NATURE | NEWS

Text-mining tool seeks out ‘hidden data’

Wide-Open checks that the data sets underlying published studies are made freely available.

Dalmeet Singh Chawla

08 June 2017

Forgotten to free your data? A tool called Wide-Open can search out instances of locked online research data sets that are supposed to be public — and it has already flagged hundreds of such instances in genetics research, according to a study[1] published in *PLoS Biology* on 8 June.

Scientists often post ‘hidden’ data online in repositories while their related studies are going through peer review, intending to make data sets public later.

Two popular repositories that offer researchers the option to keep genetics data hidden, for example, are the Gene Expression Omnibus (GEO) and the Sequence Read Archive (SRA), both run by the US National Center for Biotechnology Information. Both sites require data sets to be made open when papers are published. But in practice, scientists often forget to do this, says Maxim Grechkin, a computer scientist at the University of Washington in Seattle.

So Grechkin and his collaborators developed Wide-Open to find non-open data, focusing on GEO and SRA. The tool scans papers for mentions of unique data-set identifier codes (called accession codes) that use the GEO’s or SRA’s code format. The tool could be tweaked to query other repositories as well, notes Grechkin.

Related stories

- Our path to better science in less time using open data science tools
- Open data: towards full transparency
- Open data: curation is under-resourced

More related stories

Related stories

- Our path to better science in less time using open data science tools
- Open data: towards full transparency
- Open data: curation is under-resourced

More related stories

Once it identifies a valid code, Wide-Open trawls the relevant repository to find out whether the data set is public. It notes as “overdue” any data set that isn’t available, but should be.

Small minority

Grechkin's team ran Wide-Open on roughly 1.5 million papers in PubMed Central, an open-access database of biomedical studies. The tool identified 473 data sets missing in GEO, and 84 in SRA.

The team alerted the repositories of its finds. By the time the GEO staff checked, they found that 27 of the flagged data sets were already live — representing a short lag in posting for some publications — and they released 429 data sets that were overdue, says Tanya Barrett, GEO's group leader of curation. The remaining cases either cited incorrect codes or mentioned data sets that couldn't be made open because of privacy concerns or incomplete data submission.

"We are happy to add Wide-Open to the tools that we use," says Barrett.

Most researchers using GEO do release their data on publication, she says. GEO staff also routinely use alerts from PubMed Central and Google Scholar to keep track of published papers, she adds, but because it's a manual process some are missed.

The researchers say in their paper that they plan to work with SRA staff to ensure the release of their hidden data sets as well.

Wide-Open now trawls GEO and SRA roughly every month, and automatically updates its site with papers whose data are missing.

Bigger problem

"In my experience, people putting their data onto GEO or SRA intend it to be made public at some point," says Timothy Vines, a former managing editor of *Molecular Ecology*, who has written about the importance of data sharing.

The bigger problem is that many researchers still aren't making their data public. "Most researchers I know don't even bother to deposit data anywhere, let alone deposit and then not share," says Chris Hartgerink, a statistician at Tilburg University in the Netherlands.

Hartgerink adds that Wide-Open could be tweaked to monitor clinical-trial data sets that have clear identifiers. But it would be more difficult to apply to fields such as the social sciences, which doesn't widely use accession codes, making data sets difficult to track.

A key limitation of Wide-Open is that it can currently scan only open-access papers because the team

has not yet secured legal permissions to scan subscription content. Grechkin says that they are liaising with subscription publishers to ask for permission.

Ultimately, Grechkin thinks journals should share some of the responsibility for making sure that data sets are openly available. In future, Wide-Open might also start ranking journals on the basis of their data-sharing practices, he says.

Nature doi:10.1038/nature.2017.22132

References

1. Grechkin, M., Poon, H. & Howe, B. *PLoS Biol.* http://dx.doi.org/10.1371/journal.pbio.2002477 (2017).