Evaluation of a scientific data search infrastructure

Alexandru Iulian Orhean\textsuperscript{1} | Anna Giannakou\textsuperscript{2} | Katie Antypas\textsuperscript{3} | Ioan Raicu\textsuperscript{1} | Lavanya Ramakrishnan\textsuperscript{2}

\textsuperscript{1}College of Computing, Illinois Institute of Technology, Chicago, Illinois, USA
\textsuperscript{2}Scientific Data Division, Lawrence Berkeley National Lab, Berkeley, California, USA
\textsuperscript{3}NERSC, Lawrence Berkeley National Lab, Berkeley, California, USA

**Summary**

The ability to search over large scientific datasets has become crucial to next-generation scientific discoveries as data generated from scientific facilities grow dramatically. In previous work, we developed and deployed ScienceSearch, a search infrastructure for scientific data which uses machine learning to automate metadata creation. Our current deployment is deployed atop a container based platform at a HPC center. In this article, we present an evaluation and discuss our experiences with the ScienceSearch infrastructure. Specifically, we present a performance evaluation of the infrastructure and discuss how it can be improved to support large-scale data search.