BNL-114352-2016

Data Storage and Sharing for the Long Tail of Science

B. Zhang, L. Pouchard

Published in IEEE

November 2016

Computational Science Initiative

Brookhaven National Laboratory

**U.S. Department of Energy
USDOE Office of Under Secretary for Science (S-4)**

Notice: This manuscript has been co-authored by employees of Brookhaven Science Associates, LLC under Contract No. DE-SC0012704 with the U.S. Department of Energy. The publisher by accepting the manuscript for publication acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or any third party's use or the results of such use of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof or its contractors or subcontractors. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

Data Storage and Sharing for the Long Tail of Science

Boyu Zhang*, Line C. Pouchard[†], Preston M. Smith*, Amandine Gasc[‡], Bryan C. Pijanowski[‡]

*Information Technology at Purdue, Purdue University

{zhan2416, psmith}@purdue.edu

[†]Libraries, Purdue University

{pouchard}@purdue.edu

[‡] Forestry and Natural Resources Department, Purdue University

{gamandin, bpijanow}@purdue.edu

Abstract—Research data infrastructure such as storage must now accommodate new requirements resulting from trends in research data management that require researchers to store their data for the long term and make it available to other researchers. We propose Data Depot, a system and service that provides capabilities for shared space within a group, shared applications, flexible access patterns and ease of transfer at Purdue University. We evaluate Depot as a solution for storing and sharing multi-terabytes of data produced in the long tail of science with a use case in soundscape ecology studies from the Human-Environment Modeling and Analysis Laboratory. We observe that with the capabilities enabled by Data Depot, researchers can easily deploy fine-grained data access control, manage data transfer and sharing, as well as integrate their workflows into a High Performance Computing environment.

Index Terms—data storage, data sharing, long tail of science, Soundscapes, Data Depot

I. INTRODUCTION

Research data infrastructure is continuously evolving to satisfy the requirements of ever increasing data sizes and high performance and distributed access. In addition, storage must now accommodate new requirements resulting from trends in research data management that require researchers to make their data available for the long term and share them with other researchers and the public.

The phrase “long tail of science” comes from a concept popularized to account for Internet commerce, the economics of the long tail, that refers to the large number of small items that taken together contribute substantial economic value [1]. This has been adapted to scientific research where individual scientists lead labs centered around small projects while producing a large portion of the total scientific output. The data produced by these teams tend to be for local use, are less structured, less documented, and less easily shared than those curated in large-scale data centers.

In this definition, the long tail is contrasted with “big science” produced in the context of long-term projects by large teams often working in international collaborations, with big instruments, producing large amounts of data, collected and analyzed in well-understood workflows in agency-supported data centers [2]. Often cited examples of big science include the Large Hadron Collider and the Large Synoptic Survey

Telescope.

With regard to data, issues related to the “long tail of science” include data repositories, open access, data dissemination and sharing, preservation, and sustainability. We understand institutional repositories as comprised of technical infrastructure such as storage and servers and of services such as data policies, transfer mechanisms, support for data deposit and preservation [3]. These issues emerge from stakeholder and funding practices. Researchers in the long tail are often left to their own devices to manage their data, provide their own storage, and perform curation. They do not have many resources to make data accessible to other scientists, nor do they have the time and required resources to prepare and preserve their data for the long term. Long tail data can be found in institutional repositories, as opposed to disciplinary ones, and institutions grapple with internal funding models for their repository [4]. Libraries are playing a leading role in this context, as the organization that traditionally takes on preservation efforts.

Another aspect characterizing long tail data is its heterogeneous nature, expressed in data formats, data sources, data use policies, data selection for dissemination, and data types—different kinds of data resulting from multiple projects in different sub-disciplines that may be collected in the same repository. These attributes are often grouped together under the “Variety” aspect of Big Data [5]. Although long tail data is sometimes equated with small data [1], [2], this is an increasingly invalid assumption, as individual researchers, for instance in universities, are now accumulating and analyzing massive amounts of data. These researchers are thus faced with issues related to data volume and sharing while having scant resources to address them. Their needs have not historically included high performance computing although this is changing as well, with a growing number of research domains such as genomics and financial analysis using computational science techniques. The advent of HPC resources such as Comet provide durable storage in addition to performance storage [6].

With the emergence of new requirements for a data management plan or data sharing plan to be included in requests for funding from publicly supported agencies, researchers

and research institutions have found themselves under the obligation to make their research data publicly available. The detailed elements of a plan depend on the agency to whom researchers apply and specific divisions have additional requirements. For instance, the National Institutes of Health emphasizes increased access to scientific publications and data while safeguarding the privacy of participants and protecting confidential and proprietary data¹. The Department of Energy Public Access Plan² stresses that effective data management has the potential to increase the pace of scientific discovery, facilitate the validation of results, and broadening the value of research data to disciplines other than the originating one, requiring that research data be made available in machine-readable form. The National Science Foundation, one of the first agencies to develop guidelines for data management and sharing and request a data management plan in its solicitations, highlights preserving access³ to data to promote scientific progress and address societal needs through research and education. NSF awardees are expected to share “primary data, samples and physical collections” gathered in the course of doing research.

Scientific research has seen an explosion of data in numerous disciplines as researchers use larger and more complex data to address “grand challenge”⁴ questions in a way that is unprecedented. Examples of such challenges include the impact of climate change on regional agriculture and food supplies, the need for reliable and sustainable sources of energy, and the development of innovative methods for the treatment and prevention of infectious diseases. Multi-disciplinary teams address these challenges and countless others by collecting, generating, cross-referencing, analyzing, and exchanging datasets in order to produce and disseminate results.

The combination of these trends results in increasingly urgent needs for storing research data, collaborating while research is in progress, and for sharing data once results have been obtained. Furthermore, the inherently heterogeneous nature of research data requires technologies that are simpler and more flexible than those developed for exascale data [7]. Research data repositories, including institutional repositories deployed at individual institutions have been built to respond to funding agency requirements. Disciplinary repositories focused on branches of science or specific disciplines also make research data available. But institutional or disciplinary repositories do not always fulfill researchers’ needs with regard to sharing research data results. In some disciplines they do not exist or are not well-known by the community they serve. Institutional repositories may not yet be available as institutions are ramping up to comply with the agencies’ mandates” [8].

Repositories that fulfill the dual role of facilitating research

collaborations and sharing research results are needed to satisfy the requirements and constraints of long-term storage and data sharing for the long tail of science. In this paper, we describe an experiment at Purdue University where we have deployed the Research Data Depot⁵ (“Depot” or “Data Depot”), a storage and service platform that came online in the Fall 2014. Depot is supported by a collaboration between the Purdue Research Computing division of the Information Technology Office (RCAC) and the University Libraries. We describe the motivation and requirements for Depot, the services comprising Depot, and adoption results in Section II. We propose Soundscape Analysis as a use case to illustrate results in Section III. In Soundscape Analysis, Data Depot is used as active storage. We discuss the benefits and implications of Depot for access and sharing in Section IV. In the last section we outline how data from Depot can be migrated to the Purdue Research Repository (PURR)⁶, an existing institutional repository to provide researchers with a large-scale data storage and preservation solution.

II. DATA DEPOT

A. Motivation: Data Depot Requirements

Since 2008, Purdue has operated an award winning campus high-performance computing program known as “The Community Cluster Program” [9]. In partnership with over 185 faculty investors, this program centrally operates Top 500 high-performance computing systems for the campus. Despite a powerful campus computing environment, data services were lagging behind the explosion of data requirements brought forth by Purdue researchers. Fortress, the tape archive system saw a five-fold increase in use from 2008 until January 2013, and ten-fold by March of 2014. As shown in Figure 1, the archive took 16 years to reach its first petabyte in use, 14 months for the second, 9 months for the third, and has now exceeded a fourth.

Fig. 1: Growth of the Fortress archive system.

Prior to the deployment of Depot, several attempts for providing storage to faculty were made. In 2008, RCAC

¹http://grants.nih.gov/grants/policy/data_sharing/

²<http://www.energy.gov/datamanagement/doe-policy-digital-research-data-management>

³<http://www.nsf.gov/pubs/2015/nsf15052/nsf15052.pdf>

⁴<http://www.whitehouse.gov/administration/eop/ostp/grand-challenges>

⁵<https://www.rcac.purdue.edu/storage/depot/>

⁶<http://purrr.purdue.edu>

provided a service dubbed “Resell-a-tray”, where researchers could purchase a disk tray in a shared storage system. This solution was expensive (\$18,000 for 9 TB of storage), and suffered from a critical expectation mismatch: IT believed that the service was selling additional scratch capacity, but later experience revealed that faculty expected that they were buying a persistent network-attached storage drive.

By 2013, Purdue’s community cluster partners had two options for data storage: a scratch filesystem optimized for high-performance, but intended for data to be short-lived, and a tape-based archive system intended for permanent, un-curated archival storage of research data. These two storage tiers left a gap in between, where actively used and persistent data, software applications for a research group could be stored, and intra-research group file sharing could be performed.

Without a storage tier between scratch and the tape archive, researchers encountered many issues:

- They required more space than allocated in their cluster scratch space.
- They were responsible for installing applications for the research group to share, but lacked a place to store them.
- They struggled to work around regular scratch purging.
- They needed easy use of UNIX groups to share access to files and subsets of files.
- Scratch filesystems are not global or shared on every computing system.
- IT did not provide a storage system for purchase with grant funds.
- No storage system was suitable for use by scientific instruments.
- Without a “lab file space”, student research files frequently were held in private spaces owned by students, and would leave with them.
- Easily sharing files with collaborators, within and outside of the institution was not supported.
- Relying on private or departmental filesystems created data management challenges when moving toward HPC.

Leveraging the university’s enterprise-wide scale-out NAS system, research computing staff began to provide a group storage service from that hardware platform, with a capability to purchase shared group storage by the terabyte, and access it on all Purdue HPC resources. This service was an unexpected success, with 65 research groups buying space in the span of one year. Most remarkably—several new groups purchased access to the solution despite not being HPC users.

As the solution grew, scalability issues were encountered with the campus-wide NAS. Scalability and I/O latency were issues on clusters, to the extent that research computing staff needed to monitor for “appropriate” work. Using a campus-wide enterprise system meant that an I/O intensive job in the wrong place could potentially impact the entire campus. Unlike access to Purdue’s cluster systems, creation and management of group spaces was not automated. And finally, data on the campus NAS was protected from accidental deletion, but not against disaster to the extent expected by Purdue faculty. An HPC-scale storage system was required

for research at Purdue, and faculty were surveyed to gather requirements.

B. Aims and Objectives

The faculty survey revealed several requirements for the new storage system: protection from accidents (snapshots), protection from disaster (replicas), and affordable prices, of less than \$200 per TB/year.

- Experience informed the Research Computing team that the system must also provide:
- A high-performance filesystem resource in which research instruments may save data.
- Easy ways to share data with collaborators.
- A resource that works for both HPC and non-HPC user.

Proposals were asked of many storage vendors, requesting a solution delivering at least one petabyte of capacity that could scale to 10 petabytes; perform at 40 gigabytes per second, 100,000 I/O operations per second; deliver 300 megabytes per second to a single client; and have low latency accessing filesystem metadata. Additionally, the system must support snapshots, replication to multiple data centers, and be fully POSIX compliant, including support for parallel I/O.

C. Methods: Data Depot deployment

Purdue selected a solution from Data Direct Networks (DDN), using IBM’s General Purpose Filesystem (GPFS). The winning proposal provided nearly 5 PB of raw disk, with 2.5 PB usable capacity, with matching sets of filesystems in two separate campus datacenters. Depot is connected to the Purdue Science DMZ at 160 GB/second.

Within Depot, each lab is automatically set up with a default directory structure to enable commonly used access patterns. POSIX ACLs are used to force the default group and file modes for each file to be open to lab members, and, to the extent possible, spare researchers from having to manually set UNIX file permissions or deal with *umask* settings.

By default, each lab is set up with directories for “*apps*”, and “*data*”, along with a corresponding set of UNIX groups to manage permissions to them. HPC faculty or a designee frequently curate a centrally-managed set of application software in “*apps*”. “*Data*” replaces the shared drive that many labs already use.

Within the space for a lab’s user directories, distinct new usage patterns have emerged. Faculty have almost universally expressed interest in having a space within their Depot for each member of the research lab. To help with this, RCAC software engineers leveraged all the information known about the faculty member’s lab to auto-populate the “users” space for each lab member known to the system. At first, options to make user directories open to all members of the lab, or private to the student were enough, but several labs have articulated the need for an additional option for user directories to be open to the individual user plus the PI.

To meet researcher requirements for data sharing both within the institution and outside, having a high-performance tool for managing data transfer, research computing staff

deployed Globus as the recommended data transfer technology [25]. Globus allows for easy transfer of data among personal computers, campus supercomputers and storage systems, and national-scale cyberinfrastructure. In addition, Globus provides “Dropbox-like” file sharing capabilities for big data.

III. USE CASE: SOUNDSCAPE ANALYSIS

A. Soundscapes and Soundscape Analysis

To confront global environmental change and its effects on biodiversity, the scientific community needs to understand ecosystem and organism dynamics at large spatial and temporal scales [10], [11]. Soundscape analysis represents a new perspective to capture biological composition and dynamics over both large spatial and temporal scales. A soundscape is a collection of sounds produced by different entities in a landscape. All soundscapes are composed of three components: biological sounds that are emitted by a variety of animal taxa (biophony), sounds from geological or atmospheric events (geophony), and sounds produced by humans and human-built machines (anthrophony) [12], [13]. A soundscape is the result of the emission, transmission, and detection of all these sounds at a specific place and time. Because natural sounds are often linked with ecological functions, soundscapes are important sources of acoustic information that can be used to explore both ecological processes and the composition and dynamics of animal communities [12], [14]. Soundscape analyses are not species-centered; instead, they focus on the higher ecological levels of “community” and even “ecosystems”.

With recent technological advances, soundscapes can now be easily collected over large temporal scales with automatic sensor units [15], [16]. The sound libraries housed in museums, universities, and other research institutions around the world are continuously expanding their collections of soundscape recordings. For example, the Center for Global Soundscapes (CGS⁷) at Purdue University has already archived over 500,000 soundscapes recordings (of either 10, 15, or 60 minute durations) since 2008⁸. The massive volume of soundscape recordings poses two major challenges: first, how to store securely and organize clearly the recordings, and provide easy access via online tools [17]–[19]; and second, how to build distributed workflows to draw scientifically valid conclusions from such enormous amounts of soundscape recordings.

To address these challenges, CGS at Purdue developed a web-based software Pumilio [17] that systematically catalogs and manages the sound recordings. Pumilio stores recording metadata such as collection site descriptions (coordinates, habitat, pictures), sound descriptions (format, duration, sampling rate), and sound objects (spectrograms, oscillograms) in a structured, uniformed format. In addition, we redesign the analysis workflow from a sequential process to a large scale, parallel distributed process on HPC clusters. Both Pumilio and the distributed workflow impose requirements on the underlying storage system from different perspectives.

B. Storage System Requirements

Soundscape analysis has various requirements from all steps of data ingestion, sharing, indexing, analysis, storage etc. In this section, we discuss the requirements and how Data Depot and the services it provides fulfill the requirements.

Data Ingestion and Sharing: For the massive amount of sound recording data to be indexed, cataloged, and analyzed to gain scientific insights, the first step is to ingest the raw recording data to the storage system. In addition, to share raw data, processed data, as well as analytical results with research collaborators and citizen scientists, the data needs to be permanently stored and made available easily. To relief scientists from the burden of dealing with various connection and speed issues related to data transfer, Data Depot supports Globus Data Transfer [20]. Globus provides a graphical web tool for users to manage file transfers without worrying about the underlying technical details. The integration of Data Depot with Globus allows the researchers to authenticate their identities using their university credentials. This provides secure and easy access to the data. Moreover, Globus handles large files transfer failures transparently for users. It can be used on Windows, Linux, and Mac OS X [21].

Data Indexing and Analysis: Pumilio and the distributed workflow collaboratively solve the data indexing and analysis problems. They have different requirements from the perspective of user access control. Pumilio supports uploading and downloading sound recording files using its web interface for both research collaborators and citizen scientists. This requires flexible file and directory structures and permissions. It needs to allow Pumilio to make changes to uploaded sound files (e.g., change format, generate spectrograms, generate preview). At the same time, it needs to allow Pumilio to maintain a public directory that stores the sound recording files in mp3 format for public download. Thus, it has to provide flexible read, write, executable permissions for the directory structure. The distributed workflow runs directly on one of the community clusters (i.e., Rice⁹) with Data Depot mounted natively. It assumes the user has proper permissions to access Rice and the PBS queue on Rice to submit analysis jobs.

To fulfill the requirements imposed by both, and to provide scientists the ability to manage fine-grained access control without dealing with Linux commands, Data Depot integrates a web-based tool developed at Purdue University to manage permissions and access [22]. Research lab managers can use the web tool to grant read, write, and executable permissions to individuals by simply check the corresponding checkbox. The changes automatically populate in all the community cluster systems.

Data Storage: From the perspective of storage capability and speed, different types of data have different requirements. Specifically, raw recording data in lossless flac or wav format results in larger files. More specifically, a file contains one minute of recording in flac format at 44.1 kHz sampling rate is around 4.1 MB, and a file contains the same recording in wav

⁷<http://centerforglobalsoundscapes.org/>

⁸<http://1159sequoia05.fnr.purdue.edu/>

⁹<https://www.rcac.purdue.edu/compute/rice/>

format is around 10 MB. Compressed recording data in mp3 format for public listening and download is around 16 KB for the same recording. Currently, CGS holds 130 TB of recording data in flac format. We need to store the smaller mp3 files in a high speed file system for fast and parallel read access, and the larger flac and wav files in a storage system that may not be as fast, but need to be of large capacity.

Data Depot, which provides high speed, parallel file access, together with Fortress, which provides long term, large capacity, collaboratively meet the different capacity and speed requirements. The Fortress system is a large, long-term, multi-tiered file caching and storage system utilizing both online disk and robotic tape drives. It has over 200 TB of disk cache, and over 10 PB of tape-based long term storage space. All files stored on Fortress appear in at least two separate storage devices. One copy is permanently stored on tape, recently used or files smaller than 100 MB have their primary copy stored on a conventional spinning disk storage array (disk cache). Disk cache provides a rapid restore time. Both primary and secondary copies of larger files reside on separate tape cartridges in the robotic tape library. Researchers are able to integrate this hybrid storage system into their workflow by storing smaller files that need fast access in Data Depot, and storing larger files in the Fortress Archive system.

Scalability: Another important issue that is often encountered in big data analytics is scalability. When massive amount of data needs to be processed, many aspects of the workflow including algorithms, software stacks, and supporting hardware platforms can be the bottleneck of performance scalability. In our current Data Depot deployment, by enabling Data Depot on every node of the community clusters, users can utilize this high performance, parallel file system in their analytical workflow directly to reduce the file I/O bottleneck and to deploy the entire analytical workflow on High Performance Computing clusters such as Rice.

C. A Step Forward: Scaling Out Soundscape Analysis

Acoustic Analysis Workflow: With the support of Data Depot, we successfully scale out the soundscape analysis workflow on Rice using parallel computing techniques. The traditional workflow is a sequential process as shown in Figure 2. Recording devices allow scientist to collect sound data simultaneously from various sites and with a high temporal resolution. This results in a large number of collections of sound recording data. The recordings are stored in a set of external drives, and a powerful workstation is connected to the external drives. Scientists perform computation and analysis on the data using the workstation.

This workflow process is highly sequential in the sense that there is one single workstation. As the sound data grow to hundreds of terabytes, this workflow is not suitable due to limiting computation power and I/O performance. Here we show how to move the entire workflow to an HPC cluster to take advantage of parallel processing power, high performance file systems, and network connections.

Fig. 2: Traditional acoustic analysis workflow

In our new workflow as shown in Figure 3, the sound recording data is stored in Data Depot. The computation is done on Rice, which has Data Depot mounted. This way, we move the computation close to the data. In addition, we take advantage of the parallel computation power of the cluster. This general approach has been successfully adopted in other scientific applications, such as drug design and protein folding analysis [23], [24].

Fig. 3: Data Depot enabled acoustic analysis workflow

Our system allows users to specify options such as the locations and time periods of interest through user input. Given the user input, our Acoustic Analysis Workflow queries the Pumilio database to find the path to sound files in Data Depot. Then the workflow generates a list of PBS jobs that will be submitted to the cluster to run. Each job runs multiple Workers written in R in parallel where each Worker performs specific computation on specific dataset. More specifically, each Worker is a R script that runs on a single CPU core, processes a subset of the recording dataset, and calculates user specified α or β index. Each Worker reads input data from Depot, and writes analysis results back to Depot.

Framework in Action: The acoustic analysis presented

in this paper focuses on two types of acoustic diversity indices that quantify different acoustic activities of one or multiple sound recordings. These indices are classified in two families called α -diversity indices—indices measuring a level of acoustic activity on a single recording—and β -diversity indices—indices measuring the differences in acoustic activity between two recordings. We tested the computation of 11 α -diversity indices (Table I) and one β -diversity index.

TABLE I: List of the α -diversity indices used in this paper for performance testing. Frequency spectrum corresponds to the average of a short-term Fourier transform (STFT) with a non-overlapping Hanning window of 512 samples.

Index name	Calculation principle
Bioacoustics index Boelman et al. 2007	Area under the frequency spectrum.
Temporal entropy (Ht) Sueur et al. 2008b	Entropy of the amplitude envelope over the time units.
Spectral entropy (Hf) Sueur et al. 2008b	Entropy of the amplitude values over the frequency units. Spectrum maximized.
Acoustic entropy (H) Sueur et al. 2008b	$H_t * H_f$.
Acoustic complexity Index (ACI) Pieretti et al. 2011	Overall frequency variation contained in a spectrogram.
Acoustic evenness Index (AEI) Villanueva-Rivera et al. 2011	Evenness of the amplitude values over the frequency units.
Median of amplitude envelope (M) Depaetere et al. 2012	Median of amplitude envelope over the time units.
Normalized Difference Soundscape Index (NDSI) Kasten et al. 2012	Ratio of frequency band 1 to frequency band 2.
Acoustic Diversity index (ADI) Villanueva-Riviera et al. 2011	Entropy of the amplitude envelope over the frequency units. Spectrum not maximized.
Number of spectral peaks (NP) Gasc et al. 2013	The number of peaks on the mean frequency spectrum.
Spectral Dissimilarity (D_f) Sueur et al. 2008b	Differences of the amplitude values over the frequency units of two maximized spectrum.

Table II shows the performance results of our Data Depot supported workflow when calculating the α -diversity indices. It shows the number of files in the collection, the size of each collection, the number of nodes each job uses, and the walltime of each job. As we can see, the aggregated size of data that are processed is 44 TB, and we used 13 jobs to process them. Our system can successfully finish computation of each individual collection within 3 hours.

Figure 4 shows preliminary analytical results of the α -diversity indices. We show the 11 α -diversity indices calculated on 8 collections of distinct locations. The collections numbers shown in Figure 4 correspond to the numbers in the first column of Table II. Due to space limitation, we only showcase the indices NP and ACI. For each figure, the x-axis is the time of the day, ranging from 0 to 24. The y-axis is the index value for 8 collections with distinct locations. By looking at the NP plot, we observe that the same α -diversity index has different dynamics among the 8 collections, which indicates that the signing animal communities are different among collections. By comparing the two indices NP and ACI,

we observe that different α -diversity indices vary substantially for the same collection, which indicates that different diversity indices measure different aspects of sounds.

Fig. 4: α -diversity indices for 8 collections.

In comparison, the process of calculating the β -diversity index is more computation and data intensive than the α -diversity index calculation as we explain below. Hence it has never been tackle at this scale before in soundscape analysis.

The collection that we are considering contains 22,883 recordings of one-minute duration, sampled at 44100 Hz/16 bits. As discussed, β -diversity index calculates the differences in acoustic activity between any two recordings within or across collections. In this use case, we consider the pairwise β -diversity index within the collection. For this collection, there are 261,804,403 pairs of files that the β -diversity index need to be calculated.

To enable this workflow, the system generates individual PBS jobs that can finish using 60 Workers within 4 hours and submits to the cluster to run. The reason for choosing this size is due to the limits imposed by Rice scheduling policy: each job has a wall time limit of 4 hours. For different clusters, the size of the job may change. In this collection, our system submits 2,095 jobs. Each job runs 60 Workers in parallel, processes 125,000 lines of input. The average β -diversity index computation time for each PBS job is 2 hours and 20 seconds.

We observe that with the functionalities provided by Data Depot (i.e., high performance read and write I/O capabilities), we are able to scale the acoustic analysis workflow on Rice cluster to process hundreds of terabytes of data.

IV. DISCUSSION

The key design goal was to create in Depot a storage service for research data, with capabilities to address many real-world data use cases, and not merely be a network drive. Depot must be able to serve as a resource for many types of research data, and provide a suite of services that make it easy to grant access, read and write data, and collaborate throughout the entire data lifecycle.

In 15 months of production, the Depot has grown to serve nearly 280 research labs, from every academic college on

TABLE II: Resources required to calculate α -diversity indices on collections of sound recordings

Collection name	Short description	Num. of files	Size	Num. of nodes	Walltime (hh:mm:ss)
arizona2013	Long-term wildfire disturbance survey, desert ecosystems	121739	10 TB	24	2:18:05
arizona[1]	Short-term habitats description, desert ecosystems	531	46 GB	1	0:12:25
Borneo2014[2]	Short-term habitats description, tropical ecosystems	2663	252 GB	8	0:12:25
Chicago[3]	Long-term habitats description, urban ecosystems	53947	1.4 TB	8	2:41:05
Hoosier[4]	Short-term road disturbance survey, temperate ecosystems	15863	1.5 TB	8	0:47:09
Laselva2011[5]	Long-term habitats description, Tropical ecosystems	74204	3.7 TB	16	2:00:21
Leopold[6]	Short-term road disturbance survey, Temperate ecosystems	22401	2.4 TB	8	1:06:53
tip2008	Long-term habitats description, Temperate ecosystems	38464	5.0 TB	8	2:01:41
tip2009	Long-term habitats description, Temperate ecosystems	32425	6.5 TB	8	2:23:39
tip2010	Long-term habitats description, Temperate ecosystems	46937	5.8 TB	8	2:30:48
tip2011[7]	Long-term habitats description, Temperate ecosystems	47970	5.2 TB	8	2:32:14
tip2012	Long-term habitats description, Temperate ecosystems	22475	2.5 TB	8	1:08:14
wells[8]	Long-term habitats description, Estuary ecosystems	41847	3.3 TB	8	2:05:40

the Purdue campus, with significant growth coming from Agriculture, Humanities, and Health and Human Science. The average purchase is for nearly 10 TB, with multiple research groups investing in nearly 200 TB of storage each.

Depot supports a wide variety of domains and applications at Purdue, from traditional HPC applications, to long-tail domains. In addition to the Global Soundscapes application described in III, Depot supports many other domains illustrated in Table III.

TABLE III: Research Data Depot domains

Science Domain	Application	HPC User?
Genomics	Data buffer and storage for campus genomics core facility	HPC User
Psychology	Storing video and audio data of subject interviews, and sharing with collaborators at other institutions	Non-HPC
Nanotechnology, Chemical Engineering, Bioinformatics	Curation of lab or discipline-specific software stacks	HPC Users
Biology	Storage target for scientific instruments	HPC Users
Climate and Weather Modeling	Dataset storage, lab-wide software stacks	HPC Users
GIS	Working space for ArcGIS data	Non-HPC

V. TRANSITION TO PRESERVATION

In order to encourage researchers to make their data publicly available and comply with agency mandates, services needed to be developed to support data management that emphasize ease of use and transition between various stages of the research life cycle. One distinguishes three important stages with different requirements for providing access to data: first a collaborative stage, where relatively small teams collect, store, and share the data within the team; this is the stage that Depot is currently focused on; second, a publication stage where data is selected to become public; and third, a preservation stage where data can be migrated to archival formats.

Data in Soundscapes is segmented as follows: raw data and sound-recording samples in collections organized by

geographical area; analysis data produced by computations as described in Section III-C; and result data such as data supporting figures and tables in publications. Although agencies such as NSF require the sharing of primary data, this requirement is often disregarded for practical reasons. Instead, researchers may choose to make publicly available data that supports publications, such as result data. Advantages of doing so include supporting some measure of reproducibility if the codes used to produce these results are also available. Another is the practical concern of data size for sharing. Depot and the Soundscapes team are providing solutions to this challenge with Depot services, Globus transfer and Pumilio. In addition, Soundscapes abide by requirements when sharing raw data.

More specifically, the data management plan filed with the National Science Foundation contained some of the following data formatting, storage, archiving, and documenting and distribution protocols. First, all data will be stored in a non-commercial format (e.g., wav or FLAC) that is lossless; will exist in a variety of storage media including massive disk arrays and tape backup; all data, regardless of quality, will be archived to tape backup to the Purdue Fortress system; that the extent of the soundscape collections will be documented in a format that is accessible to the public (e.g., web site) and that files requested by the public or scientists be accessible given the constraints generally in existence with big data (e.g., they are large, complex, difficult to define boundaries). Tools that we have developed to manage and analyze the data are readily available too to the public and are accessible to common repositories such as GitHub and the R Project for Statistical Computing.

However, the preservation aspect of data sharing is currently missing with Depot. As an institutional data repository Depot lacks the ability to preserve data for the future, and to provide Digital Object Identifiers so that datasets or collections can acquire persistent identifiers. Long-term preservation was not a goal for this system and as with any university computing or storage resource, sustainability is a concern. The majority of the capital cost of Depot is expected to be recovered over the course of 5 years. However, from the researcher's perspective, having to fund most of the true cost of data storage from their own funds is a significant sustainability risk, especially

at large scales. Finding ways to centrally fund the storage system, and reduce the direct cost to researchers are areas to further explore.

As a way to investigate a solution to the preservation problem, the Depot team is exploring workflows for transitioning data to be preserved through PURR, the Purdue institutional repository [26], [27]. PURR is a web-based, collaborative system and preservation platform built on HUBzero [28], for publishing and preserving data produced by Purdue researchers. PURR is supported by a collaboration of the Purdue University Libraries, the Information Technology Office, and the Office of the Executive Vice President for Research and Partnerships. The PURR team has developed a Web interface and implemented a workflow that allows researchers working with data to collaborate in a private space. Researchers at Purdue can benefit from a 10 GB space allocation or 100 GB with a funded grant per project on PURR. The space limitation makes it an impractical solution for projects with large-volumes of data, such as Soundscapes. The value of PURR in the context of Soundscapes comes from the ability for a researcher to select datasets for publication and preservation thus reducing data sizes to be manageable in PURR. Data selected for publication acquire metadata and supporting documentation, such as citations for publications in PURR. When data is published, the attribution of a Digital Object Identifier (DOI) for the dataset increases discoverability by making the data citable [29]. PURR is committed to preserving all content for a period of ten years. PURR data is replicated on campus. For published data PURR subscribes to the services of the MetaArchive Cooperative, a 13-state, 3-countries trustworthy digital repository that uses the LOCKSS infrastructure to preserve data over a geographically distributed network [30], [31]. After 10 years, content in PURR will be weeded following collection de-accessing practices in the Purdue Libraries.

Depot and PURR are complementary systems that provide an end-to-end approach allowing researchers at Purdue University to collaborate on space optimized for labs with high-volume data (Depot), publish a selection of their data with DOIs, and preserve these data through well-established collection management and preservation practices (PURR). Each system provides a part of a comprehensive solution for collaboration and sharing with publication and preservation. Issues pertaining to the migration from one system to the other include the selection of datasets to be made available through PURR, the acquisition of machine-readable metadata, the crafting of annotations to make datasets usable and the sustainability of the platforms.

VI. CONCLUSION

We have described Research Data Depot, a high-volume collaborative platform at Purdue University, and the requirements, adoption and growth of this platform in 15 months of production. We provided a use case that illustrates how Depot is used to collaborate and serve data to a HPC cluster. We have also described how data in Depot can be made available for publication and preservation through PURR.

While the implementation of preservation workflows is part of future work, we are confident that the solution developed at Purdue University will allow researchers to experiment with new forms of data-intensive collaboration while being in compliance with funding agency requirements.

ACKNOWLEDGMENTS

The authors acknowledge the following grants and organizations for financial support for collecting data, technical and personal development: CNH-BCS NSF grant 1114945, AISL NSF grant 1323615, National Park Service, Purdue OCPR's incentive Grants Program, Department of Forestry and Natural Resources Wright Fund, College of Agriculture Information Technology. This research was supported in part through computational resources provided by Information Technology at Purdue, West Lafayette, Indiana. The authors want to thank Dante Francomano and Jarrod Doucette for their valuable inputs and discussions. Code samples and timing data are available on github¹⁰.

REFERENCES

- [1] P. B. Heidorn, "Shedding light on the dark data in the long tail of science," *Library Trends*, vol. 57, pp. 280–299, 2008.
- [2] J. C. Wallis, E. Rolando, and C. L. Borgman, "If we share data, will anyone use them? data sharing and reuse in the long tail of science and technology," *PloS one*, vol. 8, p. e67332, 2013.
- [3] S. L. Shreeves and M. H. Cragin, "Introduction: institutional repositories: current state and future," *Library Trends*, vol. 57, pp. 89–97, 2008.
- [4] C. Lynch, "Big data: how do your data grow?" *Nature*, vol. 455, pp. 28–29, 2008.
- [5] L. Pouchard, "Revisiting the data lifecycle with big data curation," *International Journal of Digital Curation*, vol. 10, pp. 176–192, 2016.
- [6] R. L. Moore, C. Baru, D. Baxter, G. C. Fox, A. Majumdar, P. Papadopoulos, W. Pfeiffer, R. S. Sinkovits, S. Strande, M. Tatineni, R. P. Wagner, N. Wilkins-Diehr, and M. L. Norman, "Gateways to discovery: cyberinfrastructure for the long tail of science," in *Proceedings of the 2014 Annual Conference on Extreme Science and Engineering Discovery Environment*, 2014, pp. 39:1–39:8.
- [7] M. A. Parsons, O. Godøy, E. Ledrew, T. F. De Bruin, B. Danis, S. Tomlinson, and D. Carlson, "A conceptual framework for managing very diverse data for complex, interdisciplinary science," *Journal of Information Science*, vol. 37, pp. 555–569, 2011.
- [8] K. Tenopir, S. Allard, K. Douglass, A. U. Aydinoglu, L. Wu, E. Read, M. Manoff, and M. Frame, "Data sharing by scientists: practices and perceptions," *PLoS one*, vol. 6, pp. 1–21, 2011.
- [9] A. G. Carlyle, S. L. Harrell, and P. M. Smith, "Cost-effective hpc: the community or the cloud?" in *Proceedings of the 2010 IEEE Second International Conference on Cloud Computing Technology and Science (CloudCom)*, 2010, pp. 169–176.
- [10] F. S. Chapin III, E. S. Zavaleta, V. Eviner, R. L. Naylor, P. M. Vitousek, H. L. Reynolds, D. U. Hooper, S. Lavorel, O. E. Sala, S. E. Hobbie, M. C. Mack, and S. Diaz, "Consequences of changing biodiversity," *Nature*, vol. 405, pp. 234–242, 2000.
- [11] A. Balmford, P. Crane, A. Dobson, R. E. Green, and G. M. Mace, "The 2010 challenge: data availability, information needs and extraterrestrial insights," *Philosophical Transactions of the Royal Society*, vol. 360, pp. 221–228, 2005.
- [12] B. C. Pijanowski, L. J. Villanueva-Rivera, S. L. Dumyahn, A. Farina, B. L. Krause, B. M. Napoletano, G. S. H., and N. Pieretti, "Soundscape ecology: the science of sound in the landscape," *BioScience*, vol. 61, no. 3, pp. 203–216, 2011.
- [13] A. Farina, *Soundscape ecology principles, patterns, methods and applications*. Springer, 2014.
- [14] J. Sueur, S. Pavoine, O. Hamerlynck, and S. Duvail, "Rapid acoustic survey for biodiversity appraisal," *PloS one*, vol. 3, no. 12, p. e4065, 2008.
- [15] M. A. Acevedo and L. J. Villanueva-Rivera, "Using automated digital recording systems as effective tools for the monitoring of birds and amphibians," *Wildlife Society Bulletin*, vol. 34, no. 1, pp. 211–214, 2006.

¹⁰https://github.rcac.purdue.edu/RCAC-Staff/sound_record

- [16] T. S. Brandes, "Automated sound recording and analysis techniques for bird surveys and conservation," *Bird Conservation International*, vol. 18, pp. S163–S173, 2008.
- [17] L. J. Villanueva-Rivera and B. C. Pijanowski, "Pumilio: a web-based management system for ecological recordings," *The Bulletin of the Ecological Society of America*, vol. 93, pp. 71–81, 2012.
- [18] J. Zhang, K. Huang, M. Cottman-Fields, A. Trusking, P. Roe, S. Duan, X. Dong, M. Towsey, and J. Wimmer, "Managing and analysing big audio data for environmental monitoring," in *Proceedings of the 2013 IEEE 16th International Conference on Computational Science and Engineering*, 2013, pp. 997–1004.
- [19] E. Baker, B. W. Price, S. D. Rycroft, J. Hill, and V. S. Smith, "BioAcoustica: a free and open repository and analysis platform for bioacoustics," *Database*, vol. 2015, 2015.
- [20] K. Chard, S. Tuecke, and I. Foster, "Efficient and secure transfer, synchronization, and sharing of big data," *IEEE Cloud Computing*, vol. 1, pp. 46–55, 2014.
- [21] I. Foster, "Globus toolkit version 4: software for service-oriented systems," in *Proceedings of the IFIP International Conference on Network and Parallel Computing*, 2006, pp. 2–13.
- [22] K. D. Colby, D. T. Dietz, P. M. Smith, and D. D. Cumberland, "Self-service queue and user management in shared clusters," in *Proceedings of the First International Workshop on HPC User Support Tools*, 2014, pp. 22–31.
- [23] T. Estrada, B. Zhang, P. Cicotti, R. S. Armen, and M. Taufer, "Reengineering high-throughput molecular datasets for scalable clustering using MapReduce," in *Proceedings of the 14th IEEE International Conference on High Performance Computing and Communications*, June 2012, pp. 351–359.
- [24] B. Zhang, T. Estrada, P. Cicotti, and M. Taufer, "Enabling in-situ data analysis for large protein folding trajectory datasets," in *Proceedings of 28th IEEE International Parallel and Distributed Processing Symposium (IPDPS)*, May 2014, pp. 221–230.
- [25] I. Foster, "Globus online: accelerating and democratizing science through cloud-based services," *IEEE Internet Computing*, vol. 15, p. 70, 2011.
- [26] M. Witt, "Co-designing, co-developing, and co-implementing an institutional data repository service," *Journal of Library Administration*, vol. 52, pp. 172–188, 2012.
- [27] C. C. Dearborn, A. J. Barton, and N. A. Harmeyer, "The Purdue University Research Repository: HUBzero customization for dataset publication and digital preservation," *OCLC Systems & Services*, vol. 30, pp. 15–27, 2014.
- [28] M. McLennan and R. Kennell, "HUBzero: a platform for dissemination and collaboration in computational science and engineering," *Computing in Science & Engineering*, vol. 12, no. 2, pp. 48–53, 2010.
- [29] Brase, "DataCite - a global registration agency for research data," in *Proceedings of the Fourth International Conference on Cooperation and Promotion of Information Resources in Science and Technology*, 2009, pp. 257–261.
- [30] K. Skinner and M. Halbert, "The MetaArchive cooperative: a collaborative approach to distributed digital preservation," *Library Trends*, vol. 57, pp. 371–392, 2009.
- [31] K. Skinner and M. Halbert, "MetaArchive: a cooperative approach to distributed digital preservation," *Against the Grain*, vol. 21, p. 6, 2013.