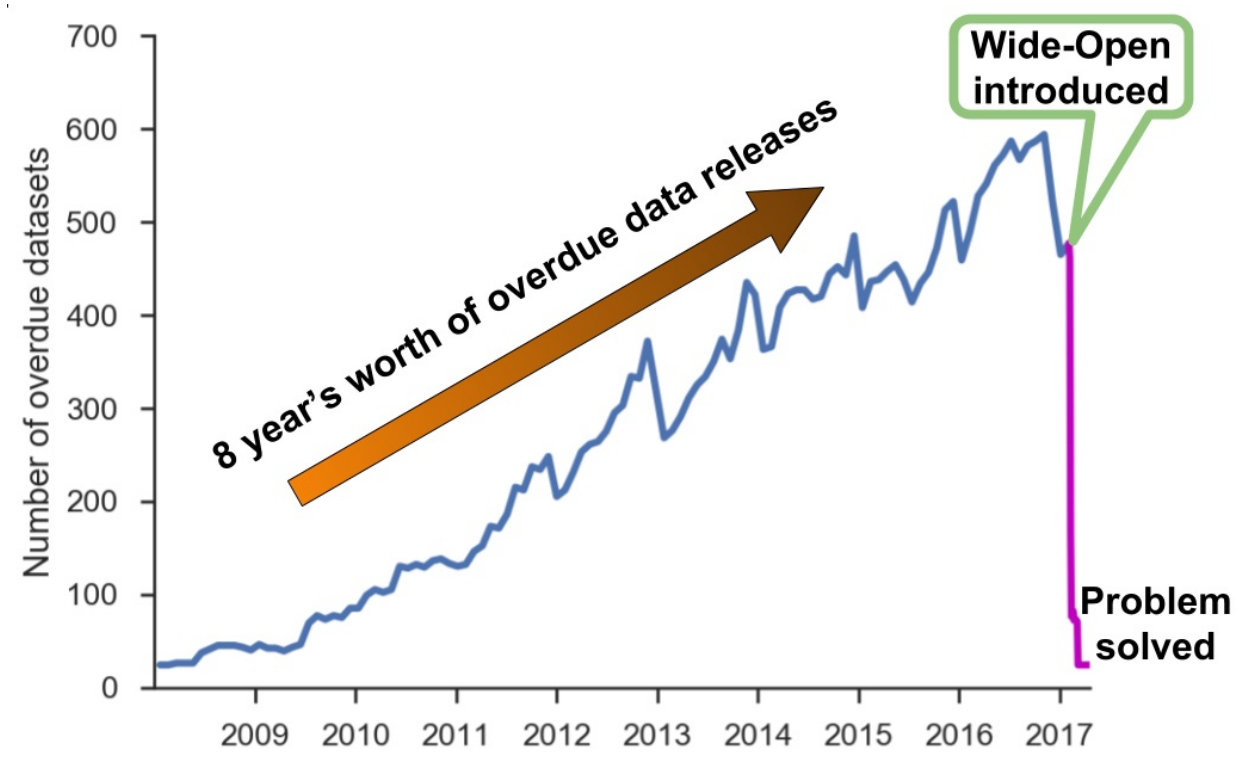
Wide-Open accelerates release of scientific data by identifying overdue datasets

June 8 2017



The use of Wide-Open on the Gene Expression Omnibus (GEO) led to the dramatic drop of overdue datasets, with 400 datasets released within the first week. Credit: Maxim Grechkin, Roli Roberts

Advances in genetic sequencing and other technologies have led to an explosion of biological data, and decades of openness (both spontaneous and enforced) mean that scientists routinely deposit data in online

repositories. But researchers are only human and may forget to tell a repository to release the data when a paper is published.

A new tool, developed by University of Washington and Microsoft researchers Maxim Grechkin, Hoifung Poon and Bill Howe, and described in a Community Page article publishing June 8 in the open access journal *PLOS Biology*, hopes to get around this problem and help advance open science by automatically detecting datasets that are overdue for publication.

Open data is a vital pillar of open science, enabling other researchers to reproduce results and use the same datasets to produce novel discoveries. While many scientific journals now require published authors to make the data underlying their findings publicly available, these policies often go unenforced. The challenge is substantial - the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus repository (GEO) alone contains 80,985 public datasets, spanning hundreds of tissue types in thousands of organisms - and the rapid growth in data makes it difficult for journals or data repositories to "police" whether datasets that should be made publicly available actually are.

The Wide-Open system is available under an open source license on GitHub; it uses text mining to identify dataset references in published scientific articles that should be publicly accessible, and then parses query results from repositories to determine if those datasets remain private.

Grechkin and his team tested their tool on two popular data repositories maintained by the NCBI - GEO and the Sequence Read Archive (SRA) . Wide-Open identified a large number of overdue datasets, which spurred repository administrators to respond by releasing 400 datasets in one week.

"We developed a simple yet effective system that has already helped make hundreds of datasets public," said lead author Maxim Grechkin. "Having an impartial and automated system enforce open data policies can help level the playing field among scientists and generate new opportunities for discovery."

More information: Grechkin M, Poon H, Howe B (2017) Wide-Open: Accelerating public data release by automating detection of overdue datasets. *PLoS Biol* 15(6): e2002477. doi.org/10.1371/journal.pbio.2002477