Data Curation with a Focus on Reuse

Maria Esteva
Texas Advanced Computing Center
Austin, Texas
maria@tacc.utexas.edu

Sandra Sweat
University of Texas at Austin
Austin, Texas
slsweat7@utexas.edu

Robert McLay
Texas Advanced Computing Center
Austin, Texas
mclay@tacc.utexas.edu

Weijia Xu
Texas Advanced Computing Center
Austin, Texas
xwj@tacc.utexas.edu

Sivakumar Kulasekaran
Texas Advanced Computing Center
Austin, Texas
siva@tacc.utexas.edu

ABSTRACT

A dataset from the field of High Performance Computing (HPC) was curated with the focus on facilitating its reuse and to appeal to a broader audience beyond HPC specialists. At an early stage in the research project, the curators gathered requirements from prospective users of the dataset, focusing on how and for which research projects they would reuse the data. Users needs informed which curation tasks to conduct, which included: adding more information elements to the dataset to expand its content scope; removing personal information; and, packaging the data in a size, a format, and at a frequency of delivery that are convenient for access and analysis purposes. The curation tasks are embedded in the software that produces the data, and are implemented as an automated workflow that spans various HPC resources, in which the dataset is generated, processed and stored and the Texas ScholarWorks institutional repository, through which the data is published. Within this distributed architecture, the integrated data creation and curation workflow complies with long-term preservation requirements, and is the first one implemented as a collaboration between the supercomputing center where the data is created on ongoing basis, and the University Libraries at UT Austin where it is published. The targeted curation strategy included the design of proof of concept data analyses to evaluate if the curated data met the reuse scenarios proposed by users. The results suggest that the dataset is understandable, and that researchers can use it to answer some of the research questions they posed. Results also pointed to specific elements of the curation strategy that had to be improved and disclosed the difficulties involved in breaking data to new users.

Keywords

Data curation; high performance computing; distributed collections architecture; data publishing and reuse.

JCDL '16, June 19 - 23, 2016, Newark, NJ, USA
Copyright is held by the owner/author(s). Publication rights licensed to ACM.ACM978-1-4503-4229-2/16/06…$15.00
DOI: http://dx.doi.org/10.1145/2910896.2910906

1. INTRODUCTION

Curation of specialized datasets obtained during multi-year research projects involve significant challenges both to the researchers that create them and to the data librarians that prepare them for preservation and publication. In this project, curation is broadly understood as the activities conducted to create data for a specific goal, and those involved in its organization, description, preservation and reuse. Researchers may wait until the research is finalized to organize and document their data, only to find out that these activities require more effort than expected, and that they need to refactor certain data creation processes to meet preservation and publishing requirements. In the case of data generated through research methods and conceptual frameworks that are only familiar to a small niche of researchers, librarians working in institutional repositories (IRs) may not have the domain knowledge to properly document it nor know which are the right venues to promote its reuse. For data that aggregates over time, currently, many IRs do not have the infrastructure to easily upload, store, and provide access to large data.

Evaluation of digital curation strategies is a multi-layered and complex process. Beyond assessing the elements of a curation program (e.g. metadata, data integrity, permanent identification, etc.) against best practices benchmarks, the impact of the strategy in relation to ease of data reuse is often evaluated after the fact, if and when the data is cited in a publication. This approach, which may entail years of waiting, is not necessarily a direct indicator of the success of the curation strategy, nor does it provide opportunities to learn what aspects of the strategy are not working and improve them while solutions can still be implemented.

We present the curation of a growing High Performance Computing (HPC) dataset in which curation tasks were determined based on best practices as well as from requirements gathered from potential users in relation to data reuse. The data is produced by a software called XALT [1], which tracks and collects information about applications used by researchers on open-science HPC systems, or supercomputers.[1] Development of the XALT software is supported by a federal grant, and the resultant data ought to be preserved and made publicly available. The case involved all the curation challenges mentioned above,

[1] HPC resources are also known as supercomputers.

which are generalized across data from various domains, as well as specific issues related to this particular data understandability and scope of content.

The research team involved curators from the project's outset. As members of the team, we could evaluate the project at an early stage, in which suggesting changes is expected and implementing them is easier to accomplish. This allowed enough time to learn about HPC concepts and to design curation tasks that were integral to the data creation workflow. A major part of the curation process involved seeking feedback from prospective users, as it became evident to the team that the questions emerging about how to curate the data would be best answered by those that would reuse the data. As a result, the users' responses informed the traditional curation tasks of organizing, documenting, and designing access and preservation strategies, as well as decisions to shape the content of the data. The curation solutions were incorporated in the open source software that generates the data, so they are available to all the software's users.

The feedback obtained from the users also helped to select the outlet for data publication, which led to a successful collaboration between The Texas Advanced Computing Center (TACC) and the Texas ScholarWorks IR [33]. We implemented a distributed collection architecture [14] that spans computational resources and policies from both organizations, within which data is gathered, curated, preserved, and published on a quarterly basis through an automated workflow. This facilitates the work of both parties, and leaves room to make improvements between data installments. Such collaboration can be replicated for future cases of large data that is generated: it can now be stored at a supercomputing center and published through an IR.

Targeted curation also led to the design of an evaluation strategy focused on data reuse. We used data mining and statistics to create proof of concept analyses following basic reuse scenarios suggested by the prospective users. Upon the analyses completion and interpretation, changes were implemented and published in the next data installment in a process of evolving curation. In fact, curation does not stop after data publication. The team continues improving the quality of the data and its documentation, conducting promotion activities to stimulate data reuse, tracking data use, and obtaining feedback until the end of the project's funding.

2. XALT: HPC TRACKING SOFTWARE AND RESULTANT DATASET

HPC resources are maintained by a handful of centers across the country with support from the National Science Foundation (NSF) [34]. Researchers in diverse domains, from Chemistry to Linguistics, and from Aerospace Engineering to Neurobiology log into the supercomputers remotely and at no cost to submit computational jobs. To run their calculations, researchers use community code applications, personal codes, or a combination of the two. Community codes are developed to solve general and specific computing problems, and they have a community of users and developers around them. Such codes are installed and maintained at supercomputer centers based on the needs of the users. HPC resources are shared by many users, and user submitted jobs wait in long or short queues, depending on how busy a supercomputer is at a given time. Because nodes and memory in supercomputers are finite resources, the community code applications installed on the systems must be selected by HPC administrators for continued use and maintenance to ensure optimal performance of the HPC resources. To understand software usage and performance in HPC resources, NSF funds the development of metrics software, one of which is XALT.

The XALT software is used to support administrative and reporting purposes on selected HPC systems. The software is designed to generate data that can enhance understanding of the HPC users' software needs, identify areas for improving their computational work, and increase the efficiency of limited resources [1]. In its current version, XALT gathers information about parallel jobs run on HPC systems through Message Passing Interface programs (MPI) only, which records jobs that are run in multiple nodes. Another way to do parallel programming is using multiple cores in one node, but, XALT version 0.7.1 does not track the non-MPI executables nor does it account for the serial jobs that are run on supercomputers.

XALT is designed to be a lightweight software because it does not add time to the jobs that are run. Therefore, the resultant data has a precise scope. For example, it shows the number of hours an application runs on an HPC system, which could lead to targeted improvements for efficiency. Open XDMoD, a metrics portal, uses such data along with analysis and charting tools to help HPC administrators identify poorly performing jobs and make decisions about efficient allocation of resources [5, 23]. The data is also used to improve systems security by tracking the shared libraries to detect changes that may point to a hacked library [1]. HPC systems tracking also helps with cost analysis based on compute usage time and number of nodes in relation to overall operational costs.

In this curation project, the data is obtained from the open science HPC system Stampede [31], which is deployed at TACC. XALT is installed at the node level, and at the start of an MPI based parallel computational job it captures job-level information and environmental variables. The resulting information includes among other things, the libraries and executables used, the amount of time the job runs, and the number of nodes and cores used for each job [1]. All collected information goes into a local JSON file; then a script gathers the data and uploads it into a MySQL database located on a Virtual Machine (VM) in Rodeo, a cloud computing system that is also maintained at TACC [30]. The database is the core instrument to organize the data gathered from the computational jobs. The resultant data is dense in content, as it records every execution that is run on Stampede, which averages 1,000 per day [31]. It is also a growing dataset, registering computational activities for Stampede from the point of XALT's development by the end of 2013 until December of 2016, with a possible extension. As it is generated, the resultant data is being documented, preserved, and publicly shared on an ongoing basis.

The XALT software is open source, so all the curation tasks that are integrated to the data generation workflow will be possible in any HPC system where the software is installed. As part of the release is a program called createDB.py, which HPC administrators can use to build the schema into a MySQL database, as well as methods that push the data into the database.

3. RELATED WORK

Concepts and practices of data curation involving overarching library principles of data citation and data sharing [26], are by now widely spread in the academic community. So are the benefits of data curation implemented as a seamless workflow from the point of data creation [14]. However, more often than not, data curation happens at particular stages of a research project's lifecycle and involves different actors. In general, day-to-day data management falls in the lap of the researchers [14], and mostly at the end of the research project. Archivists and data librarians control the selection and storage of information, by methods of "reshaping, reinterpreting, and reinventing" as part of the curation process [27]. Reshaping of data extends to curation when considering issues of anonymization and potential reuse [17].

While data reuse is a major goal in curation projects, it is also elusive, as it is often not clear how to achieve it for particular datasets [6, 13]. Specifically, in the case of computational data, Borgman explains that, "Machine-collected data tend to be consistent and structured, and to scale well, but considerable expertise is required to interpret them" [4]. Others indicate that non-standard file formats used for computational data limit interoperability, human readability, and future use [26]. Indeed, curation practices are needed as an integral part of computational data publishing in order to facilitate interpretation and to exploit reuse [11, 19]. Difficulties with niche data extend to finding an adequate IR for preservation. Currently, there are only a handful of open HPC administration datasets available for public consumption. Examples include the parallel workload archives, published as part of work on parallel job scheduling, and the XDMoD portal, a hardware focused high level overview of computing systems [9, 23], none of which constitute a permanent OAIS compliant repository or publish data with consistent metadata [21].

Evaluation of digital curation strategies often involves benchmarks to assess one or more components of the program, such as: the sustainability of the file formats selected for curation, the stability of the repository, and the completeness of the metadata [10, 18]. As suggested by Faniel and Zimmerman, in this project, we derived curation strategies from data reuse requirements and from research scenarios suggested by prospective users [8]. Such requirements became the benchmarks against which we assessed the results and corrected the curation strategy. The degree of data visibility also affects data reuse. Methods of enhancing visibility include: associating published papers with data, enhancing and using different venues for metadata registry [3, 7, 11], and facilitating encounters between data producers and users [36]. Our approach to increase data visibility uses a combination of the latter strategies and tracks interest in the data for the remainder of the project.

4. CURATION STRATEGIES

As a first step the team, including researchers and curators, conducted a curatorial analysis of the overall project. For the curators this meant that we had to understand both the goals and processes involved in the research, as well as the scope and content of the resultant dataset. The analysis included learning about the functions of the data for purposes of managing HPC resources; and, knowing the workflow steps by which the data is obtained, its contents, the size it would attain by the end of the project, and understanding its reuse prospects. In turn, the researchers learned about the process of curation and agreed to incorporate curation tasks onto the data generation software.

As a result of the analysis we identified outstanding curatorial issues. Users would have difficulties transferring and downloading data, which by the end of the project will reach ~64 GB. Also, we needed to determine a data format to facilitate reuse for analysis purposes. Adequate documentation would be needed so the data would be comprehensible to non-HPC experts. Importantly, data had to be anonymized to preserve the privacy of those running jobs on Stampede, whose identification is embedded in the data through their supercomputer accounts and the paths to the executables. Finding a permanent repository for the data was another challenge, as there are no computer science specific IRs available, and most open repositories do not normally take large, growing datasets [12, 33]. Given the narrow niche of the HPC administrative field, we considered the limited scope of the data for reuse. Upon concluding the analysis, we identified the need to complete the following curation tasks: a) document the dataset and enhance its understandability, b) expand its reuse scope, c) select a repository for its preservation and publication, d) decide the format, size and frequency of data distribution, e) integrate the processes of data creation and curation and automate the curation workflow, and f) evaluate the curatorial activities.

4.1 User Input on Data Reuse

Throughout the curatorial analysis recurrent questions were which, why, and how would users be interested in reusing this dataset [22]. We decided to seek guidance from prospective data users, for which seven researchers were recruited to answer questions and to provide feedback through the project's development. The selected group comprised of two HPC experts, two data scientists, and three social scientists. We contacted participants due to their expressed interest in the project and through referrals. While we understand that this is not a big pool of interviewees, we considered the very specialized scope of the dataset and the need to reach out to users with knowledge about the project or related interests. For example, expanding the use of HPC data to Social Sciences is a novel concept and it was not easy to find potential users for which this data would fit in their research. In fact, the social scientists we contacted had previously indicated to the XALT team that the data could be used to understand the human factors that influence computing practices. We did not pursue feedback from more HPC administrators because the data was created for HPC administrative use, so we knew that most of their requirements were being addressed. Interestingly, we later found that by pursuing curatorial changes to address social scientists we also helped HPC administrators.

Obtaining curation guidance involved finding out how users wanted to download and manipulate the data for analysis on different platforms (desktop, cloud, supercomputers, etc.) and using different software. It also included discussions about the research questions they expected to answer using the data. We designed a protocol with a few open-ended questions focusing on requirements for data reuse. The interview protocol is published along with the data in the IR [22]. The questions addressed the interviewees computing resources and tools for data analysis, to discover their needs regarding file sizes and format; data transfer modes and broadband; and, to schedule the frequency of data publication. We also asked the interviewees where they expected to find the data. Lastly, we asked them what and how would they use the data for to accomplish their research goals.

Through consensus of responses, JSON was selected as the format for data distribution, pointing to its readability by humans and machines, widespread use, robust documentation, and compatibility with the reported analytic software choices of the users. In terms of file size, most interviewees said they would use laptop or desktop computers to download and analyze the data, and the amounts they needed varied from a few thousand records, to individual months or over a year's worth of data. Other users were interested in analyzing the data in HPC resources, adding that the amount required would depend on the specific research question they wanted to answer. We thus decided to release compressed packages, each containing three individual monthly files of data per quarter, which would support the ample variation in requirements as well as easing download.

When asked where they expected to find the XALT data, interviewees said that they expected to find it via regular web searches and through links in the XALT project's website. HPC users said it would be easy for them to access the data directly from a storage resource that is accessible from other computational resources. Our access strategy, which we describe in section 4.3, incorporates all of these suggestions to accommodate users requests as well as to increase the visibility of the dataset.

Discussing research scenarios with the interviewees, the social scientists suggested that this data could be integrated with other data sources (e.g. online job postings) to identify changes in the software development industry, and to learn if academic development leads or follows industry. They also wanted to understand users' interactions with large computational systems over time and in relation to major events (e.g. weather emergencies), and study usage patterns to determine research cycles based on scientific domains.

HPC administrators wanted more granular information about how much time is spent in a library call and in the computing job to re-evaluate allocated run times. Such metrics could be compared across the network of supercomputing centers to identify optimized codes. HPC administrators also proposed analyzing the data to help long range planning for hiring staff at supercomputing centers and to create user-training tutorials. A software engineer amongst the interviewees said he would want to learn the frequency of use of the community code he developed. In turn, data scientists proposed using the dataset for teaching big data management, database systems, and data analysis using the Hadoop framework. As the curators received and evaluated the users' requirements, we decided to explore if the XALT data could be used to learn the composition of computational workflows. An additional element of the curation strategy was to expand the data visibility once published. Our data visibility strategies are described in Section 6.

4.2 Shaping Content to Address Reuse Scenarios

Understanding what users needed from and expected to do with the dataset was extremely helpful to curate its content. But, the curators in the team also had to become familiar with how users run jobs on HPC resources, what the XALT software does, and what each information element gathered in the data means. A major task was reviewing the information elements that would be part of the public data; and, include new ones to enable its application to the research problems identified in the data reuse scenarios. Studying the information elements included in the XALT data, we observed that some were redundant, others had obscure names that needed to be renamed for clarity, and others were idiosyncratic to the Stampede resource and would be of limited interest to general users. As we were becoming more familiar with the data and the HPC practices, we started preparing a data dictionary to clarify the meaning of the information elements gathered by XALT.

Stampede user names and paths to the users environments had to be anonymized for privacy reasons. It was decided that each user name would be replaced with a unique user identification number to enable analysis based on users. Applications written by users and considered personal, some of which bear personal names, were also anonymized through the use of hashes.

A new information element, field_of_science, was integrated from the supercomputing center's administrative database. When researchers request time on HPC resources through a supercomputing center's portal, they select a field of science from a list including high level fields of science created by the NSF. We considered that this information would allow learning what fields of sciences use a given resource and what types of jobs are run, and which libraries are used in relation to that field. This information could help answer the research questions posed by social scientists and could also be useful to HPC administrators.

Knowing how users run jobs on computing resources was of importance to make data content decisions. Some users design computational workflows that include executing different community code applications in one job. In other cases, to complete a workflow, users run individual jobs because they need to evaluate results before completing the next step. To understand the composition of workflows, the curators suggested adding job_id information to the data. This element, whose inclusion was not originally planned in the research project, would allow grouping information about applications executed during individual jobs. We also decided to add two other information elements in the public dataset: start_time and run_time, which can be used to calculate a job's end time. Knowing when a job starts and ends would allow identifying if individual jobs are run consecutively by a same user and help infer if those constitute a connected workflow.

Up until the first data installment was published, the team worked on the XALT software side to create clean data: removing duplicate entries, fixing bugs in the job and libraries usage counts, and on files citing null entries.

4.3 Repository Selection and Access Strategy

In selecting an open repository we considered the need for ongoing delivery of data and evaluated which would be the best fit for the data's theme. Searching for Computer Science repositories in the Registry of Research Data Repositories (re3data) and in the Open Access Tracking Project did not yield results [25, 28]. As an alternative, we considered general topics repositories and looked at different DataVerse instances and at the University of Texas DSpace based Texas ScholarWorks. The repositories present varied features: from accommodating up to 1 GB or up to 1 TB of data, they offered different levels of customer service assistance for data upload and metadata entry, and in some cases required the payment of fees [12, 35]. After considering the different options, we saw the advantage of collaborating with Texas ScholarWorks as a publishing outlet [33]. While general repositories serve a large cross-space with little delineation to promote specific topics, it made sense for this data to be amongst other UT academic work.

In coordination with the repository librarian, we implemented an automated workflow from data creation and curation, to long-term storage and publication within a distributed collection architecture that spans multiple resources and policies across two campus organizations. The data is generated, processed, and stored in the HPC resources maintained at TACC, and the Texas ScholarWorks IR is used as a data access point, DOI landing page, and as storage for the data documentation. This solution addresses the limitations in storage space presented by the repository. It also facilitates data movement and analysis across national HPC resources connected through TACC for those users interested in analyzing the data in

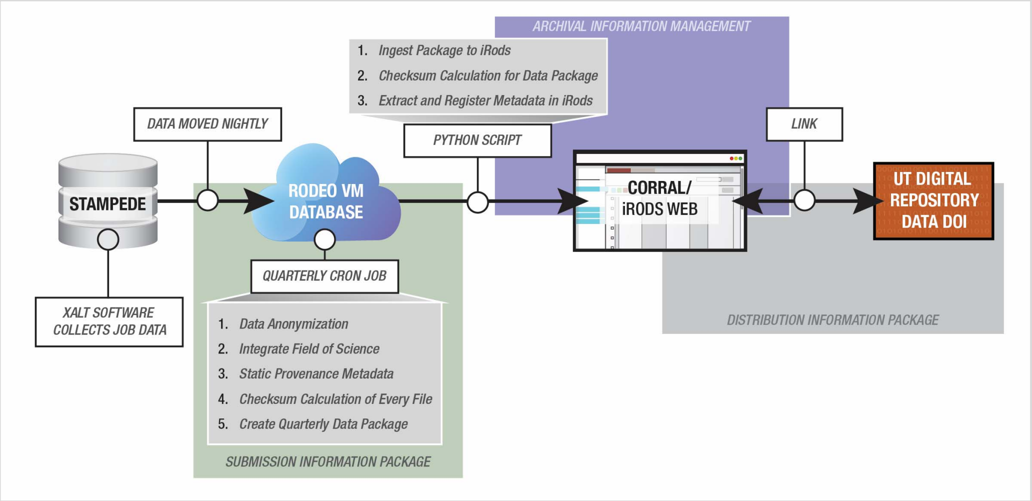
ARCHIVAL INFORMATION MANAGEMENT
1. Ingest Package to iRods
2. Checksum Calculation for Data Package
3. Extract and Register Metadata in iRods
DATA MOVED NIGHTLY
PYTHON SCRIPT
LINK
STAMPEDE
RODEO VM DATABASE
CORRAL/ iRODS WEB
UT DIGITAL REPOSITORY DATA DOI
XALT SOFTWARE COLLECTS JOB DATA
QUARTERLY CRON JOB
1. Data Anonymization
2. Integrate Field of Science
3. Static Provenance Metadata
4. Checksum Calculation of Every File
5. Create Quarterly Data Package
SUBMISSION INFORMATION PACKAGE
DISTRIBUTION INFORMATION PACKAGE

Figure 1. Automated curation workflow from data generation to publication mapped to the OAIS model.

supercomputers. In turn, data segmentation and compression facilitates download to and analysis in the users local computers. In addition, because of the evolving nature of the software development and curatorial work involved in this project, we required flexibility to make workflow adjustments between data installments and to explain those by amending the metadata when needed. The latter was easier to attain working with the local IR.

Importantly, the coordination with Texas ScholarWorks involved a commitment to comply with library standards. In designing the workflow, the stages and processes within the distributed collection architecture were mapped to the OAIS model as well as to best practices for data preservation. This collaboration, which was the first one of this kind between TACC and the UT Libraries, will be generalized to other large datasets that originate on and are stored in supercomputing centers and need to be published to an open repository.

During the interviews, the users mentioned that they expected data access from the software's public repository location and from the research project's page. Taking these suggestions, we cited the data with its DOI in the XALT software GitHub's page and on the XALT project page in TACC's website.[2] Note that the latter sites do not constitute permanent web addresses and may change.

4.4 Automated Curation Workflow within a Distributed Collection Architecture

The XALT data production and curation workflows are integrated and automated across different computational, cloud, storage, and publication platforms. Figure 1 provides an overview of the distributed collection architecture mapped to the OAIS model [15]. Data is gathered by XALT software on the HPC resource Stampede [31] and sent nightly to a database hosted on a virtual machine (VM) located on Rodeo, a cloud computing system also at TACC [30]. Inside the VM, a cron-job that specifies commands to run at the conclusion of every quarter invokes a script to query the database and writes out the three months of data in three individual JSON files. At the same time in the database, each username is converted into a unique string to anonymize the data, and the field of science table for every user is integrated from TACC's computational accounts database. Provenance metadata, such as creator and publisher, are added to the resultant JSON files which are then hashed for authenticity and time stamped. A readme file recording the three hashes is produced and included in the quarterly package. Up to this point, a script in the XALT software completes the generation of a submission information package (SIP) that complies with authenticity and documentation standards. Therefore, any supercomputing center that uses the XALT software will have this method for producing curated data that they can submit to repositories of their choice.

In the workflow implemented on Stampede, the SIP is moved automatically to an iRODS instance on Corral, TACC's High Performance Storage facility [29]. Using the open source data management broker iRODS allows automating data management tasks through actions that can be initiated from any trigger within the distributed architecture [24]. In this case, when files are scheduled for ingest into the iRODS instance on Corral, a checksum is calculated for each package, and its metadata is extracted and registered on the iRODS metadata catalogue. Using one of the various clients supported by iRODS, curators can view, annotate, change and move or delete files on the storage resource if needed. This access is not available to public users, which can only download the package containing all the files. As an archival information package (AIP), data on Corral/iRODS is redundant and geographically replicated. The resource is monitored 24/7 by systems administrators, while the curators and researchers can focus on managing the AIPs. the Corral/iRODS resource is supported by the University of Texas System to provide reliable and permanent data management and storage services across the UT System [29], assuring that the data will be permanently available. The dissemination information package (DIP) consists of the quarterly data file with embedded dates and provenance information, the metadata dictionary, a catalog of the community codes, and the XML DataCite metadata file (See section 4.5). Thus, regardless of the amount of quarterly packages obtained by a user, the citation information will always be included in the DIP. In this way, the published data preserves representation and integrity information.

For dissemination of the DIP, the project uses the iRODS web interface such that files stored under this collection are automatically available on the open web for the general public to access. A link to the data exists on the Texas ScholarWorks repository, which is the landing page for the data DOI. The

[2] GitHub: https://github.com/Fahey-McLay/xalt
TACC: https://www.tacc.utexas.edu/research-development/tacc-projects/xalt

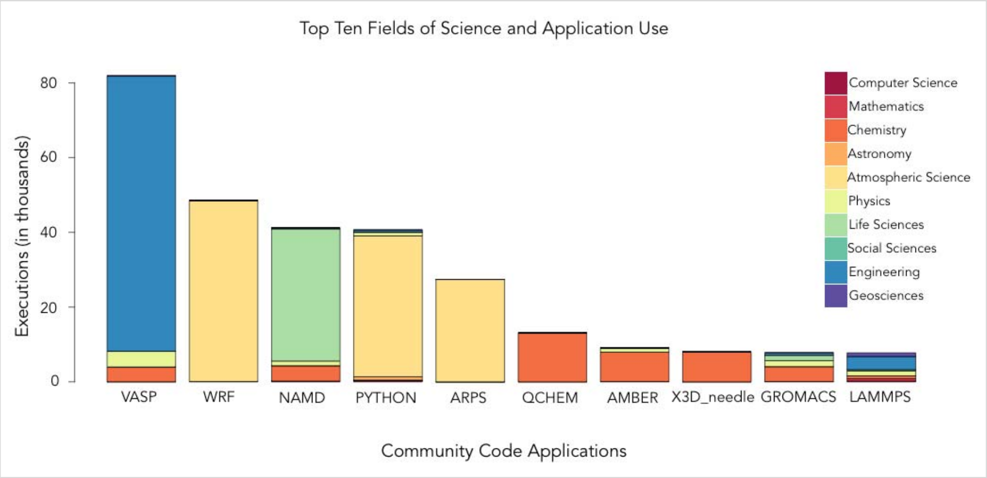

Figure 2. Distribution of users self-selected fields of science over community code applications for one year on Stampede.

landing page contains the project's Dublin Core metadata, another copy of the data dictionary, and a copy of the community code applications catalogue. Due to feedback from the interviewees, most lately we added examples of the proof of concept analysis to illustrate data reuse. Users can search for this dataset within Texas ScholarWorks, access it through its DOI from any available citation, and find it through web searches. Importantly, the distributed collection architecture complies with the functions of a trusted repository.

Creating this workflow involved significant work at the project's outset to identify and implement curation tasks, and improvements are in progress. The goal is to achieve a fully automated workflow to eliminate the need for manual curation. The data first public release was in September of 2015, but the workflow is still undergoing development, and will go into full production before the end of the grant in September 2016.

4.5 Data Understandability and Metadata

Project metadata, including a crosswalk between the DataCite and the Dublin Core schemas was created to publish the data in the Texas ScholarWorks and obtain a DOI [7]. We used the Dublin Core metadata application profile available in the IR to provide an overview of the project, explaining at a high level how the data is obtained and how it can be used [20]. The data dictionary explaining every information element included in the data is key to the dataset's understandability, as the variables recorded can differ in meaning within the field of Computer Science and are largely unknown outside the field [22].

We also created a community applications catalogue of identified community codes used on Stampede. Applications used in HPC systems are often developed by a community of users around scientific problems or domains. As such, they may be sustainable or may have a short duration depending on adoption, and there is no unified catalogue in which they are documented [22]. In the catalogue we point to the community applications projects so data users can learn about the domain that they serve and the types of computational issues they address.

5. CURATION EVALUATION VIA PROOF OF CONCEPT DATA ANALYSES

To evaluate whether our curation decisions were adequate to meet the users' research goals, we designed proof of concept analyses that use statistical and data mining methods. We based the proof of concepts on basic premises outlined in the research scenarios introduced by the prospective users that we interviewed. Our goal was not to pursue their research questions, but to find out if the information elements selected for publication in the dataset are useful to the kinds of research that the users expressed interest in conducting. We also wanted to find out if the accompanying documentation was adequate to perform analysis on the data and interpret the results. The questions selected for testing are at the tip of more complex research problems, but have the elements to determine if the content of the dataset is adequately curated and identify areas for improvement. To evaluate the analysis results, we obtained feedback from the users that introduced the research scenarios.

Almost a year's worth of data containing 6,987,901 records for the period 2014/07 to 2015/06 was used. Each record in the data file corresponds to one execution of a particular application, so there can be multiple records with the same job_id. All executables that do not correspond to a community code application were filtered out. The final dataset used for the evaluation contains 130 unique executables, 1,868 unique users, 569,811 unique job ids, and 98 unique fields of science. For purposes of showing clear results on this paper, we consolidated the 98 fields of science into ten major fields, and considered only the top ten community codes in the analyses. The proof of concepts were conducted by one of the data scientists interviewed. To conduct the tests, he used Wrangler, the data intensive computing resource at TACC and made use of the data dictionary [32].

The first proof of concept corresponds to the inquiries presented by social scientists and was also useful to HPC administrators. It aimed to answer the following questions: What fields of science are using HPC? What applications do they use? How are those applications shared across different fields of science? The questions relate HPC usage patterns to scientific domains, and allowed evaluation of the integration of the NSF fields of sciences information.

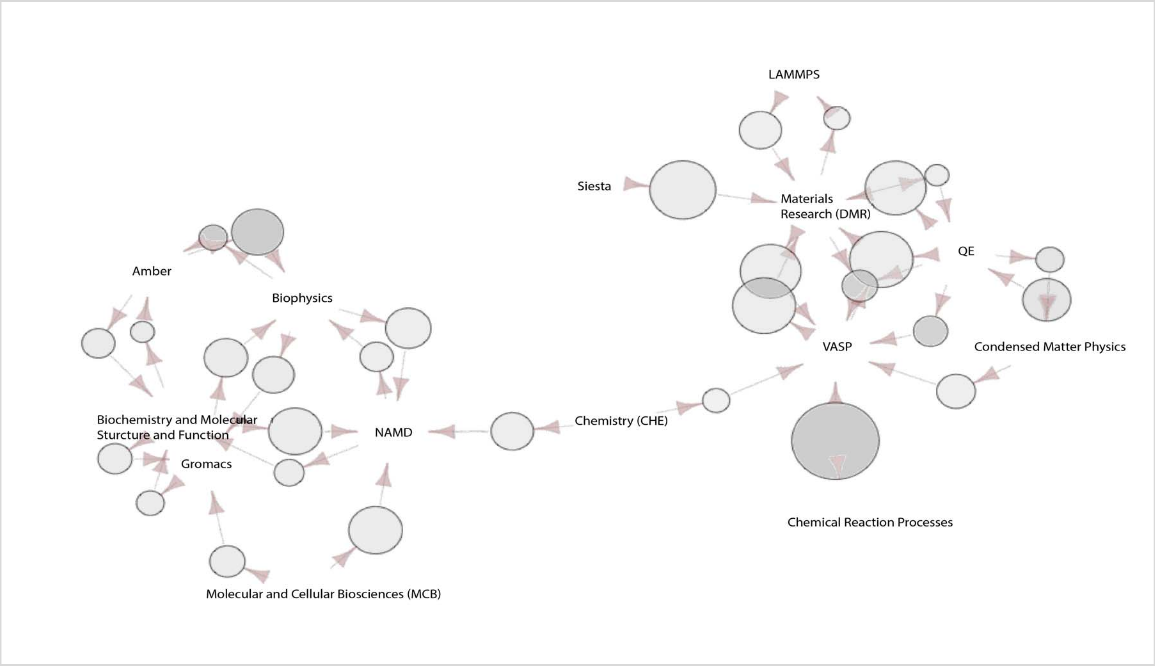

Figure 3. Visualization of association rule mining results showing inferences between fields of science and community code applications.

Figure 2 shows the statistical distribution of self-selected fields of science over the top ten community applications used on Stampede, expressed as number of executions in one year. The HPC experts that provided feedback observed that for the most part the results confirmed their intuition about the users, such as the field of Engineering dominates the usage of Stampede. At the same time, the possibility to mine bigger numbers of users and software exposed more granular information that was new to the administrators. They could now observe that applications that they considered domain specific such as VASP (Vienna Ab initio Simulation Package for Molecular Dynamics in Chemistry), Gromacs (for Biomolecular Dynamics), and LAMMPS (Large-scale Atomic/Molecular Massively Parallel Simulator for soft and hard materials Molecular Dynamics), are actually being shared across various scientific disciplines.

Results interpretation improved our understanding of the possibilities and limitations of the data. Concerning the heavy usage of Python, one of our interviewees said, “I don't know why Atmospheric Science is the top Python user. It's not wrong, necessarily, it just doesn't appear to come from any intuition I have about what gets run by whom on Stampede. We'd have to look into those cases to see if there's a common analysis package used by this community that is written in Python.” Another user said, “The large volume of Python jobs might be explained by its use in connection to other executables, or as a tool for pre or post processing if done in parallel. Since Python is just the shell to run other scripts, there is not enough information to indicate what exactly it is used for.” These comments helped us understand that the data cannot answer the question of how Python is used in the context of parallel processing using MPI.

Association analysis was used over the same data to understand the relations between scientific fields through the use of community code applications [2]. Association analysis can identify inference rules that can predict the occurrence of one set of variables based on the co-occurrence of another set of variables. The method consists of two steps. The first step identifies objects that commonly occur together as a frequent item set. The second step identifies inference rules among objects based on their presence in the frequent item set. Figure 3 shows the most relevant patterns found in the data, the fields of science and applications are connected via directional links. One field might be linked to multiple applications and vice versa. The direction of the link indicates the inference from one to the other, and in most cases there are links in both directions between pairs. However, the strength of the inferences can be different for different pairs of objects and directions of inference. The size of the circle in the middle of each link indicates the strength of the inference, the bigger the circle, the stronger the inference. The most revealing pattern shows that users from the field of Chemistry use applications that are also commonly used by Biology and Physics users. The connection is not causal and is established based on executable usage by fields of science. It shows direct connections between field of science and executable, and potential indirect connections among executables. The results indicate that Chemistry users use both VASP (for Quantum Mechanics, Chemistry), and NAMD (Nanoscale Molecular Dynamics for Biophysics), with certain level of significant frequency, and that VASP and NAMD are also significantly used by users from related fields of science. Note that Chemistry users using VASP may or may not be the same users that use NAMD. Such connections are at the tip of understanding the interdisciplinary nature of the computational research conducted on Stampede.

Feedback obtained from one of the evaluators about this scenario referred to the field of science information element. The field of science list used in HPC centers was developed by the NSF and uses a general classification scheme. When users request an allocation on HPC resources they are prompted to select a high level term that relates to their research project. In looking at the results from Figures 2 and 3, the evaluator noted that the list may not always reflect with granularity the domain involved in a research project, and that in cases of multidisciplinary research, users cannot select more than one term. Given that we cannot easily and promptly make changes to the list, we made changes in the data documentation to clarify the field_of_science provenance and scope to our users (See section 5.1).

The second proof of concept involved learning about computational job patterns, useful to understand data provenance and for HPC administration purposes. While XALT software does not have a way of capturing the relation between computational jobs to changes in the data, the assumption for tracking relationships between subsequent computing operations or unique jobs seemed a first step towards a provenance framework in HPC. First we used statistical analysis to identify job patterns. A job is the series of calculations performed on a user's data. A user can submit one job with multiple executables or can submit multiple jobs, each containing a single executable, and each job submitted by a user is identified with a job_id. The results revealed that only 7.2% (41,083 out of 569,811) of jobs use two executables, and there are no known jobs with more than two executables. This shows that the majority of users prefer one execution a time. Jobs using two executables are dominated by the use of ARPS and WRF, both weather forecasting applications, in the same job submission (40994 out of 41083, 99.8%). Another interesting observation is that more than half of the remaining jobs with two executables (54 out of 89) use Python in combination with another executable.

We then used sequential rule pattern mining to identify patterns in the order of the executions for all the jobs run by each user. The sequence of executable paths used by each user was modeled as a list ordered by the starting time of each execution. The results showed that in most cases users run the same executable repeatedly. There are only a few outliers in which different executables are used in a particular order repeatedly. This may suggest a workflow where users use different executables sequentially to achieve a desired analytic result. One example of such pattern is shown as following:

[[VASP*], [QE*], [VASP*]]

The example shows that the executables VASP and QE – Quantum Espresso, are used in alternative sequences by some users. Both executables are popular tools for molecular structure modeling in material sciences, and share similar features and functionalities. The sequential pattern may indicate that users are trying to compare the performance of the tools, or they are attempting to use complementary features from both executables. The results provide interesting observations about how parallel jobs are configured, but more investigation needs to be conducted to corroborate the use of more complex workflows on Stampede. For example, users may wrap a workflow in a personal script and we don't have a way to know the contents of that script, as XALT does not capture this information.

We did not perform HPC specific analysis because this is routinely done by Stampede administrators and we rely on their reports to evaluate the reuse value of the XALT data. At TACC, the data is used to report the top applications and libraries used per month and per year. Also, the data is integrated into TACC-Stats (the center's statistical system) to help understand performance. TACC-Stats measures performance on each node at the start, the end and 10min intervals during a job run and the XALT database is mined to learn what executable is running in the job [23]. Using this data, Stampede administrators discovered which programs were causing delays on the resource's large memory queue and moved those jobs to regular nodes, significantly increasing the efficiency of the tasks [16]. After seeing the proof of concept results, the HPC administrators requested to conduct association analysis to observe connections between libraries, executables and fields of science.

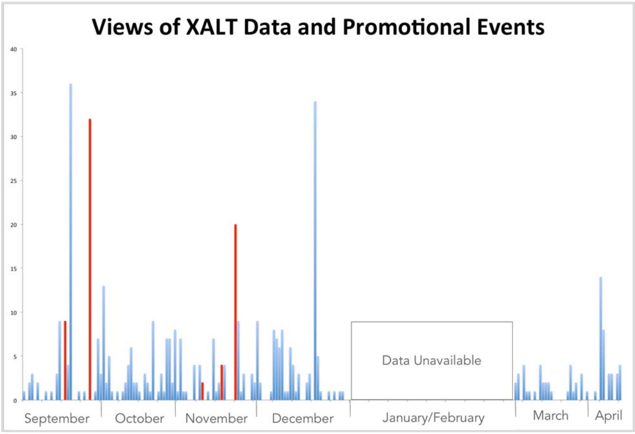

Figure 4. Dataset downloads from Corral/iRODS web-page.

5.1 Lessons Learned

As a result of the analyses we concluded that the curation strategy did expand the scope of the data, but that further explanations in the documentation were in order. To clarify the usage of the field of science information element, we added a description of its provenance in the Texas ScholarsWork landing page metadata and in the data dictionary. Suggestions by evaluators to provide images and descriptions of the proof of concept analyses as part of the data, suggested to us the complexities involved in breaking new uses of a dataset to the public. We incorporated this feedback by including a description and images of the proof of concept analysis in the landing page, obtained a DOI for the software, and furthered examples of data reuse by linking to and citing publications that refer to XALT.[3] In the next data release and in XALT announcements we will highlight the changes made in the documentation to elicit reuse for this niche dataset.

Visualizing the results of the proof of concepts provided a clearer understanding of how much we still need to know about HPC usage that the current version of the XALT data does not capture. We are in the process of writing a proposal to expand XALT to record other non-MPI parallel jobs as well as serial jobs. These changes will render data with more complete information about work and users on HPC resources.

6. TRACKING DATA VISIBILITY

To improve data visibility, we engaged in targeted promotion activities. An initial general promotion featured on the front page of the Texas ScholarWorks occurred with the publication of the data at the beginning of September of 2015. At the end of that month, we sent an email announcing the data public release to those who had expressed interest in the XALT software presentations at HPC events and had subscribed to its mailing list. Social scientists and data analysts were also included in these announcements based on referrals from colleagues that we interviewed. At the same time, we updated the different XALT project pages and included the data citation information.

To evaluate the results of this outreach strategy, we looked at the number of views of the landing page and the number of downloads of the data dictionary from the Texas ScholarWorks reports. Of course, we cannot track data downloads through IR

[3] GitHub: https://dx.doi.org/10.5281/zenodo.49772, TACC:https://www.tacc.utexas.edu/research-development/tacc-projects/xalt

metrics because the data is not hosted on the IR. After the beginning of the promotion campaign and for the first twelve days of October, we observed an increase of 12 downloads of the metadata and three data dictionary downloads. In parallel we started tracking data downloads using the aggregated server log information from the Corral/iRODS public web access page. Figure 4 shows the number of downloads from September through December of 2015, where the blue bars are day-to-day downloads and the red bars indicate downloads during a promotional activity.[4] The red bars in September belong to the first data publication and to email outreach activities. The red bars in November correspond to a presentation about the data at the 2015 Supercomputing conference and to Twitter announcements promoting the event and the release of the data. Throughout the promotional activities, the graph consistently shows increase in the number of downloads of XALT data the day of and after an event. The last significant increase in December does not correspond to a known event but to a single user downloading all the data packages. Instead, during the periods which we did not conduct promotional activities, we observed a decrease in data downloads. Overall, the results are consistent with those from the Texas ScholarWorks landing page in relation to data views and documentation downloads. The decrease in views during phases in which there are no promotion activities further demonstrates the value of the conducting outreach for niche datasets.

7. CONCLUSIONS

While data is becoming ubiquitous and IRs are emerging at a rapid pace, it is still not easy for users to find and to identify a dataset's uniqueness and reuse potential. This is exacerbated in the case of computational datasets with a narrow community of users. We curated a growing HPC dataset according to best practices and focused on the needs and suggestions of potential users. The latter led to targeted curation decisions in relation to content, format, and delivery choices to expand the data scope and facilitate its reuse. To overcome the limitations of manual curation we implemented the curation tasks as an automated workflow that spans multiple technical platforms from data generation to publication. Because we integrated the curation functions in the software that generates the data, this strategy can be replicated in other HPC resources. This workflow model can be used to curate large and growing datasets in collaboration between supercomputing centers and IRs or other publication outlets. However, when using a distributed collection architecture, making sure that the system complies with long-term preservation and access standards is key.

Curating data focusing on reuse entails having data creators, curators, and potential users go through the exercise of imagining what data reuse means for a particular dataset from the research project's outset. This strategy derived more focused curation activities and methods for evaluating impact. Data analysis proof of concepts were conducted to understand if the curated data could answer basic questions related to the potential users' research interests. The evaluation showed both the potential and the limitations of the curated dataset. While the data may not be useful to answer all the questions conceived by researchers, the data analysis results provided a flavor of how the data can be used to explore trends in computational research and interdisciplinary collaboration. But, most importantly for assessing the curation strategy, it indicated how to improve the documentation, tested the precision of the information elements, and directed future improvements to the XALT data. Finally, tracking data views and downloads in conjunction with promotional activities suggests that those are important to increase interest in the data.

The project revealed complex interrelations between understanding a dataset, designing questions and analysis to reuse it, and interpreting its results. We learned that the proof of concept analyses including the response of the users to the results, enhanced our understanding about the possibilities and limitations of the data and how to improve its curation. We also found that changes done to cater the data to a broader public were useful to its regular users. While expanding the data use scope entails an involved process, curatorial time is saved by automating tasks that enable ongoing curation and publication as data is being generated. All of this highlights the importance of implementing curation strategies that can be tested in collaboration with users and throughout the research process.

8. ACKNOWLEDGMENTS

This research was possible through the National Science Foundation Collaborative Research Grant: SI2-SSE: Understanding the Software Needs of High End Computer Users, # 1339708.

9. REFERENCES

[1] Agrawal, K., Fahey, M., McLay, R., and James, D. 2014. User environment tracking and problem detection with XALT. *Proceedings of the First International Workshop on HPC User Support Tools* (Nov. 2014), 32-40.DOI=http://dx.doi.org/10.1109/HUST.2014.6

[2] Agrawal. R. and Srikant, R. 1994. Fast Algorithms for Mining Association Rules in Large Databases. In Proceedings of the 20th International Conference on Very Large Data Bases (VLDB '94), Jorge B. Bocca, Matthias Jarke, and Carlo Zaniolo (Eds.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 487-499.

[3] Arlitsch, K.and O'Brien, P.S. 2012. Invisible institutional repositories: Addressing the low indexing ratios of IRs in Google Scholar. *Library Hi Tech* 30,1 (Mar. 2012), 60-81.

[4] Borgman, C. L. 2012. The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*. 63, 6 (Jun. 2012), 1059–1078. DOI=http://dx.doi.org/10.1002/asi.22634

[5] Browne, J., DeLeon, R., Patra, A., Barth, W., Hammond, J., Jones, M., and Wang, F. 2014. Comprehensive, open-source resource usage measurement and analysis for HPC systems. *Concurrency and Computation: Practice and Experience* 26, 13 (Sep 2014), 2191-2209. DOI=10.1002/cpe.3245

[6] CODATA-ICSTI Task Group on Data Citation Standards and Practices 2013. Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. *Data Science Journal* 12,0, (Sep. 2013), CIDCR1-CIDCR75. DOI=http://dx.doi.org/10.2481/dsj.OSOM13-043

[7] DataCite 2015. DataCite Metadata Search. http://search.datacite.org/ui

[8] Faniel, I. and Zimmerman, A. 2011. Beyond the Data Deluge: A Research Agenda for Large-Scale Data Sharing and Reuse. International Journal of Digital Curation, 6, 1 (2011) 58-69. DOI = http://dx.doi.org/10.2218/ijdc.v6j1.172

[4] The gap in the graph for the month of January 2016, reflects the logs purges, a common practice in systems administration which we failed to prevent.

[9] Feitelson, D., Tsafrir, D., and Krakov, D. 2014. Experience with using the Parallel Workloads Archive *Journal of Parallel and Distributed Computing*, 74, 10, (Oct. 2013), 2967-2982. DOI=http://dx.doi.org/10.1016/j.jpdc.2014.06.013

[10] Giaretta, D. 2007. The CASPAR Approach to Digital Preservation. *The International Journal of Digital Curation* 3,2 (July 2007), 112-121. DOI=http://dx.doi.org/10.2218/ijdc.v2i1.18

[11] Goodman, A., Pepe, A., Blocker, A. W., Borgman, C. L., Cranmer, K., Crosas, M., … Slavkovic, A. 2014. Ten Simple Rules for the Care and Feeding of Scientific Data. *PLoS Comput Biol. 10*, 4, (Apr. 2014), e1003542. DOI=http://doi.org/10.1371/journal.pcbi.1003542

[12] Harvard Dataverse Project 2015. *Dataset + File Management.* http://guides.dataverse.org/en/4.2.2/user/dataset-management

[13] Hey, T. and Trefethen, A. 2003. *Grid Computing - Making the Global Infrastructure a Reality*. West Sussex: Wiley.

[14] Higgins, S. 2008. The DCC Curation Lifecycle Model. The International Journal of Digital Curation 3,1 (July 2008), 134-140. DOI=http://dx.doi,org/10.2218/ijdc.v3i1.48

[15] International Organizational for Standardization. (2012). ISO 14721: 2012: Space data and information transfer systems – Open archival information systems (OAIS) – Reference model. Genève, Switzerland: International Organization for Standardization.

[16] James,D., McLay, R., Si Liu, R. Evans, T. Barth, W., Lamas-Linares, A., Budiardja, R., and Fahey, M. 2015. Tales from the trenches: can user support tools make a difference?. In Proceedings of the Second International Workshop on HPC User Support Tools(HUST '15). ACM, New York, NY, USA, , Article 2 , 11 pages. DOI=http://dx.doi.org/10.1145/2834996.2834998

[17] Jurczyk, P. and Xiong, L. 2009. Distributed anonymization: Achieving privacy for both data subjects and data providers. In *Data and Applications Security XXIII (Jan. 2009),* 191-207. Springer Berlin Heidelberg.

[18] Kulasekaran, S., Trelogan, J., Esteva, M., and Johnson, M. 2014. Metadata Integration for an Archaeology Collection Architecture. *International Conference On Dublin Core And Metadata Applications (*Oct. 2013*)*. Austin, TX, USA, 53-63. http://dcpapers.dublincore.org/pubs/article/view/3702

[19] Lubell, J., Rachuri, S., Mani, M., and Subrahmanian, E. 2008. Sustaining Engineering Informatics: Towards Methods and Metrics for Digital Curation. *The International Journal of Digital Curation.* 3,2 (Nov. 2008), 59-73. DOI=http://dx.doi.org/ijdc.v3i2.58

[20] Lyon, Colleen; Cofield, Melanie; Borrego, Gilbert; (2015): Reducing Metadata Errors in an IR with Distributed Submission Privileges; University of Texas at Austin. http://dx.doi.org/10.15781/T2KW2

[21] Management Council of the Consultative Committee for Space Data Systems 2012. *Reference Model for an Open Archival Information System (OAIS)* (June 2012). Washington, DC.

[22] McLay, R. and Fahey, M. R. 2015. *Understanding the Software Needs of High End Computer Users with XALT.* Texas Advanced Computing Center. Dataset. DOI=http://dx.doi.org/10.15781/T2PP4P

[23] Palmer, J., Gallo, S., Furlani, T., Jones, M., DeLeon, R., White, J., Simakov, N., Patra, A., Sperhac, J., Yearke, T., Rathsam, R., Innus, M., Cornelius, C., Browne, J., Barth, W., and Evans, R. 2015. Open XDMoD: A Tool for the Comprehensive Management of High-Performance Computing Resources. *Computing in Science and Engineering,* 17, 4, (July 2015)52-62. DOI=http://dx.doi.org/10.1109/MCSE.2015.68

[24] Rajasekar, A., Moore, R., Hou, C. Y., Lee, C. A., Marciano, R., de Torcy, A., Wan, M., Schroeder, W, Sheau-Yen, C, Gilbert, L., Tooby, P. and Zhu, B. 2010. iRODS Primer: integrated rule-oriented data system. *Synthesis Lectures on Information Concepts, Retrieval, and Services*, *2*,1 1-143. doi:10.2200/S00233ED1V01Y200912ICR012

[25] Registry of Research Data Repositories 2015. *About re3data.* http://www.re3data.org

[26] Starr, J., Castro, E., Crosas, M., Dumontier, M., Downs, R. R., Duerr, R., … Clark, T. 2015. Achieving human and machine accessibility of cited data in scholarly publications. *PeerJ. Computer Science* 1, 1 (May 2015). DOI=http://doi.org/10.7717/peerj-cs.1

[27] Schwartz, J and Cook, T. 2002. Archives, Records, and Power: The Making of Modern Memory. *Archival Science*, 2, 1-2 (Mar 2002), 1-19.

[28] Tag Team 2015. *Open Access Tracking Project.* http://tagteam.harvard.edu/hubs/oatp/items

[29] Texas Advanced Computing Center 2015a. *Corral User Guide.* http://tacc.utexas.edu/user-guides/corral

[30] Texas Advanced Computing Center 2015b. *Rodeo: General Cloud Computing and Storage.* http://tacc.utexas.edu/systems/rodeo

[31] Texas Advanced Computing Center 2015c. *Stampede User Guide*. https://portal.tacc.utexas.edu/user-guides/stampede

[32] Texas Advanced Computing Center 2015d. *Wrangler User Guide*. https://portal.tacc.utexas.edu/user-guides/wrangler

[33] Texas Scholar Works 2015. Frequently Asked Questions. http://repositories.lib.utexas.edu/pages/faq#getting_started

[34] Towns, J., Cockerill, T., Dahan, M. Foster, I., Gaither, K., Grimshaw, A., Hazlewood, V.,Lathrop, S., Lifka, D., Peterson, G.D., Roskies, R., Scott, J.R., Wilkins-Diehr, N. 2014. XSEDE: Accelerating Scientific Discovery. *Computing in Science and Engineering,* 16, 5 (Sep 2014*)*, 62-74. DOI=http://dx.doi.org10.1109/MCSE.2014.80

[35] UTDR 2015. *The University of Texas Digital Repository: About*. http://repositories.lib.utexas.edu/

[36] White, E., Baldride, E., Brym, Z., Locey, K., McGlinn, D., and Supp, S. 2013. Nine Simple Ways to Make it Easier to (Re)use Your Data. *Ideas in Ecology and Evolution,* 6, 2 (Aug. 2013) 1-10. DOI= http://dx.doi.org/10.4033/iee.2013.6b.6.f