DIGITAL DATA

Google Books, Wikipedia, and the Future of Culturomics

BOSTON—Humanities scholars gathered for an unusual talk here on 8 May at the annual meeting of the American Historical Association. Jean-Baptiste Michel and Erez Lieberman Aiden, mathematicians at nearby Harvard University, focused on issues that are standard material for historians. They discussed, for example, the period of intense censorship in Nazi Germany that began with festive book burnings in 1933 and ended with the Nazi's surrender in 1945. Some of those books were written by Jewish intellectuals, others just contained ideas that were considered "un-German." Some scholars and artists who were highly influential virtually disappeared from public discourse when the Nazi Party came to power, while those favored by Nazi propaganda vaulted to prominence.

What was unusual was the method that Michel and Lieberman Aiden used to study censorship during that period. By tracking the rise and fall of people's names in millions of German and English books, the researchers identified not only people known to have been censored but also many whose suppression was not recorded. The mathematicians did not read those millions of books, of course. Instead, they performed a quantitative analysis of data they obtained from Google Books. "This is fantastic," says Anthony Grafton, a historian from Princeton University who was sitting in the audience. Grafton stresses that the technique is "a new starting point" for historical analysis rather than a replacement. "But it is amazing that you get a coherent picture of censorship in the public sphere."

On page 176 of this issue (and published online 16 December 2010), a team led by Michel and Lieberman Aiden introduce this and many other examples of this data-intensive approach to the humanities, which they call culturomics (*Science*, 17 December 2010, p. 1600). The entire data set that underlies their study—a mapping of the words from 4% of all books ever published—is now online for any researcher to explore. "But even with all those data," cautions Michel, "you'll need to carefully interpret your results." One of the big challenges, for example, is to correctly extract the names of individual people from those 500 billion words. A potentially valuable resource for this type of analysis is Wikipedia, the massive online encyclopedia, which contains entries for nearly 750,000 people born since 1800. But its shortcomings, from the reliability of its information to the organization of its content, become quickly apparent to those who use it. Several efforts are under way to improve Wikipedia as a teaching and research tool, including one by the Association for Psychological Science (APS).

Adrian Veres, one of the co-authors of the Michel *et al*. paper, experienced these problems firsthand. Veres, a chemistry and physics undergraduate at Harvard University, has been using Wikipedia to analyze the fame of scientists whose names appear in books over the centuries. His first task was to identify who among the myriad of people mentioned in books are actually scientists. He started by creating computer algorithms that scour Wikipedia for telltale words such as biologist, chemist, and physicist, as well as keywords that identify people in subfields such as ecology, physiology, and genetics.

Wikipedia's content is created through cooperative—and often combative—editing by millions of volunteers. But even if the information is correct, "not all of it is structured data" with clear labels, says Veres. It was a serious challenge for his algorithms to extract biographical information, such as years of birth and death, profession, and nationality, from entries that are a "bag of words." An online community effort called DBpedia has been enforcing structure onto Wikipedia entries, encouraging editors to sort the information into standard fields and then collecting those data for researchers to use. But that effort is only starting.

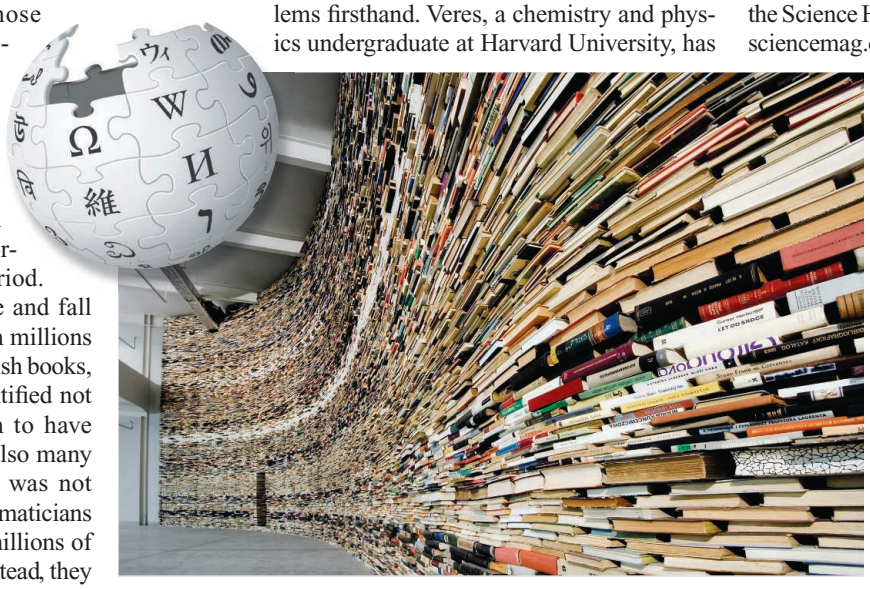
Veres whittled down about 7000 candidates to a list of 4209 physicists, chemists, biologists, and mathematicians who were alive between the years 1800 and 2000. He then ranked them by "fame"—as measured by the frequency with which those scientists' full names appear in books. He shared the early results of this analysis with *Science*. You may be surprised who tops the list. You can view the Science Hall of Fame online at http://www.sciencemag.org/content/331/6014/143.3.full.

Even so, Veres had to exclude the social sciences from his analysis because the Wikipedia entries were so unreliable. In the field of psychology, for example, Linda Bartoshuk, a psychophysicist and former APS president, was not detected as a scientist because of the paucity of details in her Wikipedia entry. (Bartoshuk was profiled in *Science* on 18 June 2010.)

Psychologists are aware of the problem. "Unfortunately, the quality of information about psychological science in Wikipedia is uneven," says Mahzarin Banaji, current APS president and a psychologist at Harvard. She is calling on psychology researchers, teachers, and students to improve Wikipedia themselves. APS is building a Web portal that will channel the effort with help from Robert Kraut and Rosta Farzan, computer scientists at Carnegie Mellon University in Pittsburgh, Pennsylvania, who specialize in human-computer interaction. The goal is to create "a more complete and accurate representation of our science," says Banaji. The Web portal is slated to debut 1 February at www.psychologicalscience.org.

When they first heard about the "culturomics" approach to the humanities, many scholars reacted "as if this were the coming of the antichrist," says Grafton. "But my reaction is, God look at this new tool!"

–JOHN BOHANNON



Working relationship. Improvements in Wikipedia will be needed for many research projects using Google's digitized books database.

CREDITS (LEFT TO RIGHT): WIKIPEDIA; (SEDIMENT) MATEJ KREN. (PHOTO) GABRIEL URBÁNEK