Developing Library Cyberinfrastructure Strategy for Big Data Sharing and Reuse
Virginia Tech Libraries
Jan 10, 2016

Abstract

Virginia Tech Libraries, in partnership with the departments of Mechanical Engineering and Computer Science, as well as the University of North Texas Department of Library and Information Sciences, propose a 2-year research grant toward an evidence-based, broadly adaptable library cyberinfrastructure (CI) strategy for big data sharing and reuse.

There are 3 major library big data sharing and reuse service patterns, whose nature and performance characteristics are far from being thoroughly understood. Intelligently matching these service patterns with appropriate CI resources constitutes an even greater challenge. In this research, we will conduct controlled experiments and benchmarking against 5 or more such operational scenarios, then generalize the insight gained into rules. We will then consolidate these rules into a widely applicable strategy, and validate it through service optimization.

The intended audience of this project is any academic, research, or public library interested in operating such services on broadly shared CI resources -- especially library strategists and IT managers leading or actively planning to deploy big data services. The strategy resulting from this research will equip these libraries with solid knowledge and techniques to leverage shared CI resources and balance desires, needs, and constraints with a clear understanding of the tradeoffs.

This research primarily addresses the IMLS agency-level goal (c), content. Our performance goal is to broaden access and expand use of big data sets by improving CI strategies for related library services. It will have the following deliverables and products: 1) A set of performance measures identified from the literature on CI strategies; 2) Documentation of experiments about 5 or more scenarios (covering three patterns of services on CI configurations using different datasets, including design, procedures, and analysis of performance); 3) A comparative analysis and summary of different CI configurations and scenarios; 4) A replicable CI strategy to mix and match these options to maximize library capacity; 5) Improved access to library big data services, and 6) A comprehensive website containing all materials relevant to this project.

Big data poses immense challenges to the library and its CI. This research furthers the thrust of library data management activities and pushes the library beyond its traditional focus on a large variety of small data sets and their preservation. It facilitates more efficient and effective use of big datasets archived at the library, and advances data intensive scientific discovery and STEM learning. In addition, library and information science faculty and students can use the outcomes of this project to enrich their curriculum on big data and library services, or to learn practical knowledge on the subject as career preparation.

Developing Library Cyberinfrastructure Strategy for Big Data Sharing and Reuse

1. Statement of Need

With an emphasis on big data sharing and reuse, this research project aims to develop an evidence-based, broadly adaptable cyberinfrastructure (CI) strategy to operate digital library services. The intended audience is any academic[1], research, or public library[2] interested in operating such services but lacking resources for exclusive, end-to-end solutions. Of particular interest are those library technology strategists and IT managers who are leading or actively planning to deploy big data services, as exemplified by members of our advisory committee[3]. The resulting strategy will equip these libraries with solid knowledge and techniques to leverage shared CI resources and balance their desires, needs, and constraints with a clear understanding of the tradeoffs.

As a key component of the nation's knowledge infrastructure, libraries must continuously reinvent themselves with the emergence and the establishment of new discovery paradigms. The recent wave of data intensive science has motivated many high-profile library big data services, notably the ambitious plan to archive all tweets at the Library of Congress [1, 2, 3], the heterogeneous and geographically replicated archival storage known as the Digital Preservation Network (DPN) [4, 5], the data mining facility at the HathiTrust Research Center (HTRC) [6 - 8], and the metadata hubs developed at the Digital Public Library of America (DPLA) [9, 10] and the SHARE initiative [11]. Many more are being developed or being planned. The scope of this project will be limited to the technical infrastructure of such services and its implications for staff training, two important components of the National Digital Platform.

Since building large, single-tenant data centers at each library would be prohibitively expensive, this project will focus on the use of widely available, shared CI resources. Supporting Document 1 provides an environmental scan that summarizes the CI choices of 16 library big data efforts, which clearly shows that the majority use either general-purpose, shared CI options (9 projects) or their exclusive clones or close relatives (6 projects). Since most of the projects we reviewed are purposefully built to be portable (only 5 out of 16 projects do not explicitly provide source code), and even easily deployable and replicable on shared CI (e.g., projects 2, 7, and 9), any library, regardless of size and budget, can be infrastructurally equipped to offer big data services when needed. The main technical obstacle is usually the lack of experience and knowhow.

Based on funding sources, the shared CI may be categorized into 4 types, each with its unique strengths, weaknesses, and challenges: 1) Institutional high-performance computing (HPC), high-throughput computing (HTC) and storage facilities, e.g., Indiana University's Big Red II, Virginia Tech's BlueRidge, etc.; 2) National HPC, HTC, and storage facilities, most notably XSEDE resources [12]; 3) National research clouds such as Chameleon Cloud, CloudLab, Open Science Data Cloud, etc.; 4) Commercial clouds, such as Amazon Web Services (AWS), Rackspace, etc. To date, there is no unified CI framework or strategy that could be referred to as best practice to pick CI for different library big data sharing and reuse situations.

Operating library big data services on shared CI resources is far from turnkey. Although some general guidelines exist [13, 14, 15], evidence-based practical advice and decision trees for library CI building are

[1] See Supporting Document 4
[2] See Supporting Document 3
[3] See Projectstaff.pdf for the full list and Resumes.pdf for their resumes

lacking. The few published cost analyses are confined to data storage or running LOCKSS boxes in the cloud [16, 17]. Hard-pressed by project schedules and deadlines, library IT managers tend to move in haste without fully evaluating the pros and cons. Even those deeply involved usually gain insights only through expensive trial-and-error cycles. Much of the experience gained, nonetheless, is ad hoc and far from being systematically benchmarked, evaluated, summarized, documented, and synthesized so as to be widely applicable.

Furthermore, big data based discovery processes are markedly different from the existing digital library workflows, leading to an urgent need to assimilate library services into active knowledge discovery processes [18]. Since a download link is no longer sufficient to provide effective access to large volumes of research data, libraries are increasingly expected to deliver not the raw data but the knowledge extracted from them (projects 1, 6-11, 14, and 15). Data mining and analysis requires significantly more processing power and higher rates when loading data from storage to processing nodes, further constraining the CI strategy [19]. Accordingly, the focus of this research goes beyond mass storage and bit rot to the CI challenges posed by big data sharing and reuse.

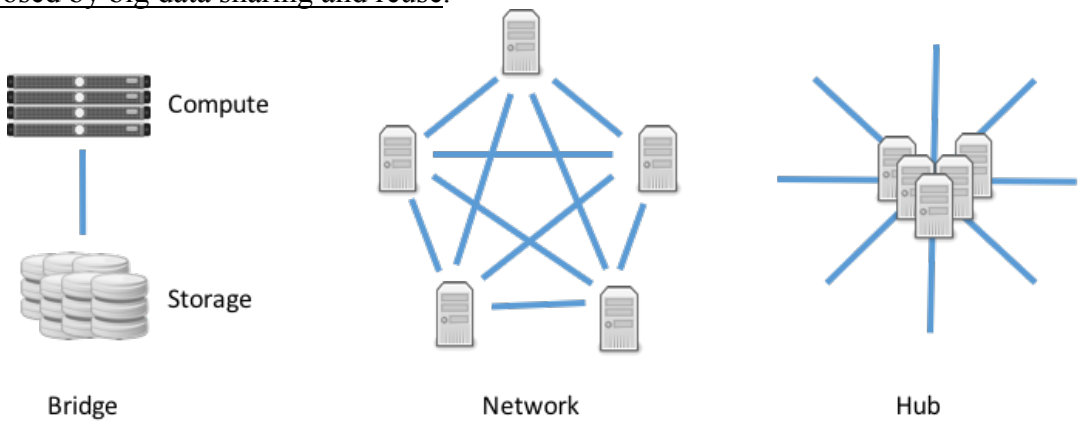


Figure 1. Three Patterns for Library Big Data Services

Conceptually, we can draw three distinct, although not mutually exclusive, service patterns from the environmental scan, schematically shown in Figure 1: 1) The Bridge Pattern, exemplified by projects 6, 9, 10, 12, 13, and 16, clearly separates the data storage and processing, and answers sporadic, on-demand, and sometimes user-specified computing needs by moving data from storage to processing nodes through the network link between them. The DPN nodes (projects 2-5), despite being primarily concerned with data storage, may be considered special cases of the Bridge Pattern. This is because the data ingestion, validation, periodic fixity checking, and refreshment are indeed on-demand data processing performed at compute nodes away from where the data is stored. 2) The Network Pattern is illustrated by projects 11, 14, and 15, and features a much tighter integration between data storage and processing. Typically involving a Hadoop cluster, this pattern uses a large number of interconnected nodes, each serving both as a storage and processing unit. These nodes intelligently replicate, balance, and co-optimize both storage and computing across the interconnections. The Network Pattern excels at MapReduce types of computation and can sustain high processing loads. However, the initial data loading stage is known to be a bottleneck [20]. As a result, data tend to be "sticky" to the CI. Once loaded, the data usually stay put. Project 12, however, is a curious exception to this general rule. 3) The Hub Pattern includes projects 7 and 8, both specialized as metadata hubs. This service pattern continuously draws live data from potentially many sources, undertakes necessary processing, then disseminates processed information to potentially large numbers of data consumers. It has higher quality of service (QoS) requirements, since downtime may lead to permanent data loss. In addition to

the performance requirements on the computing and storage nodes, it also requires stable network connections to the external systems upon which it depends.

The nature and characteristics of these service patterns are far from being thoroughly understood. The project team has conducted some preliminary work [19, 21 - 24,] and uncovered a number of CI tradeoffs, although more work is needed to consolidate them. Intelligently matching these service patterns with appropriate CI options constitutes an even greater challenge. For example, except for CI availability and staff expertise constraints, it is not clear to outsiders why many of the reviewed projects made their current CI choices or suddenly changed from choice to another. What are the key considerations and tradeoffs? How heavy is the financial or personnel burden, and how do these projects balance cost vs. functionality? Making CI decisions without a clear understanding of these issues will be costly. Consider, for example, the much delayed Library of Congress Twitter Archive, which in 2013 took 24 hours to complete a simple query [1, 2]. From the limited information available, we speculate that the service may have been conceived as a Bridge Pattern service but should be a Network or Hub Pattern service, or a combination of both. This project, consequently, is designed to aid strategizing such decision-making. This will be achieved by generalizing into rules insights gained through controlled experiments and benchmarking of one representative project in each service pattern, namely the Goodwin Hall data management system [19, 21] developed for Virginia Tech Smart Infrastructure Lab (VTSIL)[4] by Virginia Tech Libraries[5], the Event Digital Library and Archive [25 - 27] developed at Virginia Tech Digital Library Research Lab (DLRL)[6], and SHARE [11] Notify[7]; see Table 1 for details. We will then consolidate these rules to more widely applicable strategy and validate them through service optimization. It is not within the scope of this project, however, to research on scaling these library services to handle higher load on a certain type of CI.

2. Impact

This research primarily addresses the IMLS agency-level goal (c), content. Our performance goal is to broaden access and expand use of big data sets by improving CI strategy for related library services. It will have the following deliverables and products: 1) A set of performance measures identified from the literature on CI strategies; 2) Documentation of experiments about 5 or more scenarios (covering three patterns of services on CI configurations using different datasets, including design, procedures, and analysis of performance); 3) A comparative analysis and summary of different CI configurations and scenarios; 4) A replicable CI strategy to mix and match these options to maximize library capacity; 5) Improved access to library big data services, and 6) A comprehensive website containing all materials relevant to this project.

These deliverables will bridge the critical infrastructural gap between library research and service, and facilitate capacity building in a distributed, self-motivated fashion. While data consultancy services are mushrooming in university libraries [28], we need to go beyond the advocacy role and back up what we preach with solid computer systems and services. Currently most libraries set arbitrarily low size limits for data deposit, and can hardly provide much discovery capability beyond download links. This research furthers the thrust of library data management activities and pushes the library beyond its traditional focus on a large variety of small data sets and their preservation. It will not only prove that more comprehensive big data services can be provided by a library with modest means, but also show how to do this properly.

[4] http://www.me.vt.edu/vtsil/
[5] http://github.com/VTUL
[6] http://www.dlib.vt.edu/
[7] http://www.share-research.org/projects/share-notify/

Table 1. Three Target Services

	Goodwin Hall Living Lab	Event DL & Archive	SHARE Notify
Operator	VTSIL, directed by co-PI Tarazaga. Data management system developed by PI Xie	DLRL, directed by co-PI Fox	SHARE, directed by co-PI Walters
Designated CI	AWS	DLRL Hadoop Cluster	Rackspace
Service Pattern	Bridge	Network	Hub
Data Type	Vibration, temperature, and flow data collected from 300+ sensors embedded in Goodwin Hall	Webpages crawled from the web and tweets gathered from Twitter APIs	Free, open data set about research and scholarly activities gathered from various sources
Data Volume	~30TB/year or higher	~1 billion tweets & 11TB of webpages	~4 million events in 8 months
Discipline	Math, Mechanical, Systems, & Industrial Engineering, Music, Visual Arts, etc.	Computer Science, Data Science, History, Sociology, Law, Medicine, etc.	All disciplines
Usage Examples	Detect footsteps, evacuation planning, HVAC optimization, etc.	Track and analyze live events such as earthquakes, political events, community activities, and violence, crime prevention, etc	Linking publications to grants, receive real time event notifications on mobile devices, etc.
Used for STEM Learning	Graduate/Undergraduate Vibrations, System Dynamics, Mechanical Engineering Labs, Approximation of Dynamical Systems	Multimedia, Hypertext and Information Access, Computational Linguistics, Digital Libraries, and Information Retrieval	
Service Status	Development done, operate intermittently using AWS Research Grant funding	Development done, in experimental operation	Development done, in experimental operation

Taking large datasets out of the dark and putting them into active use is a critical contribution the library community can provide to data intensive science and STEM learning. See Table 1, considering as examples the 3 projects to be used in this research. Both VTSIL and DLRL have attracted sizable multidisciplinary user communities and are used in STEM learning [25 - 27] extensively, e.g., in 11 undergraduate or graduate courses. SHARE Notify, on the other hand, was prompted by the 2013 White House OSTP directive to expand access to federally funded research and are connecting more than 80 data providers with diverse users[8]. An improved CI strategy will amplify the benefits of these and many other similar projects by helping provide these services in a more efficient and effective manner.

[8] http://osf.io/share/

This project aggressively collaborates across organizational and disciplinary boundaries. It is a joint project among the data creator, the domain scientist, the computer scientist, the digital library practitioner, and the information systems evaluation expert. It embeds skilled librarians in active big data projects and integrates library CI and services more tightly with them. Even in the prior work leading to the Goodwin Hall project, library staff has been routinely placed in the research labs and worked closely with scientists from various disciplines. They programed not only databases and websites, but also data acquisition hardware and routers[9]. They encoded data, extracted technical metadata, as well as filtered signals to detect earthquakes, footsteps, and faulty wires[10]. The library has become an integral component of the research lab. The SHARE project represents another aspect of aggressive collaboration of this research. A joint project of ARL, AAU, and APLU, its target audience covers the whole higher education sector.

Still, even the best strategy would not achieve fruition without a skilled workforce to carry it out. Although not a Learning in Libraries proposal, this project contributes important knowledge on skill gaps so that future librarian can operate state-of-the-art CI and systems. Library and information science faculty and students can use the outcomes of this project to enrich their curriculum on big data and library services, or to learn practical knowledge on the subject as career preparation. More specifically, it will inform the curriculum enhancement currently underway at UNT Library and Information Science (LIS) program, and allow expansion of the NSF-funded digital library learning modules now available through Wikiversity (originated by VT Computer Science and the School of Information and Library Science at UNC Chapel Hill). Team members of this project will also communicate with the above mentioned intended audience via conferences and regular courses in LIS, as detailed in the Communications Plan section.

3. Project Design

In order to achieve our performance goal, we set the following objectives to specifically address National Digital Platform key themes of enhancing services and professional learning:

- Achieve better understanding of the relation between library big data services and the CI.
- Devise a theoretical framework based on which we can match the services with CI.
- Verify the theory with experimental service optimization.
- Provide educational and reference resources.

More specifically, we attempt to answer the following research questions related to operating library big data services on shared CI:

- What are the key technical challenges?
- What are the monetary and non-monetary (time, skill set, administrative, etc.) costs? Are there any cost patterns or correlations to the CI options?
- What are the knowledge and skill requirements for librarians?
- What are the key service and performance characteristics?
- For Bridge Pattern services, how to balance the storage, processing capacities, and network bandwidth?
- For Network Pattern services, how to quantify the data loading bottleneck?
- For Hub Pattern services, how to quantify the effects of QoS degradation?

[9] https://github.com/VTUL/VT-SIL-Collection-System
[10] https://github.com/VTUL/Goodwin_Data_Acquisition

- How to consolidate the answers to the above questions to form an easy to adapt and effective library CI strategy?

We adhere to the principle of evidence-based decision making. Since little in-depth experience and knowhow in the field is published or available for analysis, we must first gather sufficient evidence to support the strategy under development. This project team is uniquely positioned to carry out this research. See Table 1 and also refer to Supporting Document 1; we have had the rare opportunity to have led developing services covering all three patterns. The team members have already collaborated with each other extensively, as shown in the co-authorship in our resumes. All three services are at the stage of experimental operation and are amenable to exploring CI optimizations. Since these systems and services are all developed in-house, they are easier to customize in our case to customize. This allows us to perform controlled experiments, a key element for scientific discovery. All 3 services handle high volumes of data, therefore benchmarking may be run on full-size, real-world CI for sufficiently long periods of time using real data. This ensures the strategy under development will not be trivial or fall into corner cases. Furthermore, by designating this proposal as a research instead of an implementation project, we have more freedom to examine the evidence in broader contexts.

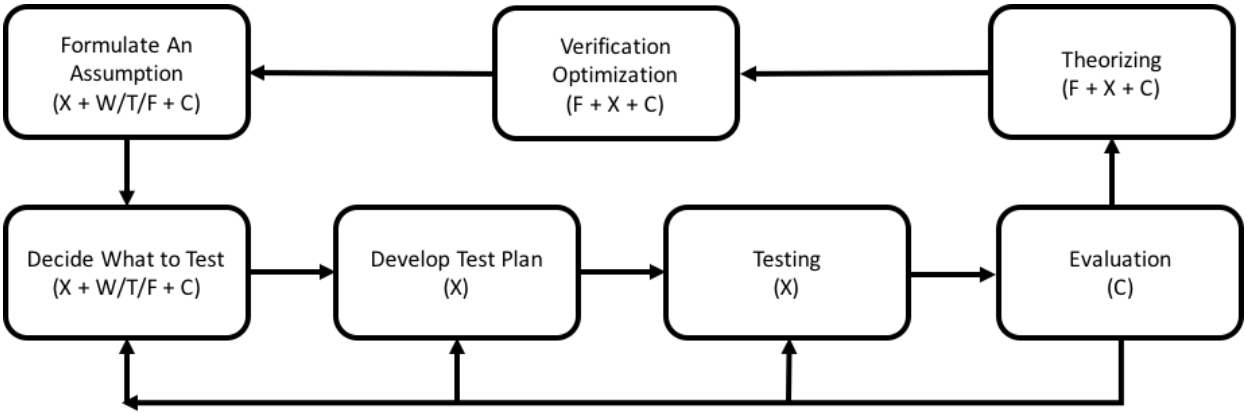


Figure 2. Activity Workflow

Since the close relations between the VT team might introduce unintended bias, we establish an iterative activity workflow with an integrated external gatekeeper, as shown in Figure 2. The uppercase letter in the parenthesis stands for the last name initial of a (co-)PI. The first (co-)PI is the lead of the activity. The two stages on the far left of the workflow usually only need to involve one service operator, depending on which pattern is to be tested. The UNT team, lead by co-PI Chen, will conduct independent evaluations and reviews on every iteration with the full cooperation from the VT team. Successful testing can move on to theorizing, otherwise immediately loop back to a stage where corrections or improvements can be made. This workflow, however, potentially creates dependencies that delay the progress. To overcome this difficulty, we will use Virginia Tech Libraries' established agile development tools, including github, confluence, and jira, to implement this workflow, where the UNT team will act as the default reviewer.

We will repeat these activities on all three target services operating on 4 different CI options, including: 1) HPC and storage provided by Virginia Tech Advanced Research Computing (VT ARC[11]), 2) HPC and storage provided by XSEDE[12] via Texas Advanced Computing Center (TACC), 3) NSF Chameleon Cloud[13],

[11] http://www.arc.vt.edu/
[12] http://www.xsede.org/

and 4) Amazon Web Services (AWS)[14]. The CI options are chosen with the consideration that they should be widely available to many libraries. XSEDE and Chameleon resources may be requested for free usage by most academic and public libraries through an application procedure. Like many other institutional CI resources, VT ARC is only open to users within the same institute, but similar CI is fairly common in higher education institutes. AWS, on the other hand, is open to anyone with a credit card.

However, not every service and CI combination is sensible, practical, or applicable for our research. Table 2 shows a matrix of test scenarios we will be performing, with the numbers denoting their priorities in this project. 5 is the highest, and 0 denotes non-applicable. Most general-purpose HPC environments do not give users the level of control required to operate Hadoop or indexing services, therefore the 4 zeros in the upper-right corner. The Goodwin Hall project was developed for AWS [19] and we have already accumulated some data. Chameleon is still under experimental operation and the approved credit may soon deplete and it may be hard to recharge. We therefore grant higher priority to the Bridge Pattern which can be tested on all 4 options. The two lower priority Chameleon test cases will only be performed if both time and service credit is available. In contrast to TACC, the resources from VT ARC is better guaranteed[15], therefore deserve higher priority.

Table 2. Test Scenario Matrix

	Goodwin Hall Living Lab	Event DL & Archive	SHARE Notify
VT ARC	5	0	0
TACC	4	0	0
Chameleon	5	3	3
AWS	0	5	5

Developing plans to test depends heavily on the assumptions made and what we want to test. As an example, we have an assumption that there should exist some kind of balance between storage, bandwidth, and computing capacities for the Bridge Pattern services. In our prior work [19, 21, 22], we have attempted to control these factors by adjusting 1) the workload, e.g., a simple ingestion task vs. heavy numeric analysis workloads, which tend to overload the computing nodes before bandwidth is saturated; 2) the computing capacity, e.g., changing the number and the type of computing nodes; 3) the bandwidth, by moving storage from co-located block storage to tape storage across the continent, etc. It is likely that a different testing plan will be needed for each test, but the general idea of controlled experiment will persist.

3.1 Evaluation

Evaluation is an integral part of this project. It will be continuous, and will permeate all stages and aspects of this research, from formulating performance goals to benchmarking design, result analysis, and theorizing and verification. In addition to the evaluation depicted in Figure 2, we will also evaluate the outcomes and the impact of the project. Both types of evaluations will be guided by a comprehensive evaluation plan to be developed in the beginning of the project.

[13] http://www.chameleoncloud.org/
[14] http://aws.amazon.com/
[15] See Partnercommitment2.pdf

The CI strategy evaluation as described in Figure 2 will examine each activity to ensure it is well grounded and conducted. We will consider system evaluation measures as we have employed in developing high performance information access systems and user-friendly systems. In the evaluation plan, we will identify types of data to be collected and methods of analysis for evaluation. For examples, we will consider the initial loading time, the response time to different discovery and access tasks, and their relationships with the types and the size of the datasets.

The project evaluation will employ basic principles and best practices for measuring success in project management, such as budget, timing, scope, quality, client/user satisfaction, project team satisfaction, and personal and professional development. Specifically, we will evaluate this project against its goals, objectives, and deliverables - the extent to which it achieves the stated goals, objectives, and generates the desired deliverables. One indicator can be the date that each proposed project activity is completed, as project goals are realized and deliverables are generated through project activities.

The evaluation will also examine the impact of the project: The outcomes and impact of research projects arise since research findings are presented and published. Accordingly, the research team will disseminate the project work via publications, conference presentations, invited talks, and news media, in a timely manner. Log analysis of the project website and altmetrics as well as citation analysis of the publications will indicate the impact of the project. Outcome indicators that this research product has successfully reached its intended audiences include:

- For leaders and IT managers of libraries and big data researchers: altmetrics, citations, and download counts of the project publications;
- For library and information science faculty and students: number of courses that use the project materials.

4. Project Resources: Personnel, Time, Budget

4.1 Personnel

PI Zhiwu Xie ([redacted] time for 2 calendar years) is an associate professor and directs the digital library development team at Virginia Tech Libraries. He leads the development of the Goodwin Hall Living Lab data management system, IMLS ETDplus Workbench, VTechData, and UWS through transactional web archiving, among others. He closely collaborates with VTSIL, DLRL, and SHARE, and served on technical committees of Fedora, APTrust, PREMIS, ResourceSync, and Altmetrics Data Quality efforts. His research extensively utilizes all types of CI options summarized in this proposal, and has been supported by Mellon, IBM, Amazon, USGS, and NSF XSEDE. Dr. Xie will oversee all aspects of the grant work and coordination between partners. He will also lead the service evaluation and porting, benchmarking, and validation.

Co-PI Tyler Walters ([redacted] time for 2 years) is the Dean of University Libraries and a professor at Virginia Tech. He serves on many professional bodies including NDSA, CNI, NISO, OR, IDCC, IJDC, MetaArchive, and Educopia. He is a founding director of the SHARE project and oversees its management and implementation. Dr. Walters contributes to project leadership, administrative support, and SHARE project support. He will also coordinate the advisory committee and lead the communication, dissemination, and digital stewardship.

Co-PI Edward Fox ([redacted] summer month per year for 2 years) is a professor in the Department of Computer Science, Virginia Tech. He is a senior computer scientist and digital library innovator. He directs

DLRL and NDLTD, and developed the VT Event DL and Archive. He chaired the IEEE Technical Committee on Digital Libraries, ACM SIGIR, and JCDL steering committee, and has been (co-)PI on 120 research grants/contracts, (co-)authored 18 books, 117 journal/magazine articles, 49 book chapters, 202 refereed conference/workshop papers, 69 posters, and over 150 other publications/reports, with an h-index of over 50. In this project, Dr. Fox will lead the theorizing and strategy consolidation, and manage the Event DL and Archive as well as the DLRL Hadoop cluster[16].

Co-PI Pablo Tarazaga ([redacted] summer month per year for 2 years) is an assistant professor in the Department of Mechanical Engineering, Virginia Tech. He founded VTSIL and VAST, through which he built the Goodwin Hall Living Lab, the world's most instrumented building for vibration. Under his lead, the Goodwin Hall project attracts a data user community with more than 40 members across many countries and disciplines from performing arts to mathematics and engineering. His research has been supported by NSF, USAF, NIST, and many corporations. Dr Tarazaga manages the Living Lab and will share with this project the vast sensor data set and liaison with the user community, against which the evidence for CI strategy will be collected.

Co-PI Jiangping Chen ([redacted] time during the summer for the 1st year, then 32% time during the summer for the 2nd year) is an associate professor in the Department of Library and Information Sciences, University of North Texas. She specializes in Information Systems Design, Analysis, Evaluation, and Usability Studies, Intelligent Information Access and Knowledge Discovery, and Digital Libraries. She is the PI of two IMLS National Leadership Grant projects and the Editor-in-Chief of The Electronic Library. Dr. Chen has extensive experience in evaluating different information systems and applications [45, 46, 47, 48]. She will lead the evaluation throughout the project period, and will communicate the skill gaps discovered in this research to the curriculum enhancement currently underway at UNT MLIS program and beyond.

Two graduate research assistants will also be supported by this project. The GRA at Virginia Tech (1.5 person-calendar year) will focus on service porting and benchmarking, and the GRA at UNT (5 person-academic month) will focus on evaluation and documentation.

An **Advisory Committee** will be established to meet every two months via conference calls. The purpose is to evaluate project progress and provide feedback regarding the applicability of this research in local contexts. No funds are requested for the committee activities. 11 committee members, listed in the project staff document[17], are either digital library strategists and innovators deeply involved or interested in providing big data services, or experts on information systems evaluation.

4.2 Facilities, Equipment, and Supplies

This project will utilize computing, storage, and network resources from Virginia Tech ARC, TACC, Chameleon, Amazon, and DLRL. Virginia Tech has committed to this project 25TB of storage and necessary computing and networking resources for the period of the award. We have also secured resources from TACC Stampede (24227 Service Units remaining as of this writing), and Chameleon (20000 Service Units), with future augmentations available and subject to XSEDE and Chameleon application and allocation cycles. We include in the budget the cost for necessary AWS resources, estimated from the AWS Solution Calculator. We request 5% machine time, or 1.2 months throughout the project period, for each AWS porting

[16] http://www.eventsarchive.org/node/12
[17] See Projectstaff.pdf

and operation. This includes all experimental design and testing, benchmarking, and validation at different stages of the research; therefore each experiment will last in order of days. The DLRL cluster, currently an exclusive CI, soon to have a 10 Gbps connection to the Internet, will also be used for comparison. Its usage is committed through co-PI Fox. Please refer to the letters of commitment and supporting documents for the details of these resources and allocations.

4.3 Travel

We also request $20,000 travel funding for five (co-)PIs and/or GAs to attend scholarly and professional meetings and conferences deliberated in the communications plan. These funds will cover conference registration, airfare, hotel, and per diem.

4.4 Timeline

The project will be carried out in 4 phases. The schedule of completion sheet[18] contains more details. After the initial 3-month planning phase, we will perform 5 testing scenarios at the pace of 3 months each. Each test will adopt the activity workflow depicted in Figure 2 and involve the full team. Tests for different scenarios may be interleaved, but we consciously avoid parallelizing them, so that insights gained in completed tests may be used as input to design the future ones. The test schedule may also be adjusted to accommodate XSEDE allocation cycles, then use AWS tests as fillers. The last 6 months will be evenly split on the strategy consolidation phase and documentation/project wrap-up/dissemination.

5. Communications Plan

Research results will be shared with key audiences interested in CI strategies for big data sharing and reuse. These key audiences include leaders and IT managers in public and academic libraries and other organizations, big data researchers and data scientists, and library and information science faculty and students. Specific efforts, such as attending professional conferences and meetings, and posting news announcements on social networking websites, will be used to reach key audiences. Scholarly and professional conferences including but not limited to JCDL, ALA Annual Conferences or ALA LITA Forums, ASIS&T Annual Meeting, DLF Forum, CNI Meeting, Open Repositories, IEEE Big Data, VLDB, ICDE, CIKM, SIGMOD, IMAC will be venues to present tutorials, papers, and demonstrations. In addition, articles derived from the research activities will be submitted to peer-reviewed scholarly and professional publications such as JASIST, IJDL, the Electronic Library, and others. Also, the research outcomes from the project will be shared with graduate students in Virginia Tech's Department of Computer Science, and the library and information science program at the University of North Texas through courses such as data modeling for information professionals and computational methods for information systems.

To sustain the benefits of this project and its influence nationwide, a project website will be established soon after the project is approved and will be continually maintained for 3 years after the project completes. All publications, presentations, posters, and related reference materials will be publically available via the project website, which will be hosted by Virginia Tech Libraries. Research datasets and appropriate project documents will be archived in VTechWorks[19] and VTechData[20] for long term access.

[18] See Scheduleofcompletion.pdf
[19] http://vtechworks.lib.vt.edu
[20] http://data.lib.vt.edu

Developing Library Cyberinfrastructure Strategy for Big Data Sharing and Reuse
Virginia Tech Libraries
Jan 10, 2016

Schedule of Completion

June 1, 2016 - May 31, 2018

The project will be carried out in 4 phases. Advisory committee will meet every 2 months. Outreach and communication will continue throughout Phase 2 until project completion.

Phase 1 (month 1-3) consists of these activities:

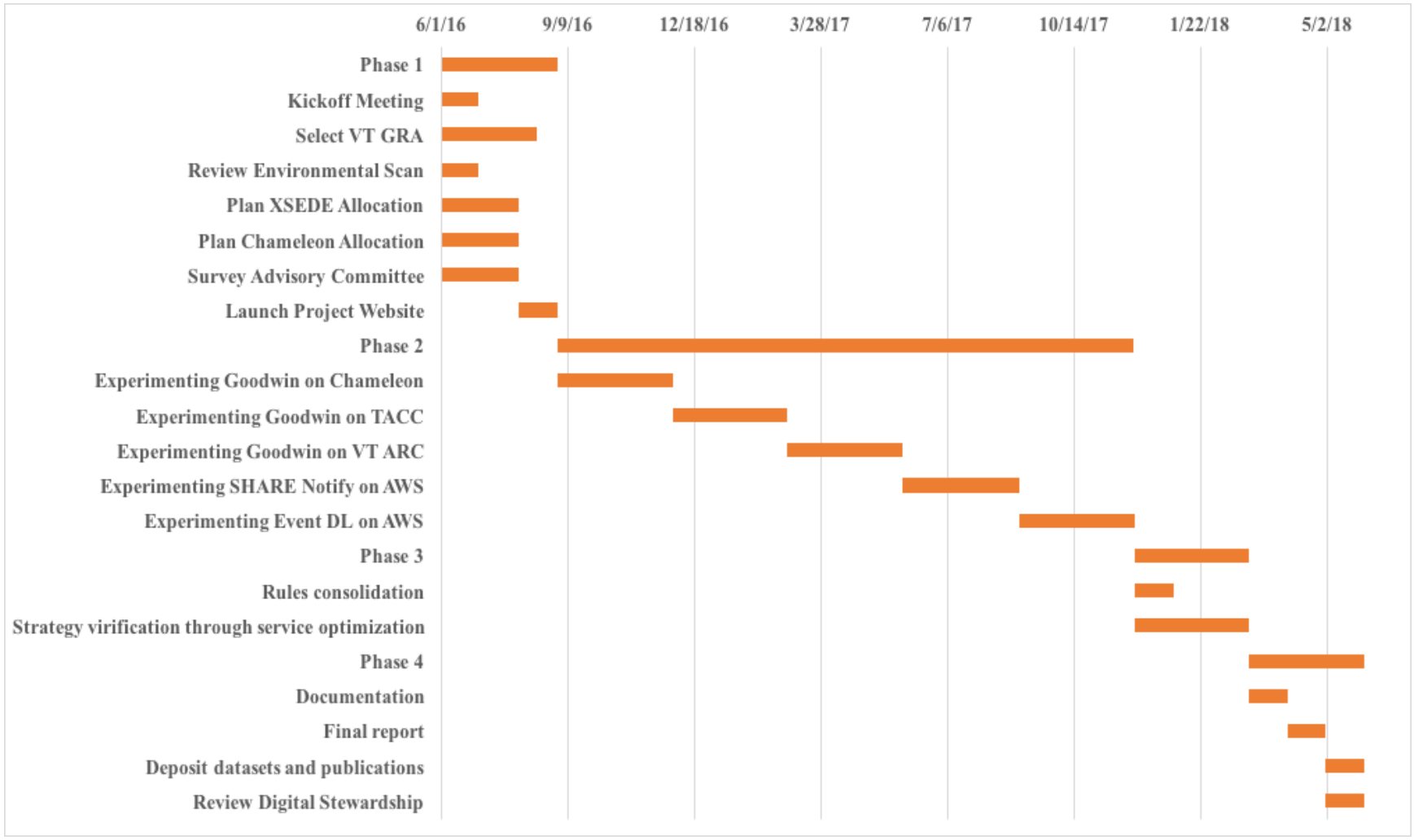
- Dr. Chen travels to VT to attend project kickoff meeting;
- Review environmental scan;
- Survey the advisory committee on key tradeoffs;
- Plan renewing XSEDE allocations;
- Select VT GRA;
- Build project website.

Phase 2 (month 4 - 18) is the period when the bulk of the controlled experiments are to be carried out, results analyzed, experiments adjusted, insights gained and documented, then rules formulated. Five major experimental scenarios that must be investigated are: running Goodwin Hall data management in 1) VT ARC, 2) TACC resources, and 3) Chameleon, respectively; 4) running Event DL and Archive on AWS; 5) running SHARE Notify on AWS. We allocate 3 months for each scenario, schematically shown in the Gantt chart. Based on environmental scan review and advisory committee survey results, we will design and carry out 2 - 4 controlled experiments for each scenario. Each experiment will be carried out iteratively, adhering to the activity workflow depicted in Figure 2 in the Narrative. However, we may interleave experiments to accommodate CI resources allocated schedules, and allow insights and rules be compared and cross-examined across CI choices. We will avoid running experiments in parallel, not only to avoid stretching the project team too thin, but also to leave sufficient time to examine the results and improve experimental design for the later experiments.

Phase 3 (month 19 - 21) focuses on synthesizing the rules verified in Phase 2. We will form the strategy, then carry out 1 to 3 more experiments attempting to optimize the CI choices.

Phase 4 (month 22 - 24) consists of documentation, final report, and data management review.

Schedule of Completion



Supporting Document 2

Bibliography of References

1. "Update on the Twitter Archive at the Library of Congress," Library of Congress, Jan. 2013. [Online]. Available: https://www.loc.gov/today/pr/2013/files/twitter_report_2013jan.pdf. [Accessed: 05-Jan-2016]
2. M. Zimmer, "The Twitter Archive at the Library of Congress: Challenges for information practice and information policy," First Monday, vol. 20, no. 7, Jun. 2015.
3. M. Raymond, "The Library and Twitter: An FAQ | Library of Congress Blog," 28-Apr-2010. [Online]. Available: http://blogs.loc.gov/loc/2010/04/the-library-and-twitter-an-faq/. [Accessed: 05-Jan-2016].
4. "DPN Services, Technical & Administrative Frequently Asked Questions." The Digital Preservation Network, 02-Jun-2015. Available: http://www.dpn.org/wp-content/uploads/2015/06/DPN_FAQ.6.2.15.pdf
5. B. Korner, "DPN: An Overview from a Technical Perspective," presented at the 2013 Preservation and Archiving Special Interest Group (PASIG) Conference, Washington, DC, 22-May-2013.
6. J. Zeng, G. Ruan, A. Crowell, A. Prakash, and B. Plale, "Cloud Computing Data Capsules for Non-consumptiveuse of Texts," in Proceedings of the 5th ACM Workshop on Scientific Cloud Computing, New York, NY, USA, 2014, pp. 9–16.
7. B. Plale, A. Prakash, and R. McDonald, "The Data Capsule for Non-Consumptive Research: Final Report," HathiTrust, Feb. 2015.
8. R. McDonald, "Elephant in the Room: Scaling Storage for the HathiTrust Research Center," presented at the 2015 Preservation and Archiving Special Interest Group (PASIG) Conference, San Diego, CA, 2015.
9. "dpla/heidrun," GitHub. [Online]. Available: https://github.com/dpla/heidrun. [Accessed: 05-Jan-2016].
10. A. Altman, G. Gueguen, and M. Breedlove, "Heiðrún: DPLA's Metadata Harvesting, Mapping and Enhancement System," presented at the 2015 Code4Lib Conference, Portland, OR, 11-Feb-2015.
11. "CenterForOpenScience/SHARE," GitHub. [Online]. Available: https://github.com/CenterForOpenScience/SHARE. [Accessed: 05-Jan-2016].
12. J. Towns, T. Cockerill, M. Dahan, I. Foster, K. Gaither, A. Grimshaw, V. Hazlewood, S. Lathrop, D. Lifka, G. D. Peterson, R. Roskies, J. R. Scott, and N. Wilkens-Diehr, "XSEDE: Accelerating Scientific Discovery," Computing in Science & Engineering, vol. 16, no. 5, pp. 62–74, Sep. 2014.
13. P. N. Edwards, S. J. Jackson, G. C. Bowker, and C. P. Knobel, "Understanding Infrastructure: Dynamics, Tensions, and Design," Working Paper, Jan. 2007. [Online]. Available: http://deepblue.lib.umich.edu/handle/2027.42/49353 [Accessed: 05-Jan-2016].
14. "Cyberinfrastructure for 21st Century Science and Engineering Advanced Computing Infrastructure Vision and Strategic Plan," National Science Foundation, Feb. 2012. [Online]. Available: http://www.nsf.gov/pubs/2012/nsf12051/nsf12051.pdf [Accessed: 05-Jan-2016].

15. G. Henry, "Core infrastructure considerations for large digital libraries," Council on Library and Information Resources, Digital Library Federation, CLIR pub 153, 2012. [Online]. Available: http://www.clir.org/pubs/reports/pub153/pub153.pdf [Accessed: 05-Jan-2016].
16. D. S. Rosenthal, D. C. Rosenthal, E. L. Miller, I. F. Adams, M. W. Storer, and E. Zadok, "The economics of long-term digital storage," Memory of the World in the Digital Age, Vancouver, BC, 2012.
17. D. S. Rosenthal and D. L. Vargas, "Distributed digital preservation in the cloud," International Journal of Digital Curation, vol. 8, no. 1, pp. 107–119, 2013.
18. T. Walters, "Assimilating Digital Repositories Into the Active Research Process," in Ray, J eds. Research Data Management: Practical Strategies for Information Professionals, Purdue University Press, 2014, pp. 189–201.
19. Z. Xie, Y. Chen, J. Speer, T. Walters, P. A. Tarazaga, and M. Kasarda, "Towards Use And Reuse Driven Big Data Management," in Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries, New York, NY, USA, 2015, pp. 65–74.
20. Y. Kargin, M. Kersten, S. Manegold, and H. Pirk, "The DBMS - Your Big Data Sommelier," in 2015 IEEE 31st International Conference on Data Engineering (ICDE), 2015, pp. 1119–1130.
21. Z. Xie, Y. Chen, T. Jiang, J. Speer, T. Walters, P. A. Tarazaga, and M. Kasarda, "On-Demand Big Data Analysis in Digital Repositories: A Lightweight Approach," in Digital Libraries: Providing Quality Information, R. B. Allen, J. Hunter, and M. L. Zeng, Eds. Springer International Publishing, 2015, pp. 274–277.
22. Z. Xie, "Benchmarking Fedora 4 Clustering: Fedora 4 Technical Working Group Assessment," presented at the Open Repositories 2015, Indianapolis, IN, 10-Jun-2015.
23. C. Brittle and Z. Xie, "Big Data Processing in the Cloud: a Hydra/Sufia Experience," presented at the Open Repositories 2014, Helsinki, Finland, Jun. 2014.
24. Z. Xie, J. Liu, H. Van de Sompel, J. van Reenen, and R. Jordan, "Poor Man's Social Network: Consistently Trade Freshness for Scalability," in Proceedings of the 3rd USENIX Conference on Web Application Development (WebApps 12), 2012, pp. 51–62.
25. R. Gruss, M. Farag, T. Kanan, M. C. English, X. Zhang, and E. A. Fox, "Teaching Big Data Through Project-based Learning in Computational Linguistics and Information Retrieval," J. Comput. Sci. Coll., vol. 31, no. 2, pp. 260–270, Dec. 2015.
26. E. A. Fox, M. Farag, S. Lee, X. Zhuang, and R. Gruss, "Conversation: Problem/project-based Learning with Big Data," presented at the 2016 Conference on Higher Education Pedagogy, Blacksburg, VA, 10-Feb-2016.
27. T. Kanan, X. Zhang, M. Magdy, and E. Fox, "Big Data Text Summarization for Events: A Problem Based Learning Course," in Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries, New York, NY, USA, 2015, pp. 87–90.
28. K. G. Akers, F. C. Sferdean, N. H. Nicholls, and J. A. Green, "Building Support for Research Data Management: Biographies of Eight Research Universities," *International Journal of Digital Curation*, vol. 9, no. 2, pp. 171–191, Jun. 2014.
29. J. Chen, and M. Namgoong. "Building a Multilingual Research Platform through Usability Testing." *International Journal of Advanced Computer Science* 4(2): 79-88, 2014.

30. J. Chen, R. Ding, S. Jiang, and R. Knudson. "A Preliminary Evaluation of Metadata Records Machine Translation," *The Electronic Library* 30(2): 264-277. 2012.
31. J. Chen, O. Azogu, W. Zhao, and M. E. Ruiz. "HeMT: A Multilingual System for Human Evaluation of Metadata Records Machine Translation." In *Online Proceedings of ASIST 2011 Annual Conference, New Orleans, LA, October 9-11*. 2011
32. J. Chen, J. "Toward a unified retrieval outcome analysis framework for cross-language information retrieval." Online proceedings of 2005 annual conference of the American Society for Information Science and Technology, Charlotte, North Carolina. October 30- November 1, 2005.
33. D. Minor, B. E. Schottaender, and A. Kozbial, "Chronopolis Repository Services," in Ray, J eds. Research Data Management: Practical Strategies for Information Professionals, Purdue University Press, 2014, pp. 239–252.
34. B. Branan and D. Minor, "Chronopolis and DuraCloud: Doing integration right," presented at the 2015 Preservation and Archiving Special Interest Group (PASIG) Conference, San Diego, CA, 2015.
35. D. Minor, B. Branan, D. Galewsky, and T. Cramer, "Ingest into the Digital Preservation Network: Standard Pipelines," presented at Open Repositories 2014, Helsinki, Finland, 10-Jun-2014.
36. D. Minor, S. Schaefer, and R. Steans, "Integrating DuraCloud with DPN at Chronopolis and the Texas Digital Library," presented at Open Repositories 2015, Indianapolis, IN, 09-Jun-2015.
37. 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, Mar. 2010.
38. "VTUL/VT-SIL-Research," GitHub. [Online]. Available: https://github.com/VTUL/VT-SIL-Research. [Accessed: 07-Jan-2016].
39. J. Schloemann, V. V. N. S. Malladi, A. G. Woolard, J. M. Hamilton, R. M. Buehrer, and P. A. Tarazaga, "Vibration Event Localization in an Instrumented Building," in *Experimental Techniques, Rotating Machinery, and Acoustics, Volume 8*, J. D. Clerck, Ed. Springer International Publishing, 2015, pp. 265–271.
40. J. M. Hamilton, B. S. Joyce, M. E. Kasarda, and P. A. Tarazaga, "Characterization of Human Motion Through Floor Vibration," in *Dynamics of Civil Structures, Volume 4*, F. N. Catbas, Ed. Springer International Publishing, 2014, pp. 163–170.
41. B. A. Jurik, A. A. Blekinge, R. B. Ferneke-Nielsen, and P. Møldrup-Dalum, "Bridging the gap between real world repositories and scalable preservation environments," Int J Digit Libr, vol. 16, no. 3–4, pp. 267–282, May 2015.
42. M. Kimpton and C. Morris, "Managing and archiving research data: local repository and cloud-based practices," in Ray, J eds. Research Data Management: Practical Strategies for Information Professionals, Purdue University Press, 2014, pp. 223–238.
43. A. Jackson, "Large-Scale Web Archive Discovery and Analytics Using Apache Solr," presented at the 2014 International Internet Preservation Consortium (IIPC) General Assembly, Paris, France, 20-May-2014.
44. "ukwa/webarchive-discovery," GitHub. [Online]. Available: https://github.com/ukwa/webarchive-discovery. [Accessed: 05-Jan-2016].

45. J. Lin, M. Gholami, and J. Rao, “Infrastructure for Supporting Exploration and Discovery in Web Archives,” in Proceedings of the 23rd International Conference on World Wide Web, Republic and Canton of Geneva, Switzerland, 2014, pp. 851–856.
46. “lintool/warcbase,” GitHub. [Online]. Available: https://github.com/lintool/warcbase. [Accessed: 05-Jan-2016].
47. J. Lin, “The Sum of All Human Knowledge in Your Pocket: Full-Text Searchable Wikipedia on a Raspberry Pi,” in Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries, New York, NY, USA, 2015, pp. 85–86.
48. R. Arora, M. Esteva, and J. Trelogan, “Leveraging High Performance Computing for Managing Large and Evolving Data Collections,” International Journal of Digital Curation, vol. 9, no. 2, pp. 17–27, Oct. 2014.

DIGITAL STEWARDSHIP SUPPLEMENTARY INFORMATION FORM

Introduction

The Institute of Museum and Library Services (IMLS) is committed to expanding public access to federally funded research, data, software, and other digital products. The assets you create with IMLS funding require careful stewardship to protect and enhance their value, and they should be freely and readily available for use and re-use by libraries, archives, museums, and the public. However, applying these principles to the development and management of digital products is not always straightforward. Because technology is dynamic and because we do not want to inhibit innovation, we do not want to prescribe set standards and best practices that could become quickly outdated. Instead, we ask that you answer a series of questions that address specific aspects of creating and managing digital assets. Your answers will be used by IMLS staff and by expert peer reviewers to evaluate your application, and they will be important in determining whether your project will be funded.

Instructions

If you propose to create any type of digital product as part of your project, complete this form. We define digital products very broadly. If you are developing anything through the use of information technology (e.g., digital collections, web resources, metadata, software, or data), you should complete this form.

Please indicate which of the following digital products you will create or collect during your project
(Check all that apply):

	Every proposal creating a digital product should complete	Part I
	If your project will create or collect	**Then you should complete**
☐	Digital content	Part II
☐	Software (systems, tools, apps, etc.)	Part III
☑	Dataset	Part IV

PART I.

A. Intellectual Property Rights and Permissions

We expect applicants to make federally funded work products widely available and usable through strategies such as publishing in open-access journals, depositing works in institutional or discipline-based repositories, and using non-restrictive licenses such as a Creative Commons license.

A.1 What will be the intellectual property status of the content, software, or datasets you intend to create? Who will hold the copyright? Will you assign a Creative Commons license (http://us.creativecommons.org) to the content? If so, which license will it be? If it is software, what open source license will you use (e.g., BSD, GNU, MIT)? Explain and justify your licensing selections.

All benchmarking results from this research will be released under the Open Data Commons Open Database Licence (ODC-ODbL) v1.0. According to DCC (http://www.dcc.ac.uk/resources/how-guides/license-research-data#x1-8000doc)

"Being written in database terms, these licences are suited to a wider range of research data than the Creative Commons equivalents. The ODC-ODbL copyleft condition is also slightly more flexible than Creative Commons' Share Alike, though the ODC attribution requirement is slightly less flexible. - See more at: http://www.dcc.ac.uk/resources/how-guides/license-research-data#x1-8000doc"

A.2 What ownership rights will your organization assert over the new digital content, software, or datasets and what conditions will you impose on access and use? Explain any terms of access and conditions of use, why they are justifiable, and how you will notify potential users about relevant terms or conditions.

We will not be producing any new digital content, only datasets. We will disseminate the results of our research in academic conferences and professional meetings. If allowed by these publications' copyright, we will deposit a copy of these publications in VTechWorks, Virginia Tech's institution repository.

A.3 Will you create any content or products which may involve privacy concerns, require obtaining permissions or rights, or raise any cultural sensitivities? If so, please describe the issues and how you plan to address them.

N/A

Part II: Projects Creating or Collecting Digital Content

A. Creating New Digital Content

A.1 Describe the digital content you will create and/or collect, the quantities of each type, and format you will use.

A.2 List the equipment, software, and supplies that you will use to create the content or the name of the service provider who will perform the work.

A.3 List all the digital file formats (e.g., XML, TIFF, MPEG) you plan to create, along with the relevant information on the appropriate quality standards (e.g., resolution, sampling rate, or pixel dimensions).

B. Digital Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan (i.e., how you will monitor and evaluate your workflow and products).

B.2 Describe your plan for preserving and maintaining digital assets during and after the award period of performance (e.g., storage systems, shared repositories, technical documentation, migration planning, commitment of organizational funding for these purposes). Please note: You may charge the Federal award before closeout for the costs of publication or sharing of research results if the costs are not incurred during the period of performance of the Federal award. (See 2 CFR 200.461).

C. Metadata

C.1 Describe how you will produce metadata (e.g., technical, descriptive, administrative, or preservation). Specify which standards you will use for the metadata structure (e.g., MARC, Dublin Core, Encoded Archival Description, PBCore, or PREMIS) and metadata content (e.g., thesauri).

C.2 Explain your strategy for preserving and maintaining metadata created and/or collected during and after the award period of performance.

C.3 Explain what metadata sharing and/or other strategies you will use to facilitate widespread discovery and use of digital content created during your project (e.g., an API (Application Programming Interface), contributions to the Digital Public Library of America (DPLA) or other digital platform, or other support to allow batch queries and retrieval of metadata).

D. Access and Use

D.1 Describe how you will make the digital content available to the public. Include details such as the delivery strategy (e.g., openly available online, available to specified audiences) and underlying hardware/software platforms and infrastructure (e.g., specific digital repository software or leased services, accessibility via standard web browsers, requirements for special software tools in order to use the content).

D.2 Provide the name and URL(s) (Uniform Resource Locator) for any examples of previous digital collections or content your organization has created.

Part III. Projects Creating Software (systems, tools, apps, etc.)

A. General Information

A.1 Describe the software you intend to create, including a summary of the major functions it will perform and the intended primary audience(s) this software will serve.

A.2 List other existing software that wholly or partially perform the same functions, and explain how the tool or system you will create is different.

B. Technical Information

B.1 List the programming languages, platforms, software, or other applications you will use to create your software (systems, tools, apps, etc.) and explain why you chose them.

B.2 Describe how the intended software will extend or interoperate with other existing software.

B.3 Describe any underlying additional software or system dependencies necessary to run the new software you will create.

B.4 Describe the processes you will use for development documentation and for maintaining and updating technical documentation for users of the software.

B.5 Provide the name and URL(s) for examples of any previous software tools or systems your organization has created.

C. Access and Use

C.1 We expect applicants seeking federal funds for software to develop and release these products under an open-source license to maximize access and promote reuse. What ownership rights will your organization assert over the software created, and what conditions will you impose on the access and use of this product? Identify and explain the license under which you will release source code for the software you develop (e.g., BSD, GNU, or MIT software licenses). Explain any prohibitive terms or conditions of use or access, explain why these terms or conditions are justifiable, and explain how you will notify potential users of the software or system.

C.2 Describe how you will make the software and source code available to the public and/or its intended users.

C.3 Identify where you will be publicly depositing source code for the software developed:

Name of publicly accessible source code repository:
URL:

Part IV. Projects Creating a Dataset

1. Summarize the intended purpose of this data, the type of data to be collected or generated, the method for collection or generation, the approximate dates or frequency when the data will be generated or collected, and the intended use of the data collected.

The intended purpose of this data to formulate cyberinfrastructure strategies. The data type is computer systems performance data, collected by running benchmarking software against a defined workload. The dates and frequency of the data collection will be documented as metadata.

2. Does the proposed data collection or research activity require approval by any internal review panel or institutional review board (IRB)? If so, has the proposed research activity been approved? If not, what is your plan for securing approval?

No

3. Will you collect any personally identifiable information (PII), confidential information (e.g., trade secrets), or proprietary information? If so, detail the specific steps you will take to protect such information while you prepare the data files for public release (e.g., data anonymization, data suppression PII, or synthetic data).

No

4. If you will collect additional documentation such as consent agreements along with the data, describe plans for preserving the documentation and ensuring that its relationship to the collected data is maintained.

n/a

5. What will you use to collect or generate the data? Provide details about any technical requirements or dependencies that would be necessary for understanding, retrieving, displaying, or processing the dataset(s).

Will use open source benchmarking software, e.g., httperf, jmeter, YCSB, etc.

6. What documentation (e.g., data documentation, codebooks, etc.) will you capture or create along with the dataset(s)? Where will the documentation be stored, and in what format(s)? How will you permanently associate and manage the documentation with the dataset(s) it describes?

We will document the data collection process and method along with the data as item descriptions, stored in VTechData, Virginia Tech's institution data repository.

7. What is the plan for archiving, managing, and disseminating data after the completion of the award-funded project?

Datasets and related publications will be archived, respectively, in VTechData and VTechWorks, Virginia Tech's institution data repository and document repository, and managed by Virginia Tech Libraries for long-term access.

8. Identify where you will be publicly depositing dataset(s): Vtechdata

Name of repository: VTechData
URL: http://data.lib.vt.edu

9. When and how frequently will you review this data management plan? How will the implementation be monitored?

This data management plan will be reviewed at the beginning and end of this project, and annually thereafter for 5 years by the repository manager.

Original Preliminary Proposal

Developing Library Cyberinfrastructure Strategy for Big Data Sharing and Reuse

Virginia Tech Libraries, in partnership with the departments of Mechanical Engineering and Computer Science, as well as the University of North Texas Department of Library and Information Sciences, propose a 2-year research grant toward a broadly adaptable library cyberinfrastructure (CI) strategy for big data sharing and reuse. The strategy will be based on intelligently matching and synthesizing five types of existing CI options against key requirements characterizing representative types of library big data services. The strategy will be validated using experimental deployments of three such services.

Statement of Need

A download link is no longer sufficient to provide effective access to terabytes of research data. Libraries are increasingly expected to deliver not the raw data but the knowledge extracted from them, e.g., running user specified algorithms against preserved data, pushing customized information from a metadata hub, or analyzing web archives. Doing so, however, requires extensive storage, processing, and associated network bandwidth. While building local, proprietary capacities at each library would be prohibitively expensive, the alternatives are far from turnkey. Further, through prior research supported by IMLS, NSF, NIH, US Air Force, Sloan, Mellon, Amazon, IBM, and VTI Instruments, we have identified other common and significant CI gaps affecting both institutional, national, and commercial CI choices. Although spontaneous and isolated efforts exist to address some of these problems, there is little systematic evaluation, comparison, and synthesis of solutions. Consequently, there is no unified CI framework or strategy that could be referred to for different library big data sharing and reuse situations.

Proposed Work Plan

This project focuses on strategies for deploying library services that leverage the five currently existing CI options and/or their combinations: 1) commercial cloud, such as Amazon or IBM cloud; 2) local infrastructure, such as Virginia Tech's Digital Library Research Lab (DLRL) Hadoop cluster; 3) institutional infrastructure, such as HPC clusters and storage at Virginia Tech Advanced Research Computing; 4) national HPC, HTC, and storage resources such as TACC Ranch long-term storage, Stampede, or Wrangler; 5) NSF research clouds such as TACC Chameleon and National Data Service Labs. We will analyze, benchmark, and compare the strength and weakness of each option's service model, performance characteristics, cost, and personnel and skill constraints.

The above research is conducted against three representative types of library big data services: 1) STEM big data reuse, exemplified by the data reuse testbed built for the Goodwin Hall Living Laboratory, the world's most instrumented building for vibration, 2) large-scale metadata aggregation and information routing, exemplified by the SHARE Notify Service, 3) large-scale integrated tweet and webpage collection, archiving, and digital library (DL) service provision, exemplified by the IDEAL event archive/DL. We choose these services because they represent distinctively varied library big data workloads. Case 1 requires balancing data movements with processing; case 2 continuously ingests, harvests, indexes, and queries large amount of metadata; and case 3 tends to lock in data and discourages movement, e.g., by incurring high initial data loading costs. In all three cases, software development has completed or is close to complete, and focus has been shifted to service deployment.

After acquiring in-depth knowledge about the CI options through active deployments and experimentation, we will generalize the insights gained and extrapolate them to widely applicable rules, based on which a CI strategy will be formed to guide the mix-and-match of these CI options for future services. This strategy will be evaluated and further validated by porting all or part of these three services from their existing infrastructure to potentially more advantageous alternatives.

Relevance to National Digital Platform and Potential Impacts

This project empowers libraries by more effectively leveraging shared CI, which can make big data more accessible and reusable. It furthers the thrust of library data management activities and pushes the library beyond its traditional focus on a large variety of small data sets, and their storage and preservation. By focusing on efficient and cost-effective library service deployment this project also helps to bridge the operational gap that divides digital library research vs. services.

This project aggressively collaborates across domains and areas of expertise. Planting skilled librarians in representative active big data research and development projects, helps to integrate library CI and services more tightly with STEM research and education.

Further, resulting improved big data services such as the Goodwin Hall Living Lab testbed will be used extensively for STEM undergraduate and graduate big data education programs. The project also exposes current and future librarians to state-of-the-art CI and reveals new skill sets required by emerging library services. This informs the curriculum enhancement currently underway at UNT MLS program, and allows extension of the NSF funded digital library learning modules now available through Wikiversity, lead by VT Computer Science and the School of Information and Library Science at UNC Chapel Hill.

Performance Goals and Outcomes

This project primarily addresses the IMLS strategic goal (c), content, by optimizing big data CI, therefore broadens access and expands use of big data sets. The outcomes of the project include: 1) Systematic benchmarking and comparison of five available CI options with respect to library big data sharing and reuse cases; 2) A replicable CI strategy to mix and match these options to maximize library capacity; 3) As a result of experimenting and service porting, improved system usability, economy, and performance of these services. 4) In terms of content, improved access to the Goodwin Hall sensor data (60TB per year, ongoing), the SHARE Notify data (2.1 million events as of this writing, less than 6 months in beta release), and the IDEAL event archive (1 billion tweets and about 15 TB of archived web pages).

Project Director and Partners

PI Zhiwu Xie has extensive research and development experience in mechanical engineering, high-performance computing, big data, and digital libraries. His research extensively utilizes all five CI options. He is also deeply involved in all three listed big data projects, and is well-situated to connect the project team. Xie leads the service deployment and benchmarking. Co-PI Tyler Walters brings the leadership and SHARE project support to the team. Co-PI Edward A Fox contributes decades of digital library research and education experience, the NSF-funded IDEAL event archive, and the DLRL Hadoop Cluster. Co-PI Pablo Tarazaga founded the Goodwin Hall Living Lab and brings to this project its large sensor data sets, data analysis use cases, and the user community. Co-PI Jiangping Chen leads the design and implementation of the evaluation plan, which includes extensive literature review, survey, automated log analysis, and usability test.

Budget

We request from IMLS a total budget of $312,129 over the 2-year period without cost sharing. This includes $210,625 direct cost and $101,504 indirect cost, calculated at Virginia Tech's negotiated rate. The direct cost is further broken down into $160,978 for the VT portion and $49,647 UNT subcontract.