SourceData is making data discoverable
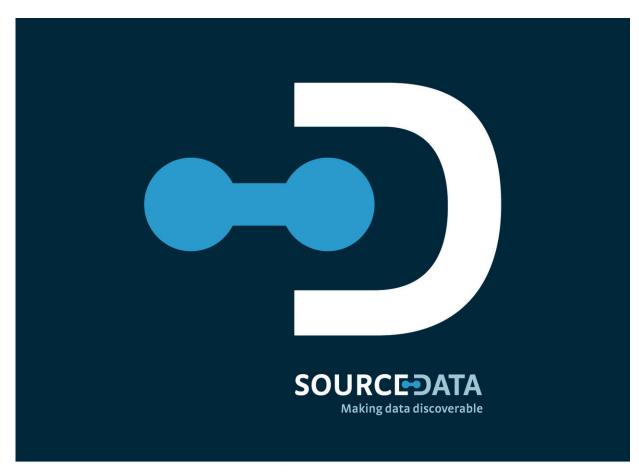
November 1 2017

The SourceData logo. Credit: EMBO

SourceData from EMBO is an award-winning open platform that allows

researchers and publishers to share figures and their underlying data in a machine-readable, searchable format, making research papers discoverable based on their data content. As highlighted in today's paper in *Nature Methods*, SourceData offers a novel method to describe research data and a suite of tools to generate, validate and use this information, providing scientists with an efficient method to find and re-use published results.

"In the biological sciences, most of the data produced by researchers is published in the form of figures. Figures are the heart of a scientific paper. However, the search tools used to find published papers are usually limited to keyword-based text searches that exclude figure contents," SourceData project leader Thomas Lemberger of EMBO explains. This can result in relevant data being missed from search results due to the lack of a consistent method for representing figures in a searchable form.

With SourceData, a machine-readable description of each figure is generated and stored in a structured database. The biological entities represented in the figure, such as genes, proteins or molecules, are linked to standardized taxonomies to avoid naming ambiguity. This means that each occurrence of a certain biological entity in a figure or result set can be quickly found within the SourceData database. SourceData also stores the direction of the relationship between entities: whether they were manipulated or observed, allowing very specific searches based on the experimental design.

Paper co-author Robin Liechti from the SIB Swiss Institute of Bioinformatics (SIB), explains "SourceData links figures to other related figures across papers and journals to build a searchable knowledge graph, which is quality-controlled by expert curators. Readers of scientific articles can use this to find the data they need in a much more efficient way."

SourceData provides a suite of applications including SmartFigures: enhanced figures containing links to related results and data that can be embedded in online publications, DataSearch: a search engine that finds published figures based on their data content, and MetaFig: a curation interface that offers computer-assisted importing of new figures into the SourceData format

The SourceData platform is currently in active development, with EMBO and SIB engaging with academic publishers to establish an open and effective standard for the discovery and reuse of figures and data.

Provided by European Molecular Biology Organization