# A Review of Data Repositories for the Long Tail of Computer Science

*by Pachev Joseph | Victor Potapenko | Naphtali Rishe | Oliver Ullrich*

*Abstract Id:* ***265*** *Submitted:* ***March 11, 2017*** *Event:* ***Conference for Undergraduate Research at FIU 2017***
*Topic:* ***Computer Science***

Computer scientists undertaking research often find themselves struggling to find or generate the right data to formulate or test hypotheses, validate models, or test algorithms – which in turn leads to greater time and effort allocated to searching for or producing data, rather than using it to perform scientific research. This data barrier is a significant impediment to scientific progress, because most of the progress occurs in the long tail of computer science: in many smaller, silo-like laboratories that rarely share data.

This paper surveys nineteen existing data repositories based on their respective feature sets. Out of the nineteen reviewed repositories, only six were found to have feature sets differentiating from standard digital libraries. Most distinguished features were found to be in-browser data preview and manipulation, community-based data curation, data access via APIs, licensing of data, real-time data publishing, data versioning, and virtual workspaces, where users can create and manage research projects with fully traceable activity logs and interlinked artifacts such as data, papers, and code. These features were found to be scattered among various repositories, and not one single repository platform was found to provide them all as a complete package. None of the surveyed repositories offered features that would allow for purchase or sale of data. Studies based on a representative sample show that providers of data in the scientific domain offer it for free approximately 80% of the time.

The survey showed that there is no single data repository that provides the right combination of features and tools geared towards the long tail of computer science. Moreover, none of the surveyed repositories provide the monetary motivational component for computer scientists to share data. We contend that a true data marketplace that implements a particular set of features would lower data-related barriers on the path to scientific progress.