CIF21 DIBBs: Building a Modular Cyber-Platform for Systematic Collection, Curation, and Preservation of Large Engineering and Science Data – A Pilot Demonstration Project

Award # 1443027

PI: Santiago Pujol, co-PIs: Ann Christine Catlin, Michael McLennan, Chungwook Sim, and Lisa Zilinsksi

Our goal was to provide a robust, scalable, easy-to-use data platform for organizing, preserving and publishing the full products of scientific and engineering research, so they can be discovered, explored, vetted, and reused by research communities worldwide. The result is DataHub https://datacenterhub.org

Challenge: Organize research datasets in a meaningful way that provides deeper insight into the investigation process, and support the interactive exploration of publicly shared datasets through intuitive and innovative interfaces that can combine and interpret heterogeneous data types.

Solution:

We introduced a new concept for the preservation of scientific research data. Datasets are organized by experiments, with a simple common structure for metadata, file collections, and structured attribute data. Experiment files are classified by use and type to help users better understand experiment content. User-defined attributes describing experiment properties and results are stored as structured data to bring key information "to the surface".

We developed an extensible, interactive web-based "dataview" technology that presents the content of datasets in tabular form, with columns that interpret structured data and file collections by type to support advanced search, filtering, previews, navigation, comparison, and launching of exploration tools. Media columns launch interactive galleries where users navigate, view, and play media files with corresponding annotations. Geospatial data launches maps with markers that link to information or files corresponding to that location. Text columns support pattern filtering and numeric columns support arithmetic filtering. Columns can launch new views that drill down to lower levels of detail for data in any column or row. Dataviews are based on "data definitions" that can be pre-defined or auto-generated. Data definitions describe content and presentation format, and can include column-based computation and queries. The extensible nature of this technology allows us to continue adding new exploration features to DataHub views. Our platform helps researchers discover and view more information more quickly, and helps research organizations arrive at standards for reporting and sharing data

Ongoing Challenge: Promote public sharing of the full products of scientific and engineering research.

We have made excellent progress in the civil engineering community. Nearly 6000 experiments have been preserved for public discovery, and more than 70,000 researchers and practitioners have created or used DataHub datasets. Our efforts are ongoing. Public data-sharing should be the natural conclusion of all science and engineering research investigations, and an infrastructure that provides easy contribution and exploration is essential. DataHub has taken an important step forward in this direction.

The Future of DataHub

Through our discussions with researchers engaged in data-driven research, we recognized the enormous contribution our infrastructure could make in supporting their investigation processes. We want to advance DataHub to enable 1) cataloguing and annotation of input to computational software and analytical tools, 2) execution of computational software on destination nodes, 3) support for multi-dimensional, hierarchical attribute data, 4) support for interactive analysis tools that operate on attribute data, 5) automatic preservation, cataloguing, and annotation of output from execution of computational software and analysis tools, and 6) tracking of research activities to ensure traceability and reproducibility. Data, computation, and research activities would be shared and supported through a single "dataset dashboard" on DataHub.