The DEEDS Platform: Structured Data Representation and Statistical Modeling Support

Andres Bejarano[†1], Robert Flynn[2], Tyler Hoskins[2], Michael Iacchetta[2], Steven Clark[1], Guneshi Wickramaarachchi[1], Sumudinie Fernando[1], Parameswaran Desigavinayagam[1], Chandima HewaNadungodage[1], Ann Christine Catlin[1]

[1] ITaP Research Computing
Purdue University
155 South Grant St
West Lafayette, Indiana 47907, USA

[2] Forestry and Natural Resources
Purdue University
715 W. State Street
West Lafayette, Indiana 47907, USA

abejara@purdue.edu†, wflynn@purdue.edu, tdhoskin@purdue.edu, miacchetta@usgs.gov, clarks@purdue.edu, gwickram@purdue.edu, swfernan@purdue.edu, pdesigav@purdue.edu, chewanad@purdue.edu, acc@purdue.edu

***Abstract*— The Digital Environment for Enabling Data-driven Science (DEEDS) project is a partnership between domain scientists and computer scientists to create a platform that offers end-to-end support for diverse scientific workflows. DEEDS provides services for organizing research activities, building file repositories, representing structured data, defining and connecting to computing tools, and analyzing results—integrated into a single, powerful dashboard. This paper focuses on an environmental science research project that relies on DEEDS to preserve, share, and analyze large volumes of interrelated measurements collected over time. The DEEDS DataTables component manages their complex data model as interactive multi-dimensional, hierarchical tables, with metadata for annotation, validation and customized viewing. The Tools component manages upload and configuration of user modeling tools and supports their launch and workflow tracking for results traceability. A discussion of DEEDS functionality is followed by a description of how this research group transitioned their scientific investigation to the DEEDS platform.**

Keywords— research life-cycle support, environmental science, data modeling, statistical analysis

I. 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.

The 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. However, 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 [1]. Research groups create shared datasets on the DEEDS platform for their projects. DEEDS provides a dataset dashboard that 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) [2].

Members of the Strategic Environmental Research and Development Program "EcoTox" project [3] are creating team-shared datasets to support their research investigation. Their dataset cases represent experimental units or aquaria defined by study properties such as animal species, chemical, and concentration. The fundamental requirement for this research groups is the systematic, reliable preservation, validation, and analysis of large volumes of collected measurements and observations. Data are modeled as hierarchical, multi-dimensional data tables to represent repeated measures for cases with multiple measurement types and multiple phases over time. The key components supporting their activities are DataTables and Tools. A brief description of these two components is given first. We then describe how EcoTox researchers transitioned their scientific workflow to DEEDS.

II. DataTables Component

Data collected during research investigations can have very complex relationships. Researchers often collect and record these measurements as sheets in Excel workbooks. When researchers explore or analyze their collected data, manually keeping track of relationships or implementing a database schema to appropriately represent the data model is a cumbersome task. The DEEDS DataTables component supports structured data representation and management, and the platform provides easy-to-use interfaces for defining complex data models by simply uploading and linking collections of spreadsheets. A DataTable is a storage unit that defines,

preserves, and manages data entries collected from research units of the project (cases or groups of cases). It is displayed as a spreadsheet with extra capabilities for representing multi-dimensional hierarchical tables (i.e., spreadsheets of spreadsheets). Users can define a DataTable by uploading a spreadsheet file in CSV format. DEEDS parses the file content and provides an initial configuration for each column. Users can edit both DataTable content and column configurations to make them suitable to their needs. Configuration metadata include data types, descriptions, labels, units, visibility, and visual appearance. DataTables also support customization for specifying blank, not applicable, or other specialized entries.

Multi-dimensional DataTables are collections of DataTables linked in parent-child fashion. A column in a parent DataTable can be defined to link to another DataTable, and this link represents an extension of the column into another spreadsheet. Such representations require that both parent and child DataTables have the same number of rows (i.e., records from the same research unit). Multi-dimensional DataTables are suitable for repeated measures or frequent data observations from the same research units, such as data collected on different days or in different experiment phases for the same specimen types.

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 component provides many user-friendly features such as copying data templates, propagating metadata, and bulk updates to make it easy to establish consistent structure and annotation. Users can easily update, browse, and query the stored data using familiar tabular interfaces presented in the DEEDS dashboard and explorer views.

III. Tools Component

DataTables and File repositories serve as the foundation for DEEDS data preservation. The Tools component provides a mechanism for users to build software applications that operate on DataTables and Files to further their research efforts. Software applications can range from simple analysis programs to highly sophisticated scientific applications written in a number of languages.

DEEDS provides a simple web form interface where users define tools and their metadata, which include input files, any number of arguments (and their data types), and any number of output files. DEEDS provides infrastructure for hosting and documenting versions of an application throughout the development lifecycle. If additional software modules are required to execute the application these may also be declared. 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. Users can request that DEEDS automatically upload any portion of the outputs to their datasets. Once uploaded, results can be examined using built-in DEEDS viewers for images, spreadsheets, text files, etc. Users can also download output from DEEDS to their local computer.

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, and DEEDS makes use of standard HUBzero technology [4] for accessing HPC clusters. DEEDS provides free access to a small queue and users can access their own queues with minimal configuration.

Once an application has been defined and configured, it is made available for launching by authorized users. Users are guided through a step by step process to specify all information required for execution based on the tool definition. Selection of input is first. 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 dataset files. If the desired input is independent of all cases, case selection can be bypassed. It is important to note that case association is maintained when results are auto-uploaded by DEEDS back to the dataset. Tool arguments are specified next. Arguments are initialized to default values and can be modified as needed. Next is the specification of output handling. Users can identify results to auto-upload, and files that are not auto-uploaded may be uploaded at a later date. Separate options are provided for auto-uploading standard output, standard error, and status files. After selection of execution resources, the user launches the tool and attention is directed to the tool workflow section.

DEEDS maintains a detailed history of computational workflows that link together input, application, and results. Additional metadata identifies when and where a tool was run and who ran it. Workflows can be shared among dataset users to promote collaborative discussion. Upon launch, an entry for the execution is made in the workflow table with unique ID and name, and users can edit names to provide more context for the workflow. Additional metadata identify execution status, input, arguments, list of output, etc. Status is updated as execution progresses, with trace information available to the user in real time. Workflow entries remain in the table unless deleted by a user, thus providing results traceability. Detailed logging of workflow information also promotes reproducibility of results.

IV. The EcoTox Study

DEEDS science domain partners are a vital force in the development of platform user interfaces and features. In this section, the EcoTox research group describes their experiences in using the platform and contributing to its design.

EcoTox projects focus on evaluating the ecological risk posed by per- and polyfluorinated alkyl substances (PFAS) [3]. PFAS are a group of contaminants of emerging concern that are slow to break down in the environment, frequently detected in water, soil, animal and human tissues, and can adversely affect the health of humans and wildlife. Thus, there is great interest in understanding PFAS toxicity from both ecological and public health perspectives, and many parties stand to benefit from our data, including academics, regulators, policymakers, and ultimately, the general public. The DEEDS platform supports our own internal workflows by providing a link between the raw data we upload and the tools we use to process and analyze these raw data. DEEDS also provides a way for interested parties to understand these links and if necessary, to reproduce our work. In doing so, it allows us to better serve other stakeholders by

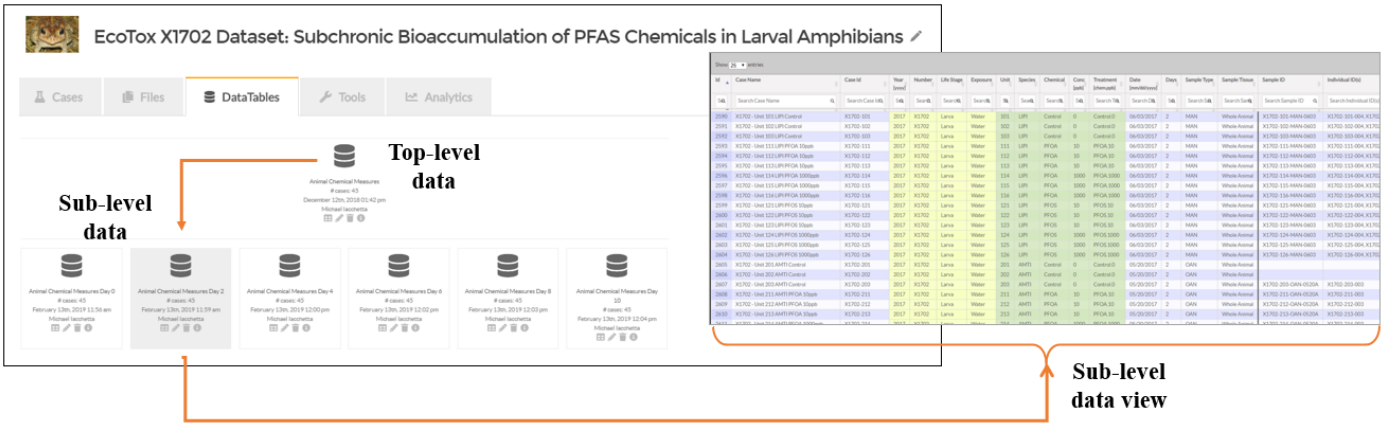

Fig. 1. Hierarchical overview of structured data in DEEDS DataTables. Starting at the top-level, Animal Chemical Measures data are composed of 6 sub-levels that correspond to animal chemical data from each of the six sampling dates. Clicking down to 'Day 2' users see the collected data for that sampling day.

ensuring the quality, integrity, transparency, and accessibility of the data we collect.

The use case presented here and selected to aid in the development and optimization of a generalizable DEEDS platform is a study determining how PFAS exposure affects the accumulation of chemicals and the health (phenotype: size, mass, developmental stage) of amphibian larvae [5][6]. Due to the nature of our study and the associated experimental constraints, our data structure is quite complex. A key element of our work on DEEDS has been the ability to add and adjust functionalities in an adaptive and iterative fashion, as new needs arose or became apparent. Thus, in addition to describing the nature of our data and its representation in DEEDS, we highlight examples of how the iterative process of building a custom data structure for our project has led to new functionalities in the DEEDS environment itself.

In aquaria, groups of salamander, frog, and toad larvae were exposed to water containing 0, 10, 100, or 1000 ppb of one of two common PFAS. The experimental units in this project are the aquaria, which are our 'cases' in DEEDS. (i.e., rows). The structure of our data and any files associated with cases (e.g. scanned datasheets, standard operating procedures, husbandry records) can all be readily assessed from this top-level display. Observations of endpoints on individual larvae subsampled from each experimental unit were recorded at six time points. From the top-level, each case is directly linked to a hierarchy of tables containing all data collected for various endpoints and at various phases of our study.

These endpoints are divided into 'phenotypic' and 'chemical' datasets. Phenotypic data were always collected at the individual level, while chemical data can represent either a single individual or a pooled sample containing multiple individuals when individual masses were too small for analysis. Additionally, we also collected chemical data quantifying levels of PFAS in the media (i.e. water) of each aquaria to ensure the quality of the data and improve transparency. Phenotypic, animal chemical, and media chemical datasets are all further structured by sampling date (i.e. Days 0-10).

A key realization as we addressed repeated measures (defined here as instances where the same measures were quantified at multiple time points, which occurred for both phenotypic and chemical data) was that independently creating tables for multiple sampling dates containing data of the same formats was not only inefficient, but had potential to lead to inconsistencies in the metadata recorded with these observations. Per our request, DEEDS developers built a feature that allowed us to build repeated measures tables once, and then clone data format, metadata requirements, and display features (like column colors and borders) to create tables for other sampling dates. This not only saved us time in data entry, but also ensured that data structure and metadata were identical across repeated measures tables. In addition, DEEDS developers were able to provide a feature that tracks when copies of tables are created and modified, to ensure traceability of any changes to data or their structure during this process.

By moving through the hierarchy of tables in DEEDS, users can easily understand how the experiment was designed. For example, we measured concentrations of the focal PFAS compounds in animals on days 0, 2, 4, and 6, 8, and 10 of exposures, which is immediately evident from the hierarchy (Fig. 1). It is then possible to enter each of these tables and access the observation level data itself. Furthermore, a file storage feature has allowed us to upload scanned raw data sheets underlying data displayed in the tables, as detailed documents describing the methods used. In short, DEEDS has enabled us to preserve and communicate a complex and hierarchical dataset, while the adaptive and collaborative approach to designing the database has led to new DEEDS functionalities that serve our needs directly.

Our dataset consists of multiple endpoints measured at both the 'case', or experimental unit level, and at the individual level across multiple sampling dates. Our primary objective is to assess whether PFAS chemical treatments result in effects on the size (mass and snout-vent length), developmental stage, and bioaccumulation of PFAS (body burden). All analysis is conducted using R scripts [7] and easily integrated into DEEDS Tools. We first test for effects of chemical treatments using linear models often incorporating random effects. Generalized and mixed effect models in the R-package nlme are flexible statistical tools that readily accommodate datasets containing not fully independent data (i.e. individuals within experimental units and sampling experimental units multiple times) and different statistical distributions. Additionally, we have several

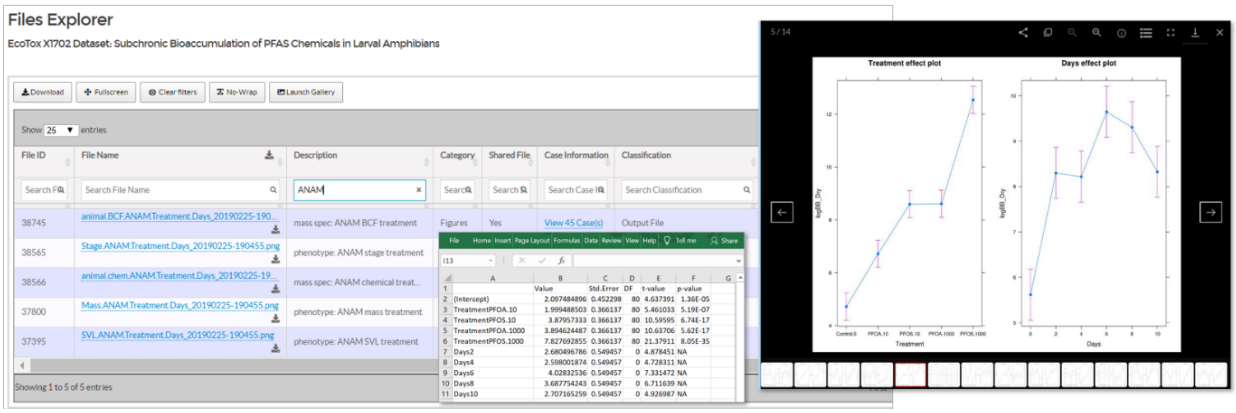

Fig. 2. Using Files Explorer to study Tool output for chemical measures after filtering for one species: 1) spreadsheet summary of the mixed linear model examining effect of chemical treatment on accumulation of PFAS across sampling dates and 2) visualization showing means and confidence intervals derived from the model.

'derived' endpoints that are calculations of the raw data, such as bioconcentration factor (BCF) and effect of chemical treatment on accumulation of PFAS (Fig. 2). DEEDS Tools allow us to maintain the original data structure, while providing the exact R-code and data used to calculate derived values. Together, Tools ensure our results are repeatable and transparent for others to easily assess how we determined effects.

As with the data tables themselves, Tool development has been an iterative and adaptive process. One hurdle that arose was that the data in DEEDS data tables is not necessarily formatted appropriately for downstream analyses. In this case, DEEDS developers provided a function allowing for the rapid conversion of DEEDS DataTables from a 'wide-format,' where each experimental unit is a row that can have multiple observations, to a 'long-format,' where each row is an observational unit. After this, it became apparent that end users unfamiliar with the statistical models we use, the code to implement them, and the underlying data used to parameterize models might have trouble reconstructing our analyses verbatim. DEEDS developers were able to build functionalities to address these issues.

After EcoTox investigators wrote and validated necessary R codes, DEEDS developers designed Tools functionalities that generalized these codes for use on the platform and directly linked to the data used to run the analyses. Ultimately, this led to Tools that replicated analyses presented in publications and other work, while allowing other users to further explore the data by downloading and modifying the analysis scripts, if desired. The result is a point-and-click interface where any of the analyses we conducted can be recreated, such that end users do not need to download datasets to local disks nor implement code in the R environment (though that functionality is retained). We see this as a major innovation relative to current practices, which often include publication of underlying statistical codes and/or raw data files in more traditional, static databases (e.g., supplemental sections of published manuscripts or open access databases like Dryad [8]). This functionality will also facilitate future analyses conducted by our group by allowing us to proceed through the entire process of data selection, analysis, and reporting without the need for downloaded data files or implementation of R code outside of the DEEDS environment, thus increasing traceability.

Creating effective visual representations of our data is also essential to the success of the project as such figures are ideal for presenting summaries of results. We generate these figures using the R-package ggplot2. Much like the analyses described above, providing open access to our data and code allows others to see exactly what data went into each figure and facilitates sharing and learning by others who may want to construct similar plots in their study. By using the open access program R, any person with access to a computer can use any of these resources we provide. Further, because DEEDS provides these as a point and click interface similar to that described above for statistical analyses, users can accomplish all of this directly within DEEDS, without downloading data or implementing R code externally, as would be required in the absence of this unique and innovative functionality.

V. Conclusion

The DEEDS platform provides integrated data and computing services to help researchers preserve, manage and share their work—with collaborators during their investigations and with global communities after publication. The interactions of use case researchers with the platform has provided vital feedback and important new ideas which have been considered and implemented in ways that are meaningful and useful to projects from many disciplines. The EcoTox experience is an example of how DEEDS supports and is supported by project researchers. New capabilities under development include ad-hoc analytics and summary statistics based on R, integration of Jupyter notebooks [9] that can access dataset data directly, and new applications for operating on and visualizing geospatial and other important data types.

Acknowledgment

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

References

[1] Catlin AC, HewaNadungodage C, Clark S, Fernando S, Wickramaarachchi G, Bejarano A, Desigavinayagam P, Patil O, “Fully Integrating Data with Compute Workflows: A Platform to Better Serve Scientific Research,” *The 13th Gateway Computing Environments Conference*, University of Texas at Austin, 2018.

[2] Catlin AC, HewaNadungodage C, Bejarano A, “Lifecycle Support for Scientific Investigations: Integrating Data, Computing, and Workflows,” *Computing in Science & Engineering*, *Special Issue: Scientific Workflows* (In Press) July/August 2019 doi: 10.1109/MCSE.2019.2901433

[3] Sepulveda MS. [Online] Available at https://www.serdp-estcp.org/index.php//Program-Areas/Environmental-Restoration/Contaminated-Groundwater/Emerging-Issues/ER-2626

[4] McLennan M, Clark S, Deelman E, Rynge M, Vahi K, McKenna F, Kearney D, "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.

[5] Abercrombie, SA, De Perre C, Jeong Y, Tornabene BJ, Sepúlveda MS, Lee LS, Hoverman JT, "Larval amphibians rapidly bioaccumulate poly- and perfluoroalkyl substances," *Ecotoxicol. Environ. Saf.* 178:137–145. Elsevier Inc. 2019.

[6] Hoover GM, Chislock MF, Tornabene BJ, Guffey SC, Choi YJ, De Perre C, Hoverman JT, Lee LS, Sepúlveda MS, "Uptake and depuration of four per/polyfluoroalkyl substances (PFASS) in Northern leopard frog Rana pipiens tadpoles," *Environ. Sci. Technol. Lett.* 2017.

[7] R Core Team "R: A Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria. 2013.

[8] Dryad Digital Repository [Online] Available at https://datadryad.org/

[9] Jupyter Development Team, "Jupyter Notebooks – a publishing format for reproducible computational workflows", *Positioning and Power in Academic Publishing: Players, Agents and Agendas* pp 87-90, 2016, doi: 10.3233/978-1-61499-649-1-87