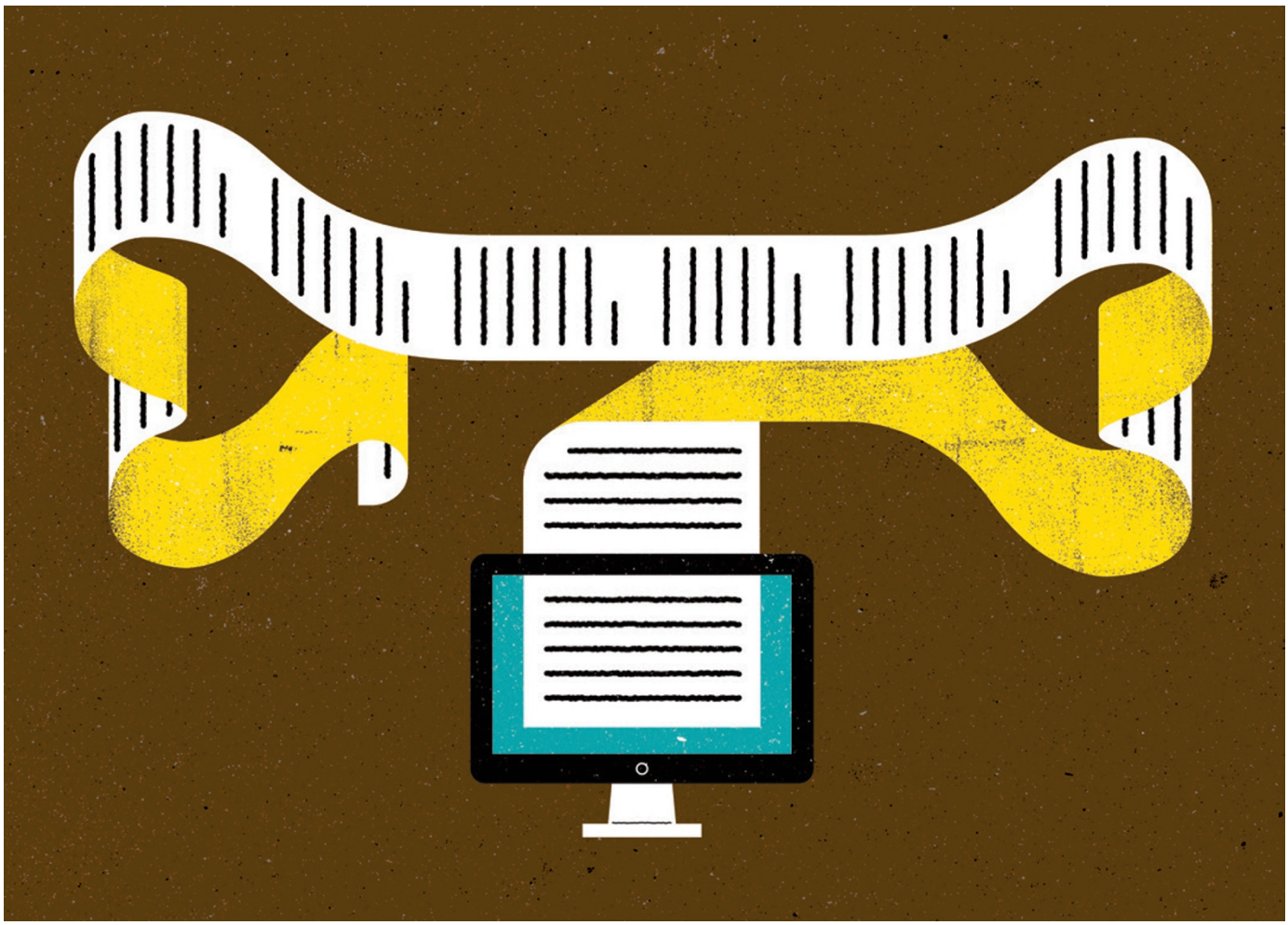
TOOLBOX

COMPUTERS READ THE FOSSIL RECORD

Palaeontologists hope that software can construct fossil databases directly from research papers.

ILLUSTRATION BY THE PROJECT TWINS



BY EWEN CALLAWAY

For a field whose raison d'être is to chronicle the deep past, palaeontology is remarkably forward-looking when it comes to organizing its data. Victorian natural history museums meticulously organized their collections with handwritten cards that survive to this day. And over the past 15 years, researchers have collectively entered records of more than a million fossils into an online database, allowing them to track broad trends in the history of life. Now, palaeontologists are exploring the use of machine algorithms to pull fossil data from their research papers automatically.

"I'm fairly convinced that this is the future, for sure," says Shanan Peters, a palaeontologist at the University of Wisconsin–Madison (UW Madison) who is co-leading an effort to use software to extract information from tens of thousands of palaeontology papers. "Building a database, per se, will be a thing of the past. Those databases will be dynamically generated based on the questions you're interested in, and the machine will do the heavy lifting."

Peters should know. He is the principal investigator of the Paleobiology Database (PBDB; paleobiodb.org), which details the age, location and identity of some 1.2 million fossils. Since it was started in 1998, researchers have spent about 80,000 hours — the equivalent of 9 continuous years — entering and opining over data from original field research and ▶

around 40,000 articles. The PBDB has produced hundreds of papers and has allowed palaeontologists to address questions that would have been otherwise unanswerable, on topics ranging from epoch-wide extinction rates to the disappearance of certain dinosaurs.

The PBDB is a database created by experts: around 380 scientists have uploaded some 560,000 published opinions on the classifications of 320,000 taxonomic names. But Peters was curious to know whether such a database could be compiled automatically by computer. So in 2013 he started a collaboration with Miron Livny and Chris Ré, then data scientists at UW Madison (Ré has since moved on to Stanford University in California). Ré had developed software called DeepDive, which mines written text (such as words in a research paper) and pulls out facts. Text mining — or content mining — is now a commonplace tool in computer science and is slowly beginning to find uses in research fields from genomics to drug discovery. Text mining palaeontology literature appealed to Ré, partly because the PBDB offers a human-curated database with which to compare a computer-generated counterpart.

PARSING THE PAST

DeepDive begins by parsing research papers in a manner that would be familiar to anyone who remembers their early grammar lessons. "It's taking those papers and converting them into text," says Ré: it is trying to determine the answer to questions such as, "What's a noun, what's a verb and how do you diagram a sentence?" Next, DeepDive attempts to predict the concepts that are stored in those sentences (such as, for palaeontology, the names of fossils and the places where they were found) and assigns a probability to each assertion. The result is software "which is usually imperfect in a lot of ways", says Ré. "That's where you get the domain scientist involved."

Peters spent about a year refining the first-pass software so that, for instance, it knows where to look in palaeontology papers for the names of new species and the geographic locations in which they were discovered. Ré describes this process as a "back and forth" with Peters that required Ré's team of data scientists to come up with custom computing solutions to make the requests feasible. "I would love to say the answer is people can press a button and use it and run it and they don't need us," Ré says — but that is a goal that his team has not yet reached.

As a proof of principle, Peters and Ré used custom software that they called PaleoDeepDive to create a text-mined, scaled-down version of the PBDB that incorporated around 12,000 papers. In some ways the computer-generated database outshines the PBDB, Peters says, because all the information in it comes with a probability assigned to it and is linked back to the original text. "The machine is really clear about uncertainty, when there's ambiguity, or differences between documents and authors," Peters says. PaleoDeepDive also managed to extract 192,000 opinions on the classification of taxonomic names from the papers, whereas the PBDB's human curators found only 80,000.

PaleoDeepDive did not do such a bad job at organizing that information either. In a December 2014 paper, Ré and Peters report that from a random sample of 100 statements drawn from the computer-generated database, 92% were correct — which they say was similar to the accuracy of the PBDB (S. E. Peters *et al. PLoS ONE* **9,** e113523; 2014). The two databases also scored similarly in a second experiment, when scientists were presented with five documents and asked to score the accuracy of facts that had been mined from them by the PBDB and by the computer.

> ***"It's a little scary, the machines are getting that good."***

And perhaps most impressively, PaleoDeepDive was used to estimate species diversity and extinction rates over the past 500 million years, coming up with measures similar to those determined by the PBDB.

"It's a little scary, the machines are getting that good. That's just something that we're going to have to get used to," says Mark Uhen, a palaeontologist at George Mason University in Fairfax, Virginia, who is on the PBDB's executive council. "I think it's one of the best innovations that palaeontology has had in a very long time," says Jonathan Tennant, a palaeontologist at Imperial College London. He uses the PBDB every day and thinks that text mining could serve as a useful way to collect a large amount of data for later manual inspection — but not as a full-on replacement for human-curated databases such as the PBDB. "I don't see machines replacing humans. I think it's important that we retain the human aspect of the analytics," he says.

John Alroy, a palaeontologist at Macquarie University in Sydney, Australia, who co-founded the PBDB but is no longer affiliated with it, is less bullish on text mining. He says that DeepDive tends to overestimate the period during which species existed, leading to mistaken estimates of species diversity. He sees speed as the only advantage of text mining. "But there is no need to be fast in this case because the PBDB is already extremely comprehensive, so pretty much any question you might want to ask can already be answered with it. That explains why it has generated so many publications," Alroy says.

TEXT-MINING FRUSTRATIONS

Peters says that he will be using the computer-generated database as a supplement to the human-generated PBDB but adds that, for now, the limited number of documents it works from make it of little added use to palaeontologists. He wanted to let PaleoDeepDive loose on a bigger set of documents, but he did not have legal permission. As other text miners have discovered, many publishers of paywalled articles are cautious about allowing researchers to text mine their papers, even if they have lawful access to the literature; publishers tend to place limits on how the results of text mining can be published and reused, and often limit the number of papers a scientist can download at any one time (see *Nature* **483,** 134–135; 2012). "I can't think of any single palaeontologist who has 40,000 papers in their own stash, at least legally acquired," says Tennant.

Peters and Livny spent months brokering a deal with one scientific publisher, Elsevier, to gain access to thousands of papers. "This is just the frustrating reality of things right now: advanced capabilities in machine reading and learning are coming out, and the bottleneck in progress is now getting documents together in one place for analysis," Peters says. He and his colleagues are working on amassing and parsing documents to feed into PaleoDeepDive and a related software tool for the geosciences literature called GeoDeepDive. Ré, meanwhile, is working with experts in other fields to apply DeepDive to drug development, genomics and human trafficking.

Many palaeontologists also want to make it easier to find the data buried in their papers, so they are calling for research papers to be described more systematically in the future. "If we start having publication where everything is standard, then it will be much easier to read and process that data," says Tennant. Uhen adds, "I think there's a sort of cultural shift going on in palaeontology, where people are interested in data aggregation, and getting more insistent about being crystal clear about where you're finding your fossils."

Despite these challenges, many palaeontologists see text mining as the way forward for their field. "It's a huge waste of time for grad students and postdocs to manually re-enter already published information into a structured database," says Ross Mounce, a palaeontologist at the Natural History Museum in London who is using text mining to track how the museum's 80-million-specimen collection is used in research papers. Peters hopes that efforts such as PaleoDeepDive will allow him and his colleagues more time to generate data instead of spending their days organizing data they already have. "I see these machine reading systems as liberating our efforts a little bit, and shifting our work back into the field and back into the museums." ■

Ewen Callaway *writes for* Nature *from London.*

CORRECTION

The Toolbox article 'How to catch a cloud' (*Nature* **522,** 115–116; 2015) gave the wrong location for the Texas Advanced Computing Center — it is in Austin not San Antonio.