visagem", com orientação do Prof Fábio Goveia, buscou-se estudar e compreender o comportamento dos elementos imagéticos onlines, publicados e compartilhados em escala exponencial, além de pensar formas de tornar esse montante de dados visíveis. Conciliando os estudos da área da Comunicação com os estudos das Ciências Exatas, foi possível utilizar ferramentas já disponíveis e criar novas para que esse montante de dados pudesse se tornar visível ao pesquisador.

Neste trabalho, buscou-se, usando da metodologia de pesquisa em imagem desenvolvida pelo Labic, apresentar o quão necessário é visualizar os dados que advém de sites de redes sociais por meio de visualizações de dados (*data visualization)*. Como dito anteriormente, os dados contemporâneos são gerados em grande fluxo por qualquer pessoa que tenha à mão intrumentos que permitam sua conectividade à internet e, de modo específico nesse trabalho, que sejam também dispositivos fotográficos como smartphones ou câmeras digitais. Além disso, esses dados trazem opiniões de determinado grupo de pessoas, criam panoramas imagéticos do discurso de lugares delimitados e contam histórias de acontecimentos como no caso do estudo de caso que é trazido à discussão: a epidemia mundial de Zika, Chikungunya e Dengue vista pelas publicações do Instagram.

A hipótese apresentada é que os dados precisam ser visualizados para que se retornem para o pesquisador o máximo de informações acerca de um determinado fenômeno ou acontecimento. Existem padrões e informações escondidas que só são revelados quando se tornar possível plotar todo o *dataset* coletado em forma de gráficos e mapas, utilizando parâmetros de ordenação baseados nos metadados das imagens. Fala-se em pesquisador, mas esse mesmo caminho pode e deve ser trilhado pelo profissionais do Jornalismo na busca e apuração de pautas em potencial vindas desse cenário digital. Com o auxílio das visualizações de dados, o trabalho jornalístico se

enriquece e consegue se aventurar para além dos meios tradicionais de se buscar notícias.

Para sustentar essa ideia, o estudo de caso utilizado nesse trabalho traz uma análise de conteúdo acerca dos tipos de imagens vinculadas às diveras hashtags sobre a epidemia mundial de Zika, Chikungunya e Dengue. A composição do dataset foi possível graças a um script denominado Leticia, criado no Labic, que coleta conteúdo publicado no site de rede social Instagram. Foram escolhidas 18 tags referentes a esse tema, como por exemplo “forazika”, “dengue” e “chikungunya”. O intervalo de tempo delimitado para coleta foi de 13 meses (março de 2015 a março de 2016) dividido em quatro períodos de tempo: março a maio; junho a agosto; setembro a novembro; e dezembro – 2015 a março – 2016. Referente a esse período foram capturadas 66.405 mídias, entre imagens e vídeos, visualizadas no formato de Nuvem de Imagens (*ImageCloud*) e as que possuíam geolocalização foram plotadas em mapa usando a ferramenta online CartoDB.

A estrutura desta monografia é composta por três capítulos: contextualização histórica e conceituamento teórico utilizado ao longo do texto; desenvolvimento da metodologia de pesquisa em imagem do Labic; e exposição do estudo de caso sobre a epidemia mundial de Zika.

O capítulo 1 trata do arcabouço teórico utilizado para dar base às hipóteses e ideias propostas no trabalho. O estudo perpassa pela criação do ciberespaço, retomando brevemente a história do desenvolvimento da Internet e a revolução da microinformática (LEMOS, 2015), e o surgimento da cibercultura e seus motores (LÉVY, 1999). Com o ambiente definido, passamos ao fenômeno contemporâneo do *Big data* exibindo o conceitos dos três V’s formulado pela empresa Gartner, e o conceito acadêmico formulado por boyd e Crawford (apud VIS, 2012). Nessa discussão Farida Vis (2012), propõe uma reformulação dos três V’s incorporando a procupação que os pesquisadores necessitam ter com o dado que trabalham. Em se tratando se visualização, o autor Lev Manovich (2012) procura distinguir os tipos de visualização que existem, formula a hipótese de que as formas geométricas dos primeiros gráficos ainda influenciam as visualizações atuais, e propõe a ideia da *media visualization* (ou visualização direta).

O capítulo 2 se resume em descrever a metodologia de pesquisa criada e desenvolvida pelo Labic para adaptar-se ao estudo de imagens. Uma linha histórica que

vem desde as manifestações de junho de 2013, referenciadas pelas hashtags #protestoes, #passelivre e #vemprarua, passando pela criação das ferramentas ALICE, AISI e *ImageCloud*, desenvolvimento de um aplicativo sobre a Copa do Mundo de 2014, experimentação com mapas do CartoDB, e culminando em possibilidades de visualizações como grafos de imagens no Gephi.

Já o capítulo 3 se refere ao estudo de caso sobre a epidemia mundial do Zika Vírus. Nesta etapa apresentamos o conceito de sites de redes sociais de Boyd e Ellison (2007); uma breve classificação de tipos de sites trazida por Recuero (2009); a reflexão de Rubinstein e Sluis (2008) sobre a massificação amadora da fotografia; e dois tipos de visualização: ImageClouds e plotagem em mapa do CartoDB. Sobre os *ImageClouds* é feita uma análise dos tipos imagéticos presentes na coleta, assim como os tipos mais frequentes (no caso, com maior quantidade de likes) no dataset, buscando contextualizar com o panorama da época. O mesmo é feito com os mapas, entretanto o fato de ser possível georreferenciar as imagens abre margem para uma análise por região, visualizando as relações entre as fotos postadas, o contexto local e a reação dos usuários perante a epidemia.

Com a análise feita, segue uma reflexão acerca da posição do Jornalismo nesse ambiente dinâmico do ciberespaço, e de como a ação de visualizar os dados pode revelar informações pertinentes sobre determinados assuntos ou acontecimentos. Para isso foi utilizado referencial em Aroso e Correia (2008), que afirmam que cada usuário agora é um transmissor de informação em potencial, e Bertocchi (2014) que apresenta como os jornalistas e canais de comunicação devem buscar o aprendizado de práticas computacionais para expandir os horizontes da profissão.

1. Contextualização teórica: Cibercultura, sociedade dos dados e visualização

O início do século XXI foi (e ainda é) intensamente marcado pelos avanços tecnológicos, pela facilidade dos internautas em se conectar com outros usuários de diferentes partes do mundo e pelo montante gigantesco de informação que se produz a cada momento nesse universo digital. A internet como conhecemos hoje é muito mais do que somente uma infraestrutura de computadores mas se tornou um meio de produção e distribuição de conteúdo, hiperconectada e alimentada por diversas redes presentes nela, como empresas, associações, universidades e mídias clássicas. Ela foi o meio pela qual foi possível o surgimento do ciberespaço e consequentemente da cibercultura, levando milhões de internautas a produzirem e compartilharem informações.

Essa sociedade contemporânea está cada vez mais imersa na cibercultura, podendo desfrutar dos constantes avanços tecnológicos, do fácil acesso a dispositivos eletrônicos que possibilitem a produção de informação (computadores pessoais, smartphones, tablets) e da interação vivida dentro dos sites de redes sociais. Nesse contexto de produção de conteúdo em larga escala, a pesquisa em Big data faz parte da considerada "futura revolução científica" (LÉVY, 2016) na qual uma quantidade absurda de dados sobre a atividade humana estão disponíveis na rede e o poder de processamento usado para coletá-los e analisá-los vem crescendo exponencialmente. Entretanto é preciso conceituar o que seria esse chamado Big data para assim poder entender e estudar a melhor forma de lidar com as informações presentes nele.

Há cada vez mais dados disponíveis para se estudar, mas não é fácil enxergá-los em meio ao fluxo veloz no qual eles são produzidos. Uma hashtag, um termo ou um assunto que se tornam populares podem atingir números impressionantes de *tweets*, posts, ou imagens relacionados a eles. Há a opção de recorte por parte do pesquisador, mas e quanto à visualização do todo? Não basta apenas ver, mas é preciso saber quais as opções que giram em torno desse ato de olhar, e é isso que as visualizações possibilitam. De acordo com o Michaelis Dicionário Brasileiro da Língua Portuguesa, visualizar significa "tornar algo ou alguém visual ou visível" ou ainda "formar uma imagem mental do que não existe".

1.1. O advento da Internet e o surgimento da cibercultura

A ideia de que a Internet seria uma "supervia de informação" fica clara no que Kahn e Cerf (1999) acreditam ser a melhor definição para o termo. Aprovada em 1995 pelo *Federal Networking Council* (FNC), consta na definição que:

> "Internet" refere-se ao sistema de informação global que
>
> i) é logicamente ligado entre si por um espaço de endereço global exclusivo com base no Protocolo de Internet (IP) ou suas extensões subsequentes;
> ii) é capaz de suportar comunicações usando uma série de Protocolos de Controle de Transmissão/Protocolos de Internet (TCP/IP) ou suas extensões subsequentes, e/ou outros protocolos compatíveis com o IP; e
> iii) fornece, utiliza ou torna acessível, tanto público quanto privado, serviços de alto nível nas camadas de comunicações e infraestruturas relacionadas descritas aqui. (Tradução própria. *Federal Networking Council*. 1995)[1]

Para eles, essa definição deixa claro a Internet enquanto organismo dinâmico, quadro para numerosos serviços e meio para a criatividade e inovação. Todavia, reduzir a Internet a uma "auto-estrada eletrônica" (LÉVY, 1999. p.124) feita de fios de cobre ou fibras óticas, não contemplaria todo o movimento de interação e relações construídas entre os usuários. A cibercultura expressa o surgimento de um novo universal sem intenção de ser totalizante. Nesse segundo dilúvio, o dilúvio informacional que segundo LÉVY (1999) não terá fim, haverá inúmeras arcas com suas próprias totalidades navegando sem parar. Não há fundo sólido sob o oceano de informações; não há repouso no Monte Ararat.

Unindo a infraestrutura técnica e material descrita na definição do FNC (1995), e o universo oceânico de informações junto com os seres humanos que navegam nele proposto por Lévy (1999), surge um novo meio de comunicação designado de "ciberespaço". O termo foi cunhado por William Gibson em 1984, no seu romance de ficção científica *Neuromancer*, e foi apropriado pelos usuários da internet e criadores de redes digitais. Lévy (1999) define ciberespaço como "espaço de comunicação aberto pela interconexão mundial dos computadores e das memórias dos computadores", que se aproxima da definição dada por Dyson et al (apud LÉVY, 1999) como "terra do saber" ou "a nova fronteira". A cibercultura seria então "o conjunto de técnicas (materiais e intelectuais), de práticas, de atitudes, de modos de pensamento e de valores que se desenvolvem juntamente com o crescimento do ciberespaço".

[1] Disponível em: <https://www.nitrd.gov/fnc/Internet_res.aspx>.

De acordo com Lévy (1999), os três princípios que nortearam o crescimento inicial do ciberespaço e que se tornaram motores dele são a interconexão, a criação de comunidades virtuais e a inteligência coletiva. A *interconexão* se refere ao preferível ato do usuário em se conectar em vez do isolamento; não haveriam mais canais específicos para veicular informação, pois todo o ciberespaço seria um grande canal interativo e sem fronteiras. As *comunidades virtuais* seriam "grupos de pessoas se correspondendo mutuamente por meio de computadores hiperconectados", formados a partir de gostos em comuns, afinidades, sem haver o problema de distâncias geográficas. Seus participantes atuam na reciprocidade, compartilhando conteúdos que interessam a todos os outros integrantes da comunidade, o que não seria possível antes do surgimento do ciberespaço propício pra esse tipo de interação. Já a *inteligência coletiva* é o que Lévy toma como a perspectiva espiritual: a interconexão entre os participantes das comunidades virtuais e entre as diversas comunidades existentes online proporciona a criação e o compartilhamento de conteúdo incessante que alimenta o conhecimento e a inteligência de todos. Juntos constroem continuamente o saber que circula pelas redes existentes.

Entretanto, a existência do ciberespaço e da cibercultura não começou de imediato. Entre 1940 e 1970, a idéia de uma rede de computadores conectados era voltada essenciamente para o contexto militar, não sendo à toa que a ARPANET, considerada a "mãe" da internet atual, foi criada pelo Departamento de Defesa dos Estados Unidos em 1969. Foi em meados dos anos 70 que a microinformática começa a se popularizar e é junto dela que o ambiente propício a cibercultura começa a surgir, tornando-se "mais que um desenvolvimento linear da lógica cibernética, surgindo como uma espécie de movimento social" (LEMOS, 2015).

Com o lema *Computadores para o povo*, a microinformática foi inventada pelos radicais californianos tendo como meta "lutar contra a centralização e posse da informação (e consequentemente do destino da sociedade informatizada) pela casta científica, econômica, industrial e militar" (BRETON apud LEMOS, 2015). Segundo Breton (apud LEMOS, 2015), ela seria advento de dois eventos marcantes no século XX: o avnanço das tecnologias digitais, com a diminuição do tamanho das peças e aumento do poder de processamento e memória; e a atitude *cyberpunk* de cunho técnico-anarquista. Nesse contexto, por exemplo, surge o primeiro computador Macintosh, criado por Steve Wozniak e Steve Jobs, que rompia com todo o caráter

militar proposto por empreendimentos como a IBM, e trazia à tona a possibilidade da popularização dos computadores aos usuários comuns e sua utililização para outros fins que não bélicos.

> A democratização dos computadores vai trazer à tona a discussão sobre os desafios da informatização das sociedades contemporâneas, já que esses não só devem servir como máquinas de calcular e de ordenar, mas também como ferramentas de criação, prazer e comunicação; como ferramentas de convívio. (...) Como sabemos a sociedade não é passível à inovação tecnológica, sendo o nascimento da microinformática um caso exemplar, mostrando a apropriação social das tecnologias para além da sua funcionalidade econômica ou eficiência técnica. (LEMOS, 2015. p. 104)

A microinformática também transforma o perfil de quem utiliza ou quer utilizar os aparelhos tecnológicos que estão surgindo. No início, era preciso que o usuário fosse alguém com domínio da área da informática: um analista, um programador ou um matemático. Com a proposta da democratização da rede dos computadores, passa-se da lógica do especialista para a lógica do usuário amador, principalmente com os avanços em dispositivos de interface, a criação da "área de trabalho" com suas janelas e menus, e inserção de anexos como mouses e teclados.

Com isso passa-se aos anos 90 e ao início do século XXI, no qual a quarta fase da informática se estabelece como sendo a dos internautas e dos computadores conectados (BRETON apud LEMOS, 2015). Nela será proposta a conexão generalizada, ou a interconexão descrita por Lévy (1999), com a proliferação de hipertextos[2], a criação das comunidades virtuais, o subsequente surgimento de sites de redes sociais, e a possibilidade de que cada indivíduo conectado em rede seja produtor e consumidor ativo de quaisquer conteúdo disponível.

Com uma conexão em rede democratizada e com o fluxo de informação seguindo o conceito de "todos-todos" (LÉVY, 1999), a quantidade de conteúdo gerado seguiu a lógica de crescimento exponencial e surgiu um novo campo de estudo e análise desses novos tipos de dados: o Big data.

1.2. A sociedade dos dados e produção de Big data

A discussão acerca do Big data se faz presente no trabalho de vários pesquisadores como Tufekci que diz que "*Big data* é um agregado de bancos de dados, em grande

[2] Hipertexto é um texto em formato digital, reconfigurável e fluido. Ele é composto por blocos elementares ligados por links que podem ser explorados em tempo real na tela. (LÉVY, 1999)

escala, de impressões acerca das atividades online e em mídias sociais" (TUFEKCI, 2013) e Diebold (2012), que considera "*Big data* não apenas como um termo seguramente estabelecido e um fenômeno contínuo, mas também uma disciplina que emerge". Essas duas definições deixam claro que além de ser uma denominação para "grande volume de dados", esse *Big data* reflete o estado atual no qual a sociedade se encontra: produção em massa de informação pelas mãos de pessoas "comuns" e como essas informações não estão alheias umas das outras, mas se comunicam, se conectam e se complementam em rede.

Adentrando um pouco mais nesse universo do *Big data*, duas outras definições propõem caracterizar mais especificamente esse tipo de dado com o qual estamos trabalhando: uma delas vem do trabalho de Danah Boyd e Kate Crawford (apud VIS, 2013), sob uma perspectiva mais acadêmica trazendo para a discussão inclusive a questão da mitologia que envolve essa aparente mina de ouro de informações; e a segunda definição, a qual sofre influência da indústria, tratando os dados de forma mais objetiva preocupando apenas com questões tecnológicas como armazenamento e processamento.

Boyd e Crawford (apud VIS, 2013) propõem três aspectos que são tidos como essenciais ao se definir exatamente o que é *Big data*. São eles:

> 1 – **Tecnologia:** maximizando o poder computacional e a precisão algorítmica para coletar, analisar, linkar e comparar grandes datasets.
>
> 2 – **Análise:** Utilizando grandes datasets para identificar padrões de modo a realizar alegações econômicas, sociais, técnicas e legais.
>
> 3 – **Mitologia:** crença generalizada de que grandes datasets oferecem uma forma superior de inteligência e conhecimento que possibilita o surgimento de insights, que seriam previamente impossíveis, cercado de uma aura de verdade, objetividade e precisão. (Tradução própria: BODY e CRAWFORD apud VIS, 2013)

Essa sensibilidade em trazer a questão do mito ao se tratar da expectativa perante tantos dados públicos disponíveis é o ponto chave da diferença entre essa visão e a visão mais industrial do termo, que será tratada logo adiante. Nesse contexto, a autora Farida Vis (2013) utiliza o conceito de mito de Roland Barthes (1993) no qual a "função chave de um mito é naturalizar crenças que são contingentes, tornando-as invisíveis e, portanto, alheias a questionamentos" para colocar um questionamento plausível sobre a forma com a qual obtemos os dados. Certamente esses dados são disponibilizados a nós

pesquisadores por meios das API[3]'s pelas próprias empresas que possuem os dados, mas quais os critérios usados para determinar as informações presentes nas API's? Seria a API um grande mito contemporâneo, imposto a nós como único modo de se obter os dados? Vis coloca que isso é um fator importante em se ter em mente quando se realiza pesquisas com esses grandes volumes de dados: não tratar de modo cético, mas estar consciente de que os campos passíveis de coleta possam ser apenas uma parte de um arquivo mais completo.

Já a respeito do olhar mais industrial sobre o *Big data*, uma definição amplamente utilizada no campo científico e tecnológico é a estabelecida pela *Gartner.Inc*, conceituada empresa americana que presta consultoria e fornece informações relacionada à Tecnologia da Informação (TI). Nesse caso, o termo é utilizado para descrever grandes volumes de informações gerados pela sociedade atual e que aumentam de forma exponencialmente a cada dia:

> *Big data* são ativos de informações que contêm grande volume, grande velocidade e grande variedade, exigindo formas de processamento inovadoras e de custo efetivo, proporcionando assim uma melhor percepção e tomada de decisão acerca dos resultados[4]. (Tradução própria. GARTNER. Acesso em: 01 jul. 2016)

Apesar de curta, essa definição abre margem para a explicação dos três aspectos citados por ela: os famosos "três Vs". São eles o "volume", a "velocidade" e a "variedade" das informações que são geradas.

O "volume" trata da escala massiva de crescimento dos dados não-estruturados (emails, *tweets*, posts no *Facebook*, vídeos, geolocalização, comportamentos, perfis) que acaba por ultrapassar a capacidade de armazenamento tradicional e esgota as possíveis soluções analíticas da época. A "velocidade" é constantemente referenciada apenas à análises em tempo real, entretanto também representa a taxa de mudanças desses dados e o link que há entre os conjuntos de informações que vêm em tamanhos e ritmos diferentes em vez de se manter constante. Já a "variedade" se concentra no formato e complexidade dos dados que surgem a todo o momento.

Tomando como base a definição de Boyd e Crawford (2012), Farida Vis (2013) procura criar "três Vs" alternativos que não fiquem restritos às preocupações

[3] API (Application Programming Interface) é um conjunto de rotinas e padrões de programação para acesso a um aplicativo de software ou plataforma baseado na Web.

[4] Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making. (GARTNER. Big Data. www.gatner.com, IT Glossary. Disponível em: <http://www.gartner.com/it-glossary/big-data/>. Acesso em: 25 ago. 2013)

mercadológicas e levem em consideração a problemática que envolve o processo de coleta e análise dos dados disponibilizados. Os três novos Vs propostos seriam: "validity" (validade), "venture" (ousadia) e "visibility" (visibilidade).

A validade seria a preocupação com a qualidade desses dados e se eles podem ser automaticamente considerados válidos para se usar de base para pesquisas e análises. A validação desses dados está intimamente ligada a um melhor entendimento do funcionamento do chamado *Data Firehose* (em português, mangueira de incêndio) e suas relações com as outras API's que os pesquisadores usam para realizar suas coletas de dados. No caso do *Twitter*, por exemplo, existem três tipos de API's possíveis: a *Twitter's Search API*, a *Twitter's Streaming API* e *Twitter Firehose*. A primeira se refere a busca retroativa de informações, usando como parâmetro de pesquisa um nome de usuário ou um termo de coleta; a segunda, coleta em tempo real uma amostra dos *tweets* que apresentarem os parâmetros previamente definidos; e a terceira, que é similar à API de Streaming, porém garante a entrega de 100% dos *tweets* que se encaixem na pesquisa e não é de graça, sendo utilizada pelas empresas provedoras de dados *Gnip* e *DataSift* (BRIGHTPLANET, 2013). Essa diferença entre os pacotes de dados entregues pela API de Streaming e pela *Firehose* devem ser levados em conta pois o modo como o *Twitter* determina a amostragem dos dados não é transparente aos pesquisadores e isso pode afetar diretamente o resultado de pesquisas.

Fred Morstatter, Jurgen Pfeffer, Huan Liu e Kathleen M. Carley (2013), em um artigo apresentado na Conferência Internacional em Weblogs e Mídias Sociais[5] (ICWSM), expuseram os resultados da comparação entre a API de Streaming e a *Firehose* usando como base um dataset referente a Síria. Os resultados obtidos foram que há sim diferenças entre os fluxos de dados que vem de cada uma: a Streaming recebe em média 43,5% de todos os dados disponíveis pela Firehouse; foi possível identificar entre 50% e 60% do top 100 atores chave em uma rede de interação usuário x usuário; e especificamente os *tweets* que possuíam geolocalização (menos de 1% do dataset inteiro) foram coletados em sua totalidade pela API de Streaming.

[5] International Conference on Weblogs and Social Media

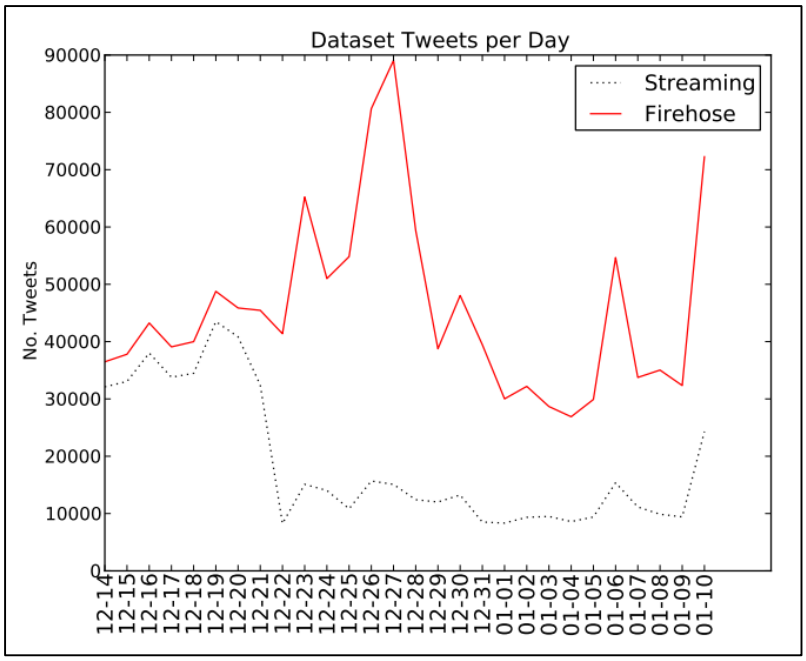

Figura 1 - Comparação de pacotes de dados entregues pela Streaming API e pela Firehose API. Fonte: MORSTATTER et all

Figura 2 - Comparação de pacotes de dados geolocalizados entregues pela Streaming API e pela Firehose API. Fonte: Fred Morstatter

São estudos como os de Fred Morstatter et al (2013) que alimentam a discussão acerca da porcentagem de dados que realmente são liberados por essas API's, e como é importante considerar essas questões metodológicas na hora de se tomar o dado como íntegro e absoluto. Mesmo assim, ainda que o dataset seja obtido em sua completude por meio da *Firehose,* deve-se tê-lo como uma representação do discurso das pessoas que necessariamente usam o site *Twitter* e seria errôneo tentar compará-lo com algum censo de uma população offline, mesmo que ambos sejam referentes a um mesmo país ou região. A questão é "sobre a compreensão de dados on-line como fundamentada em outros dados on-line, em vez de critérios off-line. O online torna-se, assim, a linha de base" (VIS, 2013).

A *venture* (em português, ousadia) seria, de acordo com Vis (2013) a ação de se aventurar em um campo de estudo que está em constante mudança e que é até então

desconhecido. É necessário ter curiosidade para pensar em visões de mundo diferentes do que já está estabelecido, e de apresentar os achados de pesquisa de modo a instigar novos olhares sobre os dados.

Já o último dos novos "três V's" é a visibilidade. Esse ponto levanta questões como a transparência acerca dos passos dados, da coleta à análise, os quais Vis (2013) diz que permanecem escondidos, podendo guardar informações cruciais ao entendimento do todo. Abarca também o processamento dos dados para formar algum tipo de visualização: o que essa visualização revela sobre o dado e o que não revela? Quais as habilidades necessárias para entender o que está sendo mostrado? Quais as diferentes formas de se visualizar um conjunto de dados e como escolher dentre elas? Essas questões estão intimamente conectadas ao modo como o dado se transforma em informação quando visualizado, portanto é pertinente saber qual conceito usar referente a "visualização".

1.3. Tornando os dados visíveis

A visualização de dados (data visualization) tem como propósito fazer com que uma determinada quantidade de informações não estruturadas, ganhem uma estrutura para assim ser possível enxergar padrões e relações escondidas no caos do montante de dados. O jornalista de dados e designer de informação David McCandless, em uma palestra dada no TED, fundação sem fins lucrativos destinada à disseminação de ideias e pesquisas, destacou a importância da visualização desses dados obtidos na internet:

> Parece que estamos sofrendo de excesso de informação ou abundância de dados. E a boa notícia é que pode haver uma solução fácil pra isso, e é usarmos mais os nossos olhos. E assim visualizando informação, para que possamos ver os padrões e conexões que importam e então projetar a informação para que faça mais sentido ou para contar uma estória ou que nos permita focar apenas na informação que for importante. (Traduzido para português. MCCANDLESS, 2010)

É nesse campo de visualização e desenvolvimento de softwares capazes de lidar com *Big data* que o pesquisador Lev Manovich, fundador e diretor do Grupo Software Studies (Universidade da Califórnia, San Diego) tem dissertado acerca da ação de se debruçar sobre esse montante de dados e quão importante é visualizá-los. Da mesma forma na qual houve a preocupação de estabelecer um conceito acerca de *Big data*, o mesmo precisa ser feito com essa visualização de dados. Manovich (2010) expõe que a definição utilizada por pesquisadores do campo de Ciência da Computação restringe

visualização ao uso de representações visuais interativas ou interfaces conduzidas por computadores, que "*InfoVis* é a comunicação de dados abstratos por meio do uso de interfaces visuais interativas" (KEIM et all. 2006)

Apesar de correta, o conceito de *InfoVis* não se resume a isso. Manovich (2010) propõe então classificar os tipos de visualizações de acordo com suas características estruturais: em um primeiro momento ele diferencia a *information visualization* (*InfoVis)* e a *scientific visualization*; depois diferencia a *InfoVis* do que ele chama de *information design*; e finaliza propondo o conceito de *media visualization*. Para fins de melhor entendimento, cada termo foi traduzido para seu correpondente em português e a expressão "*information visualization"* será tratada também por seu nome encurtado "*InfoVis"*.

Muitos pesquisadores consideram que a distinção entre visualização científica e visualização informacional é que a primeira utiliza dados numéricos, enquanto que a última usa dados não-numéricos, como elementos textuais ou redes de relações entre usuários. Entretanto, Manovich (2010) diz que essa divisão não é assertiva, pois apesar das *InfoVis* terem a possibilidade de usar outros tipos de dados que não sejam números, a base primária delas ainda continua sendo numérica.

/Para ele, a diferença entre visualização científica e *InfoVis* começa no modo como os diferentes campos de estudo (ciência e design, respectivamente) se relacionaram com o avanço tecnológico do final do século XX e com qual área da tecnologia de computação gráfica eles acabaram por se vincular. As visualizações científicas se aprimoraram na década de 80, juntamente com o advento do campo 3D, o qual requeria estações especializadas de trabalho, enquanto que as *InfoVis* só tiveram seu "boom" na década de 90, com a ascensão dos softwares gráficos no ambiente trabalho em 2D e a obtenção de computadores pessoais por profissionais da área. Outra diferença que Manovich (2010) apresenta é que as *InfoVis* usam parâmetros espaciais arbitrários, enquanto que a visualização científica se baseia em critérios fixos acerca de objetos que já têm seus limites definidos, como uma imagem 3D de um cérebro ou localizações plotadas em um mapa.

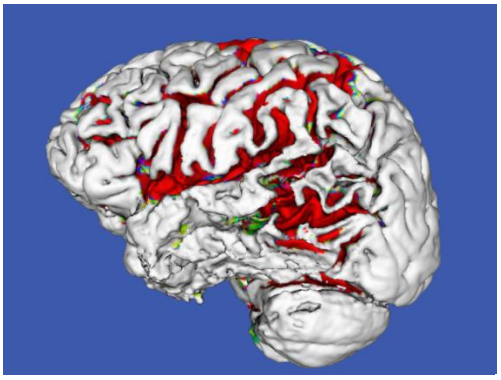

Figura 3 – Exemplo de visualização científica

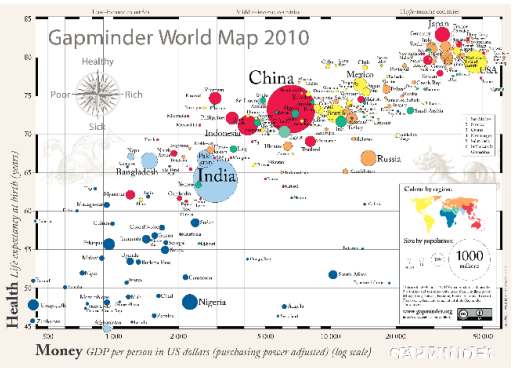

Figura 4 - Exemplo de visualização informacional

Definido isso, Manovich (2010) avança na distinção entre *information visualization* e *information design*. O design informacional trabalha com uma estrutura de dados clara, mas que precisa ser visualizada de alguma maneira (por exemplo: mapa de trens, com suas linhas, estações e localizações). Já a visualização informacional não teria uma estrutura clara de dados a priori, pois seu objetivo é justamente identificar essa estrutura quando os dados forem visualizados. De modo resumido: "Design informacional trabalha com informação; visualização informacional trabalha com dados" (MANOVICH, 2010).

O hábito de representar informações utilizando gráficos tem seu início juntamente com o desenvolvimento mais profundo do pensamento estatístico por volta do século XVII e foi acompanhado também

> por um aumento também no pensamento visual: diagramas foram usados para ilustrar provas matemáticas e funções; nomogramas foram desenvolvidos para auxiliar nos cálculos; diversas formas gráficas foram inventadas para tornar as propriedades dos números empíricos - suas inclinações, tendências e distribuições - mais facilmente de serem comunicadas ou acessíveis à investigação visual. (...) Mais recentemente, os avanços na computação estatística e display gráfico forneceram ferramentas de visualização de dados que seriam impensáveis meio século atrás. Da mesma forma, os avanços na interação humano-computador têm criado novos paradigmas para explorar informações gráficas de uma forma dinâmica, com flexível controle pelo usuário. (Tradução própria. FRIENDLY, 2009)

Os primeiro diagramas aos quais Friendly (2009) se refere são os mais simples gráficos estatísticos que ainda hoje são utilizados para transformar os dados numéricos em representações visuais. John Arbuthnot (1667-1735) e William Playfair (1759-1823) foram os idealizadores dos estilos de gráficos mais comumente usados nos dias de hoje (FRIENDLY, 2009): Arbuthnot foi responsável pela criação do primeiro gráfico de linha que se tem notícia, no ano de 1711, enquanto que Playfair idealizou o conceito de gráfico de barras (1786) e gráfico de pizza (1801). Manovich (2010) coloca que os dois

principais conceitos-chave na criação das *InfoVis* advêm dessas representações estatísticas: o conceito de “redução” e o conceito de“espaço”.

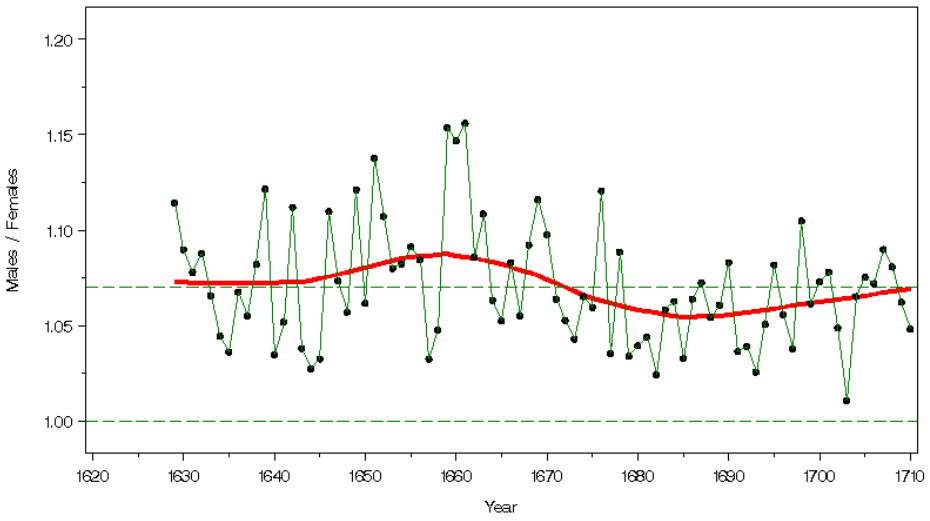

Figura 5 - Primeiro teste de significância estatística baseado no desvio entre dados observados e uma hipótese nula (John Arbuthnot - 1711)

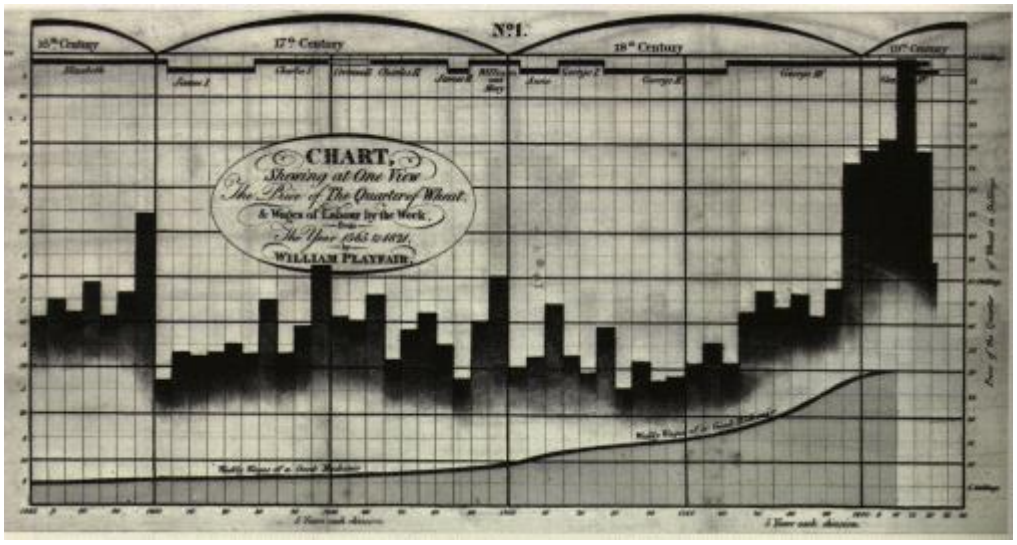

Figura 6 - Gráfico de barras e gráfico de linha usando dados econômicos (William Playfair - 1786)

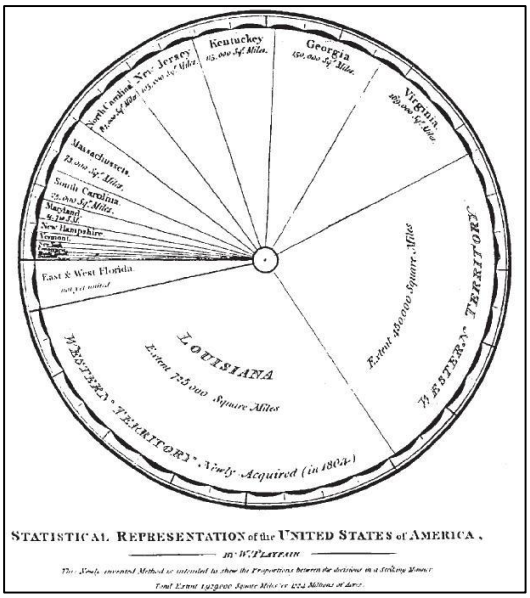

Figura 7 - Invenção do gráfico de pizza usado para ilustrar as relações da parte com o todo (William Playfair - 1801)

Para Manovich (2010), as *InfoVis* reduzem os dados a formas geométricas, linhas e curvas para transformá-los em objetos e assim mostrar a relação entre eles. Apesar dessa prática possibilitar a visualização de padrões, o preço que se paga é a extrema esquematização. Toma-se um dado complexo e, nesse processo de redução, ele se simplifica, perdendo a maioria das especificidades que o tornam único. A hipótese proposta por Manovich (2010) é de que a prática de redução visual se tornou comum pois seguiu o pensamento científico no qual tudo no universo pode ser reduzido a partículas pequenas (no caso da física, átomos; biologia, células; química, elementos químicos) logo o mesmo poderia ser feito com as interações sociais. O paradigma é que se espera que ao reduzir os dados, complexas estruturas de interação e comportamento possam se revelar a partir de elementos simples.

Acerca da questão do "espaço", Manovich (2010) que as visualizações informacionais de hoje privilegiam as variáveis espaciais em detrimento de outras dimensões visuais, tais como cor, saturação, padrões de sombra, texturas, entre outras, e coloca o exemplo do gráfico de dispersão (*scatter plot*), cujos pontos representam as informações e a distância entre eles implica na semelhança ou discrepância entre elas. Nesse caso, como no próprio software Gephi usado no Labic, as dimensões visuais são relegadas à função de apenas categorizar em vez de transparecer mais informações sobre o próprio dado. De acordo com Manovich (2010), o privilégio das variáveis espaciais em detrimento de outras dimensões visuais se dá devido a necessidade do cérebro humano separar espacialmente os objetos presentes no nosso mundo para uma

melhor compreensão do ambiente em que vivemos. Identificar os diferentes tamanhos, formas e posições dos objetos em relação ao nosso corpo possuem papel importante no reconhecimento do espaço a nossa volta. Logo, as propriedades cromáticas de uma cena não exercem tanta influência na nossa vivência do dia a dia se comparadas com as variáveis espaciais.

Com essas questões em mente, Manovich (2010) propõe que o modo que mais resguarda a complexidade e singularidade dos dados seria o que ele chama de *media visualization*, considerada como uma "visualização direta" ou uma "visualização sem redução". Nesse tipo de visualização, os dados são reorganizados preservando o máximo da sua forma original: o que está sendo visualizado não é substituído por novos objetos, como por exemplo formas geométricas ou linhas, apresentando modificações quantitativas porém sem grandes perdas qualitativas. Pensando nas linhas de estudo de semântica e estudo de imagem, visualizações que seguem o estilo proposto por essa *media visualization* são respectivamente o *WordCloud* (Nuvem de Palavras) e o próprio *ImageCloud* (Nuvem de Imagens). Neles, o pesquisador está visualizando o próprio dado quantificado: a palavra é representada pela palavra, enquanto que a imagem é representada pela própria imagem em si.

Figura 8 - Exemplo de WordCloud

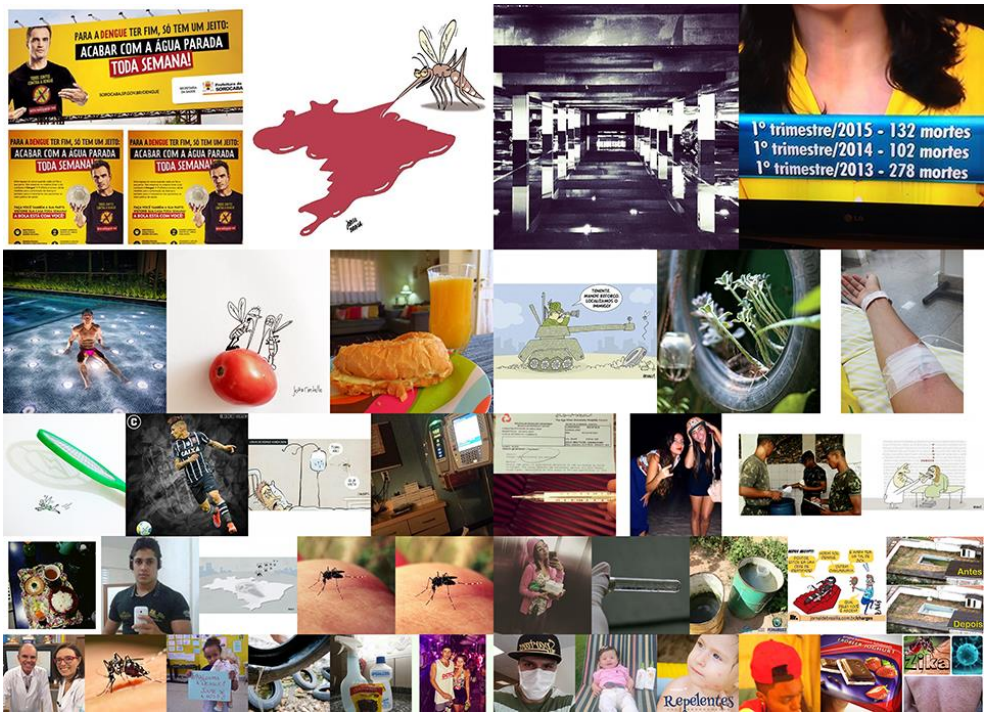

Figura 9 - Exemplo de ImageCloud

2. Laboratório de Estudos sobre Imagem e Cibercultura: um retrospecto do laboratório

Criado em 2007 como projeto de extensão do Professor Fábio Malini, o Laboratório de Estudos sobre Imagem e Cibercultura tem como objetivo produzir experimentalmente ferramentas de uso digital e promover pesquisas e atividades de extensão no campo da comunicação vinculado à cultura digital. Multidipliscinar, o laboratório faz parte do Departamento de Comunicação Social, com associação ao Programa de Pós-Graduação em Ciência da Computação, e concilia computação, design, antropologia e as teorias da comunicação para a realização de suas pesquisas.

O laboratório atualmente possui três linhas de pesquisa: a linha de Modelagem e Análise de Redes com foco em semântica, a linha de Análises de Imagens, e a linha de Desenvolvimento de Softwares, específicos para aplicação digital. Há ainda uma atuação no campo de estudos de contextos políticos e sociais, especialmente protestos.

2.1. Labic e primeiros estudos sobre Imagem: coleta da hashtag #protestoes

Estudar as questões relacionadas à internet permite a análise das diversas possibilidades de formas de expressão dos usuários na rede. Ao falar de redes sociais digitais, dando destaque às mais populares como *Facebook, Twitter* e *Instagram*, o pesquisador se vê diante de uma vasta gama de caminhos a seguir. Para além dos estudos de redes de interação, no Labic houve busca por visualização das informações disponibilizadas pelos usuários da rede por meio de imagens.

Diante disso, o laboratório iniciou sua proposta de coleta e análise de imagens, tendo sua primeira experiência motivada pelas manifestações ocorridas em todo território brasileiro no período entre os dias 17 e 24 de junho de 2013. A marcha dos 100 mil, em Vitória - ES, que atravessou a Ponte Deputado Darcy Castello de Mendonça, conhecida como Terceira Ponte, de modo simbólico, foi no estado capixaba o ato que mais reuniu manifestantes, resultando também em forte produção de conteúdo publicado nas redes sociais. Com a célebre frase “Não é só por 20 centavos”, os protestos foram amplamente divulgados não só pela mídia tradicional, mas por mídias alternativas (como a Mídia Ninja ES) e pelos próprios manifestantes que portavam equipamentos como câmeras e celulares com acesso à internet.

A linha de estudos em semântica do Laboratório já possuía metodologia avançada em coleta de *tweets* e percebeu-se que, dentro do dataset vinculado a hashtag #VemPraRua, existia uma série de outras tags específicas das localidades nas quais aconteciam manifestações, como por exemplo #protestosp, #protestorj e #protestomg.

No caso do movimento do "Vem Pra Rua" em solo capixaba, algumas das tags regionais foram "#protestoes", "#protestovitória" e "#protestovix", sendo que a maior parte dos conteúdos se concentravam indexados pela hashtag #protestoes. Tendo isso em conta, os pesquisadores do laboratório optaram por um recorte de pesquisa, se concentrando em realizar, ainda que de forma manual, a coleta das imagens produzidas e compartilhadas somente da hashtag #protestoes. Na época, o laboratório ainda não possuía metodologia de coleta automatizada para imagens e as poucas ferramentas que entregavam algum tipo de dado eram ou pagas ou restringiam consideravelmente a amostra de conteúdo a ser entregue. Os sites de redes sociais escolhidos foram o *Facebook* (492 imagens) e *Instagram* (500 imagens) cujas metodologias de captura basearam-se no funcionamento próprio do motor de busca de cada site. A coleta manual realizou-se então no ato rudimentar de salvamento de imagens (botão direito, salvar como..., armazenar numa pasta do computador), pois a elaboração de um script não seria viável, devido a uma questão de tempo e necessidade de analisar as imagens conforme as manifestações ocorriam.

Como o *Facebook* apresenta as suas informações usando como parâmetro o histórico de visualizações, comentários e opções de curtir do usuário logado, as imagens que apareciam pela busca da hashtag, não seguiam padrões cronológicos nem eram dispostas por quantidade de curtidas ou compartilhamentos. A conclusão seria de que se a pesquisa fosse realizada por um outro usuário, a ordem e as informações que apareceriam seriam outras, resultando em diferentes tipos de datasets que gerariam diferentes resultados de análises. Outro ponto a se notar é de que as imagens postadas e compartilhadas nesse site de rede social não seguem um padrão de tamanho como as do Instagram, que na época eram reformatadas para o formato 4:3 semelhante ao da Polaroid[6].

[6] Estilo de câmera fotográfica que revela instantaneamente a imagem num formato 4:3. Dá-se o nome de Polaroid pela Empresa Polaroid ter sido quem popularizou o modelo no mercado.

Para salvar as imagens do *Instagram*, foi utilizado o visualizador online *Webstagram*[7], por dispor uma maior quantidade de publicações por página e facilitar seu download. Como os padrões de data seguem a lógica do tempo que passa, as imagens foram organizadas em dois grupos, semana 1 (17/06 – 22/06) e semana 2 (23/06 – 24/06), já que não era possível precisar o dia exato no qual elas foram postadas (restrição do *Instagram*).

Para criar a visualização das imagens do Facebook e Instagram, que continham a hashtag "#protestoes", foi utilizado o *software ImageJ*[8], tendo como opção macro de visualização desenvolvida especificamente para o programa chamada de *ImagePlot*[9]. Com as imagens já coletadas, o *ImageJ* foi utilizado principalmente por ser recomendado àqueles que precisam lidar com uma grande quantidade de imagens juntas.

2.2. Métodos de visualização: *ImageJ* e *ImagePlot*

O *ImageJ* é um programa escrito em linguagem de programação *java*, de domínio público, utilizado para o processamento de imagens, desenvolvido pela *Research Services Branch, ramificação do National Institute of Mental Health* localizado em Maryland, nos Estados Unidos. O programa possui código aberto (open source), fornecendo extensibilidade por meio da criação de novos plug-ins e novas macros *java*, e possibilitando a resolução de possíveis problemas de análise e visualização encontradas pelo pesquisador.

Em se tratando da macro *ImagePlot*, as ferramentas de visualização disponíveis mostram os dados das imagens como pontos, linhas e barras. As visualizações do *ImagePlot* mostram as imagens reais, que podem ser redimensionadas em qualquer tamanho e organizadas em qualquer ordem – de acordo com as respectivas datas, conteúdo, características visuais. Como o vídeo digital é apenas um conjunto de

[7] Visualizador online de publicações do *Instagram*. Disponível em: <https://websta.me/>.

[8] O ImageJ pode exibir, editar, analisar, processar, salvar e imprimir magens de 8 bits, 16 bits e 32 bits. Ele pode ler vários formatos de imagem, incluindo TIFF, GIF, JPEG, BMP, DICOM, FITS e "RAW". Ele suporta "stacks" ("pilhas"), uma série de imagens que partilham de uma única janela. É multitarefa, assim operações demoradas, como a leitura do arquivo de imagem pode ser realizada em paralelo com outras operações. (...) Ele suporta padrões de funções de processamento de imagem como manipulação de contraste, nitidez, suavização, detecção de bordas e filtragem mediana" (Traduzido do http://rsb.info.nih.gov/ij/docs/intro.html)

[9] Macro desenvolvida pelo Grupo de Estudos no Software, compatível com o ImageJ, que possibilita a plotagem e visualização de múltiplas imagens em um gráfico com Eixo X e Eixo Y.

imagens estáticas individuais, a macro também consegue explorar padrões em filmes, animações, jogos, e outros dados de imagem em movimento.

Ambos são utilizados pelo Grupo de Estudos do Software[10], dirigido pelo professor e pesquisador Lev Manovich, e referência internacional no quesito de visualização de dados. Desde 2007, esse grupo desenvolve pesquisas sobre análise de cultura relacionada às interações na internet, bem como visualizações de *Big Data* (como por exemplo a de 4535 capas da Times Magazine [11]dispostas por brilho médio e variação do padrão de brilho).

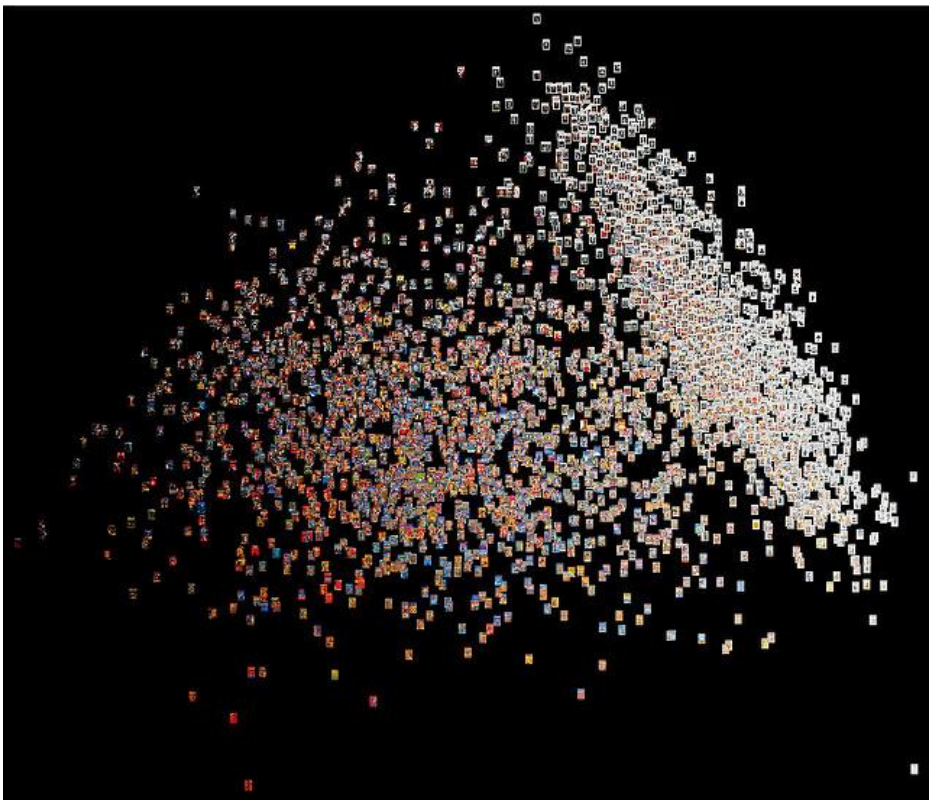

Figura 10 - 4535 capas da Times Magazine: Eixo X Brilho médio vs Eixo Y Variação do padrão de brilho

A macro *ImagePlot* permite a visualização de grupos de imagens com base em uma série de parâmetros como matiz, brilho, saturação e outros dados que podem ser posteriormente adicionados às colunas do arquivo da planilha correspondente ao *dataset*. No caso das imagens coletadas como primeira etapa do projeto "Visagem", os dados foram organizados de modo a transparecer, além das relações entre as propriedades numéricas da imagem, possíveis análises do campo da teoria das imagens.

[10] Disponível: http://lab.softwarestudies.com/
[11] Disponível em: <http://www.flickr.com/photos/culturevis/3951496507/in/set-72157624959121129>.

As 500 imagens capturadas do *Instagram* e as 492 imagens do *Facebook* foram organizadas em um dataset contendo “Nome do Arquivo”, “ID”, “Brilho Médio”, “Saturação Média”, “Cor Média”, “Legenda” e “Usuário”. O *dataset* foi criado no excel, sendo cada uma dessas divisões uma coluna na tabela, e posteriormente transformado para a extensão de texto separado por tabulações (.txt) para que pudesse ser lido pela macro *ImagePlot*. As medições para descobrir brilho, saturação e cor de cada imagem foram conseguidas através de outra macro chamada “measurements”, disponibilizada junto com o *ImageJ*.

A coluna “Nome do Arquivo” consta como a identificação da imagem no computador. É por meio dessa coluna que o programa relaciona cada linha da tabela com determinada imagem na pasta. A linha da tabela contém as informações como brilho, saturação e cor da imagem correspondente a ela. Apesar de ser uma classificação fácil, feita através de números, uma dificuldade percebida foi que ao numerar as imagens sequencialmente (1.jpg; 2.jpg; 3.jpg e assim por diante) o *ImageJ,* ao rodar a macro “measurements”, organizava os resultados de forma confusa, seguindo o raciocínio de que as primeiras imagens eram aquelas que começavam com o número um (1.jpg; 10.jpg; 11.jpg; (…) 100.jpg). Sendo assim, uma forma de contornar o problema foi nomear as imagens com uma quantidade de zeros proporcional ao número de imagens (da imagem 001.jpg à imagem 099.jpg), gerando as informações na ordem correta das imagens.

Os valores para “Brilho Médio”, “Saturação Média” e “Cor Média” foram obtidos, como dito anteriormente, por meio da macro “measurements” e depois anexados à nova tabela de metadados que estava sendo construída pelos pesquisadores.

As colunas “Legenda” e “Usuários” estão estritamente ligadas. Os usuários são os perfis (ou páginas, no caso do *Facebook*) que postaram ou compartilharam as imagens que continham a hashtag “#protestoes”. Já a “Legenda” foi uma maneira encontrada para tornar viável a utilização dos usuários como um dado na criação das visualizações, já que a inserção de texto como um parâmetro dentro do *ImageJ* não era possível.

No momento de gerar os gráficos alguns problemas ficaram evidentes. De modo geral, as visualizações tiveram de ser geradas utilizando computadores mais potentes (por exemplo, o computador utilizado possui 32 GB de RAM e processador Intel Core 7), devido à necessidade de grande quantidade de memória e de maior eficácia no processamento das imagens. Um detalhe que pôde ser percebido nas imagens do

Instagram foi que a utilização de filtros em preto e branco, alterava a composição da imagem transformando-a em *Grayscale* (LUT) e acarretando um erro de leitura na macro. A solução encontrada foi converter todas as que continham esse metadado para RGB, não alterando as características básicas da imagem, mas viabilizando a renderização dos gráficos.

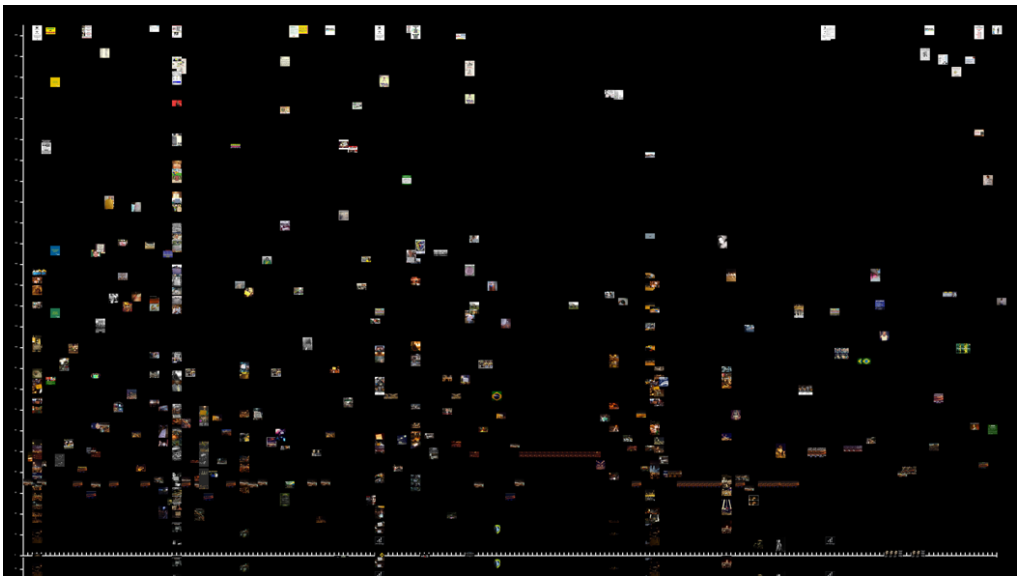

Figura 11 - 492 imagens do Facebook: Eixo X Usuário vs Eixo Y Brilho Médio

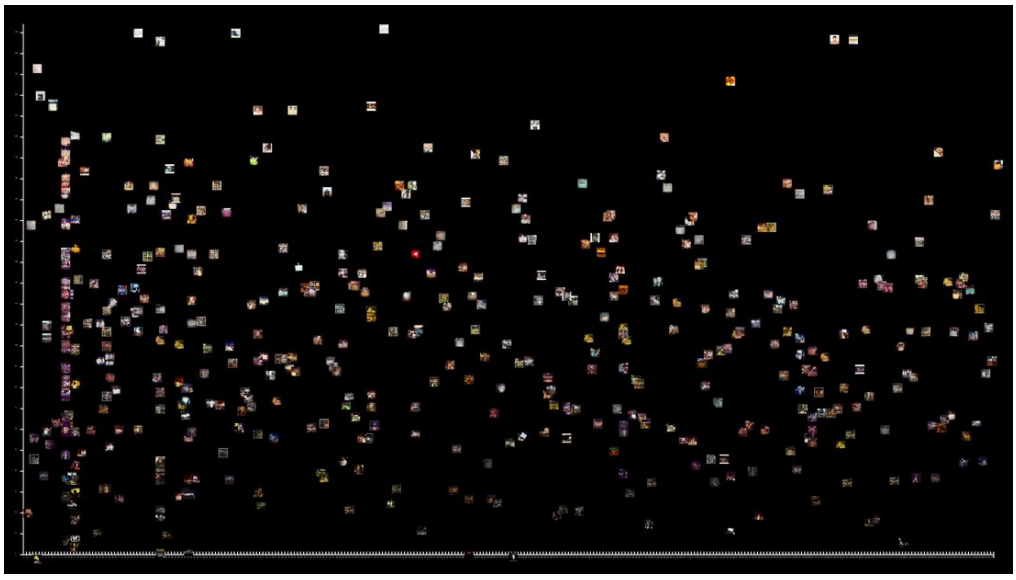

Figura 12 - 500 imagens do Instagram: Eixo X Usuário vs Eixo Y Brilho Médio

Figura 13 - Montagem das 492 imagens doFacebook #protestoes

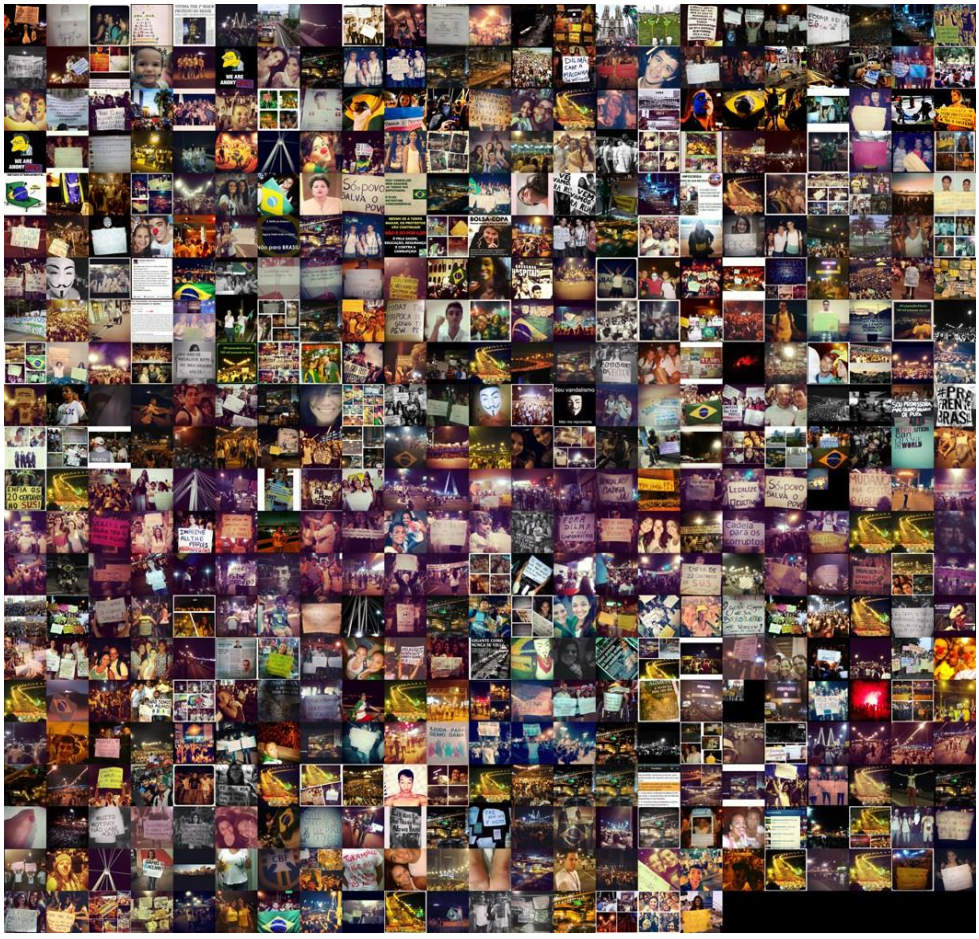

Figura 14 - Montagem das 500 imagens do Instagram #protestoes

A dinâmica de coleta e de criação do dataset sobre a *hashtag* #protestoes revelou uma série de questões acerca da necessidade de obter suporte da tecnologia para automatizar processos, como a coleta das imagens em si e solução de erros em vários arquivos de uma só vez. Foi na convergência entre humanidades e tecnologias que se desenvolveu a primeira versão do script de coleta de imagens do Labic cujo nome dado foi *Crawler*[12].

2.3. O Movimento Passe Livre e as imagens do *Twitter*: *Crawler* e *ImageCloud*

Conforme os dias passavam, a pauta das manifestações de 2013 ainda predominavam e traziam para discussão pontos de diversas agendas dos movimentos sociais. Um dos movimentos que obteve bastante destaque durante os protestos foi o MPL (Movimento Passe Livre), que historicamente luta pela tarifa zero no transporte público. O intuito do movimento era trazer para discussão as políticas públicas de transporte coletivo, nas quais o translado fosse parte do direito de ir e vir do cidadão, e não sendo tratado apenas de forma monetarizada.

O protesto contra o reajuste das tarifas do transporte no estado de São Paulo, organizado pelo Movimento Passe Livre (MPL) no início de junho de 2013, teve grande repercussão em todo o Brasil. Pessoas de diferentes estados do país começaram a aderir à causa e, assim, iniciou-se a jornada de manifestações ocorridas ao longo de junho e julho. A mobilização do MPL foi o estopim para que outras pautas fossem colocadas em vista além da oposição às altas tarifas estabelecidas pelo governo.

Diferentemente do #protestoes, de cunho mais regional, a hashtag #passelivre apresentava maior alcance territorial, sendo utilizada na maioria dos estados nos quais estavam acontecendo manifestações. Por isso, para ampliar a visão sobre os acontecimentos daquele ano, buscamos entre o período de 15 de junho a 15 de julho as publicações que foram feitas no site de rede social *Twitter* relacionadas a essa tag do movimento.

Por se tratar do *Twitter*, primeiro era necessário coletar todos os *tweets* que continham a tag do #passelivre para depois verificar quais postagens possuíam links, o que seria um indício de presença de imagens. Através do software

[12] Script escrito em linguagem de programação *java* cuja entrada é uma tabela em formato de arquivo .csv. Ele busca dentro da tabela os tweets que têm links, abre esses links, e salva as imagens que estiverem contidas neles.

yourTwapperKeeper[13], o montante de *tweets*, delimitado pelo período temporal já citado, foi coletado e passou a compor o *dataset* que seria trabalhado e analisado. Após o término da extração, o programa entrega um arquivo com extensão .csv (comma separated values[14]) que contém as informações disponibilizadas pela API[15] do *Twitter*, como a data de criação do *tweet*, número de retweets, o usuário que originou o *tweet*, entre outras.

Com esse arquivo em mãos, o script *Crawler* separa os *tweets* com links daqueles sem links, e inicia o processo de coleta de imagens. Cada link de cada *tweet* é visitado, salvando-se a página da internet em uma tabela excel e realizando-se o download das imagens contidas nela. Para que não houvesse uma grande ocorrência de imagens desnecessárias, como peças publicitárias de borda de página ou elementos gráficos do design do site, foram estabelecidos padrões para que as imagens fossem consideradas "válidas": as imagens deveriam ter um tamanho mínimo de 200x200 pixels e um tamanho em disco de, pelo menos, 15 kb. Também só seriam aceitas imagens nos formatos .bmp; .jpg; .jpeg; .tiff e .tif para a extração.

No final da coleta fica visível a diferença entre a coleta manual e a coleta automatizada. Enquanto que na primeira experiência do #protestoes o total de imagens coletadas foi de 942 mídias, levando cerca de duas semanas para o pesquisador atingir esse número, na coleta automatizada do #passelivre o montante foi de 6.638 imagens coletadas, com maior precisão e rapidez.

[13] yourTwapperkeeper é a versão open source de uma ferramenta que permite aos investigadores controlar, arquivar e compartilhar conjuntos de dados de *tweets* relacionados com várias palavras-chave. Disponível em: < https://github.com/540co/yourTwapperKeeper>.
[14] Formato de aquivos de tabela no qual os valores de cada coluna são separados pele caractere de vírgula ",". Esses arquivos são lidos por programas como o Microsoft Excel ou LibreOffice Calc.
[15] API (Application Programming Interface) é um conjunto de rotinas e padrões de programação para acesso a um aplicativo de software ou plataforma baseado na Web.

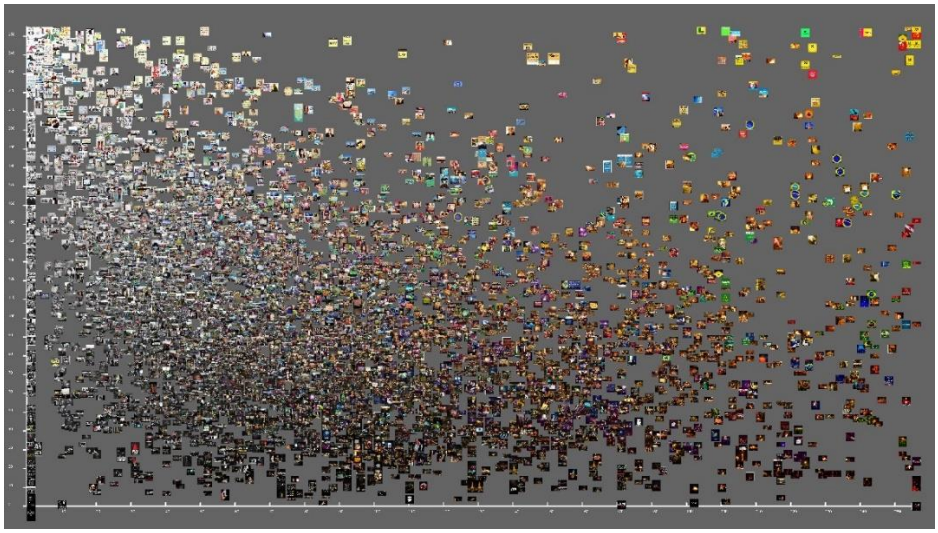

Figura 15 - 6638 imagens da #passelivre: Eixo X Saturação média vs Eixo Y Brilho Médio

Com o sucesso da automatização, novos horizontes de pesquisa se expandiram. Ficou claro que capturar um grande volume de informação produzido nos sites de redes sociais não é impossível e que há maneiras de tornar esse processo cada vez mais rápido, auxiliando o pesquisador a estudar a discussão em rede de modo quase simultâneo.

Juntamente com as hashtags anteriores, a #vemprarua foi uma tag muito ativa durante as manifestações de junho e as subsequentes. Esse lema que o MPL já havia entoando desde 2005, durante a Revolta da Catraca em Florianópolis e outras capitais, viralizou por meio de uma apropriação feita pelo comercial da montadora de veículos Fiat. O "Vem Pra Rua" da Fiat tinha em seu significado trazer as pessoas para a rua para torcer pelo país durante a Copa das Confederações realizada no Brasil em 2013, mas acabou por inflamar os ânimos dos cidadãos que protestavam, além de propagar o levante do movimento "Vem Pra Rua" que buscava a diminuição das tarifas de transporte e posteriormente, abarcou pautas mais abrangentes como corrupção e desigualdade social.

Essa popularidade da tag pode ser notada por meio da quantidade de linhas que o arquivo .csv possui ao ser gerado pelo script *Crawler*. Cada linha representa um *tweet* com link que foi encontrado dentro do arquivo inicial gerado pelo programa *YourTwapperKeeper*. No caso da hashtag #VemPraRua, foram encontrados mais de 190 mil links entre o período do dia 15 de junho a 18 de julho de 2013.

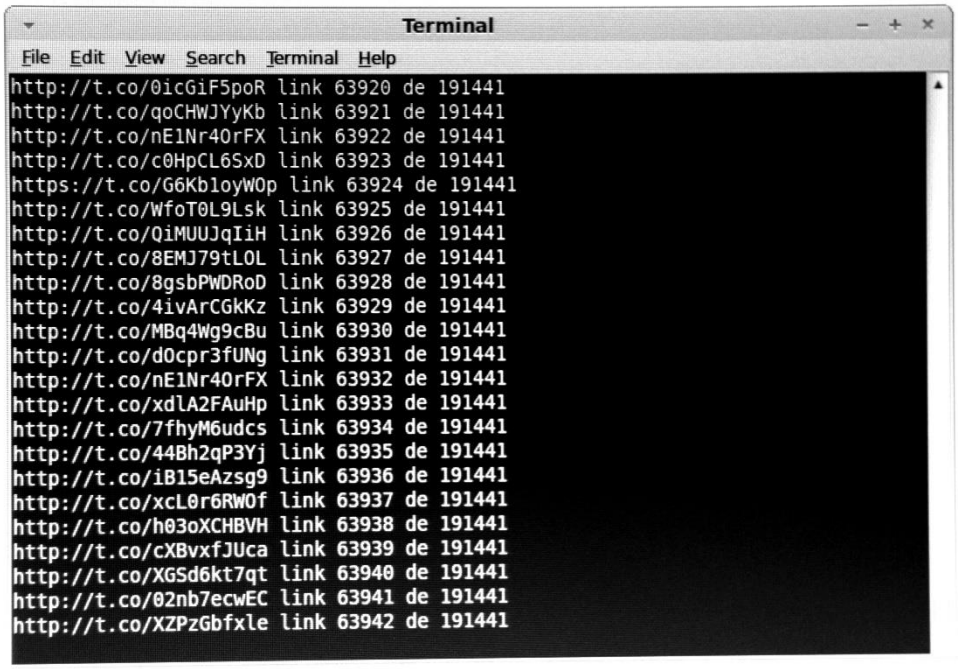

Figura 16 - Terminal de comando no qual o script Crawler rodava o #vemprarua

Desses 190 mil links contidos nos tweets, 85.595 imagens foram coletadas, retratando uma pluralidade de elementos como imagens de cartazes convocatórios, protestos diurnos e noturnos, fotos aéreas, fotos selfies, entre outros. Para visualizar essa grande quantidade de imagens na tela do computador, recorreu-se mais uma vez ao *ImageJ/ImagePlot* para organizar e exibir as informações em tela e ser possível enxergar padrões por meio do gráfico criado.

Utilizando informações que a API do *Twitter* disponibiliza (como a data e hora na qual o *tweet* foi publicado) juntamente com parâmetros de cor, brilho e saturação de cada imagem (gerados após a coleta no *software ImageJ*), quatro modos de visualização foram realizados: saturação x brilho médio, tempo x cor média, tempo x brilho médio, e a Nuvem de Imagens.

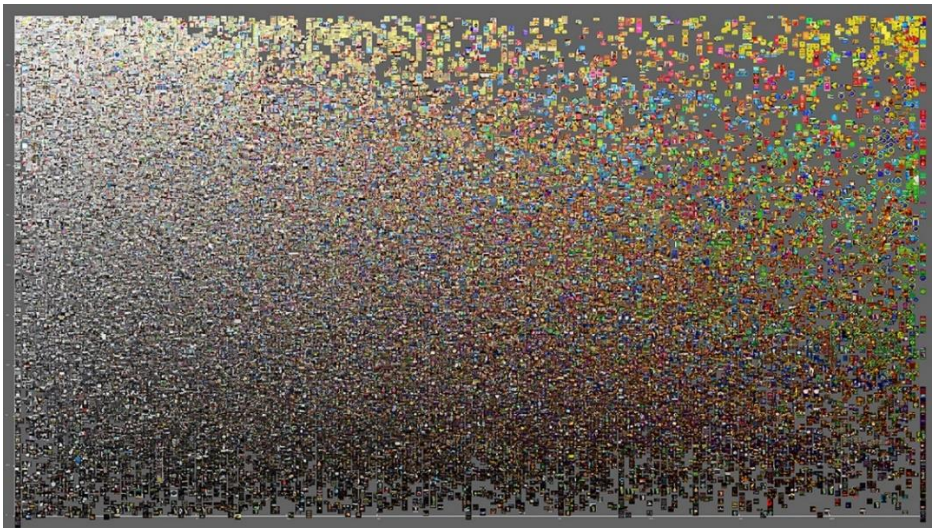

Figura 17 - 85 595 imagens do #vemprarua, entre 15 de junho à 18 de julho: Eixo X Brilho vs Eixo Y Saturação

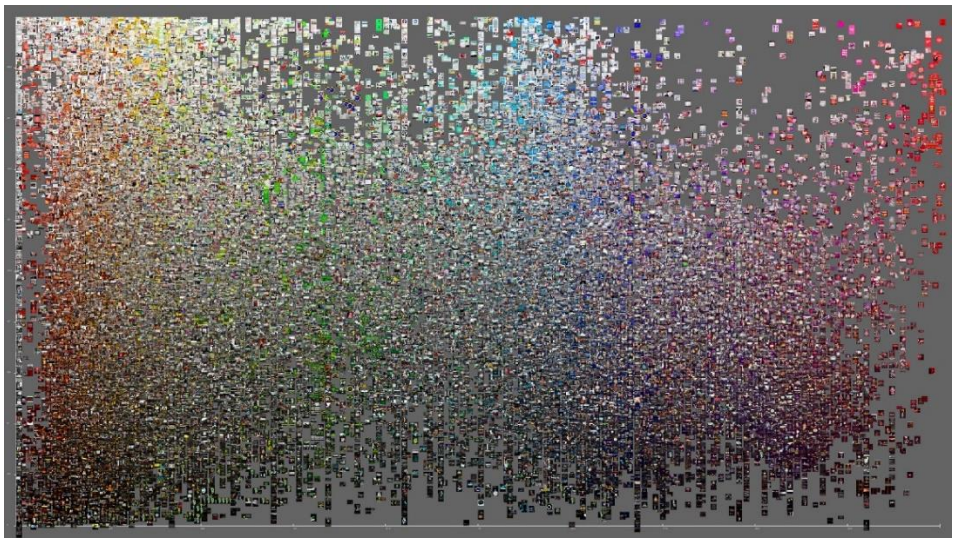

Figura 18 - 85 595 imagens do #vemprarua, entre 15 de junho à 18 de julho: Eixo X Cor vs Eixo Y Brilho

Figura 19 - ImageCloud das 85 595 imagens do #vemprarua: da esquerda para a direita, de cima pra baixo, ordem decrescente de *retweets*

O problema mais comum que a plotagem do *ImageJ/ImagePlot* apresenta é a sobreposição das imagens dentro do gráfico. O programa trabalha com eixo X e eixo Y no plano cartesiano, sendo que a posição da imagem nesse espaço é calculada baseando-se em dois parâmetros pré determinados pelo pesquisador. Dessa forma é muito provável que duas imagens tenham posições muito próximas (ou até iguais) e acabem umas sobre as outras, resultando na não utilização nem visualização da imagem que ficar por baixo. Como o gráfico final é estático, não há qualquer tipo de interatividade que possibilite rearranjar as imagens ou mesmo separar por camadas. Imagens sobrepostas são perda de informação imagética e resulta em alterações e mudanças nos resultados das análises.

Levando esse fato em consideração, o *ImageCloud* foi desenvolvido e pensado com a proposta de plotar as imagens baseando-se em um único parâmetro, gerando uma visualização linear mais simplificada porém não menos reveladora. Da esquerda para direita (e de cima para baixo) temos as imagens mais *retweetadas* (caso o parâmetro definido seja quantidade de *retweets*). Dessa forma, é fácil visualizar quais são as imagens mais relevantes do dataset.

2.4. Ferramentas idealizadas: ALICE e AISI

No quesito de interatividade, o *ImageCloud* não vai muito além do que o *ImageJ/ImagePlot* propõe: o produto final é estático e não permite modificações de parêmetros após sua renderização. Pensando nisso, o Labic propôs o desenvolvimento de um software que contemplasse em suas funções a escolha e posterior mudança de parâmetros, bem como o possível recorte de um grupo específico dentro de um conjunto inteiro de imagens. Esse é o conceito por trás do ALICE (*Analytical Laboratory for Image Collections as Entities*): um laboratório que carrega determinado banco de imagens existentes no computador do pesquisador. A linguagem de programação utilizada é *java* com biblioteca *javax.swing* e roda em qualquer sistema operacional desde que tenha o *java* instalado. Essa ferramenta ainda está sendo aprimorada, mas sua versão inicial já soluciona o empecilho da interatividade. Contudo há sérios problemas de processamento que precisam ser solucionados para que seu funcionamento tenha a eficiência esperada.

Quando se faz pesquisa na área de Big Data em sites de redes sociais, faz-se necessário perceber determinados comportamentos que acabam por refletir na análise final de qualquer dataset, como, por exemplo, a postagem de imagens iguais por

diferentes usuários. Aparentemente isso não seria problemático, não fosse a possibilidade de interferir e alterar a estatística de frequência por *retweets* no final da coleta. Por exemplo, a imagem A tem 100 *retweets* num único link, enquanto que a imagem B tem 100 *retweets* divididos em 10 links diferentes. O correto seria dizer que a imagem A e a imagem B possuem a mesma quantidade de *retweets* no total, porém quando a imagem B é replicada em links diferentes o programa a enxerga também como imagens diferentes, não reconhecendo que é uma repetição da mesma imagem.

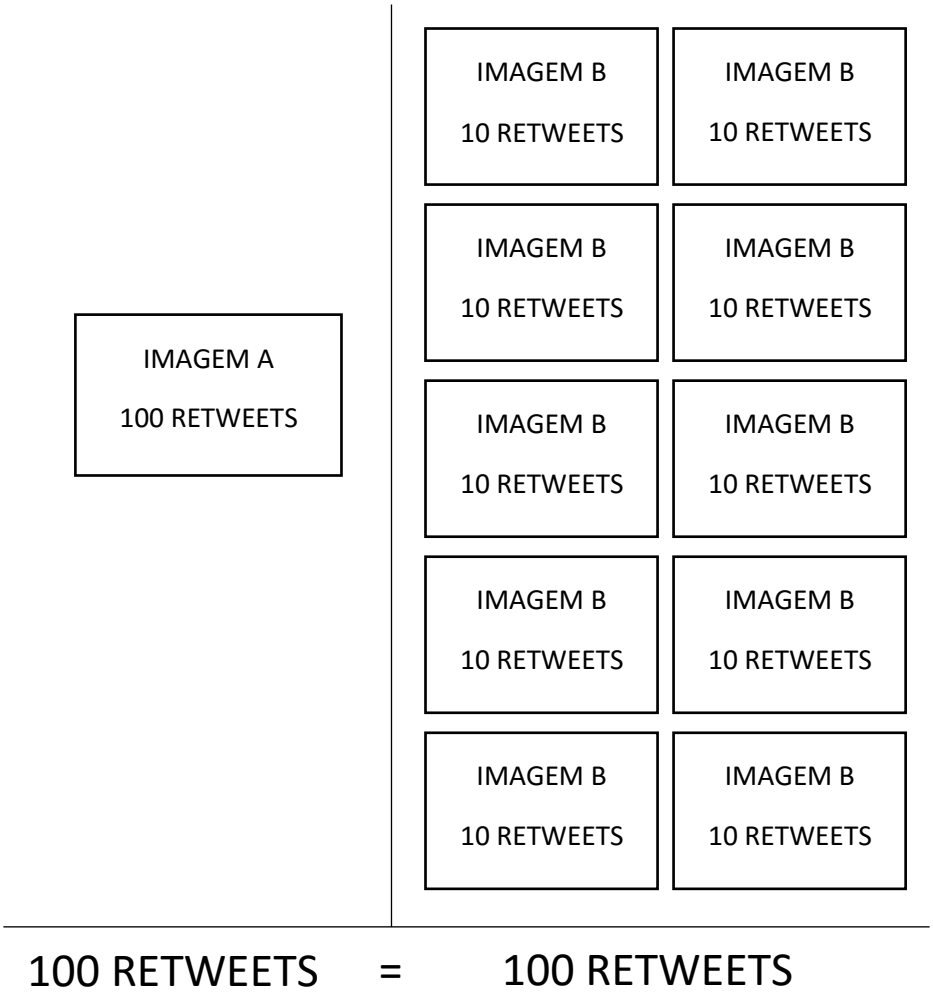

Figura 20 – Esquema ilustrativo "Imagem A e Imagem B". Fonte: a autora

Sendo assim, viu-se a necessidade de desenvolver um script que levasse em conta a presença no *dataset* de imagens iguais, porém em links diferentes. De nome AISI[16] (*Automatic Identifier of Similar Images*), o script desenvolvido no software *Matlab*[17]

[16]Em português: "Identificador de Imagens Similares"
[17] A plataforma MATLAB é otimizada para resolução de problemas de engenharia e científicos. A linguagem MATLAB baseada em matrix é a maneira mais natural do mundo para expressar matemática computacional. É usado para o aprendizado de máquina, processamento de sinais, processamento de imagem, visão computacional,

(*Matrix Laboratory*), tem o objetivo de identificar imagens similares dentro de um banco de imagens por meio da comparação entre cada uma delas usando como parâmetro suas variáveis numéricas, como a média da saturação, média das variâncias, média dos histogramas de cada camada de cor, RGB, entre outros. O algoritmo tem como saída um arquivo de texto em formato ".txt" contendo o nome da imagem e um número que varia entre 1 e a quantidade de imagens que o *dataset* possui no total. Quando o script compara os parâmetros visuais já citados e define que uma imagem é similar a outra, ele repete essa variável de número.

Figura 21 - Exemplo de similaridade entre imagens

2.5. Aplicação da metodologia de coleta de imagem: app Cores da Copa

Em junho de 2014, o Brasil foi palco de outro grande evento que já vinha ganhando destaque tanto na imprensa internacional quanto nos questionamentos trazidos pelas manifestações de junho de 2013. A Copa do Mundo 2014 aconteceu no Brasil cercada de controvérsias e discussões sobre a capacidade do país de receber um evento desse porte mediante todos os problemas estruturais visíveis na prestação de serviços básicos à população brasileira (saúde, educação, saneamento básico, salário mínimo justo). Entretanto, conforme a data oficial da abertura da Copa se aproximava, o furor futebolístico foi tomando conta das redes declinando a hashtag #NãoVaiTerCopa e ascendendo sua antônima #VaiTerCopaSim.

A abertura da Copa aconteceu no dia 12 de junho de 2014 e seu encerramento no dia 13 de julho. Durante esses 32 dias, o laboratório reuniu uma equipe responsável por monitorar e produzir relatórios diários utilizando como base as imagens compartilhadas no *Twitter*. Os termos de coleta utilizados para compor o *dataset* foram "copa", "copa do mundo", "copa2014", "brasil2014", "worldcup"; hashflags representantes de cada

comunicações, finanças computacionais, design controle, robótica, e muito mais. Disponível em: <http://www.mathworks.com/index.html>.

país (#BRA; #ITA; #POR; #USA); partidas que iriam acontecer (#BRAxGER; #BRAxCRO); e nomes de jogadores icônicos como Messi, Neymar e Robben. O horário de extração foi das 20h até as 20h do dia seguinte, totalizando assim 24h de coleta.

A coleta dos *tweets* publicados nesse período foi a partir de um script chamado "Marcus", desenvolvido em linguagem *Python*[18] e em parceria com o cientista da computação André Panisson. O programa busca os tweets que contém os termos previamente selecionados e os armazena em um banco de dados do MongoDB[19], instalado em um servidor remoto. A cada 15 minutos o script Marcus coletou os *tweets* e o script de imagem Crawler[20] foi programado, nesse caso especificamente, para entrar no banco de dados, selecionar os *tweets* desse período, eliminar links externos ao *Twitter* (links de outros sites) e elencar os 100 links mais "frequentes" (nesse caso, quantidade de *retweets*) que foram postados diretamente no *Twitter* (pic.twitter.com). Depois desse procedimento, o script entrava em cada um desses 100 links, salvava a imagem e gerava uma tabela em formato *.csv* relacionando a url e a imagem salva, de forma a manter a conexão entre eles e documentar as informações para as futuras análises. Esse processo se repetiu a cada 15 minutos durante todo o período de 24h: se havia novos *tweets* entre os links mais frequentes, eles eram coletados; se fossem os mesmos, o script verificava se havia mudança nos retweets e atualizava a coluna que armazenava o valor de *retweet.*

Quando o processo de coleta acima se encerrava, o AISI varria a pasta em que se encontravam as imagens salvas e comparava os histogramas de cada uma, a fim de identificar imagens similares. Identificando essa semelhança, o script as compreendia como uma única imagem e somava suas frequências (número de *retweets*) de modo a levar em consideração a republicação de uma mesma imagem em diferentes links. Sem esse procedimento, cada uma dessas imagens possuiria menor peso na rede, ainda que tivessem grande ocorrência.

Ao longo de toda a duração do evento foi coletado um montante de 42.522 imagens (17.473 imagens únicas, após o AISI juntar imagens semelhantes) provenientes de cerca de dois milhões de links compartilhados diretamente no *Twitter*, considerando apenas as

[18] Linguagem de programação.
[19] Aplicação de código aberto, de alta performance, sem esquemas, orientado a documentos. Banco de dados utilizado pelo Labic. Disponível em: <http://www.mongodb.org/>.
[20] Disponível em: https://github.com/ufeslabic/crawler

imagens compartilhadas nos 100 tweets mais compartilhados a cada 15 minutos (TopTweets). Além dos relatórios diários contextualizando as imagens mais retweetadas do dia, o *dataset* da Copa rendeu a experimentação de novas formas de visualizações dinâmicas, conciliando tecnologia com *webdesign*. Os parâmetros considerados para o desenvolvimento das visualizações foram os tipos cromáticos (cor predominante, saturação e brilho), o número de *tweets* e *retweets* de cada imagem, e o horário no qual ela foi postada.

Com esses dados, foi desenvolvido o aplicativo nomeado "Cores da Copa[21]", que torna visível o ritmo cromático e foto-afetivo das imagens que circularam no *Twitter* durante a Copa do Mundo de 2014, com as imagens coletadas sendo dispostas em quatro tipos de visualização: "Calendário Cromático", "Timeline Cromática", "Mosaico Cromático" e "Mosaico de Imagens". Os dois primeiros são gráficos dinâmicos ativados pelo clique do mouse nos quais cada círculo colorido é uma imagem. Os parâmetros usados nesses casos são a gama de cor predominante (posição no eixo X e cor do círculo), o compartilhamento acumulado ao longo de todo o período (tamanho do círculo), e o compartilhamento em determinado horário (posição no eixo Y ou "altura" no gráfico). A diferença entre eles é que o "Calendário Cromático" apresenta uma visualização para cada dia do mês com intervalos de hora, enquanto que a "Timeline Cromática" é uma visualização do mês inteiro com intervalos de dias.

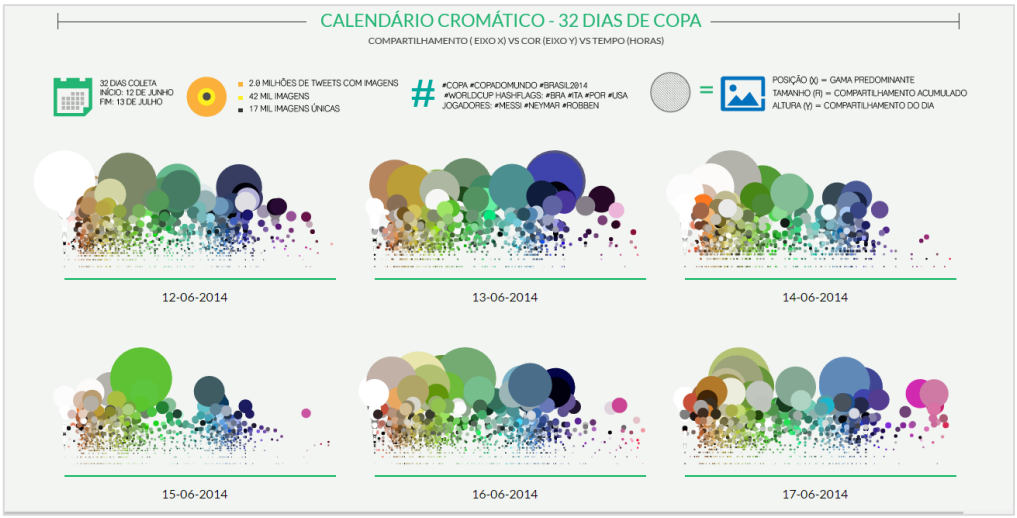

Figura 22 – Recorte do Calendário Cromático do app Cores da Copa. Fonte: Labic

[21] Disponível em: <http://www.labic.net/coresdacopa/>.

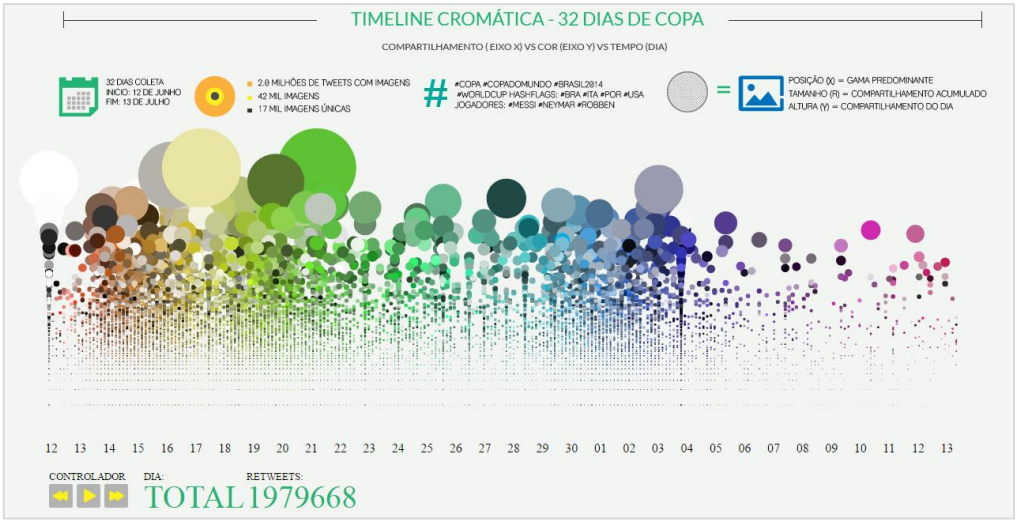

Figura 23 - Recorte do Timeline Cromática do app Cores da Copa

Já o "Mosaico Cromático" e o "Mosaico de Imagens" não possuem essa informação de tempo dando prioridade à cor e à interatividade, possibilitando que o usuário veja cada uma das imagens do mosaico com um passar de cursor por cima delas.

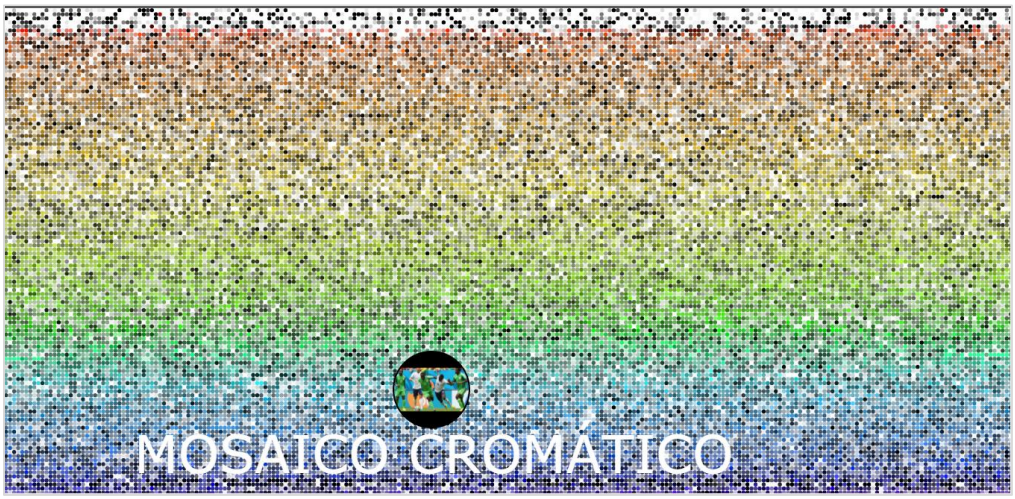

Figura 24 - Recorte do Mosaico Cromático do app Cores da Copa

2.6. Coleta e visualização de imagens do *Instagram*: Leticia e CartoDB

Nesse estudo sobre as imagens da Copa deu-se prioridade ao *Twitter*, haja visto o avançado estágio de desenvolvimento nas técnicas de coleta e processamento de dados oriundos desse site de rede social. Entretanto, a principal base do *Twitter* é semântica, sendo conhecido pelas suas postagens limitadas aos famosos 140 caracteres: há a possibilidade de se inserir imagens, mas o texto é o foco principal nesse caso. Com isso em mente, a linha de pesquisa em imagem do Labic decidiu abranger os estudos para

outros sites de redes sociais nos quais a imagem fosse protagonista nas narrativas. Assim, o *Instagram* foi escolhido como próximo passo para o desenvolvimento de uma metodologia de coleta própria, além de ser um exemplo da potencialidade de criação e replicação de imagens pelos perfis e páginas de usuários.

O primeiro passo seria verificar se já existiam maneiras de se coletar o conteúdo postado e entender como esses processos se realizavam. Dentro do próprio site oficial do *Instagram* existe uma aba específica para que desenvolvedores autônomos pudessem ter acesso aos chamados *endpoints* da API, que são pontos de entradas para algum serviço específico que seja requisitado pelo código do programador. Nesse caso, o *endpoint* é um url na qual é possível passar parâmetros aos quais esse comando retornará os dados pedidos. Um exemplo de *endpoint* seria "https://api.instagram.com/v1/tags/{tag-name}/media", no qual o conteúdo entre as chaves {} é substituído por qualquer hashtag que se deseje coletar.

Partindo da possibilidade de trabalhar diretamente com informações oficiais entregues pela própria API do site foi desenvolvido um script denominado "Letícia", que requisita a essa API uma hashtag pré-determinada pelo usuário e recebe de retorno todas as imagens que estiverem vinculadas a essa tag de coleta. O programa foi escrito em linguagem de programação java podendo funcionar nos sistemas operacionais Windows e Linux, rodando a partir de uma linha de comando no terminal. As limitações dele são de uma única hashtag por coleta e a falta de liberdade que o usuário tem em decidir quais campos serão coletados ou em delimitar o recorte de tempo visado.

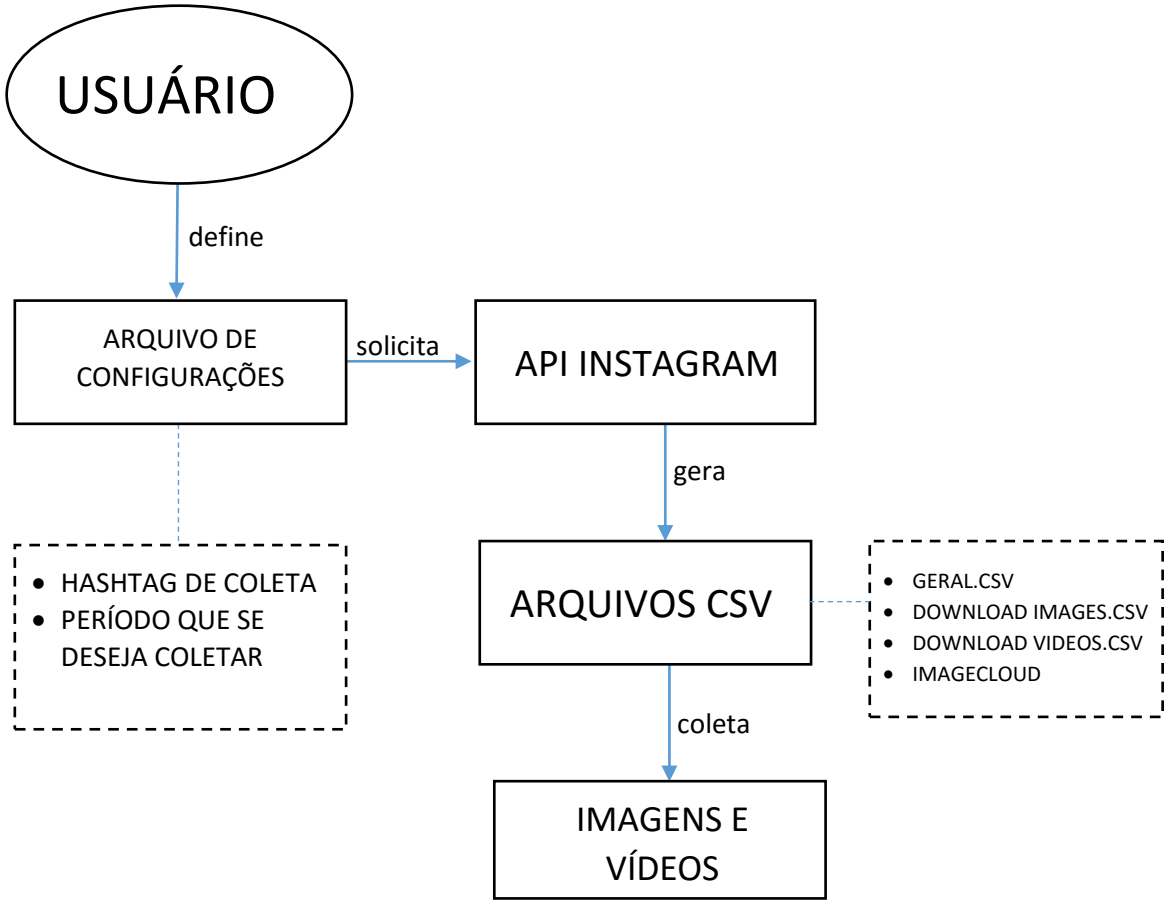

Figura 25 - Esquema de funcionamento da primeira versão do Leticia. Fonte: a autora

As limitações citadas acima e as necessidades de pesquisa que foram surgindo ao decorrer de seu uso motivaram o constante aperfeiçoamento da ferramenta.O script então foi remodelado em linguagem *python* com o objetivo de primeiro armazenar informações retornadas da API em um banco de dados e a partir disso iniciar a coleta das imagens. Algumas novas funcionalidades incluídas nessa atualização, como a coleta de múltiplas hashtags, maior facilidade de trabalhar com recortes no dataset coletado (faixas de datas específicas, quantidade superior a x de likes, determinado filtro, existência de geolocalização, entre outros), e a recuperação da coleta caso haja quedas de conexão.

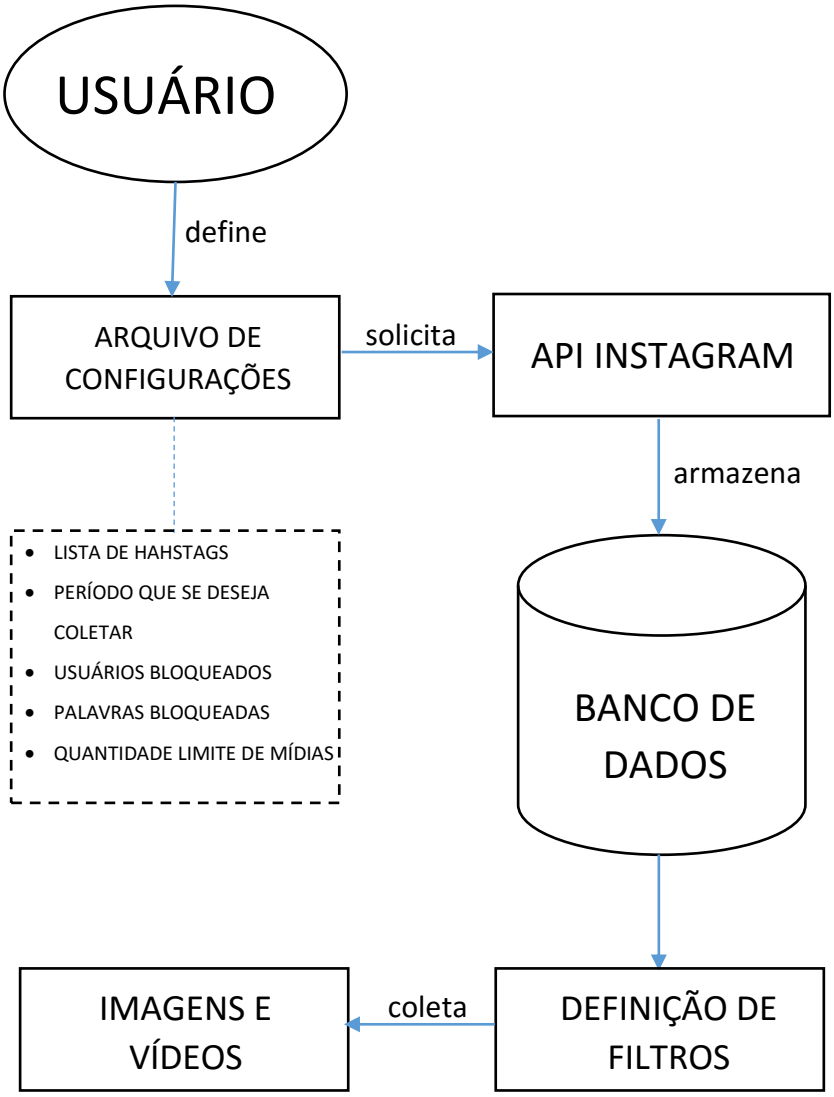

Figura 26 - Esquema de funcionamento da versão atualizada do Leticia. Fonte: a autora

Em comparação com os datasets advindos do *Twitter*, é possível notar que os usuários do *Instagram* tendem a vincular mais frequentemente a sua geolocalização às suas publicações. Essa prática abre leque para possibilidades de visualizações com mapas que envolvam o parâmetro de localização por região, dando respaldo a análises que incluam contexto geográfico e/ou cultural.

Para plotar os dados com geolocalização coletados buscou-se ferramentas que possibilitassem leituras de tabelas *.csv* e que apresentassem uma interface amigável e de fácil manuseio. Uma das ferramentas encontradas que atendeu as necessidades da pesquisa foi o editor CartoDB que, de acordo com a definição encontrada no site oficial,

é uma "ferramenta de auto-atendimento de mapeamento e análise que combina uma interface intuitiva com recursos de descoberta poderosos[22]". Ainda em sua descrição, consta que você pode mesclar e combinar seus datasets para obter novos *insights* sobre suas visualizações e não precisa ser um especialista para começar a mapear os dados, pois a interface simples de *point and click* lhe permite fazer tudo desde o design até a publicação do trabalho. Por se tratar de um produto de uma empresa privada, o potencial completo de seus recursos só é acessado mediante pagamento de planos contratados, entretanto para pesquisas pontuais e pequenos datasets, a versão grátis atende as expectativas.

Nessa conciliação da geolocalização, usada pelos usuários do *Instagram* e fornecida pela API oficial, com referencial geográfico visual é possível identificar em quais regiões determinada hashtag é mais popular, como foi seu comportamento com o passar do tempo e que tipo de conteúdos ou estilos de fotos são vinculados ao termo pesquisado. A primeira experimentação dessa técnica de visualização em mapas foi aplicada durante o acompanhamento da realização do Exame Nacional do Ensino Médio (Enem) no ano de 2014, realizado nos dias 08 e 09 de novembro.

O mapa foi criado com as postagens publicadas com a hashtag #enem no dia anterior à prova e nos dois dias de aplicação das provas. Cada círculo representa uma imagem postada e as cores variam entre os três dias coletados. Em um primeiro momento, ao olhar o mapa fica visualmente claro em quais regiões do país a tag foi mais utilizada, espalhando-se pela região litorânea e, no caso dos estados mais interioranos, concentrando-se pontualmente em suas capitais. A dinamicidade do mapa permite a comparação entre os três dias coletados, por meio dos filtros disponíveis, além de utilizar as funções do mouse para revelar informações como a foto em si, quantidade de *likes* e filtro usado.

[22] https://carto.com/

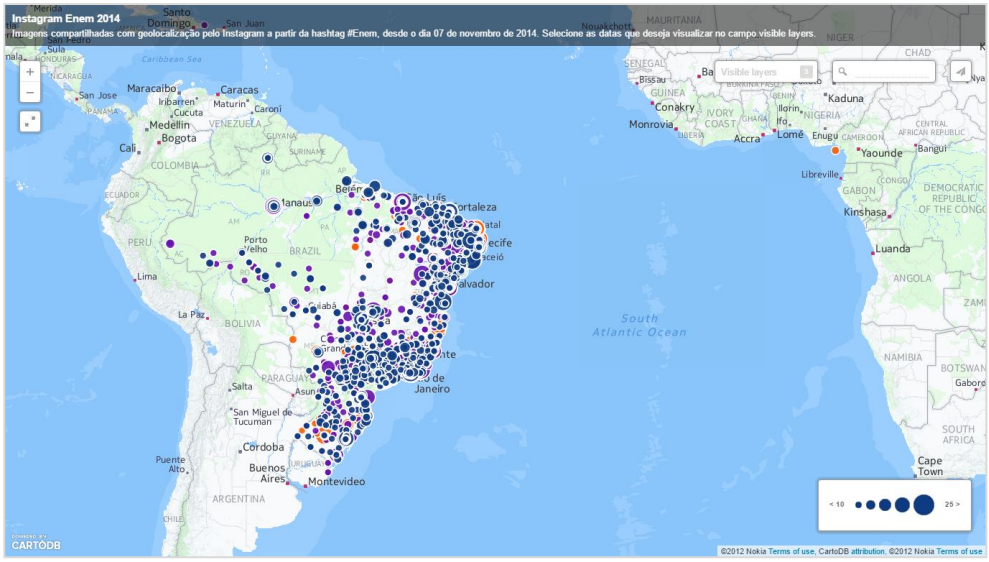

Figura 27 - Mapa criado com a ferramenta CartoDB com os tweets da #enem

A aparência final da plotagem em mapa é consequência direta do recorte do *dataset* com o qual o pesquisador decide trabalhar. O alcance da tag #enem se restringe ao território nacional, pois é bastante específica da área de educação do Brasil, aparecendo em postagens de outros países pontualmente ou relacionada a universidades estrangeiras que aceitam o exame como porta de entrada em seus cursos.

Diferentemente disso, outra pesquisa realizada pelo Labic, em parceria com o Fundo das Nações Unidas para a Infância (UNICEF), revelou o panorama da epidemia mundial da Zika, Chikungunya e Dengue nos sites de redes sociais. Por serem três doenças distintas porém transmitidas por um único vetor (mosquito Aedes Aegypti) e atingir diversos países em graus de intensidade diferentes, a visualização em mapa vai além de ser apenas mais um modo de mostrar dados brutos, mas torna-se essencial para explorar e manter a relevância da informação contida nas postagens. Padrões imagéticos e assuntos de destaque em regiões específicas do globo são realçados pelas postagens dos usuários do *Instagram* que constroem juntos a narrativa da tag. A análise desse dataset, bem como os produtos de visualização advindos dele serão melhor tratados no capítulo 3 desse trabalho.

Figura 28 - Mapa sobre Zika, Dengue e Chikungunya criado com a ferramenta CartoDB

2.7. Possibilidades futuras

Com o avanço nas técnicas de coleta e na visualização dos montantes de imagens publicados e republicados no *Twitter* e *Instagram*, foi posto em discussão o que mais seria interessante estudar ou obter para enriquecer futuras análises em *Big Data* e revelar novos vieses de pesquisas. Esse *brainstorm* gerou três processos metodológicos que foram idealizados porém ainda não postos em prática efetivamente: a coleta automatizada de textos vinculados à imagens; categorização de imagens usando *MatLab*; e criação de grafos cujo nós sejam imagens.

Utilizando-se a linguagem de programação Java e uma biblioteca externa (Jsoup), o script de coleta automatizada de texto e imagem acessa o código-fonte de uma página *web* em HTML, verifica quais textos se encontram mais próximos de onde a imagem está localizada, os coleta e os vincula àquela foto ou figura. Por fim, o *script* exporta um arquivo .csv com o nome da imagem na primeira coluna e o conteúdo do texto coletado na segunda coluna. Esse processo ainda se encontra em processo de desenvolvimento e ainda não foi aplicado à datasets coletados.

Muito se fez pensando na visualização das grandes quantidades coletadas e de como é relevante ter uma visão do todo para enxergar padrões outrora escondidos, logo é válido fazer o caminho inverso e se debruçar sobre cada imagem para pensar em modos de categorizar perante o conjunto inteiro. A categorização se faz importante como maneira de organizar as informações das imagens e possibilita recuperar essas informações, através de buscas com os termos das categorias. Além de podermos

analisar todas as imagens de um dataset, podemos com isso analisar as imagens de uma determinada categoria e ver quais as semelhanças e disparidades presentes nelas.

De modo arcaico é possível abrir cada arquivo de imagem no visualizador padrão de cada sistema operacional, ao mesmo tempo em que se edita uma tabela no excel, porém é um processo muito suscetível a erros humanos e ineficiente perante muitas imagens a serem categorizadas. Pensando nisso, foi criado um *script* para MatLab no qual as categorias definidas previamente pelos pesquisadores são inseridas no código vinculadas a uma letra do teclado. Quando se inicia o processo de categorização, o programa abre uma janela contendo uma imagem do dataset e o pesquisador precisa apenas apertar a tecla que corresponda a categoria desejada e pressionar "enter". A categoria então é inserida automaticamente à uma tabela do Excel criada pela própria ferramenta. O processo não deixa de ser manual, porém, muito mais ágil que o anterior.

Em paralelo ao estudo de imagens, a linha do laboratório que foca na parte semântica dos dados (tweets, posts de facebook) utiliza de um programa denominado *Gephi*[23] para a confecção de grafos de interação em rede. Infelizmente esse programa só suporta a entrada de informações textuais para criar as relações de nó e aresta, mas em 2014 foi lançado um plugin chamado *ImagePreview* no qual se tornou possível representar o nó por meio de uma imagem.

Com o intuito de testar suas capacidades, foi realizada uma extração através do *Netvizz*, ferrmenta de coleta integrada ao *Facebook*, dos 100 últimos posts da página Mídia Ninja nesse site. Desses posts, foram selecionados apenas os que continham fotos e, com a ajuda de dois outros plugins – os algoritmos de distribuição *Circular Layout* e *Noverlap*[24] , o grafo se organiza pela quantidade de compartilhamentos de cada post (conforme evidenciado pela espessura de suas arestas) e depois é exportado com as fotos a que cada nó faz referência. O objetivo agora seria encontrar formas de anexar diversas informações, como o número de curtidas, comentários e compartilhamentos a cada um dos nós, além da hora e o dia em que foram postados.

[23] Gephi é um software open-source para visualização e análise de redes. Ele auxilia analistas de dados a intuitivamente revelar padrões e tendências, realça discrepâncias e conta histórias com seus dados. Usa uma máquina de renderização em 3D para mostrar volumosos grafos em tempo-real e para acelerar a exploração. Disponível em: <https://gephi.org/about/>.

[24] O algoritmo "Circular layout" ordena os nós baseando-se em um de seus atributos; o "Noverlap" lê as posições x/y dos nós e o tamanho deles a fim de evitar sobreposições no grafo.

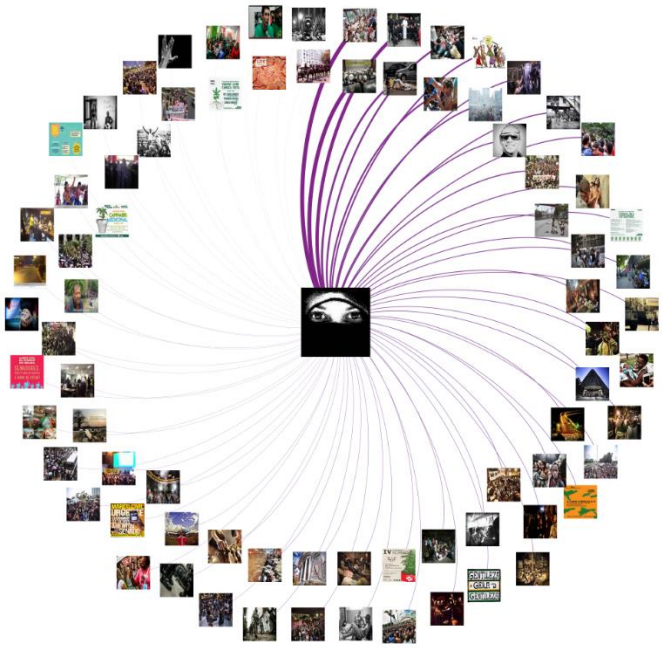

Figura 29 - Teste de Grafo de Imagens feito com o software Gephi

3. Estudo de caso: visualização da epidemia mundial do Zika no site *Instagram*

Tomando como contexto a quarta fase da informática proposta por Breton (apud LEMOS, 2015) como sendo a fase dos "internautas" e dos computadores conectados, juntamente com a a formação das comunidades virtuais e criação da inteligência coletiva (LÉVY, 1999), pode-se inferir que esse ambiente foi propício para a popularização e proliferação dos sites de redes sociais como *Orkut*, *Fotolog*, *MySpace*, e os gigantes ocidentais da atualidade *Facebook*, *Twitter*, *Snapchat*, *LinkedIn*, entre outros.

Com o sistema de produção de conteúdo transformado em uma relação "todos-todos" (LÉVY, 1999) e a facilidade de acesso aos dispositivos como câmeras e smartphones, bem como conectividade à rede online, cada internauta se torna um contador de histórias em potencial. A vantagem é poder compartilhar suas opiniões e cultura, de modo público, com o resto do mundo numa relação recíproca de aprofundamento de conhecimento. Com o foco no usuário, a cibercultura e o ciberespaço se tornaram ambientes propícios para o surgimento e proliferação dos chamados sites de redes sociais, sendo o *Instagram* o site escolhido para realizar um estudo de caso de aplicabilidade de visualizações de dados.

3.1. Sites de Redes Sociais: definição e primeiros sites

Os sites de redes sociais foram definidos por boyd e Ellison (2007) como

> serviços baseados em web que permitem aos indivíduos (1) contruir um perfil público ou semi-público dentro um sistema de fronteiras, (2) articular uma lista de outros usuários com os quais eles compartilham alguma conexão, e (3) visualizar e percorrer sua lista de conexões e as feitas por outros no mesmo sistema. (BOYD e ELLISON, 2007)

Para eles o que torna os sites de redes sociais únicos é a possibilidade do usuário conseguir visualizar e articular sua própria rede de amizades, criando um perfil próprio que exiba isso em rede. A maioria dos sites de redes sociais tem esses perfis online como suas espinhas dorsais, sendo cada uma original em si mesma, pois é onde o usuário pode "escrever-se para existência" (SUNDÉN apud BOYD e ELLISON, 2007).

Essa definição de sites de redes sociais é compartilhada também pela pesquisadora Raquel Recuero (2009) que ainda propõe duas categorias de sites de redes sociais: os propriamente ditos, e os apropriados. Os propriamente ditos seriam aqueles "cujo foco principal está na exposição pública das redes conectadas aos atores, ou seja, cuja finalidade está relacionada à publicização dessas redes", enquanto que os sites de redes sociais apropriados "não eram, originalmente, voltados para mostrar redes sociais, mas que são apropriados pelos atores com este fim".

De acordo com Boyd e Ellison (2008), a primeira SNS (*social network site*) que se encaixa na definição proposta foi o *SixDegree.com*[25], em 1997, com suas funções de criar perfis, listar amigos e navegar pelas listas de outros amigos. Anteriormente, outros sites já apresentavam a chance de se criar perfis (como sites de encontro), ou de listar amigos (como o famoso ICQ), porém o *SixDegree.com* foi o primeiro a combinar as três funcionalidades em um sistema só. Entre 1997 e 2002, várias ferramentas de comunidade foram surgindo: *AsianAvenue*[26], *BlackPlanet*[27], *MiGente*[28] e o mais significativo da época *Friendster*[29], considerado por Chafkin (apud BOYD e ELLISON, 2008) "uma das maiores decepções da história da internet". A partir de 2003, a maioria dos sites de redes sociais mais conhecidos foram lançados: *LinkedIn*, *MySpace*, *Last.FM* e *Orkut* (2003); *Flickr (2004); YouTube* (2005); *Facebook* e *Twitter* (2006); e no final de 2010, o *Instagram*.

[25] http://sixdegrees.com/
[26] http://www.asianave.com/ - Não compatível com Google Chrome
[27] http://www.blackplanet.com/
[28] http://www.migente.com/
[29] http://www.friendster.com/ - Desativado desde 2015

3.2. Aplicabilidade das *InfoVis* no caso Zika Virus do *Instagram*

Criado em outubro de 2010[30] e à época só compatível com o sistema operacional IOS, o site de rede social *Instagram* é voltado para publicação de imagens e vídeos por seus usuários, podendo haver aplicação de filtros fotográficos e compartilhamento em outros sites de redes sociais, como *Facebook* ou *Twitter*.

A cibercultura (LÉVY, 1999) não só permitiu que os internautas pudessem emitir suas opiniões como também pudessem compartilhar seus olhares, suas visões de mundo, por meio de seus registros fotográficos. O que antes era considerado uma prática exclusiva de profissionais da área da fotografia, se transformou em algo alcançável pelo amador que possui uma câmera portátil ou mesmo um smartphone com câmera.

> A massificação amadora[31] da fotografia, e sua renovada visibilidade online, sinaliza uma mudança na valorização da cultura fotográfica. Se no passado o espaço da fotografia pública era dominado por praticantes profissionais, atualmente o trabalho desses especialistas aparece lado a lado de imagens produzidas por indivíduos que não possuem o mesmo investimento profissional em fotografia (Tradução própria. RUBINSTEIN e SLUIS, 2008)

Conciliando essa prática comum da fotografia amadora, o ato de publicar e compartilhar conteúdo online e a presença de perfis e redes conectadas de seguidores, o *Instagram* se tornou um dos maiores sites de redes sociais voltado para conteúdo imagético atingindo recentemente[32] a marca de 500 milhões de usuários ativos – desses, 300 milhões utilizam o *Instagram* diariamente.

O modelo dos posts desse site de rede social é formado basicamente por fotos postadas no perfil do usuário, que também aparecem no feed de seus seguidores. A "página" principal de cada perfil é composto por um mosaico de fotos quadradas (recentemente, foi liberada a postagem de fotos em formatos retangulares) que possuem ou não uma legenda, hashtags e geolocalização. Cada foto postada é um discurso, algo que o usuário deseja mostrar para seus seguidores:

> no nível mais básico, cada ação de dar upload em uma imagem contém uma recompensa em potencial – sempre existe a possibilidade de que alguém vai vê-la e apreciá-la; a recompensa é entregue em forma material se outro usuário deixar um comentário ou marcar a imagem como favorita (BURGESS apud RUBINSTEIN e SLUIS, 2008)

[30] http://blog.instagram.com/post/8755272623/welcome-to-instagram
[31] *Mass amateurization* (SHIRKY, Clay. 2008)
[32] http://blog.instagram.com/post/146255204757/160621-news

A metodologia utilizada para coleta foi com base no script “Leticia”, citado no capítulo 2 deste trabalho, que captura imagens e vídeos do *Instagram*. Para a execução do programa, primeiro deve se configurar um arquivo de texto padrão que contém os parâmetros de coleta que o pesquisador deseja utilizar para montar seu *dataset*.

Alguns dos campos de parâmetros a serem configurados são:

- *Block_users*: uma lista de usuários do site que são bloqueados na hora da coleta. Muito útil pra evitar *flood* de postagens de perfis considerados como *bots*.
- *Minutes*: até quantos minutos para trás o pesquisador gostaria que o script recolhesse as imagens e vídeos.
- *Tags*: lista de tags de coleta.
- *Users_search*: lista de usuários a serem coletados.
- *Block_words*: lista de palavras bloqueadas que apareçam na legenda ou comentário de qualquer mídia.
- *Max_collect*: delimita a quantidade de mídias a serem coletadas.

Após a configuração dos parâmetros, o script começa a coleta e gera um arquivo .csv. No caso deste trabalho, o objeto escolhido para pesquisa foram as imagens acerca da epidemia mundial de Dengue, Zika e Chikungunya. Foram escolhidas 18 tags referentes a esse tema: “forazika”, “forazikavirus”, “zikazero”, “zikazerobrasil”, “zika”, “zikavirus” “microcefalia”, “dengue”, “denguemata”, “denguenao”, “aedesaegypt”, “aedesaegypti”, “chikungunya”, “aedes”, “combateaedes”, “guillainbarre”, “GuillainBarreSyndrome” e “fightaedes”. Em um intervalo de tempo de 13 meses (março de 2015 a março de 2016), delimitado para fins da pesquisa, foram capturadas 66 405 mídias entre imagens e vídeos. Esses 13 meses estão divididos em quatro períodos de tempo: março a maio; junho a agosto; setembro a novembro; e dezembro – 2015 a março – 2016.

3.2.1. ImageClouds e CartoDB

Com os arquivos armazenados na memória do computador é possível então criar os *ImageClouds*, o que possibilita ao pesquisador ver o conjunto como um todo e analisar os tipos imagéticos e possíveis discursos que ali aparecem.

Figura 30 - ImageCloud de Março a Maio de 2015, ordenado por quantidade de curtidas

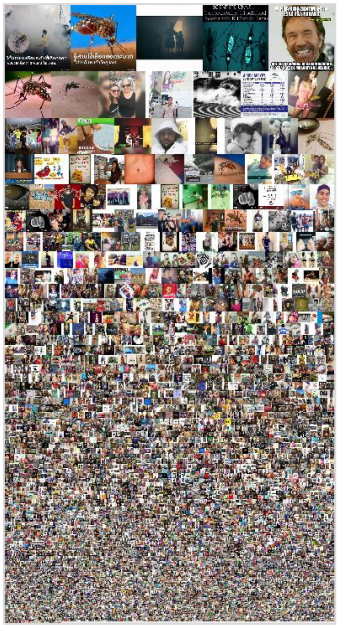

Figura 31 - ImageCloud de Junho a Agosto de 2015, ordenado por quantidade de curtidas

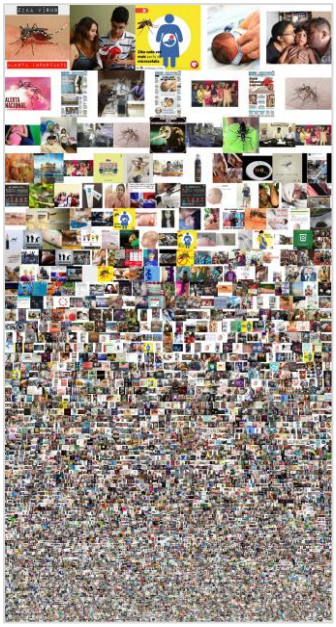

Figura 32 - ImageCloud de Setembro a Novembro de 2015, ordenado por quantidade de curtidas

Figura 33 - ImageCloud de Dezembro (2015) a Março (2016), ordenado por quantidade de curtidas

As visualizações acima permitem identificar rapidamente quais imagens são as mais curtidas de qualquer *dataset*, além de oferecer um panorama do que mais se fala sobre aquele assunto naquele determinado período de tempo. No contexto das doenças transmitidas pelo mosquito Aedes Aegypti, é importante ressaltar o volume das imagens nesses quatro recortes de tempo definidos. Nos três primeiros períodos foram coletadas cerca de 10 mil imagens em cada um, enquanto que no último período (dezembro 2015 a março 2016) foram coletadas mais de 40 mil. Esse aumento significativo no número de publicações se deve ao início do verão no hemisfério sul, juntamente com o início do surto epidêmico do Zika vírus. Com o número de casos aumentando, os órgaos responsáveis pela saúde começaram a se manifestar, conscientizando a população sobre os perigos da doença.

No mosaico predominam fotos do mosquito em si (característico pelo seu corpo com listras brancas), informacionais que vão desde como evitar focos da doença nas residências a lista de sintomas que aparecem em pessoas contaminadas, fotos de ações conscientizadoras como palestras ou reuniões, fotos de ações em campo de grupos de agente da saúde, selfies de pessoas hospitalizadas, e as mais populares são de artistas brasileiros aderindo à campanha #NãoFicoParado, da Ambev, que incentiva as pessoas a virarem garrafas para evitar criadouros do Aedes.

A geolocalização disponibilizada pelo *Instagram* é um recurso útil para identificar de que locais são postadas as imagens capturadas (desde que o usuário tenha ativado essa função. Do montante de 66 405 mídias capturadas, 12 540 imagens possuíam o georeferenciamento vinculado às fotos, possibilitando assim a visualização em mapa utilizando o CartoDB, ferramenta online (citada no capítulo 2) que permite a inserção de dados e a subsequente criação de mapas dinâmicos.

Para fins de organização da análise, o mapa final[33] foi dividido em regiões para que se pudesse descobrir as temáticas presentes nelas. As regiões definidas foram “Brasil”, “América do Sul (Exceto Brasil)”, “América Central” e “América do Norte”.

[33] Disponível em: <https://unicef-zika.cartodb.com/viz/1769f152-dfba-11e5-ad87-0e3a376473ab/public_map>.

3.2.2. Mapa de Imagens: Brasil

Figura 34 - Recorte do Brasil (CartoDB)

A região Norte foi a que teve a menor quantidade de postagens dentre as regiões brasileiras, com as imagens se concentrando nas capitais de cada estado (Manaus, Boa Vista, Macapá, Belém, Porto Velho, Rio Branco e Palmas). O conteúdo das imagens variam entre a arte da campanha #ZikaZero lançada pelo Governo Federal, e fotos da população aderindo à luta contra o mosquito. Fotos de palestras com a comunidade, campanhas de conscientização envolvendo crianças e selfies feitas pelo grupos de combate ao mosquito (Marinha e Exército) também aparecem com frequência na região.

Figura 35 - Exemplo de postagem *Instagram*: Região Norte[34]

A segunda região com menos postagens no *Instagram* foi a região Centro-Oeste. Apesar da quantidade de posts ainda se concentrar nas capitais dos estados (Goiânia, Brasília, Campo Grande e Cuiabá), vemos que pelo interior também se espalham

[34] https://www.instagram.com/p/BCBGPCdnN4y/

postagens sobre a mobilização da população contra a proliferação da doença. Aparecem imagens do combate em campo, como intervenções em terrenos baldios e visita das equipes de saúde aos moradores da região; orientações e recomendações sobre como evitar a presença de focos do mosquito nas residências; e algumas declarações de pessoas que já foram diagnosticadas com dengue, mostrando os braços com intravenosas de soro.

Especificamente em Brasília, algumas fotos foram publicadas por perfis de senadores, deputados e ministros (ex.: @eduardoamorimse, @fabioreis1515, @antoniobritobahia, @luciano_s_castro, @ministro_george_hilton, @clarissagarotinho), e pelo próprio perfil oficial do Ministério da Saúde (@minsaude). A maioria são fotos de falas em audiências públicas sobre a dengue e a zika, e reuniões para tomadas de medidas perante a situação (ex.: a Reunião Bilateral Brasil - EUA realizada dia 18 de fevereiro).

Figura 36 - Exemplo de postagem *Instagram*: Região Centro-Oeste[35]

O tipos de postagens que aparecem nas regiões citadas anteriormente também se repetem nos estados do Nordeste. A adesão à campanha do #ZikaZero é visível, com as comunidades se organizando para combater os focos do mosquito, por meio de denúncias de terrenos abandonados ou ações conscientizadoras em palestras. Um assunto destaque nessa região do país é o da microcefalia. Estão presentes publicações de fotos de bebês que nasceram com a deficiência, explicações didáticas sobre o que é a microcefalia e que tipo de efeitos ela causa no feto, recomendações de proteção às gestantes (como uso de repelentes) e informações sobre a possível relação entre a doença e o Zika vírus.

[35] https://www.instagram.com/p/BB71s6FjuH5/

Figura 37 - Exemplo de postagem *Instagram*: Região Nordeste[36]

Como região mais populosa do país (5.668.232 habitantes[37]), a Região Sudeste também apresenta a maior quantidade de imagens publicadas na rede do *Instagram*, especificamente na Região Metropolitana de São Paulo. Devido a esse grande volume de posts, a variedade de conteúdo também aumenta abarcando todos os tipos já mencionados até então: informacionais, institucionais, recomendações e orientações, metódos de prevenção como a utilização de repelentes, selfies de quem recebeu o diagóstico de dengue ou zika, informações sobre a microcefalia, e a divulgação da campanha nacional de combate. De peculiar nessa região foi a aparição de alguns memes de internet fazendo piada a respeito da situação (ex: um mosquito fala para outro "Transmito Dengue, Zika e Chikungunya. Já posso pedir música no Fantástico?") e pontuais preocupações durante o período de carnaval, com dúvidas sobre a possível transmissão pela saliva.

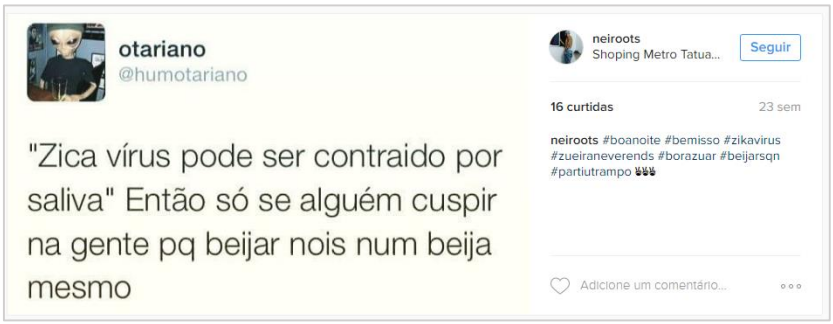

Figura 38 - Exemplo de postagem *Instagram*: Região Sudeste[38]

Já na Região Sul, o padrão é similar ao da Região Norte (imagens concentradas nas capitais), entretanto também se propagam ao longo do litoral. O conteúdo dos posts é o mesmo de todas as outras regiões brasileiras. Entretanto, o termo Zika se encontra ligeiramente mais "poluído", ou seja, com outros significados à palavra "Zika". Estes

[36] https://www.instagram.com/p/-uasV_PIQI/
[37] http://www.censo2010.ibge.gov.br/sinopse/index.php?dados=8
[38] https://www.instagram.com/p/BBqh5niAYJL/

referenciando-se ao surto da doença, ou como gíria popular em tags. Este último comportamento pode contaminar o conjunto de dados imagéticos, sendo necessário um trabalho de aprofundamento analítico, contextualizando e delimitando refinamento da tag.

Figura 39 - Exemplo de postagem *Instagram*: Região Sul[39]

Se o recorte da região do Brasil for comparado com o restante do mapa, se torna claramente visível que a quantidade majoritária de postagens referentes às tags de coletas delimitadas foram publicadas em território brasileiro, local no qual o surto epidêmico teve maior visibilidade, atraindo a atenção da mídia mundial. Claro que se deve levar em consideração a população do país, bem como a quantidade de usuários do site de rede social *Instagram*, mas ainda assim a observação de que no Brasil a tag foi mais popular é louvável. Um fator que pode estar associado a isso é o fato de a Dengue ser uma preocupação recorrente nos verões brasileiros, sendo transmitida pelo mesmo vetor do Zika e Ckikungunya: o Aedes Aegypti.

[39] https://www.instagram.com/p/0F0YwygkNJ/

3.2.3. Mapa de Imagens: América do Sul (exceto Brasil)

Figura 40 - Recorte da América do Sul (CartoDB)

No Uruguai, a quantidade de posts é pequena (menos de 20 fotos em todos o território uruguaio), e fazem menção tanto à dengue quanto ao zika. Sobre a dengue, o Uruguai confirmou seu primeiro caso autóctone em fevereiro[40] e, como atual presidente do Mercosul, convocou uma reunião em Montevideo[41] com os Ministros Latino-americanos da Saúde, representantes da CELAC e funcionários da Organização Pan-Americana da Saúde para discutir sobre o surto do Zika vírus.

Na Argentina e no Chile, a maioria das imagens postadas fazem menção apenas à dengue e ao mosquito Aedes, não havendo um destaque ao Zika e a Chikungunya. Essa falta de presença das duas últimas doenças se deve ao fato de que a Argentina vive um surto de dengue, com mais de 10 mil casos confirmados[42], enquanto que os números de casos de zika (autóctones e importados) permanecem baixos se comparados com outros lugares. O mesmo pode ser dito do Peru, cujas imagens mais marcantes foram a de medidas preventivas como o fumigamento de áreas com focos do mosquito Aedes Aegypti.

[40] http://saude.terra.com.br/uruguai-confirma-primeiro-caso-de-dengue-autoctone-no-pais,b295d5daeee002f0dba549b215dfb802n392gdrd.html
[41] https://www.instagram.com/p/BBVHVY6qZ7j/
[42] http://saude.terra.com.br/sobe-para-10-mil-o-total-de-casos-confirmados-de-dengue-na-argentina,319e934df1b81be2afd8cdfa9157afdfwnzex503.html

Na fronteira entre Paraguai e Argentina, mais precisamente na província argentina de Misiones, duas cidades chamam atenção pela concentração de pontos: Posadas e Puerto Iguazú. As duas cidades são as que mais reportaram casos confirmados de dengue na região: de 2450 casos, 60% aconteceram em Posadas e 30% em Puerto Iguazú[43], explicando o porquê delas estarem marcadas no mapa.

Na parte superior da América do Sul, os países que mais interagiram com as tags coletadas foram Venezuela, Colômbia e Equador. As imagens se proliferam em uma faixa que vai desde Caracas e Maracaibo (Venezuela), pasaando por Bogota e Cali (Colômbia) e chegando em Quito e Guavaquil (Equador).

No Equador, é possível identificar no mapa alguns pequenos aglomerados de imagens em Quito, capital do Equador, Guayaquil, maior cidade e principal porto do país, e Portoviejo. Nas duas últimas cidades foram detectados os dois primeiros casos autóctones[44] de zika no Equador, e em Quito haviam sido confirmados dois casos importados da doença[45].

Figura 41 - Exemplo de postagem *Instagram*: Equador[46]

De acordo com boletins epidemiológicos[47] liberados pelo Instituto Nacional de Salud (INS) da Colômbia, até o momento foram notificados 17.898 casos de chikungunya, sendo desses 17.707 confirmados (98,9%). No dataset, percebe-se que diferentemente de outros países a doença mais citada na Colômbia é a chikungunya, que

[43] http://www.clarin.com/sociedad/Murio-mujer-dengue-confirmaron-zika_0_1517848587.html
[44] http://www.eluniverso.com/noticias/2016/01/15/nota/5346097/detectan-dos-casos-autoctonos-zika-guayaquil-portoviejo
[45] http://www.eluniverso.com/noticias/2016/01/10/nota/5338744/msp-confirma-dos-casos-importados-zika-quito
[46] https://www.instagram.com/p/BB-AznCCvEb/
[47] http://www.ins.gov.co/boletin-epidemiologico/Boletn%20Epidemiolgico/2016%20Bolet%C3%ADn%20epidemiol%C3%B3gico%20semana%2028.pdf p. 88 a 91.

chegou antes do Zika e já habia deixado o país em estado de alerta. Com a proliferação de Zika em países vizinhos, reforçou-se o combate do Aedes por ele ser o transmissor das duas doenças além da dengue. A confirmação de circulação do Zika vírus no país se deu na semana do dia 04 ao dia 10 de outubro de 2015 (semana epidemiológica 40) e até a presente semana foram notificados 8.826 casos confirmados e 90.895 casos suspeitos em todo território colombiano. Um destaque é que foram reportados 5.904 casos confirmados de zika em mulheres grávidas, e 11.826 sob suspeita, desde o início do período epidêmico, com confirmação de 21 casos de bebês com microcefalia.

Figura 42 - Exemplo de postagem *Instagram*: Colômbia[48]

As imagens publicadas na Venezuela são em sua maioria de ações preventivas contra os focos do mosquito, como o fumigamento nas ruas e dentro das casas, e informacionais sobre o avanço do surto de zika na América Latina e sobre os sintomas que pessoas infectadas apresentam.

O país vive no momento uma "crise humana de saúde", declarada pela Assembléia Nacional da Venezuela, e "sofre com a escassez de remédios, insumos médicos e infraestrutura humanitária"[49]. Somando a isso está o fato de que o governo venezuelano não divulga os boletins epidemiológicos desde 04 de julho de 2015[50], dificultando análises mais profundas da quantidade de casos oficiais reportados.

3.2.4. Mapa de Imagens: América Central

[48] https://www.instagram.com/p/BBlTE0lsuHS/
[49] http://internacional.estadao.com.br/noticias/geral,assembleia-nacional-declara-crise-humana-de-saude-na-venezuela,1825959
[50] http://www.npr.org/sections/goatsandsoda/2016/02/02/465246295/so-how-many-zika-cases-are-there-in-venezuela-4-000-or-400-000

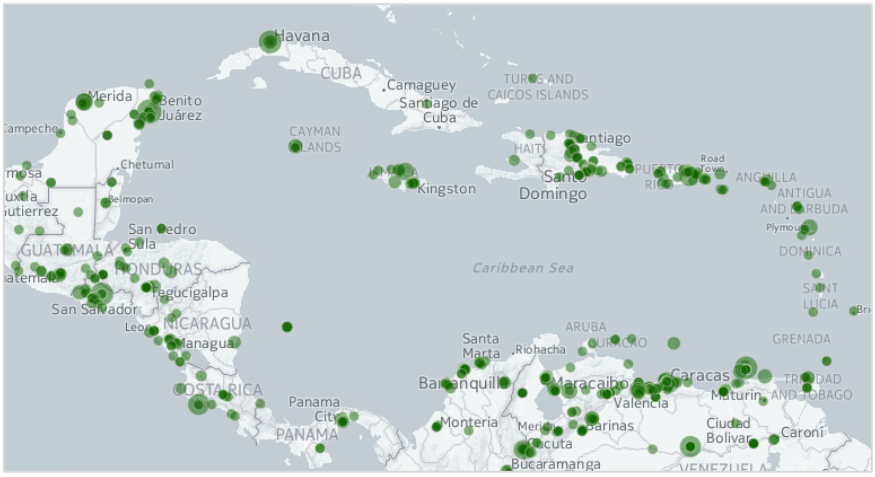

Figura 43 - Recorte da América Central (CartoDB)

Nas ilhas do Mar do Caribe, de Cuba a Trinidad e Tobago, as imagens refletem nações e populações preocupadas com o surto epidêmico na América Latina. Por se tratar de uma área no qual o turismo possui papel fundamental na economia local, é de interesse comum que todas as medidas preventivas sejam tomadas. Ações comunitárias para encontrar e eliminar focos do mosquito estão sendo realizadas, além do uso de repelente e de mosquiteiros em volta das camas.

A parte continental reúne a Guatemala, Belize, Honduras, Nicaragua, Costa Rica e Panamá, que possuem imagens com temas bem semelhantes entre si. Além da conscientização da população em eliminar os focos da dengue e se proteger com repelentes e mosquiteiros, há também o relato de viajantes e turistas que se inteiram sobre o assunto e postam fotos mostrando a situação pela qual estão passando em solo estrangeiro.

Figura 44 - Exemplo de postagem *Instagram*: Caribe[51]

[51] https://www.instagram.com/p/_m9I-smBgW/

3.2.5. Mapa de Imagens: México e Estados Unidos

Figura 45 - Recorte do México (CartoDB)

Houve 65 casos de Zika confirmados em solo mexicano sendo 35 em Chiapas, 21 em Oxaca, quatro em Nueva Leon e Jalisco, Sinaloa, Guerrero, Veracruz e Yucatan apresentaram um caso cada um. O *Mexico Tourism Board* divulgou um mapa[52] apontando que áreas são consideradas passíveis de se adquirir a doença, e afirmou que altitudes mais altas não são ambientes comuns do Aedes, sendo esses locais menos prováveis das pessoas serem infectadas. Em questão de imagem, o México não teve muita representatividade no mapa durante o período pesquisado: além das muitas selfies não relacionadas ao tema do Zika, havia fotos de pessoas tomando soro com medicamentos, imagens informacionais sobre os sintomas do zika, chikungunya e dengue e fotos de mosquitos, mortos ou ilustrativas.

Figura 46 - Recorte do Estados Unidos (CartoDB)

[52] http://www.travelweekly.com/uploadedFiles/PDFs/T0215MEXICOZIKAMAP.pdf

Até o momento foram relatados 1.658 casos de zika em solo americano, todos importados de zonas de surto, não sendo autóctones[53]. As regiões com mais posts publicados são o lado leste dos Estados Unidos, e a Califórnia. Quatro conglomerados ficam mais visíveis: nas cidades de Nova York, Filadélfia, e Washington D.C; a península da Flórida; o estado do Texas e o litoral da Califórnia.

As três primeiras cidades são grandes pólos urbanos, e consequentemente, com maiores tendências a serem mais ativas nas redes. Instituições como o National Pediatric Center, a Organização Panamericana de Saúde e a University of the Science divulgaram informações sobre a proliferação da doença assim como orientações para a população. A maioria das imagens não trata diretamente dos casos de americanos que contraíram a zika, mas fala sobre a situação nos países da América Latina e o perigo que existe da doença se espalhar pelos Estados Unidos.

Figura 47 - Exemplo de postagem *Instagram*: Estados Unidos[54]

A Florida foi o segundo estado americano que mais apresentou casos de zika importados: 307 casos relatados na região. Assim como nas cidades anteriores, também há a preocupação em informar sobre as doenças que o aedes transmite, para isso emissoras de TV buscam médicos especialistas para comentarem sobre o assunto. No Texas, o avanço nas pesquisas sobre o Zika virus seguem na busca por uma vacina que seja eficaz no combate de uma possível epidemia: visivelmente essa é a maior preocupação dos americanos. 76 casos foram confirmados no estado e doutores e professores de universidade são convidados de programas de TV e entrevistados nos jornais locais, sendo questionados sobre a epidemia global

[53] http://www.cdc.gov/zika/geo/united-states.html
[54] https://www.instagram.com/p/BBIFmMlGrCs/

Figura 48 - Exemplo de postagem *Instagram*: Florida[55]

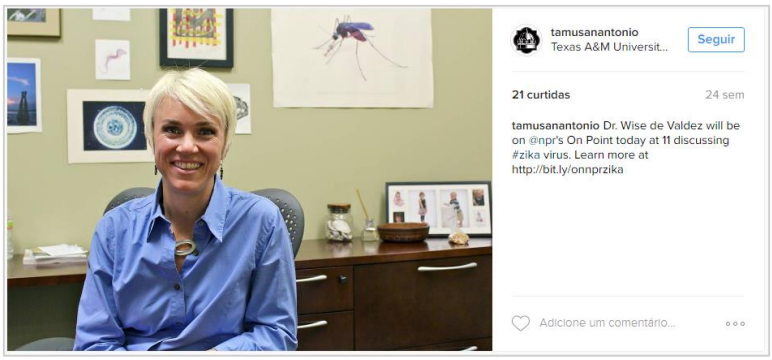

Figura 49 - Exemplo de postagem *Instagram*: Texas[56]

O último conglomerado, o da Califórnia, reflete o que já vem sido mostrado no restante do país: preocupação com a zika, especialistas sendo procurados, palestras informacionais sendo dadas, entre outros. Nota-se que não há ações comunitárias nem medidas combativas como fumigamento nas ruas e casas. O que há é a recomendação de não produzir focos do mosquito dentro de casa, e ter atenção redobrada em viagens às áreas de risco.

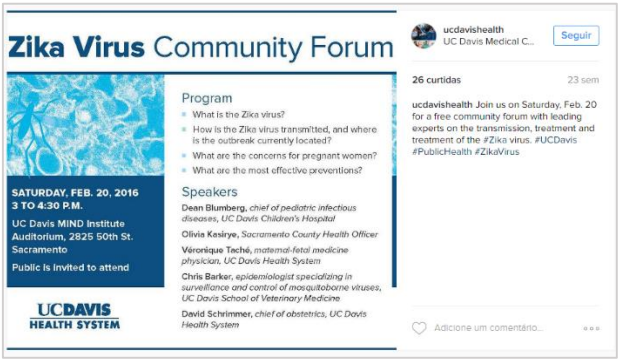

Figura 50 – Exemplo de postagem *Instagram*: Califórnia[57]

[55] https://www.instagram.com/p/BBS461gD3c4/
[56] https://www.instagram.com/p/BBnEVqvCg7Z/

CONSIDERAÇÕES FINAIS

"Se antigamente, o poder de transmitir informação estava reservado apenas a um pequeno nicho de entendidos, atualmente, esta pertence a todos quantos tiverem disponibilidade e vontade de informar" (AROSO e CORREIA, 2007). Levando esse trecho em consideração, percebe-se que na conjuntura contemporânea, o papel de contador de história do jornalista não é o mesmo de 50 anos atrás. Hoje, a produção de conteúdo está na mão de cada pessoa que deseje produzir ou que deseje compartilhar a sua visão do mundo com a sua rede de amigos e conhecidos. Entretanto, isso não significa que o jornalismo perdeu sua função principal como provedor de informação, mas apenas que, com o advento da cibercultura (LÉVY, 1999), há muito mais lugares para se olhar e para se apurar em buscar das notícias. E é preciso saber como fazer isso.

O *Big data* pode parecer apenas uma expressão para se designar uma grande quantidade de dados, mas é muito mais que isso. Essa produção em massa de conteúdo produzido pelos internautas e compartilhado nos sites de redes sociais abre um leque de oportunidades para serem trabalhadas e exploradas pelos jornalistas de modo a possibilitar a compreensão dos discursos nas redes digitais e até a descoberta de pautas guiadas pelos padrões enxergados na forma como os dados se organizam.

A fronteira entre as pesquisas das ciências humanas e das ciências exatas não mais é nitidamente definida, pois há agora um território que busca conciliar as habilidades de cada uma delas, trabalhando em conjunto para se debruçar sobre esse ambiente dinâmico e vasto que é o ciberespaço – trata-se de um novo campo do conhecimento chamado Humanidades Digitais. Nota-se que o jornalismo de dados ainda não é uma prática comum na formação dos jornalistas e é nesse aspecto que a profissão precisa ser direcionada para entender e contar histórias de um mundo conectado. Há uma infinidade de dados sendo produzidos diariamente cujo potencial de contar histórias não é tão bem aproveitado por não serem coletados, processados e visualizados de uma forma que seu discurso seja compreendido.

No estudo de caso da epidemia mundial de Zika, Chikungunya e Dengue presente nesse trabalho percebe-se o quão rico é trabalhar com dados de sites de redes sociais quando visualizados e agrupados por determinado parâmetro para se enxergar possíveis padrões. O que antes eram apenas várias imagens postadas por diferentes usuários de

[57] https://www.instagram.com/p/BB8k_leAwu4/

diferentes localidades do mundo, são agora parte de um banco de dados que visualizado em forma de mapa, permite que seja feita uma análise panorâmica (ou até mesmo profunda) por cada região. Foi na criação de visualizações que surgiu a possibilidade de enxergar que o assunto da microcefalia foi destaque na região Nordeste, diferentemente das outras regiões, pois foi nela que houve maior ocorrências de bebês afetados pela enfermidade. Ou mesmo que na Argentina e no Chile, o Zika Virus não era um assunto de extrema urgência, pois esses países viviam um surto de dengue mais preocupante que as novas doenças transmitidas pelo Aedes.

Não é apenas a visualização em mapa que revela novas informações, mas existe uma gama de outros parâmetros que podem ser aproveitados na descoberta de relações entre os dados. As nuvens de imagens, por exemplo, permitem identificar quais os tipos imagéticos que foram mais populares (maior número de likes, no caso do *Instagram*) além de mostrar ao pesquisador ou jornalista qual a a aparência que os usuários do site deram àquele assunto. Os gráficos do *ImageJ/ImagePlot* conseguem processar quantidades muito mais volumosas de informação do que programas mais simples, eliminando a dificuldade que é trabalhar com o o fluxo giganteco do que vem da internet. Quando se pode enxergar os dados com os quais se trabalha, o surgimento de hipóteses e relações de interação ficam mais ao alcance de serem reveladas e recontadas por meio de notícias ou mesmo de infográficos.

Anderson, Bell e Shirky (apud BERTOCCHI, 2014) afirmam que

> "temos um panorama mediático no qual mais técnicas serão adotadas na produção de notícias: análises algorítmicas de base de dados, visualização de dados, solicitações de conteúdos por parte de amadores, produção automatizada de narrativas, criação de narrativas baseada em dados, entre outros"

Cabe ao jornalista aprender a lidar com esse novo tipo de dado digital cujo formato é intimamente ligado à linguagem de programação contemplada pelas ciências exatas. A questão é buscar ao menos uma base de informática que permita uma busca simples pelas ferramentas grátis disponíveis, ou se aprofundar nesse universo com auxílio de programadores ou especialistas que saibam trabalhar com API's e criação de scripts próprios de coleta.

REFERÊNCIAS

AROSO, Inês; CORREIA, Frederico. **A Internet e os novos papéis do jornalista e do cidadão**. In Revista Eletrônica Temática, [2007]. Disponível em.: <http://www.insite.pro.br/2007/35.pdf>. Acesso em: 28 junho 2016

BERTOCCHI, Daniela. **DOS DADOS AOS FORMATOS:** o sistema narrativo no jornalismo digital. In: ENCONTRO ANUAL DA COMPÓS, 23, 2014, Universidade Federal do Pará. Anais.... Belém: [2014]. Disponível em: <http://compos.org.br/encontro2014/anais/Docs/GT10_ESTUDOS_DE_JORNALISMO/bertocchi_daniela_compos2014_menor_2232.pdf>. Acesso em: 15 junho 2016

BOYD, danah; ELLISON, Nicole. **Social Network Sites: Definition, History, and Scholarship**. [2008]. In Journal of Computer-Mediated, Volume 13, [2007]: p.210-p.230. Disponível em: <http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.2007.00393.x/abstract>. Acesso em: 28 junho 2016

BRIGHTPLANET. **Twitter Firehose vs. Twitter API**: What's the difference and why should you care?. Texto disponibilizado em 25 jun. 2013. Disponível em: < https://brightplanet.com/2013/06/twitter-firehose-vs-twitter-api-whats-the-difference-and-why-should-you-care/>. Acesso em: 02 julho 2016

DIEBOLD, Francis X. **On the Origin(s) and Development of the Term 'Big Data'**. [2012]. Disponível em: <http://ssrn.com/abstract=2152421>. Acesso em: 27 junho 2016

FRIENDLY, Michael. **Milestones in the history of thematic cartography, statistical graphics, and data visualization**. In: Gallery of Data Visualization. [2009]. Disponível em: <http://www.math.yorku.ca/SCS/Gallery/milestone/milestone.pdf>. Acesso em: 03 julho 2016

GARTNER. **Big Data**. www.gartner.com, IT Glossary. Disponível em: <http://www.gartner.com/it-glossary/big-data/>. Acesso em: 01 julho 2016

KAHN, Robert E.; CERF, Vinton G. **What Is The Internet (And What Makes It Work)**. [1999]. Disponível em: < http://www.cnri.reston.va.us/what_is_internet.html>. Acesso em: 15 junho 2016

KEIM, Daniel; MANSMANN, Florian; SCHNEIDEWIND, Jörn; ZIEGLER, Harmut. **Challenges in Visual Data Analysis**. Estados Unidos: IEEE Computer Society Washington, [2006]. Disponível em: <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=1648235>. Acesso em: 28 junho 2016

LEMOS, André. **Cibercultura:** tecnologia e vida social na cultura contemporânea. Porto Alegre: Sulina, [2015]. 295 p.

LÉVY, Pierre. **Cibercultura**. Tradução de Carlos Irineu da Costa. 2. ed. São Paulo: Ed. 34, [1999]. 260 p. (Coleção Trans).

MANOVICH, Lev. **What is Visualization?.** [2010]. Disponível em: < http://manovich.net/content/04-projects/064-what-is-visualization/61_article_2010.pdf> Acesso em: 01 julho 2016

PIERRE Lévy – O Big Data e a próxima revolução científica. Fronteiras do Pensamento: [2016]. 2:25. Disponível em: <https://www.youtube.com/watch?v=W5hIcxKPVRw>. Acesso em: 29 junho 2016

RECUERO, Raquel. **Redes sociais na internet**. Porto Alegre: Sulina, 2009. 191 p.

RUBINSTEIN, Daniel; SLUIS, Katrina. **A LIFE MORE PHOTOGRAPHIC:** Mapping the networked image. In Photographies, [2008]. Disponível em: <http://dx.doi.org/10.1080/17540760701785842>. Acesso: 27 junho 2016

THE beauty of data visualization – David McCandless. TEDGlobal 2010, [2010]. 17:56. Disponível em: < http://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization>. Acesso em: 24 junho 2016

TUFEKCI, Zeynep. **Big Data:** Pitfalls, Methods and Concepts for an Emergent Field. [2013]. Disponível em: <http://ssrn.com/abstract=2229952>. Acesso em: 10 junho 2016

VIS, Farida. **A critical reflection on Big Data:** Considering APIs, researches and tools as data makers. First Monday, [2013]. Disponível em: < http://firstmonday.org/ojs/index.php/fm/article/view/4878/3755>. Acesso em: 26 junho 2016

VISUALIZAR. Michaelis Dicionário Brasileiro da Língua Portuguesa. Melhoramentos, [2015]. Online. Disponível em: <http://michaelis.uol.com.br/busca?id=zaZ3M>. Acesso em: 28 junho 2016.