The DEEDS Platform:
Support for Integrated Data and Computing across the Research Lifecycle

Chandima HewaNadungodage†, Ann Christine Catlin, Andres Bejarano, Steven Clark,
Guneshi Wickramaarachchi, Sumudinie Fernando, Parameswaran Desigavinayagam

ITaP Research Computing (Rosen Center for Advanced Computing)
Purdue University
155 South Grant St
West Lafayette, Indiana 47907, USA

chewanad@purdue.edu†, acc@purdue.edu, abejara@purdue.edu, clarks@purdue.edu,
gwickram@purdue.edu, swfernan@purdue.edu, pdesigav@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, 4) access toolkits for ad-hoc analytics, and 5) publish and share the investigations with broader research community. All interfaces are available from a single dashboard so that the workflow between data and tools is smooth and intuitive. This paper highlights the DEEDS platform capabilities for providing integrated data and computing services throughout the entire research investigation lifecycle and presents some examples from use cases that contributed to the platform development.

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

1. Introduction

Scientific investigations are complex processes that require research groups to make decisions about the methods they will use to preserve their data, build software for analysis, connect data to analysis tools, and share data and results. In most cases, these decisions are made in an ad-hoc way, so that researchers responsible for different areas of the project (collecting data, writing code, analyzing results) operate in different environments. Project data, code, analysis, and results are thus fragmented, which complicates preservation, sharing, interoperability, results traceability, and reuse.

Research cyberinfrastructures are most often created either to support data preservation and sharing [1 - 3] or to provide computing services and workflow support [4-9]. Even those infrastructures that support both data preservation and computations do not effectively integrate them; for example, 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 [10 - 15]. Science gateways (or virtual research environments) [16, 17] have been proposed to bridge the gaps and provide end-to-end support for research investigations; however, most of the time these platforms are heavily customized for the targeted research domain and difficult to be used by other disciplines.

The Digital Environment to Enable Data-driven Science (DEEDS) project recognized the need for an end-to-end solution, where the platform and its interactive interfaces provide the essential services required by researchers for representing and managing their collected data, defining metadata, exploring data collections, running analyses with selected data, tracking user workflows for reproducibility of results, and sharing all elements of the investigative process. Requirements and use of these essential services differ widely among research projects, even within the same science domain. The major challenge of implementing the DEEDS platform was to generalize the requirements from different science domains and put together a framework that would satisfy the data and computing needs of researchers from any discipline. DEEDS merged requirements for data management, computing services, and user interfaces into a single platform through a collaborative

effort that engaged researchers in the fields of chemistry, nutrition science, environmental science, agriculture, electrical engineering, and civil engineering [18].

DEEDS 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.

In this paper, we present the DEEDS platform innovations for integrating data and computing to provide end-to-end support for research investigation lifecycle. We describe the platform architecture and implementation and present the use cases from the science domains that defined features, services, and usability requirements for DEEDS.

2. DEEDS Platform Capabilities for Research Lifecycle Support

From the data upload point of view, DEEDS operates on a collection of "*datasets*". A dataset can be considered as a project or study conducted by a group of researchers. A project or study may have more than one dataset, and each dataset is managed through the DEEDS dataset *dashboard*. For data discovery, DEEDS allows more fine-grained access which facilitates easy search, sort, and comparison capabilities across related datasets.

The DEEDS dashboard provides a comprehensive set of features to encapsulate the end-to-end research investigation lifecycle. It controls data flow, metadata, and operations through a sequence of tabs that manage dataset Cases (organization of research activities), Files (repository management), DataTables (structured data management), Tools (computing services and workflows) and Analytics (built-in ad-hoc data analysis). Each tab provides interfaces for data acquisition and management related to the corresponding stage, and the platform maintains the interconnectivity between the elements. Figure 1 depicts the layout of the DEEDS dashboard.

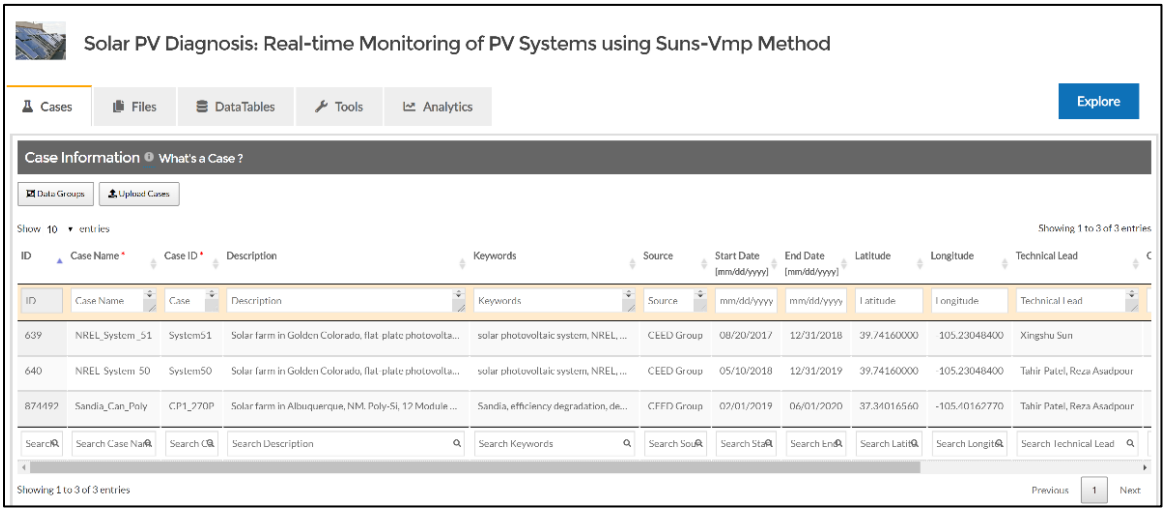

Figure 1. DEEDS Dashboard Layout.

2.1. Data Collection and Organization

The structure of a DEEDS dataset is organized according to a case-based approach. A "Case" represents how a researcher defines the basic unit of investigation for their research; it can be an experiment, study unit, site, specimen, or something else depending on the research domain. On the DEEDS platform, cases are the basic building blocks of a dataset and all the other elements in the dashboard (Files, DataTables, Tools, and Analytics) are linked together using the cases organization.

A dataset can consist of a few cases or hundreds of cases and can be expanded over the course of the investigation. DEEDS gives users the freedom of organizing their dataset into cases that best represent their research. However, it is also important to represent these datasets in a common structure that facilitates ease of discovery and re-usability by the broader research community once they are published. To address that need, DEEDS requires cases to be associated with a standard set of metadata such as name, description, keywords, bibliographic, spatial, and temporal information that complies with the Dublin Core Metadata Initiative (DCMI) metadata terms [19]. Having this standard allows DEEDS to assemble contents from a wide range of research domains under one platform for preservation, publication, discovery, and exploration.

Research groups use different ways to collect the measurements, observations, and other information related to their research; how the investigation process is carried out depends on the format of the input data collected. Based on the discussions with the domain scientists, following two formats that are widely used in our use case groups are currently supported for collecting and organizing research data.
(a) File Collections: Some research groups use collections of files as the input to their research investigations. They need interfaces to analyze these files using computational software and generate outputs.
(b) Complex Structured Data: Some research groups have complex data models collected using spreadsheets and need interfaces to interactively explore and update these data, and carryout statistical analysis using these data as input.

The “Files” tab in the DEEDS dashboard provides interfaces for uploading, managing, and sharing files collected/produced throughout the research investigation process. These files are organized into different categories for ease of exploration and reuse. Users can select a file category from the system defined categories (Reports, Data, Media, Figures, Other), or define custom categories as required (e.g. instrument-generated files, cell images, etc.). At upload, DEEDS captures and stores file metadata such as file name, category, size, MIME type, uploaded case(s), uploader and timestamp; when applicable, thumbnails and previews are also generated by the system. After upload, users can further classify and annotate the 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. Currently DEEDS supports file uploads directly from the web interface or using an SFTP client.

Data collected during research investigations can have very complex relationships. Researchers often collect and record these measurements as spreadsheets [20, 21]. For most of our use case groups, data sharing usually has meant exchanging the spreadsheet files themselves (email, drop box, share point) or translating spreadsheets to web forms. With both methods, research teams lose the flexibility, efficiency, and familiarity of shared interactive spreadsheet operations. Also, when researchers explore or analyze their collected data, manually keeping track of the relationships or implementing a database schema to appropriately represent the data model is a cumbersome task. The “DataTables” tab in the DEEDS dashboard provides 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. 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. Once the spreadsheets are uploaded and their connections are defined by the user, DEEDS creates and manages all necessary data models, database schemas, and connections. The DataTables interface provides many user-friendly features such as copying data templates, propagating metadata (data type, label, units, description, formatting, and visibility), and bulk updates to make it easy to establish consistent structure and annotation. Users can easily update, browse, and query the stored data at all levels using the familiar tabular interfaces presented in the DataTables dashboard and DEEDS explorer views. More implementation details on complex structured data support is provided in section 3.2.

In the future DEEDS data acquisition mechanisms will be expanded to provide APIs for importing files from external repositories, and to support collecting and managing live streaming data.

2.2. Computational Workflows and Data Analytics

DataTables and File repositories serve as the foundation for DEEDS data preservation. The Tools and Analytics tabs provides mechanisms for users to run computations and analysis on the data collected in the DataTables and Files to further their research efforts. Using the DEEDS dashboard files, data tables, and computational tools are directly connected to each

other and the relationships between inputs, computations, and outputs can be easily captured and recorded as workflows. This organizational structure makes it easier for researchers to understand, interpret, and use a dataset, allowing smooth interaction and transition between the elements.

Software applications can range from simple analysis programs to highly sophisticated scientific applications written in a number of languages. Sometimes, setting up and running these programs require substantial background work and domain knowledge. The " Tools" tab provides features to define, set up, store, and manage metadata describing computational programs, manage tool executions, and capture computing workflows. Once configured, the software can be launched and executed from the DEEDS dashboard with just few clicks. This make it easier for research groups to keep track of their programs and provides smooth knowledge transfer when a new member joins a team, or a current member leaves the team. 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. Researchers can also publish their tools and make them available to be used by other interested research groups.

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. More implementation details on computational workflow support is provided in section 3.3.

Data collection mechanisms in DEEDS are structured to follow the case-based layout for ease of uploading the data and maintaining the connections between dashboard elements. However, sometimes researches require more relaxed and ad-hoc mechanisms to analyze their data, without restricting to a specific structure. The "Analytics" tab in the DEEDS dashboard is designed to provide these ad-hoc analysis capabilities by allowing users to slice and dice the stored data as they want. DEEDS analytics interface is based on R and will provide features to compare data across cases, join data from multiple data tables, compare across related datasets, generate graphs and reports, run statistical toolkits, and if necessary, save the generated results and attach them back to cases. From the Analytics dashboard it is possible to invoke Jupyter notebooks (https://jupyter.org/), this allows more experienced users to carry out any sophisticated analysis on their data collections. DEEDS analytics capabilities are still under construction.

2.3. Exploration and Discovery

While the DEEDS dashboard provides feature-rich interfaces for data collection, computations, analysis, management, and sharing throughout the stages of the research investigation, the DEEDS Explorer provides a uniform interface for search and discovery across all the areas of the DEEDS dashboard. Cases, file collections, multi-level data spreadsheets, tool definitions, workflows, analytics results, and countless variations of these are available for interactive exploration by users.

We have created a general, extensible, "data-definition" language that accesses the data stored in a MySQL database and presents them as interactive tabular "dataviews." Our language defines each column of the display by specifying the database table and the field for the source of the data, and then 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 dataviews. Dataview layout can be pre-defined (e.g., case information) or dynamically defined in real-time (e.g., data tables). All dataviews 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 geospatial data, timelines for temporal data, and graphs for visualizing measurements and statistics. DEEDS explorer capabilities are not limited to single dataset level but can be expanded across datasets. Comprehensive metadata based on standards for vocabulary and semantics is key means for ensuring consistency in interpretation, use, and exchange of data across the datasets. Once the datasets are published, the broader research community will also be able to discover and compare the data across datasets using the DEEDS explorer interfaces.

2.4. Access Control

DEEDS datasets provide fine-grained access control. The dataset owner can share the dataset with the other members of the research team to collaboratively conduct the investigation. It is possible to assign different levels of permissions to different members of the team (data entry only, view only, manager, etc.) to control which elements they can access or which actions they can perform. Once the investigation is complete, the dataset can be published for discovery and reuse by the broader research community. Users will be able to publish the entire dataset or only publish selected components from the dataset, and also impose access controls on who will be able to view/reuse the published data (available to general public, available for registered users, restricted access for a certain group of users, etc.). The DEEDS platform will also provide traceability and tracking capabilities, an activity log is maintained to record user actions for each dataset and can be monitored by authorized users.

3. Platform Architecture and Implementation

3.1. System Architecture

The architecture and design of the DEEDS platform is built on top of the HUBzero™ cyberinfrastructure [22] and is an extension of DataCenterHub [23], a platform designed to preserve and publish scientific data for discovery and exploration. DEEDS has transformed DataCenterHub from a data preservation platform to a system that supports the end-to-end research investigation process. The DEEDS platform is implemented using the LAMP stack (Linux, Apache, MySQL, PHP) and JavaScript/jQuery for the front-end. DEEDS utilizes standard features provided by the HUBzero framework for user registration and management, authentication, system security, deployment, and issue tracking. We also use the HUBzero "submit package"[1] to provide high performance computing support on remote and local clusters; however, this package is customized to provide a web-based compute API to support DEEDS requirements. All the other features are implemented and maintained by the DEEDS R&D team. Figure 2 shows the high-level architecture of the DEEDS platform. The Platform layer consists of storage units (database, repository, definitions, templates, etc.) and CMS components that provide functionality to support the lifecycle of a research investigation. The User Interface layer provides easy to use interfaces for data collection, management, computations, analysis, exploration, and discovery.

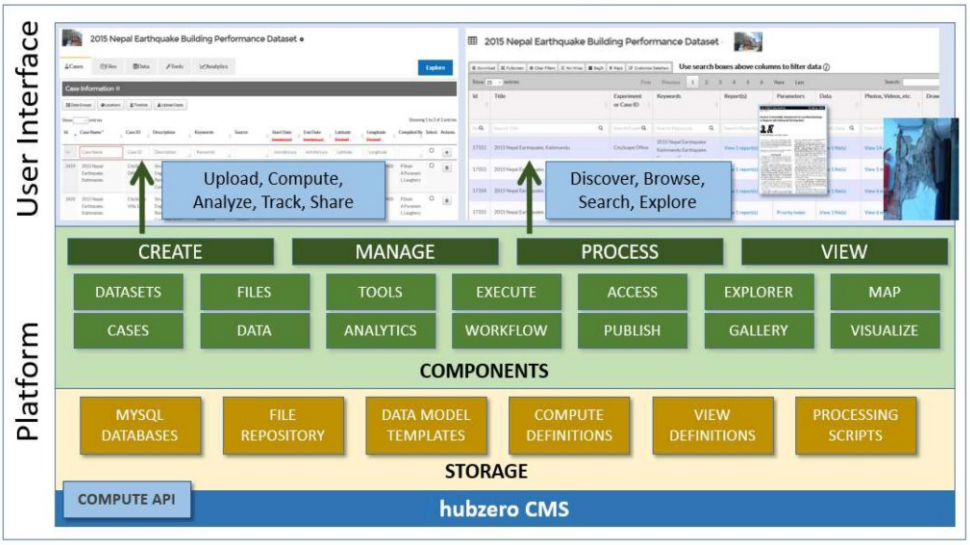

Figure 2. High-level Architecture of the DEEDS Platform.

Figure 3 shows the DEEDS platform database and repository organization. DEEDS database schema consists of a main database and a set of dataset specific databases. A new MySQL database is created for each new dataset, and entries relevant only to that dataset are mapped using MySQL views linked to their respective tables in the main database.

[1] https://help.hubzero.org/documentation/210/installation/installdeb/submit (

Datasets, cases, files, data models, and tools are assigned identifiers that are unique across the entire data platform; these identifiers are used to locate content in the database and link the repository components.

Files and file metadata are stored by datasets, file category, and cases. File storage for each case is “flattened” and therefore filenames within a single case directory must be unique. The flattened structure makes the assignment of file metadata more user-friendly in the data collection interface, and also ensures that data exploration and discovery is more efficient and effective. An additional directory above the case level supports the sharing of files across the cases. These files are physically stored in the “Shared Files” directory and linked symbolically from the case directories.

Data models also stored by datasets, each dataset directory has a collection of JSON format templates that describes the data models defined for that dataset. Data template metadata is saved in the main database and linked to dataset specific databases using MySQL views; actual data values are stored only in the data tables created in the dataset specific databases Further details on the DEEDS data model is given in section 3.2. Tools are defined within a dataset; however, tools can be imported to other datasets by authorized users. Therefore, Tool definitions are stored outside the datasets and shared across all datasets. Further details on tools and workflow support is given in section 3.3.

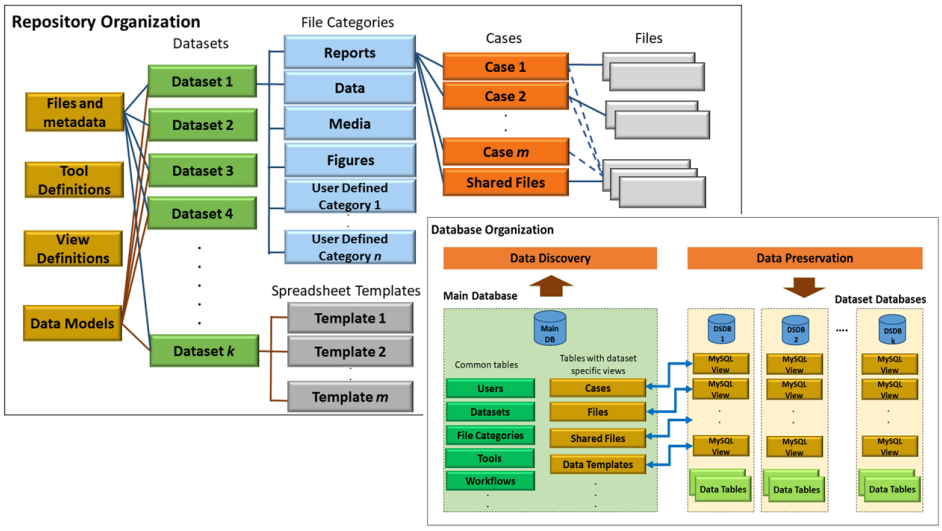

Figure 3. Database and Repository Organization.

3.2. The DEEDS Data Model

Spreadsheets have long been the preferred way for researchers to collect measurements, observations, and other data [20. 21]. The tabular format of the spreadsheets allows direct manipulation to data items by specifying its row and column indexes in the table. Furthermore, such nomenclature allows the definition of relations between data elements, rows or columns by using arithmetic operations, functions or algorithms. The relations may either modify existing data entries or generate new ones according to the relation specification.

Despite such flexibilities, spreadsheets come with drawbacks that must be addressed by the software application or user. The content of a spreadsheet is represented as strings in a file, usually separated by a character not used as data content; the actual data types for the spreadsheet columns must be either guessed or specified before defining the relations or saving the data in a data base engine. Some office suite software (e.g., Microsoft Excel and Libre Office Calc) can guess the data types and granulate them to cells; meanwhile, relational data base engines (e.g., Oracle and MySQL) require specifying the column types before inserting the data. The latter forces all data in a column to have the same data type, and this constraint applies when editing or adding table entries.

3.2.1. Support for Complex Structured Data

The tabular nature of spreadsheets allows them to be used as the fundamental store unit following a structured data model. DEEDS represents each spreadsheet using the notion of data templates. A “data template” is a storage unit designed for tabular data with the scope of hierarchical multi-dimensional dependency representation, along with complementary

information defined by both users and the DEEDS platform (technical specifications are discussed in section 3.2.2). Data templates are independent from each other (in terms of data content) but could be linked by connecting table columns to other data templates. Links between data templates follow a parent-child relationship: a parent data template must be generated before linking a sub level data template (child) to one of its columns. A link in a data template represents an extension of the column into another spreadsheet; such representation requires that both parent and child data tables to have the same number of rows (i.e., keep records from the same research units). The level of a data template is the number of parent data templates that must be drilled down before it is reached, this is analogous to the level of a node in a tree structure. The set of data templates without parent is named "top-level" data templates. This set specifies the starting points in the hierarchical branches of the data in a DEEDS dataset, therefore their respective level is zero. After the user specifies the linked columns and sub-level data templates, the connections between the data templates and all underlying database schemas are created and maintained by DEEDS the platform making the setup process transparent for the user.

Although data normalization is often encouraged, achieving a good schema design requires a thorough understanding of the data and the implications behind each dependency link in order to avoid duplications. There are situations where such duplications facilitate querying over the data, as well as gaining insight of what the data represents (e.g., periodical repeated measures obtained from tests on experimental units). Furthermore, the structure of some measurements or observations might be better represented as extensions over a table in a new dimension. Multi-dimensional data templates are suitable for repeated measures or frequent data observations from the same research units (e.g., same data collected through different days from the same specimen).

The entries in the data templates are associated with the research units defined in the Cases dashboard tab (as specified in section 2.1). Some datasets may require their data templates to be linked with a subset of cases rather than with the entire set, for this reason DEEDS defines the concept of "data groups": a subset of cases to be associated with specific data templates. Data groups are useful when the collected data is not uniform for all cases but rather a subset of them (e.g., specific data collected from one specimen that cannot be obtained from a different one). Each data template in a dataset is associated with a data group. Data groups are defined in the Cases tab, where users establish group membership. If there are no explicit data groups defined by the user, DEEDS assumes that *all* cases in a dataset are associated with *all* data templates defined in that dataset.

Once the template is defined, users can easily update, browse, query, and analyze the stored data using the dashboard interfaces. It is possible to edit cell values directly using the tabular interface or modify or add additional data using spreadsheet upload. DEEDS will warn the users if the changes are not consistent with the specifications of the respective data template. Furthermore, data templates enable additional operations (e.g., filtering, drilling down, and download) using the explorer dataviews, a DEEDS module built for interacting with data templates without compromising its data. Finally, this representation enables posterior data analysis (e.g., machine learning and data mining methods) over the data by saving data snapshots as data frames using the Analytics interface (under development). DEEDS saves a history log containing user actions (e.g., creation, data editing, and metadata editing), this helps dataset managers with the traceability of the data and metadata after the creation of the data templates.

3.2.2 Data Infrastructure and Technologies

The information for every DEEDS dataset is saved in a relational database engine, the system generates one database per dataset. The list of tables related to the DataTables interface are shown in Figure 4. DEEDS is responsible for the direct management of the tables and their inter-connections, users only provide the dataset information and populate the tabs (Cases, Files, DataTables, Tools, and Analytics) with the respective content.

A DEEDS data template (DT for abbreviation) is a tuple of tables $\{D, M\}$. Table D stores the data of the DT, the first three columns are reserved for the row id, dataset id, and case id respectively, regular data columns are identified as $D.col_i, i = 3,4,5 \dots, n$ where n is the number of columns in the table (as indicated in the original spreadsheet file). Table M contains the metadata for each column $D.col_i$. For each column DEEDS stores its name (as given by the user in the initial spreadsheet file), description, data type, units, data template link id (in case the entire column links to another DT), aesthetics attributes, creation data, and last modification data. Additionally, the basic information from each DT (name, description, the name of the data and metadata tables, creation data and last modification data) is stored in the data_templates table. Tables following the [meta]data_<ds_id>_<dt_#> format are tables that exist only in the respective

dataset specific database, the remaining tables are views from the main DEEDS database containing the information pertinent to the respective dataset. This representation is required for data exploration after a dataset is published in DEEDS.

From the user's perspective, generating a data template is a process guided by the system at every step. DEEDS allows users to generate one data template at a time. The process starts by requesting the name and description (optional) of the data template. Next, the user is asked to upload a csv file (DEEDS assumes the file contains only the name of the columns and their respective data) or to copy an existing data template. If the user uploads a spreadsheet file, then its content is analyzed in order to check its correctness and make an initial guess of the data types for each column. The system then displays the content from the respective source. At the header of each column the user can change the attributes of the column format, options include: data type, data format, aesthetic values, visibility (show or hide the column), among others. All format attributes can be changed later except for the data type (due to data base and linking restrictions). A column linking to another data template is specified in its respective data type option, the link value is set blank at this step and left for the user to define after the data template is saved in the system. The last step requires the user to confirm the data template, if confirmed then DEEDS saves the $\{D, M\}$ tuple in the database, registers a new entry in the *data_templates* table with the basic information of the data template, and finally defines the data-definition file (an internal visualization format for tabular data displaying) for the data template. Data template generation diagram is presented in Figure 5.

Sub-level data templates can be specified after generating the respective parent data template. Using the format options for each column, the user can indicate that a column links to another data template if its data type is set to link. In this case the system offers options to either automatically set the link to either an existing data template or a new data template using the guided process described above. The link represents an extension of the data template into another data template (following the principle of "spreadsheet of spreadsheets"), drilling down through data templates is possible using the DEEDS dataview module. On the displayed view, users can interact with the tabular content by filtering values (both globally or by column), interact with media content or visualize in a map (if cell data correspond to such) and download the data as a csv spreadsheet file.

Tables_in_deeds_ds_<ds_id>
data_<ds_id>_<dt_1>
data_<ds_id>_<dt_2>
data_<ds_id>_<dt_3>
data_template_links (view)
data_templates (view)
metadata_<ds_id>_<dt_1>
metadata_<ds_id>_<dt_2>
metadata_<ds_id>_<dt_3>

Figure 4. Database Structure for the Data Templates in a Dataset.

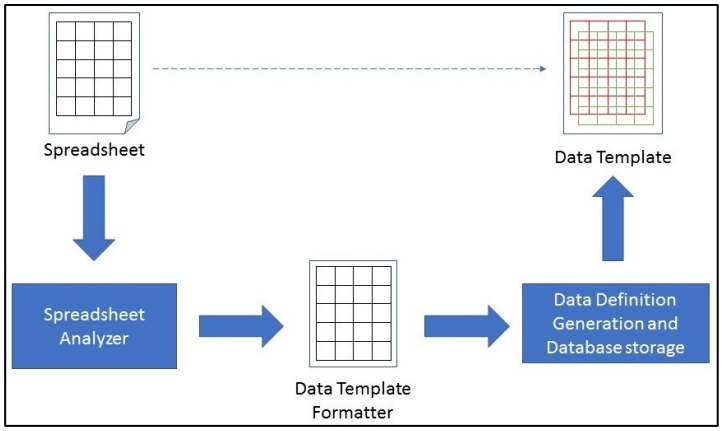

Figure 5. Data Template Generation Diagram.

3.3. Support for Tools and Computing Workflows

Research groups use different computational software and programs to conduct their investigations. DEEDS provides a simple web form interface where users can define computational tools and their metadata, and also provides infrastructure for hosting and documenting versions of an application throughout the development lifecycle. A new web-based API for HUBzero "submit" was developed to manage compilation, installation, launch, and execution of tools from the DEEDS dashboard.

3.3.1. Tool Definition

As shown in Figure 6(LHS), tool definition is a semi-automated process. DEEDS provides a step by step dialog to guide the user through defining a tool and collects the metadata (listed in Figure 6) required by the submit API for installing and launching the tool. When defining a tool, users can either upload the source code for a program they developed or choose

to execute a pre-installed licensed software package. Resource requirements of the application dictate where it will be executed. Short duration executions requiring low core count can be executed on DEEDS hardware on demand – no queues required. Long running or multi-core applications are better suited for execution on cluster hardware; DEEDS currently supports executing tools on Purdue University HPC clusters, in the future tool execution in external clusters will be supported. DEEDS provides free access to a small queue and users can access their own HPC queues with minimal configuration. The list of possible execution sites and queues should be specified in the tool definition along with the expected walltime and number of cores (if applicable). If any additional software modules are required to execute the application these may also be declared for each execution site. Users can also set up output upload options to automatically upload any portion of the outputs to their datasets. Users are encouraged to add user manuals and readme files as part of the tool definition to make it easy for other researchers to use the tool later.

Once the user enters the required metadata, DEEDS stores these information in the database and saves the source code in the DEEDS repository (if applicable). Then the system generates submit configuration files required for compiling and establishes the execution support environment. The configuration files contain information on programming language, executable format (compiled from source code or pre-installed executable), execution sites, required libraries and modules, and operational rules. To ensure a secure environment, DEEDS then sends a tools configuration request to a tool administrator who will manually inspect and approve any user-written code and scripts. The administrator also inspects the auto-generated configuration files and adds any extra information which needs to be added manually. If the software passed the security requirements, the administrator invokes submit API calls to compile (if applicable) and install the tool. The submit API parses the configuration files and establishes an execution environment in each site specified by the tool definition. During the configuration process, the tool is in “Pending” status. If the configuration is successful, the tool status is updated to “Ready” and the tool is added to the “launch” section in the Tools tab. If any errors occur during compilation or installation, the tool status is set to “Error” and the user is notified to check and correct the errors. Successfully configured tools in “Ready” status can be imported to other datasets by authorized users. It is possible to add new execution sites or queues to an existing tool definition; this will generate a new configuration request for that site/queue.

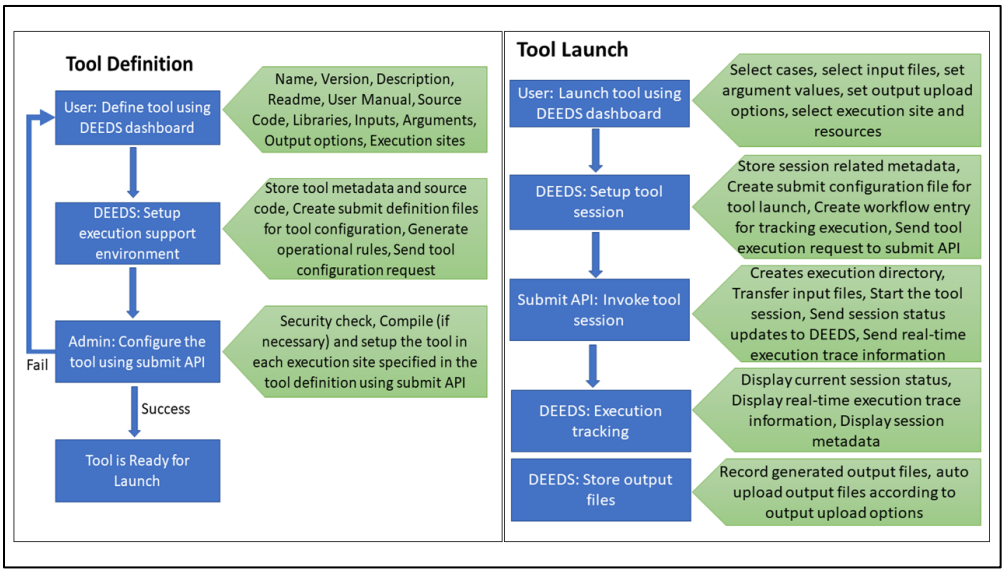

Figure 6. Steps for Tool Definition (LHS) and Tool Launch (RHS).

3.3.2. Tool Launch

Once a tool is defined or imported to a dataset, it can be launched from the dashboard Tools tab by any user who is authorized to access that tool. Tool launch steps are shown in Figure 6(RHS). Using the tool definition information, DEEDS guides the user through the tool launch steps - selecting cases and input files, specifying arguments, specifying output upload options, and selecting an execution site from the pre-configured list of sites. Files uploaded to a dataset may be associated with particular cases or they may be independent of cases. If the desired input is associated with one or more cases the user selects such cases to filter the list of files. If the desired input is independent of all cases, case selection can

be bypassed and files that are common to the entire dataset will be available for input selection. To make it easy for the user to select the input data, DEEDS filters the current files uploaded to the selected cases based on the specifications in the tool definition and presents a list of candidate files. Tool launch interface also displays the default values for the input arguments and output auto upload based on the tool definition; these values can be modified at launch if necessary. This process makes it easier for any user to invoke the tool with just few clicks, without worrying about the underlying complexities of configuration.

When the user submits a tool launch request, DEEDS saves the metadata related to that tool launch to the database, creates a tool session execution directory and symlinks the selected input files from the file repository. Using the launch information entered by the user, DEEDS generates a JSON formatted launch configuration file which will be parsed by the submit API to invoke the tool. This file contains the information on tool and version, input files, command line arguments, execution site and queue, resources (serial/parallel, number of cores, walltime), and trace files to be tracked during execution. A computational workflow entry to track the execution progress is created and displayed in the "workflows" section in the Tools tab (see section 3.3.2). Then DEEDS sends a tool execution request to the submit API to parse the launch configuration file and set up the necessary submit execution directories. For execution on remote clusters, input files are copied to the remote machine, then submit generates the tool invoke command according to the input files and other command line arguments specified in the launch configuration file. In addition to output files produced by the application, DEEDS captures text written to standard output and standard error, and a separate record is made of resource requirements for executing the application. Session status updates and execution information (stdout, stderr, execution stats, any other program generated trace files) are displayed in the workflow section of the Tools dashboard in real-time. Once the execution is complete, depending on output upload options specified at tool launch, DEEDS will automatically annotate, classify, and upload the generated output files to the cases selected at launch. Files that are not auto-uploaded can be manually uploaded later.

3.3.2. Workflow Tracking

A fundamental principal of the DEEDS platform is traceability and reproducibility of results. To fully capture computational workflows, DEEDS encapsulates all the information related to a tool session end-to-end and displays them in one place, maintaining the interconnectivity between the cases, input files and arguments, computational tools, and produced output. The workflow information includes workflow name, tool and version, selected input files, argument values, execution status, execution metrics, trace files (stdout, stderr, etc.), outputs generated, user who launched the tool, and launched timestamp. Any input or output files connected to the workflow are directly accessible from the workflow display; execution status, metrics, and trace information can be directly viewed from the dashboard. Users can view, search, and compare workflows for different tool sessions. Workflows can be shared among dataset users to promote collaborative discussion.

Currently, the definition and management of pre-processing and post-processing code is not fully implemented in the DEEDS interface. In the future, the Tools tab will support the definition and management of pre and post-processing code, and the definition, execution, and tracking of Pegasus enabled computing workflows using HUBzero submit [8]. It is also worth to note that the applications that require interactive user input via GUIs cannot be launched as DEEDS tools and automatically tracked. However, users can run these software outside of the DEEDS platform and create a user defined workflow to manually integrate inputs and outputs.

4. Use Cases

The DEEDS R&D team partnered with research groups from different 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. A pilot version of the DEEDS platform is available for our use case researchers at DataCenterHub (www.datacenterhub.org). Our partners are relying on DEEDS to preserve, use, 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. In the following paragraphs we present some examples on how our use case researchers utilize the DEEDS platform to conduct their research.

Environmental Science [EcoTox]: This research develops amphibian toxicity reference values for ecological risk assessment in sites contaminated with poly- and perfluoroalkyl chemicals (PFAs). Reference values aid in making decisions on exposure mitigation and federal regulations for pollution control [24].
The EcoTox team members built the following DEEDS datasets to support their research investigations,

EcoTox X1702 Dataset: Subchronic Bioaccumulation of PFAS Chemicals in Larval Amphibians
EcoTox X1703 Dataset: Amphibian PFAS Adult Dermal Exposure Sublethal Effects
EcoTox X1901 Dataset: Toxicity Reference Values for Aquatic Exposure of PFAS Mixtures Using Mesocosms

These datasets demonstrate the use of DEEDS DataTables interfaces for structuring, standardization, preservation, and annotation of collected quantitative & qualitative data, including repeated measures for phenotype and chemical data. DEEDS provides data curation, validation, and quality control for all the chemical data uploaded into these data tables. Figure 7 shows some example operations in the edit mode for the "Animal Phenotype Measures" data template from the EcoTox X1702 dataset. This team also use DEEDS datasets Files interfaces to organize, annotate, and share protocols, control forms, and other vital documents related to the research. EcoTox team uses R scripts for data analysis and plot generation; these are defined as DEEDS tools in dataset dashboard and all the computation workflows are tracked end-to-end ensuring traceability and reusability of produced results.

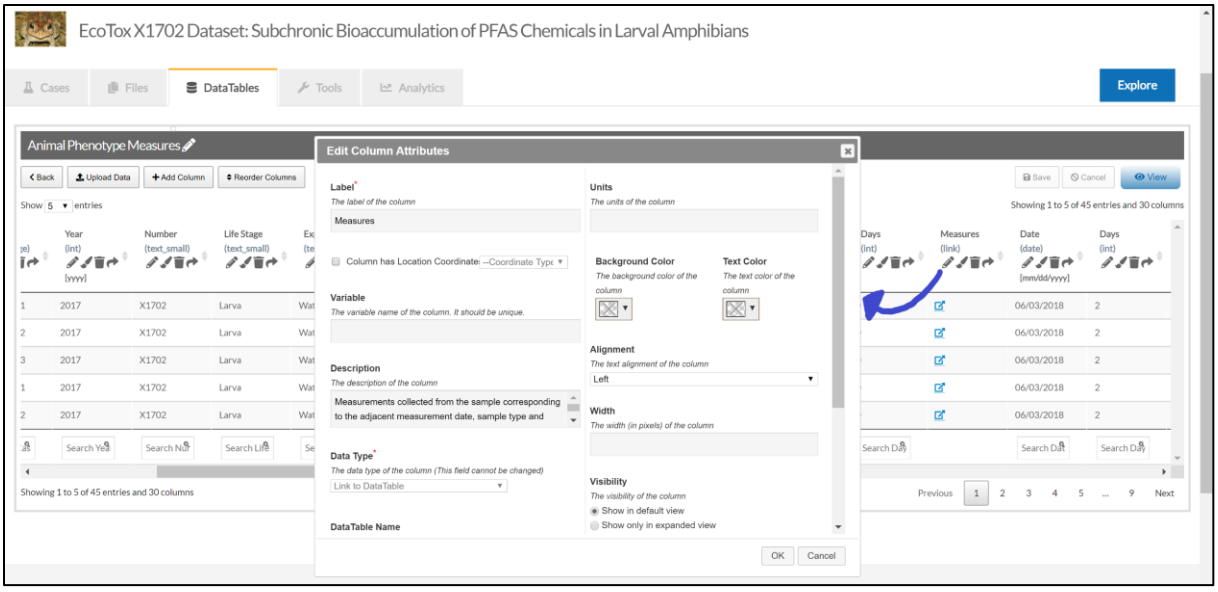

Figure 7. Edit Mode of "Animal Phenotype Measures" Data Template from EcoTox X1702 Dataset.

Nutrition Science [Berries & Bone]: This research studies blueberry intake at different dose levels added to regular diet and measures the effect on net bone calcium retention and biochemical markers of polyphenol and bone metabolism in postmenopausal women [25]. This team built the "Berries & Bone Dataset" to support their research investigation.
This DEEDS dataset demonstrates structuring and standardization of DEEDS DataTables for preservation and annotation of collected quantitative and qualitative data, including specimen, cross sectional, and other repeated measures. This team made the best use of DEEDS DataTable capabilities to put special emphasis on customization of data tables for viewing and exploration of extensive and complex measurements. DEEDS Files interface capabilities are used for collection, classification, and annotation of subject-based data files (e.g., scans, survey data) and dataset-based files such as protocols, IRBs, clinical data timelines, and questionnaires. This dataset adheres to HIPAA requirements for data privacy (DEEDS is a HIPAA compliant platform). Berries and Bones team also uses R scripts for data analysis, data comparison, and plot generation; these are defined as DEEDS tools in dataset dashboard and all the computation workflows are tracked end-to-end ensuring traceability and reusability of produced results.

Electrical Engineering [SolarPV]: This research models the efficiency of solar photovoltaic (PV) systems by coupling data for weather, manufacturer-specific PV technology, and solar farm health to diagnose efficiency degradation and predict system lifetime [26]. The SolarPV team built the "Solar PV Diagnosis Dataset: Real-time Monitoring of PV Systems using Suns-Vmp Method" to support their research investigations. They use DEEDS Files interfaces to organize, annotate, and share field

data and initial parameter input files corresponding to each solar farm case under investigation. DEEDS Tools capabilities are used to define and maintain the versioning of their Suns-Vmp MATLAB research code. This tool uses DEEDS HPC services for tool launch. DEEDS automatically captures all the execution workflows with auto-upload of generated output for preservation, exploration, and comparison across cases. Use of DEEDS to support their research is acknowledged in press releases [26] about their work and in conferences and papers.

Computational Chemistry [Quantum]: This research studies spectroscopy, kinetics, and photochemistry of transient species to optimize molecular structure, predict properties, and provide reference data to guide experimental search for these species [27]. The Quantum team members built the "Quantum Post- Lock and Key Descriptor Dataset" to support their research investigation. This dataset used DEEDS Tools interfaces to define the algorithms that will be executed using the Gaussian software packages [28]. These tools can be launched from the DEEDS dashboard and executed on HPC cluster services provided by DEEDS. A post-processor software was developed by the Quantum team and is attached to Gaussian executions through DEEDS. This code is automatically invoked after successful Gaussian executions to generate csv files and figures which describe charge, thermochemical, and vibrational data. DEEDS captures the end-to-end workflow and auto uploads .log and .chk files generated by the Gaussian software and also all the output files generated by the post-processor code back to the dataset and annotate and associate them with cases for exploration and comparison. Figure 8 shows the tool definition, tool launch, and workflow tracking interfaces of the computational chemistry dataset.

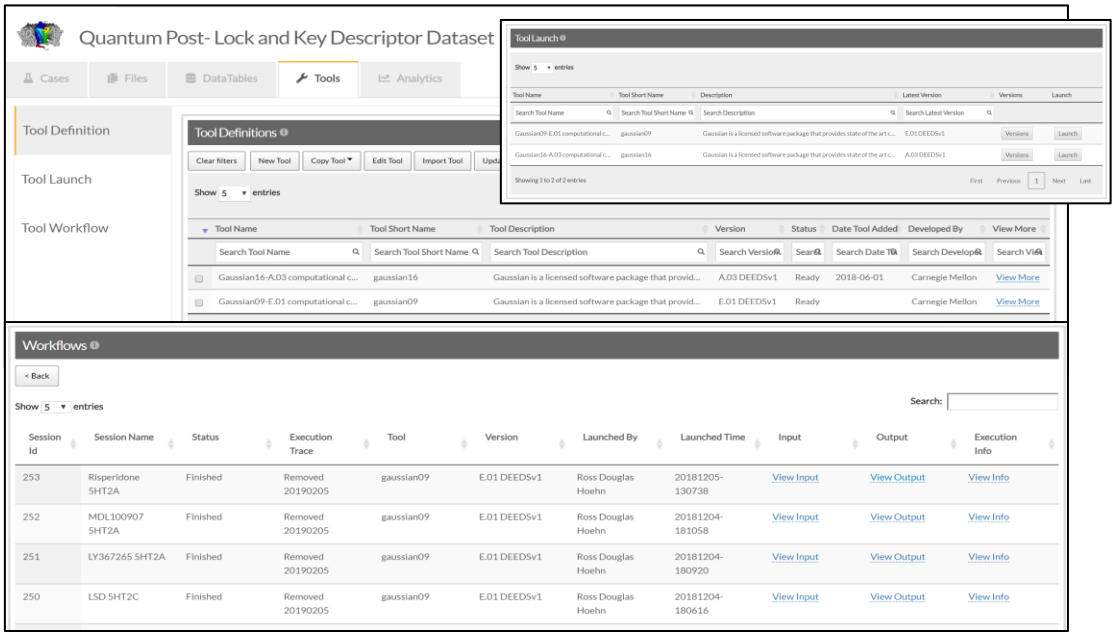

Figure 8. Tools Dashboard of the Quantum Dataset.

Civil Engineering [SMARTI]: This research studies aging rural bridge infrastructure by integrating existing datasets and data collected using next-generation health monitoring technologies and assess socio-technical impacts associated with potential decisions in bridge construction [29]. This team built the "SMARTI National Bridge Inventory (NBI) Dataset" to support their research investigation. This DEEDS dataset uses DEEDS DataTables interfaces for structuring, standardization, preservation, and annotation of quantitative and qualitative data collected by third party (government) agencies over the past 30 years for tens of thousands of bridges in Nebraska. Using the DEEDS dashboard features, this team also paid special attention on customization of data tables for viewing and exploration of extensive and complex measurements. DEEDS Files interface capabilities are used to organize, annotate, and preservation nearly a quarter of a million bridge-based files (reports, images, figures). These files are attached to bridge "cases" and internally linked to their inspection data tables for exploration, comparison, and analysis. This team uses maps features in DEEDS dataview interfaces for geospatial exploration of their bridge data. SMARTI team created Python scripts as DEEDS tools for data analysis, data comparison, and plot generation. These execution workflows are automatically tracked in Tools interface providing traceability and reusability of the generated results. They also use Jupyter notebooks capabilities provided by Analytics dashboard to perform bridge data analysis.

Forest Science [GFBI]: This team is part of the Global Forest Biodiversity Initiative (GFBI)[2] which supports research and policy-making in forest science and related areas [30]. The "GFBI Dataset: Positive Biodiversity–Productivity Relationship Predominant in Global Forests" was built for preservation, sharing, and use of forest inventory data for productivity analysis. This dataset utilizes DEEDS DataTables interfaces for organization, preservation, annotation of quantitative and qualitative forest "plot" data from worldwide contributions paying special attention on customization of data tables for viewing and exploration of nearly a million data table "plot" cases. A DEEDS tool is defined to analyze inventory data and compute biodiversity–productivity relationships. This tool can be executed from the DEEDS dashboard using the HPC services, DEEDS automatically captures the workflows and upload generated output back to the dataset for exploration, comparison, and analysis. GBFI team uses maps features in DEEDS dataview interfaces for geospatial exploration of their forest plot data (Figure 9). Use of DEEDS platform to support their research is acknowledged and DEEDS explorer interface for GFBI data is made publicly available in their website[3].

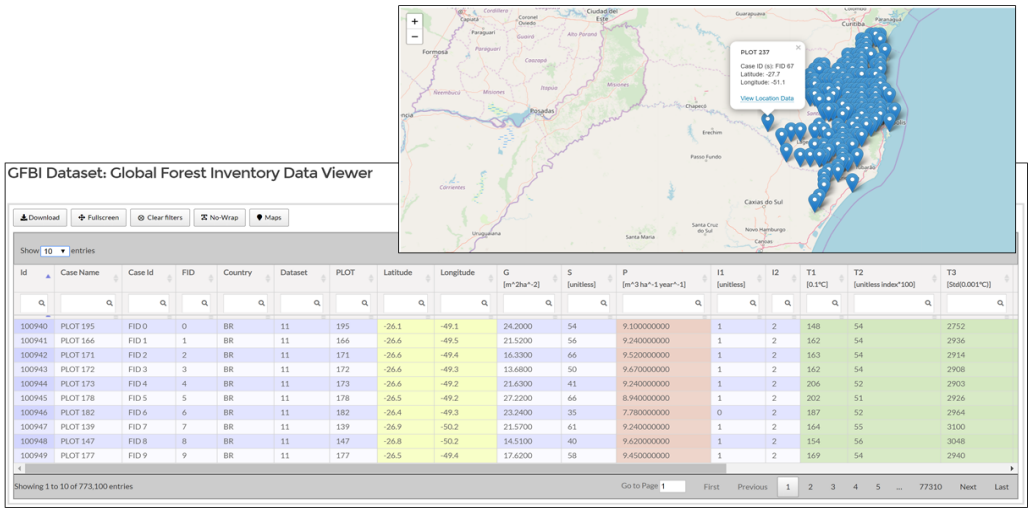

Figure 9. Map Interface of the GFBI Dataset.

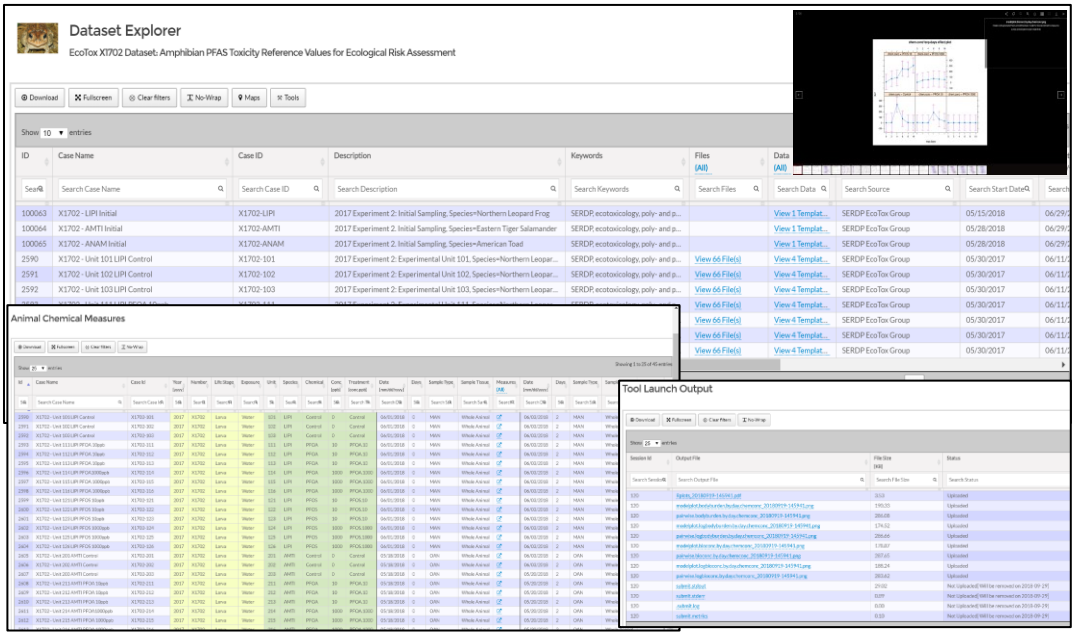

Figure 10. Some Examples of Explorer Dataviews for the EcoTox X1702 Dataset.

[2] https://www.gfbinitiative.org/
[3] https://ag.purdue.edu/facai/data/gfbi/

All of our use case research groups use the visualization, search, sort, discover, and comparison capabilities offered by explorer dataviews to understand and further analyze their results. For a given dataset, cases, files, data tables, tools, workflows, and results generated by tool executions are all connected and are accessible by following the drill-down links from the Dataset Explorer dataviews. Once published, broader research community can also use these DEEDS explorer interfaces to understand the research investigation process and browse the input data, computational workflows, and generated outputs. Since all elements of the entire research lifecycle are interconnected and available in one place, DEEDS ensures provenance, traceability, and reusability of the produced outcomes. Figure 10 shows some examples of the explorer dataviews for the EcoTox X1702 dataset.

5. Related Work

Currently there are many data sharing platforms that support components of the scientific workflow. Most existing research cyberinfrastructures are created either to support data preservation and sharing or to provide computing and data analytic capabilities. Although there are some discipline specific platforms that support end-to-end research investigations, these platforms are heavily customized for the targeted research domain and cannot be used by other disciplines. In this section we briefly discuss several widely used systems and compare their features to that offered by the DEEDS platform.

Zenodo [1] is a platform which provides features for research data sharing and publication. It stores research data from any research domain and offers citable sources for its collection of research outputs. Figshare [2] is another discipline neutral platform that offers features for research data preservation, sharing, and search. Dryad [3] is a platform where users can make research data publicly available and link those data to published literature. Data packages are assigned keywords, abstracts, and spatial and temporal information. SQLShare [10] provides its users with interfaces to upload, share, and manipulate research data collected in the form of spreadsheets. Once uploaded, users can write and run SQL queries on these spreadsheets and share the results. DataHub [11] from MIT is an installable framework for research data sharing. Users can install, manage, and run their own data sharing platform using this framework. Users view data in tabular formats and can write, save, run, and share SQL queries. It also provides analytical tools and charts for data visualization.

There are several differences between DEEDS and the above data preservation platforms. DEEDS capabilities are not limited to preserving file collections, we offer support for preserving complex structured data in the form of multi-dimensional hierarchical data tables. Some of the above platforms [1,2,3] only allows users to download the preserved files; whereas DEEDS offers interactive interfaces for online exploration of the preserved files and structured data. Some systems [10, 11] have limitations on the file types supported by the system (E.g. CSV, text, or JSON files); whereas, DEEDS places no restrictions on file types which can be uploaded, shared, and explored.

CKAN [13] and DKAN [14] are two popular open source data management platforms. Users can install and customize these platforms to build their own data management websites. Both platforms offer a full suite of cataloging, publishing, and visualization features that allows users to easily share data with the public. Dataverse [15] is another popular platform which provides excellent features for sharing, preserving, citing, exploring, and analyzing research data. User can set up a personal dataverse and make their research data more discoverable to the broader research community. SciServer [12] is a fully integrated cyber-infrastructure system which enables researchers to share and analyze big data. It allows researchers to work with Terabytes or Petabytes of scientific data, without needing to download large datasets. This system provides significant data analytics capabilities. However, in these platforms the connection between the input data, computational workflows, and analytical results is not straightforward. Also, sometimes users need specialized knowledge in computer science algorithms and SQL syntax to use the tools and querying features. With DEEDS, cases, file collections, data models, computations, and results are linked to each other in a clear way, with instant, direct access to each element, and user-friendly querying, exploration, and comparison capabilities.

There is a wide range of high-performance computing (HPC) platforms [4] and workflow management systems (WMS) [5] like Pegasus [6], Kepler [7], and Taverna [8]. These infrastructures provide interfaces to create, execute, and share workflow models across a broad range of scientific and engineering disciplines. These systems automatically map the high-level workflow descriptions onto distributed resources and automatically locate the input data and computational resources necessary for workflow execution. Although these systems provide support for sophisticated workflow models, the "execution workflows" supported by these systems are just one part of the of the research investigation lifecycle and are

not fully integrated with other important elements (such as research unit composition, metadata, protocols, publications) to represent the whole picture of the investigation. The DEEDS platform fully integrates all the elements of a research investigation in one dashboard and provide effortless navigation and interconnections between the elements.

Growing number of science gateways or virtual research environments (VREs) [16, 17] are been implemented to provide the research scientists seamless access to data, software, and computing services they require in performing their research activities without having to worry about underlying complexities and technical details. However, most of these platforms are ad-hoc systems built to serve the needs of a specific community and difficult to be used by a different discipline. Recently, gCube system [31] was proposed to enable the creation and operation of an innovative typology of hybrid data infrastructures by aggregating resources from other infrastructures and offer VREs as-a-Service. While gCube and DEEDS have similar objective of providing a discipline neutral platform to support research investigations, we believe there are some fundamental differences. gCube allows the users to create their own VRE by combining components from list of available software and resources and provide services supporting the entire data management lifecycle. DEEDS allows users more flexibility in organizing and conducting their research, users can define research units and data models in a way that best suit their investigation. Also, once published, DEEDS explorer interfaces make it possible for the broader research community to easily discover, browse, understand, and reuse the research outcomes as the entire investigation process is interconnected and available in one place.

The above-mentioned platforms have diverse focus, goals, and capabilities, and serve the scientific community in different capacities. What makes the DEEDS platform unique from the above systems is our goal of providing end-to-end support for the research investigation lifecycle with the flexibility of a self-serve platform that can be used by wide range of disciplines. DEEDS offers a full range of services, from research data acquisition to publication of the outcomes – supporting, integrating, and capturing all the intermediate steps, all connected and accessible from one dashboard, to ensure traceability, reproducibility, discoverability, and ease of use.

6. 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. When the investigation is complete, research groups can publish the research outcomes along with data, algorithms, and workflows, using the same platform.

In this paper, we discussed the concepts, capabilities, architecture, and implementation details of the DEEDS platform and presented the use cases that contributed to building the platform. Currently the pilot version of the DEEDS platform deployed in DataCenterHub is only accessible by our use case research groups. We continue to improve and evolve DEEDS to support research projects across a broad spectrum of scientific disciplines. The system will be available for the general public in October 2019. In the future we are planning to make DEEDS CMS components available as part of the HUBzero open source service for no-cost download and installation by the user communities.

ACKNOWLEDGMENTS

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

REFERENCES

[1]. Zenodo website. https://zenodo.org (accessed 30 January 2019).
[2]. Figshare website. https://figshare.com (accessed 30 January 2019).
[3]. Dryad website. http://datadryad.org/ (accessed 30 January 2019).
[4]. Calegari, P., Levrier, M., Balczyński, P., "Web Portals for High-performance Computing: A Survey," ACM Transactions on the Web (TWEB), 13(1):5, 2019.
[5]. Liew, C. S., Atkinson, M. P., Galea, M., Ang, T. F., Martin, P., Hemert, J. I. V., "Scientific workflows: Moving across paradigms," ACM Computing Surveys (CSUR), 49(4):66, 2017.
[6]. Ludäscher, B. et al., "Scientific workflow management and the Kepler system," Concurrency and Computation: Practice and Experience, 18(10):1039-1065, 2006.
[7]. Wolstencroft, K. et al., "The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud." Nucleic acids research, 41(W1):W557-W561, 2013.
[8]. McLennan, M. et al. "HUBzero and Pegasus: Integrating Scientific Workflows into Science Gateways," Concurrency and Computation: Practice and Experience, 27(2):328-343, 2015.
[9]. De Roure, D., Goble, C., Stevens, R., "The design and realisation of the My Experiment Virtual Research Environment for social sharing of workflows," Future Generation Computer Systems, 25(5):561-567, 2009.
[10]. SQLShare website. https://sqlshare.uw.edu/ (accessed 30 January 2019).
[11]. DataHub website. https://datahub.csail.mit.edu/ (accessed 30 January 2019).
[12]. SciServer website. http://www.sciserver.org/ (accessed 30 January 2019).
[13]. CKAN website. https://ckan.org/ (accessed 30 January 2019).
[14]. DKAN website. http://getdkan.com/ (accessed 30 January 2019).
[15]. Dataverse website. https://dataverse.org/ (accessed 30 January 2019).
[16]. Candela, L., Castelli, D., Pagano, P., "Virtual research environments: an overview and a research agenda," Data Science Journal, GRDI-013, 2013.
[17]. Barker, M. et al., "The global impact of science gateways, virtual research environments and virtual laboratories," Future Generation Computer Systems, 95:240-248, 2019.
[18]. Catlin, A. C., HewaNadungodage, C., Bejarano, A., "Lifecycle Support for Scientific Investigations: Integrating Data, Computing, and Workflows," Computing in Science & Engineering, 21(4):49-61, 2019.
[19]. Weibel, S., Kunze, J., Lagoze, C., Wolf, M., "Dublin core metadata for resource discovery," Internet Engineering Task Force RFC, 2413(222), 132, 1998. http://dl.acm.org/citation.cfm?id=rfc2413.
[20]. Birch, D., Lyford-Smith, D., Guo, Y., "The future of spreadsheets in the Big Data Era." arXiv preprint arXiv:1801.10231, 2018. https://arxiv.org/ftp/arxiv/papers/1801/1801.10231.pdf
[21]. Buys, C. M. and Shaw, P. L., "Data Management Practices Across an Institution: Survey and Report," Journal of Librarianship & Scholarly Communication, 3(2), 2015.
[22]. McLennan, M., and 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.
[23]. Catlin, A.C. et al., "A Cyber Platform for Sharing Scientific Research Data at DataCenterHub," IEEE Computing in Science and Engineering, 20(3): 49-70, 2018.
[24]. Abercrombie, S. A. et al., "Larval amphibians rapidly bioaccumulate poly-and perfluoroalkyl substances," Ecotoxicology and environmental safety, 178:137-145, 2019.
[25]. Weaver, Connie. http://www.purdue.edu/newsroom/releases/2014/Q3/purdue-receives-3.7-million-to-study-blueberries-and-bone-health.html (accessed 30 January 2019).
[26]. Alam, Ashraf. https://www.ibj.com/articles/21663-purdue-aims-to-boost-solar-progress (accessed 30 January 2019). https://www.purdue.edu/newsroom/releases/2018/Q3/physics-model-acts-as-an-ekg-for-solar-panel-health.html
[27]. Hoehn, R., Nichols, D., Neven, H., Kais, S., "Status of the Vibrational Theory of Olfaction," Frontiers in Physics, 2018, https://doi.org/10.3389/fphy.2018.00025.
[28]. Frisch, M. et al. Gaussian 16, Revision A.03, Gaussian Inc., Wallingford CT. http://gaussian.com/g16new/ (accessed 30 January 2019).
[29]. Gandhi, Robin. https://engineering.unl.edu/smarti-project-rural-bridge-health-management-partnership-between-uno-unl-1m-nsf-grant/ (accessed 30 January 2019).
[30]. Steidinger, B. S. et al., "Climatic controls of decomposition drive the global biogeography of forest-tree symbioses," Nature, 569(7756):404, 2019.
[31]. Assante, M. et al., "The gCube system: delivering virtual research environments as-a-service," Future Generation Computer Systems, 95:445-453, 2019.