TOOLBOX

MINING THE SECRETS OF COLLEGE SYLLABUSES

The creators of the Open Syllabus Project hope that sharing data can both improve and reward teaching.

ILLUSTRATION BY THE PROJECT TWINS

BY ANNA NOWOGRODZKI

Despite a growing movement to glean insights from scholarly materials that are available online — from articles and data sets to conference presentations and lectures — one kind of academic document remains little examined. And that is the syllabus: a document that lays out the reading materials, topics and expectations of college courses.

That, at least, was the case until January this year, when data scientists, sociologists and digital-humanities researchers at Columbia University in New York City launched a tool called the Open Syllabus Explorer. This integrates more than 1 million publicly available syllabuses and lays open their data in a conveniently searchable format. A version containing three times as many syllabuses is scheduled to launch in January 2017.

The team behind the tool, the Open Syllabus Project (OSP), hope to nudge universities towards making more syllabuses public. They argue that doing so could aid textbook authors, instructors and course developers, and would reward the design of effective teaching materials, which is largely overlooked by conventional measures of academic effort.

"Syllabuses are among the most important documents written by scholars which are not yet widely shared, and they ought to be," says Peter Suber, director of the Harvard Open Access Project and the Harvard Office for Scholarly Communication in Cambridge, Massachusetts, who serves on the OSP advisory board. "They reflect serious scholarly judgements about what's worth teaching."

Such judgements can be welcome news to textbook authors. Stuart Russell, a computer scientist at the University of California, Berkeley, didn't realize until *Nature* interviewed him for this article that his 1995 book *Artificial Intelligence* (Prentice Hall), co-authored with Peter Norvig, was the most highly assigned text in the field of computer science. "I was definitely surprised," he says.

Beyond stoking professional pride, such information could strengthen tenure and promotion packages. Authoring a textbook, no matter how useful and informative it might be, generally yields few citations in scholarly papers, so its academic impact is likely to be low. The OSP could help to shift the balance. "We're at a point in time when I think faculty have to take more ownership of their whole record of scholarship, of impact, of influence," says Amy Brand, director of the MIT Press. Hard data on syllabus usage, she says, could empower faculty members "to tell their own story about what their work is doing in the world". ▶

At present, Open Syllabus Explorer searches more than 1 million syllabuses dating back to 2000, cross-referenced with 20 million texts, to produce data on how often a text is taught. Users can search those data by author, title, institution and academic discipline. The tool also reports which textbooks are commonly used together, and ranks each text on how frequently it is taught (see 'Top texts').

An updated version, due to become available on 21 January 2017, the Explorer's first anniversary, will feature 3 million syllabuses cross-referenced with about 150 million texts; these will include titles from the arXiv preprint server, CrossRef and the Virtual International Authority File — which links together identical bibliographic records from different national-library catalogues. The update will include new search options, such as the ability to search by date or type of institution, says Joe Karaganis, the OSP's project director. The new version will also incorporate better Canadian and UK data, information about where to find materials and, eventually, full-text syllabuses, if the authors have given permission to reproduce them.

"We have some big ambitions," Karaganis says. "All the techniques are very crude at present but they're all improvable, and the data science is only getting better."

FISHING FOR CITATIONS

The OSP is based at the American Assembly, a public-policy institute at Columbia University, and is funded by the Sloan Foundation and the Arcadia Fund. It was inspired by a search engine called Syllabus Finder, which scraped the public web for syllabuses from 2002 (the year it was built) until 2009. That tool was created by Dan Cohen, then a historian at George Mason University in Fairfax, Virginia, who is now executive director of the Digital Public Library of America. It amassed what Cohen says was then the largest collection of syllabuses ever assembled, comprising about 1 million documents. He released the URLs as a database in 2011.

Unlike the OSP, Cohen's tool provided links to the full text of each syllabus. But it included only courses run up to 2009, when he had to retire the tool because of changes to Google's programming interface — a move that vexed Cohen's colleagues, including his wife, an early-childhood educator. "I still get e-mails begging me to turn the Syllabus Finder back on," he says.

When the OSP began in 2014, the team built tools to scrape the public Internet — including the links used by Cohen, who had lost a portion of the data owing to a coding error. But, as Cohen was, the team is limited to publicly accessible syllabuses: about 6 million of an estimated 80 million to 120 million syllabuses in the United States alone, by Karaganis's reckoning. Syllabuses sealed behind the walls of private course-management software, such as Blackboard, remain out of reach. "Columbia, for example, is sitting on 80,000 syllabuses from the last 12 or 13 years," says Karaganis. "A large state school could have two, three times that."

TOP TEXTS

The 5 most-taught scientific texts according to OSP.

Textbook (author)	Syllabuses
Biology: Concepts and Connections (N. A. Campbell *et al.*)	2,196
Fundamentals of Anatomy and Physiology (F. Martini *et al.*)	752
Chemistry (R. Chang)	612
Human Anatomy & Physiology (E. N. Marieb and K. Hoehn)	605
Human Anatomy (E. N. Marieb *et al.*)	591

Data filtered by fields: Astronomy and Astrophysics, Biology, Chemistry, Computer Science, Earth Sciences, Engineering, Psychology, Sociology.

SOURCE: OSP

The OSP team then had to build tools to extract what those syllabuses contained. Citations, for instance, had no consistent structure, says David McClure, the project's technical director. The tool searched for titles by cross-referencing each syllabus against a database of 20 million titles — 11 million from Harvard LibraryCloud and 9 million from JSTOR. A matching title and author counted as a citation. "We built in different techniques for allowing fuzziness, like allowing the word 'by' in between the author and title," says McClure.

A NEW METRIC

The OSP distils those data down to a single metric called the teaching score, which indicates how often a text is assigned in syllabuses. It can take any value from 1 (rarely taught) to 100 (frequently taught).

According to Suber, teaching scores are an alternative to conventional metrics of scholarly impact. They reflect the burgeoning 'alternative metrics' ethos, which aims to quantify the whole of a person's research output. "I think this teaching score can take part in the new alt-metrics movement and give us a more sensitive measurement of the impact of texts," he says.

Already, a handful of researchers and universities are using the data to do just that. The University of Kentucky in Lexington issued a press release when it discovered that a paper by Edward Morris, one of its faculty members, ranked 46 out of 13,225 sociology-related texts. It now ranks 371 out of 53,177, and Morris plans to use the figure to support his promotion to full professor.

US universities aren't the only ones paying attention. Most of the roughly 1,000 visits to the OSP each day are from the United States, says Karaganis, but significant traffic comes from Ukraine, Russia and Egypt as well.

Other researchers have used the data to compile lists of widely taught graphic novels and comics, for instance, or to quantify the fraction of frequently taught sociology texts authored by women. Melanie Martin, a post-doc at Yale University in New Haven, Connecticut, used the Syllabus Explorer to identify the most commonly taught texts in her field, evolutionary anthropology. But, because there is no way to search the database by subfield — for instance, limiting biology results to such subdisciplines as neuroscience or genomics — she had to scan the 16,000 anthropology titles manually. "Without better filtering, I think it's limited," she says.

BUILDING ON PEER EXPERTISE

Another possible application of OSP data involves course design. By enabling faculty members — particularly junior ones — to build on the knowledge of their peers, the OSP could help them to teach more creatively, such as by identifying new ways to present teaching material. "This could go a long way to improving the quality of instruction," says Russell. It would also improve efficiency, leaving faculty members more time for other activities such as research and mentoring.

However, it is important not to over-interpret the data, says Lisa Janicke Hinchliffe, a specialist in information literacy at the University of Illinois at Urbana–Champaign. The project's sample set might not be a good proxy for all syllabuses, even at a particular institution. For instance, the second most-assigned text at Harvard, according to the Explorer, is 'Letter from Birmingham Jail' by Martin Luther King Jr. But about 80% of the OSP's Harvard syllabuses come from the John F. Kennedy School of Government, Karaganis says (although the OSP doesn't publicly list its sources in this much detail). So it's not possible to conclude how popular this text is at Harvard overall.

For Hinchliffe, the value of the OSP lies in its ability to reveal the breadth of resources that instructors use. "I don't need a definitive 'These are the top-six taught books,'" she says. "I want to see the variety."

Such information could go a long way towards simplifying course design, a notoriously time-consuming process. Just ask Suber, who has been teaching philosophy for 21 years. "Whenever I knew a new course was coming, I would try to start preparing it at least a year in advance," he says. "Writing 40 lectures is a huge job; it's harder than writing a book."

The OSP's data could ease that burden. Plus, says Suber, the data are fun to explore, sometimes revealing unexpected pairings. His legal philosophy text, *The Case of the Speluncean Explorers* (Routledge, 1998), for instance, has been taught alongside Sappho's lyric poetry. "There are partners or juxtapositions that I never would have guessed," he says. ■

CORRECTION

The Toolbox article 'Democratic databases: science on GitHub' (*Nature* **538,** 127–128; 2016) misstated how the Git software records changes in files. It does in fact maintain multiple versions of the files.