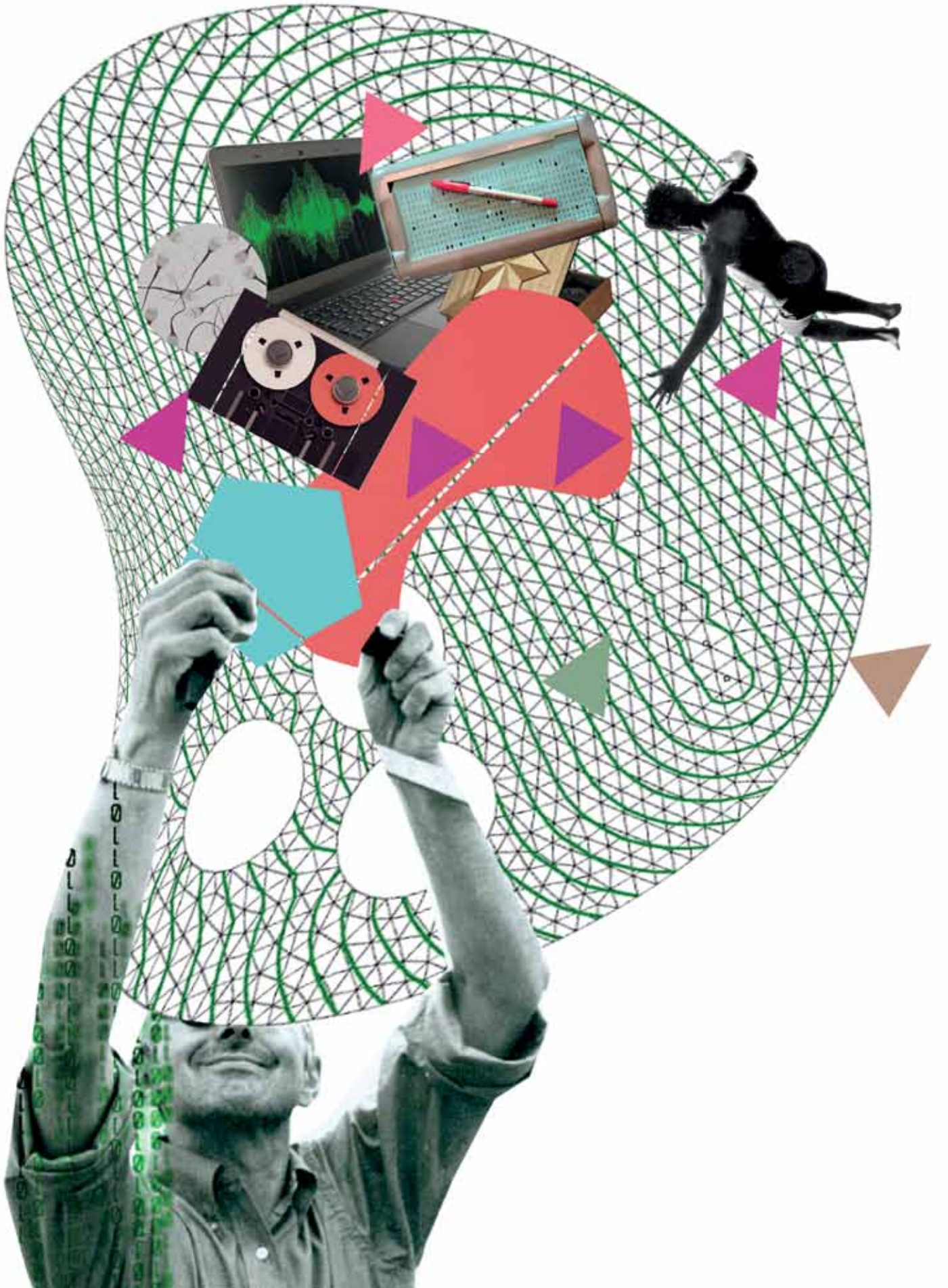
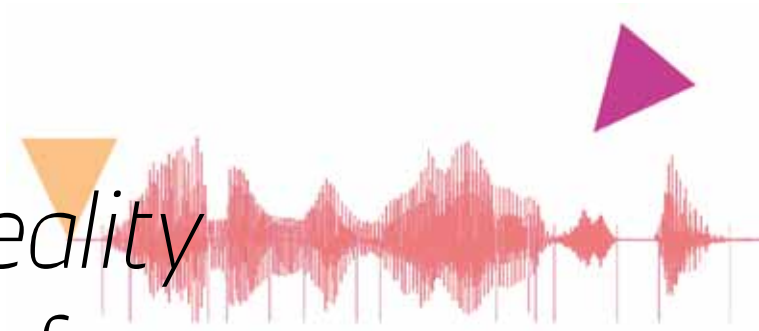




The reality emerging from an **avalanche of data**

Digital humanities are spreading through a variety of disciplines, influencing the training of researchers and driving public policy

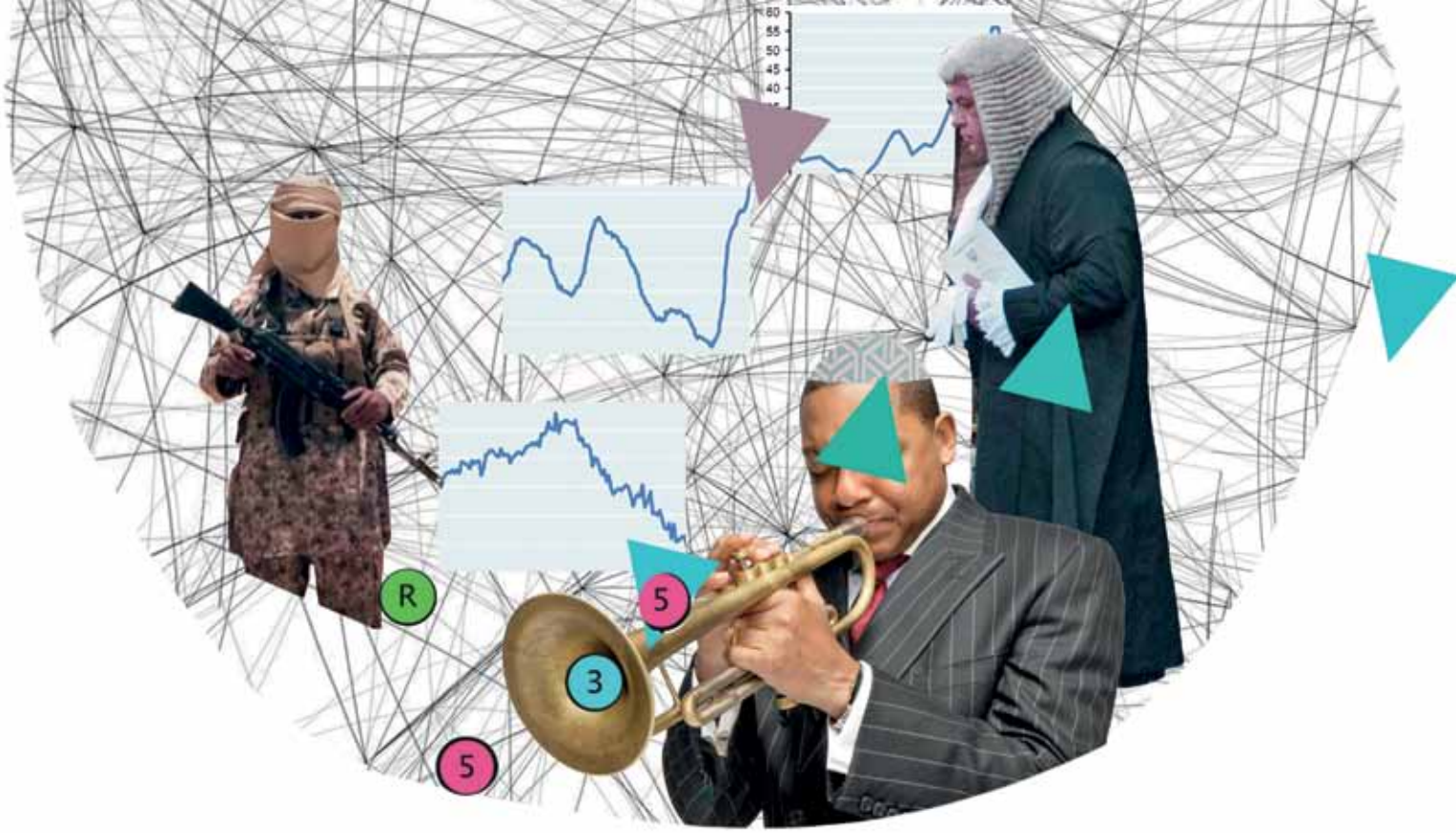
Fabrício Marques

PUBLISHED IN MAY 2017

Computers are tools used in the work of researchers in all fields of knowledge, but in the communities of the humanities and social sciences, the digitization of artistic and historical collections and the input of economic and social information into giant databases have opened up new frontiers for observing phenomena and analyzing trends. This expansion has rather naturally developed into a closer relationship with computer scientists, whose Big Data research studies have multiplied the ways to organize and analyze information, giving rise to an interdisciplinary field known as the digital humanities. "The term was coined to define research that uses computational technology to study the humanities, but it also refers to the research that uses the humanities to study digital technology and its influence on culture and society," explained Brett Bobley, director of the Office of Digital Humanities at the National Endowment for the Humanities (NEH), a U.S. government funding agency. "This is not a new field," said Bobley, "but rather a range of activities that can include the use of aerial photographs by archeologists to scan sites, the development of data analysis techniques that help linguists study old newspapers, and the study of the ethics of the technology by philosophers, to give just a few examples."

One of the NEH-funded projects in digital humanities recovered the field diaries of British explorer David Livingstone (1813-1873). Historical accounts of his 1871 voyage to Central Africa were written on old newspapers because no paper was available. Over time, the ink faded, and the writings in which Livingstone recorded his impressions about the dynamics of the slave trade, among other observations, were rendered illegible. Between 2013 and 2017, a group of humanities and computer science researchers from the United States and the United Kingdom were able to recover the writings by using spectral imaging

ILLUSTRATIONS ZÉ VICENTE



photographic techniques that permitted retrieval of information invisible to the human eye.

Another example was the collaboration between historians from several parts of the world in organizing records about nearly 36,000 slave ship voyages that took place between 1514 and 1866, carrying more than 12 million slaves from Africa. The effort, begun in the 1990s by American historian David Eltis at Emory University, resulted in the Trans-Atlantic Slave Trade Database, available online since 2007 at slavevoyages.org. The analysis of the data, assembling records in several languages and encompassing the activities of the ports through which the vessels passed, has offered the historians new insights on how Africans experienced and resisted deportation and enslavement and revealed new transatlantic connections in the slave trade.

An initial compilation was released as a CD-Rom in 1999, but the collaborative effort to obtain data about the voyages put together a more complete picture of the slave trade. During its initial phase, it is estimated that Brazil took in nearly 3.6 million slaves, but documents showed that this contingent was closer to 5 million—for a total of 10.7 million Africans deported to the Americas. "The initiative had a considerable impact on the research about slavery," said Manolo Florentino, a professor at the Federal University of Rio de Janeiro (UFRJ) in charge of the Brazilian arm of the project. Chief among them was the fact that it replaced estimates with solid data obtained from primary sources. Another impact was that it displayed Brazil's prominence in the slave trade. "A large number of the documents obtained through the project are written in Portuguese, a sort of lingua franca of the slave trade," said Florentino, who in recent years has embarked on efforts to translate the entire site into Portuguese. Florentino said that the collection of data on the deportation and enslavement of the Africans now provides information for a less-explored line of research involving the paths the slaves took inside Brazil after they arrived in the ports.

Database of slave trafficking demonstrated Brazil's prominence in the slave trade

A VARIETY OF PROJECTS

The results of a recent international call for proposals has demonstrated the diversity of the digital humanities. One hundred and eight proposals by interdisciplinary teams from 11 countries were submitted during the fourth edition of what is known as the "Digging into Data Challenge," and 14 were approved. The initiative is part of the Trans-Atlantic Platform (T-AP), a collaboration in the humanities and social sciences that

is bringing together 16 funding agencies from Europe and the Americas, including FAPESP. "We saw a noticeable increase in the number of countries taking part, which in previous calls for proposals had numbered only four. The surge in new collaborations is making a big difference," said Brett Bobley, who devised the idea for the Digging into Data program in 2008. Approved projects encompass disciplines, such as musicology, linguistics, history, political science and economics, and they will receive investments totaling $9.2 million, equivalent to R$29 million. One of the proposed projects involves researchers from the United States, Germany and The Netherlands and will focus on three databases that make up the written and oral records of folklore from several corners of Europe. The goal is to identify patterns that reappear over time in different places to show which beliefs were common in the past based on the stories told and the spreading of legends and tales of supernatural occurrences.

Another example, led by economists and computer scientists from the United States, Canada and The Netherlands, plans to cross-reference information about price variations of products sold on the Internet all over the world, continuously collected by the Billion Prices project at the Massachusetts Institute of Technology (MIT), with economic data that can be used to produce research studies on inflation, purchasing power, and standards of living in several countries. There is also an initiative to analyze 70 years of press coverage of terrorist attacks in a search for patterns regarding what constitutes a responsible approach to the problem. Another project will investigate the melodic structures of jazz recordings in an attempt to connect them to the development of the historical and social context in which the songs emerged.

To select the 14 included projects, more than 200 experts evaluated the 108 proposals. "The variety of issues covered shows that there is a huge potential to be developed in the field of digital humanities in Brazil," said Claudia Bauzer Medeiros, a professor at the Institute of Computing at the University of Campinas (UNICAMP) and FAPESP representative on the T-AP. Medeiros took part in the entire process, from drafting the call for proposals to selecting the projects. "The field is under-explored in Brazil because there is still so little collaboration among researchers from the humanities and social sciences and computer science. They're gradually realizing that this interaction is possible. Researchers in the humanities and social sciences don't have to understand computing to work well in this field, but they do have to collaborate with experts on

London in the fight against crime

Tools explore data on 197,000 trials

Records on 197,000 trials conducted between 1674 and 1913 by London's Central Criminal Court, commonly referred to as Old Bailey, which is the name of the street on which the court is located, were made available for consultation on the Internet back in 2003 on oldbaileyonline.org. The challenge posed by the task of identifying phenomena and trends buried in a volume of information approaching 127 million words mobilized researchers from the United Kingdom and the United States to develop ways to tap textual data that were much more sophisticated than performing a search of the repository.

The project, known as "Data Mining with Criminal Intent," funded in 2009 under the initial call for proposals for the Digging into Data project, scoured the records of Old Bailey with the help of a combination of digital tools. One of them was Zotero, which allowed for the collection and organization of information, and the other was a portal called TAPoR that helped users analyze writings through a variety of software. The strategy has led to some interesting results. It was possible to see, for example, that the word "poison" was much more commonly associated with "coffee" than with "food," indicating how Londoners were murdered by poisoning.

By the same token, one notes that punishments for bigamists became less severe throughout the 19th century. According to Stephen Ramsay, a professor of English at the University of Nebraska-Lincoln, one of the leaders of the initiative, the project's contribution was not limited to obtaining previously unnoticed historical evidence. "The stories of Old Bailey express the darker motivations behind the human condition, such as revenge, dishonor and loss, which is the raw material of the humanities," he said, according to *The Chronicle of Higher Education*.

aspects of computing," said the researcher who is also the coordinator of the FAPESP Research Program on eScience.

Brazilians are participating in one of the projects selected under the Digging into Data Challenge. It involves a collaboration among researchers from France, Argentina and Brazil studying how opinions spread in society and how the process has changed as a result of advances in information technology. The study will analyze two databases to map the establishment of networks of relationships among groups of individuals; such connections will be represented in visual structures (graphs). In one collection by the *New York Times* newspaper, the objective will be to analyze reports about Brazil published over the course of 70 years to map the relationships between groups of individuals and entities mentioned in the pieces that mentioned Brazil. "The plan is to understand where they came from and how the ideas and opinions reproduced in the texts were related, especially those regarding political and economic topics, and how this has changed over time. We also want to determine the possible influence that news by foreign correspondents published in that newspaper had on the formation of public opinion in Brazil," explained Maria Eunice Quilici Gonzalez, researcher and head of the Brazilian group that is taking part in the project and a professor in the Department of Philosophy of the School of Philosophy and Sciences at São Paulo State University (UNESP), Marília campus.

The second database is a collection of Twitter postings on electoral processes. The idea is to show how opinions form and grow stronger in the virtual environment. "We would like to analyze the dynamics of how opinions spread through social media. The more extensive the relationships, the tighter are the network connections represented in the graphs. The trend

How São Paulo became urbanized

Platform will assemble geo-referenced data about the transformation of São Paulo's capital city from 1870 to 1940

São Paulo urbanized at a faster rate than other cities, growing from only 30,000 inhabitants in 1870 to one million in 1940. The study of the city's transformations during this period will be supported by a platform of geo-referenced information, supplied by numerous sources, such as theses, reports and maps. Any researcher who has data and can relate them to an address in the São Paulo capital is invited to include them in the Pauliceia 2.0 platform, whose design was opened to suggestions from potential users on April 4, 2017.

The project, which brings together researchers from the Federal University of São Paulo (UNIFESP), the National Institute for Space Research (INPE), the São Paulo State Public Archives and Emory University, is funded by the FAPESP research program in eScience. "Anyone who has studied São Paulo's hotels could add information about them to the addresses. Anyone who has studied crimes committed in the city can do the same for that data. Any information that can be referenced in the space can be added to the platform," said historian Luis Ferla, the UNIFESP professor who coordinated the project.

There is one project team that is dedicated to developing a database of the numbering on buildings of that time to ensure that data localization is reliable. "It is such complex work that it is first being tested in a pilot area, in downtown São Paulo," Ferla explained. A preliminary version of the platform will be available for testing in July 2018. "Anyone who wants to study this period will find a lot of material on the platform to use in their analyses. The project seeks to curate knowledge about the city's urbanization." More information is available at unifesp.br/himaco.

The city in the 1940s, when it reached its first million inhabitants

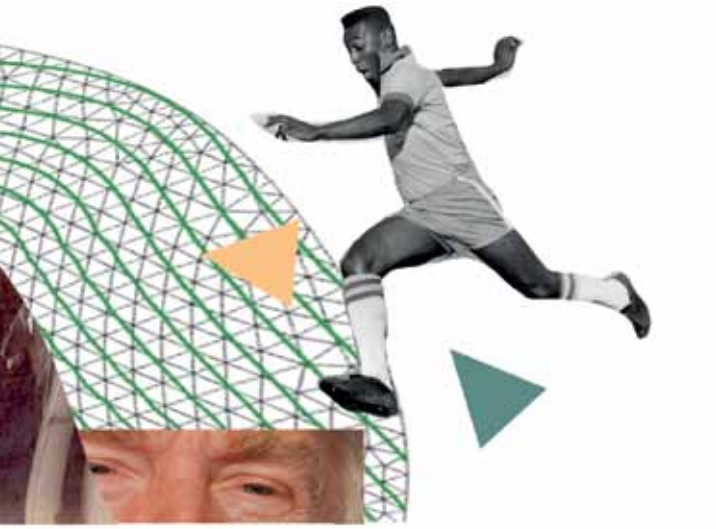


PHOTO HILDEGARD ROSENTHAL / MOREIRA SALLES INSTITUTE COLLECTION

is for them to take center stage and inhibit the growth of other connections, thus showing the pathway to how opinions are formed," Gonzalez reported. One of the group's interests lies in studying the formation of politically polarizing environments on social media. "Groups that once were isolated are now able to reinforce their opinions and gain followers, feeding off of communications on social media," Gomez said. "This happened recently, for example, with groups for or against impeachment in Brazil." In addition to specific objectives, the project has more general ambitions, including assessing possibilities for creating models to study social practices and investigate the potential ethical consequences of using Big Data analysis on processes of social self-organization, which are those that emerge from spontaneous interactions among various social actors—leaderless and without interference from an organized center.

The project will be carried out in partnership with researchers from the universities of Cergy-Pontoise in France and Buenos Aires in Argentina. The team is critical of the idea that it is possible to shape behaviors or guide the formation of opinions by manipulating trends obtained through analysis of Big Data alone. "It would be an exaggeration to say that Donald Trump was elected president and that the British voted to leave the European Union solely because the respective campaigns hired the political marketing firm Cambridge Analytics to utilize data and social media tools to manipulate voters' wishes and fears," Gonzalez said. "The study of Big Data can identify trends, but it is far from capable of explaining human nature. Its use will only be efficient if it is accompanied by the study of the attitudes of certain groups, which in the case of the United States and the United Kingdom were related to the preponderance of nationalism and an aversion to multiculturalism."

With an undergraduate degree in physics, a master's degree in philosophy and a PhD in linguistics and cognitive science, Gonzalez will also contribute to the project, with the support of a team of Brazilian researchers, by providing ideas concerning the ethics involving individuals' actions on social media. "The concept of privacy, for example, is changing. Some of the notions of privacy held by my generation do not apply to people on social media who systematically expose their personal details. There is also the issue of individuals who create false profiles, altering their personal characteristics, socioeconomic status and even their gender in an effort to virtually interact with others," she said. In her view, if at home, many people have to maintain an identity they do not like, so they can live out their fantasies on social media without any apparent family pressures. "Their identity is fictitious, but the interaction that it provides can to some extent be real. They are able to use it to create a relationship with virtual partners, which in the past was not possible." To address situations such as this, the Brazilian group will think about how Big Data analysis can help in the understanding of new patterns of behavior and the dynamics of formulating public opinion.

TOPICS AND ADVANCES

The next scheduled edition of the Digital Humanities conference in August 2017, which will bring nearly 1,000 researchers from several countries together in Montreal, Canada, gives us some idea of the scope of the topics and technological advances that have established bridges between computer scientists and professionals in the humanities and social sciences. Workshops will address topics, such as research applications in the humanities for computer vision tools, a concept used mainly in robotics through which artificial systems are able to extract information about images, simulating the functioning of the human vision system. They may also raise questions about ethical and legal problems related to the use of digitized data that could expose an individual's privacy. Honored at the conference in Montreal will be those responsible for the Text Encoding Initiative (TEI) project, a consortium that has developed and maintained a standard for the representation of texts in digital format since the 1980s, making them machine readable, and driving studies in the human sciences, especially in linguistics. "In the last 15 years, we've had a qualitative change in the volume of textual data available, which has radically changed the possibilities of research," said Karina van Dalen-Oskam, chair of the Steering Committee of the Alliance of Digital Humanities Organizations (ADHO), the entity that organizes the conference.

A historical corpus of the Portuguese language

Database containing 3.3 million words assembles annotations on writings from various eras

A collaboration with computer scientists has occurred more naturally in some fields of the humanities than in others. One example involves studies about changes in the use of language. Charlotte Galves, a professor at the Institute of Language Studies of the University of Campinas (IEL-UNICAMP), often said that she became devoted to the digital humanities long before she knew there was such a thing. In 1998, she began to compile 16th- to 19th-century writings to put together a historical corpus of the Portuguese language, a database of texts with morpho-syntactic annotations of words and sentences that had already served as a basis for a series of studies about the history of the Portuguese language in Portugal and Brazil. "It is now possible to observe how the language has changed over the centuries, particularly in Brazil, which has increasingly distanced itself from European Portuguese as a result of its contact with other languages, despite being influenced by it again during the second half of the 19th century," said Galves.

Writings by Father Antônio Vieira (1608-1697) are part of the collection

The database has continued to grow and now contains 3.3 million words from 76 original documents. Named Corpus Tycho Brahe, in reference to the 16th-century Danish astronomer who documented the movement of the planets, the collection used its first word-labeling tools developed by computer scientist Marcelo Finger, a professor at the Institute of Mathematics and Statistics of the University of São Paulo (IME-USP). The database grew slowly; corrections to the automatic notations were made by Galves herself, with the help of postdoctoral researchers and students she advised. "I learned a lot about Big Data, but I couldn't do without the help of computer scientists," she said. The next step is to make the database fully accessible on the Internet. It is possible to download the collection at: www.tycho.iel.unicamp.br/corpus, but it is not currently possible to search online.

The same model of historical Portuguese is now being used by Galves and Filomena Sandalo, also a professor at UNICAMP, for the study of an indigenous language, *Kadiwéu*, spoken by an ethnic group in the Brazilian state of Mato Grosso. Oral accounts by indigenous people were collected and are being converted into annotated texts. "The idea is to use the same platform to create the corpora for other languages, using the same tools," Galves explained.

A professor of computational literary studies at the University of Amsterdam in The Netherlands, van Dalen-Oskam points to the progress that new approaches have made in researching literature, such as the concept of remote scanning, which analyzes large volumes of data related not only to the work being studied but also to the entire historical context in which it was produced, or to the field of stylometry that enables attribution of authorship to works of doubtful authenticity. "These approaches allow us to learn more about the development of literary genres and even about factors that make a particular text a best seller or not," she said.

The growth of this interdisciplinary field is accompanied by criticism that the digital humanities have generated more headlines than solid advances in knowledge and that they compete with traditional humanities in terms of the allocation of research funding. In an article published in *The New York Times* in 2015, Armand Marie Leroi, a professor of evolutionary biology at Imperial College London, called into doubt digital humanities' capacity to produce innovative analyses of literature. He said that converting art into data does make it possible to look for new meanings in a work through new algorithms. "But it would have to create a very smart algorithm capable of flagging irony in the work of Jane Austen," he wrote. "The truth we talk about in art criticism is not the same as scientific truth."

Researchers in this field respond with the argument that the digital humanities offer only an extension of traditional methods and skills and are not intended to replace them. *Digital Humanities* (MIT Press, 2012) states in its first chapter that the digital humanities "do not obliterate the ideas of the past, but rather supplement the commitment by the humanities to academic interpretation, informed research, organized argument and dialogue between the communities that practice it."

Political scientist Eduardo Marques, a professor at the University of São Paulo School of Philosophy, Literature and Human Sciences (FFLCH-USP), pointed out that the approaches used by computer science and human and social sciences within the digital humanities come from different sources. "There was a meeting of two movements. One came from the hard sciences, with the development of data mining tools that enabled the production of information about the social world and the generation of new empirical fields. The human sciences, however, made use of existing statistical tools to study social phenomena," he explained. Since the rationales are different, it is difficult to bring them together, Marques noted. "While the computer scientists

PHOTO REPRODUCTION OF OIL ON CANVAS, BY AN EARLY 18TH-CENTURY UNKNOWN AUTHOR / WIKIMEDIA COMMONS

Courses and disciplines on quantitative and ethical analysis of the use of data are gaining ground

are looking for patterns in large volumes of data in order to raise research questions, the social scientists are working from theoretical assumptions and are using digital tools to test their validity," he said. "There is a lot of dialogue, but it is hard to bring together different ways of approaching the issues."

This dialogue has influenced the training of researchers. In the humanities and social sciences, courses and disciplines in quantitative methods and analysis are gaining ground. "This is good news because the social sciences have always had a huge weakness in this field in Brazil, which also extends to qualitative analysis and studies with small samples," Marques explained, referring to initiatives, such as the Summer School in Concepts, Methods and Techniques in Political Science and International Relations offered by the International Political Science Association (IPSA), the Department of Political Science at FFLCH-USP and the Institute of International Relations at USP. Also growing in importance are disciplines on the ethical use of data. "It is an emerging issue and does not just look at how to prevent the dissemination of confidential patient data or sensitive public safety information," added Claudia Bauzer Medeiros. There is a risk of producing biased analyses because many computer programs "learn" as the data are processed. Software is being developed to identify long-term patterns and incorporate them into their analytical capacity. "There have been situations in which the learning inadvertently reproduced biases," she said. "In the United States, it was discovered that a program used experimentally by judges in some cities to expedite rulings dealt more stringently with blacks and Latinos because it used as a lesson data from previous rulings."

The development of computational tools that help analyze large volumes of data about health, demographics and violence is used in studies of social processes that are then applied in public policies. "Socioeconomic and demographic data analyses are often used in urban planning strategies. Digitization of data on migratory waves feeds studies that help us understand future trends in immigration," said the IC-UNICAMP researcher.

An example of the growing involvement of the social sciences in Big Data in Brazil can be seen at the Center for Metropolitan Studies (CEM), one of the Research, Innovation and Dissemination Centers (RIDCs) funded by FAPESP. One focus of the center is to produce and disseminate geo-referenced data on Brazilian cities. Public agencies generated data that were not made available, and the information was appropriated by companies, which charged to provide them. The CEM purchased several databases and digitized others, making them available on its website (fflch.usp.br/centrodametropole). At first, the collections were not large enough to be associated with the notion of Big Data. This changed a few years ago when the center developed a database tailored toward a large research effort on the study of patterns of inequality in the last 60 years. Significant work was required to provide consistency to questionnaires and correct the gaps in a 1960 Census sample, whose punch cards had been lost, and to reorganize the information from five later censuses to generate comparable data. "This generated a multi-terabyte database of information, at a volume much larger than what is traditionally seen in Brazil's social sciences," said Eduardo Marques, who was CEM director from 2004 to 2009. The effort led to the book entitled *Trajetórias das desigualdades – Como o Brasil mudou nos últimos 50 anos* (Editora UNESP, 2015) [Paths of Inequality in Brazil: A Half-Century of Change], edited by current CEM Director, Marta Arretche, containing chapters written by experts on topics such as education and income, demographics, labor markets and political participation. Each chapter required a specific processing of data. ■