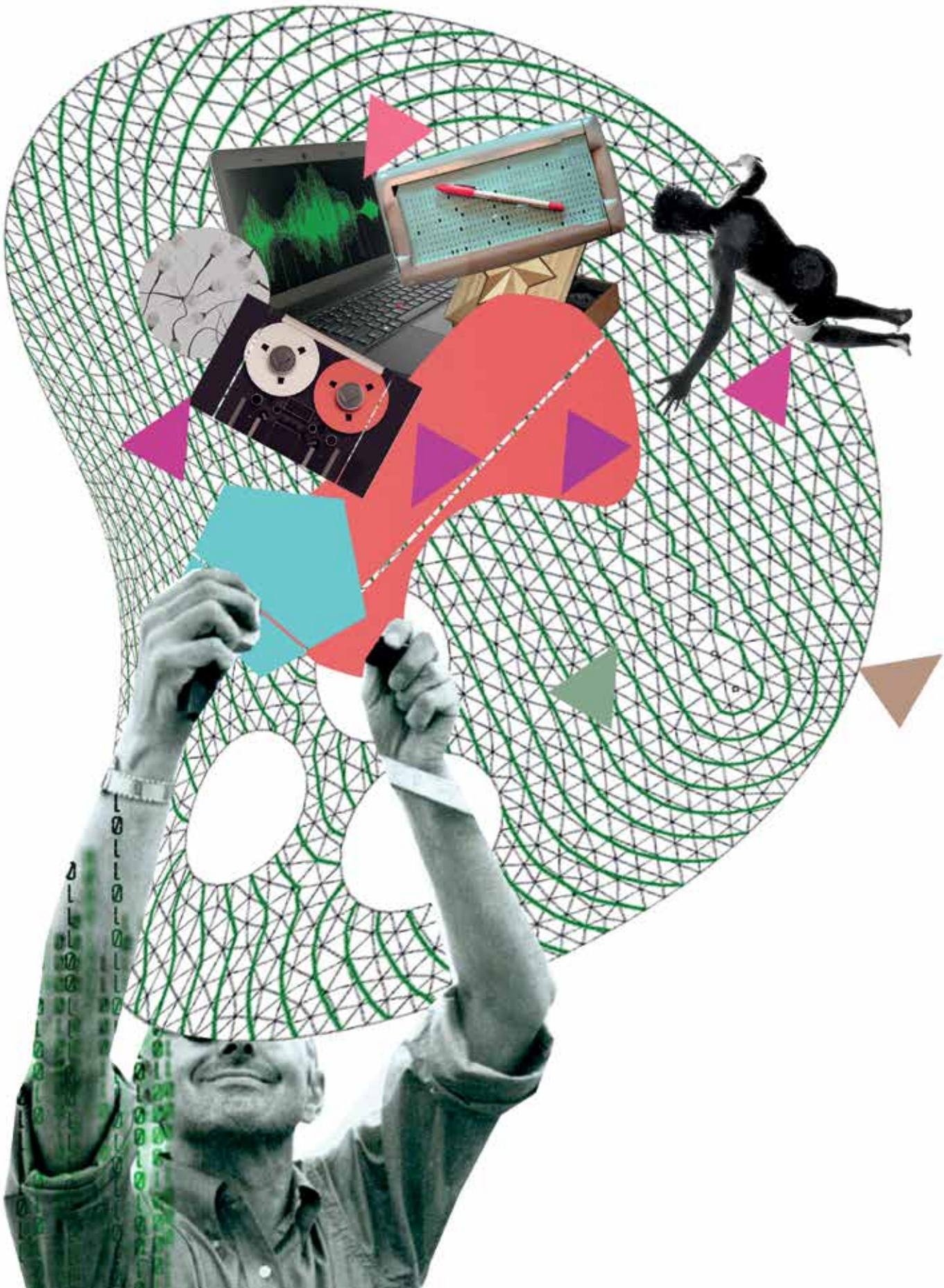


La réalité qui émerge

de l'avalanche de données

Les humanités numériques se diffusent dans plusieurs disciplines, influencent la formation de chercheurs et inspirent des politiques publiques

Fabrício Marques

PUBLIÉ EN MAI 2017

Les ordinateurs sont un outil de travail pour les chercheurs de tous les domaines de la connaissance. Mais dans le cas de la communauté des sciences humaines et sociales, la numérisation d'archives artistiques et historiques et l'offre de gigantesques banques de données d'informations économiques et sociales ont ouvert la voie à de nouvelles observations de phénomènes et d'analyse de tendances. Il y a eu un grand rapprochement avec les informaticiens, dont les recherches sur le Big Data ont multiplié les types d'organisation et d'analyse des informations, et à la croisée de ce rapprochement est né un champ interdisciplinaire : les humanités numériques. D'après Brett Bobley, directeur du Bureau des Humanités Numériques de la National Endowment for the Humanities (NEH), agence de soutien à la recherche du gouvernement nord-américain, « le terme a été créé pour définir la recherche qui incorpore la technologie informatique à des études en sciences humaines et sociales, mais aussi celle qui utilise les sciences humaines et sociales pour étudier la technologie numérique et son influence sur la société et sur la culture ». Il ne s'agit pas selon lui d'un nouveau domaine de connaissance, mais d'un éventail d'activités qui peut englober notamment l'utilisation de photographies aériennes par des archéologues pour scanner des sites, le développement de techniques d'analyse de données qui aident des linguistes à étudier des journaux anciens ou encore l'étude de l'éthique de la technologie par des philosophes.

Un des projets financés par la NEH en humanités numériques a travaillé sur les carnets de note de l'explorateur britannique David Livingstone (1813-1873). Des récits de son voyage en Afrique Centrale, en 1871, ont été écrits sur de vieux journaux par manque de papier disponible. Avec le temps, l'encre a commencé à s'effacer et les impressions de Livingston sur la dynamique du commerce d'esclaves, par exemple, sont devenus illisibles. Entre 2013 et 2017, un groupe de chercheurs en sciences sociales et humaines et en informatique des États-Unis et du Royaume-Uni a réussi à récupérer les écrits en employant des techniques de photographie d'image spectrale, capables de récupérer des informations invisibles à l'œil nu.

ILLUSTRATIONS ZÉ VICENTE

Un autre exemple fut la collaboration d'historiens de plusieurs parties du monde pour organiser les registres de près de 36 000 voyages de navires négriers entre 1514 et 1866, qui ont transporté plus de 12 millions d'esclaves d'Afrique. Débuté dans les années 1990 par l'historien nord-américain David Eltys de l'Université Emory, le projet a donné lieu à une banque de données sur le commerce d'esclaves, disponible sur Internet depuis 2007 à l'adresse slavevoyages.org. L'analyse des données réunit des registres dans plusieurs langues et évoque les ports par où sont passés les navires. Elle a permis aux historiens d'en savoir plus sur la manière dont les Africains ont vécu et résisté à la déportation, et révélé aussi de nouvelles connexions transatlantiques dans le commerce d'esclaves.

Un premier relevé a été lancé en 1999 sous la forme d'un CD-Rom, mais le travail de collaboration pour obtenir des données sur les voyages a ensuite permis de retracer un portrait beaucoup plus vaste du commerce d'esclaves. La première estimation d'esclaves envoyés au Brésil était d'environ 3,6 millions d'esclaves, mais des documents ont montré que ce nombre était plus proche des 5 millions – sur un total de 10,7 millions d'Africains déportés vers les Amériques. D'après Manolo Florentino, professeur de l'Université Fédérale de Rio de Janeiro (UFRJ) et responsable de la branche brésilienne du projet, l'initiative a eu différents impacts sur les recherches sur l'esclavage. Le principal a été de remplacer des estimations par des données solides, issues de sources primaires. Un autre a été de montrer la prépondérance brésilienne dans le commerce d'esclaves : « Une grande partie des documents obtenus par le projet est écrite en portugais, une sorte de langue véhiculaire de la traite négrière ». Au cours de ces dernières années, Florentino s'est employé à traduire tout le site en portugais.

À présent, la collection de données sur la déportation et l'esclavage des Africains permet d'alimenter une source de recherche moins explorée : les trajectoires empruntées par les esclaves dans le pays, après leur arrivée dans les ports brésiliens.

Une banque de données sur la traite négrière a montré la place importante occupée par le Brésil dans la commerce d'esclaves

DIVERSITÉ DE PROJETS

Les résultats d'un récent appel à projets international ont montré la diversité des humanités numériques. 108 propositions d'équipes interdisciplinaires de 11 pays ont été soumises à la 4e édition de la *Digging Into Data Challenge*, et 14 ont été approuvées. L'initiative fait partie de la Plateforme Transatlantique (T-AP), une collaboration en sciences humaines et sociales qui réunit 16 agences de soutien à la recherche d'Europe et des Amériques, parmi lesquelles la FAPESP. Selon Bobley, « il y a eu une augmentation significative de pays participants, qui n'étaient que 4 aux appels antérieurs. Cela fait une grande différence, avec l'apparition de nouvelles collaborations ». Bobley a idéalisé le programme *Digging Into Data Challenge* en 2008. Les projets approuvés se répartissent par disciplines telles que la musicologie, la linguis-

tique, l'histoire, la science politique et l'économie, et ils vont recevoir des investissements de l'ordre de 9,2 millions de dollars US. L'une des propositions réunit des chercheurs des États-Unis, d'Allemagne et de Hollande qui vont analyser trois banques de données rassemblant des registres écrits et oraux du folklore de différentes parties de l'Europe. L'objectif est d'identifier les modèles qui se répètent au cours du temps dans des endroits différents et qui aideront à identifier les croyances anciennes les plus communes, sur la base des histoires qui se racontaient et sur la propagation de légendes et de cas surnaturels.

Un autre exemple, mené par des économistes et des informaticiens des États-Unis, du Canada et de Hollande, entend croiser des informations sur la variation des prix de produits vendus sur Internet dans le monde entier, collectées en continu à travers le projet *Billion Prices* du Massachusetts Institute of Technology (MIT), avec des données économiques pour produire des recherches sur l'inflation, le pouvoir d'achat et le niveau de vie dans plusieurs pays. Il y a aussi une initiative qui analysera 70 ans de couverture médiatique des attaques terroristes, en quête de modèles sur ce que serait une approche responsable du problème. Une autre étudiera les structures mélodiques d'enregistrements de jazz pour tenter de les associer à l'évolution du contexte historique et social dans lequel les chansons sont apparues.

Pour sélectionner les 14 projets retenus, plus de 200 spécialistes ont évalué les 108 propositions. Claudia Bauzer Medeiros, professeure de l'Institut Informatique de l'Université d'état de Campinas et représentante de la FAPESP au niveau de la T-AP, a participé à tout le processus depuis l'appel d'offres jusqu'à la sélection des projets : « La diversité de problèmes abordés montre qu'il y a un grand potentiel à développer dans les humanités numériques au Brésil. [...] Ce domaine est peu exploré dans le pays parce qu'ici il y a encore peu de collaboration entre les chercheurs en sciences humaines et sociales et ceux de l'informatique. Ils perçoivent peu à peu que cette interaction est possible. Le chercheur en sciences humaines et sociales n'a pas besoin d'être un connaisseur en informatique pour travailler dans ce domaine, mais il faut qu'il collabore avec des spécialistes pour les aspects informatiques ». La chercheuse est aussi coordinatrice du programme de Recherche en eScience de la FAPESP.

Un des projets sélectionnés dans le *Digging Into Data Challenge* compte sur la participation de Brésiliens. Il s'agit d'une collaboration entre des chercheurs de France, d'Argentine et du Brésil pour étudier comment se diffusent les opinions dans la société et comment le processus

Londres contre le crime

Des outils exploitent des données de 197 000 procès

Des registres de 197 000 procès réalisés entre 1674 et 1913 à la Haute Cour criminelle de Londres, plus connue sous le nom de Old Bailey (qui est le nom de la rue où se trouve la Cour), sont depuis 2003 accessibles sur Internet à l'adresse oldbaileyonline.org. L'identification de phénomènes et de tendances parmi un volume d'informations de plus de 127 millions de mots a mobilisé des chercheurs du Royaume-Uni et des États-Unis, qui ont développé des types d'exploitation de données textuelles beaucoup plus sophistiquées que la recherche disponible dans les archives.

Financé en 2009 par le premier appel à projets Digging Into Data, le projet *Data Mining with Criminal Intent* a examiné les registres de Old Bailey avec différents outils numériques. L'un d'eux est le logiciel Zotero, qui permet de collecter et d'organiser des informations ; et l'autre le portail de recherche en analyse de textes, TAPoR, qui aide les usagers à analyser des textes en utilisant différents logiciels. La stratégie a permis d'obtenir des résultats singuliers. Par exemple, le mot « poison » était beaucoup plus souvent associé à « café » qu'à « nourriture », ce qui a permis de comprendre de quelle manière les londoniens étaient empoisonnés.

D'autre part, il a été possible d'observer que les punitions pour les bigames sont devenues moins sévères au cours du XIX^{e} siècle. D'après Stephen Tamsay, professeur d'anglais de l'Université de Nebraska-Lincoln, et un des leaders de l'initiative, le projet ne sert pas seulement à obtenir des évidences historiques qui n'étaient pas perçues auparavant. Dans *The Chronicle of Higher Education*, il écrit : « Les histoires de Old Bailey expriment les motivations plus profondes de la condition humaine, comme la revanche, le déshonneur et la perte, qui sont la matière première des humanités ».

a subi des transformations avec l'avancée de la technologie de l'information. La recherche va analyser deux banques de données pour recenser la construction de réseaux de relations entre des groupes d'individus – ces connexions seront représentées sous forme de structures visuelles, les graphes. Dans les archives du journal The New York Times, le but est d'analyser des reportages sur le Brésil publiés pendant 70 ans afin de recenser les relations entre groupes d'individus et les entités mentionnées dans ces textes qui parlent du pays. Maria Eunice Quilici Gonzalez, leader du groupe brésilien qui participe au projet et professeure du Département de Philosophie de la Faculté de Philosophie et Sciences de l'Université d'état de São Paulo (Unesp – campus de Marília), explique : « l'intention est de comprendre d'où viennent et comment se relient les idées et les opinions reproduites dans les textes, principalement sur des thèmes politiques et économiques, et comment cela a évolué dans le temps. Et aussi de vérifier l'influence éventuelle des informations publiées dans le journal par des correspondants étrangers dans la constitution de l'opinion publique dans le pays ».

La deuxième banque de données est une collection de messages sur les processus électoraux du réseau social Twitter. L'idée est de montrer

comment des opinions se constituent et se consolident dans le milieu virtuel. « Nous voulons analyser la dynamique de propagation d'opinions sur des réseaux sociaux », déclare Gonzalez. « Plus les relations sont fréquentes, plus les nœuds des réseaux représentés dans les graphes deviennent denses. La tendance est qu'ils deviennent centraux et inhibent la croissance des autres nœuds, montrant ainsi le parcours de la formation d'une opinion ». Un des intérêts est d'étudier la formation de milieux de polarisation politique sur des réseaux sociaux. « Des groupes qui avant étaient isolés réussissent à fortifier leurs opinions et à

L'urbanisation de São Paulo

Une plateforme va réunir des données géoréférencées sur la transformation de São Paulo entre 1870 et 1940

São Paulo s'est urbanisée plus vite que d'autres métropoles, elle est passée d'à peine 30 000 habitants en 1870 à 1 million en 1940. L'étude des transformations de la ville pendant cette période va être visible sur une plateforme d'informations géoréférencées, qui sera alimentée par plusieurs sources dont des thèses de doctorat, des rapports et des cartes. Tout chercheur possédant des données en lien avec une adresse de la ville est invité à les inclure sur la plateforme Pauliceia 2.0., dont le projet a été présenté à des usagers potentiels le 4 avril 2017, en quête de suggestions.

Le projet réunit des chercheurs de l'Université Fédérale de São Paulo (Unifesp), de l'Institut National des Recherches Spatiales (Inpe), des Archives Publiques de l'état de São Paulo et de l'Emory University, des États-Unis. Il est financé par le programme FAPESP de Recherche en eScience. Comme l'indique Luis Ferla, historien, professeur de l'Unifesp et coordinateur du projet, « Celui qui a étudié les hôtels de São Paulo pourra alimenter les adresses avec des informations sur chacun d'eux. Celui qui a étudié les crimes commis dans la ville aussi. Toute information pouvant être située dans l'espace peut alimenter la plateforme ».

Au sein du projet, une équipe développe une banque de données avec la numération de constructions de l'époque, pour garantir que la localisation des informations soit la plus fidèle possible. Ferla explique que « c'est un travail tellement complexe qu'il est d'abord testé dans une zone pilote, au centre de São Paulo ». Une version préliminaire de la plateforme sera disponible pour des tests en juillet 2018. « Ceux qui veulent étudier cette période trouveront beaucoup de matériel sur la plateforme pour produire des réflexions. Le projet souhaite organiser la connaissance sur l'urbanisation de la ville ».
Plus d'informations seront disponibles à l'adresse unifesp.br/himaco.

La ville dans les années 1940, quand elle a atteint son premier million d'habitants

PHOTO HILDEGARD ROSENTHAL / ARCHIVES INSTITUTO MOREIRA SALLES

faire de nouveaux adeptes en s'alimentant des communications sur les réseaux sociaux. Cela s'est passé récemment, par exemple, avec les groupes favorables et contraires à la destitution de la présidente au Brésil ». En plus des objectifs spécifiques, le projet a des ambitions plus générales telle que la création de modèles pour étudier des activités sociales et l'analyse d'éventuelles conséquences éthiques de l'utilisation de l'analyse de Big Data dans des processus d'auto-organisation sociale, ceux qui émergent de l'interaction spontanée entre plusieurs auteurs sociaux, sans leadership ou interférence d'un centre organisateur.

Le projet sera réalisé en partenariat avec des chercheurs des universités de Cergy-Pontoise et de Buenos Aires. L'équipe critique la thèse selon laquelle il est possible de modeler des comportements ou orienter la formation d'opinions en manipulant des tendances seulement obtenues à travers l'analyse de Big Data. Selon Maria Eunice Gonzalez, « il est exagéré d'affirmer que Donald Trump est devenu président et que les Britanniques ont voté pour la sortie de l'Union Européenne seulement parce que les campagnes respectives utilisaient les services d'une entreprise de marketing politique, la Cambridge Analytica, qui aurait utilisé des données et des outils des réseaux sociaux pour manipuler les peurs et les désirs des électeurs [...] L'étude de Big data peut montrer des tendances, mais elle est loin d'expliquer la nature humaine. Son utilisation ne sera efficace que si elle s'accompagne de l'étude des dispositions de certains groupes, qui dans le cas des États-Unis et du Royaume-Uni étaient liées à la prépondérance d'un nationalisme et d'une aversion du multiculturalisme ».

Détentrice d'un premier cycle en physique, d'un master en philosophie et d'un doctorat en linguistique, Maria Eunice Gonzalez va aussi travailler sur le projet et réfléchir, avec le soutien d'une équipe de chercheurs brésiliens, à l'éthique qui concerne les actions d'individus sur les réseaux sociaux : « Le concept de privacité, par exemple, est en train de changer. Certaines notions de privacité de ma génération ne s'appliquent pas aux sujets sur les réseaux sociaux, qui exposent systématiquement des détails personnels. Il y a aussi le problème des individus qui créent de faux profils, modifient leurs caractéristiques personnelles, leur situation socio-économique et même leur genre pour interagir virtuellement avec d'autres. Si chez elle la personne est très souvent tenue de maintenir une identité qui ne lui plaît pas, sur les réseaux sociaux elle peut réaliser ses fantasmes sans pressions familiales éventuelles. « L'identité est fictive, mais l'interaction qu'elle offre peut être réelle dans un certain sens. Par son biais il est possible de créer une relation avec des partenaires virtuels, ce qui n'existait pas auparavant ». Pour traiter les situations de ce type, le groupe brésilien va penser à la manière dont l'analyse de Big Data peut aider à comprendre les nouveaux modèles de conduite et la dynamique de formation de l'opinion collective.

THÈMES ET AVANCÉES

La programmation de la prochaine édition de la conférence Digital Humanities, qui réunira en août 2017 près de 1 000 chercheurs de plusieurs pays à Montréal, donne une idée de l'étendue des thèmes et des avancées du travail entre informaticiens et professionnels des sciences humaines et sociales. Des groupes de travail vont traiter de topiques tels que l'application dans des recherches en sciences humaines d'outils de vision informatique, un concept surtout utilisé en robotique et par le biais duquel des systèmes artificiels sont capables d'extraire des informations d'images en simulant le fonctionnement de la vision humaine. Ou engager des discussions sur les problèmes éthiques et légaux liés à l'emploi de données numérisées qui peuvent exposer la vie privée des individus. Lors de la conférence, un hommage sera rendu aux responsables du projet *Text Encoding Initiative* (TEI), un consortium qui depuis 30 ans développe et maintient un modèle de codification de textes en format numérique lisible pour des machines, et qui a suscité des recherches en sciences humaines, en particulier dans le domaine de la linguistique. Karina van Dalen-Oskam, présidente de l'Alliance des Organisations en Humanités Numériques (ADHO), l'entité qui organise la conférence, affirme qu'au cours des 15 dernières années il y a eu « un changement qualitatif du volume de données textuelles disponibles, ce qui a changé radicalement les possibilités de recherche ». Professeure d'études littéraires computationnelles de l'Université d'Amsterdam, Dale-Oskam met en avant le progrès de nouvelles approches pour la recherche en littérature, comme le concept de lecture distante, qui analyse de grands volumes de données liées aussi bien à l'ouvrage étudié qu'à tout le contexte dans lequel il a été produit ; ou le domaine de la stylométrie, qui permet de reconnaître les auteurs de textes apocryphes : « Ce genre d'approches permet d'en savoir plus sur le développement de genres littéraires et même sur des facteurs qui font d'un texte un best-seller ou non ».

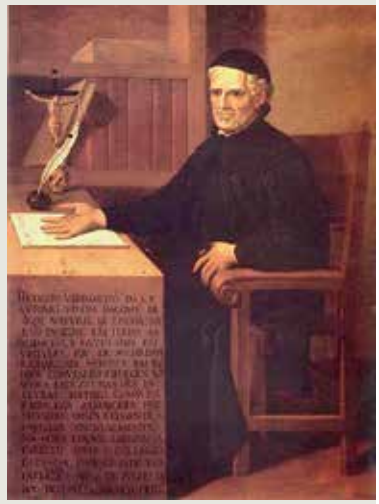
La croissance de ce champ interdisciplinaire s'accompagne aussi de critiques : pour leurs détracteurs, les humanités numériques produiraient plus des manchettes dans les journaux que des avancées solides de la connaissance. En plus,

Un *corpus* historique de la langue portugaise

Une banque de données de 3,3 millions de mots réunit des annotations sur des textes de différentes époques

Dans certains domaines des sciences humaines et sociales, la collaboration avec les informaticiens s'est faite de manière plus naturelle que dans d'autres. À titre d'exemple, les études sur les transformations dans l'utilisation de la langue. Charlotte Galves, professeure de l'Institut des Études du Langage de l'Université d'état de Campinas (IEL-Unicamp) a l'habitude de dire qu'elle se consacrait aux humanités numériques bien avant de savoir que la dénomination existait. En 1998, elle a commencé à compiler des textes des XVI[e] au XIX[e] siècles pour composer un corpus historique de la langue portugaise, une banque de textes avec des annotations morphosyntaxiques de mots et de phrases, qui a déjà servi de base à une série d'études sur l'histoire du portugais au Portugal et au Brésil : « Il est possible d'observer comment la langue s'est transformée au fil des siècles, en particulier au Brésil, où elle s'est éloignée du portugais européen sous l'effet du contact avec d'autres langues, même si elle a à nouveau subi son influence dans la deuxième moitié du XIX[e] siècle ».

La banque de données a grandi et compte aujourd'hui 3,3 millions de mots de 76 textes originaux. Elle a été baptisée Corpus Tycho Brahe en référence à l'astronome danois du XVI[e] siècle qui s'est proposé de cataloguer le mouvement des planètes. Ses premiers outils pour étiqueter des mots ont été développés par l'informaticien Marcelo Finger, professeur de l'Institut des Mathématiques et Statistiques de l'USP. L'évolution a été lente – les corrections des annotations automatiques ont été faites



Les écrits du Père Antônio Vieira (1608-1697) font partie du corpus

personnellement par Charlotte Galves, avec l'aide d'étudiants. « J'ai appris beaucoup de choses sur le Big Data », déclare-t-elle, « mais je ne pourrais pas me passer de l'aide des informaticiens ». La prochaine étape est de rendre la banque de données intégralement accessible via Internet – actuellement, il est possible de télécharger le matériel à l'adresse tycho.iel.unicamp.br/corpus, mais pas de faire des recherches en ligne.

Le même modèle du portugais historique est à présent utilisé par Charlotte Galves et Filomena Sandalo, également professeure de l'Unicamp, pour l'étude d'une langue indigène, le kadiwéu, parlé par une ethnie qui habite dans l'état du Mato Grosso. Des comptes rendus oraux d'indigènes ont été collectés et sont convertis en textes écrits avec des annotations. « L'idée est de créer un ensemble d'informations sur d'autres langues sur la même plateforme, en utilisant les mêmes outils », explique Charlotte Galves.

elles rivalisent avec les champs traditionnels des sciences humaines et sociales dans le partage du financement de la recherche. Dans un article du The New York Times de 2015, Armand Marie Leroi, professeur de biologie évolutionnaire de l'Imperial College de Londres, met en doute la capacité des humanités numériques à produire des analystes novatrices de littérature. Pour lui, « changer de l'art en données permet de rechercher de nouveaux sens dans une œuvre par le biais de nouveaux algorithmes : « Mais il faudra créer un algorithme très intelligent, capable de signaler l'ironie dans l'œuvre de Jane Austen. [...] La vérité de la critique d'art n'est pas du même type que la vérité scientifique ».

En réponse, les chercheurs concernés affirment que les humanités numériques offrent seulement une extension des méthodes et des compétences traditionnelles, sans intention de les remplacer. Écrit par un collectif d'auteurs, le livre *Digital Humanities* (MIT Press, 2012) explique dans son premier chapitre que les humanités numériques « ne suppriment pas les idées du passé mais complètent l'engagement des sciences humaines vis-à-vis de l'interprétation universitaire, de la recherche informée, de l'argument structuré et du dialogue entre les communautés qui la pratiquent ».

Le politologue Eduardo Marques, professeur de la Faculté de Philosophie, Lettres et Sciences Humaines de l'Université de São Paulo (FFLCH-USP), précise que les approches de la science informatique et des sciences humaines et sociales ont, dans les humanités numériques, des origines différentes : « Il y a eu la rencontre de deux mouvements. Un est venu des sciences dures, avec le développement d'outils d'exploration de données qui ont permis de produire des informations sur le monde social et de générer de nouveaux champs empiriques. Et dans les sciences humaines, il est parti de l'utilisation déjà existante d'outils statistiques pour étudier des phénomènes sociaux ». Et comme les logiques sont différentes, il est difficile de les unifier : « Tandis que les informaticiens recherchent des modèles dans les grands volumes de données pour susciter des questions de recherche, les professionnels des sciences sociales partent de présupposés théoriques et utilisent des outils numériques pour tester leur validité. Le dialogue est grand, mais il est difficile d'unifier des formes différentes d'approche des questions ».

Ce dialogue influence la formation de chercheurs. Dans le cas des sciences humaines et sociales, des cours et des disciplines sur les méthodes et l'analyse quantitative augmentent. Marques estime que « c'est une bonne nouvelle, parce que les sciences sociales ont toujours connu

PHOTO REPRODUCTION À L'HUILE SUR TOILE D'AUTEUR INCONNU, DÉBUT DU XVIIIE SIÈCLE / WIKIMEDIAS COMMONS

Les cours et les disciplines sur l'analyse quantitative et l'éthique dans l'utilisation des données augmentent

une grande fragilité dans ce domaine au Brésil, qui s'étend aussi à l'analyse qualitative et à des études avec de petits échantillons ». Il fait ici référence à des initiatives telle que celle de l'École d'Été de Concepts, Méthodes et Techniques en Science Politique et Relations Internationales offerte par l'Association Internationale de Science Politique (IPSA), le Département de science politique de la FFLCH-USP et l'Institut des Relations Humaines de l'USP. Les disciplines sur l'utilisation éthique de données sont aussi davantage mises en évidence. « C'est un sujet émergent et il ne cherche pas seulement à prévenir la diffusion de données confidentielles sur des patients ou d'informations sensibles à la sécurité publique », prévient Claudia Bauzer Medeiros. Le risque est de produire des analyses erronées parce beaucoup de programmes informatiques « apprennent » avec les données traitées. Les logiciels sont développés pour identifier des modèles au fil du temps et les incorporer à leur capacité d'analyse. « Il y a déjà eu des situations où l'apprentissage a reproduit sans le vouloir des préjugés. Aux États-Unis, on a découvert qu'un programme utilisé expérimentalement par des juges dans certaines villes pour accélérer des décisions était plus rigoureux avec les Noirs et les Latino-Américains parce qu'il prenait comme référence des décisions où ils étaient cités comme référence négative ».

Le développement d'outils informatiques qui aident à analyser de grands volumes de données sur la santé, la démographie et la violence alimente des études sur des processus sociaux qui sont appliqués dans des politiques publiques. La chercheuse de l'IC-Unicamp cite comme exemple : « Il est habituel d'utiliser des analyses de données socioéconomiques et démographiques dans des stratégies de planification urbaine. La numérisation de données sur les vagues migratoires alimente des études qui aident à comprendre les tendances futures en matière d'immigration ».

Un exemple de l'échange croissant entre les sciences sociales et le Big Data au Brésil est le Centre d'Études sur la Métropole (CEM), un des Centres de Recherche, Innovation et Diffusion (Cepid) financés par la FAPESP. Un des objectifs du centre est de produire et de diffuser des données géoréférencées sur les métropoles brésiliennes. Les organismes publics produisaient des données appropriées mais qui n'étaient pas mises à disposition. Les entreprises intéressées devaient les acheter. Le CEM a acheté plusieurs bases de données et en a numérisé d'autres, puis les a mises à disposition sur son site (fflch.usp.br/centrodametropol). Au début, les collections n'étaient pas assez grandes pour faire partie du concept de Big Data. Mais cela a changé il y a quelques années, quand le centre a développé une grande banque de données pour l'étude des modèles d'inégalité des 60 dernières années. Un travail intense a été nécessaire pour perfectionner les questionnaires et corriger les lacunes d'un échantillon restant du Recensement de 1960, dont les fiches se sont perdues, et réorganiser les informations des cinq recensements suivants pour produire des données comparables. Eduardo Marques a été directeur du CEM entre 2004 et 2009 : « Cela a généré une banque de plusieurs téraoctets d'informations, un volume beaucoup plus grand que celui traditionnel des sciences sociales dans le pays ». Le travail a donné lieu au livre *Trajectoires des inégalités – Comment le Brésil a changé au cours des 50 dernières années* [titre original : *Trajetórias das desigualidades – Como o Brasil mudou nos últimos 50 anos*] (éd. Unesp, 2015), coordonné par la directrice actuelle du CEM, Marta Arretche. Les chapitres ont été écrits par des spécialistes en éducation et revenu, démographie, marché du travail, participation politique et autres. Chaque chapitre a exigé un traitement spécifique de données. ■