Fully Integrating Data with Compute Workflows: A Platform to Better Serve Scientific Research

[*] Ann Christline Catlin
Reearch Computing
Purdue University
West Lafayette, IN USA
acc@purdue.edu

Chandima HewaNadungodage
Steven Clark
Research Computing
Purdue University
West Lafayette, IN USA
chewanad@purdue.edu
clarks@purdue.edu

Sumudinie Fernando
Guneshi Wickramaarachchi
Research Computing
Purdue University
West Lafayette, IN USA
swfernan@purdue.edu
gwickram@purdue.edu

Andres Bejarano
Parameswaran Desigavinayagam
Omkar Patil
Department of Computer Science
Purdue University
West Lafayette IN USA
abejara@purdue.edu
pdesigav@purdue.edu
patilo@purdue.edu

***Abstract*—The NSF Office of Advanced Cyberinfrastructure has recognized the emerging and evolving need for platforms that fully integrate data and computing workflows, and is calling for research to deliver systems that provide a full spectrum of data services and also offer a coherent coupling with computing software. The Digital Environment to Enable Data-driven Science (DEEDS) project has created a cross-domain, self-serve platform for data and computing that supports the entire end-to-end research investigation process. DEEDS offers interactive interfaces to 1) collect, manage, and explore data, 2) define and launch tools, 3) track computational workflows, and 4) access toolkits for ad hoc analytics. All interfaces are available from a single dashboard so that the workflow between data and tools is smooth and intuitive. In this paper, we describe DEEDS innovations for handling data and computational workflows, and we present the use cases from four science domains that defined features, services, and usability requirements for DEEDS.**

Keywords—data collection, data sharing, interactive data exploration, computing services, computational workflow

I. Introduction

Research cyberinfrastructures are most often created either to support data preservation and sharing or to provide computing services. Even those infrastructures that support both data and computation do not effectively integrate them – computing services are not directly connected to the interactive interfaces that manage the data used or generated by computational tools, and valuable metadata describing relationships between input, computation, and output cannot be captured. The NSF Office of Advanced Cyberinfrastructure has recognized the emerging and evolving need for platforms that fully integrate data and computing, and is calling for research to deliver production-level systems that provide a full spectrum of data services and also offer a coherent coupling with computational software.

The Digital Environment to Enable Data-driven Science (DEEDS) project offers a cross-domain, self-serve data and computing platform that supports the entire end-to-end research investigation process. Our platform makes it easy for research groups to define and organize their research activities as shared DEEDS datasets. Researchers can upload, annotate, and manage data; define tools and computing resources; and connect data to computational and statistical tools for execution. DEEDS automatically captures, uploads, and classifies output, and also tracks and annotates research workflows to support traceability of results. Our platform offers innovative technologies for handling research data stored as file collections or as complex hierarchical data tables, and it integrates ad hoc analytics and visualization of data as part of dataset support. When the investigation is complete, DEEDS makes the publication of data, algorithms, and workflows seamless. Research communities can explore published datasets with interactive viewers that interpret data by type and use for advanced navigation and search. DEEDS is a novel and effective platform that provides reliable, systematic, and user-friendly services to preserve and share data, tools, and computational workflows.

II. Platform Innovation to Support Data-Driven Science

A. Platform Architecture

DEEDS is built on top of the hubzero™ cyberinfrastructure [1] and is an extension of DataHub [2], a platform designed to preserve and publish scientific data for discovery and exploration. DEEDS is transforming DataHub from a data-focused platform to a full research support environment. A number of capabilities from existing DataHub components were re-used, but the DEEDS platform was re-designed and re-implemented from the ground up to ensure smooth and intuitive transitions between data and computing services. DEEDS is implemented using the LAMP stack (Linux, Apache, MySQL, PHP) and Javascript/JQuery for the front-end. Fig. 1 shows the high-level architecture of the DEEDS platform.

B. Dataset Organization and the DEEDS Dashboard

A DEEDS dataset is organized as a collection of "cases". Cases represent experiments, study units, sites, specimens, or research activities that define and clarify how the research investigation was carried out. This organizational structure makes it easier for researchers to understand, interpret, and use a dataset, since data and files are more directly connected to the activities that produced them. Cases correspond to observations, measurements, input/output files, figures, reports, properties,

This material is based upon work supported by the National Science Foundation **CIF21 DIBBs: EI: Award #1724728**.

Presented at Gateways 2018, University of Texas, Austin, TX, September 25–27, 2018

methods, outcomes, and any other data that are collected or generated throughout the research investigation.

The DEEDS dashboard provides a user-friendly interface to upload, update, annotate, and share data. Computational tools and resources used during the investigation are defined and launched from that same dashboard. It also displays computational workflows (user-selected input, tools, execution information, and generated output) which are automatically captured by DEEDS and stored to the dataset. The dashboard consists of five sections; Cases, Files, Data, Tools, and Analytics. Each section provides key functionality to support different stages of the research investigation process.

Cases: Define cases (or study units) and associate them with metadata such as description, keywords, bibliographic, spatial, and temporal information. Cases can be defined/updated one at a time using the interactive web interface, or they can be uploaded/updated in bulk using the csv file upload feature.

Files: Upload and associate file collections to cases. Files are organized into categories for ease of use and discovery. There are four system-defined categories: Reports, Data, Media, and Figures, and users can define custom categories (e.g., device-generated files). At upload, DEEDS captures metadata that includes size, timestamp, and username. Thumbnails and previews for media, figures, and reports are also generated. Users can classify and annotate files to make it easy to search, select, use, and explore them. A shared upload feature helps users associate common files with many (or all) cases. Interactive interfaces are built into DEEDS to view, search, and explore file collections.

Data: Define complex structured data models describing measurements, observations, outcomes, and other data collected throughout the research investigation. Our unique "spreadsheets of spreadsheets" approach for defining, uploading, updating, and viewing data models is described in Section C.

Tools: Define and upload computational software, select execution resources, launch tools, and track execution. Output generated by tool execution is automatically collected, annotated, and uploaded to the dataset cases. Tools and workflow tracking are described in Section D.

Analytics: Carry out ad hoc analysis of dataset data stored as Data or Files. Users will be able to build custom reports and graphs, compute statistics, audit for missing data, and compare information across cases. DEEDS Analytics is still under construction.

A notable capability of our DEEDS dataset is its full-featured support for the preservation, sharing, and interactive exploration of *both* file collections *and* complex structured data.

C. *Complex Structured Data*

Spreadsheets have long been the preferred way for researchers to collect measurements, observations, and other data. Data sharing usually has meant exchanging the spreadsheet files themselves or translating spreadsheets to web forms. With both methods, research teams lose the flexibility, efficiency, and familiarity of shared interactive spreadsheet operations. DEEDS has developed a novel and powerful interface that lets users define, upload, update, view, and explore spreadsheets as interactive data tables, with an interconnected multi-level spreadsheet capability that can support the complex data models needed for representing research data and measurements.

Using the dashboard Data interface, researchers upload their spreadsheets to build multi-dimensional data tables and define relationships among columns. A top-level spreadsheet is uploaded first (with or without data) and users can then connect sub-level spreadsheets to columns of the top-level spreadsheet. We call our approach "spreadsheets of spreadsheets" since a dataset can have any number of top-level data spreadsheets and any number of sub-level spreadsheets connected to parent columns up to a depth of five levels.

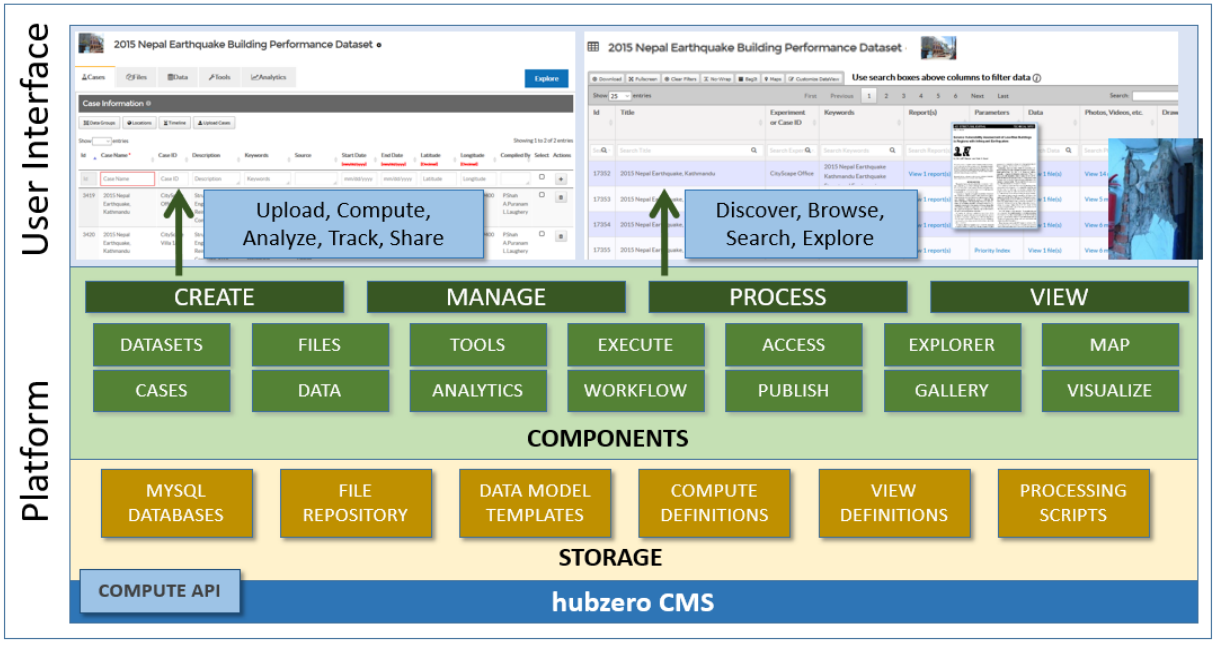

Fig. 1. DEEDS platform architecture.

DEEDS parses uploaded spreadsheets and creates the underlying definitions and database tables. Users can then click on interactive spreadsheet columns to define column metadata such as data type, label, units, description, formatting (width, color, font), and visibility. Users can also define computed columns based on data from other columns. Sub-level spreadsheets are defined using the data type 'spreadsheet' for the parent column, and a spreadsheet for that column can then be uploaded. Data values can be updated one at a time using the web interface or they can be updated in bulk using a csv file. The Data interface supports "drill-down links" to access, update, and explore data at all levels. Fig.2 shows how users can define data spreadsheets using the Data tab of the DEEDS dashboard.

In some instances, data collected for dataset cases do not fit to a single model, with some cases having entirely different data from other cases. To support diverse data among cases, users can divide cases into "data groups" and define different models (top- and sub-level spreadsheets) to represent the data for each group.

D. Computational Tools and Workflow Traccking

Files, Data, and Tools are managed from the same DEEDS dashboard so that interactions between them can be handled easily and smoothly. From the Tools interface, users can define, launch, and monitor their computational programs. Tool definitions include software name, version, and description; input/output (files, formats, command line arguments); execution resources; and access restrictions. For user written programs, source code is uploaded for compilation (if necessary) and installation on target destinations (local server, HPC clusters). Users can also execute licensed software or other open source software that are installed locally or on HPC clusters.

A web-based API for hubzero submit [3] was developed to handle installation, launch, and execution of tools invoked from the DEEDS dashboard. Submit rules are created by DEEDS based on tool definitions. To ensure a secure environment, an administrator inspects and approves user-written code and scripts. Once installed, all authorized users can click to launch dataset tools – first selecting cases and input files; specifying arguments; and directing which generated output should be returned to the dataset. Real-time execution tracking is displayed on the dashboard, and when execution ends, DEEDS automatically annotates, classifies, and uploads output to the selected cases.

DEEDS captures all computational workflows end-to-end. This is a key innovation that makes it possible to offer full traceability of research results, as well as enable more accurate interpretation, vetting, and re-use of those results. Captured workflows are displayed in the Tools tab of the dashboard, and include tool execution information (tool, version, arguments, resources, start time, trace); user-selected input; generated output; user data; and other metadata – with clickable links to browse file content and explore further workflow details.

E. Data Exploration

The customizable display of selected data from the MySQL database is a fundamental need across all areas of the DEEDS dashboard. Cases, File collections, multi-level Data spreadsheets, Tool definitions, workflows, Analytics results, and countless variations of these must be displayed for interactive exploration by users.

We have created a general, extensible, "data definition" language that accesses data stored in a MySQL database and presents them as interactive tabular "data views." Our language defines each column of the display by specifying the database table and the field for the source of the data, and applying display rules for types, properties, formats, and operations, which are given as arguments to the column. Extensible data typing allows us to attach applications and operations to the columns as needed, such as media viewers and drill-down links to the new data views. Data view layout can be pre-defined (e.g., Cases) or dynamically defined in real-time (e.g., Data). All data views have interactive exploration features that are automatically part of every tabular display, such as search, sort, filter, link, and download. Our language also provides type-specific exploration tools such as maps for spatial data, timelines for temporal data, and graphs for visualizing measurements and statistics.

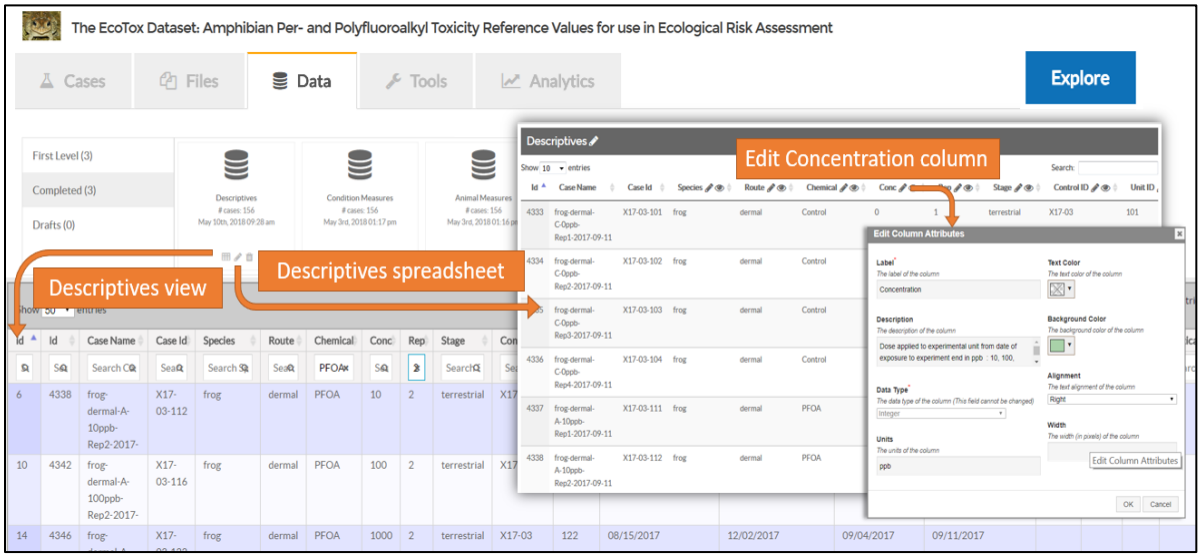

Fig. 2. Defining data spreadsheets for the EcoTox dataset using the DEEDS dashboard Data interface.

III. Use Cases

The DEEDS R&D team partnered with research groups from four science domains to jointly define requirements for user interfaces, features, functionality, and usability. Specific projects were used for characterizing types and forms of data; computational code and execution methods; data flows and computing workflows; and wishlists for ad hoc analytics such as visualization and customized reporting. Our partners are relying on DEEDS to preserve and share data and tools for their funded projects, and their datasets will be published for global discovery and holistic re-use when project investigations are complete. Use cases that guided and validated platform requirements are:

- Environmental Science [EcoTox]: Develop amphibian toxicity reference values for ecological risk assessment in contaminated sites. This research, funded by the Dept. of Defense, aids in making decisions on exposure mitigation and federal regulations for pollution control [4].
- Nutrition Science [Bone]: Study effects of blueberries added to regular diet on net bone calcium retention and on biochemical markers of bone metabolism in postmenopausal women. This research is funded by the NIH [5].
- Electrical Engineering [SolarPV]: Investigate efficiency of solar photovoltaic (PV) systems by coupling data for weather, manufacturer-specific PV technology, and solar farm health to determine efficiency degradation and predict system lifetime. This research is funded by NSF [6].
- Chemistry [Molecule]: Study spectroscopy, kinetics, and photochemistry of transient species in the gas phase to optimize molecular structure, predict properties, and provide reference data to guide experimental search for these species. This research is funded by NSF [7].

Table I summarizes the dataset composition for these use cases.

A. Research Computing: SolarPV & Molecule Datasets

These two research groups use DEEDS to 1) upload, classify, annotate, manage, and share collections of files and computational code, 2) launch and track tool executions on HPC clusters with user-selected input and auto-uploaded output, 3) share workflows, and 4) explore, analyze, and graph results.

The chemistry group wrote post-processing code to filter the large Gaussian output and generate customized reports for thermochemical, vibrational, orbital, and charge data. Spectrum images (Raman, IR, depolarization) are also generated. These valuable results are now automatically part of any Gaussian execution using DEEDS. The SolarPV group will use Analytics for function approximation to compare efficiency degradation for normalized data across solar farms.

B. Statistical Modeling: EcoTox & Bone Datasets

These two research groups use DEEDS to 1) define data models for collected measurements and observations using the "spreadsheets of spreadsheets" interface, 2) update, track, and view data as interactive data tables, 3) define and launch statistical modeling code, 4) share computational workflows, and 4) explore, analyze, and graph results.

The Ecotox group will use Analytics for statistical analysis of body burden, bioconcentration factor, and animal measures over time by chemical and dose. The Bone group will use Analytics for function approximation of calcium retention during treatment and recovery phases, and for exploration of relationships between retention and subject characteristics, polyphenol levels, and diet.

IV. Conclusions

Innovations in cyberinfrastructure impact all areas of scientific research. In particular, the evolution of data-focused platforms toward unified platforms that integrate data and computing is critical to the advancement of data-driven science. The concept, design, and implementation of DEEDS is answering the challenge for unified platforms. DEEDS is a self-serve, cross-domain platform that provides a full range of services for data preservation, sharing, exploration, and discovery, and also offers comprehensive support for computational tools and research workflows. Interactive data interfaces are directly connected to the launch, execution, tracking, and output management of tools for research computing and statistical modeling. The DEEDS dashboard provides a unified, user-friendly interface to create datasets, define research activities, collect files and structured data, execute computational software, return and annotate results, review computational workflows, and further analyze and compare the results. DEEDS viewers facilitate easy exploration and discovery of datasets and their heterogeneous content. We continue to improve and evolve DEEDS to support research projects across a broad spectrum of scientific disciplines.

TABLE I. Datasets for DEEDS Use Cases

Use Case	Dataset, Cases	File Collections and Data Models	Tools
SolarPV (Research Computing)	One dataset, each case represents one solar farm	File Collection: input: raw & curated weather data, PV module parameters output: efficiency data & circuit parameters over time, graphs for efficiency degradation	MATLAB research code with ongoing algorithm advances
Molecule (Research Computing)	Dataset for each new study, e.g. the receptor proteins project, each case is a receptor class	File Collection: input: geometry basis set and optimization parameters output: optimized geometry, vibrational frequencies, energies	Gaussian computational chemistry software
EcoTox (Statistical Modeling)	One dataset, each case is one aquarium defined by amphibian, chemical, concentration, exposure	Data Model: acquarium descriptives, condition measures (temp, humidity, mortality, mass spec), animal measures (snout length, weight, gosner stage, mass spec samples) all measures over time	R codes to model, analyze, and report
Bone (Statistical Modeling)	One dataset, first group of cases represents participants, second group represents diet intervention	Data Model: participant descriptives, repeated specimens (serum, urine, feces), repeated measures (anthropometrics, compliance, nutrient intake) over baseline, treament, & recovery phases	R codes and SAS to model, analyze, and report

ACKNOWLEDGEMENT

The DEEDS project is supported by the National Science Foundation CIF21 DIBBs: EI: #1724728. We would like to acknowledge the work of our co-PIs Ashraf Alam, Marisol Sepulveda, Connie Weaver, and Joseph Francisco, and we are grateful for the efforts of their post-doctoral fellows and graduate students, who have worked closely with us on DEEDS.

REFERENCES

[1] McLennan M, Kennell R. "HUBzero: A Platform for Dissemination and Collaboration in Computational Science and Engineering." IEEE Computing in Science and Engineering, 12(2):48-53, 2010.

[2] Catlin AC, HewaNadungodage C, Pujol S, Laughery L, Sim C, Puranam A., Bejarano, A. "A Cyber Platform for Sharing Scientific Research Data at DataCenterHub." IEEE Computing in Science and Engineering, Vol. 20 (3): May/June, 2018. https://datacenterhub.org.

[3] McLennan M, Clark S, Deelman E, Rynge M, Vahi K, McKenna F, Kearney D, Song C. "HUBzero and Pegasus: Integrating Scientific Workflows into Science Gateways." Concurrency and Computation: Practice and Experience, 27(2): 328-343, doi: 10.1002/cpe.32. 2015.

[4] Sepulveda, Marisol. https://www.purdue.edu/newsroom/releases /2016/Q2/purdue-researchers-awarded-2.5-million-to-study-effects-of-perfluoroalkyl-substances-on-amphibians.html

[5] Weaver, Connie. http://www.purdue.edu/newsroom/releases/2014/Q3/purdue-receives-3.7-million-to-study-blueberries-and-bone-health.html

[6] Alam, Ashraf. https://www.ibj.com/articles/21663-purdue-aims-to-boost-solar-progress

[7] Hoehn, R., Nichols, D., Neven , H., Kais, S. Status of the Vibrational Theory of Olfaction, Front. Phys., 19 March 2018 https://doi.org/10.3389/fphy.2018.00025