Fermilab

Managed by Fermi Research Alliance, LLC for the U.S. Department of Energy Office of Science

SIMONS FOUNDATION DATA MANAGEMENT

Cooperative Research and Development Final Report

CRADA Number: FRA-2016-0010

Fermilab Technical Contact: Robert Illingworth

Summary Report
26 January 2017

Fermi National Accelerator Laboratory / Kirk and Pine Street / P.O. Box 500 / Batavia, IL 60510 / 630.840.3000 / www.fnal.gov / fermilab@fnal.gov
Managed by Fermi Research Alliance, LLC for the U.S. Department of Energy Office of Science

NOTICE

In accordance with Requirements set forth in Article X.A(2) of the CRADA document, this document is the final CRADA report, including a list of Subject Inventions, to be forwarded to the Office of Science and Technical Information as part of the commitment to the public to demonstrate results of federally funded research.

CRADA number: FRA 2016-0010

CRADA Title: SIMONS FOUNDATION DATA MANAGEMENT

Parties to the Agreement: Simons Foundation and and Fermi Research Alliance

Abstract of CRADA work:
The Simons Foundation is a non-profit organization that funds scientific research. Their Autism Research is generating significant amounts of genomic data, which they need to both archive and make available for researchers around the globe. The Foundation already stores data at Fermilab, as per agreement FRA-2015-0006. With this CRADA, the Fermilab Scientific Computing Division has developed a prototype solution for the data management needs of the Foundation, modifying services developed for High Energy Physics. The tools allow researchers to select datasets from a metadata catalogue of Foundation data, transfer selected files or datasets from Fermilab to remote storage, and enable the Foundation to manage access and authorization of the files. The prototypical solutions developed through this CRADA is considered the Phase I of a body of work that can be expanded and transitioned to production quality in a subsequent phase II, to be determined with the Foundation.

Summary of Research Results:
The following deliverables have been defined in the CRADA Statement of Work and have been completed by the close-out date.

1. A metadata catalogue for storing information about the content and locations of Simons Foundation files for the Diversity Project data. This catalogue is based on the existing Fermilab SAM data management system with any High Energy Physics specific features removed. Fermilab has adapted and deployed a SAM metadata catalogue dedicated to the Simons Foundation. The catalogue is accessible through the SAM Client tools.
1.1. The catalogue supports generic key/value fields where the key names are specified by the Simons Foundation. It was designed to include metadata specific to the domain of Autism research.
1.2. It is possible to store one or more checksums for each file. The system supports checks of md5 checksums after the files are transferred from the Fermilab Active Archive Facility to local storage for data analysis. The integrity check is performed by the SAM Client tools, comparing the checksum on record in the SAM system and the one calculated locally.

1.3. The catalogue stores provenance information for each file – the application used to create it and the identity of any other catalogued files from which it was derived.
1.4. Client interfaces allows adding metadata to the catalogue, modifying it, and returning all metadata for any given file to the end user
2. An interface to perform flexible search queries for files based on their metadata and to return an access URL for each file. The SAM Client tools allow the transferring of the files identified through the metadata queries.
3. The capability to create datasets based on metadata queries and store them for later retrieval. The SAM Client tools support the creation of named datasets based on metadata queries. The dataset names are used to initiate data transfers.
4. Authentication based on CILogon certificates, including OpenID based access using Google accounts. The email addresses of all users requesting access is recorded.
5. The tools to perform end-to-end integrity checking for copies of the data using the checksums stored in the catalogue. This was a feature added to the SAM Client tools to facilitate usage by the Simons Foundation. See description at point 1.2 above.
6. Design and deliver an initial version of a network probe tool that can diagnose network connectivity and performance issues at a client site. Fermilab has developed a tool that tests the network connectivity to the major components of the SAM system and the Active Archival Facility. It presents error conditions and possible remediation actions
7. A client package for Red Hat Linux-compatible systems. This package is the SAM Client tools. In addition, the SAM services provide web interfaces to monitor the status of the data transfers.

Subject Inventions listing: None

Report Date: 01/26/2017

Technical Contact at Fermilab: Robert Illingworth

This document contains NO confidential, protectable or proprietary information.