2024

Introdução à análise de dados on-line

Cecom - Centro de Estudos do Campo da Comunicação

Labidecom - Laboratório de Inovação, Desenvolvimento e Pesquisas em Educomunicação

Escola de Comunicações e Artes

Universidade de São Paulo

Sumário

Módulo 1

Pesquisa e métodos digitais

Objetivos de aprendizagem:

- Entender a relevância da internet na pesquisa social
- Familiarizar-se com as discussões sobre a pesquisa na internet e os métodos digitais
- Verificar o uso de métodos digitais em pesquisas científicas

A pesquisa na sociedade em rede

KopiteCowboy (2015), CC BY-SA 4.0

Os computadores e a internet na pesquisa já não são novidades. Iniciantes no mundo acadêmico mal podem imaginar a época em que teses eram redigidas em máquinas de escrever ou quando alguém, num mundo sem bases digitais conectadas, ferramentas de busca de literatura científica e livrarias on-line, podia demorar meses para efetuar revisões de literatura. Houve um tempo em que as submissões de artigos a revistas e congressos eram feitas pelo correio e foi um avanço quando, em vez de textos impressos, eram enviados disquetes. Dispositivo que muitos, hoje, nunca viram.

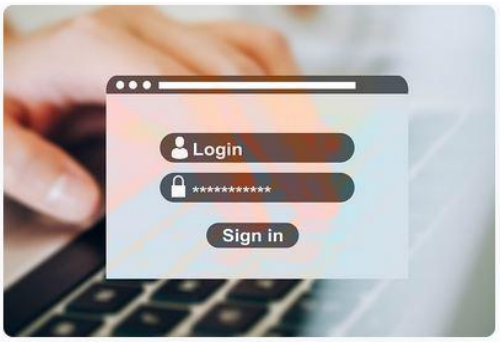
As alterações no modo de vida das sociedades, advindas dessa veloz inserção de tecnologias digitais, desde meados da década de 1990, ocorrem não somente no campo da investigação científica, mas em todas as dimensões. Daí, a criação de termos para nomear o atual momento histórico caracterizado pela emergência de uma sociedade global interconectada por redes de informação, como o de **sociedade em rede** (Castells, 2010/1996).

Entretanto, como as ciências sociais sempre tiveram como objeto a realidade de seu tempo, sendo admitido que seu surgimento decorre das transformações da primeira revolução industrial, elas são, ao mesmo tempo, afetadas pelas transformações, internamente em seu modo de trabalho, e convocadas para compreender as mudanças na sociedade.

É provável que você que está iniciando esse curso já tenha ou esteja desenvolvendo alguma pergunta de pesquisa que possa estar relacionada ao universo da internet. Como pretende realizar a pesquisa, que dados utilizará?

A mídia digital como objeto de pesquisa

geralt (2019), Pixabay

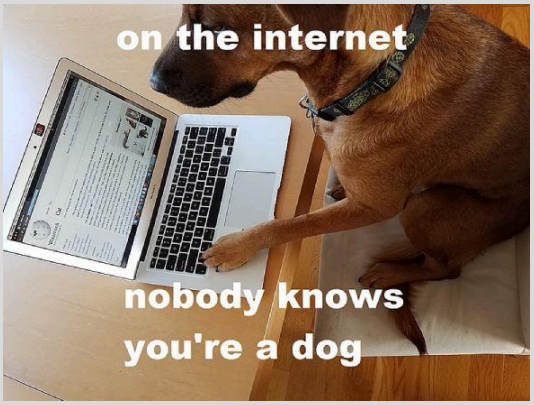


No contexto social mencionado, há o debate sobre a emergência da **ciência social digitalizada** (Witte, 2012) ou da **sociologia digital** (Nascimento, 2020). A discussão ocorre em várias disciplinas, com propostas de especialidades voltadas ao digital (e.g., "história digital") ou para a organização do modo de produção de conhecimento de maneira mais geral. Isso poderia ocorrer, por exemplo, por meio da pesquisa em campos interdisciplinares, como o das "humanidades digitais".

Outros autores defendem o estudo do papel da mídia digital na sociedade como um tópico relevante de pesquisa, mas não como nova disciplina ou campo inter, multi ou transdisciplinar. O **problema apontado** (Fuchs, 2019) é a hiperfragmentação do mundo acadêmico em subcampos e subdisciplinas cada vez mais especializados, promovendo uma diversidade sem unidade.

Seja como for, a pesquisa social já tem contribuído para aumentar a compreensão e derrubar mitos sobre as práticas digitais. Desse modo, assim como a novidade da internet se esvaiu, também o folclore, por vezes divertido, sobre ela foi superado por conhecimentos mais rigorosos. Um exemplo é o da preocupação sobre a suposta **ampla falsificação de identidades** que a rede fomentaria, com animais se passando por pessoas.

Internet e identidades



O meme acima é um dos que foi inspirado pelo **cartum**, hoje clássico, de Peter Steiner, publicado em 1993, quando a questão das identidades on-line era um tema de questionamento.

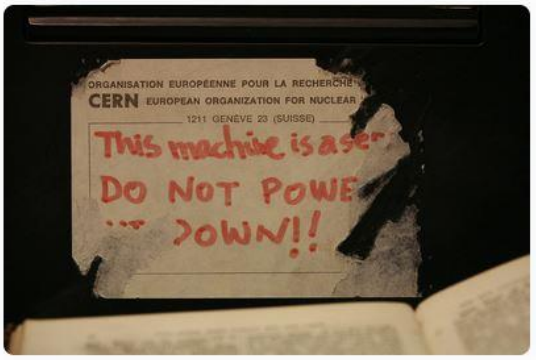
O contínuo imbricamento entre o mundo virtual e o off-line reforçou a importância social da internet, que é cada vez mais utilizada pelas pessoas. Desse modo, é possível indagar (Kozinets, 2014/2009, p. 11):

> "Existe uma distinção útil entre a vida social on-line e os mundos sociais da 'vida real'? Cada vez mais, a resposta parece ser não. As duas se mesclaram em um mundo: o mundo da vida real, como as pessoas o vivem. É um mundo que inclui o uso da tecnologia para se comunicar, debater, socializar, expressar e compreender."

Ainda que essa convergência possa ser debatível, ela mesma é um tema de pesquisa, a respeito de diferentes situações e objetos de investigação. Em sentido mais geral, as questões de conhecimento que envolvem a internet foram amplamente expandidas. A discussão sobre os dados on-line ou dados digitais e o uso deles associados aos métodos digitais decorre desse cenário.

Pesquisa on-line e métodos digitais/virtuais

Robert Scoble (2008), CC BY 2.0



O termo **pesquisa on-line** (*online research*) é corrente na literatura de língua inglesa e seu uso está relacionado ao uso da internet em investigações. Isso começou a ocorrer desde que a web se tornou pública, sendo comum, nos **debates da época** (Hooley et al., 2018), a distinção entre a pesquisa que examina a internet e a que utiliza a pesquisa on-line. A ideia de **métodos on-line** ou **digitais** se aplica à última noção, ainda que tais métodos sejam naturalmente mais utilizados quando os fenômenos em estudo estão associados à rede.

Apesar de não existir consenso a respeito do que são métodos digitais (*digital methods*), é possível delinear dois entendimentos principais: um geral ou amplo e outro específico, mais restrito. No primeiro caso, o termo indica (Snee et al., 2016, p. 1) o

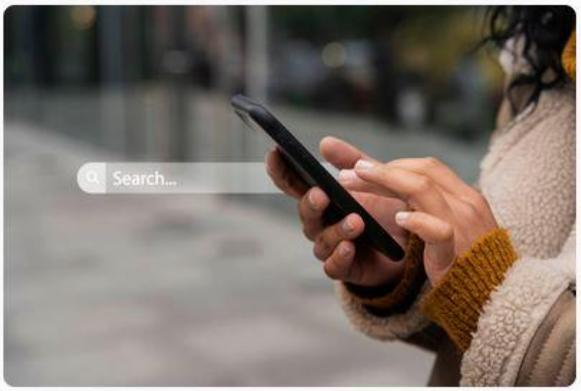
> "*uso de tecnologias on-line e digitais para coletar e analisar dados de pesquisa* ... [envolvendo] não apenas a pesquisa que explora fenômenos on-line, mas também com interesse mais amplo na utilização de métodos digitais para abordar todos os aspectos da vida social contemporânea".

Nessa perspectiva, o "digital" na expressão é visto como um **termo genérico** (Dawson, 2020, p. 1) que descreve métodos de pesquisa que usam aplicativos e procedimentos relacionados a computadores. É comum, a partir desse entendimento, que se discuta a transposição de métodos tradicionais das ciências sociais para o digital, entre outros, os levantamentos (*surveys*), entrevistas e grupos focais, etnografias e experimentos.

Por vezes, para diferenciar essa abordagem daquela que veremos a seguir, fala-se em métodos virtuais com respeito a esse entendimento ampliado sobre métodos digitais.

Métodos digitais em sentido estrito

Freepik



O entendimento restrito sobre métodos digitais é baseado nas propostas de Richard Rogers (2013), que delimita o domínio deles às pesquisas que utilizam, primariamente, dados da própria internet, procurando refletir sobre eles e "seguir o método do meio".

> Isso significa levar em consideração as especificidades da mídia digital investigada: estudar a sociedade a partir da investigação de um dispositivo, aplicativo ou plataforma digital implica também estudar essas instâncias. Em síntese, essa abordagem metodológica emprega ferramentas e objetos digitais para investigar os traços da vida social que emergem das interações on-line.

A adesão aos postulados teóricos de cada proposta, bem como os problemas e questões de pesquisa ao qual alguém irá se dirigir são os fatores que geralmente presidem a escolha ou ênfase em determinada concepção.

De fato, os entendimentos expostos sobre os métodos digitais não são incompatíveis. Uma pessoa com uma concepção mais ampla sobre o tema pode julgar relevante entender como os dados de mídia digital que utiliza podem ser afetados por processos técnicos. Em contrapartida, outra que estude uma plataforma a partir de dados internos a ela poderá optar por complementar seu estudo com entrevistas on-line.

A divisão dual da pesquisa com métodos digitais é, em certa medida, uma estratégia de exposição didática, porém, é possível perceber e refletir sobre como investigações envolvendo a internet podem aderir mais a uma ou outra abordagem.

Veja, no próximo tópico, a linha do tempo sobre o desenvolvimento de tecnologias, os impactos sociais da internet e a pesquisa on-line.

Internet, pesquisa e sociedade

Dave Braunschweig (2016), CC BY-SA 4.0

Linha do tempo - Tecnologias, sociedade e pesquisa da internet

Alguns marcos na trajetória de desenvolvimento de tecnologias, impactos sociais e investigações ligadas à internet.

Linha do tempo elaborada com informações de **Craig** (2023), Hooley et al. (2012), Litman e Robinson (2020), **Science and Media Museum** (2020) e **Zimmermann e Emspak** (2022).

1950 — 1960

Origens

A criação da internet se relaciona ao contexto, na década de 1950, da Guerra Fria, nos Estados Unidos. A base de conhecimentos do país no setor da computação e o receio de que a União Soviética pudesse atacar o país impulsionaram a ideia de criar um sistema de comunicações que fosse imune a um ataque nuclear soviético. A URSS também fez planos de interligação dos computadores do país, no entanto, **não teve sucesso**.

1958

ARPA

A criação da ARPA (Agência de Projetos de Pesquisa Avançada), pelo governo dos EUA, reunindo cientistas e militares teve como objetivo principal fazer com que o país estivesse à frente do inimigo soviético no campo tecnológico. Entre os projetos da ARPA estava testar a viabilidade de uma rede de computadores em grande escala, ideia que resultou na ARPANET.

1962

O conceito da Rede Galáctica

O cientista estadunidense Joseph Carl Robnett Licklider, do MIT, propôs pela primeira vez uma rede global de computadores, a "Galáctica". Como chefe da Diretoria de Técnicas de Processamento de Informações da ARPA, ele estava ligado à agência desde essa data até 1964. E sua abordagem ajudou a dar forma à ARPANET, a precursora da Internet atual.

1965 — 1974

ARPANET

A troca rudimentar de mensagens entre dois computadores, em 1965, deu início à ARPANET, que começa efetivamente a se estabelecer. No entanto, à medida que ela crescia, um conjunto de regras para lidar com pacotes de dados precisava ser elaborado. Em 1974, os cientistas da computação Bob Kahn e Vint Cerf inventaram um novo método chamado protocolo de controle de transmissão, popularmente conhecido como TCP/IP, que permitia que os computadores se comunicassem, o que fez com que a ARPANET crescesse rapidamente, se tornando uma pioneira rede global interconectada de redes, ou "Internet".

1973

Primeira chamada por celular

O engenheiro e inventor Martin Cooper, então na empresa Motorola, realiza a primeira chamada por um telefone móvel.

Introdução à Análise de Dados On-Line

1979 — 1981
Redes de telefonia móvel
A primeira rede celular automatizada (1G) analógica comercial foi lançada no Japão pela Nippon Telegraph and Telephone em 1979. E, em 1981, houve o lançamento simultâneo na Dinamarca, Finlândia, Noruega e Suécia do sistema Nordic Mobile Telephone (NMT).

1980 — 1990
Programas para análise de dados
Computadores foram usados por pesquisadores que faziam análises de conteúdo quantitativas desde a década de 1960. Porém, os softwares, principalmente para análise de dados qualitativos, começam a ser lançados somente a partir dos anos de 1980.

1986
Survey on-line
Kiesler e Sproull realizam um levantamento amostral on-line – Response effects in the electronic survey, *Public Opinion Quarterly*, *50*(3): 402-413. https://doi.org/10.1086/268992

1989 — 1990
World Wide Web
Tim Berners-Lee faz uma demonstração da World Wide Web (WWW), em 1989, e no ano seguinte ela é lançada ao público, com a ideia básica de "mesclar as tecnologias em evolução de computadores, redes de dados e hipertexto em um sistema de informações global poderoso e fácil de usar", conforme o **laboratório CERN**, onde esse físico e cientista da computação britânico trabalhava.

1990
Internet e WWW
Embora os termos "world wide web" e "internet" sejam geralmente utilizados de maneira intercambiável, há diferenças entre eles. A internet é a infraestrutura de rede que conecta os dispositivos, já a World Wide Web é uma forma de acessar informações por meio da internet. Nesse sentido, a criação do HTTP (Hypertext Transfer Protocol), do sistema URI (Universal Resource Identifier) ou URL, que fornece endereço exclusivo às páginas da rede, e da linguagem de marcação de textos HTML, por Tim Berners-Lee, caracterizaram a web como uma estrutura de comunicação eficiente que, nos anos seguintes, passaria a ter ampla adoção nas sociedades.

1993 — 1994
Navegadores
Em 1993 é lançado o primeiro navegador (*browser*) gráfico e multiplataforma, o Mosaic, e no ano seguinte o navegador Netscape, popular nos primeiros anos da internet, principalmente até o lançamento, em 1995, do Internet Explorer. Veja uma história dos navegadores, no **site da fundação Mozilla**.

1994
Entrevistas on-line
Foster realiza entrevistas on-line assíncronas, usando o e-mail – Fishing with the net for research data, *British Journal of Educational Technology*, *25*(2): 91–97. https://doi.org/10.1111/j.1467-8535.1994.tb00094.x. Também em 1994, Brotherson efetua a primeira discussão metodológica sobre entrevistas em grupo de foco on-line – Interactive focus group interviewing: A qualitative research method in early intervention, *Topics in Early, Childhood Special Education*, *14*(1): 101-118. https://doi.org/10.1177/027112149401400110

1994
Buscadores
Lançamento das primeiras ferramentas de busca na internet, como Alta Vista, Lycos, Excite e Yahoo.

1995
Journal of Computer-Mediated Communication
A **revista científica** começa a ser publicada neste ano, evidenciando o interesse acadêmico no tema.

1995
Etnografias e experimentos na internet

Correll escreve sobre “etnografias da internet” – The ethnography of an electronic bar: The lesbian café, *Journal of Contemporary Ethnography*, *24*(3): 270–98. https://doi.org/10.1177/089124195024003002. No mesmo ano, são conduzidos e publicados resultados dos primeiros experimentos feitos on-line.

1995
Videoconferência on-line
Primeira ocorrência pública de videoconferência pela internet.

1995 — 2008
Criptomoedas
Desde 1995, houve a criação de criptomoedas, no entanto, a popularização da ideia se deu, principalmente, a partir do surgimento do Bitcoin em 2008, tendo a primeira transação comercial com seu uso ocorrido dois anos depois.

1995
Comércio on-line
O ingresso na internet da Amazon e do eBay é um marco do comércio on-line. No Brasil, uma iniciativa relevante, de 1999, foi a do site Submarino. Com o tempo, a **Amazon ampliaria seu horizonte de atuação**, se tornando uma das grandes empresas do setor digital, fazendo parte do grupo **GAFAM**, acrônimo de empresas digitais gigantes: Google, Apple, Facebook, Amazon e Microsoft.

1996
Ética na pesquisa on-line
O debate sobre a ética na pesquisa on-line tem forte impulso. O número 2, do volume 12, da revista ***The Information Society*** documenta essas discussões.

1996
Serviços de mensagens instantâneas
Neste ano, foram lançados os primeiros serviços de mensagens instantâneas, como o ICQ.

1996
Webmail
O lançamento do então HoTMaiL (as letras maiúsculas eram uma homenagem ao HTML) marcou o início dos serviços de correio eletrônico gratuitos (a **história do e-mail**, em geral, é maior). A empresa foi comprada no ano seguinte pela Microsoft.

1997
Google
Lançamento do então apenas buscador Google.

1997
Weblogs
Atribui-se a Jorn Barger, com seu blog Robot Wisdom Web, o primeiro uso desse formato, popular nos anos seguintes, dando origem a plataformas que facilitaram a criação de weblogs.

1997
Wi-fi
A tecnologia Wi-Fi tem uma longa história de desenvolvimento, mas esse ano pode ser considerado o de sua criação oficial, pois foi nele que o padrão IEEE 802.11 foi aprovado por órgão regulatório dos EUA.

1997
Rede social: Six Degrees
É lançado o SixDegrees.com, considerado por muitos o primeiro site de rede social.

1997
Netflix

A Netflix é fundada por Reed Hastings e Marc Randolph, inicialmente enviando DVDs pelo correio. Em 2022, já como plataforma audiovisual, possuía 200 milhões de assinantes. Ao lado, o logotipo original da companhia, utilizado até 2000.

1997 — 1998
Recrutamento de pessoas e coleta massiva de dados
Em 1997 a internet foi usada, pela primeira vez, para recrutar participantes para um estudo, no trabalho de Smith e Leigh – Virtual subjects: Using the Internet as an alternative source of subjects and research environment. *Behavior Research Methods*, *29*(4), 496-505. https://doi.org/10.3758/BF03210601. No ano seguinte teve início o **Project Implicit** que, em um ano e meio, obteve mais de 600 mil respostas on-line para um teste que visava mensurar o nível de preconceito implícito das pessoas que participaram.

1998 — 2000
Literatura internacional: pesquisas reflexivas
Entre esse anos, foram publicadas várias pesquisas no ambiente on-line, que possuem importantes reflexões sobre métodos, entre elas:

- ***Life Online: Researching Real Experience in Virtual Space*** (1998), de Annette N. Markham;
- ***Tune In, Log On: Soaps, Fandom, and Online Community*** (2000), de Nancy K. Baym;
- ***The Internet: An Ethnographic Approach*** (2000), de Daniel Miller e Don Slater;
- ***Virtual Ethnography*** (2000), de Christine Hine.

1999 — 2003
Literatura internacional: discussão metodológica
Vários livros metodológicos sobre a pesquisa on-line, que se tornaram referências, foram publicados, entre outros:

- ***Doing Internet Research: Critical Issues and Methods for Examining the Net*** (1999), editado por Steven Jones;
- ***Internet Communication and Qualitative Research: A Handbook for Researching Online*** (2000), de Chris Mann e Fiona Stewart;
- ***Online Social Sciences*** (2002), organizado por Bernad Batinic, Ulf-Dietrich Reips e Michael Bosnjak;
- ***Internet Research Methods*** (2003), de Claire Hewson, Carl Vogel e Diana Laurent.

1999
Association of Internet Researchers
Fundação formal da **Association of Internet Researchers**.

2000
Pesquisa de mercado
A partir dos anos 2000, a pesquisa de mercado passou a utilizar intensivamente a internet, principalmente pela obtenção de respostas a pesquisas em amostras de conveniência. Tais amostras são compostas pelo recrutamento de respondentes a partir de convites em anúncios pop-up que surgem enquanto as pessoas navegam na internet, bem como pelo uso de plataformas que permitem criar questionários (como a SurveyMonkey). Com frequência são oferecidos incentivos, como a possibilidade de participar de sorteios ou ganhar recompensas.

2001
Wikipedia
A enciclopédia global exemplificou a tendência iniciada na época de produção de conteúdo digital coletivo.

2003 — 2004
Redes sociais: Myspace e Facebook
Em 2003 é lançado o site da rede social Myspace e, no ano seguinte, o Facebook.

2004
Web 2.0
O termo, criado e divulgado pela empresa O'Reilly Media, se popularizou a partir desse ano. Busca caracterizar o que seria uma segunda geração de comunidades e serviços relacionados à internet, como redes sociais e blogs.

2004
VoIP
Desenvolvimento dos serviços de telefonia a partir de VoIP (Voice over Internet Protocol).

2005 — 2006
YouTube
Criado em 2005, o site de vídeos foi comprado pelo Google em 2006, por 1,65 bilhão de dólares.

2006
Twitter
Lançado nesse ano, o Twitter foi adquirido em 2022 pelo bilionário Elon Musk, que mudou o nome da rede social, no ano seguinte, para X.

2006
ABCiber
Fundação da **ABCiber** (Associação Brasileira de Pesquisadores em Cibercultura), reunindo investigadores do Brasil voltados ao tema.

2007
iPhone
É lançado, pela Apple, com várias inovações, como a tela sensível ao toque, se tornando um dos **produtos mais rentáveis de todos os tempos**.

2008
Google Chrome
O navegador da empresa, que já conquistara dimensão extraordinária, é lançado, passando a obter cada vez mais usuários, com o tempo.

2009 — 2014
WhatsApp
Criado por dois programadores, ex-funcionários do Yahoo!, o aplicativo foi lançado em 2009, tendo como objetivo mostrar se uma pessoa estava disponível para receber ligações telefônicas. Só passou a ganhar maior adesão quando, em versões posteriores, adquiriu o caráter de app de mensagens instantâneas. Em 2014, o conglomerado Meta Platforms Inc. (ainda sob o nome Facebook Inc.) comprou o WhatsApp, que tinha então cerca de 400 milhões de usuários, por mais de US$ 19 bilhões.

2010
Redes sociais: Pinterest e Instagram
O Instagram inicialmente era voltado à edição e compartilhamento de fotos, antes de ser comprado pelo Facebook, em 2012, por cerca de um bilhão de dólares.

2011
Métodos de pesquisa para internet
***Métodos de Pesquisa para Internet*,** de Suely Fragoso, Raquel Recuero e Adriana Amaral, foi, no contexto local, um livro pioneiro na discussão da pesquisa no ambiente on-line.

2012
SOPA-PIPA
Os projetos de lei "Stop Online Piracy Act" e "Protect Intellectual Property Act", conhecidos por seus acrônimos, que transitaram no Congresso estadunidense, tinham como objetivo combater a circulação e o comércio ilegal de materiais protegidos por copyright na internet. As propostas eram apoiadas por grandes empresas de entretenimento dos EUA. O teor restritivo e punitivo das medidas, entretanto, levou a protestos de cidadãos e também de grandes empresas digitais e organizações. Diante disso, essas propostas foram arquivadas.

2013
Vigilância governamental e espionagem
Edward Snowden, um ex-funcionário da CIA, vazou informações sigilosas do governo dos Estados Unidos, revelando detalhes de alguns dos programas de vigilância que o país – utilizando servidores de empresas como

Google, Apple e Facebook – realiza interna e externamente, em vários países da Europa e da América Latina, entre eles o Brasil. Snowden vive exilado desde essa época na Rússia.

2015
Transmissões de vídeo ao vivo
O formato de transmissão de vídeos ao vivo (*live streaming*) pela internet se tornou mais comum quando o Facebook lançou o Facebook Live e o Twitter comprou o Periscope, um aplicativo de transmissão ao vivo.

2016
TikTok
Criada pela empresa ByteDance, na China, a plataforma de vídeos curtos passou a se expandir globalmente no ano seguinte, sendo o aplicativo mais baixado na App Store em 2019.

2016
Pokemon Go
Fenômeno global do ano, evidenciou as possibilidades de combinação entre internet e realidade aumentada.

2018
LGPD
Em um contexto global de discussão regulatória sobre proteção à privacidade, quanto ao uso/tratamento de dados pessoais e a segurança de dados compartilhados com terceiros, principalmente empresas digitais, o Brasil aprovou a **Lei Geral de Proteção de Dados Pessoais** (Lei nº 13.709/2018), em 2018. No mesmo ano, a Europa e os Estados Unidos aprovaram legislações com objetivos similares.

2018
IoT
Há aumento significativo de diferentes aparelhos conectados, na chamada Internet das Coisas (IoT). Eram previstos cerca de sete bilhões de dispositivos até o final deste ano.

2018
Cambridge Analytica
O escândalo Cambridge Analytica, relacionado à coleta e uso indevido de dados pessoais do Facebook por essa empresa, com provável influência na eleição presidencial dos EUA em 2016, foi revelado numa **reportagem** do *Guardian*, de 2018. A partir daí, essa plataforma e outras adotaram políticas de uso de dados mais restritivas. Um líder da pesquisa da área usou o termo "APIcalypse", num **artigo** em que criticou as normas adotadas então e que se mantiveram até hoje. O trocadilho remete ao bloqueio de acesso às APIs (Interface de Programação de Aplicações) das plataformas para coleta de dados (ponto explicado no Módulo 3).

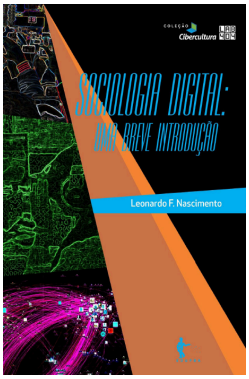
2019 — 2022
5G
Em 2019, o 5G, padrão de tecnologia de quinta geração para redes móveis e de banda larga, foi lançado para o público dos países avançados. No Brasil, os testes começaram em 2020 e a implantação definitiva, dois anos depois.

2020 — 2021
Covid e tecnologias
A epidemia global de Covid levou diversos países a adotarem medidas de distanciamento social. Nesse sentido, atividades de trabalho, estudo e socialização em geral passaram a ser realizadas com o uso de tecnologias e a internet. Um destaque foi para os aplicativos de videochamadas, como o Zoom e o Google Meet, que permitiam realizar reuniões virtuais, tanto pelo celular quanto pelo computador.

2022
ChatGPT
Recebida com entusiasmo, surpresa e também preocupações, a plataforma de Inteligência Artificial ganhou muitos usuários em todo o mundo e obrigou os concorrentes a agilizarem o lançamento de produtores similares.

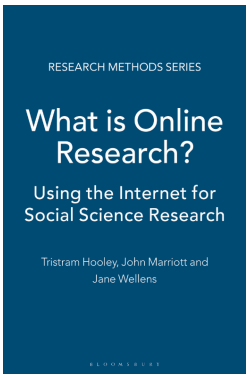
REAs de aprofundamento

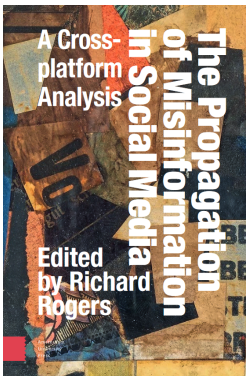
Materiais para estudos após o curso - Módulo 1

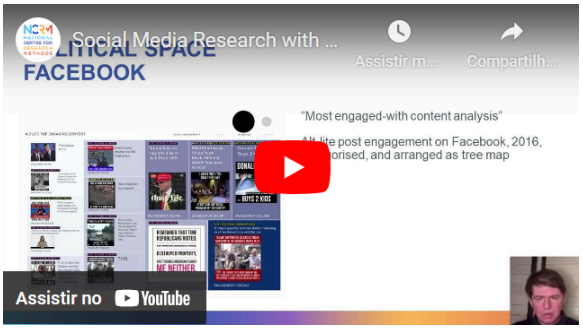


Os e-books indicados possibilitam aprofundar conhecimentos sobre temas estudados no Módulo. Em **Sociologia Digital: Uma Breve Introdução**, Leonardo F. Nascimento apresenta discussões e debates essenciais sobre o assunto; já **Métodos Digitais: Teoria-Prática-Crítica**, organizado por Janna Joceli Omena, reúne capítulos sobre aspectos conceituais e práticos da perspectiva metodológica iniciada por Rogers.

Embora um tanto antigo, publicado em 2012, o **livro sobre a pesquisa on-line** retrata bem a trajetória e a discussão relacionada com a compreensão ampliada a respeito de métodos digitais, até a data de sua publicação. Já o **livro editado por Richard Rogers** apresenta vários artigos, alguns dos quais exemplificam a abordagem mais específica relacionada aos métodos digitais. Ambos os livros são abertos, clique nos links no topo das imagens para acessá-los.







Neste vídeo, de 2021, Richard Rogers fala sobre métodos digitais, em particular na sua aplicação ao estudo da mídia social.

Módulo 2

Dados digitais

Objetivos de aprendizagem:

- Compreender o papel dos dados digitais na sociedade atual
- Conhecer a noção de “dados” nas ciências sociais
- Reconhecer o uso de dados on-line em pesquisas
- Refletir sobre a importância da crítica dos dados na pesquisa

A centralidade dos dados na sociedade

Outra noção gestada a partir do reconhecimento da importância de tecnologias digitais nas últimas décadas foi a de **sociedade conduzida e orientada por dados** (*data-driven society*). Ela penetrou na consciência popular primeiro a partir do **discurso jornalístico** e de gestores, administradores e políticos, sendo usada muitas vezes com um pendor tecnicista e otimismo ingênuo.

Alguns autores procuram dar contorno reflexivo a essa noção, destacando dimensões como a quebra de privacidade, vigilância e controle associadas a ela ou mesmo a **ameaça à democracia pela ação das plataformas digitais.** Outros, sugerem conceitos diferentes, como **datacracia** ou **colonialismo de dados**. Em todas essas noções, cuja discussão escapa ao escopo do curso, há destaque ao papel dos dados na conformação do mundo contemporâneo.

Mas o que são **dados**? Como explica um **especialista** (Kitchin, 2014, p. 2) no tema:

> “Etimologicamente, a palavra dados (*data*) é derivada do latim *dare*, que significa ‘dar’. Nesse sentido, os dados são elementos brutos que *podem ser* abstraídos de (dados por) fenômenos – medidos e registrados de várias maneiras. Entretanto, no uso geral, os dados se referem aos elementos que *são* capturados, extraídos por meio de observações, cálculos, experimentos e gerenciamento de registros”.

A citação expressa um sentido geral para o termo. Outra definição, a partir de uma **perspectiva crítica** (Couldry & Mejias, 2019, p. XIII) sobre o modo como eles se transformaram num dos pilares do sistema capitalista atual, é a de dados como fluxos de informações que transitam da vida humana para as infraestruturas de coleta e processamento. Assim, abstraem a vida, convertendo-a em informações que podem ser armazenadas e processadas por computadores.

O adjetivo **digital** não é utilizado, mas a discussão está relacionada sobretudo a esse tipo de dado: logs de acesso a espaços na internet e informações sobre o tempo de visualização de páginas, interações com pessoas e conteúdos de redes sociais (curtidas, compartilhamentos, comentários etc.), informações sobre o uso de produtos que possuem recursos de IoT (internet das coisas), localizações de GPS, entre outros.

Esses dados e sua agregação em escalas até então inimagináveis (*big data*) foram impulsionados por interesses do mundo das finanças, dos negócios e do marketing, mas logo chamaram a atenção dos cientistas sociais, em particular daqueles ligados às abordagens quantitativas e que já usavam computadores. A feitura de análises comportamentais a partir de dados e metadados de indivíduos sem que, em tese, ocorressem vieses de observação representava, para alguns, a principal inovação da pesquisa on-line e seu método por excelência.

Rastros, traços e pegadas digitais

"Your recent Amazon purchases, Tweet score and location history makes you 23.5% welcome here."

Vários termos são usados para destacar a singularidade dos dados das pessoas que utilizam a internet, como **rastros**, **traços** ou **pegadas digitais**. Eles representam marcas de atividades que elas realizam e que poderão ser submetidas a análises com diferentes objetivos. Algumas avaliações sobre o papel e as possibilidades desses dados nas ciências sociais são **entusiásticas** (Venturini & Latour, 2019, p. 43):

> "Graças à rastreabilidade digital, os pesquisadores não precisam mais escolher entre precisão e alcance em suas observações: agora é possível seguir uma multiplicidade de interações e, simultaneamente, distinguir a contribuição específica que cada uma delas tem para a construção de fenômenos sociais. Concebidas em uma era de escassez, as ciências sociais entram em uma época de abundância."

Porém, há alertas sobre o risco de que a pesquisa social seja soterrada pela avalanche de dados desse tipo. Além disso, as pessoas que conduzem investigações científicas não têm muitas vezes controle sobre eles, nem exato **conhecimento das categorias** (Witte, 2012) a partir das quais são produzidos. Há ainda o risco de que a prevalência do chamado **positivismo digital** (Fuchs, 2019), alicerçado nas ciências da computação, colonize as ciências sociais e humanas, em detrimento de abordagens interpretativas e críticas.

Os computadores e a internet promoveram novas práticas de investigação social, bem como revitalizaram estratégias tradicionais. Ao facilitarem a coleta e tratamento de dados complexos deram novo vigor, por exemplo, à Análise de Redes Sociais (ARS) e à produção de visualizações para identificar padrões ou configurações relevantes.

Isso ocorreu e continua a ocorrer a partir de propostas, críticas e debates sobre a validade desses procedimentos, bem como pela adaptação de métodos e o uso de certos dados habituais à investigação social. Entretanto, em paralelo, se desenvolvem também perspectivas e dados **nativamente digitais**, ampliando o leque de possibilidades para a pesquisa que se dirige aos fenômenos on-line.

Dados nativamente digitais e dados digitalizados

Reproduzido do site Trump Twitter Archive

Os dados **nativamente digitais** estão associados à perspectiva restrita sobre métodos digitais, sendo geralmente derivados do uso das mídias digitais, por exemplo, metadados de usuários, visualizações de vídeos do YouTube, números de "curtidas" numa postagem do Facebook ou de seguidores de um perfil do Instagram.

Nem todo dado que podemos obter pela internet é nativamente digital, pois há inúmeras informações que foram inseridas na rede por questões de acessibilidade, mas que podem também estar e ter sido geradas em outros locais, como estatísticas, documentos e atas de governos, empresas e organizações, livros e decisões judiciais. Esses seriam, portanto, dados **digitalizados**.

Observa-se ainda a existência de um terceiro tipo de dado na internet: aquele que consiste em um **objeto digital que passa a ser arquivado** (Rogers, 2015) e isso tem efeito em sua natureza, como o arquivo de tweets de Trump. Isso pode ocorrer também com páginas da internet, metadados de usuários e outros materiais que sejam arquivados.

Outros tipos de dados

Além de dados nativamente digitais e digitalizados, há aqueles que são produzidos a partir de técnicas tradicionais das ciências sociais (pela pesquisa on-line) adaptadas ao contexto da rede. A entrevista telefônica ou presencial é com frequência, hoje, realizada a partir de dispositivos digitais e pela internet; os questionários em papel dão, cada vez mais, lugar aos formulários on-line e grupos pesquisados podem ser submetidos a certas condições experimentais apoiadas por computadores ou pela internet.

Nesses casos, o papel da rede é acessório a uma investigação e a natureza do dado pode ser afetada, com vantagens e desvantagens, pela condição de coleta. Uma vantagem costumeiramente mencionada é que a discussão de temas sensíveis pode ser favorecida pela distância que há na interação pela internet. Seja como for, essa é uma instância de reflexão de qualquer pesquisa.

Como conclusão parcial, é interessante notar a possibilidade de atualizar o dito do poeta Stéphane Mallarmé (1842-1898) que, a seu tempo, afirmava que tudo existiria no mundo para chegar a um livro. Atualmente, tudo existe para terminar na internet.

Taxonomia de dados

Além da distinção entre digitais e digitalizados, os dados podem ser categorizados de outras maneiras. Uma diferenciação comum entre eles comum é quanto aos **qualitativos**, relacionados a textos (escritos, mas também imagéticos ou audiovisuais, por isso alguns os caracterizam como não numéricos), e os **quantitativos**, associados a números. Ambos os tipos podem ser utilizados em uma investigação e estão ligados aos **formatos dos dados**, como: texto simples, numérico, percentual, valor monetário, data/horário e hiperlink.

Há outros modos de categorizar os dados, como uma maneira para entendê-los melhor. Kitchin (2024), além de distingui-los também entre quali e quantitativos, no que seria a **forma** dos dados, elabora as seguintes categorias:

Estrutura	• *Estruturados:* organizados, armazenados e exportados em um modelo de dados específico, como números/texto em tabelas ou bancos de dados com formato consistente (por exemplo, nome, data de nascimento, endereço, gênero etc.). • *Semiestruturados:* pouco estruturados, sem um modelo ou esquema de dados predefinido. Apesar da estrutura irregular, seus campos são parcialmente consistentes. Um exemplo disso, são páginas da web com XML (*Extensible Markup Language*), que codifica os documentos. • *Não estruturados:* nenhum modelo ou estrutura comum é identificável. Cada elemento pode ter uma estrutura ou um formato próprio. São dados, geralmente, de natureza qualitativa, como o de um conjunto de postagens do Facebook.
Fonte	• *Capturados:* diretamente por meio de alguma forma de medição, como observação, pesquisas, experimentos de laboratório e de campo, manutenção de registros, câmeras, scanners e sensores. • *Excedentes:* gerados por um dispositivo ou sistema, como um subproduto de sua função principal. A nota eletrônica para processar pagamentos também pode servir para controles de estoque ou medida de desempenho de quem faz determinadas vendas, por exemplo. • *Transientes:* nunca examinados ou processados e simplesmente descartados, por serem muito volumosos, não estruturados, caros para processar e armazenar ou, ainda, por faltarem técnicas para extrair valor deles, sendo de pouca utilidade estratégica ou tática. • *Derivados:* produzidos por meio de processamento ou análise adicional de dados capturados.
Quem fez	• *Primários:* gerados pelo indivíduo que faz uma pesquisa, a partir dos instrumentos usados e dentro de um projeto de pesquisa próprio. • *Secundários:* gerados por outra pessoa e disponibilizados para reutilização e análise para outros indivíduos. • *Terciários:* uma forma de dados derivados, como contagens, categorias e resultados estatísticos. São geralmente divulgados por órgãos estatísticos, para garantir a confidencialidade com relação a quem os dados se referem.
Tipo	• *Indexicais:* permitem a identificação e a vinculação (por exemplo, nomes, endereços de IP, números de passaporte e de cartão de crédito, números de série do fabricante). • *Atributivos:* representam aspectos de um fenômeno, mas não são indexicais. A impressão digital é indexical, mas as informações dos atributos de idade, sexo, altura, peso, tipo sanguíneo, não o são. • *Metadados:* dados sobre dados, como nomes e descrições de campos específicos em uma planilha. Podem se referir ao conteúdo dos dados ou a todo o conjunto de dados.

Uma diferenciação menos intuitiva é entre **dados criados e dados existentes** (Laville & Dione, 1999). No primeiro caso, os dados são produzidos a partir de uma intervenção explícita de quem realiza a pesquisa no objeto de estudo. Esse dado é característico da pesquisa experimental, na qual alguma ação deliberada busca provocar uma mudança que será investigada. Nas ciências sociais e da comunicação, essa forma de pesquisa é menos comum e com mais tradição no contexto estadunidense. A ação deliberada pode ser de vários tipos: pedir para que o

grupo em estudo leia, veja ou faça algo, por vezes, comparando o efeito disso num grupo não submetido à intervenção.

Por sua vez, os dados existentes (termo um tanto ambíguo) estão associados a todos os outros tipos de dados que a pessoa que realiza uma investigação elabora utilizando as técnicas da pesquisa social, como a observação, as entrevistas e os questionários. Na verdade, se o termo **dado** não fosse consagrado pelo uso, seria melhor substituí-lo por outros, como **apreendido** (De Bruyne et al., 1991/1976) ou **capturado** (Kitchin, 2014).

Dados proprietários e dados abertos

Jonathan Gray (2011), Domínio Público

Os dados variam também quanto ao nível de acesso e finalidade. Alguns são totalmente públicos e podem ser utilizados livremente, inclusive em pesquisas científicas. Num polo oposto, há dados privados e restritos a ambientes internos, como intranets, e não disponíveis para uso externo. Há, é claro, dimensões intermediárias entre esses dois tipos.

Uma vez que há custos envolvidos na produção e disponibilização de dados, bem como por eles poderem ter valor na produção de conhecimento sobre o mundo, o acesso a eles tem sido frequentemente restrito. Isso é feito, por exemplo, limitando o acesso a pessoas que pagam ou recebam alguma aprovação ou limitando a forma como os dados podem ser usados. Nessa perspectiva, os termos de serviço das plataformas digitais são geralmente restritivos, tratando os dados como ativos da empresa, o que gera dificuldades para quem investiga esses espaços.

Por outro lado, as possibilidades de democratização do conhecimento pela internet favoreceram as iniciativas de **dados abertos** (*open data*), ou seja, os esforços voltados à divulgação e compartilhamento de conteúdo de maneira ampla, sem restrições de uso. A ideia está alinhada com movimentos como o do software open source, da publicação científica e da ciência abertas, preconizando aumento da transparência e controle social sobre a informação.

Dados abertos: Iniciativas globais e locais

Há iniciativas globais, de sites, como o **Open Data Inception** e o **Data Portals**, ou repositórios, como o do **Banco Mundial**, reunindo dados de muitos países. No caso do Brasil, o governo federal possui um **Plano de Dados Abertos** e algumas iniciativas locais são as seguintes:

- Portal Brasileiro de Dados Abertos (dados do governo federal e governos locais)
- IBGE - Dados abertos
- Ipeadata (dados econômicos e sociais brasileiros)
- Portal de Dados Abertos do TSE
- DivulgaCandContas - TSE (dados de candidaturas e contas eleitorais)
- Portal de Dados do Cetic.br (dados de uso da internet no Brasil, entre outros)

A expressão dados **digitais** abertos é redundante, pois os defensores dos open data geralmente colocam a disponibilização deles na internet como uma de suas características. Fala-se também em **Linked Open Data** (LOD), termo que recebe diferentes traduções (dados abertos interligados/conectados/vinculados), cujo significado remete tanto a certas práticas para publicar e conectar dados estruturados na web quanto aos conjuntos de dados desse tipo, como os da **The Linked Open Data Cloud**.

Há também várias fontes de dados que, embora não sejam estritamente abertos, podem, muitas vezes, ser utilizados como dados secundários de alguma investigação. O InternetLab, por exemplo, elaborou **Um guia da dieta de mídia digital brasileira** com vários levantamentos desse tipo. Outra fonte relevante de dados para a pesquisa da comunicação são os **acervos digitais de periódicos ou de coleções**. Por vezes, o acesso é condicionado à assinatura da publicação.

Acervos digitais de periódicos

Acima é mostrada a tela da ferramenta de pesquisa da **Hemeroteca Digital** da Biblioteca Nacional, cuja coleção abrange periódicos, principalmente, do século XIX, mas também possui jornais importantes na história da imprensa moderna do país, como **Última Hora**, **Jornal do Brasil** e **O Pasquim**. Outros acervos digitalizados de jornais e periódicos brasileiros são os seguintes:

- Arquivo Público do Estado de São Paulo - coleção com muitos periódicos dos séculos XIX e XX, infelizmente sem ferramenta de busca interna deles;
- Veja;
- O Globo;
- Folha de S.Paulo;
- O Estado de S. Paulo.

Em termos de acervos de audiovisual digitalizados, talvez a principal iniciativa local seja a da Cinemateca Brasileira, em seu **Banco de Conteúdos Culturais**.

Crítica dos dados

Google Maps

> "Embora nem todas as formas de conhecimento estejam firmemente enraizadas em dados – por exemplo, conjecturas, opiniões e crenças –, os dados são claramente um material de base fundamental para a forma como entendemos o mundo." (Kitchin, 2014, p. 12)

Como a citação destaca, os dados possuem um papel relevante na vida social. No entanto, a naturalização e o uso mal-intencionado ou ingênuo deles são frequentes. As polêmicas sobre os tratamentos durante a epidemia de Covid-19 foram estimuladas por dados, interpretações e por **pesquisas sem rigor científico** (Hellmann & Homedes, 2022). Os políticos muitas vezes publicam postagens comemorativas em redes sociais quando atingem determinado número de seguidores, com o desejo de se autocongratular por popularidade expressiva. Mas quantos desses são robôs ou foram "comprados"?

Os dados podem ser controversos e estão relacionados à fonte que os produziu: um mapa-múndi feito na Argentina provavelmente chamará as ilhas que o país disputa com o Reino Unido de "Malvinas", enquanto o nome "Ilhas Falkland" aparecerá em outros (na dúvida, o Google Maps, usa ambos). O modo como as coisas do mundo são nomeadas e descritas podem variar, conforme os valores de quem produz a informação.

O que importa destacar, assim, é que a produção e o uso de dados deve ser feito, na pesquisa científica, de maneira questionadora e crítica. O próprio jornalismo de qualidade procura agir desse modo ao fazer, por exemplo, **distinções sobre métricas de redes sociais**.

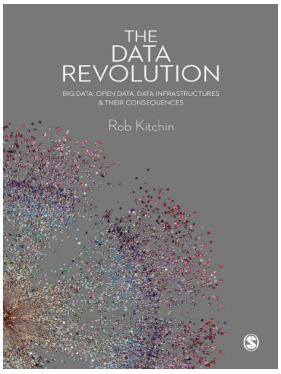
Uma cientista de dados, ao abordar alguns **aspectos que introduzem imperfeições e vieses em dados**, oferece uma perspectiva básica para a crítica a eles. Conforme nota a autora:

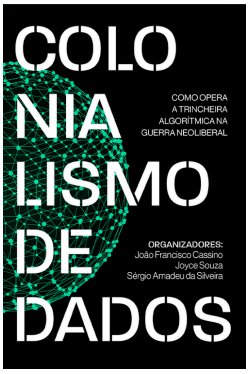
> "Precisamos questionar os dados, em vez de presumir que, só porque atribuímos um número a algo, de repente isso é a verdade pura e simples. Ao se deparar com um estudo ou conjunto de dados, pergunte: O que pode estar faltando? Qual é a outra maneira de considerar o que aconteceu? E o que essa medida específica considera, exclui ou incentiva?" (para. 41).

A partir dos estudos, espera-se que seu conhecimento sobre os dados e sua relevância na investigação científica tenha aumentado. A atividade que encerra o módulo propõe, para consolidar aprendizados, uma revisão do que foi visto até aqui, com a preocupação, também, de apresentar o uso de dados em pesquisas de comunicação.

REAs de aprofundamento

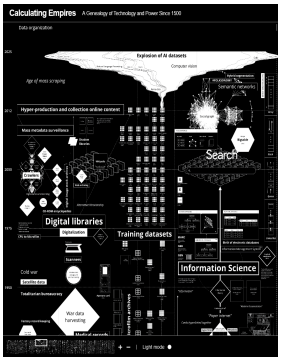
Materiais para estudos após o curso - Módulo 2

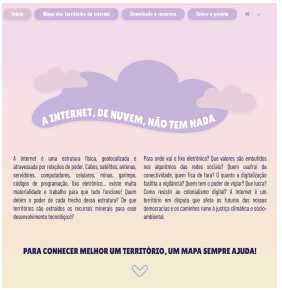


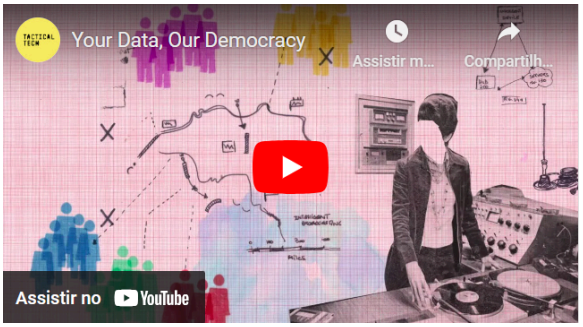


O excelente **primeiro capítulo** (“Conceptualising Data”) do livro **The Data Revolution** é disponibilizado pela editora. Mesma situação do livro digital **Colonialismo de Dados**, organizado por João Francisco Cassino, Joyce Souza e Sérgio Amadeu da Silveira, que aprofunda o conceito mencionado no curso.

A produção de dados digitais está ligada a infraestruturas materiais. Alguns projetos na internet, como **Calculating Empires** e **Cartografias da Internet**, destacam esse aspecto.







O minidocumentário **Your Data, Our Democracy**, da organização **Tactical Tech**, aborda questões críticas envolvendo a influência do uso de dados na sociedade.

Módulo 3

Fluxo de trabalho: coleta de dados

Objetivos de aprendizagem:

- Conhecer etapas comuns do trabalho com dados em pesquisas sociais
- Reconhecer estratégias para a coleta de dados on-line
- Refletir sobre preocupações relacionadas à coleta de dados
- Aplicar o conhecimento na coleta de algum conjunto de dados (*dataset*)

Trabalho com dados digitais

Ciclo de trabalho com dados

O diagrama com um fluxo, na forma de ciclo, de trabalho com dados em pesquisas científicas é uma exposição sintética e geral de um procedimento que pode diferir. Alguém pode, por exemplo, ter acesso a uma base de dados coletada por outrem e a partir dela gerar uma indagação que demande análises; ou perceber, após ter feito a primeira coleta e análise de dados, que precisará coletar outros. Há casos em que as coletas e análises sugerem a necessidade de mudança no problema de investigação. Nada impede, ainda, que alguém colete dados com fins basicamente exploratórios, como insumo à reflexão e construção de questões de pesquisa posteriormente.

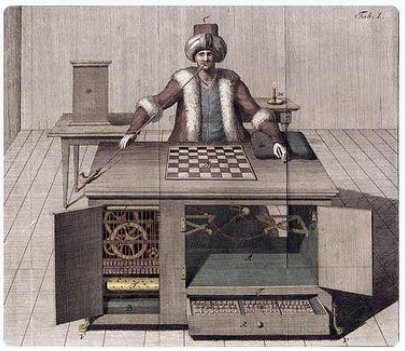
Por outro lado, a figura indica uma cronologia com etapas que se sucedem – com o potencial iterativo mencionado – e na qual, ao fim, o resultado regressa ao ponto de partida. Em outras palavras, as questões de conhecimento que deram início e

conduziram o trabalho serão idealmente esclarecidas, em maior ou menor medida, pelo que apurou.

Neste módulo vamos conhecer estratégias para coleta e tratamento dos dados e, no seguinte, algumas possibilidades analíticas.

Coleta de dados e amostragem

Joseph Racknitz (1789), Domínio Público



Os dados coletados terão íntima relação com a amostra (ou conjunto de amostras) escolhida para a pesquisa. A amostra é unidade básica da investigação empírica, composta por grupos, de pessoas ou de outro tipo (como sites ou documentos), a partir dos quais serão obtidas as informações/dados de uma pesquisa.

Uma questão crítica sobre as amostras, apontada por **vários autores** (e.g., Kitchin, 2014), diz respeito à **representatividade amostral**, ou seja, a capacidade dos dados capturarem adequadamente o fenômeno que buscam representar, gerando resultados extrapoláveis para a população em geral. Esse ponto é particularmente complexo quando os dados são obtidos pela internet. Um dos **principais desafios** (Rogers, 2013), nesses casos, é saber se determinada amostra da web representa um país ou um tipo específico de conteúdo.

A **construção de amostras** (Burgess & Bruns, 2018) (na forma de corpura) pelo agrupamento de conteúdo sobre algum tópico, principalmente via hashtags, se tornou uma estratégia comum no estudo da mídia social, assim como abordagens envolvendo a seleção de conteúdos de atores vistos como relevantes para o problema da pesquisa nesses espaços.

Os dados decorrem da amostra e ambas as dimensões estão ligadas às questões de pesquisa, fundamentalmente, além de aspectos éticos e práticos do desenvolvimento da investigação. A reflexão prévia sobre isso, portanto, se impõe.

Sobre a pesquisa on-line de maneira geral, é interessante notar que, no contexto anglo-saxão, a utilização de amostras de pessoas que recebem pequenos

pagamentos, a partir do recrutamento em plataformas digitais como a **Amazon Mechanical Turk**, tem sido comum. Inicialmente essa prática foi vista por alguns como **renovadora da pesquisa quantitativa** (Litman & Robinson, 2020). No entanto, são apontados problemas nessa opção, relacionados à validade dos resultados e à **questão ética** (Samuel, 2018) de produzir conhecimento com base em uma força de trabalho mal remunerada.

Estratégias para a coleta de dados na internet

Basicamente há duas formas principais de coletar dados na internet: a comunicação direta com as APIs de sites ou plataformas e a técnica de raspagem de dados (*scraping*), com características, vantagens e limitações específicas, como notam van der Vlist e Helmond (2023), veja a seguir.

Coleta via APIs	• Usa pontos de acesso a bancos de dados (“back-end”). • Os dados ficam ocultos no servidor de um site (ou seja, não são visíveis no navegador da web). • Os resultados geralmente estão na forma de dados estruturados.
Raspagem de dados	• Usa páginas da web e sites renderizados (“front-end”). • Os dados são visíveis em seu navegador. • Páginas e sites diferentes podem ter estruturas e formatos diversos. • Os resultados geralmente estão na forma de dados semiestruturados.

Aponta-se, por vezes, o uso de serviços de empresas que coletam dados, como um terceiro modo, mas essas empresas utilizam, de fato, alguns dos procedimentos mencionados.

A utilização da Interface de Programação de Aplicações (*Application Program Interface* ou API) para coletar dados digitais favorece a obtenção de dados. A API é uma ferramenta de software que permite a interação entre alguém, a partir de algum aplicativo de coleta, e os dados de determinado site ou plataforma, de maneira

gratuita ou sob certas condições ou taxas. A **partir de APIs** (Monaco & Amaudo, 2020) podem ser obtidos dados em grande quantidade, bem estruturados e sem que, em tese, existam questões éticas. Isso ocorre, pois a extração é regulamentada pelas próprias plataformas e os dados são compartilhados sem violação a direitos autorais.

Já o método de raspagem está ligado à extração de dados de páginas e sites a partir do código-fonte deles, num processo automatizado, no qual o aplicativo que faz a coleta é configurado para capturar determinados conteúdos marcados por certa codificação.

A raspagem de dados tem analogia com a ideia da cópia e colagem de conteúdos da web e, do mesmo modo que esse método, possui caráter mais controverso, tendo em vista que a cópia pode estar em desacordo com os termos de serviço de determinado espaço digital ou ferir a privacidade de quem produziu algo que está sendo copiado. Nesse sentido, é importante refletir criticamente a respeito do **caráter ético do projeto** (Peeters & Borra, 2020) em determinado contexto para decidir sobre quando e como usar a estratégia de raspagem de dados e outros procedimentos metodológicos.

Reflexão sobre a coleta

Numa perspectiva mais geral, vários aspectos merecem ser pensados, previamente à coleta e trabalho com dados digitais. Em primeiro lugar, o papel desses dados na elucidação das questões de pesquisa. Mas há também questões práticas: os aplicativos on-line podem ter problemas, a instalação de softwares pode falhar e as APIs das plataformas podem mudar, geralmente se tornando mais restritas, e isso é bastante frustrante para quem faz pesquisa.

A prudência recomenda, assim, que o planejamento da investigação leve em conta eventualidades: se alguma forma gratuita ou pouco onerosa de coleta de dados se tornar inviável, seria possível fazer de outra forma? O quanto isso afetará o cronograma do trabalho? Possuo ou tenho tempo suficiente para aprender o

conhecimento técnico requerido para utilizar determinada estratégia de coleta de dados? Questões desse tipo devem ser levadas em consideração.

Já durante o processo de coleta de dados digitais, van Es et al. (2017) sistematizam vários pontos a serem pensados, conforme se segue.

Questões a serem consideradas ao se coletar dados para pesquisa

- Quais considerações éticas foram levadas em conta ao coletar os dados da pesquisa?
- Que tipo de dados está sendo usado?
- Como os dados foram coletados? Quais ferramentas ou softwares foram usados, ou quem forneceu os dados?
- Quais critérios foram usados para selecionar o conjunto de dados? Quem está incluído ou excluído do conjunto de dados?
- Quais são as limitações desses métodos de coleta de dados? Qual é o grau de confiabilidade do método de coleta utilizado?
- Quais metadados o conjunto de dados contém (por exemplo, local, hora, data de um tweet)?
- Ao combinar conjuntos de dados, quais vieses podem resultar dos diferentes contextos de origem dos dados?

No tópico seguinte serão expostos procedimentos, ferramentas e exemplos relacionados à coleta de dados, conforme a categorização mostrada abaixo. Em cada grupo, são descritas características básicas de aplicativos. É possível notar que a ordem de exposição vai de estratégias mais simples às que demandam mais trabalho.

Coleta manual e importação de dados

- Não envolve, a rigor, programas, porém o uso de softwares para o manejo de dados em tabelas, como o Google Planilhas, é geralmente recomendável.

Baixar e arquivar páginas da web

Mozilla Firefox (navegador multiplataforma)
- Download.
- Permite copiar páginas web (ou partes delas) em formato de imagem ou PDF.

HTTrack Website Copier (programa open source, para Windows e Linux)
- Download.
- Copia as páginas de algum website para um computador.
- Manual.
- Tutorial em vídeo (em inglês).

A1 Website Download (programa proprietário para Windows e Mac)
- Download.
- Possui versão paga e gratuita. Os primeiros 30 dias de uso, após a instalação, permitem utilizar todas as funcionalidades do programa. Depois disso, caso ele não seja comprado, os recursos diminuem.
- Faz o mesmo que o programa anterior (cópia de páginas), mas é mais rápido, o que pode ser útil para sites complexos, com muitas imagens e páginas secundárias.
- Tutorial em vídeo (em inglês)

Conifer (serviço on-line)
- Permite criar contas onde serão estocados os sites que forem copiados pelo aplicativo.
- Os dados podem ser baixados e o arquivo de cada pessoa é de 5 GB.
- Tutorial em vídeo (em inglês).

WayBack Machine (serviço on-line)
- Permite verificar o conteúdo e aparência anterior de páginas web, conforme elas tenham sido copiadas pelo Internet Archive. O que for localizado pode ser copiado ou baixado e arquivado pelos métodos anteriores.

Aplicativos on-line pagos

Apify (site de serviço)
- Aplicativo pago, mas que permite coletas de dados gratuitas, até determinados montantes.
- Exige a criação de conta e possui uma interface intuitiva.
- Extrai dados, que podem ser exportados em diferentes formatos, principalmente, de redes sociais.
- Tutoriais em vídeo (em inglês).

PhantomBuster (site de serviço)
- Mesmas características do anterior.
- Tutoriais em vídeo (em inglês).

Raspagem de dados

Instant Data Scraper (plugin para o navegador Chrome)

- Download.
- Extrai dados de páginas da web e os exporta como arquivos Excel ou CSV.
  Utiliza uma IA para buscar os possíveis conteúdos relevantes, que quem utiliza o plugin poderá selecionar.
- A interface é intuitiva e fácil de usar, porém, os recursos são limitados.
- Tutorial em vídeo (em inglês).

Data Miner (plugin para o navegador Chrome)

- Download.
- Além de baixar o plugin, é necessário criar uma conta no aplicativo.
- Permitir fazer a raspagem de dados, a partir de "receitas" preexistentes ou criadas por quem usa.
- Os dados capturados podem ser exportados em diferentes formatos tabulares.
- O serviço possui versão paga, que torna mais fácil o trabalho, porém, estudando os procedimentos para criar as receitas, é possível extrair muitos tipos de dados da web com a versão gratuita.
- Tutoriais em vídeo (em inglês).

Screaming Frog SEO (programa multiplataforma)

- Download.
- Aplicativo com versão gratuita e paga. A primeira com limitações.
- Embora voltado principalmente a profissionais da comunicação digital, ao fazer a recuperação automática de dados na web, em particular as ligações entre páginas (links), pode ser útil a pesquisas acadêmicas.
- Os dados obtidos são exportados em diferentes formatos de tabela.
- A própria empresa disponibiliza guias de uso.
- Tutoriais em vídeo (em inglês).

Aplicativos gratuitos

Media Cloud (plataforma)

- A plataforma Media Cloud é um projeto open source que permite que se recuperem notícias sobre determinado assunto, exportando os dados em diferentes formatos.
- É necessário criar uma conta no site, cujo uso é relativamente simples.

Facepager (aplicativo multiplataforma)

- Download.
- Aplicativo para a recuperação automatizada de dados de plataformas, como Facebook e YouTube, desenvolvido por Jakob Jünger e Till Keyling (2019).
- Possui um conjunto de pré-configurações (*presets*) que facilita fazer a solicitação para a coleta de dados.
- Os desenvolvedores criaram um site bastante explicativo sobre o programa.
- Tutoriais em vídeo (em inglês).

YouTube Data Tools (aplicativo on-line multiplataforma)

- Desenvolvido por Bernhard Rieder (2015), no âmbito da Digital Methods Initiative, o aplicativo compreende um conjunto de seis ferramentas on-line para extrair dados de

vídeos, canais, redes de canais e comentários em vídeos do YouTube.
- A interface é simples e intuitiva.
- Tutoriais em vídeo (em inglês).

4CAT (aplicativo open source multiplataforma)
- Download.
- Programa que coleta dados de várias plataformas de mídia digital (Facebook, TikTok, X/Twitter, Telegram, entre outras), por vezes, em combinação com o plugin do navegador Firefox Zeeschuimer.
- A instalação envolve a criação de um Docker e configurações especiais no computador, já que o programa é executado remotamente. É recomendável que as pessoas interessadas leiam com atenção o tutorial e vejam o vídeo sobre como instalar o 4CAT.
- Há um paper (Peeters & Hagen, 2020) que descreve o desenvolvimento do software.
- No site do projeto, vários tutoriais explicam como fazer coletas e outras questões.
- Tutoriais em vídeo (em inglês).

Ao baixar algum programa, leia atentamente as instruções e verifique se seu computador possui os requisitos necessários. Não instale nenhum programa sem ter testado outro recém-inserido, pois se ocorrer algum erro, não saberá qual o causou. Caso note problema no computador, desinstale o aplicativo.

Coleta de dados on-line: possibilidades, ferramentas e exemplos

Várias estratégias podem ser utilizadas para a coleta de dados on-line. Veremos aqui, resumidamente, explicações sobre coletas com o uso de ferramentas digitais e softwares. Alguns são gratuitos, como o Facepager, o 4CAT e o YouTube Data Tools ou pagos (mas, por vezes, com planos gratuitos), como o Apify.

É possível, ainda, em certas circunstâncias, coletar dados digitais estritos ou que apenas estão na internet, por meio de raspagem de dados ou abordagens manuais, conforme será mostrado.

Coleta manual e importação de dados

Embora se pense, frequentemente, no uso de ferramentas digitais para coletar dados on-line, é usual que sejam extraídos dados a partir da cópia (CRTL-C / CRTL-V) de textos ou imagens. Conforme a pesquisa, isso pode ser suficiente, em particular se o escopo do trabalho é pequeno.

O primeiro vídeo tutorial exemplifica um procedimento “manual” de coleta de dados de seguidores do Instagram, que podem ser utilizados para a construção de uma rede entre seguidos/seguidores.

O segundo vídeo mostra como importar dados tabulares para uma planilha do GDocs. Nesse caso, está sendo feita uma “raspagem” simples de dados. Veremos raspagens mais complexas, em outro momento.

Diferentes dados abertos podem ser importados. Mas é válido organizá-los. O terceiro tutorial, a partir do site do TSE, exemplifica uma organização comparativa, antes de baixar os dados.

Por fim, demonstra-se como coletar e organizar dados do arquivo digital de um jornal.

Todos os dados coletados poderão ser trabalhados depois. Vale notar, assim, a importância de entender o uso de planilhas e, por vezes, fórmulas.

Coleta manual

O **vídeo** exemplifica como coletar manualmente dados de perfis seguidos por uma conta do Instagram. A lógica é a mesma para a coleta de seguidores. As fórmulas mostradas são: =MOD(ROW();2) e =INDIRECT("A" & (2*ROW()))

Coleta com raspagem simples

O **vídeo** demonstra como importar dados de uma tabela na web para uma planilha do GDocs. A fórmula que deve ser utilizada está abaixo. Quando se copia uma lista, é necessário usar “list” no lugar de “table”. E o número desses elementos é relativo a quantos a página web possui.

=IMPORTHTML("endereço_do_site"; "table"; número_da_tabela)

Organizar e importar dados

No **vídeo** é mostrada a estruturação para extrair dados do site do TSE com dados de prestações eleitorais, sendo escolhidos as cinco candidaturas que obtiveram mais votos para a Câmara Federal pelo estado de São Paulo, na eleição de 2022. Os dados foram importados em arquivo de planilha.

Coleta de dados de arquivo digital de jornal

O **vídeo** mostra como foram coletados dados de unidades informativas contendo determinado termo (“Marielle”) em um jornal, a partir do acervo digital da empresa

noticiosa. A coleta foi manual, sendo planejada a partir de variáveis (os títulos das colunas) relacionadas a interesses da pesquisa.

Síntese

- Coletas manuais são práticas e válidas quando a escala dos dados a serem capturados é pequena.
- Os dados podem ser inseridos em diferentes formatos e programas no computador de quem faz a coleta.
- Quando a técnica de copiar e colar gera dados desorganizados, pode ser tentada a raspagem simples de dados tabulares, para que se importe o conteúdo de uma tabela em uma planilha do Google.
- É válido importar dados de fontes abertas já com algum tipo de organização que seja coerente com o propósito da pesquisa.
- É possível submeter o dado a tratamentos posteriores.

Baixar e arquivar páginas da web

A cópia de componentes de páginas web ou de sites como um todo pode se relacionar a diferentes objetivos de pesquisa. Rogers (2013), por exemplo, associa o desenvolvimento de sua perspectiva sobre métodos digitais a uma reportagem investigativa com a análise temporal de sites de extrema direita dos Países Baixos, arquivados no Internet Archive, para perceber se havia transformação no discurso.

Em termos acadêmicos, também é útil arquivar páginas que possam desaparecer, para citá-las.

Sendo assim, é válido saber como copiar/gravar páginas atuais ou do passado.

Com o navegador **Mozilla Firefox** é possível copiar, como imagem, uma página web ou fragmento dela, assim como salvá-la como arquivo PDF, como mostrado a seguir.

Aplicativos para a cópia de páginas, como **A1 Website Download**, **HTTrack Website Copier** e **Conifer**, permitem baixar sites inteiros ou páginas específicas. O último distingue-se por permitir que se crie uma conta on-line, com 5GB, na qual é inserido o conteúdo clonado.

Já com o site **WayBack Machine**, podemos recuperar o estado anterior de páginas web. Talvez elas tenham dados relevantes que possam se combinar a dados atuais. Outro uso acadêmico do primeiro serviço é permitir gerar referências estáveis.

Mozilla Firefox

Com o uso das combinações **CTRL+clique no botão direito do mouse** (Windows) ou **CMD+clique** (Mac), selecione em “Capturar tela” e escolha, na janela de opção, se copiará toda a página ou só um fragmento. Veja o **vídeo**.

É possível salvar uma página como PDF, escolhendo essa opção no campo “Destino”, depois de executar os comandos para a impressão no navegador.

Cópia de página ou site

O programa **A1 Website Download** foi utilizado no primeiro exemplo do **vídeo**, sendo feita a captura de página do site **Social Blade** com métricas do Twitter de Jair Bolsonaro. Mostra-se a coleta e a abertura do arquivo, no navegador. A interface do programa **HTTrack Website** é similar, até mais simples, do que a do software utilizado.

Depois, utilizando o serviço **Conifer**, demonstra-se como baixar um site.

WayBack Machine

Este **vídeo** demonstra como copiar uma página para esse arquivo, inserindo-a também na conta criada nele e gerando um link estável.

Veja também um tutorial detalhado sobre esse procedimento.

Outro tutorial do **WayBack** explica como ver diferentes estados de uma página.

Síntese

- Dada a natureza inconstante e mutável da internet – conforme estudo do Pew Research Center, cerca de 38% das páginas web que existiam em 2013 não estavam disponíveis uma década depois – a gravação de páginas e sites pode ser interessante para determinadas investigações.
- É possível também ver e coletar estados anteriores de páginas web, em ferramentas como o **Wayback Machine**, de modo a compor bases de dados mais robustas ou fazer comparações diacrônicas.

Aplicativos on-line pagos

Há vários aplicativos na internet que oferecem serviços de coleta de dados da web, particularmente de sites de rede social. Com eles, são extraídos dados como os de postagens e comentários, dos membros de grupos e comunidades, dos seguidores etc. Aqui são indicados dois que possuem planos de teste ou possibilitam usar o programa para coletas de menor quantidade de maneira gratuita: o **Apify** (veja **tutoriais** da empresa) e o **PhantomBuster** (**tutorial**).

Apify

É necessário criar conta nesses aplicativos de coleta, depois, na área interna, fazer a escolha do que se pretende extrair. No **vídeo**, é mostrada a zona interna do **Apify**, exemplificando como se pode escolher algum tipo de opção para capturar dados on-line de algum espaço. O vídeo exibe a coleta de postagens do Instagram do político Guilherme Boulos, que depois foi transferida para uma planilha do GDrive para trabalho posterior.

Síntese

- Dependendo do tipo de dado que se deseja, o uso de aplicativos para coletas pode ser uma alternativa vantajosa, principalmente pela rapidez e simplicidade.
- Se o projeto possui pequena escala, é provável que o plano gratuito seja suficiente. Porém, caso isso não ocorra, a avaliação sobre a relação entre o custo e o benefício de pagar pelos dados pode ser feita.
- Como sempre, os dados poderão ser tratados após a coleta, conforme os tipos de análise que se pretenda fazer.

Raspagem de dados

Coletar a informação diretamente do código-fonte de uma página web, na chamada técnica de "raspagem de dados" (*scraping*), é algo feito, muitas vezes, com o uso de programas no formato de plugins de navegadores, como os que serão apresentados.

Uma variante da raspagem é o rastreamento (*crawling*), método pelo qual são extraídas as URLs existentes em um site. Também será indicado um programa que realiza essa ação, entre outras.

Aplicativos com exemplos a seguir:

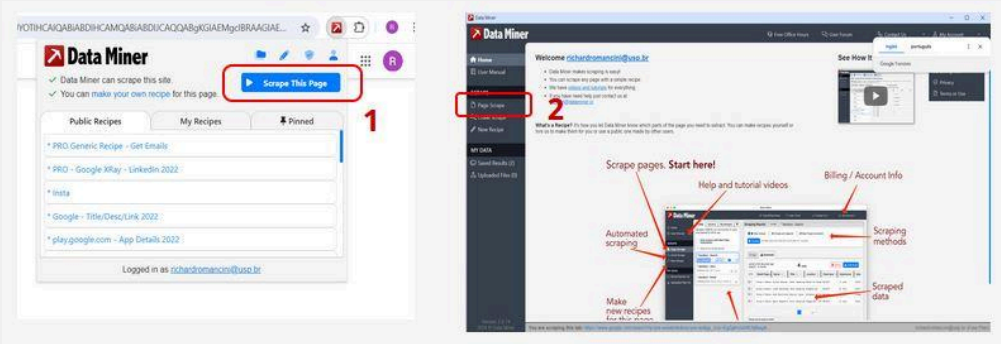
- **Instant Data Scraper:** boa opção, pela praticidade da coleta, quando os dados estão bem estruturados. Quem o utiliza, simplesmente deve clicar no botão para verificar qual tabela o programa seleciona. Uma vez que os dados sejam percebidos pela Inteligência Artificial do plugin, é possível baixá-los em diferentes formatos (e posteriormente fazer ajustes, como também será mostrado).
- **Data Miner:** plugin mais sofisticado do que o anterior, que permite usar as chamadas "receitas" para a extração de dados. É possível utilizar alguma receita existente ou elaborar uma.
- **Screaming Frog SEO Spider:** programa proprietário que pode ser utilizado em versão gratuita, com algumas limitações, como a do número de coletas que podem ser feitas.

O **vídeo** mostra a coleta de dados de uma página do Google Scholar de uma revista científica brasileira de comunicação, utilizando o **Instant Data Scraper**. O plugin foi capaz de coletar os dados de modo coerente. Inclusive, foram extraídos, dessa forma, dados de todas as revistas Qualis A da área. Essas coletas foram inseridas numa planilha, na qual é feito um tratamento inicial de dados, como será visto no próximo módulo.

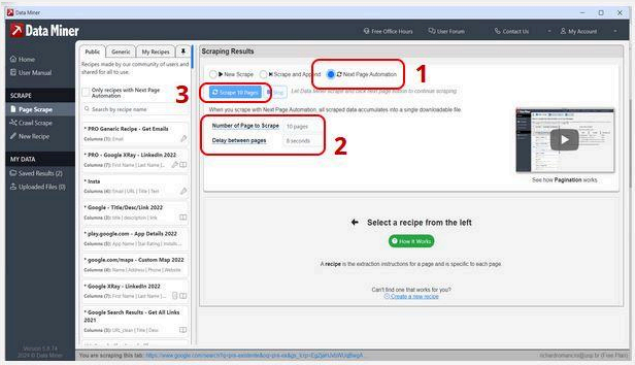
Neste **vídeo**, mostra-se uma coleta de dados, com o aplicativo **Data Miner**, dos resultados da ferramenta Google para uma consulta com o termo "Marielle Franco". Depois de fazer a busca, o plugin foi aberto e foi usada uma receita preexistente que captura os dados e os insere numa tabela com o Título, a Descrição e o Link de cada resultado. Foi feita também

uma configuração, mostrada na tela seguinte, para que o plugin coletasse resultados de outras páginas além da inicial.

No Data Miner, após estar com a página pronta para a raspagem, clicando-se em “Scrape This Page”, a parte interna do programa será mostrada e, com a escolha de “Page Scrape”, serão visualizadas as opções para a raspagem de múltiplas páginas (algo útil, por exemplo, para o resultado de buscas), como mostrado.



Na sequência da automatização da coleta de múltiplas páginas, em “Next Page Automation” na área interna do **Data Miner**, é preciso escolher, em “Number of Page to Scrape”, quantas páginas serão raspadas, a partir do comando para isso.



A opção mais complexa de construir uma receita própria para a raspagem envolve a seleção das linhas (rows) e colunas (cols) que a tabela que receberá os dados irá ter. A empresa que desenvolveu o plugin produziu material (**vídeo 1**, **vídeo 2**) que procura explicar como fazer isso.

Com o programa **Screaming Frog SEO Spider**, como mostra o **vídeo**, foram coletadas URLs do site da Associação Nacional dos Programas de Pós-Graduação em Comunicação (Compós), depois exportadas para uma planilha XLS. O mesmo foi feito (mas não mostrado no vídeo) com o site da Intercom (Sociedade Brasileira de Estudos Interdisciplinares da Comunicação), de modo que, no próximo módulo, poderemos mostrar como seria possível comparar os sites dessas associações em relação a esse parâmetro.

Síntese

- Utilizar a raspagem ou rastreamento de dados para coletas, dependendo dos

objetivos da pesquisa e do conteúdo on-line de interesse, pode ser válido. Porém, é necessário ser cauteloso, sendo recomendável a leitura dos Termos de Uso dos sites, que podem, legal ou tecnicamente, impedir ou limitar a extração de dados.

- Deve-se ter em mente também questões ligadas aos direitos autorais, bem como ter cuidados em relação à privacidade e à confidencialidade de certos tipos de dados que se pretenda obter/utilizar.
- Por vezes, os sites podem limitar o número de dados capturados, a cada solicitação. Assim, uma boa prática é coletar apenas os dados que se quer usar.
- Em relação às limitações de coleta do programa **Screaming Frog SEO Spider**, na versão gratuita, uma estratégia é fazer múltiplas extrações e, depois, unir as planilhas.

Aplicativos gratuitos

Há plataformas on-line e programas gratuitos, voltados a acadêmicos, jornalistas e estudantes, para a coleta e, por vezes, análise de dados. Veremos, a seguir, os seguintes:

- **Media Cloud:** plataforma para a coleta de dados de notícias on-line de veículos ou grupos de periódicos geograficamente diferenciados.
- **Facepager:** software que permite extrair dados de algumas plataformas.
- **YouTube Data Tools:** conjunto de ferramentas on-line, bastante intuitivas, para obter dados do YouTube por meio da API da plataforma.
- **4CAT:** ferramenta de pesquisa que busca facilitar a captura e a análise de dados de várias mídias sociais.

O **vídeo** demonstra como foi feita a coleta de notícias que possuíam a palavra “Marielle” em veículos on-line brasileiros, utilizando a plataforma **Media Cloud**. Ao fim da recuperação de dados, eles foram exportados no formato CSV.

O **Facepager** permite fazer coletas de dados de diferentes plataformas. É possível escolher pré-configurações, explicadas na interface do software, e ajustá-las, para tanto. Geralmente, as coletas são limitadas pelas APIs dos serviços. Este **vídeo** mostra um exemplo de coleta de dados de verbetes da Wikipédia em português que possuem o termo “Marielle Franco”.

As ferramentas do **YouTube Data Tools** (YTDT) possibilitam: visualizar o ID e outras informações de um canal; obter informações e estatísticas sobre canais; extrair dados de uma rede de canais ligados por indicações; criar dados para uma rede de vídeos, com base na noção de “co-comentário”, e obter informações básicas sobre um vídeo, a partir do seu ID, com dados da seção de comentários.

No **vídeo**, o criador do programa apresenta um panorama das ferramentas.

Vamos ver, a seguir, alguns exemplos de uso do **YTDT**.

O primeiro módulo do **YTDT** (Channel Info) permite conhecer o ID de um canal e outras informações, como número de vídeos publicados e visualizações deles. O módulo Video List extrai uma lista de todos os vídeos do canal, com estatísticas. Nesse caso, é possível gerar um arquivo CSV que pode ser importado para um programa de planilha, como o Excel. O **vídeo** exemplifica essas ações.

No módulo Video Coments do **YTDT**, podem ser coletados vários dados, no formato de planilha, entre eles, os comentários de algum vídeo. Outra possibilidade, no módulo Video List, é recuperar uma lista de vídeos, a partir de certo termo.

Essas duas operações são feitas neste **vídeo**.

O **4CAT: Capture and Analysis Toolkit** requer baixar um software (o *docker*), para a execução remota do aplicativo e configurações de instalação. Este **vídeo** tutorial explica como fazer isso. No entanto, é válido notar que a instalação mostrada é feita em um Mac, e no Windows o código deve ser levemente diferente. Assim, no Prompt de Comando desse sistema, ao verificar os arquivos baixados do 4CAT, é necessário usar “dir” em vez de “ls”.

A estrutura para localizar os arquivos deve seguir o padrão do Windows, e o Prompt precisa ser executado como “administrador”. Para tanto, clique no ícone com o botão direito do mouse e escolha essa opção.

O **vídeo tutorial** exemplifica uma coleta de dados utilizando o **4CAT**, no TikTok. A partir do uso de termos de busca, são recuperados conteúdos sobre a polêmica da eleição para prefeito em São Paulo de 2024, a respeito da acusação a um dos candidatos por uso de droga ilícita.

Síntese

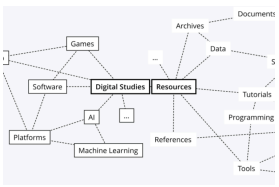
- Voltados sobretudo a plataformas de mídias sociais, os aplicativos vistos são opções práticas para coletas.
- A amplitude das extrações está relacionada ao grau de abertura das APIs. Nesse sentido, o **YouTube Data Tools** tira proveito da API menos restritiva dessa plataforma.
- A instalação do **4CAT** exige atenção e cuidado. Além disso, vale notar que programas de segurança do computador podem impedir o acesso ao *local host*, assim como o uso de versão desatualizada do Docker pode ter efeito similar.

REAs de aprofundamento

Materiais para estudos após o curso - Módulo 3

Os dois manuais, **Fluxo do Trabalho com Dados** e **Análise de Dados para o Monitoramento de Redes Sociais**, e o site indicado, **Caixa de Ferramentas do Jornalismo de Dados**, evidenciam o interesse do campo jornalístico pelo trabalho com dados, fazendo discussões e oferecendo recomendações (como de ferramentas, no caso do site) que podem ser úteis também para os indivíduos que realizam pesquisas acadêmicas.

A ideia de capacitar jornalistas, ativistas e pessoas comuns para realizar investigações com impacto social é objeto de projetos como **Exposing the Invisible**, que produziu um **kit informativo** sobre como fazer pesquisas com essa preocupação. Uma **parte do site desse kit** aponta métodos para a coleta de dados na internet.



Há vários sites e páginas da web (**Médialab SciencesPo**, **Social Media Research Toolkit**, **TDM** e **Digital Studies Resources**) que organizam listas de ferramentas, muitas ligadas à coleta e análise de dados, para a pesquisa on-line. Alguns, como a da importante e pioneira **Digital Methods Initiative**, dão acesso aos programas desenvolvidos pelo próprio grupo. Como são projetos acadêmicos, normalmente as indicações são de recursos sem ônus. Em tempo: alguns itens compilados podem hoje não existir ou não serem mais funcionais.

Módulo 4

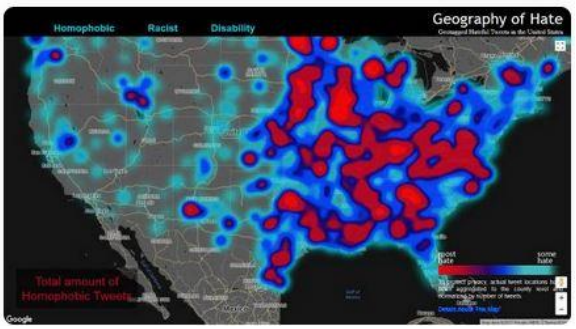
Tratamento dos dados

Objetivos de aprendizagem:

- Perceber o papel do tratamento e organização dos dados para as etapas posteriores da pesquisa
- Conhecer dimensões relacionadas à qualidade dos dados
- Familiarizar-se com o padrão *tidy data*, para adquirir a capacidade de estruturar dados nesse formato
- Aprender a realizar operações de limpeza, refino e conversão de dados

Estruturação e arranjo dos dados

Monica Stephens (2013)



Os dados poderão ser obtidos ou organizados a partir de diferentes modos, relacionados geralmente ao formato que possuem. Isso será relevante para dar legibilidade e favorecer a recuperação das informações, além de por vezes permitir tratamentos adicionais.

A maneira mais comum de estruturar dados é a tabular, com o uso de programas de edição de planilhas, como o Microsoft Excel, o LibreOffice Calc e o Google Sheets. Com mais frequência, as tabelas de planilhas contêm textos e números, mas é possível também reunir dados multimídia (imagens, gráficos, URLs). Diferentes formatos de dados podem ser combinados em planilhas. A imagem de abertura deste tópico deriva provavelmente de planilha com dados de coordenadas geográficas que se associam a dados sobre a quantidade de tweets com alguma palavra relacionada ao ódio. Depois, algum programa deu a forma gráfica que se visualiza.

Convém notar que, embora corriqueiramente os termos **base de dados** (*database*), **conjunto de dados** (*dataset*) e **estrutura** ou **quadro de dados** (*dataframe*) sejam utilizados de modo intercambiável, o primeiro é mais próprio de grandes quantidades de dados estruturados, enquanto os seguintes seriam mais aplicáveis a uma tabela simples.

Arquivos de textos, por vezes antes trabalhados em programas de planilha, podem conter dados que deem origem a **redes** ou **grafos**. Dados textuais complexos (entrevistas ou textos de páginas web) podem ser armazenados em programas de edição de texto, posteriormente tendo algum tipo de organização diferente, como a de nuvem de palavras. Pode ser relevante, conforme a investigação, salvar ou arquivar uma página (ou páginas) da web, a partir inclusive de programas, como já mostrado.

É possível ainda criar uma pasta na estrutura de arquivos do computador em que sejam inseridos dados coletados, com diferentes formatos, como vídeos, áudios, fotos, memes ou charges.

Um aspecto que favorece a confiabilidade da pesquisa é a possibilidade de outras pessoas, eventualmente e seguindo preceitos éticos, verificarem os chamados **dados brutos** – que podem ainda ser dados secundários de outras pesquisas. Nesse sentido, a reflexão sobre como manter e arquivar os dados coletados é importante.

Planilhas
Se você chegou até aqui, para a continuidade, um conhecimento básico sobre a manipulação de planilhas é relevante. Caso não o tenha, recomenda-se a consulta ao seguinte **manual**.

Limpeza e refino dos dados

PAN XIAOZHEN (2017), Unsplash

Antes de iniciar a análise de dados, com frequência é necessário realizar operações de limpeza e refino do que foi capturado. Uma recomendação importante é que seja

feita uma cópia do arquivo com os **dados brutos**, de modo que, se ocorrer algum problema durante o trabalho, seja possível regressar ao que se coletou originalmente. Assim, na cópia, o trabalho de **limpeza dos dados** (*data cleaning/cleansing*) (cf. Hall, 2024 e MacDonald, 2024) pode começar.

Tarefas de limpeza dos dados:

- Remover dados irrelevantes;
- Eliminar dados duplicados redundantes;
- Reparar erros estruturais;
- Resolver casos de dados ausentes;
- Filtrar dados discrepantes;
- Verificar a precisão, a consistência e a uniformidade dos dados;
- Validar se os dados estão corretos.

Mesmo antes da limpeza, pode ser necessário fazer algum outro tipo de tratamento nos dados, para efeito de legibilidade, por exemplo: **importar um arquivo em CSV** ou **JSON** para o Excel. Há programas open source, como o **OpenRefine**, que fazem limpezas de dados. O próprio Google Planilhas possui ferramenta para isso, no caminho **Dados > Limpeza de dados**.

Questões técnicas, como a duplicação indevida de dados durante a captura, podem ser corrigidas pelos métodos expostos. Os dados devem ser padronizados e consistentes, o que pode exigir revisões nos formatos de dados de células ou uniformizar a grafia de palavras (por vezes, caracteres especiais geram informações com ruído). Em outras situações, a correção ou refino dos dados depende de uma avaliação mais criteriosa, levando em consideração aspectos como a efetiva necessidade de certos dados para a análise ou a pertinência ética de alterar ou excluir dados confidenciais de pessoas.

As práticas e reflexões feitas neste momento têm como objetivo assegurar a **qualidade dos dados**, geralmente, desdobrada em dimensões como:

Completude	Proporção, com relação aos dados obtidos, entre os que atendem aos requisitos desejados e os que têm falha ou ausência. *Exemplo:* Se num questionário aplicado 10% das pessoas deixaram de responder a uma pergunta, isso diminui a completude dos dados.
Unicidade	Inexistência de registros múltiplos sobre algo. *Exemplo:* Um formulário respondido e registrado duas vezes por uma pessoa afeta negativamente essa dimensão.
Atualidade	Grau em que os dados representam a realidade em determinado momento no tempo. *Exemplo:* A atualidade dos dados de pesquisas eleitorais tende a decair rapidamente, diferentemente de dados sobre preferências por times de futebol.
Validade	No sentido do dado se enquadrar (ser válido) quanto ao parâmetro esperado de alguma definição. *Exemplo:* Se uma pergunta sobre quanto tempo alguém usa a internet por dia tem como resposta "esporte", o dado é inválido.
Acurácia	Capacidade do dado representar adequadamente o aspecto do mundo que procura refletir. *Exemplo:* A data de nascimento de alguém reflete acuradamente a idade dessa pessoa.
Consistência	Dados consistentes são corroborados internamente, por outros dados do conjunto, ou por dados existentes em outros locais. *Exemplo:* Se uma planilha possui dados em colunas com a data de nascimento e idade das pessoas, eles devem se confirmar mutuamente. Se não, há inconsistência.

A estrutura de tabela Tidy Data

Logotipo do pacote TidyR

Uma noção importante relacionada à ordem de dados tabulares é a de **dados organizados** (*tidy data*), proposta por Hadley Wickham, um conhecido desenvolvedor da linguagem de programação estatística e gráfica R. A ideia básica é organizar os dados num padrão com **três regras inter-relacionadas**:

- Cada variável deve ter sua própria coluna.
- Cada observação deve ter sua própria linha.
- Cada valor deve ter sua própria célula.

Uma variável é qualquer característica ou medida de um fenômeno (como “nome”, “peso” e “altura”). As observações remetem a todos os valores descritos/medidos em uma mesma unidade (uma pessoa, um dia, uma nacionalidade etc.). Essas descrições são chamadas de valores, enquanto componente de alguma célula.

Embora nem toda tabela precise utilizar esse padrão, um ponto forte dessa proposta é a ligação entre as estruturas de organização físicas e semânticas dos dados. Essa padronização tem **várias vantagens** (Costa, 2021): o tempo de limpeza e organização de dados tende a ser menor, além disso, os dados no formato tidy são mais facilmente compreendidos e reproduzíveis por pessoas que compreendem a lógica do formato, sendo compatíveis com as ferramentas tradicionais de análise e produção de visualizações de dados utilizadas em linguagens como R e Python.

Entretanto, a plena compreensão do formato não é imediata, principalmente para quem está acostumado em trabalhar com planilhas eletrônicas, como observam certos **autores** (e.g., Goedhart, 2017). Vamos ver um exemplo comparado de tabela em formato de planilha tradicional e em tidy.

Tabela norma e tabela tidy

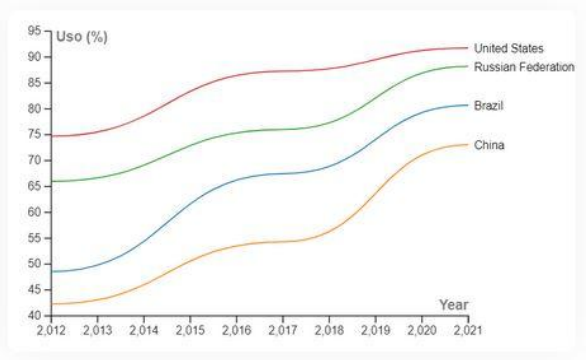
Country Name	2012	2017	2021
Brazil	48,56	67,47	80,69
China	42,30	54,30	73,05
Russian Federatic	66,00	76,01	88,21
United States	74,70	87,27	91,75

Dados tabulares comuns

Country Name	Year	Uso (%)
Brazil	2012	48,56
China	2012	42,30
Russian Federation	2012	66,00
United States	2012	74,70
Brazil	2017	67,47
China	2017	54,30
Russian Federation	2017	76,01
United States	2017	87,27
Brazil	2021	80,69
China	2021	73,05
Russian Federation	2021	88,21
United States	2021	91,75

Dados tabulares em tidy

As duas tabelas foram elaboradas com os mesmos dados, obtidos do repositório de dados abertos do World Bank, mostrando índices percentuais da população que usa a internet em diferentes países. A tabela da esquerda não está em tidy, por colocar os anos e os indicadores percentuais como observações e não como variáveis. A tabela da direita gerará melhor plotagem (criação de imagem), como a mostrada a seguir, porém, poderia ser menos facilmente compreensível no corpo do texto de um artigo científico. Em suma, dependendo do objetivo de exposição do dado, o uso do formato tidy pode ou não ser adequado.



Pacotes de linguagens como R e Python possuem estratégias para a transformação de dados para o padrão tidy. No entanto, esse tópico não é abordado por este curso. Assim, é possível sugerir o auxílio de chatbots, nesse caso, verificando o acerto do resultado. Outra possibilidade é combinar o uso de recomendações de alguma IA com a conversão da tabela para o formato tidy, usando um programa como o Planilhas Google. Veja um **exemplo**.

Coleta e tratamento de dados como prévia das análises

Mika Baumeister (2018), Unsplash

As etapas de trabalho com dados não são estanques. Ao coletar e tratar dados, muitas vezes, começamos a planejar e mesmo a fazer, embrionariamente, a análise. O vídeo **deste tutorial** mostra a consolidação e tratamento dos dados coletados sobre revistas científicas, numa Planilha Google, destacando isso. Observe, ainda, as recomendações gerais para a organização de dados em planilhas.

A partir de agora será feita referência a dados coletados até o momento, para os exemplos didáticos. Você deve ter conseguido coletar dados, como foi proposto, no entanto, se desejar poderá utilizar algum dos conjuntos abaixo para realizar exercícios práticos.

Conjuntos de dados coletados

1. **Perfis seguidos por organizações feministas no Instagram** (coleta manual)
2. **Principais influenciadores brasileiros** (raspagem simples)
3. **Comparativo de recursos e gastos eleitorais de candidatos** (dado importado do TSE)
4. **Matérias jornalísticas do Acervo da Folha de S.Paulo sobre Marielle Franco** (coleta manual)
5. **Postagens de Guilherme Boulos no Instagram** (coleta com o aplicativo Apify)
6. **Dados das revistas Qualis A de Comunicação, Google Acadêmico (2024/h5)** (raspagem com Instant Data Scraper)
7. **Resultados de busca no Google sobre "Marielle Franco"** (raspagem com Data Miner)
8. **Links dos sites das associações científicas Intercom e Compós** (extração com Screaming Frog SEO Spider)
9. **Notícias on-line sobre Marielle Franco em veículos brasileiros (2021-2024)** (coleta com Media Cloud)
10. **Verbetes da Wikipédia em língua portuguesa sobre Marielle Franco** (coleta com Facepager)
11. **Lista de vídeos do canal do YouTube do Instituto Marielle Franco** (coleta com YouTube Data Tools)
12. **Comentários no vídeo do canal do YouTube do Instituto Marielle Franco com mais interações deste tipo** (coleta com YouTube Data Tools)
13. **Vídeos do YouTube com o termo "Marielle"** (coleta com YouTube Data Tools)
14. **Postagens do TikTok com os termos "Boulos" e "cocaína"** (coleta com aplicativo 4CAT)

REAs de aprofundamento

Materiais para estudos após o curso - Módulo 4

O tratamento de dados pode ser, no todo ou em parte, realizado em programas como o Excel e o Planilhas Google. A indicação à esquerda é de uma **postagem** que fala de algumas operações de tratamento de dados no segundo aplicativo e a da direita de um **curso aberto da UFPel** sobre ele.

Os dois tutoriais da Escola de Dados ajudam a aprofundar aspectos relacionados com a limpeza e organização de dados. O **primeiro**, além de abordar o tópico principal, também ensina como importar dados abertos utilizando a API da Câmara de Deputados. Já o **segundo** é um resumo do seminal artigo em que Hadley Wickham propôs e exemplificou a noção de tidy data. Caso deseje, consulte também outro **tutorial do OpenRefine**.

Uma via de aprofundamento no trabalho com dados é a utilização de linguagens como R e pacotes para o tratamento de dados, como tidyverse. O canal do YouTube Prática de Dados possui vídeos sobre o **primeiro** e o **segundo** tópicos.

Módulo 5

Análise e visualização de dados

Objetivos de aprendizagem:

- Entender a relação entre a análise e a visualização de dados na pesquisa
- Conhecer ferramentas para a realização de análises e visualizações
- Familiarizar-se com diferentes tipos de gráficos, de modo a favorecer escolhas
- Aplicar o conhecimento na elaboração de análise que utilize os recursos estudados

Análises de dados

rawpixel.com, Freepik

O adágio “os dados não falam por si” tem como corolário a centralidade da análise de dados nas investigações. Entre os objetivos dessa etapa, conforme diferentes **estudos** (Fry, 2008; van Es et al., 2017), estão: descrever e hierarquizar os dados, destacando características relevantes, de modo a revelar padrões e, ao mesmo tempo, evidenciar relações entre as várias dimensões deles. Isso ocorre, internamente a um conjunto de dados e entre diferentes *datasets* e observações de uma investigação.

As possibilidades analíticas são bastante variadas e os critérios de escolha estão ligados às opções paradigmáticas e teóricas de quem pesquisa, bem como aos problemas de investigação enfrentados. Uma categorização comum é entre as abordagens analíticas voltadas a dados quantitativos (estatística inferencial ou descritiva) e qualitativos (análises de conteúdo e discurso).

Ao longo deste módulo, abordaremos, de maneira introdutória, as ferramentas analíticas mostradas a seguir. Por meio delas, será possível apresentar

possibilidades de análises relacionadas a dados quantitativos e produção de visualizações (4CAT, Tableau e serviços web), dados textuais, tratados quantitativamente (4CAT, AntConc e Voyant Tools) e Análise de Redes (Flourish, VOSviewer e Gephi).

4CAT

- O aplicativo online, cujas características para a coleta de dados de plataformas digitais foram vistas, possui os chamados "processadores" analíticos.
- Eles permitem obter dados organizados sob diferentes parâmetros, de maneira simples, como mostra esse vídeo.

Tableau

- Software e plataforma proprietários para a feitura de análises visuais de dados.
- A relativa facilidade de uso é um dos pontos fortes desse programa.
- Possui diferentes versões: comercial, de teste, para uso de estudantes e educadores, on-line e desktop. Aqui, vamos explorar o gratuito **Tableau Public**, que requer apenas registro na plataforma e pode ser usado on-line.
- A empresa oferece bons materiais de treinamento, com vídeos e manual digital.
- Há uma comunidade interessada de pessoas que compartilham produções na plataforma da empresa, bem como explicações e tutoriais no YouTube e outros locais da internet.
- Veja essa breve descrição sobre uso do programa.

Serviços on-line para a criação de gráficos

RAWGraphs

- Download.
- Aplicativo on-line open source, que permite a criação de diferentes tipos de gráficos.
- Possui interface simples e tutoriais explicativos sobre a criação dos gráficos, veja esse exemplo.

Flourish

- Requer a criação de conta e possui plano pago e gratuito.
- Tem como diferencial a possibilidade de criar visualizações dinâmicas para a web.
- Permite a criação de gráficos de rede.
- O uso é relativamente intuitivo, porém é mais complexo do que o anterior.

Datawrapper

- Outro serviço para a criar gráficos na web, com serviço por assinatura e uso gratuito.
- O grau de dificuldade e de recursos fica entre os dois serviços já mostrados.

Análise textual

Voyant Tools

- Aplicativo on-line gratuito com várias ferramentas de análise de dados textuais, capaz de fornecer o índice de legibilidade, as palavras usadas com frequência (e nuvem de palavras), frases-chave, entre outras informações.
- Fácil de usar, com interface amigável.
- Tutorial em português.
- Como é um serviço on-line, se os dados são sensíveis e exigem confidencialidade, não é uma boa opção.

AntConc

- Software gratuito multiplataforma criado pelo linguista Laurence Anthony, com ferramentas para a análise textual.
- Possui manual e vídeos explicativos feitos pelo autor. Na internet há também materiais de ensino do programa em português (como esse manual ou as úteis postagens de Tarcízio Silva). No entanto, é preciso ter atenção sobre o relacionamento entre a versão do programa usada e o material de estudo obtido.

Análise de redes

VOSviewer

- Programa multiplataforma especializado na produção de redes bibliográficas.
- Manual em português.

Gephi

- Programa open source multiplataforma e gratuito para a produção de redes.
- O uso é relativamente simples e o site do software possui muitos tutoriais, introdutórios ou mais aprofundados, como esse, assim como as pessoas que o utilizam e produzem explicações, em vários espaços, como o YouTube.

Visualização de dados

Choong Deng Xiang (2022), Unsplash

Como notam vários **trabalhos** (Kennedy & Allen, 2017; Kirk, 2019), a visualização de dados na pesquisa pode ser entendida sob dupla perspectiva: método de

investigação e meio para comunicar resultados no âmbito acadêmico e para o público em geral.

A produção de visualizações e as análises podem, por vezes, estar bastante ligadas, principalmente em termos de análises exploratórias dos dados. Há a ressalva, porém, sobre a necessidade do cuidado para que a organização visual dos dados não induza, de maneira equivocada, a análise. Em outras circunstâncias, pode haver uma dissociação entre a produção de visualizações, a partir de análises.

Mas o que é exatamente uma **visualização de dados**? Este **vídeo** explica didaticamente o assunto.

As visualizações e os dados que as informam tendem a ser percebidos como objetivos. Isso ocorre, pois os números, historicamente, são vistos como confiáveis. Eles sugerem universalidade, neutralidade e ligação com a ciência. Além disso, as convenções consolidadas ao longo do tempo sobre as visualizações colaboram para que sejam vistas como neutras, meras janelas para os dados. No entanto, essa é uma concepção ingênua, uma vez que as visualizações, assim como os dados, são produzidas a partir de escolhas, decisões sobre o que mostrar e priorizar. Os mesmos dados, sob diferentes perspectivas, podem conduzir a diferentes propostas visuais e mensagens.

O número de possibilidades de produção de visualizações é significativo, mas não ilimitado. A conhecida norma **APA** (American Psychological Association) descreve, entre os elementos que compõem o trabalho científico escrito, além do texto, as **tabelas** e **figuras**. As primeiras possuem um componente visual e podem ser elaboradas de diversas formas, mas com aparência relativamente similar. No caso das figuras, entretanto, há mais diferenciação, e o termo engloba, para esta norma, os gráficos, diagramas, fotografias, desenhos e qualquer outra forma de representação ou ilustração não textual.

De maneira geral, os gráficos e diagramas, a partir de agora referidos pelo primeiro termo, são a forma de visualização mais usual. O conhecimento sobre essas visualizações está relacionado à possibilidade de produzir materiais mais adequados do ponto de vista da comunicação científica. Quanto maior a compreensão das possibilidades, dos pontos fortes e limitações de cada possível forma, maior será a chance de boas escolhas.

Intenções de quem produz e experiências de quem vê

Em relação à experiência que a visualização proporcionará, um **especialista** (Kirk, 2019) no tema, comenta que há três intenções principais:

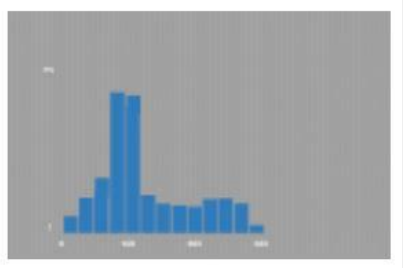
Explicativa: com a peça procurando fornecer um retrato visual dos dados, destacando os principais significados que se busca transmitir.
Exploratória: nesse caso, as pessoas que veem o material são mais livres, o que é favorecido por produções digitais, interativas e participativas que permitem a manipulação dos dados.
Expositiva: simples exibição visual de dados, cuja interpretação dependerá fundamentalmente de quem vê. Assim, é mais adequada em trabalhos voltados a públicos com conhecimento do assunto que podem fazer sua própria interpretação, por vezes apoiada em explicações fornecidas em outro lugar, como um texto ou uma apresentação.

Geralmente, em artigos científicos e outras formas de comunicação internas ao ambiente acadêmico, o uso de visualizações possui objetivos explicativos.

Para aprofundar o entendimento sobre como os gráficos podem ter esse teor, vamos examinar características de categorias, grupos ou famílias dessas visualizações. Como existem muitos tipos de gráficos, as categorizações, a partir das características comunicativas deles, são úteis. Kirk (2019) fez a proposta, sumarizada a seguir, de descrever os gráficos em cinco grupos.

Categórico

Singular Fact (2020), LottieFiles



Enostra Agency (2022), LottieFiles

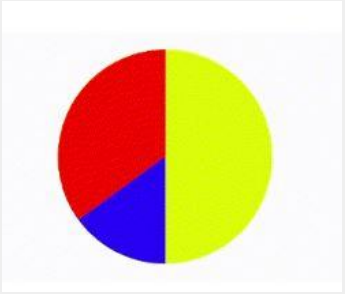
Gráficos deste grupo servem para comparar categorias e distribuições de valores quantitativos.

Alguns gráficos da família:

- Gráficos de barras: horizontais, verticais, agrupadas e empilhadas;
- Gráfico polar;
- Gráfico de radar;
- Gráfico de pontos;
- Historiograma;
- Nuvem de palavras.

Hierárquico

Mukhammad Ato'illah putra (2024), LottieFiles



Pallavi (2021), LottieFiles

Servem para destacar o relacionamento entre o todo e suas partes, bem como hierarquias.

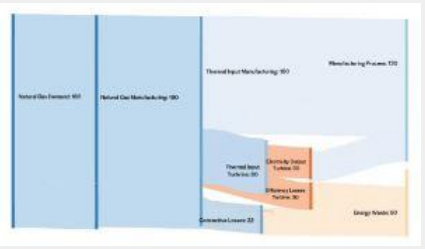
Alguns gráficos da família:

- Gráficos de setores (também chamado de gráficos de pizza), com a variação do gráfico de rosca ou donut;
- Gráfico mapa de árvore (treemap);
- Gráfico de waffle;
- Dendrograma;
- Diagrama de Venn.

Relacional

Dat esh (2022), LottieFiles

Evan Skjel (2020), LottieFiles



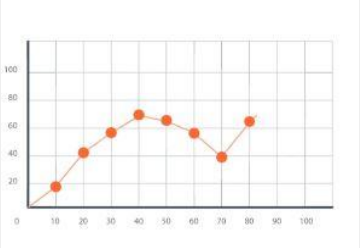
São úteis para explorar correlações e conexões.

Alguns gráficos da família:

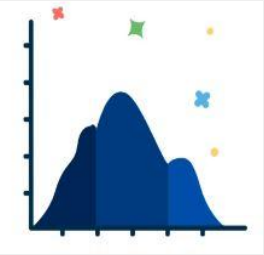
- Gráficos (ou visualizações) de rede;
- Gráfico aluvial (ou sankey);
- Diagrama de corda;
- Gráfico de dispersão;
- Gráfico de bolhas.

Temporal

Priyanshu Rijhwani (2024), LottieFiles



Canberk (2021), LottieFiles



Representam graficamente tendências e intervalos ao longo do tempo.

Alguns gráficos da família:

- Gráfico de linha ou gráfico de ranking (bump chart);
- Gráfico de inclinação;
- Gráfico de área;
- Gráfico de fluxo contínuo;
- Gráfico de Gantt;
- Gráfico de instância.

Espacial

Flavia Bernárdez (2020), LottieFiles

Zack Rodger (2020), LottieFiles

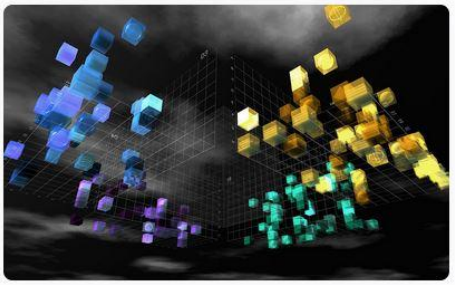
Essas visualizações produzem o mapeamento de padrões espaciais por meio de sobreposições e distorções.

Alguns gráficos da família:

- Mapa com pinos;
- Mapa de fluxo;
- Mapa de conexão;
- Mapa isoplético;
- Mapas de densidade de pontos;
- Mapa coropléticos;
- Cartograma.

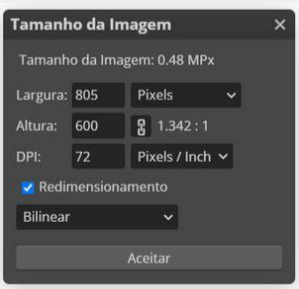
No próximo tópico, serão mostradas recomendações para produzir gráficos com qualidade, com tutoriais que exemplificam o uso dos programas mencionados.

Produção e leitura de visualizações



As decisões de design afetam a eficácia das visualizações de dados. Sosulski (2019) procura, recorrendo a diferentes especialistas, sugerir padrões essenciais de design aplicáveis às visualizações, de maneira geral. O conhecimento desses dez padrões, colabora na produção de gráficos de mais qualidade.

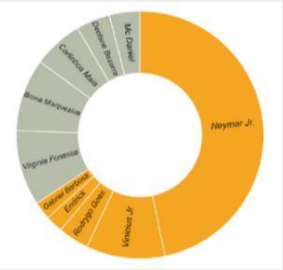
#1 FORMATO DE ARQUIVO DO GRÁFICO

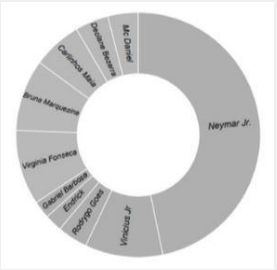


A legibilidade de um gráfico está diretamente ligada à **resolução** e ao **formato** do arquivo. Para impressões de qualidade em papel, o ideal é 300 pontos por polegada (dpi), e para web de 150. Acima está uma caixa de opção de programa de edição, mostrando onde alterar esse parâmetro.

Formatos de arquivo usuais para o primeiro meio são TIFF, EPS e PSD. Já para o segundo, JPG, PNG e GIF. O formato **SVG possui vários diferenciais** interessantes, principalmente o fato de ser escalável, o que o torna um arquivo de trabalho bastante útil. As imagens dos gráficos podem ser retrabalhadas, com alterações de cores em vários programas on-line como **Photopea** e **Boxy SVG**.

#2 COR

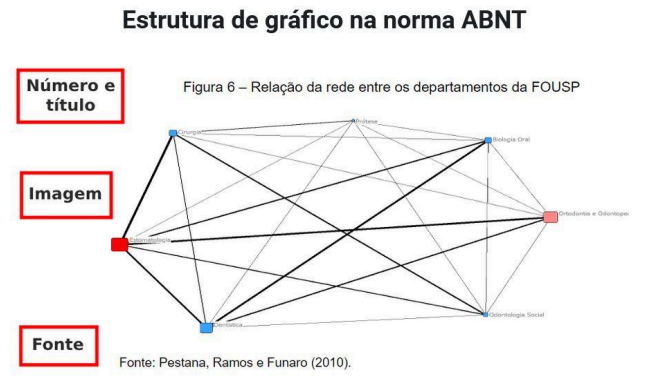


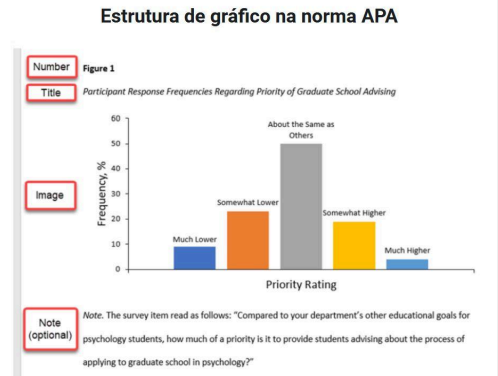


Cores devem ser usadas apenas quando corresponderem a diferenças nos dados. Por vezes, podem ser utilizadas quando se quer destacar apenas um aspecto do gráfico, como uma barra ou linha específica. No exemplo acima, de um gráfico de rosca com dados de seguidores no Instagram de dez influenciadores brasileiros, a cor assinala os jogadores de futebol. Porém, é importante garantir contrastes de cor que facilitem a visualização também em escala de cinza. Esse aspecto é prejudicado, no caso, e o destaque desejado é perdido. O **valor simbólico e cultural das cores** é outro aspecto que merece reflexão.

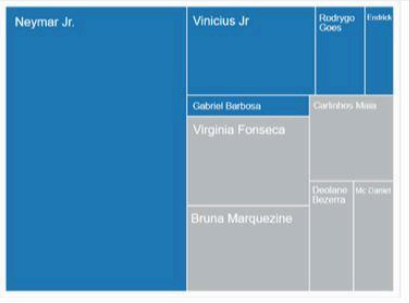
#3 ESTRUTURA E TEXTO

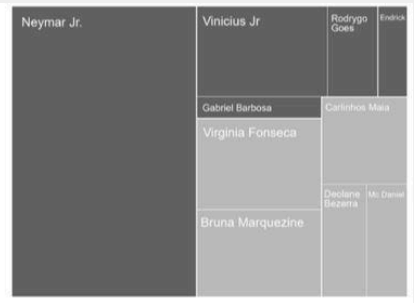
Geralmente quando inseridos em textos acadêmicos, os gráficos possuem numeração sequencial e títulos descritivos. No entanto, a forma exata depende do padrão utilizado por alguma publicação ou da norma que deve ser utilizada. Veja como se estruturam gráficos nas normas **ABNT** e **APA**. Nesse aspecto, é importante garantir a uniformidade formal, ao longo de um trabalho.





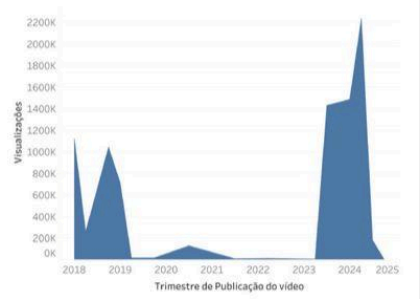
#4 LEGIBILIDADE





Aspectos como o esquema de cores, o tamanho, a família tipográfica e a direção do texto afetam a capacidade de leitura de um gráfico. Textos na horizontal são mais fáceis de ler. O uso excessivo de fontes em itálico e negrito também deve ser evitado, por razões de legibilidade. Todos os elementos textuais do gráfico (rótulos de eixo, escalas, rótulos de dados etc.) devem ser legíveis. No exemplo acima, em gráfico de árvore a partir dos mesmos dados sobre influenciadores locais, há dificuldade de leitura nos nomes de influenciadores que não são jogadores de futebol, devido a um problema de contraste.

#5 ESCALAS



Os eixos **x** e **y** dos gráficos devem possuir incrementos lógicos (0, 1, 2, 3, 4...; 0, 2, 4, 6, 8...; 0, 10, 20, 30, 40...; 0, 50, 100, 150, 200, 250... etc.), mas não necessariamente iniciando em zero.

É recomendável que o valor final do eixo y esteja próximo, em alguma medida superando, do maior valor de algum dado neste eixo (como nos casos acima). Veja outro **exemplo**.

Os dois gráficos de área mostrados acima foram construídos com os **mesmos dados** sobre visualizações de vídeos no YouTube que mencionam o termo “Marielle”. No entanto, o período de agregação dos dados do gráfico da esquerda foi anual e o outro, trimestral. Isso gerou a mudança na forma geral que é visualizada. Dependendo do que se quer destacar num trabalho, um ou outro gráfico poderá ser mais adequado.

#6 INTEGRIDADE DOS DADOS

O livro influente de Edward Tufte **The Visual Display of Quantitative Information** (2007/1983) introduziu noções, como a de **integridade gráfica** e **fator de mentira** (*lie factor*), relevantes para a discussão sobre como a apresentação visual pode induzir interpretações enganosas dos dados. A forma principal de manipulação, intencional ou não, é quando a codificação visual distorce o tamanho da correspondência entre os valores. Esse é o caso, bastante evidente, do gráfico acima, discutido numa **postagem** sobre o tema.

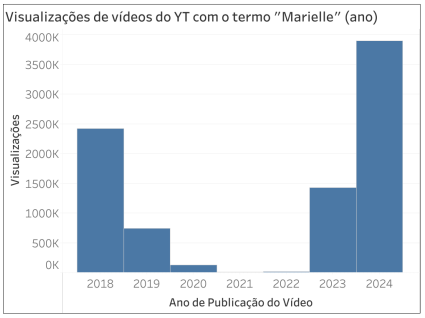
A representação seletiva de dados ou períodos de tempo relacionados a eles, o uso de eixos não rotulados ou enganosos, a apresentação de gráficos 3D que confundem proporções são alguns outros pontos que prejudicam a integridade visual dos dados.

A escolha de um modelo inadequado de gráfico pode também prejudicar a interpretação dos dados. Os dois gráficos mostrados no item 5 possuem problema. Examine o ***dataset*** e reveja-os. **Qual o problema?** Veja a resposta, a seguir

O site **VisLies** apresenta galerias anuais com visualizações que induzem a erros de interpretação.

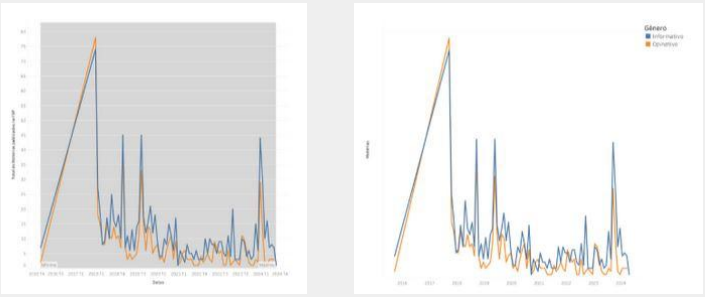
Resposta: Gráficos de linha sugerem uma tendência e uma continuidade temporal que não podem ser inferidas dos dados. Em outras palavras, não é possível dizer que os vídeos publicados em 2018 foram vistos neste ano ou em ano posterior. O *dataset* apenas informa o número de visualizações de cada vídeo, mas sem informar quando elas ocorreram. Por conta disso, uma visualização mais adequada dos dados seria a partir de um gráfico de barra, como abaixo, por exemplo.

Escolha adequada de gráfico



Gráficos de linha sugerem uma tendência e uma continuidade temporal que não podem ser inferidas dos dados. Em outras palavras, não é possível dizer que os vídeos publicados em 2018 foram vistos neste ano ou em ano posterior. O *dataset* apenas informa o número de visualizações de cada vídeo, mas sem informar quando elas ocorreram. Por conta disso, uma visualização mais adequada dos dados seria a partir de um gráfico de barra, como acima, por exemplo.

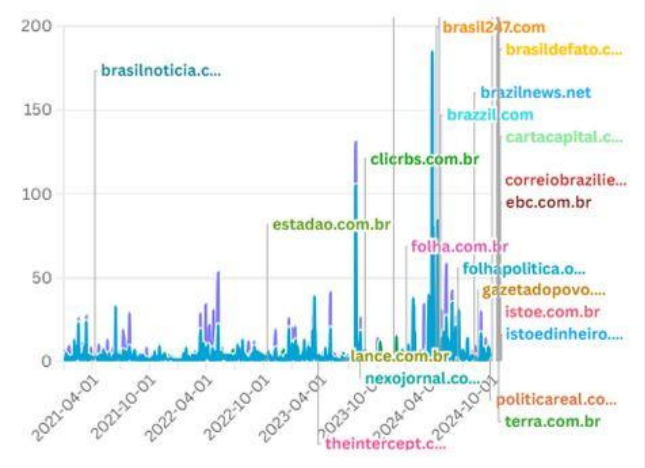
#7 RUÍDO VISUAL

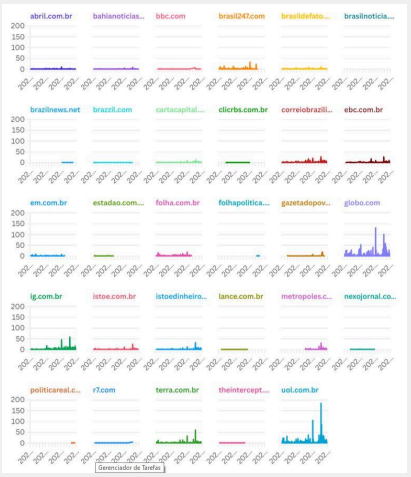


Os dois gráficos de linha acima mostram os mesmos dados referentes a matérias que mencionam Marielle Franco publicadas no jornal **Folha de S.Paulo**. O da esquerda dificulta a leitura dos dados, pelo excesso de grafismos. Elementos gráficos meramente decorativos, redundantes ou desnecessários nas visualizações desviam o foco da exibição dos dados. Desse modo, prejudicam a eficácia dos gráficos em análises de dados.

Vale a pena ver as transformações, em termos de eliminação de excessos visuais, em gráficos de barra, de pizza, tabelas, e mapas, produzidos pela empresa Darkhorse Analytics.

#8 DENSIDADE DE DADOS

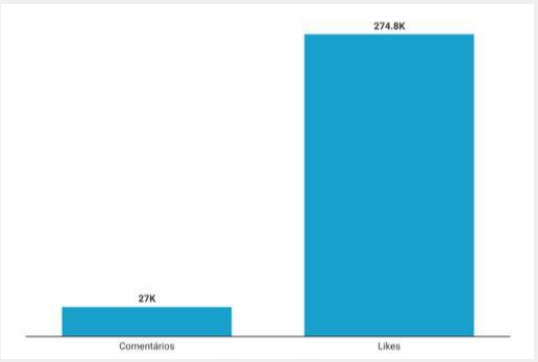




A noção de densidade de dados remete à quantidade de elementos (linhas, pontos, tipos etc.) inseridos no gráfico. Deve-se buscar uma relação adequada entre o que se mostra e a capacidade de identificar o que é relevante.

No gráfico acima, à esquerda, elaborado a partir das notícias com o termo “Marielle” publicadas em veículos on-line locais, há uma excessiva densidade. Isso torna inviável perceber os aspectos importantes que o gráfico poderia comunicar, por exemplo, qual veículo publicou mais. Resolver esse problema, a cada situação, poderá envolver a retirada de elementos redundantes (como no exemplo anterior), aumento do tamanho do gráfico ou escolha de outro tipo de visualização. Para gráficos de linha, uma possibilidade é o uso dos gráficos de **Pequenos Múltiplos** (*Small Multiples*), modelo proposto por Tufte. Em relação ao exemplo, no gráfico à direita, se torna claro que os veículos **UOL** e **O Globo** publicaram mais matérias sobre o tema.

#9 RIQUEZA DE DADOS



Este padrão está relacionado à **qualidade** e ao nível de **granuralidade** (detalhamento) dos dados. O primeiro aspecto se associa, além dos aspectos discutidos no Módulo anterior, a questões como: grau de confiabilidade da fonte dos dados, possuir descrições sobre a metodologia para a obtenção dos dados, suas variáveis e dimensões, bem como informar a data em que foram coletados. O nível de granuralidade depende do objetivo da visualização.

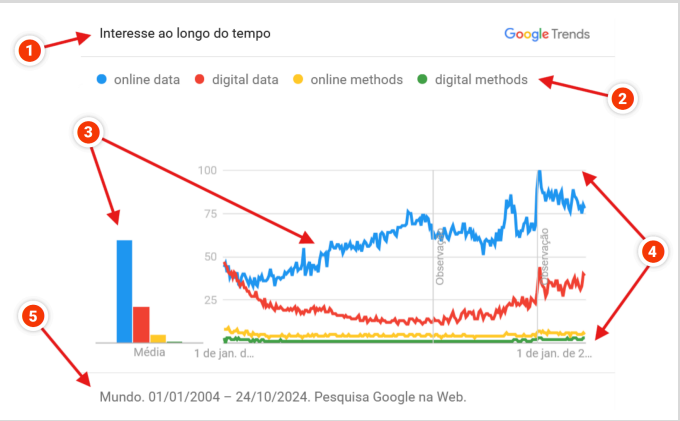
No entanto, note que o gráfico acima, elaborado a partir da contagem das interações (likes e comentários) de vídeos do YouTube que mencionam Marielle Franco possui menor granularidade do que **essa visualização**, que separa essas variáveis pelas categorias de vídeos, adicionando informação. Além disso, nesse modelo de gráfico dinâmico para web (conforme o mouse passa pela barra, são mostradas informações numéricas) é possível inserir link para os próprios dados com os quais foi produzido o gráfico.

#10 ATRIBUIÇÃO

É praxe, sobretudo na comunicação científica, indicar a fonte dos dados de gráficos, inclusive quando os dados foram produzidos por quem fez a pesquisa. Daí, isso é informado. Por vezes, podem ser inseridas informações relevantes para a compreensão da visualização: certo tipo de tratamento nos dados ou alguma opção de visualização (como a distribuição escolhida para os gráficos de rede) e eventualmente o próprio software utilizado.

A leitura de gráficos é uma habilidade associada à capacidade de produzi-los. A análise crítica preocupada com o modo como eles aparecem, em trabalhos acadêmicos e em geral, pode ajudar. É, inclusive, uma possível forma de inspiração para ajudar alguém a elaborar visualizações mais interessantes. A respeito da produção e leitura de gráficos, o projeto **Seeing Data**, que reúne várias pessoas que desenvolvem pesquisas na área de visualização de dados, possui um material de qualidade para estudar o assunto. Um dos conteúdos, **adaptado e mostrado a seguir**, sugere que se enfatize a leitura de cinco aspectos de qualquer gráfico.

Cinco aspectos de um gráficos para se prestar atenção



1. Título

- Existe?
- Fornece informação sobre o que esperar do gráfico?
- O gráfico efetivamente mostra o que o título afirma?

Comentário: na versão on-line desta visualização, há uma explicação sobre a elaboração do gráfico, notando que “os números representam o interesse de pesquisa relativo ao ponto mais alto no gráfico de uma

determinada região em um dado período. Um valor de 100 representa o pico de popularidade de um termo. Um valor de 50 significa que o termo teve metade da popularidade. Uma pontuação de 0 significa que não havia dados suficientes sobre o termo". Essa observação baliza a compreensão de aspectos como a escala (eixo y) e o teor dos dados.

2. Legenda ou rótulo

- O gráfico requeria esse elemento?
- As cores e formas utilizadas possuem algum significado especial?
- Se existe, colabora com a leitura e interpretação do gráfico?

3. Dados

- Em gráficos de linha, é válido entender o que significa o movimento para o espaço superior ou inferior dela. Algo positivo ou negativo? Percebe-se algum padrão numa linha ou entre elas? Existe alguma tendência sazonal?
- Já para gráficos de barra, a comparação coloca questões como: quais as maiores e menores categorias? Como elas se comparam?

4. Eixos e Escalas

- Quais são os eixos?
- Iniciam em zero?
- Qual é o intervalo representado na visualização?
- Qual é o significado dos valores **maiores** e **menores**?

No caso deste gráfico, ver o comentário no campo do título.

5. Fonte

- Os dados provêm de onde?
- A fonte é confiável?
- Se não há fonte, como saber se o gráfico é acurado e veraz?
- Informa algum tipo de especificidade ou tratamento dos dados?

No caso deste gráfico, ver o comentário no campo do título.

Concluindo esse tópico, você poderá ver, a seguir, pequenos tutoriais em vídeo que exemplificam a construção de gráficos com os programas e serviços mencionados.

4CAT - Mural de imagens do TikTok

Com um dos processadores analíticos do aplicativo foi feita a contagem de hashtags e produzido um mural de imagens das postagens.
Veja o tutorial

RAWGraphs - Gráficos de rosca, pizza e dendrograma

Os dois primeiros usaram os dados de influenciadores brasileiros e o dendrograma foi construído a partir dos links dos sites das entidades científicas da comunicação.
Veja os tutoriais: gráfico de rosca, gráfico mapa de árvore e dendrograma

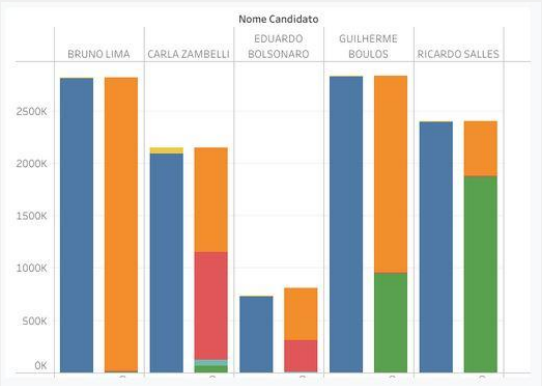
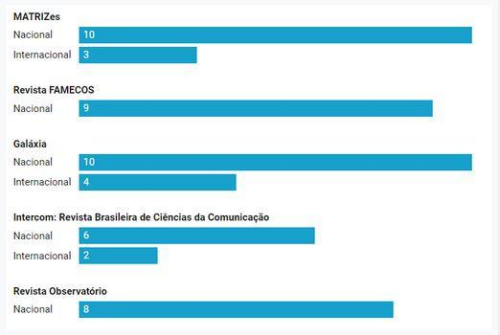


Tableau - Gráficos de barra, linha e área

O primeiro tutorial utiliza os dados do TSE de gastos e recursos de candidatos; o segundo, os das matérias de um jornal com o termo “Marielle”. Já o terceiro, o número de publicações de vídeos em um canal do YouTube.
Veja os tutoriais dos gráficos: barra, linha e área

Flourish - Gráficos de pequenos múltiplos (linha) e de barra dinâmico

O primeiro gráfico utiliza dados de notícias sobre Marielle Franco em veículos on-line e o outro, explorando as possibilidades interativas dos gráficos digitais, explora novamente os dados do TSE sobre gastos e recursos de candidatos.
Veja os tutoriais dos gráficos: pequenos múltiplos e dinâmico de barra



Datawrapper - Gráficos de barras horizontais agrupadas e dinâmico

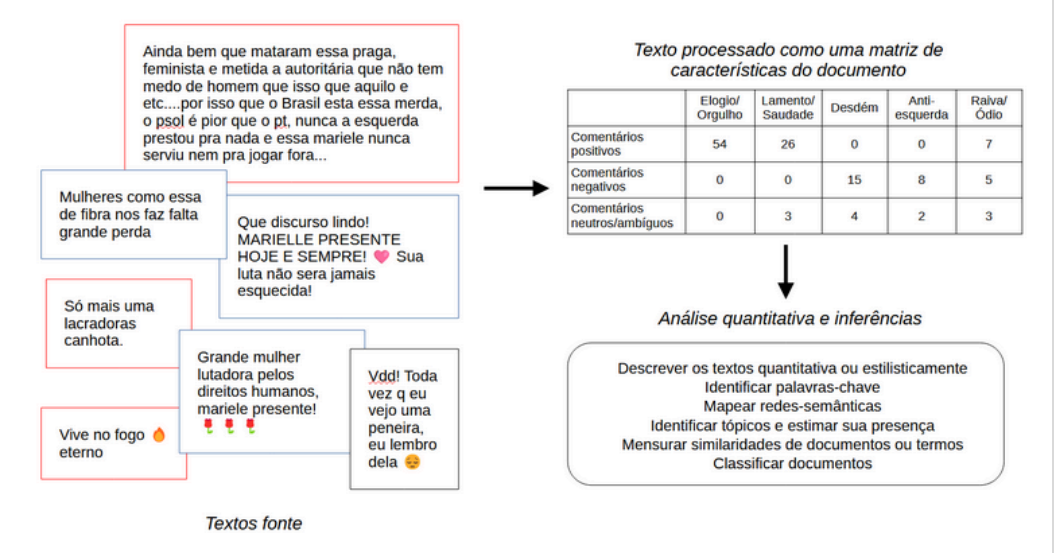
O primeiro destes gráficos de barras, permite comparar, entre um conjunto de revistas, o número de textos publicados de autoria com vínculo institucional no Brasil ou exterior, enquanto o segundo apresenta uma comparação entre interações de pessoas que viram vídeos sobre Marielle no YouTube e a categoria do conteúdo postado.
Veja os tutoriais de gráficos de barras: horizontais agrupadas e dinâmico

Análise textual

Romain Vignes (2015), Unsplash

Todos os tipos de texto contêm informações que podem ser tratadas como uma **forma de dado** em pesquisas, observa um **autor** (Benoit, 2020), que também destaca que isso significa transformá-los em dados mais estruturados, sintéticos e quantitativos. Desse modo, os textos podem ser utilizados em abordagens tradicionais de análise de dados. Geralmente, sobretudo na pesquisa quantitativa, isso envolve a extração de **características** de um texto, depois tabuladas e contadas. Veja o exemplo, a seguir.

Transformação de textos em dados quantitativos



Exemplo de estratégia geral de conversão de dados textuais em quantitativos para análises. Adaptado de Benoit (2020).

As ciências sociais possuem uma longa tradição de análise de textos para obter informações, a partir da codificação humana, em categorias elaboradas no processo de pesquisa. Porém, como argumentam certos **pesquisadores** (Izumi & Moreira, 2018), os textos eram, geralmente, usados com parcimônia devido à dificuldade de trabalhar com eles em larga escala. Isso muda a partir do advento e disseminação da internet, das ferramentas de análise computacionais e com os desenvolvimentos metodológicos associados. Nesse contexto, há diversidade de perspectivas e possibilidades analíticas no trabalho com o texto como dado. Ao mesmo tempo, a grande variedade entre os tipos de texto inviabiliza uma abordagem metodológica única, pois

> "o conteúdo que gostaríamos de extrair de um texto se estivermos interessados em conhecer seu tópico é qualitativamente diferente do conteúdo que extrairíamos se estivéssemos interessados em conhecer seu sentimento. Identificar a ideologia de um texto é bem diferente de identificar seu autor (uma tarefa do campo da estilometria). Os tipos de quantidades que os cientistas sociais esperam extrair dos textos são diversos e estão em constante crescimento" (Grimmer et al., 2022, p. 65).

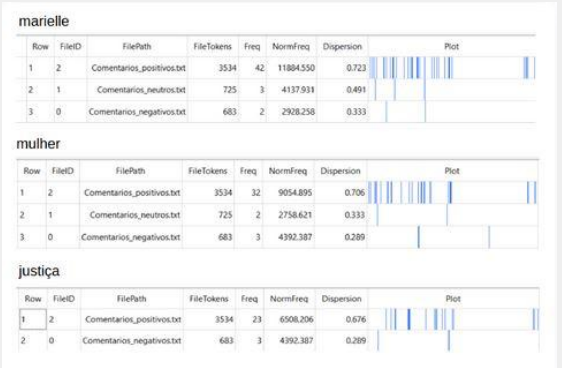
Embora não exista restrição quanto aos tipos de textos que possam ser analisados como dados, bem como quanto às formas de coleta, alguns **autores** (Balestrini et al., 2023) argumentam que um ponto forte da abordagem é sua associação com meios de obtenção de dados em larga escala e de maneira não reativa, como nas produções (postagens, comentários etc.) que as pessoas publicam na internet. Isso seria um aspecto que poderia contornar vieses de observação.

Dentre as abordagens de estudo do texto como dado, a linguística de corpus (LC) tem ganhado atenção, além de sua área disciplinar de origem, a Linguística. A LC se desenvolveu, desde a década de 1960, a partir da linguística computacional, adquirindo contornos mais específicos nas décadas seguintes. Embora seu estatuto, enquanto metodologia ou teoria de estudo, seja **debatido até hoje** (McEnery & Hardie, 2012), em seu campo de origem, suas técnicas de pesquisa foram adotadas por várias áreas e disciplinas das ciências sociais interessadas em reconhecer padrões de uso de palavras, para inferir o significado dos dados linguísticos.

Para **alguns** (e.g., Kutter, 2018), as limitações da análise automatizada fazem com suas técnicas sejam válidas, principalmente, em análises exploratórias que levem à geração de hipóteses e indagações a serem exploradas por outros métodos de análise textual. Nesses casos, como mostram **metanálises** (Pérez-Paredes & Curry, 2024), a LC é usada como complemento de análises em pesquisas que utilizam métodos mistos. Assim, ela pode estar conjugada a análises de discurso, conteúdo e outras estratégias.

A seguir, são expostas características de algumas técnicas usuais da LC, utilizando exemplos de uma análise no programa AntConc, tendo como corpus os comentários do vídeo com maior número de visualizações no canal do Instituto Marielle Franco. Esses comentários foram codificados em termos de exposição de sentimentos negativos, positivos e neutros/ambíguos em relação à Marielle e ao vídeo.

Frequência e dispersão

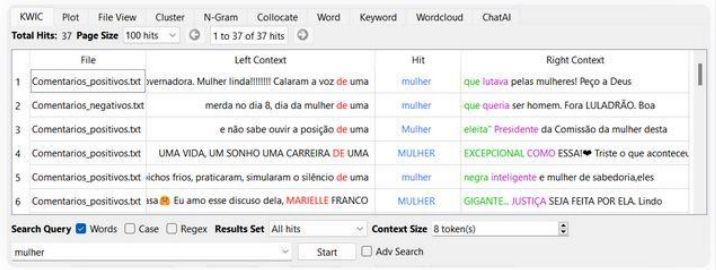


No AntConc, é possível fazer o processamento analítico de diferentes corpura textuais ao mesmo tempo. Assim, os arquivos com comentários positivos, negativos e neutros foram, como mostra o resultado da imagem, verificados em termos da **frequência** e **dispersão** de palavras. Essas medidas fornecem informações básicas sobre a importância de palavras em textos. No caso, a análise se concentra nas três palavras com mais ocorrências no todo, mostrando como aparecem em cada um dos corpura.

Os dados das medidas de frequência de palavras indicam ocorrências absolutas e relativas, sendo a última mais adequada para comparações. Uma noção importante é a de **token**, que significa cada conjunto contíguo de caracteres. Vale notar que é possível fazer com que o programa não recupere palavras comuns, mas sem significado analítico (“e”, “a”, “o”, “para” etc.). Além disso, pode ser necessário acrescentar tokens no programa para que sejam contados certos caracteres que tenham valor para uma pesquisa, mas que por padrão são ignorados (consulte o Manual do AntConc sobre isso), o que ocorre com os sinais gráficos de hashtag e arroba.

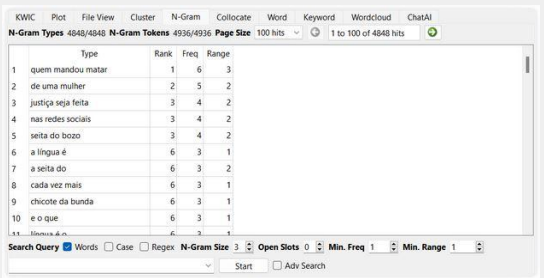
A dispersão de palavras descreve a distribuição de algum termo no texto ou documento. Como os três corpura foram compostos por comentários do YouTube, com ordem textual em uma temporalidade do período mais recente (2024) ao início dos comentários (2022), é possível inferir que a preocupação com “justiça” esteve mais presente nos comentários positivos, no início e desde a metade do tempo até o período mais recente das publicações. A imagem do Plot favorece a percepção disso.

KWIC



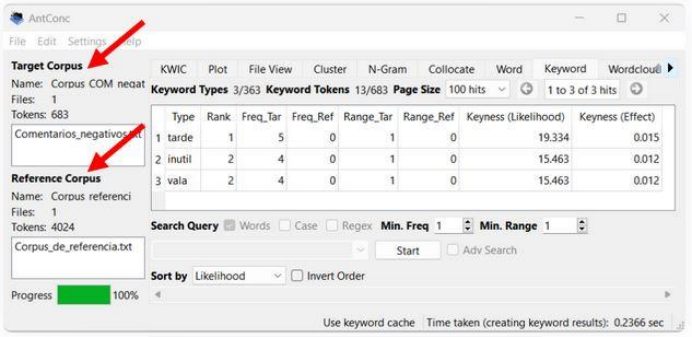
A investigação de **palavra-chave em contexto** ou KWIC, conforme o acrônimo em inglês, examina os padrões de coocorrência de palavras adjacentes ou próximas, sendo frequentemente usada em análises exploratórias para entender quais tipos de palavras se agrupam nas proximidades de certo termo. A palavra-chave no centro de um quadro contextual é conhecida como palavra-nó (**Hit**, no termo do AntConc, como mostra a imagem acima) e as adjacentes ajudam a compreender o sentido mais exato de uso dela.

N-grams



Os padrões de **colocação entre palavras contíguas**, cujo número é chamado **n-grama**, são úteis para perceber grupos de palavras semanticamente importantes. Tais padrões podem ser localizados sem a especificação de alguma palavra-chave, como no caso do exemplo da imagem, em uma exploração puramente indutiva. O número de palavras contíguas a serem recuperadas é uma escolha de quem faz a pesquisa.

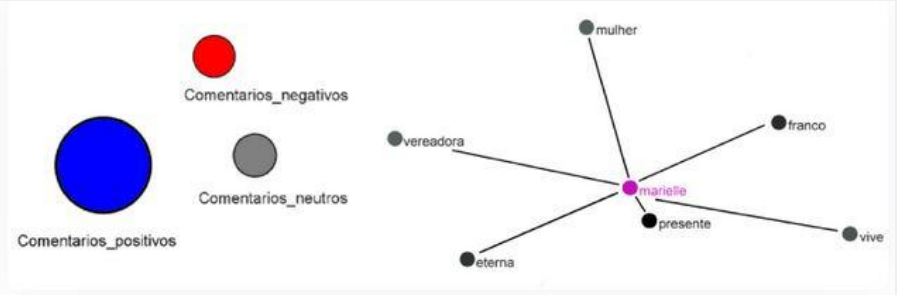
Palavras-chave



A identificação de palavras-chave também examina um tipo de colocação ou coocorrência entre corpura. O objetivo é localizar, a partir da comparação, palavras que aparecem no corpus analisado (“Target Corpus”) em taxas muito maiores ou menores do que seria esperado, a partir dessa comparação com um corpus de referência (“Reference Corpus”).

Quando determinada palavra ocorre em um documento com frequência significativamente maior ou menor do que a esperada com base nas frequências observadas desse tipo de palavra em um ou mais documentos diferentes, isso tem valor analítico. É importante que o **corpus de referência**, usualmente pelo menos cinco vezes maior que o outro, tenha justificativa lógica. Por exemplo, no caso mostrado na imagem, o corpus em análise, consistindo dos comentários negativos ao vídeo sobre Marielle, foi comparado com uma coleção muito maior de comentários em outros vídeos envolvendo a ex-vereadora.

Visualização



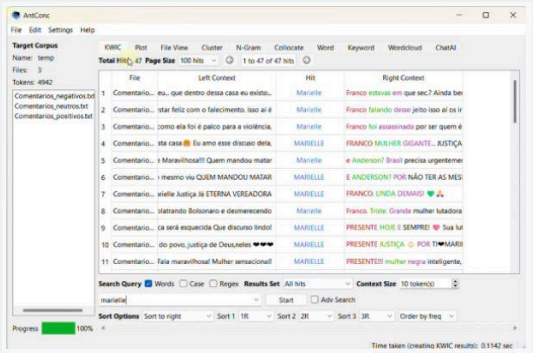
As **nuvens de palavras** são, provavelmente, o formato de visualização mais diretamente associado aos trabalhos que usam textos como dados. Entretanto, como discute Laurence Anthony (2018), criador do software AntConc, muitas outras visualizações, como os gráficos de barra, linha, mapas de calor, são também utilizadas, nos estudos da LC, para dar expressão visual às análises de

aspectos já discutidos. Isso é natural, devido ao teor quantitativo da abordagem.

Como nota o autor, cada um dos métodos de visualização têm pontos fortes e fracos, por isso é importante entendê-los antes de escolher uma visualização apropriada para a análise. Há **discussões** (Allen, 2018) sobre esse aspecto e trabalhos que fazem uso mais sofisticado de visualizações para comunicar dados textuais. Veja um **exemplo**.

O AntConc gera nuvens de palavras e plots. O **#Lancsbox**, um programa similar, é um pouco melhor nesse aspecto, já que possui um **módulo dedicado a visualizações**, no qual foram produzidas as figuras que ilustram a discussão.

Finalizando esse tópico, você poderá ver os tutoriais com um panorama do AntConc e sobre como produzir uma nuvem de palavras, além de ver outras informações com o aplicativo on-line Voyant Tools.



AntConc - Panorama geral

O tutorial apresenta um breve panorama do programa voltado à Linguística de Corpus, mostrando como importar dados e realizar algumas análise nele.
Veja o tutorial

Voyant Tools - nuvem de palavras e outros recursos

O aplicativo on-line tem uso intuitivo e permite elaborar nuvens de palavras, escolhendo o nível de detalhamento, bem como outras informações sobre textos.
Veja o tutorial

Análise de redes sociais

A Análise de Redes Sociais (ARS), de acordo um **pesquisador** (Scott, 2019, p. 85) reconhecido no campo, é

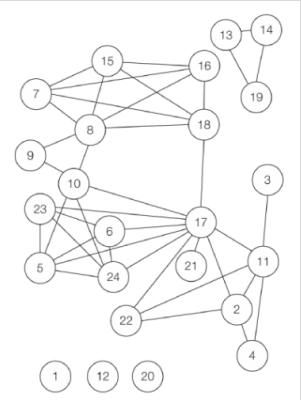
> "um conjunto de conceitos, medidas e técnicas de análise relacional. Trata-se de uma abordagem especificamente concebida para apreender as características mais importantes das estruturas sociais ... pode ser usada para explorar as relações sociais em si e também as estruturas culturais de normas e ideias que ajudam a organizar essas relações".

História da ARS

A abordagem possui uma longa história nas ciências sociais, que remonta aos primórdios de disciplinas como a sociologia e a antropologia. Na área da comunicação, nas últimas décadas, houve forte crescimento de seu uso, principalmente nos **estudos da mídia social e comunidades on-line** (cf. Quan-Haase et al., 2024),

Já nas primeiras décadas do século XX, a sociologia alemã, com autores como Simmel (1858-1918), destacava que a sociedade era constituída por meio das interações entre indivíduos. Essa ideia é formalizada, em décadas posteriores, principalmente na psicologia social. Jacob Moreno (1889-1974), por exemplo, criou os chamados **sociogramas** (como o mostrado a seguir), como uma forma de representar visualmente as redes sociais com padrões de pontos e linhas. A abordagem desenvolvida, com influência da **teoria do grafo**, é chamada de **sociometria**, e passa a ser utilizada em estudos de "dinâmicas de grupo" e de comunidades maiores.

Sociograma de relacionamento de Moreno



Reproduzido de **What is Social Network Analysis?**.

De modo simultâneo, nos Estado Unidos, métodos formais, principalmente os das teorias de conjuntos algébricos, foram utilizados no desenvolvimento de um paradigma para a ARS. Houve, assim, certa complementaridade entre as perspectivas: enquanto a teoria do grafo utilizada nos estudos sociométricos se concentrava nas interações entre indivíduos, a teoria dos conjuntos destacava as posições, funções e papéis ocupados por eles na estrutura social revelada.

Afirma-se (Scott et al., 2024) que a ARS não é, precisamente, uma abordagem teórica, mas sim uma orientação teórica geral (paradigma) que enfatiza os relacionamentos entre atores, cujos métodos têm sido utilizados no desenvolvimento de teorias sociais específicas. Por vezes, ela dialoga com teorias sociais, com as quais compartilha algumas afinidades, apesar de diferenças, como a **Teoria Ator-Rede** (Venturini et al., 2018).

Como apontam algumas **autoras** (Quan-Haase et al., 2024), várias razões tornam a ARS atraente no estudo da mídia digital: as redes sociais na internet possuem grande número de pessoas que interagem entre si, a abordagem favorece a compreensão sobre esse objeto ao destacar a estrutura de relacionamentos; além disso, a abordagem é produtiva, devido a seu foco ser em como os recursos fluem em uma rede. O último ponto pode ser relevante, por exemplo, para examinar a disseminação de certo conteúdo.

As mesmas autoras notam que, do ponto de vista metodológico, a delimitação das fronteiras do estudo é feita geralmente a partir de uma abordagem **nominalista**. Desse modo, a natureza da pergunta de pesquisa indica quem será incluído na rede. Uma estratégia bastante utilizada é o agrupamento a partir de alguma hashtag.

Embora válida, essa estratégia requer precauções, já que, entre outros pontos, uma discussão similar pode envolver o uso de várias hashtags.

A ideia de entender uma estrutura social como uma rede está no centro da ARS, por isso, a importância de técnicas que transformem dados, digitais ou não, em visualizações de rede (**grafos**). A seguir, serão apresentados alguns conceitos básicos sobre os grafos e como se pode construir uma visualização de rede com diferentes programas.

Estrutura: nós e arestas

	João	Teresa	Raimundo	Maria	Joaquim	Lili
João	0	0	0	0	0	0
Teresa	1	0	0	0	0	0
Raimundo	0	1	0	0	0	0
Maria	0	0	1	0	0	0
Joaquim	0	0	0	1	0	0
Lili	0	0	0	0	1	0

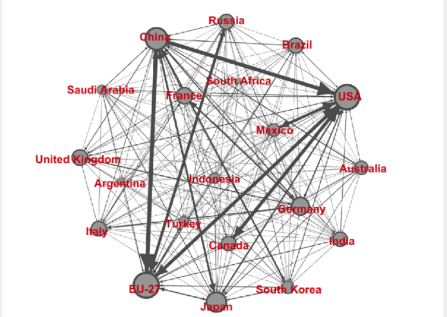
Source	Target
João	Teresa
Teresa	Raimundo
Raimundo	Maria
Maria	Joaquim
Joaquim	Lili

Gráficos de rede ou grafos dão forma visual a estruturas de relacionamento, mais ou menos explícitas, entre atores, representados nos chamados **nós** ou **vértices**. A ligação ou conexão entre os nós, que podem ser pessoas ou entidades como países, empresas, produtos e citações (dimensão comum em análises bibliométricas), é chamada de **aresta**. O direcionamento da relação é indicado por setas e nesse caso o grafo é chamado de **direcionado**.

O grafo acima é uma representação da primeira parte do famoso poema "Quadrilha", de Carlos Drummond de Andrade. A aresta representa a relação de "amor" e não possui, em nenhum caso, reciprocidade. Se houvesse, deveria haver duas arestas ou uma aresta com duas setas. Ao lado do grafo, há a representação da rede de relacionamento no formato de matriz e num padrão lido pelo programa Gephi.

Os nós e arestas podem ter algum **peso**, que expresse alguma medida e seja representado de modo numérico ou visual. Veja esse mostra o exemplo a seguir:

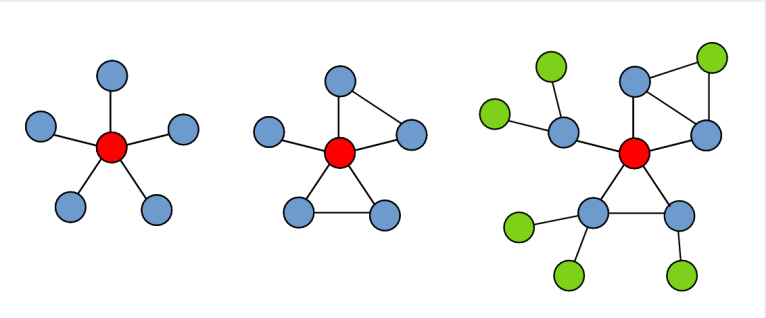
Grafo do comércio internacional dos países do G-20



O grafo exemplifica a ideia do peso para nós e aresta, explicando que: “As arestas direcionadas (links) entre quaisquer dois nós (países) são dadas pelo valor total do comércio entre esses países. A espessura das arestas e o tamanho de suas setas indicam as magnitudes dos fluxos comerciais em ambas as direções. Os tamanhos dos nós são definidos de acordo com o logaritmo do PIB de um país. É possível ver claramente como os nós Europa, China e Estados Unidos formam um grande triângulo global de comércio e como outros países se conectam a ele”. Reproduzido de **Netconomics: Novel Forecasting Techniques from the Combination of Big Data, Network Science and Economics**.

Os grafos podem ser organizados a partir de **estruturas egocêntricas** ou sem essa característica, nos gráficos inteiros. No primeiro caso, o grafo é organizado a partir de um nó central, o “ego”, e os seus contatos imediatos são chamados de “alters”.

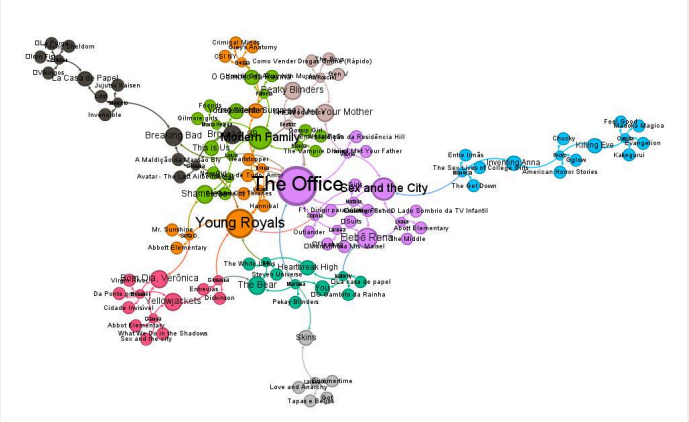
Tipos de grafos com estrutura egocentrada



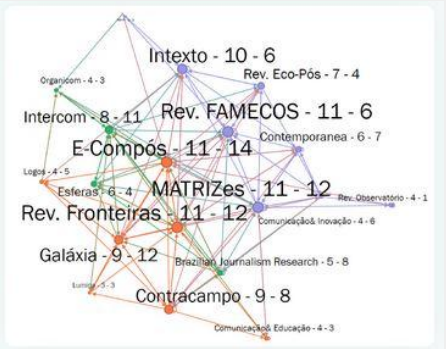
A complexidade dos grafos egocentrados pode variar, conforme mostram os exemplos.

É mais comum que as redes sejam elaboradas com um único tipo de ator como nó. No entanto, nas **redes de dois modos**, chamadas por vezes de **redes de afiliação**, as ligações ocorrem entre dois conjuntos distintos de entidades. Um exemplo, mostrado a seguir, é uma rede em que um conjunto seja de pessoas e o outro de séries de televisão preferidas por elas. Embora a ideia possa ser interessante, para determinado estudo, observa-se (e.g., Tomaél & Marteleto, 2013) que os conceitos analíticos da ARS são voltados a redes de um único modo.

Rede de dois modos



Centralidade



A **centralidade do grau** mede o número de conexões de um nó, somando as arestas conectadas a ele. Em gráficos direcionados, há um **grau de entrada** (*in-degree*) que corresponde às arestas que chegam ao vértice, e um **grau de saída** (*out-degree*), a partir do número de arestas que partem dele. A medida geral do grau de centralidade é a soma das anteriores.

O grafo acima mostra os graus de entrada e de saída, isto é, o quanto os artigos publicados pelas revistas mostradas receberam ou fizeram citações a trabalhos de outros periódicos do conjunto. Por isso, a partir de configurações no software que gerou a visualização, as revistas com maior grau ocupam posição mais central e possuem nós maiores.

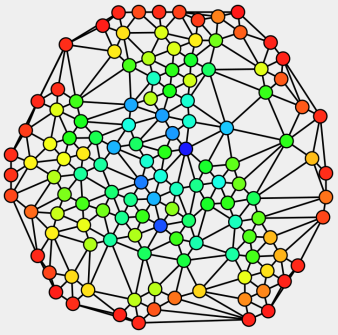
A centralidade de grau pode se associar, dependendo do tipo de relação mapeada, a aspectos como relevância, influência ou popularidade do ator representado em um nó.

Além desse tipo de centralidade (grau), há também as seguintes medidas:

- **Centralidade de intermediação:** verifica os nós que atuam como “pontes” na rede. O valor é dado pela contagem do número de vezes que determinado nó percorre o trajeto mais curto até os outros. Nós com alta centralidade de intermediação controlam fluxos de informação e recursos. Em redes de transporte, as estações centrais têm maior escore nesse quesito.
- **Centralidade de proximidade:** mede a proximidade de um nó em relação a todos os outros. Um nó com alta centralidade de proximidade tem acesso rápido aos demais nós, sendo capaz de se comunicar diretamente ou por poucos intermediários com o resto da rede. Alto escore nessa métrica se associa a nós com relevância, por exemplo, para a propagação de mensagens.
- **Centralidade de autovetor:** identifica as conexões diretas de um nó associadas à centralidade de seus vizinhos. Nós com alta centralidade de autovetor podem ser vistos como poderosos ou com prestígio, pois, ainda que não estejam conectados a muitos nós, suas conexões ocorrem com vértices importantes da rede. O algoritmo do buscador Google utiliza uma variante dessa medida.

Para entender melhor a centralidade de intermediação, veja o grafo a seguir.

Exemplo de centralidade de intermediação



A tonalidade (do vermelho=0 ao azul=máx.) indica a centralidade de intermediação de cada nó. Grafo de Claudio Rocchini, CC BY 2.5.

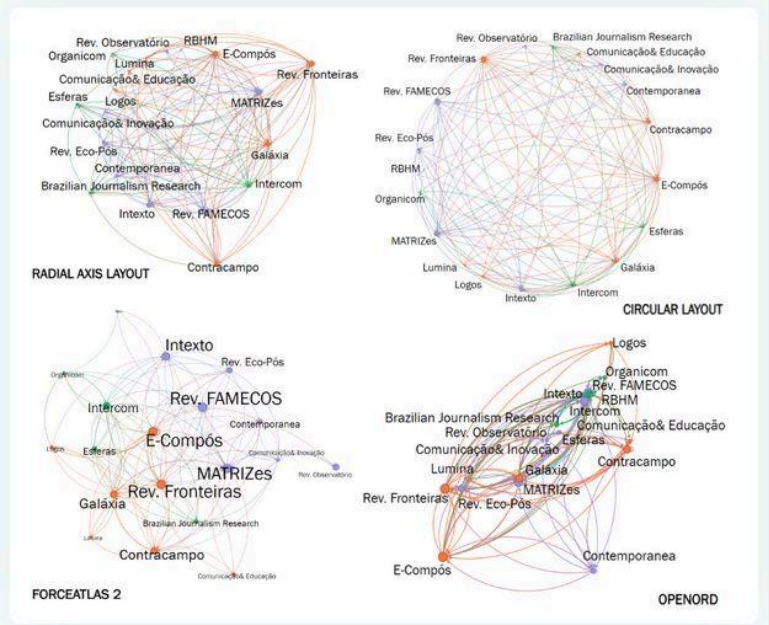
Programas como o Gephi realizam o cálculo de diferentes tipos de centralidade que poderão estar associadas à visualização.

Clusterização

O **coeficiente de clusterização** é a métrica da tendência dos nós de uma rede formarem clusters ou grupos. Está associada à **modularidade** de uma rede, ou seja, à propensão de determinados nós estabelecerem conexões com outros, formando grupos.

No grafo acima, com os mesmos dados de revistas, mas com design mais elaborado (sem números de grau e com arestas curvas), a cor dos nós e arestas (roxo, laranja e verde) está relacionada aos clusters formados a partir das conexões entre nós.

Distribuição

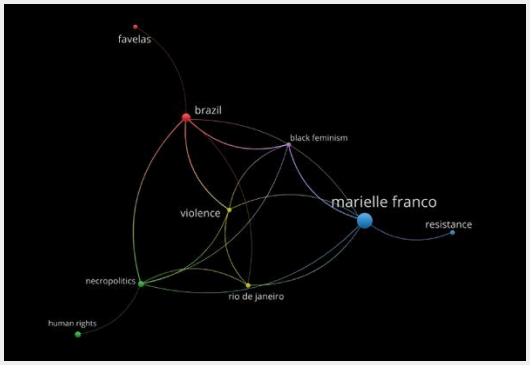


Os grafos podem ser desenhados de modo manual, em programas de desenho. No entanto, tem se tornado frequente o uso de aplicativos que produzem visualizações elaboradas, seguindo métricas como as vistas aqui. Geralmente essas ferramentas possuem os chamados **algoritmos de distribuição** que irão, a partir de suas características específicas, distribuir nós e arestas para produzir uma visualização. Há, entretanto, margem para personalizações, durante o uso do programa.

Como escolher uma distribuição/visualização? Em primeiro lugar, é essencial que a representação dos dados seja legível, além disso, a topologia do grafo deve destacar os aspectos para os quais se pretende apontar. Como se pode ver, acima, os mesmos dados serviram para elaborar grafos bastante diferentes, a partir dos algoritmos, indicados abaixo das imagens. Desse modo, conhecer as características dos algoritmos é importante, consultando materiais sobre o assunto, como **esse**.

Como bem observa Recuero (2017): "A visualização é uma forma de mostrar aquilo que as métricas calculam, e não uma justificativa *per se*. Ela deve ser, portanto, visualmente informativa daquilo que as métricas demonstram. Por conta disso, é sempre importante descrever quais métricas e algoritmos foram utilizados para a visualização" (p. 62).

Para concluir o estudo sobre a análise de redes, você pode ver os tutoriais, exemplificando o uso de programas na produção de grafos.

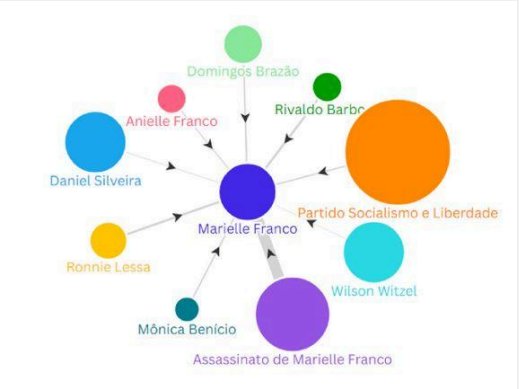


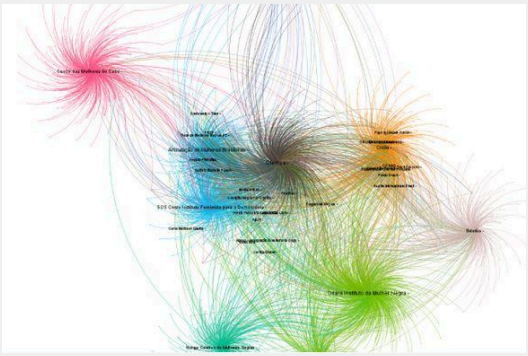
VOSviewer - Grafo bibliográfico

Exemplo de como produzir um grafo com palavras-chave de trabalhos acadêmicos sobre Marielle Franco, coletados do Scopus (veja esse tutorial sobre como fazer isso). Em programas que editam arquivos SVGs, como o on-line Boxy, é possível apagar o logotipo do VOSviewer. Entretanto, é válido informar o uso dele, em textos que utilizam grafos feitos com ele.
Veja o tutorial

Flourish - Grafo egocêntrico

É possível construir gráficos neste programa, como mostra o exemplo, a partir de verbetes da Wikipédia relacionados a Marielle Franco.
Veja o tutorial





Gephi - Grafos de relacionamento entre perfis do Instagram e entre citações de revistas

O primeiro tutorial utiliza os dados de perfis seguidos por organizações feministas brasileiras e o segundo, os dados de revistas científicas.
Em relação à primeira visualização, vale a pena notar o uso de filtros no Gephi, de modo a diminuir o número de rótulos de nós mostrados, bem como a aplicação da cor aos clusters do grafo.
Veja os tutoriais dos grafos: perfis, revistas

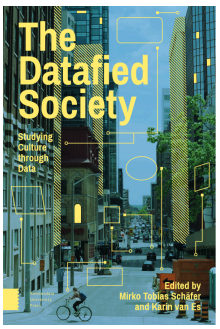
Você aprendeu muito sobre análises de dados, mas terá uma visão mais sólida sobre o tema, para efetuar seus próprios trabalhos, à medida em que ler criticamente estudos que utilizam metodologias como as expostas. Nesse sentido, sugere-se o exame dos seguintes artigos, cada um utilizando uma das estratégias de análise destacadas:

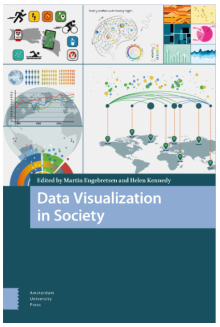
- "#PraCimaDeles: O Humor na Construção da Identidade Política de Guilherme Boulos", de Richard Romancini, Viviane Barbosa Marques e Fernanda Castilho Santana – **Palabra Clave**, **27**(4), e2749, 2024. Utiliza análises descritivas de dados do Instagram.
- "Ni Una Menos: A Luta pelos Direitos das Mulheres na Argentina e Suas Representações no Facebook", de Rodrigo Esteves de Lima-Lopes e Maristella Gabardo – **Revista Brasileira de Linguística Aplicada**, **19**(4), 801-824, 2019. Aplica a Linguística de Corpus, no trabalho metodológico.
- "A Endogamia da Comunicação: Redes de Colaboração na CSAI", de Marco T. Bastos, Gabriela Zago e Raquel Recuero – **Revista Famecos**, **23**(2), ID21459, 2016. Faz uso da ARS.

Uma observação final importante é que as metodologias ou técnicas descritas podem ser conjugadas entre si e a outras, como exemplifica a discussão de Recuero (2018) sobre a combinação de ARS e análise de conteúdo.

REAs de aprofundamento

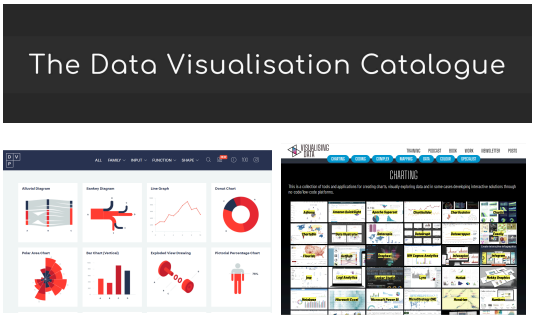
Materiais para estudos após o curso - Módulo 5

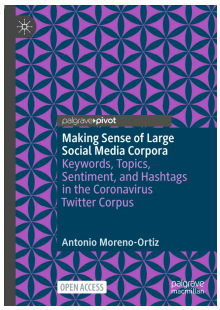


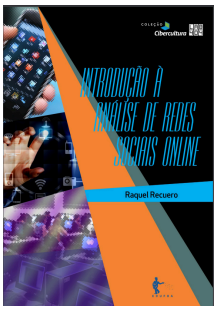


Os dois livros abertos, ambos coletâneas, abordam pontos relevantes a respeito do trabalho com dados. Um destaque de **The Datafied Society** é o capítulo "Towards a Reflexive Digital Data Analysis", de Karin van Es, Nicolás López Coombs e Thomas Boeschoten. Já **Data Visualization in Society** discute vários aspectos da *visualização de dados*.

Há sites na internet, como **The Data Visualisation Catalogue** e **Data Viz Project**, que têm a proposta de explicarem, ilustrando, diferentes tipos de gráficos. Já o **site do designer e pesquisador Andy Kirk** possui uma coleção de recursos para a feitura de visualizações.







As duas publicações abertas indicadas, **Making Sense of Large Social Media Corpus** e **Introdução à Análise de Redes Sociais**, permitem aprofundar conhecimentos sobre os temas estudados da análise textual de dados e a ARS.

Módulo 6

Ética em pesquisa com dados on-line

Objetivos de aprendizagem:

- Conhecer aspectos básicos ligados à ética na pesquisa e especificidades da investigação com dados on-line nesse contexto
- Refletir sobre práticas éticas em situações controversas envolvendo a pesquisa digital
- Perceber caminhos de aprofundamento de estudos dos temas do curso
- Revisar o que foi aprendido

Temas gerais

Freepik

A ciência objetiva o bem social, por isso preocupações éticas devem nortear seus procedimentos. Em pesquisas, isso significa incorporar o tema desde a elaboração das perguntas à publicização de resultados. As investigações que utilizam métodos digitais, em sentido amplo ou restrito, possuem questões éticas parecidas àquelas que não envolvem tecnologias e o ambiente da rede. Entretanto, há especificidades que merecem atenção.

Pontos centrais da ética na pesquisa envolvendo seres humanos são expostos na **Lei nº 14.874**, de 28 de maio de 2024, que instituiu o Sistema Nacional de Ética em Pesquisa com Seres Humanos. A Lei, em seu Art. 3, destaca as seguintes **exigências éticas e científicas** para as pesquisas com seres humanos:

> "I - respeito aos direitos, à dignidade, à segurança e ao bem-estar do participante da pesquisa, que deverá prevalecer sobre os interesses da ciência e da sociedade;
> II - embasamento em avaliação favorável da relação risco-benefício para o participante da pesquisa e para a sociedade;

> III - embasamento científico sólido e descrição em protocolo;
> IV - condução de acordo com protocolo aprovado pelo CEP;
> V - garantia de competência e de qualificação técnica e acadêmica dos profissionais envolvidos na realização da pesquisa;
> VI - garantia de participação voluntária, mediante consentimento livre e esclarecido do participante da pesquisa;
> VII - respeito à privacidade do participante da pesquisa e às regras de confidencialidade de seus dados, garantida a preservação do sigilo sobre sua identidade;
> VIII - provimento dos cuidados assistenciais necessários em casos que envolvam intervenção;
> IX - adoção de procedimentos que assegurem a qualidade dos aspectos técnicos envolvidos e a validade científica da pesquisa;
> X - condução da pesquisa em plena compatibilidade com as boas práticas clínicas".

Os riscos de prejuízo a pessoas e grupos envolvidos em pesquisas são diversos. Ao mesmo tempo, o balanço entre riscos e benefícios é complexo, pois nem sempre quem pesquisa tem clareza sobre os potenciais danos de sua investigação. Desse modo, reflexões cuidadosas ao longo do processo de pesquisa se impõem.

Um aspecto central da ética na pesquisa é o âmbito metodológico, pois, como nota um **autor** (de La Taille, 2008), quem conduz a investigação, geralmente, irá interferir na vida dos participantes da pesquisa e isso pode, de algum modo, ferir a dignidade dessas pessoas. O mesmo estudioso observa que

> "Não existe 'risco zero'. Todo e qualquer método pode ser prejudicial para o sujeito da pesquisa, pois um mero questionário pode, por exemplo, desencadear angústias imprevisíveis em quem o responde. Quiséssemos o 'risco zero', não faríamos pesquisa (e nem entraríamos em qualquer interação humana!). Todavia, há métodos que, mais do que outros, apresentam claramente riscos" (p. 275).

Controvérsias específicas da pesquisa digital

Ruth Suehle (2010), CC BY-SA 2.0

Um **autor** (Nunes, 2019) observa, com propriedade, que há três aspectos que geram reflexões e preocupações específicas, em relação à ética, na pesquisa on-line:

1. A compreensão entre o que é público ou privado

Usualmente, interações sociais em locais públicos podem produzir dados de pesquisa. Isso não exime quem pesquisa de preocupações éticas, mas a distinção tradicional entre ambientes off-line públicos (ruas, praças, praias etc.) e privados colocava um limite claro entre uma situação com menos formalidade na regulação da pesquisa e outra diferente.

No entanto, o contexto on-line problematiza a distinção rígida entre essas esferas. As pessoas que utilizam a internet e as redes sociais, ainda que interajam ou façam publicações em modo "público", costumam expressar preocupação sobre como o que produzem será utilizado por outros indivíduos.

Nesse sentido, é válido refletir sobre o uso de dados pessoais da internet, ainda que eles possam ser, em tese, públicos. É nessa linha que vai, por exemplo, o documento com **Diretrizes Éticas** da Association of Internet Researchers (AoIR). **Autoras** (Markham & Buchanan, 2017) influentes na discussão do assunto defendem, assim, localizar as questões éticas nos contextos específicos das investigações, e

> "uma abordagem baseada em casos que reconheça e considere as tensões éticas como conflitos com considerações legais, disciplinares, institucionais e culturais .... Ao colocar questões éticas de forma consistente e refletir sobre o processo de pesquisa, quem investiga irá equilibrar melhor suas diferentes obrigações" (p. 204).

2. A necessidade e a forma de obtenção do consentimento livre e esclarecido

A noção de "consentimento informado" decorre da ética médica e, com o tempo, passou a ser parte central da ética da pesquisa de diversas disciplinas quando as investigações envolvem seres humanos. Geralmente, o uso de um **Termo de**

Consentimento Livre e Esclarecido (TCLE) operacionaliza o conceito na prática. Conforme define a **Lei nº 14.874**, o TCLE é o "documento no qual é explicitado o consentimento livre e esclarecido do participante da pesquisa, ou do seu responsável legal, de forma escrita, com todas as informações necessárias, em linguagem clara e objetiva, de fácil entendimento, para o completo esclarecimento sobre a pesquisa da qual se propõe participar" (Art. 2º, LIII).

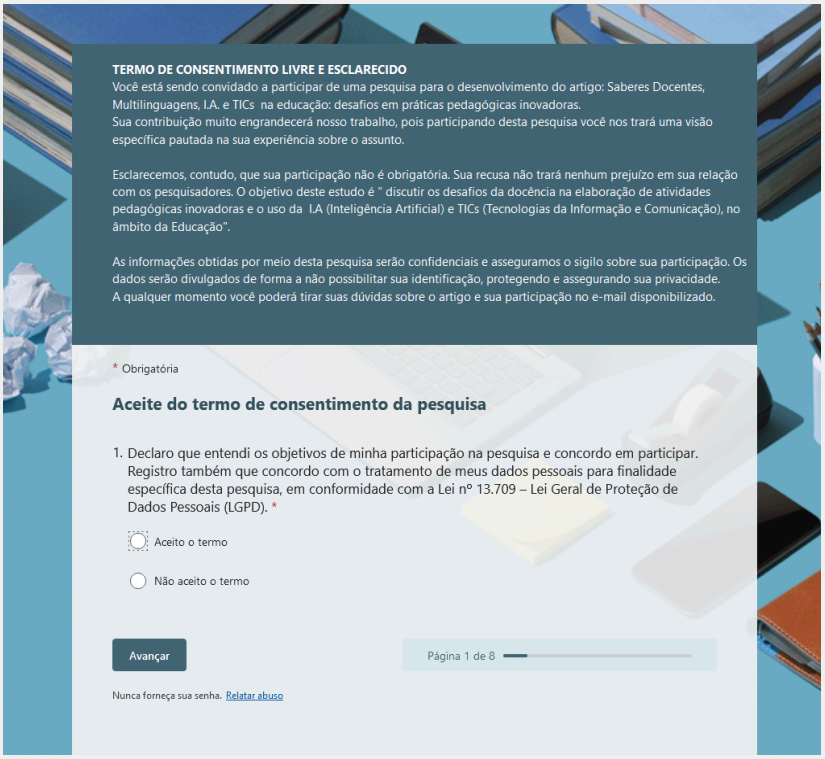
O modo como se dá o aceite deste termo, na maior parte das vezes, evidencia uma diferenciação entre a pesquisa off-line e a que utiliza o ambiente digital, particularmente a quantitativa, pois, como nota uma **estudiosa** (Flick, 2016, para. 6):

> "Enquanto a ética tradicional da pesquisa envolvendo seres humanos geralmente envolve o contato face a face entre quem pesquisa e o indivíduo pesquisado, permitindo que ocorra uma conversa, a tecnologia elimina esse contato, diluindo significativamente a capacidade de quem solicita o consentimento de avaliar a autonomia, a competência e a compreensão de quem poderá consentir, e destas pessoas de entender as minúcias da explicação".

A **Resolução CNS Nº 510/2016** nota que o processo de consentimento ou assentimento livre e esclarecido pode acontecer em qualquer fase da pesquisa. O momento em que isso ocorre depende da pesquisa, pois, embora o receio de não obter o consentimento após terem sido feitas observações seja compreensível, há também o risco de que essa solicitação possa perturbar e enviesar o contexto de pesquisa, como o de algum grupo on-line. Assim, o momento em que o pedido será feito é uma decisão que envolve ponderação. Entretanto, em pesquisas de levantamento utilizando a internet como meio de coleta de dados, a praxe tem sido explicar textualmente a pesquisa, mostrando os termos de consentimento (veja exemplos de TCLEs desse tipo, a seguir), que deve ser aceito antes do início efetivo da coleta de dados.

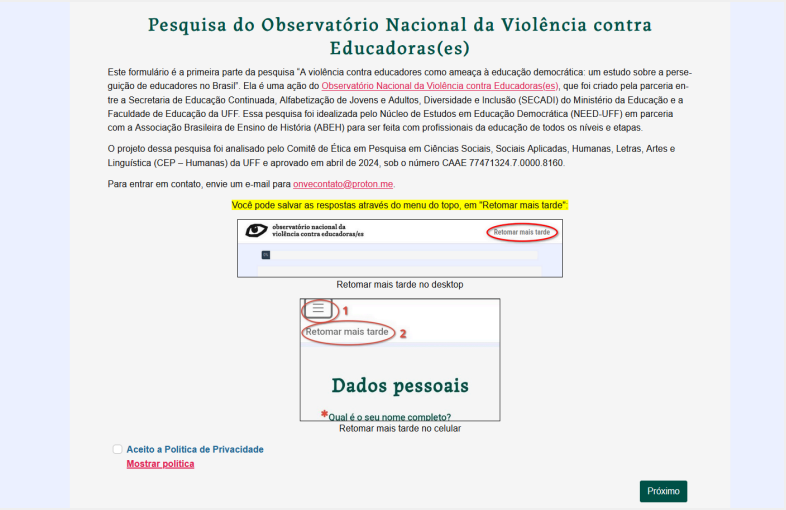
Por outro lado, nem sempre é possível obter o consentimento informado de um número grande de pessoas, como ocorre na pesquisa com *big data*. A tendência internacional tem sido tratar esses dados a partir de agregações das informações e sem que exista possibilidade de identificação individual de qualquer pessoa. A Resolução mencionada, no parágrafo único de seu Art. 1, ao notar que pesquisas desse tipo não precisam ser registradas e avaliadas por Comitês de Ética abona essa prática.

Exemplo 1



TCLE da pesquisa **Saberes docentes, multilinguagens, IA e TICs na educação**, de Rosália M. N. Prados, Fernanda Castilho, Denise Maria Martins e Rodrigo Avella Ramirez.

Exemplo 2



Tela inicial do instrumento de coleta de dados digital da da pesquisa **A violência contra educadores como ameaça à educação democrática**, do Observatório Nacional da Violência contra Educadora(es). A seguir, a exposição da Política de Privacidade e o TCLE dessa investigação.

Política de privacidade

Seja bem-vindo/a à **Pesquisa do Observatório Nacional da Violência contra Educadoras(es)**!

No **Observatório Nacional da Violência contra Educadoras(es)**, priorizamos sua privacidade. Esta política de privacidade explica como lidamos com dados pessoais, como eles são processados e para que finalidades suas informações são coletadas. Deixamos claro quais são as obrigações e direitos que correspondem a você

Buscamos total transparência sobre a finalidade da pesquisa e como ela afeta os dados que você nos fornece. Nosso dever legal é informar, e você, como usuária/o civilmente capaz, ao aceitar nossos termos e condições, declara estar devidamente informado e ciente.

É importante destacar que a **Pesquisa** está em conformidade com a legislação vigente em matéria de proteção de dados, incluindo a Lei Geral de Proteção de Dados Pessoais (baseada na lei europeia), a Lei Brasileira n.º 12.965/2014 e o Regulamento Geral de Proteção de Dados da União Europeia (RGPD). A coleta e tratamento de dados pessoais são realizados com seu consentimento expresso e dentro dos limites por você concedidos para o funcionamento do site.

1. Dados de Identificação

Antes de abordarmos os dados coletados, fornecemos as informações de contato do responsável pelos dados: Prof. Dr. Fernando de Araújo Penna, Diretor da Faculdade de Educação da UFF e coordenador da Pesquisa, endereço de e-mail: onvecontato@proton.me.

Os dados pessoais que você fornecer serão armazenados e processados de acordo com os fins descritos nesta Poiítica de Privacidade, até que você solicite a sua remoção. Além disso, reservamos o direito de exigir que determinados dados sejam obrigatórios para a participação de nossa pesquisa. A falta de fornecimento de dados exigidos por lei pode impedir o acesso a certas seções.

Informamos que esta Política de Privacidade pode ser alterada a qualquer momento para se adequar a novas leis ou mudanças em nossas atividades, com notificação expressa em caso de alterações.

2. Condições de Uso

Para sua tranquilidade, sempre solicitamos seu consentimento expresso para a coleta de dados, especificando a finalidade em cada caso. Ao acessar e utilizar esta página da web, você aceita sua condição de usuário com direitos e responsabilidades correspondentes.

O **ONVE** nunca solicitará informações pessoais, a menos que sejam necessárias para fornecer os serviços oferecidos. Não compartilharemos informações pessoais com terceiros indiscriminadamente, exceto quando exigido por lei. Excepcionalmente, os dados da pesquisa serão compartilhados com o CEBRAP – Centro Brasileiro de Análise e Planejamento, instituição parceira do Observatório Nacional da Violência contra Educadoras/es para análise descritiva dos dados.

3. Informações Coletadas na Pesquisa

Coletamos os dados que serão listados a seguir, por meio de preenchimento de formulário, utilizando para tal fim de hospedagem os servidores da empresa LimeSurvey GmbH, localizada em Papenreye 63, 22453 Hamburgo, Alemanha.

Os dados coletados serão: nome completo do respondente, local de residência do respondente (estado e município), unidade da federação onde está situada a instituição na qual o respondente atua hoje (estado e município), instituição de ensino onde atua, endereço de e-mail para contato, número de telefone celular (com WhatsApp) com DDD, seu sexo biológico e identificação de gênero, sexualidade, cor ou raça, idade/faixa etária, suas condições de saúde e deficiências físicas/cognitivas caso tenha, sua religião/crença, área e nivel de atuação dentro da docência, situações de violência experimentadas e seus impactos na vida profissional e pessoal, informações sobre a instituição onde trabalha ou trabalhava quando da ocorrência de situações relatadas, e informações de perfil profissional.

Além disso, durante a duração da pesquisa, seu endereço IP será coletado para fins de segurança, auxiliando na identificação de possíveis ataques contra a integridade dos dados. Após o fechamento da pesquisa os endereços IP serão excluídos.

4. Autorização legal para coleta de dados

A Lei Geral de Proteção de Dados Pessoais (LGPD) estabeleceu regras específicas para o tratamento de dados pessoais com finalidade acadêmica. O *Guia Orientativo sobre Tratamento de Dados Pessoais para Fins Acadêmicos*, lançado pela Autoridade Nacional de Proteção de Dados (ANPD), esclarece dúvidas relacionadas a essas regras e visa garantir que o tratamento de dados pessoais, quando associado à produção e disseminação do conhecimento, seja realizado com segurança jurídica e respeito aos direitos dos titulares.

Os pontos centrais do guia versão sobre: os principais pontos abordados pelo guia: Regime Jurídico Especial: A LGPD instituiu um regime jurídico especial mais flexível para o tratamento de dados pessoais com finalidade acadêmica e para a realização de estudos e pesquisas. Esse regime visa equilibrar a proteção de dados pessoais com a liberdade acadêmica e o fluxo de informações necessário para a produção do conhecimento.

O guia esclarece os conceitos relacionados ao tratamento de dados pessoais para fins exclusivamente acadêmicos. Isso inclui a definição e o alcance desse tipo de tratamento. Também aborda os critérios para qualificar uma entidade como "órgão de pesquisa" e as hipóteses legais que autorizam o tratamento de dados pessoais por esses órgãos.

Dentre as hipóteses legais da LGPD encontra-se previsto o tratamento de dados pessoais para realização de estudos por órgãos de pesquisa. Esta hipótese alcança, inclusive, o tratamento de dados pessoais de natureza sensível, independentemente de consentimento pelo titular dos dados.

5. Finalidade dos Dados

A cada função do formulário ou seção que você acessa, solicitamos apenas os dados necessários para as finalidades específicas descritas. Você deve dar seu consentimento expresso ao nos fornecer informações pessoais.

As finalidades específicas dos dados são: nome completo para identificação dos respondentes; e-mail para validação e contato (eventual contato para segunda fase da pesquisa, assim como contato para chamar o respondente a finalizar a resposta).

As demais perguntas têm finalidade estatística, já que serão utilizadas também para a produção de dados amplos sobre o objeto da pesquisa. Esses dados estabelecerão perfis de vítimas de violência que poderão, aí sim, ser contatadas individualmente pela equipe do projeto numa segunda fase da pesquisa para serem convidados para entrevistas individuais.

Além disso, a pesquisa é feita em parceria com o MEC e, portanto, produzirá subsídios para uma potencial política pública de defesa de educadoras e educadores.

Como já dito, a lei permite que o tempo de armazenagem dos dados seja indeterminado, por se tratar de pesquisa acadêmica. Caso você deseje que seus dados sejam excluídos do nosso banco de dados, basta enviar e-mail para: onvecontato@proton.me

6. Uso das Informações

Podemos usar as informações coletadas para envio de comunicações relacionadas a nossa pesquisa, envio de newsletter e afins.

7. Proteção de Dados

Garantimos que todas as informações pessoais que você fornece são armazenadas de forma segura e acessíveis apenas a pessoal autorizado e diretamente vinculado à execução da pesquisa. Empregamos medidas de segurança físicas, eletrônicas e administrativas adequadas para proteger suas informações pessoais contra acesso não autorizado ou divulgação.

Sendo assim, os seguintes protocolos são adotados por recomendação da empresa consultora em segurança da informação e cuidados digitais Mycelium Tecnologia: medidas de proteção física e lógica (PIN / senhas de acesso) aos dispositivos utilizados; softwares antivírus e antimalware; utilização de criptografia de arquivos (criptografia de disco); controles de acesso às contas vinculadas à pesquisa (senhas únicas e fortes, e autenticação em duas etapas).

8. Compartilhamento de Dados

Garantimos que terceiros não terão acesso a seus dados pessoais, exceto conforme indicado nesta Política. Não vendemos ou alugamos suas informações pessoais a terceiros.

Os dados serão acessados dentro do projeto que realiza a pesquisa, e terão acesso compartilhado com o CEBRAP – Centro Brasileiro de Análise e Planejamento, instituição contratada pelo Observatório Nacional da Violência contra Educadoras/es para fazer a análise descritiva dos dados.

9. Cookies utilizados

YII_CSRF_TOKEN: este cookie contém um código aleatório aleatório utilizado para identificar cada sessão de interação com o formulário e impedir fraudes / ataques CSRF. Não armazena quaisquer identificadores pessoais ou informações sensíveis de participantes da pesquisa.

PR-RaMTK, ou LSC-RaMTK: este cookie é utilizado para armazenar variáveis de sessão como o status login na plataforma LimeSurvey. Não armazena identificadores pessoais ou informações sensíveis de participantes da pesquisa.

10. Exatidão e veracidade dos dados coletados

Como usuário, você é o único responsável pela veracidade e alteração dos dados que enviar através do formulário de pesquisa. Cabe única e exclusivamente a você garantir e responder em qualquer situação, a exatidão, vigência e autenticidade dos dados pessoais fornecidos, e se comprometer a manter esses dados devidamente atualizados, para atualização enviar os novos dados para - onvecontato@proton.me. De acordo com o texto desta Política de Privacidade, você concorda em fornecer informação completa e correta no formulário de contato.

11. Cancelamento de inscrição e revogação de autorização

Como titular dos dados que nos foi fornecido, você pode exercer em qualquer momento seus direitos de acesso, retificação, cancelamento e oposição, enviando-nos um e-mail para Fernando Penna, coordenador da pesquisa, no endereço de e-mail onvecontato@proton.me. Resta necessário que você comprove sua identidade anexando uma fotocópia do seu documento de identidade como prova válida.

12. Alterações na Política de Privacidade

Reservamo-nos o direito de atualizar esta Política de Privacidade periodicamente. Recomendamos que você revise esta política toda vez que tiver dúvidas sobre como estamos protegendo suas informações. Lembramos que qualquer mudança sempre será notificada via e-mail.

13. Contato

Caso tenha alguma dúvida ou preocupação relacionada a esta política, entre em contato conosco através do e-mail: onvecontato@proton.me

Esta Política de Privacidade foi atualizada em 10/05/2024 pela Mycelium Tecnologia

Aceitar Fechar

Termo de Consentimento

Para poder começar o preenchimento da pesquisa, é necessário também aceitar o Termo de Consentimento:

O(a) Sr(a) está sendo convidado(a) a participar da pesquisa "A violência contra educadores como ameaça à educação democrática: um estudo sobre a perseguição de educadores no Brasil", do Observatório Nacional da Violência contra Educadoras(es). Sua participação não é obrigatória. A qualquer momento você pode desistir de participar e retirar seu consentimento. Você tem plena autonomia para decidir se participa ou não, bem como retirar sua participação a qualquer momento. Além disso, serão garantidas a confidencialidade e a privacidade das informações prestadas por você. Caso não queira participar, ou desistir, você não será penalizado. Sua participação é muito importante para a execução da pesquisa. Por favor, leia este documento com bastante atenção antes de declarar que concorda em participar. Caso haja alguma palavra ou frase que o(a) senhor(a) não consiga entender, converse com o pesquisador responsável pelo estudo ou com um membro da equipe desta pesquisa para esclarecê-lo(a).

Você responderá agora um questionário que busca analisar o fenômeno da perseguição a educadores(as) a partir das perspectivas dos(as) profissionais da educação do Brasil. A pesquisa será realizada por meio de um questionário online, o qual você acessará na página seguinte a essa, que está hospedado na plataforma LimeSurvey. Se o(a) Sr.(a) aceitar responder o questionário, seus dados serão analisados para entendermos o fenômeno da perseguição e violência contra educadoras(es).

Essa pesquisa via questionário online é constituída por 81 (oitenta e uma) perguntas, relacionadas ao fenômeno da perseguição à educadores(as) no Brasil. Estima-se que você precisará de aproximadamente 15 (quinze) minutos para responder o questionário completo. A precisão de suas respostas é determinante para a qualidade da pesquisa. O questionário estará disponível para ser respondido entre os dias 20 de maio e 21 de agosto de 2024.

A análise dos dados envolve o risco de quebra de confidencialidade (algum dado que possa identificar o(a) Sr.(a) ser exposto publicamente). Para minimizar esse risco, seu nome e qualquer material que indique sua participação serão ocultados ou excluídos e não serão liberados sem a sua permissão. Além disso, a equipe de coordenação da pesquisa implementou uma política de segurança voltada à segurança dos dados. Todo material coletado será armazenado em local seguro. Não compartilharemos informações pessoais com terceiros indiscriminadamente, exceto quando exigido por lei. Excepcionalmente, os dados da pesquisa serão compartilhados com o CEBRAP – Centro Brasileiro de Análise e Planejamento, instituição contratada pelo Observatório Nacional da Violência contra Educadoras/es para consultoria em pesquisa quantitativa e análise descritiva dos dados, que também possui política de segurança própria.

A qualquer momento, durante a pesquisa, ou posteriormente, você poderá solicitar do pesquisador responsável informações sobre sua participação e/ou sobre a pesquisa, o que poderá ser feito através dos meios de contato explicitados neste termo. Somente os resultados gerais da pesquisa serão divulgados em apresentações, relatórios individuais, artigos científicos, e outros materiais com fins científicos ou educacionais.

Contudo, sua participação também pode trazer benefícios. Os possíveis benefícios resultantes da participação na pesquisa são: (i) a criação de mecanismos de acolhimento para esses profissionais; (ii) proteção e fortalecimento social da identidade docente e da autoestima dos(as) educadores(as). Isso não apenas fortalece o profissionalismo dos(as) mesmos(as), mas também ressalta a importância da reconstrução da legitimidade social da escola e dos(as) educadores(as) como central para o desenvolvimento democrático da sociedade. Num contexto de violências e perseguições, é fundamental preservar a integridade e a dignidade destes(as).

Sua participação nesta pesquisa é totalmente voluntária, não implicando em nenhum custo, assim como nenhum tipo de compensação em dinheiro pela sua participação.

Para ter sua via deste TCLE, você poderá imprimi-lo, ou gerar uma cópia em pdf, ou solicitar que seja enviado ao seu e-mail uma versão deste documento. Os contatos do pesquisador responsável estão disponíveis no final desse termo, onde você poderá tirar suas dúvidas sobre o projeto e sua participação, agora ou a qualquer momento. Além disso, este termo possui o telefone e endereço do Comitê de Ética em Pesquisa que autorizou a pesquisa. Qualquer questionamento quanto aos aspectos éticos desta pesquisa favor entrar em contato com o comitê.

Pesquisador Responsável: Fernando de Araujo Penna
Email: onvecontato@proton.me
Pesquisa apoiada pelo MEC

Comitê de Ética em Pesquisa em Ciências Sociais, Sociais Aplicadas, Humanas, Letras, Artes e Linguística (CEP – Humanas)
Endereço: Rua passo da pátria, nº 156 – São Domingos – Niterói
Campus da Praia Vermelha da UFF
Instituto de Física (torre nova – 3º andar)
Telefone: (21) 2629-5119
Email: eticahumanas.comite@id.uff.br

Cópia em PDF / para impressão

Assinale todas as que se aplicam
É obrigatório aceitar o Termo de Consentimento para prosseguir

☐ **Declaro que entendi os objetivos e condições de minha participação na pesquisa e concordo em participar. Recebi um exemplar deste termo de consentimento livre e esclarecido e me foi dada à oportunidade de ler e esclarecer as minhas dúvidas. Concordo em participar da pesquisa: "A violência contra educadores como ameaça à educação democrática: um estudo sobre a perseguição de educadores no Brasil".**

Essa opção, de maneira geral razoável em relação à pesquisa com dados existentes, é controversa na investigação científica que envolve a manipulação de participantes. O **notório experimento** realizado no Facebook, em 2012, com cerca de 700 mil usuários da plataforma com o objetivo de avaliar o "contágio emocional", a partir da alteração do que elas viam em seu feed, é um exemplo. A publicação do artigo provocou **ultraje público** e **críticas no campo científico** (e.g., Flick, 2016), em particular pela inexistência de algum tipo de consentimento das pessoas. Embora a empresa tenha alegado que a pesquisa cumpria suas regras, foi apenas quatro meses depois desse experimento que os **termos de uso do Facebook foram alterados**, passando a incluir a informação sobre a coleta de dados das pessoas que usam o serviço para estudos científicos.

3. A garantia de anonimato e confidencialidade

Por vezes, quem faz a investigação obtém dados sem identificação, anônimos, por exemplo, ao coletar dados a partir de um questionário on-line que não solicita isso, nem quaisquer informações indexicais. Mas é frequente que se saiba quem forneceu os dados. Então, é preciso torná-los anônimos, ou seja, não associados a uma pessoa, para proteger a privacidade de quem colaborou.

No entanto, no contexto digital, isso é também problemático, na medida em que citações literais podem ser rastreadas, por exemplo, quando inseridas em buscadores da internet que apontam suas fontes. Por isso, quando a identificação de quem participou da pesquisa for indesejável, o uso de citações diretas ou informações e dados, como fotos e imagens, que permitam identificar pessoas deve ser evitado. Na etapa de publicação de resultados, para garantir o sigilo de quem deu informações, é comum o uso de pseudônimos ou identificações genéricas.

A confidencialidade dos dados está relacionada também à manipulação e compartilhamento dos dados brutos, por isso quem faz a pesquisa deve estabelecer procedimentos que protejam a identidade dos participantes nessas ocasiões.

Vale notar que pessoas e instituições públicas não requerem, de maneira geral, as mesmas preocupações relacionadas às dimensões de anonimato e confidencialidade na pesquisa científica que indivíduos comuns.

Veremos, a seguir, como as preocupações discutidas têm gerado propostas de quadros de referência reflexivos para ajudar a tomada de decisões éticas durante pesquisas no ambiente digital.

Enquadramentos reflexivos

Quadros de referências ou guias sobre questões éticas podem favorecer opções contextualizadas nas investigações envolvendo a mídia digital. Há algumas propostas convergentes sobre o tema, como as de Williams et al. (2017), Townsend e Wallace (2016) e Fuchs (2018).

A primeira proposta recomenda que, durante a investigação, sejam feitas considerações sobre o modo de publicação de determinado conteúdo (público ou não) e as características do produtor (indivíduo público ou não, vulnerável ou não). A partir daí, quem pesquisa deve decidir se irá solicitar a autorização para o uso do conteúdo na investigação e mesmo se deverá utilizar esse material no trabalho.

O guia de Townsend e Wallace (2016), por sua vez, localiza as questões éticas num contexto amplo, sugerindo a reflexão sobre três dimensões principais:

Aspectos legais	• Os termos e condições da plataforma foram consultados? • As diretrizes da disciplina acadêmica, dos agentes de fomento, legais ou institucionais relevantes foram consultadas?
Privacidade e risco	• Quem utiliza a mídia social pode esperar razoavelmente ser observado por estranhos? • Os participantes da pesquisa são vulneráveis (crianças ou adultos vulneráveis, por exemplo)?
Reuso e publicação	• Quem utiliza as mídias sociais será anonimizado nos resultados publicados? • Será possível publicar ou compartilhar a base de dados?

Em linha com as propostas anteriores, Fuchs (2018) discute particularmente os dilemas éticos relacionados aos estudos qualitativos, recomendando uma **ética de pesquisa on-line crítico-realista**. A discussão do autor, no trabalho, significativamente intitulado “Caro Sr. Neonazista, Você Pode Me Dar Seu

Consentimento Informado para que Eu Possa Citar Seu Tweet Fascista?", exemplifica a proposta.

Desse modo, o dilema exposto é equacionado a partir do argumento de que, quando alguém publica algo usando hashtag, numa mídia social, se engaja numa discussão pública, portanto, sabe que o que publicou será lido por outras pessoas, e até deseja isso, não possuindo expectativa de privacidade. Assim, o uso desse dado, sem pedido de autorização, seria abonado. Ao mesmo tempo, o autor nota que, mesmo utilizando conteúdo produzido por pessoas desconhecidas (não personalidades públicas), anonimiza os dados.

A seguir são mostrados alguns "estudos de caso", retirados do trabalho de Townsend e Wallace (2016), com situações de pesquisa que demandam decisões e suas propostas de encaminhamento ético.

ESTUDOS DE CASO

Estudo de caso #1

Contexto

Alguém deseja estudar narrativas pró-legalização do uso da maconha. Os dados serão coletados a partir do Twitter, portanto são dados públicos abertos. A pesquisa será feita com a coleta de dados, postagens, publicadas com as hashtags *#cannabis*, *#legalize* e *#ismokeit*, durante os últimos 7 dias.

Preocupações

De saída, o assunto é sensível porque se refere a uma atividade ainda ilegal no Reino Unido. Em segundo lugar, pode haver usuários com menos de 18 anos de idade contribuindo para o debate. Por isso, quem faz a pesquisa deve trabalhar de modo a tratar os dados adotando procedimentos de proteção do anonimato.

Encaminhamento ético

Quem pesquisa decide que os dados são públicos, porque são postados no Twitter (plataforma na qual a configuração padrão para postagens é pública); a maioria dos perfis são públicos e podem ser vistos e seguidos por qualquer um. Além disso, o uso de hashtags implica que os usuários estão interessados em contribuir em uma comunidade ou debate e, portanto, esperam um número ainda maior de pessoas vendo seus dados. O tema é sensível, e pode haver dados de menores, assim, há risco considerável de danos. A autoria decide que convém acessar os dados e apresentar resultados a partir de dados agregados, mas não é correto publicar um conjunto de dados (proibido pelo Twitter de qualquer forma) ou republicar citações diretas que levarão pessoas interessadas ao perfil do usuário, comprometendo o anonimato. Quem faz a pesquisa apresentará, portanto, citações parafraseadas (removendo os identificadores) para refletir os temas que surgirem, e fornecerá detalhes sobre como outros poderão recriar os dados da pesquisa. Algumas citações diretas podem ser usadas com o consentimento informado do usuário da plataforma, mas o pesquisador sabe que deve tomar medidas para garantir que o usuário tem mais de 18 anos de idade.

Estudo de caso #2

Contexto
Uma pesquisadora deseja explorar os temas dominantes nas publicações de mídia social de atletas olímpicos em seus perfis de mídia social. Os perfis são públicos e normalmente têm centenas de milhares de seguidores. As plataformas sob escrutínio incluem Twitter e Facebook.

Preocupações
A pesquisadora pode considerar estas postagens públicas, e é ético publicar seus dados textualmente?

Encaminhamento ético
É razoável que a pesquisadora considere estes dados públicos, porque o esportista está publicando em um perfil público com o objetivo de divulgar sua conta de forma mais geral e com a intenção de alcançar o maior número de pessoas possível. Nesse caso, também é razoável que a pesquisadora republique esses dados – o esportista tem uma grande expectativa de que (um grande número de) estranhos estarão vendo seus dados e, de fato, isto é frequentemente desejado. Portanto, é pouco provável que os dados sejam sensíveis. Também dado o tamanho de seu público, é pouco provável que o pesquisador represente um dano potencial ao esportista, além de qualquer risco potencial que ele coloque sobre si mesmo. As citações podem ser republicadas em sua forma original.

Estudo de caso #3

Contexto
Um pesquisador conduz uma análise crítica do discurso de uma base de dados de tweets usando as hashtags *#DonaldTrump*; *#TrumpTrain*; *#VoteTrump2016*; *#AlwaysTrump*; *#MakeAmericaGreatAgain* ou *#Trump2016*. Os tweets são analisados a fim de descobrir como os apoiadores de Trump argumentam em prol de seu candidato no Twitter.

Preocupações
Podemos considerar estes dados públicos? Há algum problema de sensibilidade ou risco de dano? Precisamos buscar o consentimento informado antes de citar estes tweets diretamente?

Encaminhamento ético
Os apoiadores do Trump usam estas hashtags para alcançar um público amplo e convencer outras pessoas a votar em Trump. Portanto, é razoável supor que tais tweets tenham caráter público: os autores esperam e querem ser observados por estranhos a fim de defender um ponto de vista político que eles querem que outros leiam. O pesquisador pode, portanto, citar diretamente tais tweets sem ter que obter o consentimento informado. No entanto, é uma boa prática apagar as IDs dos usuários comuns, que não são eles mesmos figuras públicas.

Comitês de Ética e LGPD

No contexto brasileiro, como já observado, a **Lei nº 14.874**, de 28 de maio de 2024, instituiu o Sistema Nacional de Ética em Pesquisa com Seres Humanos. Os Comitês de Ética em Pesquisa (CEP) possuem papel central nesse Sistema, sendo responsáveis por avaliar projetos. No Parágrafo único de seu Art. 1, a **Resolução CNS Nº 510/2016**, dirigida especificamente às ciências sociais e humanas, informa os casos em que as investigações **não** precisam ser registradas e avaliadas por CEPs:

> "I - pesquisa de opinião pública com participantes não identificados;
> II - pesquisa que utilize informações de acesso público, nos termos da Lei n o 12.527, de 18 de novembro de 2011;
> III - pesquisa que utilize informações de domínio público;
> IV - pesquisa censitária;
> V - pesquisa com bancos de dados, cujas informações são agregadas, sem possibilidade de identificação individual; e
> VI - pesquisa realizada exclusivamente com textos científicos para revisão da literatura científica;
> VII - pesquisa que objetiva o aprofundamento teórico de situações que emergem espontânea e contingencialmente na prática profissional, desde que não revelem dados que possam identificar o sujeito; e
> III (sic) - atividade realizada com o intuito exclusivamente de educação, ensino ou treinamento sem finalidade de pesquisa científica, de alunos de graduação, de curso técnico, ou de profissionais em especialização.
> § 1 o Não se enquadram no inciso antecedente os Trabalhos de Conclusão de Curso, monografias e similares, devendo-se, nestes casos, apresentar o protocolo de pesquisa ao sistema CEP/CONEP;
> § 2 o Caso, durante o planejamento ou a execução da atividade de educação, ensino ou treinamento surja a intenção de incorporação dos resultados dessas atividades em um projeto de pesquisa, dever-se-á, de forma obrigatória, apresentar o protocolo de pesquisa ao sistema CEP/CONEP."

Quem elabora uma proposta de investigação deve ter em mente essas exceções para avaliar se seu projeto de pesquisa envolvendo seres humanos deve ser enviado a um CEP. Ao mesmo tempo, é importante saber que a Lei nº 14.874 estabelece, em seu Art. 14, que a "análise ética de pesquisa, realizada pelo CEP, com emissão do parecer, não poderá ultrapassar o prazo de 30 (trinta) dias úteis da data de aceitação da integralidade dos documentos da pesquisa, e essa aceitação, ou sua negativa, deverá ser feita pelo CEP em até 10 (dez) dias úteis a partir da data de submissão". Um parâmetro como esse deve, portanto, ser levado em conta no planejamento e cronograma de investigações que requeiram avaliação.

Outra legislação relacionada à regulação da pesquisa é a **Lei Geral de Proteção de Dados Pessoais** (LGPD). Devido à **redação ambígua**, há certa controvérsia se ela se aplica diretamente à investigação científica conduzida em instituições sem fins lucrativos. Porém, existe alinhamento claro entre o espírito da Lei e certas regras preconizadas com relação às práticas tradicionais da ética na pesquisa. Além disso, a própria Lei nº 14.847 faz referência à LGPD, em particular, no que diz respeito à proteção e ao anonimato de dados pessoais das pessoas que participam de pesquisas (ver Artigo 69).

Na LGPD, a noção de "tratamento de dados" é definida como:

> "toda operação realizada com dados pessoais, como as que se referem a coleta, produção, recepção, classificação, utilização, acesso, reprodução, transmissão, distribuição, processamento, arquivamento, armazenamento, eliminação, avaliação ou controle da informação, modificação, comunicação, transferência, difusão ou extração" (Art. 5, inciso X).

Tendo em vista a discussão prévia, as preocupações éticas do pesquisador devem estar em todas essas etapas.

Conclusão e desenvolvimentos

BoliviaInteligente (2024), Unsplash

Neste curso, abordamos conceitos relacionados aos dados on-line e à pesquisa digital, bem como estratégias de coleta, tratamento e análise de dados. A natureza introdutória da formação fez com que o conteúdo se limitasse a algumas abordagens, se preocupando em desenvolver um lastro para aprofundamento futuro de quem realizaria o curso.

Foram indicados materiais para avanços posteriores que podem corresponder a uma primeira etapa de novos estudos. No entanto, outras dimensões podem ser sugeridas. Nesse sentido, duas podem ser recomendadas, desde já: uma de teor específico e outra de caráter geral, sendo que ambas podem colaborar para tornar alguém mais competente na pesquisa envolvendo dados on-line.

A geral é a contínua leitura crítica de pesquisas e discussões que envolvam dados e métodos digitais, pois isso favorece o desenvolvimento do senso crítico e científico quanto aos dados. As revistas científicas de comunicação apresentam, quase sempre a cada edição, trabalhos assim, e os livros organizados por Silva e Stabile (2016), Silva et al. (2018) reúnem muitos estudos e discussões metodológicas interessantes.

Outras leituras válidas, nessa perspectiva, são o trabalho de Lopes e Freire (2015), discutindo implicações de ferramentas, métricas e monitoramento de conteúdos produzidos por fãs em redes sociais e, no âmbito dos métodos digitais, o de Rogers (2019), abordando as possibilidades de uso de métricas alternativas para estudar o engajamento social em questões problemáticas, no que chamou de “análise crítica”.

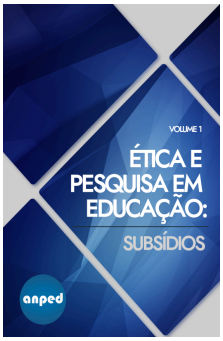
A sugestão mais restrita está ligada ao fortalecimento de uma atitude reflexiva frente aos instrumentos técnicos de coleta, organização e análise de dados. Como argumentam van Es et al. (2021), as ferramentas fazem um trabalho epistêmico, de modo que suas premissas e adequação para fins de pesquisa precisam ser avaliadas criticamente. Isso envolve, entre outros pontos, a busca por conhecimento sobre as características, as limitações e o papel que elas exercem no

desenvolvimento da pesquisa. É preciso evitar uma atitude acrítica e ingênua sobre a influência que os instrumentos exercem sobre os dados obtidos a partir deles e sobre o processo de investigação como um todo.

Um provável desenvolvimento das ferramentas, por sinal, deverá ser a associação com softwares de inteligência artificial, o que já foi vislumbrado, embora de maneira tímida, em determinados momentos desta formação. Mas essa é uma área que deve crescer e que exigirá também reflexividade na adoção por parte de quem faz pesquisas.

REAs de aprofundamento

Materiais para estudos após o curso - Módulo 6



Desde 2019, a Associação Nacional de Pós-Graduação e Pesquisa em Educação (ANPEd) tem publicado **e-books abertos** abordando várias dimensões da ética na pesquisa. Embora os trabalhos privilegiem a área da Educação, são úteis também para pesquisadores de outras disciplinas. Já o **código elaborado pela Fundação de Amparo à Pesquisa do Estado de São Paulo** (Fapesp) apresenta discussões importantes sobre o tema.

O **vídeo do site Poder 360** explica o que é Lei Geral de Proteção de Dados, enquanto o **tutorial da professora Camila Mendes** demonstra como configurar um Formulários Google, de modo a que o uso dele como instrumento de pesquisa científica esteja de acordo com as recomendações da LGPD para a coleta de dados.

Tendo chegado ao fim do curso, numa trajetória de aprofundamento, você poderá apreciar as "receitas" de pesquisa com métodos digitais, de certa complexidade, conforme as propostas feitas pelo **Public Data Lab** e por ele em parceria com outras organizações, na série "**Digital Investigation Recipes**", com **vídeo** como o mostrado ao lado.

Atividade - Revisão geral

O curso do qual este material é proveniente contém várias atividades, no entanto, elas fazem mais sentido no contexto da formação. Este questionário, porém, talvez seja válido para as pessoas que estudaram o conteúdo. As respostas estão no fim do documento.

1) A proposta de "seguir o método do meio" caracteriza:

a) o entendimento restrito sobre métodos digitais
b) o entendimento amplo sobre métodos digitais
c) o entendimento geral sobre métodos digitais
d) Nenhuma das alternativas

2) Os programas para a análise de dados quantitativos e qualitativos começaram a ser utilizados, respectivamente, nas décadas de:

a) 1970 e 1990
b) 1960 e 1980
c) 1950 e 1970
d) 1980 e 1990

3) O escândalo Cambridge Analytica está ligado à qual plataforma:

a) Google
b) Facebook
c) Twitter/X
d) TikTok

4) Publicado em 2014, o artigo "Experimental Evidence of Massive-Scale Emotional Contagion Through Social Networks" (https://www.pnas.org/doi/10.1073/pnas.1320040111) ganhou notoriedade pela forma polêmica de produção dos dados da investigação. Em parceria com a rede social Facebook, o grupo que conduziu o trabalho manipulou o Feed de Notícias de grande quantidade de usuários para perceber se havia mudança no estado emocional deles. Nesse sentido, é possível dizer que a pesquisa possui uma abordagem ... , com o uso de dados ... .

a) experimental – existentes
b) quantitativa – existentes
c) experimental – criados
d) quantitativa – criados

5) Quando utilizados em pesquisas de outras pessoas, os dados abertos são:

a) Dados primários
b) Dados secundários
c) Dados terciários
d) Nenhuma das alternativas

6) Os IDs de usuários de serviços digitais e o número de vídeos de canais do YouTube são dados, respectivamente:

a) Atributivos e metadados
b) Metadados e indexicais
c) Indexicais e estruturados
d) Indexicais e atributivos

7) No artigo "Fascism 2.0: Twitter Users' Social Media Memories of Hitler on his 127th Birthday" (https://doi.org/10.1163/22116257-00602004), Christian Fuchs construiu uma base de dados com 4.193 tweets publicados em 20 de abril de 2016, utilizando hashtags como #Hitler, #AdolfHitler, #HappyBirthdayAdolf, #HappyBirthdayHitler. A preocupação do autor era discutir, a partir da perspectiva da teoria crítica, o modo como o fascismo se atualiza nos dias de hoje. Previamente, o autor identificou hashtags relevantes, com uma ferramenta on-line, antes de coletar as postagens – o que foi feito com o uso de outra ferramenta digital por meio da qual era acessada a ... do Twitter (atual X).

a) API
b) conta de Elon Munsk
c) estrutura de perfis
d) busca

8) As técnicas de coleta de dados digitais a partir de "scraping" e "crawling" caracterizam-se, respectivamente, por:

a) Extração de URLs e coleta de informações de páginas web a partir do código-fonte
b) Dados do código-fonte e dados abertos
c) Extração de URLs e de dados do código-fonte
d) Coleta de informações de páginas web a partir do código-fonte e extração de URLs

9) Em "Collective Memory and Social Media: Fostering a New Historical Consciousness in the Digital Age?" (https://doi.org/10.1177/1750698017750012), Thomas Birkner e André DonkView investigam a polêmica em torno da proposta de mudança no nome de uma praça numa cidade alemã, à luz de descobertas históricas que associavam a figura homenageada – um ex-presidente da República de Weimar – à ascensão do nazismo. Com interesse nos tópicos e na qualidade do discurso produzido por uma esfera conservadora de contrapúblico, efetuam a análise de postagens de uma página do Facebook. Compõem então uma amostra, referente ao período ligado aos interesses da pesquisa, de 231 postagens com 1733 comentários.

Os autores, entretanto, não apontam nenhuma ferramenta específica de coleta de dados. Isso é uma indicação de quê?

a) A seção de métodos do artigo falhou em não apontar esse aspecto
b) Dependendo da proposta da pesquisa é possível prescindir de ferramentas específicas para a coleta de dados
c) Um assistente coletou os dados
d) Nenhuma das alternativas

10) As coletas de dados digitais a partir de APIs e por meio de "raspagem" geralmente estão associadas a:

a) Dados não estruturados e semiestruturados
b) Dados estruturados e não estruturados
c) Dados estruturados e semiestruturados
d) Dados semiestruturados e estruturados

11) Valerie Lookingbill e Kimanh Le analisam, em "There's Always a Way to Get Around the Guidelines: Nonsuicidal Self-Injury and Content Moderation on TikTok" (https://doi.org/10.1177/2056305124125437) como os usuários com histórico de automutilação não suicida (ANS) utilizam o algoritmo do TikTok para se envolver com conteúdo voltado a esse tema. Para tanto, entrevistam oito usuários do TikTok e realizam uma análise de conteúdo de 150 vídeos da plataforma. Esses vídeos foram localizados a partir do uso de hashtags, previamente vistas como importantes no âmbito da ANS, inseridas na ferramenta de busca "Discover" da plataforma. Esta estratégia aponta para a possibilidade de que uma investigação?

a) Em determinadas circunstâncias, utilizar ferramentas de busca internas a uma plataforma.
b) Faça uso de amostras de dados on-line aleatórios
c) Utilize uma abordagem de coleta de dados via API
d) Nenhuma das alternativas

12) NÃO é uma das regras para a construção de tabelas no formato tidy:

a) Cada variável deve ter sua própria linha
b) Cada observação deve ter sua própria linha
c) Cada valor deve ter sua própria célula
d) Cada variável deve ter sua própria coluna

13) O uso de visualizações no trabalho científico geralmente possui objetivo:

a) Explicativo
b) Exploratório
c) Expositivo
d) Nenhuma das alternativas

14) A noção de "densidade de dados" remete à:

a) Quantidade de variáveis numa tabela
b) Quantidade de elementos inseridos em um gráfico
c) Quantidade de nós em um grafo
d) Nenhuma das alternativas

15) Os infográficos e as visualizações de dados diferem, pois:

a) Os infográficos são mais simples do que as visualizações
b) As visualizações possuem menos edição do que os infográficos
c) Os infográficos procuram apresentar dados brutos com um mínimo de edição
d) As visualizações procuram elaborar narrativas, ao contrário dos infográficos

16) Os gráficos de pizza e de linha têm como ponto forte, respectivamente, destacar:

a) Tendências temporais e hierarquias
b) Hierarquias e conexões
c) Hierarquias e tendências temporais
d) Distribuições e hierarquias

17) A transformação de textos em dados quantitativos envolve, num primeiro momento:

a) Extrair características de um texto
b) Contar as letras de um texto
c) Contar as palavras de um texto
d) Extrair tokens desnecessários

18) A noção de "clusterização" está relacionada a qual característica dos grafos:

a) Tamanho dos nós
b) Direcionamento das arestas
c) Modularidade
d) Centralidade de grau

19) As redes elaboradas a partir de dois conjuntos diferentes de atores são chamadas de:

a) Egocêntricas
b) De dois modos
c) Direcionadas
d) Bidirecionais

20) A Resolução CNS Nº 510/2016 indica casos em que as investigações não precisam ser registradas e avaliadas por CEPs. Qual situação NÃO está nesse âmbito:

a) Pesquisa que utilize informações de domínio público
b) Pesquisa com grupos vulneráveis
c) Pesquisa censitária
d) Pesquisa com bancos de dados, cujas informações agregadas, sem possibilidade de identificação individual

Referências

Módulo 1

Castells, M. (2010). **A sociedade em rede: A era da informação: economia, sociedade e cultura** (v. 1). Paz e Terra. Obra original publicada em 1996.

Dawson, C. (2020). **A-Z of digital research methods**. Routledge.

Fuchs, C. (2019). What is critical digital social research? Five reflections on the study of digital society. **Journal of Digital Social Research** , **1**(1), 10-16. https://doi.org/10.33621/jdsr.v1i1.7

Hooley, T., Marriott, J., & Wellens, J. (2012). **What is online research? Using the internet for social science research**. Bloomsbury Academic. http://doi.org/10.5040/9781849665544

Kozinets, R. V. (2014). **Netnografia: Realizando pesquisa etnográfica online**. Penso. Obra original publicada em 2009.

Nascimento, L. F. (2020). **Sociologia digital: Uma breve introdução**. EDUFBA. https://repositorio.ufba.br/bitstream/ri/32746/5/SociologiaDigitalPDF.pdf

Rogers, R. (2013). **Digital methods**. The MIT Press. https://doi.org/10.7551/mitpress/8718.001.0001

Snee, H., Hine, C., Morey, Y., Roberts, S., & Watson, H. (Eds.). (2016). **Digital methods for social science: An interdisciplinary guide to research innovation**. Palgrave Macmillan. https://doi.org/10.1057/9781137453662

Witte, J. C. (2012). A ciência social digitalizada: Avanços, oportunidades e desafios. **Sociologias**, **14**(31), 52-92. https://doi.org/10.1590/S1517-45222012000300004

Módulo 2

Couldry, N., & Mejias, U. A. (2019). **The costs of connection: How data is colonizing human life and appropriating it for capitalism**. Stanford University Press.

De Bruyne, P., Herman, J., & De Schoutheete, M. (1991). **Dinâmica da pesquisa em ciências sociais: Os polos da prática metodológica** (5a ed.). Francisco Alves. Obra original publicada em 1976.

Fuchs, C. (2019). What is critical digital social research? Five reflections on the study of digital society. **Journal of Digital Social Research** , **1**(1), 10-16. https://doi.org/10.33621/jdsr.v1i1.7

Hellmann, F. , & Homedes, N. (2022). Uma pesquisa clínica não ética e a politização da pandemia da COVID-19 no Brasil: O caso da Prevent Senior. **Developing World Bioethics**, 1-14. https://doi.org/10.1111%2Fdewb.12369

Jones-Rooy, A. (2019, 24 de julho). I'm a data scientist who is skeptical about data. **Quartz**. https://qz.com/1664575/is-data-science-legit

Kitchin, R. (2014). **The data revolution: Big data, ppen data, data infrastructures & their consequences**. Sage. https://doi.org/10.4135/9781473909472

Laville, C., & Dionne, J. (1999). **A construção do saber: Manual de metodologia da pesquisa em ciências humanas**. Editora UFMG.

Rogers, R. (2015). Digital methods for web research. In R. A. Scott & S. M. Kossiyn (Eds.), **Emerging trends in the social and behavioral sciences** (pp. 1-22). John Wiley & Sons, Inc. https://doi.org/10.1002/9781118900772.etrds0076

Spagnuolo, S. (2022, 23 de agosto). Para analisar as redes, ignore bots e olhe engajamento. **Núcleo**. https://nucleo.jor.br/linhafina/2022-08-23-ignore-bots-olhe-engajamento/

Venturini, T., & Latour, B. (2019). O tecido social: Rastros digitais e métodos quali-quantitativos. In J. J. Omena (Ed.), **Métodos digitais: Teoria-prática-crítica** (pp. 37-46). ICNOVA – Instituto de Comunicação da Nova. https://colecaoicnova.fcsh.unl.pt/index.php/icnova/issue/view/22

Witte, J. C. (2012). A ciência social digitalizada: Avanços, oportunidades e desafios. **Sociologias**,**14**(31), 52-92. https://doi.org/10.1590/S1517-45222012000300004

Módulo 3

Bounegru, L., & Gray, J. (Eds.). (2021). **The data journalism handbook: Towards a critical data practice**. Amsterdam University Press. http://doi.org/10.5117/9789462989511

Burgess, J., & Bruns, A. (2018). Abordagens e métodos para o estudo das mídias sociais na comunicação política. **Aurora: revista de arte, mídia e política**,**10**(30), 129-146. https://revistas.pucsp.br/index.php/aurora/article/view/35869

da Costa, A. B. F. (2020). **Fluxo do trabalho com dados: Do zero à prática**. Open Knowledge Brasil. https://escoladedados.org/wp-content/uploads/2021/03/livrov2.pdf

Jünger, J., & Keyling, T. (2019). **Facepager. An application for automated data retrieval on the web**. https://github.com/strohne/Facepager/

Kitchin, R. (2014). **The data revolution: Big data, open data, data infrastructures & their consequences**. Sage. https://doi.org/10.4135/9781473909472

Litman, L., & Robinson, J. (2020). **Conducting online research on Amazon Mechanical Turk and beyond**. Sage. https://doi.org/10.4135/9781506391151

Monaco, N., & Amaudo, D. (2020). **Análise de dados para o monitoramento de redes sociais**. NDI. https://bit.ly/3X9GuUy

Peeters, S. & Borra, E. (2020). **Capturing data: Scraping and formatting data from web sites**. Department of Media Studies, University of Amsterdam. https://tinyurl.com/y3pjcm2t

Peeters, S., & Hagen, S. (2022). The 4CAT Capture and Analysis Toolkit: A modular tool for transparent and traceable social media research. **Computacional Communication Research**, **4**(2), 571-589. https://doi.org/10.5117/CCr2022.2.007.hAgE

Rogers, R. (2013). **Digital methods**. The MIT Press. https://doi.org/10.7551/mitpress/8718.001.0001

Samuel, A. (2018, 15 de maio). Amazon's Mechanical Turk has reinvented research. **JSTOR Daily**. https://daily.jstor.org/amazons-mechanical-turk-has-reinvented-research/

van Es, K.; López Coombs, N.; & Boeschoten, T. (2017). Towards a reflexive digital data analysis. In K. van Es & M. Schäfer (Eds.), **The datafied society: Studying culture through data** (pp. 171-180). Amsterdam University Press. http://doi.org/10.5117/9789462981362

Módulo 4

Cassel, P. E., & Peterossi, H. G. (2020, 11 a 12 de novembro). **Considerações sobre o impacto da Lei Geral de Proteção de Dados na Pesquisa** [artigo de evento]. XV Simpósio dos Programas de Mestrado Profissional. São Paulo. https://bit.ly/3XMhfJk

Costa, D. (2021, 20 de agosto). Organizando banco de dados: Uma introdução ao conceito de Tidy Data. **Datapsico**. https://bit.ly/4eaojEK

Goedhart, J. (2017, 6 de outubro). Converting excellent spreadsheets to tidy data. **The Node**. https://bit.ly/4efbnyc

Hall, D. (2024). Data cleaning: Definition, techniques & best practices for 2024. **Technology Advice**. https://technologyadvice.com/blog/information-technology/data-cleaning/

MacDonald, L. (2024). 7 essential data cleaning best practices. **Monte Carlo**. https://www.montecarlodata.com/blog-data-cleaning-best-practices/

van der Vlist, F., & Helmond, A. (2023a). **Reference worksheet I: Data management**. Department of Media Studies, University of Amsterdam. http://bit.ly/msrw-1

Wickham, H., Çetinkaya-Rundel, M., & Grolemund, G. (2017). **R for Data Science: Import, tidy, transform, visualize and model data** (2a ed.). R4DS. https://r4ds.hadley.nz/

Módulo 5

Allen, W. (2017). Making corpus data visible: Visualising text with research intermediaries. **Corpora**, **12**(3), 459-482. https://doi.org/10.3366/cor.2017.0128ope

Anthony, L. (2018). Visualisation in corpus-based discourse studies. In C. Taylor & A. Marchi (Eds.), **Corpus approaches to discourse: A critical review** (pp. 197-223). Routledge.

Balestrini, D. P., Stoeger, H. & Ziegler, A. (2023). Quantitative text analysis in gifted and talented research. **High Ability Studies**, **34**(2), p. 189-228. https://doi.org/10.1080/13598139.2023.2167812

Benoit, K. (2020). Text as data: An overview. In L. Curini & R. Franzese (Eds.), **Handbook of research methods in political science and international relations**. (pp. 461-497). Sage. https://kenbenoit.net/pdfs/CURINI_FRANZESE_Ch26.pdf

Grimmer, J., Roberts, M. E, Stewart, B. M. (2022). **Text as data: A new framework for machine learning and the social sciences**. Princeton University Press.

Kennedy, H., & Allen, W. (2017). Data visualisation as an emerging tool for online research (2a ed.). In N. G. Fielding, R. M. Lee & G. Blank (Eds.), **The SAGE handbook of online research methods** (pp. 307-326). Sage.

Kirk, A. (2019). **Data visualisation: A handbook for data driven design** (2a ed.). Sage.

Izumi, M., & Moreira, D. (2018). O texto como dado: Desafios e oportunidades para as ciências sociais. **BIB - Revista Brasileira de Informação Bibliográfica em Ciências Sociais**, (86), 138-174. https://bibanpocs.emnuvens.com.br/revista/article/view/455

Quan-Haase, A., Foisey, L., & McLaughlin, R. (2024) Social media and digital network. In J. McLevey, J. Scott & P. J. Carrington (Eds.), **The Sage handbook of social network analysis** (2a ed.) (pp. 297-308). Sage.

McEnery, T. and Hardie, A. (2012). **Corpus Linguistics: Method, theory and practice**. Cambridge University Press.

Pérez-Paredes, P., & Curry, N. (2024). Epistemologies of corpus linguistics across disciplines. **Research Methods in Applied Linguistics**, **3**(3). https://doi.org/10.1016/j.rmal.2024.100141

Recuero, R. (2017). **Introdução à análise de redes sociais**. EDUFBA. https://repositorio.ufba.br/bitstream/ri/24759/4/AnaliseDeRedesPDF.pdf

Recuero, R. (2018). Estudando discursos em mídia social: Uma proposta metodológica. In T. Silva, J. Buckstegge & P. Rogedo (Orgs.), **Estudando cultura e comunicação com mídias sociais** (pp. 13-30). IBPAD. https://bit.ly/40Sdhke

Scott, J. (2012). **What is social network analysis?** Bloomsbury Academic. http://doi.org/10.5040/9781849668187

Scott, J., McLevey, J., & Carrington, P. J. (2024). Introduction. In J. McLevey, J. Scott & P. J. Carrington (Eds.), **The Sage handbook of social network analysis** (2a ed.) (pp. 1-16). Sage.

Sosulski, K. (2019). **Data visualization made simple: Insights into becoming visual**. Routledge.

Tomaél, M. I., & Marteleto, R. M. (2013). Redes sociais de dois modos: Aspectos conceituais. **Transinformação**, **25**(3), 245-253. https://www.scielo.br/j/tinf/a/L7QwLS5RZ5JwffJ5Bxrzc4v/

Tufte, E. R. (2007). **The visual display of quantitative information** (2 ed.). Graphics Press. Obra original publicada em 1983.

van der Vlist, F., & Helmond, A. (2023b). **Reference worksheet III: Data visualisation**. Department of Media Studies, University of Amsterdam. http://bit.ly/msrw-3

van Es, K., & Schäfer, M. (Eds.). (2017). **The datafied society: Studying culture through data**. Amsterdam University Press. http://doi.org/10.5117/9789462981362

van Es, K., & Verhoeff, N. (Eds.). (2023). **Situating data: Inquiries in algorithmic culture**. Amsterdam University Press. http://doi.org/10.5117/9789463722971

Venturini, T., Munkl, A., & Jacomy, M. (2018). Ator-rede *versus* Análise de Redes *versus* Redes Digitais: Falamos das mesmas redes. **Galáxia**, (38), 5-27. http://dx.doi.org/10.1590/1982-2554236645

Módulo 6

de La Taille, Y. (2008). Ética em pesquisa com seres humanos: Dignidade e liberdade. In I. C. Z. Guerriero, M. L. S. Schmidt & F. Zicker (Orgs.), **Ética nas pesquisas em ciências humanas e sociais na saúde** (pp. 268-279). Aderaldo & Rothschild. https://bit.ly/4hIODIS

Flick, C. (2016). Informed consent and the Facebook emotional manipulation study. **Research Ethics**, **12**(1), 14-28. https://doi.org/10.1177/1747016115599568

Fuchs, C. (2018). "Dear Mr. Neo-Nazi, can you please give me your informed consent so that I can quote Your Fascist tweet?": Questions of social media research ethics in online ideology critique. In G. Meikle (Ed.), **The Routledge companion to media and activism** (pp. 385-394). Routledge. https://bit.ly/3VBcoZh

Lopes, M. I. V., & Freire, C. (2015). A dimensão epistemológica do monitoramento on-line: Para um estudo crítico das técnicas de pesquisa na internet. In M. Ledo Andión & M. I. V. Lopes. (Orgs.), **Comunicación, cultura e esferas de pode***r* (pp. 229-251). USP-ECA/USC-GEA/AssiBERCOM/AGACOM. https://www.eca.usp.br/acervo/producao-academica/002798786.pdf

Markham, A., & Buchanan, E. (2017). Research ethics in context: Decision-making in digital research. In K. van Es & M. Schäfer (Eds.), **The datafied society: Studying culture through data** (pp. 171-180). Amsterdam University Press. http://doi.org/10.5117/9789462981362

Nunes, J. B. C. (2019). Pesquisas online. In Comissão de Ética em Pesquisa da ANPEd (Org.), **Ética e pesquisa em Educação: Subsídios (vol. 1)** (pp. 146-154). ANPEd. https://anped.org.br/wp-content/uploads/2024/05/eticaANPED.pdf

Rogers, R. (2019). Engajados de outra maneira: As mídias sociais — das métricas de vaidade à análise crítica. In J. J. Omena (Ed.), **Métodos digitais: Teoria-prática-crítica** (pp. 73-96). ICNOVA – Instituto de Comunicação da Nova. https://colecaoicnova.fcsh.unl.pt/index.php/icnova/issue/view/22

Silva, T., Buckstegge, J., & Rogedo, P. (Orgs.). (2018). **Estudando cultura e comunicação com mídias sociais**. IBPAD. https://bit.ly/40Sdhke

Silva, T., & Stabile, M. (Orgs.). (2016). **Monitoramento e pesquisa em mídias sociais: Metodologias, aplicações e inovações**. Uva Limão. https://bit.ly/3O9QSaL

Townsend, L., & Wallace, C. (2016). **Social media research: A guide to ethics**. The University of Aberdeen. https://www.gla.ac.uk/media/Media_487729_smxx.pdf

Williams, M. L., Burnap, P., & Sloan, L. (2017). Towards an ethical framework for publishing Twitter data in social research: Taking into account users' views, online context and algorithmic estimation. **Sociology**, **51**(6), 1149-1168. https://doi.org/10.1177/0038038517708140

Créditos, agradecimentos, licença e citação

Conteudista e produtor: Richard Romancini

Revisor de texto: XXXX XXXXXX

Câmera e som nos vídeos *live action*: Mário Rocha (Labidecom)

Foto da capa: chaylek, Vecteezy

Cabe agradecer aos estudantes da turma de graduação do curso **CCA0305 - Procedimentos Educomunicativos em Educação a Distância II** (2024), da Licenciatura em Educomunicação da ECA/USP, que analisaram a versão prévia deste curso, fazendo observações para ajustes, bem como aos participantes da disciplina de pós-graduação **CCA5962 - Movimentos Sociais, Comunicação e Educação** (2024), do PPGCOM/USP, que foram cursistas da turma piloto, também realizando comentários críticos.

Forma de citação deste material:

ABNT ROMANCINI, Richard. *Introdução à análise de dados on-line*. São Paulo: Cecom, Labidecom-ECA/USP, 2024.

APA Romancini, R. (2024). *Introdução à análise de dados on-line*. Cecom, Labidecom-ECA/USP.

Respostas da Atividade de Revisão: 1-a, 2-b, 3-b, 4-c, 5-b, 6-d, 7-a, 8-d, 9-b, 10-c, 11-a, 12-a, 13-a, 14-b, 15-b, 16-c, 17-a, 18-c, 19-b, 20-b