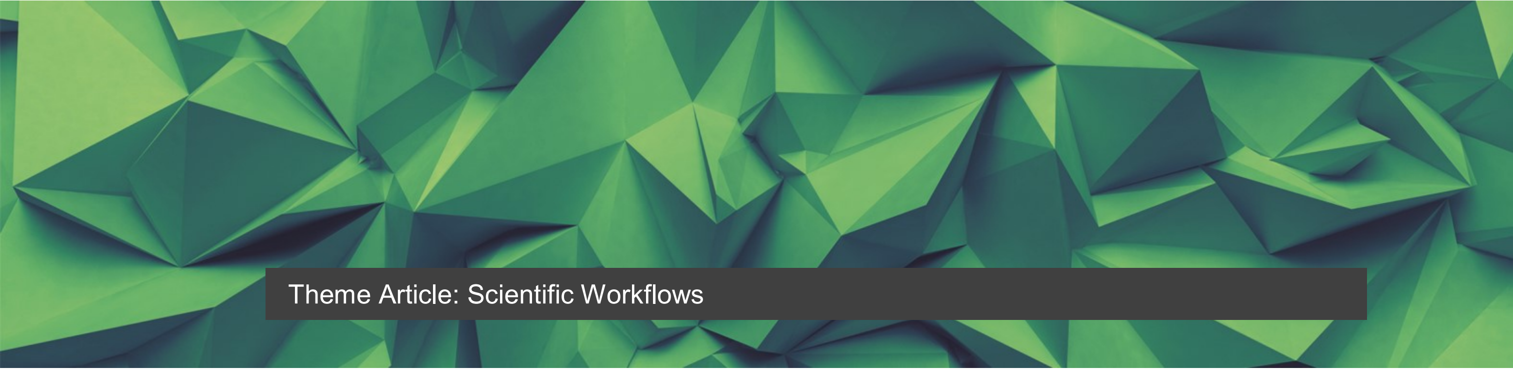


Lifecycle Support for Scientific Investigations: Integrating Data, Computing, and Workflows

Authors

Ann Christine Catlin

Chandima HewaNadungodage

Andres Bejarano

Scientific workflows have emerged as a model for representing the complex processes carried out by scientists throughout their investigations, encompassing research activities corresponding to data collection, data flow, computation, output analysis, and all the ways these are used together to produce results. Existing infrastructures support elements of the workflow, such as data repositories or computing services, but these are not integrated as interactive environments that provide full investigation lifecycle support. The Digital Environment for Enabling Data-driven Sciences (DEEDS) project brought together domain scientists and computer scientists to create a platform that provides interactive end-to-end support for diverse scientific workflows. Key among requirements were preservation, provenance, coupling of data and computing, results traceability, collaborative sharing, exploration, and publication of the full products of research. This paper highlights use cases that contributed to DEEDS development, and concludes with lessons learned from a process that joined experiences and perspectives from diverse science domains.

Keywords: Knowledge management applications, Database management, Workflow management, Information technology and systems, Information sharing, Interactive data exploration and discovery

INTRODUCTION

Workflows capture the complex processes carried out by scientists throughout their research investigations, encompassing their data, computations and results analysis, as well as the dependencies that connect them and the human decision-making factors that control them. From an abstract perspective, workflows can be considered views or representations of scientific work, and thus workflows have emerged as a model for representing and managing scientific activities [1]. Workflow models are based on the formalization and organization of scientific processes, together with standards for the metadata that describe them and the operations that are applied to them. As the sciences become increasingly data and compute intensive – and the mechanisms required to carry out workflows more complex – domain scientists are looking for advances in support of workflow models as imperative for advancing their own work. In response, hundreds of workflow models have been implemented as packages, frameworks, platforms or infrastructure components [2] to serve the needs of scientific research. Some focus on executing computational workflows that consist of large-scale, distributed computing tasks and exchange of data [3]. Others provide customized platforms for discipline-specific workflows, such as the rich environments for bioinformatics researchers [4].

For a great many domain scientists, however, existing packages and infrastructure components do not offer a complete solution, so they continue to use a combination of manual and ad hoc methods throughout their investigations [2]. Relying on ad hoc methods means that scientists responsible for different parts of the research effort (collecting data, writing code, analyzing results) operate in disconnected environments. This leaves data, computations, workflows and results fragmented and unsuitable for preservation, replication, sharing, and reuse.

The Digital Environment for Enabling Data-driven Sciences (DEEDS) project brought together scientists from computational chemistry, electrical engineering, environmental science, and nutrition science to articulate, clarify, and formulate requirements for workflow support. The partnership between domain scientists and computer scientists created DEEDS, a cross-domain, web-based platform that provides end-to-end support for diverse scientific workflows, incorporating data management services, computing services that connect data to computational tools, and services for analyzing and visualizing results.

The DEEDS workflow model uses an experimental or case-based approach, and represents the joined experiences and perspectives of science domain partners. The platform provides a framework for defining and organizing research activities as a shared DEEDS dataset. Scientists can upload, annotate, and manage data, where data can be defined as collections of files or as hierarchical spreadsheet-like data tables. They can define computing tools along with their required input and execution resources. DEEDS guides them through launch and execution of tools, and automatically uploads, annotates, and classifies generated output. Computing workflows are captured, preserved, and annotated end-to-end.

Requirements of the domain scientists made it clear that platforms supporting workflow models should, in addition, support a variety of other models in order for the working environment to provide value and benefit. Key among these are preservation, provenance, collaborative sharing, and results traceability.

One model not commonly associated with workflow support is a model for publication and dissemination of research data, tools, and workflows. In general, venues that publish the products of research are disconnected from the platform where the research was conducted. As a result, the process for publication cannot make use of the platform where data, computations, and workflows were assembled, preserved, and shared throughout the investigation. DEEDS datasets support both day-to-day investigative workflows and their publication for discovery and reuse. This offers significant advantages for the completeness and integrity of the published dataset.

In this paper, the diverse nature of the use cases that contributed to DEEDS requirements is described, followed by a discussion of the DEEDS workflow model and corresponding standards for metadata and operations. We share some important lessons learned during the development process and conclude with a brief look at implementation.

USE CASES

The partnership of computer scientists and researchers from diverse science domains was key to the development of a platform that provides support throughout the investigation lifecycle for distinct scientific workflows. Specific research projects were used for characterizing types and forms of data, data flows, computing software, results analysis, and shared interactions. Use cases that guided and validated DEEDS requirements are:

- **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 [5].
- **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 [6].
- **Electrical Engineering [Solar PV]:** 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 [7].
- **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 [8].

These research groups had used mostly ad hoc methods to conduct their investigations. The expectation of concrete improvements to their scientific workflows fostered a constructive environment for the exchange and unification of ideas and experiences. These would define the features and functionality of DEEDS, as well as test its completeness and usability.

Spreadsheet Data Collection and Statistical Modeling

The **EcoTox** and **Berries & Bones** investigations follow a common scientific workflow where statistical modeling is applied to large volumes of collected data corresponding to experimental units or cases that parameterize their study. Investigations begin with experiment design and data gathering protocols. Measurements and observations for each case are collected into Excel spreadsheets from which data are selected and extracted as input for modeling. Researchers interpret results and identify approaches for further modeling and analytics, leading to outcomes described by plots and statistics.

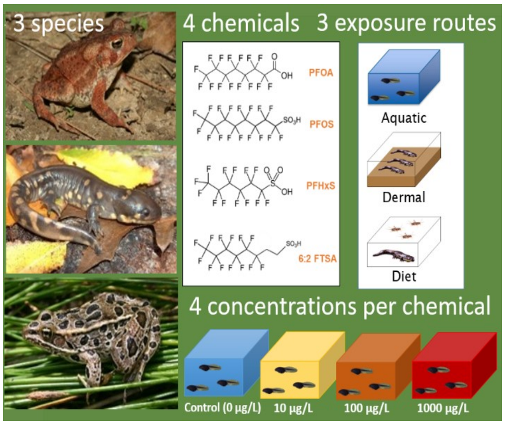


Figure 1(a). Mass spectrometry and phenotype measures are repeated over the course of the EcoTox experiments for each aquarium.

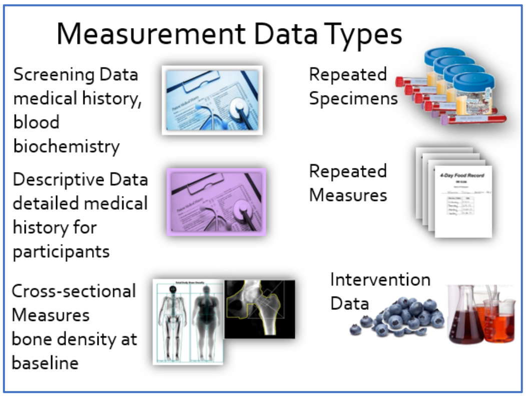


Figure 1(b). Heterogeneous data types are collected for Berries & Bone subjects within and across multiple treatment intervention and recovery phases.

EcoTox experimental units are aquariums parameterized by species, chemical, concentration, and exposure route, including replication within and across repeated experiments (Figure 1a). The Berries & Bone clinical trials collect data per study participant, repeated within and across multiple treatment and recovery phases (Figure 1b). Over the course of their investigations, tens of thousands of data points are collected into spreadsheets. Study spreadsheets had previously been shared manually via Email, Sharepoint, or Dropbox. These methods do not allow spreadsheets to be updated, tracked, validated, searched, or explored in an interactive, access controlled, team-shared way. EcoTox modeling scripts are written and executed by team analytsts (using R). Berries & Bone analysts use SAS. Analysts previously assembled input and ran statistical models on their own, and then shared results via Email or reports. This process precludes the automatic tracking and preservation of input selection, model execution, and linkage to generated output, which are vital ingredients of results traceability. The full research team could not interactively review, validate, explore, or reproduce computing workflows and their output, since the collected measurements, algorithms, workflows and output remained untracked and in separate environments. The full products of research could not be published for scientific communities to explore and reuse, since scientific activities carried out throughout the investigation had not been effectively captured and preserved.

The DEEDS partnership with the EcoTox and Berries & Bone research groups is advancing their day-to-day scientific workflow with a team-shared dataset where researchers can

- Upload, annotate, and preserve all files produced by their research effort (e.g., protocols, reports, diagrams, device-generated mass spectrometry files)
- Define, update, validate, and explore multi-dimensional heterogenous spreadsheet-based data as online, interactive data tables
- Upload and preserve modeling code
- Launch and execute models, including input selection, execution trace, and automatic retrieval, upload, and annotation of generated output
- Capture, preserve, and view computing workflows that link input, models, and output for results traceability and data provenance
- Publish the full products of research for public or access-controlled discovery and re-use

For modeling software that is uploaded and executed through DEEDS, researchers benefit from DEEDS services that capture computing workflows end-to-end. However, Berries & Bone researchers are still able to link their input, models, and output for results traceability by uploading SAS results to DEEDS and then constructing their own workflows in DEEDS to connect input and output.

As an example of how science domain requirements shaped the DEEDS platform, let's look at the spreadsheet-based data model – a central supporting structure for such scientific workflows. Spreadsheets assembled during an investigation represent a hierarchical, multi-dimensional data model comprising classes of measurement variables, their relationships, and their repetitions. DEEDS needed to represent spreadsheet-based data in a way that clarified the structure and relationships of the data and also connected data to all necessary metadata, both for annotation and customized viewing. Our domain scientists wanted comprehensive operations to support interactive web-based upload, update, view, filter, and exploration. In fact, they wanted DEEDS to retain all the flexibility, efficiency, and familiarity of Excel spreadsheets, and to this was added: 1) data validation, 2) access-controlled team sharing, and 3) tracking of data selected for model input.

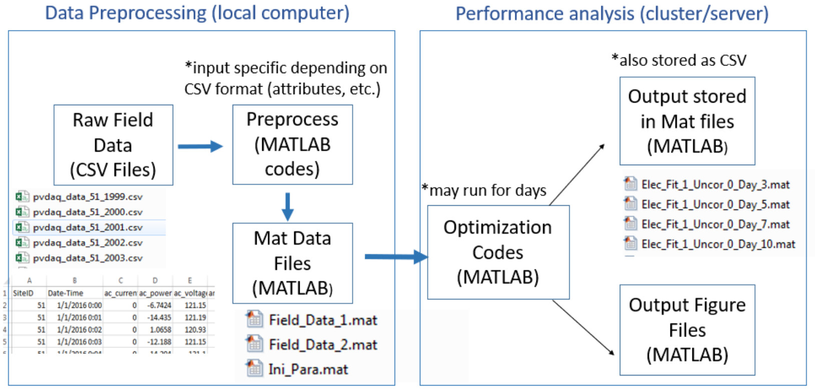
The result is the DEEDS "DataTables" component that represents spreadsheets as interactive data tables. Scientists define data tables by uploading research spreadsheets, with or without data. Researchers assign column properties for annotation (e.g., description, units) and custom viewing (e.g., labels, color, display format). Data values can be updated by interactively editing cells or uploading new spreadsheets. To support relationships between spreadsheets of data, DEEDS data tables can be hierarchical. For example. a data table representing a class of measurements can define separate columns to represent each set of measurements collected per day. Each linked column connects to a new spreadsheet-based data table containing the measurement variables and their values. Researchers can display customized, searchable views of data tables, they can pass between data tables, and they can drill down to any level of the hierarchy to view linked data tables. Data can be selected interactively and stored as DEEDS files for input to models.

File Collections and Research Computing

From a high-level perspective, the **Solar PV** and **Quantum** investigations follow similar scientific workflows. Researchers identify and characterize cases for the investigation and classification of physical properties. They prepare large collections of input files describing initial parameters, conditions, and other data needed to study each case, and these input files are fed to algorithms that require high-performance scientific computing. Output data are examined, executions are repeated with varied input, and results are collected for interpretation and analysis.

The Solar PV research group is diagnosing the health of solar farms characterized by panel technology and field data. Their algorithms use physics-based degradation models to analyze panel performance, and input to the algorithm is constructed by extracting and correcting raw field data from the National Renewable Energy Laboratory (NREL). Their ad hoc workflow (Figure 2a) begins with creation of input files assembled by individual researchers and shared via Email or USB flash drives. Researchers advance and test computing code in separate environments, without systematically preserving code versions and version-specific input and output. Executions are launched manually on HPC clusters, with manual transfer of input and output. Results are shared via Email and reports. Collection of files (raw data, corrected data, output), computational software, and computing workflows had not been preserved or published.

Quantum chemistry researchers optimize molecular structures and investigate their properties using licensed Gaussian^{TM} computational chemistry software [9] installed on target HPC clusters. One use case is investigating protein receptor activation, where receptors are classified by the chemical species which bind to them. Initial structures are defined for twelve molecules across three classes and used as input for Gaussian optimization (see Figure 2b for class 5-HT$_{2A}$). Results data are extracted from Gaussian LOG files for analysis. Ad hoc workflows require researchers to manually launch Gaussian, demanding knowledge about HPC resources and operation. Researchers transfer input and output files manually. Results and interpretations are shared via Email and reports; program input and output had not been preserved or published.


Data Preprocessing (local computer)
Performance analysis (cluster/server)
*input specific depending on CSV format (attributes, etc.)
Raw Field Data (CSV Files)
Preprocess (MATLAB codes)
pvdaq_data_51_1999.csv
pvdaq_data_51_2000.csv
pvdaq_data_51_2001.csv
pvdaq_data_51_2002.csv
pvdaq_data_51_2003.csv
Mat Data Files (MATLAB)
Field_Data_1.mat
Field_Data_2.mat
Ini_Para.mat
*may run for days
Optimization Codes (MATLAB)
*also stored as CSV
Output stored in Mat files (MATLAB)
Elec_Fit_1_Uncor_0_Day_3.mat
Elec_Fit_1_Uncor_0_Day_5.mat
Elec_Fit_1_Uncor_0_Day_7.mat
Elec_Fit_1_Uncor_0_Day_10.mat
Output Figure Files (MATLAB)

	A	B	C	D	E
1	SiteID	Date-Time	ac_curren	ac_power	ac_voltage
2	51	1/1/2016 0:00	0	-6.7424	121.15
3	51	1/1/2016 0:01	0	-14.435	121.19
4	51	1/1/2016 0:02	0	1.0658	120.93
5	51	1/1/2016 0:03	0	-12.188	121.15

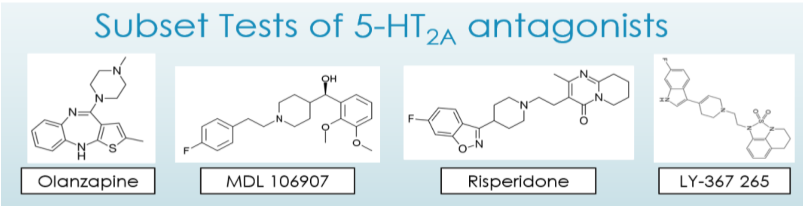


Figure 2(a). To fully capture Solar PV scientific workflows, the dataset encapsulates all processes, data, and metadata – ensuring preservation and provenance of data, data flows, computing code, and study methods (top).

Figure 2(b). The Quantum dataset cases come naturally from classes of receptor proteins, and span the types of activity (agonist, antagonist). One case is the 5-HT$_{2A}$ antagonists with four example molecules (bottom).

DEEDS support for tools and their execution was guided by our research computing use cases, which had explicit and comprehensive requirements for user interfaces, data flows, and execution support. The result is the “Tools” component where scientists can define and interactively launch computing tools from the dataset dashboard. Each tool execution is captured, preserved, and displayed with full metadata for results traceability including input selected and output generated. The tool definition interface and versioned repository allow researchers to define tools, enter information for input, output, and execution resources, and (where applicable) upload code archives with programming language and library requirements so that DEEDS can prepare code for execution.

Tools defined to DEEDS may be under development (Solar PV algorithms) or may exist as packages at target sites (Gaussian). With permission, DEEDS tools can be imported to any dataset for execution by other scientists and can also be downloaded by new research groups to modify for their own research. In fact, researchers can attach code such as post-processing scripts to existing DEEDS tools to enhance functionality and add value. Quantum researchers added a post-processor which extracts key results from Gaussian LOG files into CSV files (e.g., vibrational data) and figures (e.g., P depolarization spectrum). Post-processed output is now automatically retrieved and uploaded by DEEDS for all Gaussian users, who can then use DEEDS applications to explore them.

Research computing use cases depend on DEEDS for systematic preservation of input and output files, which can be viewed and explored within their team-shared dataset. Solar PV researchers also use DEEDS to share pre-processing methods for correcting raw NREL data. This code is solar farm specific and thus is not defined as a tool, but the site-specific code has been fully documented with example files and uploaded to their dataset to provide valuable insight into algorithm input.

Table 1. DEEDS datasets provide a "live" interactive shared environment for use case researchers to work with their data and computing tools throughout their investigation lifecycle.

Use Case	**Dataset, Cases**	**File Collections and Data Models**	**Tools**
Solar PV	One dataset, each case represents one solar farm defined by PV technology and NREL-based field data	File collection: Input: field data, solar panel parameters Output: predicted efficiency & circuit parameters over time (CSV, figures)	Research code (MATLAB) with ongoing algorithm advances
Quantum	Dataset for each new study, e.g., protein coupled receptor study where each case is a receptor binder class of selected molecules	File collection: Input: model chemistry, molecular system Output: optimized geometry, vibrational frequencies, charge data, thermochemical data, orbital information (CSV, figures)	Gaussian computational chemistry software
EcoTox	Dataset for each new study, cases for all studies represent aquariums parameterized by species, chemical, concentration, exposure route, replication	Data model: Aquarium descriptives, mass spectrometry measures (media, animal) animal phenotype (snout length, body mass, stage), mortality counts, all measures over time Output: reference data, plots, statistics	R codes to model, analyze, and report
Berries & Bone	One dataset, cases define either study subjects or treatment interventions	Data model: Subject descriptives, repeated specimens (serum, urine, feces), repeated measures (anthropometrics, compliance, nutrient intake) across treatment & recovery phases Output: blinded outcomes, plots, statistics	SAS to model and analyze

DEEDS PLATFORM STANDARDS

Organization and Structuring of Scientific Activities

The DEEDS workflow model is visually represented by the DEEDS dashboard, which characterizes research activities and data flows at a high level. Dashboard tabs identify workflow elements: Cases, Files, DataTables, Tools, and Analytics. At a lower level, each tab encapsulates the metadata, operations, and dependencies corresponding to each element. The DEEDS dashboard is the key enabler for interactively capturing, preserving, and connecting scientific work.

Scientists begin by creating a team-shared DEEDS dataset. The dataset integrates their data, tools, computational workflows, and results analysis – and also provides the interactive environment that supports their day-to-day research activities (Table 1).

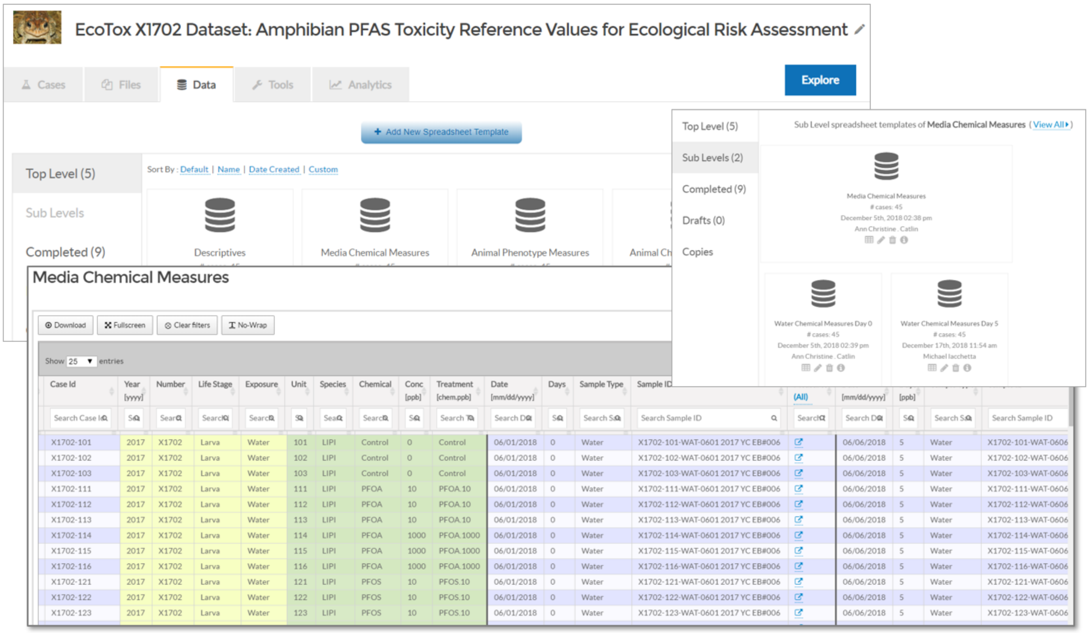
The DEEDS workflow model enforces a case-based approach, requiring scientists to first organize their investigation as cases, which can be modified and expanded over the course of the investigation. **Cases** can be experimental units, specimens, molecules, subjects (for clinical trials), buildings (for earthquake reconnaissance), plots (for crop studies), solar farms, or other study units that clarify how the research was conducted. This interpretive framework makes datasets easier to understand and use, since other elements of the scientific workflow correspond to the framework, connecting them directly to the research activities that produced them. Case metadata consists of case name, id, description, discovery keywords, spatial and temporal information, source (e.g., research laboratory), and DEEDS compilation information (who, when). Case operations are simple. Cases can be entered and edited on the Cases web-form, or they can be updated by spreadsheet uploads (when large numbers of cases need to be added or modified).

Capturing Data for Scientific Workflows

DEEDS supports two distinct forms of data used in scientific investigations: files and data tables. Dashboard management for files and data tables requires and consumes DEEDS metadata for preservation and provenance across the scientific workflow.

File Collections

Operations on file collections include upload, annotate, categorize, classify, download, manage, explore and transfer for execution. **Files** uploaded by users can represent information for one case or many cases. For example, input files for statistical models could represent data for a single case, a selected subset of cases, or all cases. Files are categorized at upload. Users select a file category from Reports, Data, Media, Figures, or Other, and can create new categories as needed to clarify the purpose or content of their files. Use case scientists might create a special category for mass spectrometry files or CT scans. File categories are controlled by assigned MIME types and permissible file extensions. This ensures the efficacy of operations to generate thumbnails or explore and visualize files with DEEDS applications. Users annotate and classify their files to provide better search and filtering capabilities and to provide better dataset interpretation for reuse when the dataset is published. File metadata consists of file name, description, category, classification, case mapping, status as a shared file, MIME type, user and timestamp for the upload, size, and generated thumbnails.



EcoTox X1702 Dataset: Amphibian PFAS Toxicity Reference Values for Ecological Risk Assessment

Sub Level spreadsheet templates of Media Chemical Measures (View All)

Media Chemical Measures
# cases: 45
December 5th, 2018 02:38 pm
Ann Christine . Catlin

Water Chemical Measures Day 0
# cases: 45
December 5th, 2018 02:39 pm
Ann Christine . Catlin

Water Chemical Measures Day 5
# cases: 45
December 17th, 2018 11:54 am
Michael Iacchetta

Media Chemical Measures

Case Id	Year [yyyy]	Number	Life Stage	Exposure	Unit	Species	Chemical	Conc [ppb]	Treatment [chem.ppb]	Date [mm/dd/yyyy]	Days	Sample Type	Sample ID	(All)	[mm/dd/yyyy]	[ppb]		
X1702-101	2017	X1702	Larva	Water	101	LIPI	Control	0	Control	06/01/2018	0	Water	X1702-101-WAT-0601 2017 YC EB#006	↗	06/06/2018	5	Water	X1702-101-WAT-0606
X1702-102	2017	X1702	Larva	Water	102	LIPI	Control	0	Control	06/01/2018	0	Water	X1702-102-WAT-0601 2017 YC EB#006	↗	06/06/2018	5	Water	X1702-102-WAT-0606
X1702-103	2017	X1702	Larva	Water	103	LIPI	Control	0	Control	06/01/2018	0	Water	X1702-103-WAT-0601 2017 YC EB#006	↗	06/06/2018	5	Water	X1702-103-WAT-0606
X1702-111	2017	X1702	Larva	Water	111	LIPI	PFOA	10	PFOA.10	06/01/2018	0	Water	X1702-111-WAT-0601 2017 YC EB#006	↗	06/06/2018	5	Water	X1702-111-WAT-0606
X1702-112	2017	X1702	Larva	Water	112	LIPI	PFOA	10	PFOA.10	06/01/2018	0	Water	X1702-112-WAT-0601 2017 YC EB#006	↗	06/06/2018	5	Water	X1702-112-WAT-0606
X1702-113	2017	X1702	Larva	Water	113	LIPI	PFOA	10	PFOA.10	06/01/2018	0	Water	X1702-113-WAT-0601 2017 YC EB#006	↗	06/06/2018	5	Water	X1702-113-WAT-0606
X1702-114	2017	X1702	Larva	Water	114	LIPI	PFOA	1000	PFOA.1000	06/01/2018	0	Water	X1702-114-WAT-0601 2017 YC EB#006	↗	06/06/2018	5	Water	X1702-114-WAT-0606
X1702-115	2017	X1702	Larva	Water	115	LIPI	PFOA	1000	PFOA.1000	06/01/2018	0	Water	X1702-115-WAT-0601 2017 YC EB#006	↗	06/06/2018	5	Water	X1702-115-WAT-0606
X1702-116	2017	X1702	Larva	Water	116	LIPI	PFOA	1000	PFOA.1000	06/01/2018	0	Water	X1702-116-WAT-0601 2017 YC EB#006	↗	06/06/2018	5	Water	X1702-116-WAT-0606
X1702-121	2017	X1702	Larva	Water	121	LIPI	PFOS	10	PFOS.10	06/01/2018	0	Water	X1702-121-WAT-0601 2017 YC EB#006	↗	06/06/2018	5	Water	X1702-121-WAT-0606
X1702-122	2017	X1702	Larva	Water	122	LIPI	PFOS	10	PFOS.10	06/01/2018	0	Water	X1702-122-WAT-0601 2017 YC EB#006	↗	06/06/2018	5	Water	X1702-122-WAT-0606
X1702-123	2017	X1702	Larva	Water	123	LIPI	PFOS	10	PFOS.10	06/01/2018	0	Water	X1702-123-WAT-0601 2017 YC EB#006	↗	06/06/2018	5	Water	X1702-123-WAT-0606

Figure 3. From the DEEDS dashboard, the DataTable element of the scientific workflow represents and manages data models for complex hierarchical multi-dimensional spreadsheet data.

Complex Structured Data

The **DataTables** element manages hierarchical, multi-dimensional data models that represent spreadsheet-based collections containing study measurements and observations. A DEEDS data model can consist of any number of data tables which can be simple (single spreadsheet) or hierarchical and multi-dimensional (columns of a top-level spreadsheet connected to sub-level spreadsheets). Any number of top-level data table columns can link to sub-level data tables at the next level of hierarchy, and this can repeat. To define their data model to DEEDS, scientists upload spreadsheets to create data templates. Column data is processed to determine column data types, and researchers are presented with an interactive data table interface where they can set properties for annotation and customized viewing. A column data type can be set as Link to Spreadsheet Template in order to link that column to a new sub-level spreadsheet.

The DataTables element provides full representation and management of scientific spreadsheet-based data models for use in scientific workflows (Figure 3). Operations include define (via templates), validate (for consistency of user data with assigned type), annotate, update, preview (to assess property settings before saving), view, explore and download. For update, DEEDS presents editable data table forms, where users can edit data cells, reset properties, or upload updated spreadsheets to add or modify data values. DEEDS guides users through definition and update of their data model, with rules for defining or replacing sub-level templates, managing empty cells, handling inconsistent data, and other conditions that enforce best practices but maintain usability. For exploration, users can click to display top-level views, with column drilldown links to display and explore lower levels. DEEDS column properties are useful for presentation and interpretation of the data model, since colors and vertical borders can be used to highlight the nature of repeated measures. Metadata for this workflow element include 1) case mapping, which is established at the top-level and inherited for all linked sub-levels, 2) column data type, and 3) column metadata for customization of viewing, such as label, description, units, color, width, alignment, display format. Consistent, compatible representation across the entire scientific workflow model is guaranteed by validating user actions at all levels, in particular delete for cases or spreadsheet templates.

Computing Workflows

The DEEDS **Tools** element represents, manages, and controls tools, executions, and captured computing workflows. Tools can either be defined within the dataset – making the user defining the tool its owner – or they can be imported from the full collection of DEEDS tools (access-controlled by owner). Tool definition requires substantial metadata, including detailed specification of input and output (description, format, extensions), arguments, requested execution resources, version, developers, date uploaded, and (if applicable) code archive, programming language, libraries, and other information needed for DEEDS to compile and establish execution support structures.

After tools are defined for the dataset, they can be launched from the Tools dashboard by authorized users. The DEEDS launch operation guides users through selection of cases under analysis, helps users select appropriate input and execution resources, and lets users identify which output should be uploaded automatically. When a tool is launched, a computational workflow is created and displayed. Workflow metadata includes execution status, tool and version, selected input, execution tracking data, trace files (e.g., stdout), output files generated by the tool, files uploaded to the dataset by DEEDS, and user and timestamp for the launch. All files connected to the workflow are directly accessible from the workflow display.

Exploration and Publication for Reuse

The DEEDS model for **Exploration** is applied uniformly across all elements of the scientific workflow, with consistent display formats and interactive user operations (Figure 3) for Cases, File collections, DataTables, Tools, and Workflows. The tabular display – called a “dataview” – presents element-specific rows (identified by case names, file names, tool names, or workflow names) and columns displaying data and metadata. Column metadata include customizable labels and hover over description. Tabular cells conform to viewing metadata such as color and number format. Cells can take users to new displays that are type dependent, such as drilldown to new tabular displays representing data in the next level of hierarchy, geospatial displays, or media galleries. The DEED exploration model is based on an extensible “data definition” language [10] that accesses data and metadata from the database and presents them as interactive views. The language defines columns of the display by identifying database table and field and applying display rules for type, property, formats, and operation, which are given as arguments to the column. Extensible data typing allows DEEDS to attach applications and operations to columns. Dataview layout can be pre-defined (e.g., Cases) or defined in real-time (e.g., DataTables).

The DEEDS model for dataset **Publication** operates across all elements of the dataset’s scientific workflow. The dataset creator selects which portions of the dataset to publish and can establish access restrictions for specific content. DEEDS applies the exploration model to published datasets, and datasets can also be downloaded as a “BagIt” bag using the file packaging format introduced by the Library of Congress for transfer of preservation quality digital content [11]. Packaging of dataset Files corresponds to the DEEDS repository structure, and multi-dimensional DataTables are packaged as CSV format files. All metadata documenting dataset content, formats, file categories, and hierarchical data table organization are stored in XML format following the Dublin Core Metadata Initiative [12].

Ongoing Work

The DEEDS project has completed its first year of NSF-funded effort. Throughout the next three years of the project, collaboration between computer scientists and domain scientists will continue in order to extend and validate the completeness and usability of DEEDS. A key focus during this period will be the development of DEEDS Analytics interfaces and operations, which will be based on creation of R data frames from CSV files or DataTables and application of R functions for ad hoc analysis and visualization.

A second major focus will be the extension of data sharing to web-based repositories and archives external to DEEDS for both import and export of files. To support interoperability across systems, DEEDS will first address the completeness and compatibility of its Files and DataTables metadata with respect to community standards and established FAIR principals [13]. Configurable data sharing options will be made available for transformations between the DEEDS metadata and file packaging formats and those of widely-used repositories such as Dataverse [14] and Zenodo [15].

LESSONS LEARNED

Before their partnership with DEEDS, use case researchers relied on ad hoc methods to manage data and computing. Their methods did not adhere to rigorous data standards, neither were their collected study data and computing workflows suitable for dissemination and reuse by broader scientific communities. Their previous methods, however, were well understood by them, and DEEDS would necessarily enforce new and unfamiliar standards, vocabularies and operations on their investigative processes.

Requirements analysis brought together different perspectives which were based on significantly different data, workflows, methods, experiences, and expectations. Some expectations were incompatible with each other and some were inconsistent with our proposed standards for a discipline-neutral environment. Many of the lessons learned during the development of DEEDS are based on merging, managing, and in some cases, capitalizing on these differences. Three specific examples are given, followed by a collection of general principles.

Example 1: Enforcing Standards

Datasets created in DEEDS have two fundamental aims: 1) support for end-to-end investigations that encompass data, tools, workflows, and results, and 2) publication for reuse. Previous work [10,16] has shown that the most effective datasets for reuse are organized according to case-based frameworks, so that files, data, and results can be connected to the activities that produced them. Some of our science domain analysts felt that the most useful data support would simply offer shared collections of thousands of measurements, without a framework describing how measurements were organized. This “measurements format” was most often used as input to their analysis tools.

In response, we discussed the advantages of a framework that clarifies experiment design and classifies measurements types, in particular to help with interpretation of dataset content, both for use within the project and later for reuse by scientific communities. While use case requirements were intended to have the highest priority for DEEDS design, we found the need to establish overriding principles that could not be compromised, such as the case-based framework. To satisfy “measurement format” needs, we proposed a feature for Analytics that lets users toggle case-based data into measurements format. But we recognized the need for education and training to emphasize the value of standards for effective dataset interpretation and reuse. This remains an ongoing endeavor.

Example 2: Results Traceability

A key goal of the DEEDS platform is the traceability of scientific results, which is implemented by capturing computational workflows that identify input, computations, and generated output. The end-to-end workflow is controlled by DEEDS. Analysts, however, often employ SAS and other packages that cannot be controlled from DEEDS. In this case too, researchers want to connect input to output for better interpretation and validation of results.

We determined that it was important to support variations of the original workflow concept, and this led to the creation of the DEEDS workflow "ratings system" to distinguish workflows with guaranteed DEEDS traceability from workflows with user-assigned connections from input to output. In this way, the original concept of "guaranteed traceability" remains immutable, but extensions and variations of workflow linkage are permissible since they are designated as such. Workflow discussions also led to the creation of a second ratings system that allows users to classify workflows (e.g., test, ad hoc, standard). This became key for identifying usage properties of workflows, such as which workflows to publish and which to use internally for educating researchers new to a project.

Example 3: Integrating Existing Packages

Workflows for research computing necessarily depend on underlying support for the execution of computational software, including the transfer and handling of input and output files. A command line driven "execution submit package" [17] was integrated into DEEDS to support the execution of tools. The submit package (hereafter referred to as submit) had been developed and advanced for nearly 20 years. It is robust and full-featured, but was not designed to support interactive user controls such as 1) interfaces where users select input and execution resources, 2) automatic upload of output, and 3) capture of execution workflows.

We underestimated the level of effort for integrating submit into the DEEDS platform in a way that satisfied all requirements for traceable, interactive workflows. This effort included: re-inventing the submit component as a web service, building an API for submit to communicate with DEEDS to interactively track execution, and creating session working directories with appropriate security features for submit operations. These capabilities had to be designed in ways that satisfied demands of both submit and computational software. To these efforts, we also added design iterations to ensure ease-of-use. The submit component was an existing and valuable asset for building DEEDS workflow support. However, the effort expended for its transition to usage in an interactive environment required an unexpected level of effort.

Collection of Lessons Learned from the DEEDS Experience

Based on our DEEDS experience, some recommended actions to help guide the design and development of cross-domain platforms for data and computing are listed below.

- Form a multidisciplinary team for requirements analysis and platform evaluation to ensure that all workflows and experiences are addressed directly and openly (different science domains, different kinds of users). Be prepared for inconsistent and incompatible feature requests. Platforms that provide data and computing services across science domains need to manage and merge conflicting experiences and expectations based on different workflow perspectives. Team members should change over time to address specific challenges in the development of workflow support.
- As part of the implementation process – at all levels of development – set up a way for domain scientists to be able to demonstrate to the full group how the system under development will be used for their data and workflows. This will help forestall costly mistakes in design and lead to valuable, eye-opening new ideas.
- Make time for selected R&D team members to work with domain scientists to gain as full an understanding of their data and computations as possible. Require these members to use the platform as the scientist would. This effort will have a huge payback in understanding researcher needs, including introduction of the right features and operations for ease-of use throughout the scientific workflow.
- Establish immutable principles to which the platform must adhere. Platform concepts and features will be challenged, but user perspectives that run counter to overriding principles need to be addressed without compromising the platform's guiding principles.
- Do not underestimate the effort for integrating existing packages that fulfill the need for specific workflow functionality. Even in an optimal situation where a needed component is already part of the foundation cyberinfrastructure, new requirements can impose time-consuming development tasks before the package is ready to become part of the new platform.

- Educate and train researchers in the benefits of standards, results traceability, and dataset reuse. Transforming ad hoc workflows followed by most research groups will not be easy, and ongoing educational workshops are needed for platforms like DEEDS to succeed.
- Do not overlook the value of commercially available solutions to help guide design of components needed to support scientific workflows. Commercial applications can provide important answers for specific components that are not addressed by past or ongoing research projects.
- During the long development process driven by use case input from domain scientists, continue to give presentations about future features (both immediate and long-term). It's easy to get caught up in deadlines for current implementation and testing tasks – and use case researchers in particular are forced to wait patiently for a usable first version. It's important and uplifting to make sure the end goals stay at the forefront for the full project group.

A BRIEF LOOK AT DEEDS IMPLEMENTATION

This section provides a brief look at how the DEEDS platform, dataset elements, and scientific workflows are implemented. DEEDS is developed on top of the Hubzero^{TM} cyberinfrastructure [18] using the LAMP stack (Linux, Apache, MySQL, PHP), with Javascript/JQuery for the front-end and CMS components for back-end management of interfaces and processing for dataset elements. Support for DEEDS Cases and Files is based on their implementation in DataHub, a platform designed to preserve and publish file repositories [10]. Representations, interfaces, processing, and metadata for interactive multi-dimensional hierarchical data tables in the context of scientific workflows is an innovative concept requiring foundational work to be described in a future publication.

The Tools component provides interfaces and services for defining and launching tools from the DEEDS dashboard and for tracking, preserving and managing execution workflows. Tools metadata complies with requirements of the "submit" package. A DEEDS API creates the submit execution support environment – these are indexed formatted text files required and parsed by submit to describe the tool (scripting language, source code to be compiled, existing executable), input requirements, execution sites/resources, library/module requirements, expected output, and operational rules. When a tool is launched, DEEDS communicates with submit to initiate and track its operations. Hubzero and the submit package carry out all the work for tool compilation/verification and tool setup/execution [18]. DEEDS currently executes tools locally on the web server (in session containers) and on Purdue University HPC clusters (serial, parallel modes). Currently available sites and manner of tool execution are based on use case needs, but DEEDS plans to take advantage of the full range of submit capabilities, for example execution on any accessible target site and support for Pegasus enabled computing workflows [17].

To capture the full scientific workflow, DEEDS defines metadata not only for individual dataset elements (repository files, data tables, tools, execution workflows) but also for the complex network of dependencies among elements as their use, status, and linkage continue to change throughout the course of the investigation.

Access to DEEDS is currently limited to our science domain partners for a production version deployed on the Hubzero infrastructure at https://datacenterhub.org. We expect to release DEEDS for public access in Fall 2019. We also plan to package DEEDS for deployment on any Hubzero infrastructure, whether hosted at Purdue University or built from the open source code at https://hubzero.org/services/opensource.

CONCLUSION

The DEEDS project brought together researchers from diverse science domains to formulate requirements for supporting their scientific workflows, and our partnership created DEEDS, a cross-domain interactive web-based platform that supports the full investigation lifecycle. DEEDS provides data services for file repositories and multi-dimensional data tables, it provides computing services that connect tools to data and execution resources, it tracks workflows to link data, algorithms and results, and it publishes the full products of scientific research for discovery and reuse.

Our paper highlights the use cases that contributed to building the workflow model, and describes standards for DEEDS operations, data flow, and metadata. Specific examples of the impact of use case requirements on DEEDS design are given, in particular support for spreadsheet-based data tables and results traceability. For each use case, previous ad hoc workflows are compared to DEEDS platform support across the investigation lifecycle. The DEEDS project plans to engage with new scientific communities and use cases to further advance workflow support and the project will continue to promote the use of platforms that help scientists preserve, share, disseminate, and reuse scientific research.

ACKNOWLEDGEMENTS

DEEDS is funded by NSF CIF21 DIBBs award #1724728. We thank NSF program director Amy Walton; co-PIs Ashraful Alam, Joseph Francisco, Marisol Sepulveda and Connie Weaver; science domain post-doctoral researchers and graduate students led by Tahir Patel, Reza Asadpour, Xingshu Sun, Mike Iacchetta, Robert Flynn, Ross Hoehn, Joanna Hodges, and Jonathan Monical; and members of R&D with major contributions: Steven Clark, Guneshi Wickramaarachchi, Sumudinie Fernando, and Parameswaran Desigavinayagam.

REFERENCES

1. NSF Workshop on the Challenges of Scientific Workflows, May 1-2, 2006. [Online]. https://confluence.pegasus.isi.edu/download/attachments/2031787/NSFWorkflow-Final.pdf?version=1&modificationDate=1254437518000&api=v2
2. Singh M. et al. “Scientific workflows: scientific computing meets transactional workflows,” NSF Workshop on Workflow and Process Automation in Information Systems: State-of-the-Art and Future Directions, May 1996, pp. 28-34.
3. Deelman E. et al. "The future of scientific workflows," *The International Journal of High Performance Computing Applications*, 32(1):159-175, April 2018. doi: 10.1177/1094342017704893.
4. Leipzig J. "A review of bioinformatic pipeline frameworks," *Briefings in bioinformatics*, 18(3):530-536, May 2017. doi: 10.1093/bib/bbw020.
5. Sepulveda M. [Online] https://www.purdue.edu/newsroom/releases/2016/Q2/purdue-researchers-awarded-2.5-million-to-study-effects-of-perfluoroalkyl-substances-on-amphibians.html
6. Weaver C. [Online] http://www.purdue.edu/newsroom/releases/2014/Q3/purdue-receives-3.7-million-to-study-blueberries-and-bone-health.html
7. Sun X. et al. “Real-time monitoring and diagnosis of photovoltaic system degradation only using maximum power point—the Suns-Vmp method,” *Wiley Online Library*, July 2018. doi: 10.1002/pip.3043.
8. Hoehn R et al. “Status of the Vibrational Theory of Olfaction,” *Front. Phys.,* 6:25, March 2018. doi: 10.3389/fphy.2018.00025
9. Frisch M et al. Gaussian 16, Revision A.03, Gaussian, Inc., Wallingford CT, 2016.
10. Catlin AC. et al. “A Cyber Platform for Sharing Scientific Research Data at DataCenterHub.*”* *IEEE Computing in Science and Engineering*, 20(3):49-70, May/Jun 2018. doi: 10.1109/MCSE.2017.3301213
11. Boyko A. et al. “NDIIPP Content Transfer Project:The BagIt File Packaging Format” 2009 [Online]. Available: https://confluence.ucop.edu/display/Curation/BagIt
12. Weibel A, et al. “Dublin core metadata for resource discovery. RFC 2413, IETF.” Sep 1998. [Online]. Available: dl.acm.org/citation.cfm?id=rfc2413
13. Wilkinson M. et al. “The FAIR Guiding Principles for scientific data management and stewardship.” *Sci. Data,* 3:160018 2016 doi: 10.1038/sdata.2016.18
14. King G. “An Introduction to the Dataverse Network as an Infrastructure for Data Sharing.” *Sociological Methods and Research*, 36:173–199, 2007.
15. CERN, [Online] https://zenodo.org/
16. Catlin AC., et al. “Fully Integrating Data with Compute Workflows: A Platform to Better Serve Scientific Research,” *Science Gateways 2018*, Sep 2018.
17. McLennan M. et al. “HUBzero and Pegasus: Integrating Scientific Workflows into Science Gateways,” *Concurrency and Computation: Practice and Experience*, 27(2):328-343, 2015. doi: 10.1002/cpe.32
18. McLennan M. et al. “HUBzero: A Platform for Dissemination and Collaboration in Computational Science and Engineering,” *IEEE Computing in Science and Engineering*, 12(2):48-53, 2010.

ABOUT THE AUTHORS

Ann Christine Catlin is a Senior Research Scientist in the Rosen Center for Advanced Computing at Purdue University. Her research is dedicated to the advance of web-based data platforms and the underlying infrastructure and technologies needed to support the acquisition, preservation, sharing, exploration, analysis and reuse of scientific and medical research data. She has collaborated with hundreds of researchers in communities of research and practice, both nationally and internationally, to architect and implement collaborative data and computing environments. Catlin has an MS in mathematics from Notre Dame University. Contact her at acc@purdue.edu.

Chandima HewaNadungodage is a Lead Software Engineer in the Rosen Center for Advanced Computing at Purdue University. She received her PhD in Computer Science from Purdue University. Her research interests include database management, data mining, and web-based cyber-infrastructures. She has contributed to design and development of data management and computing platforms for medical and scientific research groups. Contact her at chewanad@purdue.edu.

Andres Bejarano received a BSc and MSc in Systems Engineering and Computation from Universidad del Norte, Colombia, and a MSc in Computer Science from Purdue University, IN. He is currently a PhD candidate in Computer Science at Purdue University. His research interests are computer graphics, geometry, scientific visualization and UI/UX design. He is with the Rosen Center for Advanced Computing as a Research Assistant. Contact him at abejara@purdue.edu.