

# Pesquisa FAPESP

MAIO DE 2017 | ANO 18, N. 255

WWW.REVISTAPESQUISA.FAPESP.BR

86% das decisões  
do STF são de  
recursos internos  
ao tribunal

LITERATURA

Inventário  
encontra

inesperada

DEMOGRAFIA

diversidade de

espécies

de animais

e plantas em

São Paulo

# HUMANIDADES DIGITAIS

Como classificar

os dinossauros

e dividir os

paleontólogos

ENTREVISTA

Neurocientista

Roberto Lent

propõe usar

o conhecimento

científico para

melhorar o ensino

ÉTICA

GRANDES BANCOS DE DADOS E ACERVOS DIGITALIZADOS

AMPLIAM A CAPACIDADE DE ANÁLISE DAS CIÊNCIAS

HUMANAS E SOCIAIS COM AUXÍLIO DA COMPUTAÇÃO

Há 100 anos

Comissão Rondon

fazia o primeiro

documentário

etnográfico

Biometria inclui

identificação

de voz, rosto e

vasos sanguíneos

LINGÜÍSTICA

ARQUEOLOGIA

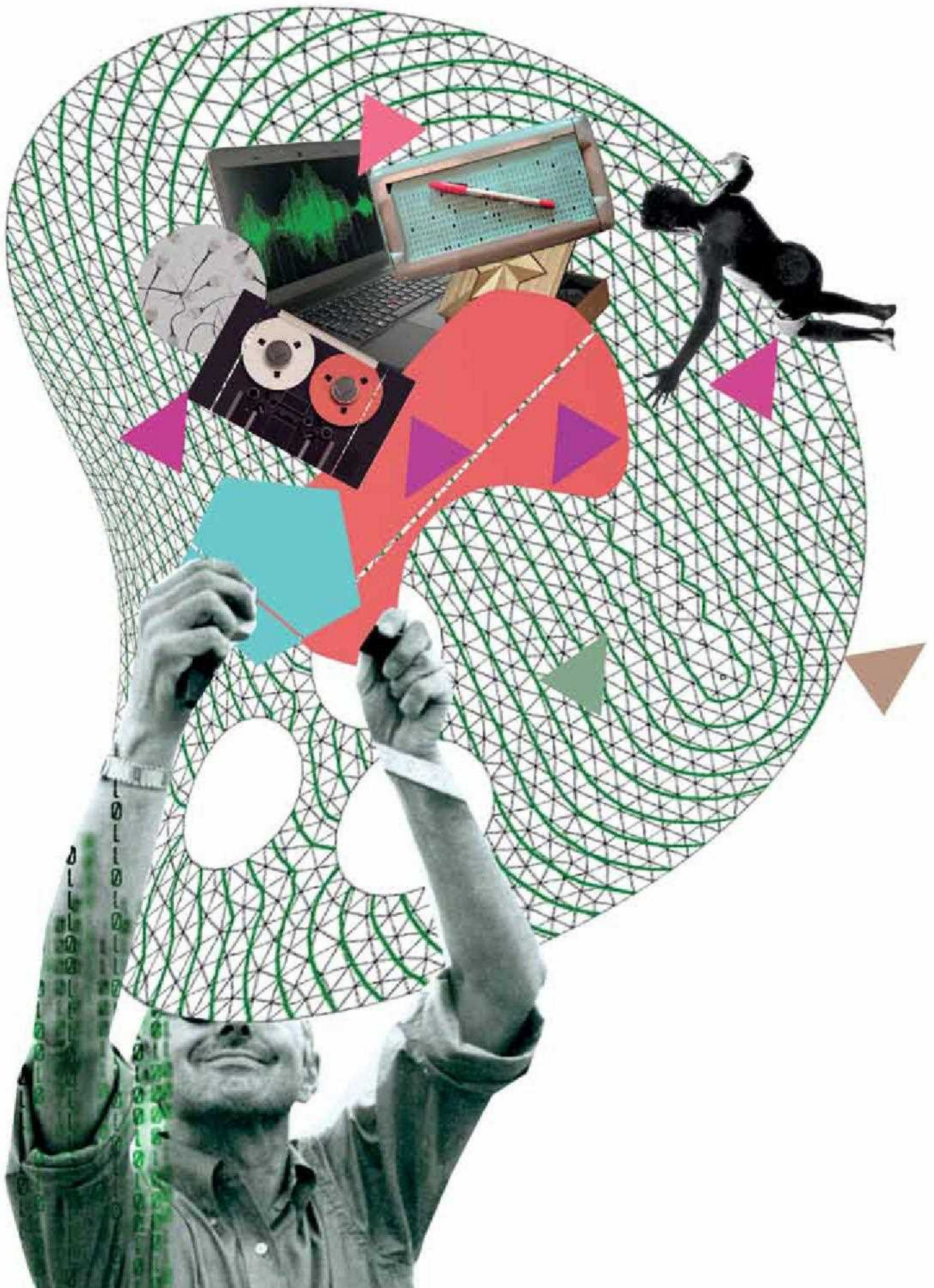
HISTÓRIA

CIDADES

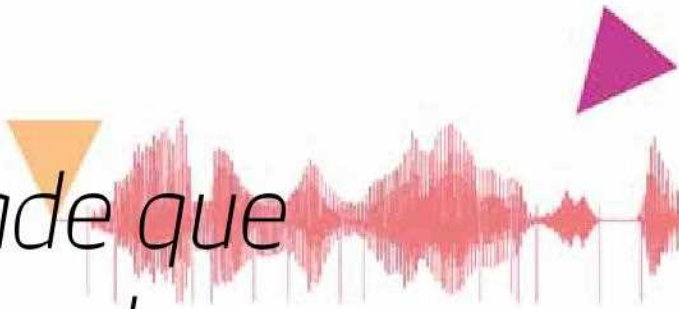
> OLFATO MUDA EM RESPOSTA A ESTÍMULOS DO AMBIENTE AO LONGO DA VIDA



## CAPA







# *A realidade que emerge da* **avalanche de dados**

Humanidades digitais se disseminam por várias  
disciplinas, influenciam formação de  
pesquisadores e inspiram políticas públicas

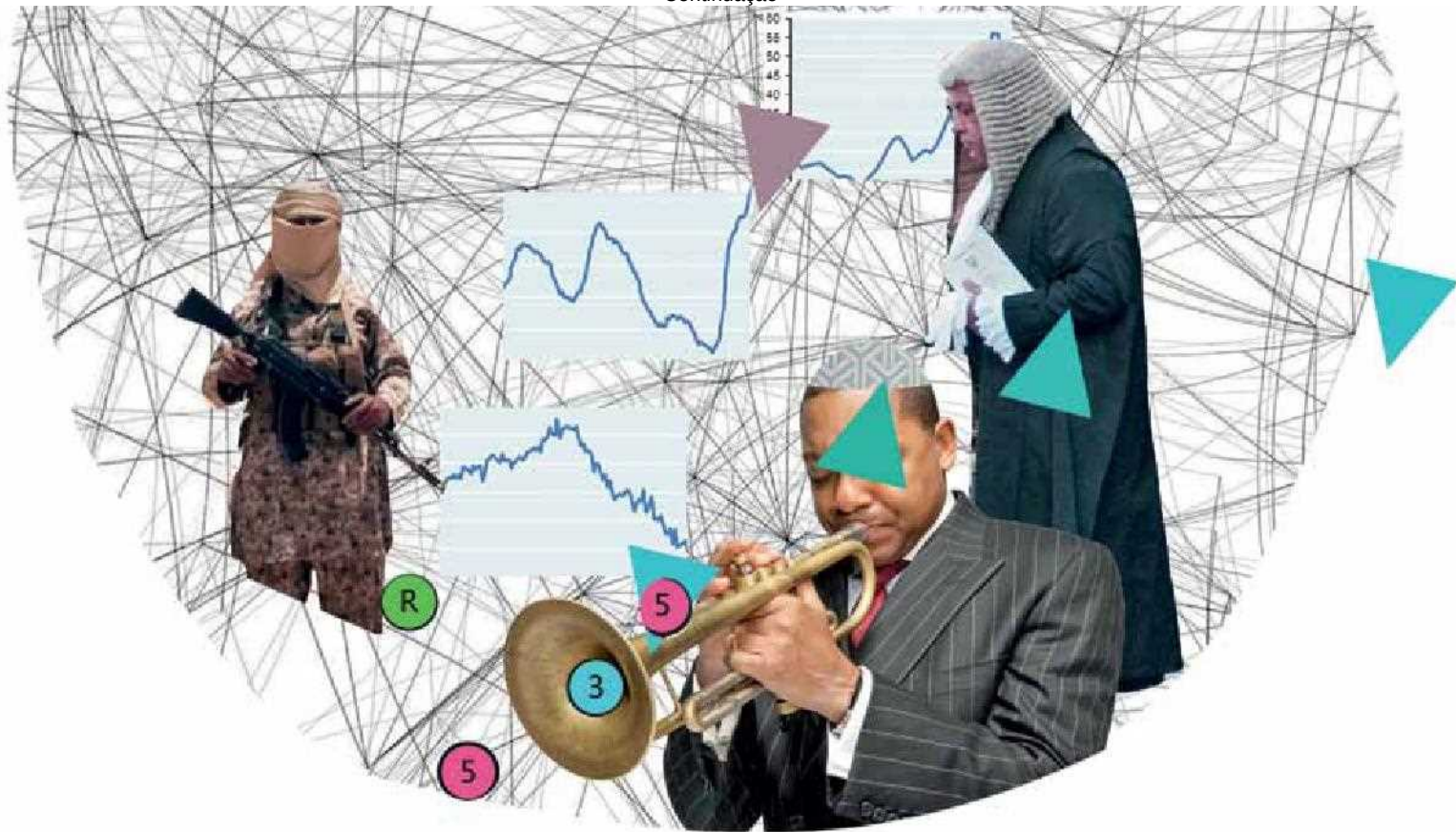
**Fabício Marques**

**C**omputadores são uma ferramenta de trabalho para pesquisadores de todas as áreas do conhecimento, mas, no caso da comunidade das ciências humanas e sociais, a digitalização de acervos artísticos e históricos e a oferta de gigantescos bancos de dados com informações econômicas e sociais abriram novas frentes de observação de fenômenos e de análise de tendências. Houve uma natural aproximação com os cientistas da computação, cujas pesquisas em Big Data multiplicaram as formas de organizar e analisar informações, dando origem a um campo interdisciplinar: as humanidades digitais. “O termo foi cunhado para definir a pesquisa que incorpora a tecnologia computacional a estudos em humanidades, mas também aquela que usa as humanidades para estudar a tecnologia digital e sua influência na sociedade e na cultura”, explica Brett Bobley, diretor do Escritório de Humanidades Digitais da National Endowment

for the Humanities (NEH), agência de fomento do governo norte-americano. Não se trata, segundo ele, de uma nova área do conhecimento, mas de uma gama de atividades que pode abranger o uso de fotografias aéreas por arqueólogos para escanear sítios, o desenvolvimento de técnicas de análise de dados que ajudam linguistas a estudar jornais antigos, o estudo da ética da tecnologia por filósofos, entre outros exemplos.

Um dos projetos financiados pela NEH em humanidades digitais resgatou os diários de campo do explorador britânico David Livingstone (1813-1873). Relatos de sua viagem à África Central em 1871 foram escritos em jornais velhos, por falta de papel disponível. Com o tempo, a tinta esmaeceu e ficaram ilegíveis os textos em que Livingstone registrou suas impressões sobre a dinâmica do comércio de escravos, entre outras observações. Entre 2013 e 2017, um grupo de pesquisadores de humanidades e de ciências da computação dos Estados Unidos e do Reino Unido conseguiu res-





gatar os escritos utilizando técnicas de fotografia de imagem espectral, capazes de recuperar informações que não são visíveis ao olho humano.

Outro exemplo foi a colaboração de historiadores de várias partes do mundo para organizar registros de cerca de 36 mil viagens de navios negreiros entre 1514 e 1866, que levaram mais de 12 milhões de escravos da África. O esforço, iniciado nos anos 1990 pelo historiador norte-americano David Eltis, da Universidade Emory, resultou no Banco de Dados sobre o Comércio de Escravos, disponível na internet desde 2007 no endereço [slavevoyages.org](http://slavevoyages.org). A análise dos dados, que reúne registros em vários idiomas e abrange a movimentação nos portos por onde passaram os navios, apontou para os historiadores novas dimensões sobre como os africanos experimentaram e resistiram à deportação e à escravização e revelaram novas conexões transatlânticas no comércio de escravos.

Um primeiro levantamento foi lançado em 1999 sob a forma de CD-ROM, mas o esforço colaborativo para obter dados sobre as viagens conseguiu traçar posteriormente um retrato mais amplo do comércio de escravos. Na primeira fase, estimou-se que o Brasil havia recebido cerca de 3,6 milhões de escravos, mas documentos mostraram que esse contingente chegou a 5 milhões – num total de 10,7 milhões de africanos deportados para as Américas. A iniciativa produziu impactos diversos nas pesquisas sobre escravidão, diz Manolo Florentino, professor da Universidade Federal do Rio de Janeiro (UFRJ) e responsável

## Banco de dados sobre tráfico negreiro mostrou a proeminência do Brasil no comércio de escravos

pelo braço brasileiro do projeto. A principal delas foi substituir estimativas por dados consolidados, obtidos de fontes primárias. Outra foi mostrar a proeminência brasileira no comércio de escravos. “Boa parte dos documentos obtidos pelo projeto é escrita em português, uma espécie de língua franca do tráfico negreiro”, conta Florentino, que se empenhou em anos recentes em traduzir todo o site para o português. A coleção de dados sobre a deportação e a escravização dos africanos permite agora, segundo o historiador, alimentar uma frente de pesquisa menos explorada, que são as trajetórias cumpridas pelos escravos depois que chegaram aos portos, pelo interior do Brasil.

### DIVERSIDADE DE PROJETOS

Os resultados de uma recente chamada internacional de projetos mostraram a diversidade das humanidades digitais. Cento e oito propostas feitas por equipes interdisciplinares de 11 países foram submetidas à quarta edição da chamada Digging Into Data Challenge, e 14 foram aprovadas. A ini-





ciativa é parte da Plataforma Transatlântica (T-AP), colaboração em ciências humanas e sociais que reúne 16 agências de fomento da Europa e das Américas, entre as quais a FAPESP. “Tivemos um aumento expressivo de países participantes, que eram apenas quatro em chamadas anteriores. Isso faz uma grande diferença, com o surgimento de novas colaborações”, diz Brett Bobley, que idealizou o programa Digging Into Data em 2008. Os projetos aprovados se distribuem por disciplinas como musicologia, linguística, história, ciência política e economia e vão receber investimentos que somam US\$ 9,2 milhões, o equivalente a R\$ 29 milhões. Uma das propostas contempladas reúne pesquisadores dos Estados Unidos, da Alemanha e da Holanda e vai se debruçar sobre três bancos de dados que agrupam registros escritos e orais sobre folclore em vários pontos da Europa. A meta é identificar padrões que se repitam ao longo do tempo em lugares diferentes e ajudem a mostrar quais eram as crenças comuns no passado, com base nas histórias que se contavam e na dispersão de lendas e casos sobrenaturais.

**O** outro exemplo, liderado por economistas e cientistas da computação dos Estados Unidos, Canadá e Holanda, pretende cruzar informações sobre a variação dos preços de produtos vendidos pela internet no mundo inteiro, coletadas continuamente pelo projeto Billion Prices, do Massachusetts Institute of Technology (MIT), com dados econômicos para produzir pesquisas sobre a inflação, o poder de compra e o padrão de vida em vários países. Há também uma iniciativa que analisará 70 anos de cobertura na imprensa de ataques terroristas, em busca de padrões sobre o que seria uma abordagem responsável do problema, e ainda outra que investigará estruturas melódicas de gravações de jazz, tentando associá-las à evolução do contexto histórico e social em que as canções surgiram.

Para selecionar os 14 contemplados, mais de 200 especialistas avaliaram as 108 propostas. “A diversidade de problemas abordados mostra que há um potencial grande a desenvolver nas humanidades digitais no Brasil”, conta Claudia Bauzer Medeiros, professora do Instituto de Computação da **Universidade Estadual de Campinas (Unicamp)** e representante da FAPESP na T-AP, que participou de todo o processo, desde o preparo do edital até a seleção dos projetos. “Esse campo é pouco explorado no país porque aqui há ainda pouca colaboração entre pesquisadores das ciências humanas e sociais e da computação. Eles estão aos poucos percebendo que essa interação é possível. Não é preciso que o pesquisador em ciências humanas e sociais seja um entendedor de computação para trabalhar nessa área, mas é

## Londres contra o crime

Ferramentas exploram dados sobre 197 mil julgamentos

Registros sobre 197 mil julgamentos realizados entre 1674 e 1913 no Tribunal Criminal Central de Londres, mais conhecido como Old Bailey, que é o nome da rua em que fica a corte, foram disponibilizados para consulta na internet a partir de 2003 no endereço [oldbaileyonline.org](http://oldbaileyonline.org). O desafio de identificar fenômenos e tendências em meio a um volume de informações que chega a 127 milhões de palavras mobilizou pesquisadores do Reino Unido e dos Estados Unidos, que desenvolveram formas de explorar dados textuais bem mais sofisticadas do que a busca disponível no repositório.

O projeto Data Mining with Criminal Intent, financiado em 2009 pela primeira chamada Digging Into Data, esquadrinhou os registros de Old Bailey com o auxílio de uma combinação de ferramentas digitais. Uma delas é a Zotero, que permite coletar e organizar informações, e a outra, um portal chamado TAPoR, que ajuda os usuários a analisar textos utilizando diferentes softwares. A estratégia permitiu chegar a resultados curiosos. Foi possível ver, por exemplo, que a palavra “veneno” tinha associações muito mais frequentes com “café” do que com “comida”, numa indicação da forma como os londrinos eram assassinados por envenenamento.

Da mesma forma pode-se observar que as punições para bigamos se tornaram menos severas ao longo do século XIX. Segundo Stephen Ramsay, professor de inglês da Universidade de Nebraska-Lincoln, um dos líderes da iniciativa, a contribuição do projeto não se limita a obter evidências históricas que não eram percebidas anteriormente. “As histórias de Old Bailey expressam as motivações mais densas da condição humana, como a vingança, a desonra e a perda, que são matéria-prima das humanidades”, disse, segundo o *The Chronicle of Higher Education*.



necessário que colabore com especialistas nos aspectos computacionais”, afirma a pesquisadora, que é coordenadora do Programa de Pesquisa em eScience da FAPESP.

Um dos projetos selecionados no Digging Into Data Challenge tem a participação de brasileiros. Trata-se de uma colaboração entre pesquisadores da França, da Argentina e do Brasil que busca estudar como as opiniões se difundem na sociedade e como o processo sofreu transformações com o avanço da tecnologia da informação. A pesquisa vai analisar dois bancos de dados para mapear a construção de redes de relações entre grupos de indivíduos – tais conexões serão representadas em estruturas visuais, os grafos. Em um dos acervos, o do jornal *The New York Times*, o objetivo será analisar reportagens sobre o Brasil publicadas ao longo de 70 anos a fim de mapear as relações entre grupos de indivíduos e entidades mencionadas nesses textos que falaram sobre o país. “A intenção é compreender de onde vinham e como se relacionavam as ideias e opiniões reproduzidas nos textos, principalmente sobre temas políticos e econômicos, e como isso evoluiu no tempo. E também verificar a possível influência das notícias publicadas no jornal por correspondentes estrangeiros na constituição da opinião pública no país”, explica a pesquisadora Maria Eunice Quilici Gonzalez, líder do grupo



brasileiro que participa do projeto e professora do Departamento de Filosofia da Faculdade de Filosofia e Ciências da Universidade Estadual Paulista (Unesp), *campus* de Marília.

O segundo banco de dados é uma coleção de postagens sobre processos eleitorais da rede social Twitter. A ideia é mostrar como opiniões se constituem e se consolidam no ambiente virtual. “Queremos analisar a dinâmica de disseminação de opiniões em redes sociais. Quanto mais frequentes são as relações, mais densas se tornam os nós das redes representadas nos grafos.

## Como São Paulo se urbanizou

Plataforma vai reunir dados georreferenciados sobre a transformação da capital paulista entre 1870 e 1940

São Paulo urbanizou-se em velocidade superior à de outras metrópoles, saindo de apenas 30 mil habitantes em 1870 para 1 milhão de pessoas em 1940. O estudo das transformações da cidade nesse período terá o respaldo de uma plataforma com informações georreferenciadas, que será abastecida por inúmeras fontes, como teses, relatórios ou mapas. Qualquer pesquisador que tiver dados e puder relacioná-los a um endereço da capital paulista está convidado a incluí-los na plataforma Pauliceia 2.0, cujo projeto foi apresentado a potenciais usuários em 4 de abril, em busca de sugestões.

O projeto, que reúne pesquisadores da Universidade Federal de São Paulo (Unifesp), do Instituto Nacional de Pesquisas Espaciais (Inpe),

do Arquivo Público do Estado de São Paulo e da Emory University, dos Estados Unidos, é financiado pelo Programa FAPESP de Pesquisa em eScience. “Quem estudou os hotéis de São Paulo poderá alimentar os endereços com informações sobre cada um deles. Quem estudou os crimes cometidos na cidade também. Qualquer informação que possa ser situada no espaço pode alimentar a plataforma”, diz o historiador Luis Ferla, professor da Unifesp que coordena o projeto.

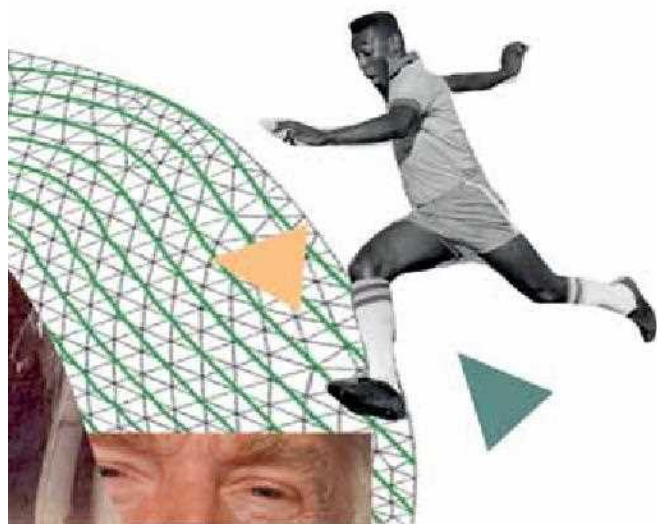
Há uma equipe dentro do projeto dedicada a desenvolver um banco de dados com a numeração de edificações da época, para garantir que a localização das informações seja fidedigna. “É um trabalho tão complexo



que está sendo testado primeiro numa área-piloto, no centro de São Paulo”, explica Ferla. Uma versão preliminar da plataforma estará disponível para testes em julho de 2018. “Quem quiser estudar esse período vai encontrar muito material na plataforma para produzir suas reflexões. O projeto quer fazer uma curadoria do conhecimento sobre a urbanização da cidade.” Mais informações estarão disponíveis no endereço [unifesp.br/himaco](http://unifesp.br/himaco).

A cidade na década de 1940, quando alcançou seu primeiro milhão de habitantes





A tendência é que eles ganhem centralidade e inibam o crescimento de outros nós, mostrando o percurso da formação de uma opinião”, informa Eunice. Um dos interesses é estudar a formação de ambientes de polarização política em redes sociais. “Grupos que antigamente estavam isolados conseguem fortalecer suas opiniões e conquistar adeptos, alimentando-se das comunicações nas redes sociais. Isso aconteceu recentemente, por exemplo, com os grupos favoráveis e contrários ao impeachment no Brasil.” Além de objetivos específicos, o projeto tem ambições mais gerais, entre as quais a de avaliar a possibilidade de criar modelos para estudar atividades sociais e investigar possíveis consequências éticas do uso da análise de Big Data em processos de auto-organização social, aqueles que emergem da interação espontânea entre vários atores sociais, sem liderança ou interferência de um centro organizador.

O projeto será realizado em parceria com pesquisadores das universidades de Cergy-Pontoise, na França, e de Buenos Aires, na Argentina. A equipe critica a tese de que é possível moldar comportamentos ou direcionar a formação de opinião manipulando tendências obtidas apenas através da análise de Big Data. “É exagerado afirmar que o Donald Trump se elegeu presidente e os britânicos votaram pela saída da União Europeia exclusivamente porque as respectivas campanhas usaram serviços de uma empresa de marketing político, a Cambridge Analytica, que teria utilizado dados e ferramentas das redes sociais para manipular medos e desejos de eleitores”, diz Eunice. “O estudo de Big Data pode apontar tendências, mas está longe de explicar a natureza humana. Seu uso só será eficiente se vier acompanhado do estudo das disposições de certos grupos, que no caso dos Estados Unidos e do Reino Unido estavam relacionadas à preponderância de um nacionalismo com aversão ao multiculturalismo.”

Graduada em física, com mestrado em filosofia e doutorado em linguística e ciência cognitiva, Eunice também vai contribuir para o projeto, com o apoio de uma equipe de pesquisadores brasileiros, com reflexões sobre a ética que envolve as ações de indivíduos em redes sociais. “O conceito de privacidade, por exemplo, está mudando. Algumas das noções de privacidade da minha geração não se aplicam aos sujeitos nas redes sociais, que expõem sistematicamente detalhes pessoais. Há, também, o problema de indivíduos que criam falsos perfis, alterando suas características pessoais, situação socioeconômica e até seu gênero para interagir virtualmente com os outros”, diz. Segundo ela, se em casa muitas vezes a pessoa tem que manter uma identidade que não lhe agrada, nas redes sociais suas fantasias podem ser realizadas sem supostas pressões familiares. “A identidade é fictícia, mas a interação que ela proporciona pode ser real, em algum sentido. Por meio dela, é possível criar uma relação com parceiros virtuais, o que não existia antigamente.” Para tratar de situações desse tipo, o grupo brasileiro irá refletir sobre como a análise de Big Data pode ajudar na compreensão de novos padrões de conduta e da dinâmica de formação da opinião coletiva.

#### TEMAS E AVANÇOS

A programação da próxima edição da conferência Digital Humanities, que reunirá em agosto cerca de mil pesquisadores de vários países na cidade de Montreal, Canadá, dá a dimensão dos temas e dos avanços que estabeleceram pontes entre cientistas da computação e profissionais das ciências humanas e sociais. Workshops vão tratar de tópicos como a aplicação em pesquisas de humanidades de ferramentas de visão computacional, conceito usado principalmente em robótica por meio do qual sistemas artificiais são capazes de extrair informações de imagens simulando o funcionamento da visão humana. Ou levantar discussões sobre problemas éticos e legais relacionados ao uso de dados digitalizados que podem expor a privacidade de indivíduos. Serão homenageados na conferência de Montreal os responsáveis pelo projeto Text Encoding Initiative (TEI), consórcio que há 30 anos desenvolve e mantém um padrão para a codificação de textos em formato digital que os torna legíveis por máquinas e que impulsionou pesquisas em ciências humanas, principalmente na área de linguística. “Nos últimos 15 anos, tivemos uma mudança qualitativa no volume de dados textuais disponíveis, o que mudou radicalmente as possibilidades de pesquisa”, afirma Karina van Dalen-Oskam, presidente da Aliança das Organizações em Humanidades Digitais (ADHO), entidade que organiza a conferência. Professora de estudos literários computacionais da Universidade de Amsterdã,



## Um *corpus* histórico da língua portuguesa

Banco de dados com 3,3 milhões de palavras reúne anotações sobre textos de várias épocas

Em algumas áreas das humanidades, a colaboração com os cientistas da computação aconteceu de forma mais natural do que em outras. Um exemplo são os estudos sobre as transformações no uso da língua. Charlotte Galves, professora do Instituto de Estudos da Linguagem da **Universidade Estadual de Campinas** (IEL-Unicamp), costuma dizer que se dedicava às humanidades digitais muito antes de saber que a denominação existia. Em 1998, ela começou a compilar textos dos séculos XVI ao XIX para compor um *corpus* histórico da língua portuguesa, um banco de textos com anotações morfosintáticas de palavras e de frases, que já serviu de base para uma série de estudos sobre a história do português em Portugal e no Brasil. "Está sendo possível observar como o idioma se transformou ao longo dos séculos, em particular no Brasil, onde vem se distanciando do português europeu sob o efeito do contato com outras línguas, apesar de voltar a sofrer a sua influência na segunda metade do século XIX", conta Charlotte.

O banco de dados foi crescendo e hoje conta com 3,3 milhões de palavras de 76 textos originais. Batizado de Corpus Tycho Brahe, em referência ao astrônomo dinamarquês do século XVI que se propôs a catalogar o movimento dos planetas, o acervo teve suas primeiras ferramentas para etiquetar palavras desenvolvidas pelo cientista da computação Marcelo Finger, professor do Instituto de Matemática e Estatística da USP.



Escritos do Padre Antônio Vieira (1608-1697) fazem parte do acervo

A evolução foi lenta – as correções das anotações automáticas foram sendo feitas pessoalmente por Charlotte, com o auxílio de pós-doutores e orientandos. "Aprendi muita coisa sobre Big Data, mas não poderia prescindir da ajuda dos cientistas da computação", afirma ela. O próximo passo é tornar o banco de dados integralmente acessível via internet – atualmente, é possível fazer o download do acervo no endereço [tycho.iel.unicamp.br/corpus](http://tycho.iel.unicamp.br/corpus), mas não pesquisas on-line.

O mesmo modelo do português histórico está sendo utilizado agora por Charlotte e Filomena Sandalo, também professora da **Unicamp**, para o estudo de uma língua indígena, o idioma kadiwéu, falado por uma etnia que habita o Mato Grosso. Relatos orais de indígenas foram coletados e estão sendo convertidos em textos escritos com anotações. "A ideia é criar corpora de outras línguas dentro da mesma plataforma, usando as mesmas ferramentas", explica Charlotte.

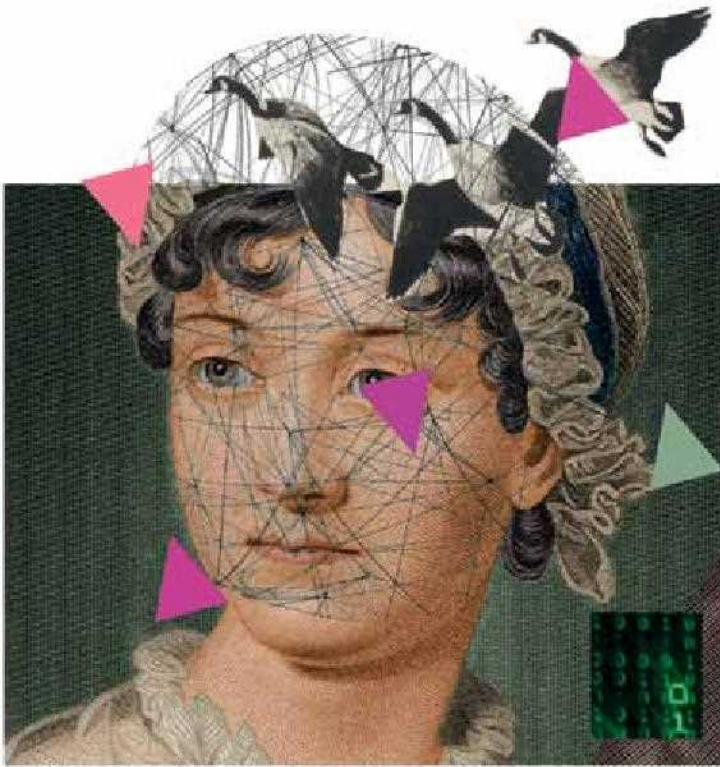
Holanda, Dale-Oskam destaca o progresso de novas abordagens para a pesquisa em literatura, como o conceito de leitura distante, que analisa grandes volumes de dados relacionados não somente à obra estudada, mas a todo o contexto histórico em que ela foi produzida, ou a área de estilometria, que permite reconhecer a autoria de textos apócrifos. "Tais abordagens permitem saber mais sobre o desenvolvimento de gêneros literários e até mesmo sobre fatores que fazem um texto se tornar ou não um best seller", diz.

O crescimento desse campo interdisciplinar convive com críticas de que as humanidades digitais produziram mais manchetes do que avanços robustos do conhecimento e também que rivalizam com os campos tradicionais das humanidades na divisão do financiamento à pesquisa. Em um artigo publicado no diário *The New York Times* em 2015, Armand Marie Leroi, professor de biologia evolucionária do Imperial College de Londres, Reino Unido, pôs em dúvida a capacidade de as humanidades digitais produzirem análises inovadoras de literatura. Segundo ele, o expediente de converter arte em dados torna possível procurar novos significados em uma obra por meio de novos algoritmos. "Mas será preciso criar um algoritmo muito esperto capaz de sinalizar a ironia na obra de Jane Austen", escreveu. "A verdade da crítica de arte não é do mesmo tipo da verdade científica."

**O**s pesquisadores da área respondem com o argumento de que as humanidades digitais oferecem apenas uma extensão dos métodos e habilidades tradicionais, sem a ambição de substituí-los. Escrito por um conjunto de autores, o livro *Digital humanities* (MIT Press, 2012) sustenta em seu primeiro capítulo que as humanidades digitais "não obliteram as ideias do passado, mas suplementam o compromisso das humanidades com a interpretação acadêmica, a pesquisa informada, o argumento estruturado e o diálogo entre as comunidades que a praticam".

O cientista político Eduardo Marques, professor da Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo (FFLCH-USP), ressalta que as abordagens da ciência da computação e das ciências humanas e sociais dentro das humanidades digitais têm origens diferentes. "Houve o encontro de dois movimentos. Um veio das ciências duras, com o desenvolvimento de ferramentas de mineração de dados que permitiram produzir informações sobre o mundo social e gerar novos campos empíricos. Já nas ciências humanas, partiu do uso já existente de ferramentas estatísticas para estudar fenômenos sociais", explica. Como as lógicas são distintas, há dificuldade em unificá-las, observa Marques. "Enquanto os cientistas da computação buscam padrões nos grandes volumes de dados





## Cursos e disciplinas sobre análise quantitativa e ética no uso de dados ganham espaço

para levantar perguntas de pesquisa, os cientistas sociais partem de pressupostos teóricos e usam ferramentas digitais para testar sua validade. O diálogo é grande, mas há dificuldade de unificar formas diferentes de se aproximar das questões.”

Esse diálogo vem influenciando a formação de pesquisadores. No caso das ciências humanas e sociais, cursos e disciplinas em métodos e análise quantitativa ganham mais espaço. “É uma boa notícia porque as ciências sociais sempre tiveram uma grande fragilidade nesse campo no Brasil, que se estende também à análise qualitativa e a estudos com amostras pequenas”, avalia Marques, referindo-se a iniciativas como a Escola de Verão em Conceitos, Métodos e Técnicas em Ciência Política e Relações Internacionais oferecida pela Associação Internacional de Ciência Política (Ipsa), o Departamento de Ciência Política da FFLCH-USP e o Instituto de Relações Internacionais da USP. Também ganham importância as disciplinas sobre o uso ético de dados. “É um tema emergente e não busca apenas prevenir a divulgação de dados sigilosos sobre pacientes ou informações sensíveis à segurança pública”, acrescenta Claudia Bauzer Medeiros. Há o risco de produzir análises enviesadas porque muitos

programas de computador “aprendem” com os dados processados. Os softwares são desenvolvidos para identificar padrões ao longo do tempo e incorporá-los à sua capacidade de análise. “Já houve situações em que o aprendizado inadvertidamente reproduziu preconceitos. Nos Estados Unidos, descobriu-se que um programa utilizado experimentalmente por juizes em algumas cidades para agilizar decisões tratava com mais rigor negros e latinos, porque usava dados de decisões anteriores tomando-os como lição.”

O desenvolvimento de ferramentas computacionais que auxiliam na análise de grandes volumes de dados sobre saúde, demografia e violência alimenta estudos sobre processos sociais que ganham aplicações em políticas públicas. “É comum utilizar análises de dados socioeconômicos e demográficos em estratégias de planejamento urbano. A digitalização de dados sobre ondas migratórias abastece estudos que ajudam a compreender tendências futuras em imigração”, exemplifica a pesquisadora do IC-Unicamp.

Um exemplo do envolvimento crescente das ciências sociais com o Big Data no Brasil pode ser visto no Centro de Estudos da Metrópole (CEM), um dos Centros de Pesquisa, Inovação e Difusão (Cepid) financiados pela FAPESP. Uma das vertentes do centro é produzir e disseminar dados georreferenciados sobre as metrópoles brasileiras. Órgãos públicos produziam dados, que acabavam não sendo disponibilizados e eram apropriados por empresas, que cobravam para fornecê-los. O CEM comprou várias bases de dados e digitalizou outras, disponibilizando-as em seu site ([fflch.usp.br/centrodametropole](http://fflch.usp.br/centrodametropole)). A princípio, as coleções não eram grandes o suficiente para se enquadrarem no conceito de Big Data. Isso mudou há alguns anos, quando o centro desenvolveu um banco de dados talhado para um grande esforço de pesquisa sobre o estudo dos padrões de desigualdade nos últimos 60 anos. Foi necessário um intenso trabalho para dar consistência a questionários e corrigir lacunas de uma amostra remanescente do Censo de 1960, cujos cartões perfurados se perderam, e reorganizar as informações dos cinco recenseamentos posteriores para gerar dados comparáveis. “Isso gerou um banco de muitos terabytes de informação, em um volume muito maior do que o tradicional nas ciências sociais no país”, afirma Eduardo Marques, que foi diretor do CEM entre 2004 e 2009. O esforço produziu o livro *Trajetórias das desigualdades – Como o Brasil mudou nos últimos 50 anos* (Editora Unesp, 2015), coordenado pela atual diretora do CEM, Marta Arretche, com capítulos escritos por especialistas em temas como educação e renda, demografia, mercado de trabalho e participação política. Cada capítulo exigiu um processamento específico de dados. ■