Improving Fedora to Work with Web-scale Storage and Services

The Digital Curation Innovation Center (DCIC) at the University of Maryland's College of Information Studies (iSchool) requests $240,138 to research ways to improve the performance and scalability of the Fedora Repository for the Fedora user community. The work will be carried out over the 2-year period (October 1, 2017 to September 30, 2019) without cost sharing. This includes $157,986 in direct costs and $82,153 in indirect costs, calculated at UMD's negotiated rate. The direct cost is further broken down into $120,118 for salaries and wages, $27,868 for fringe benefits, $10,000 for travel.

1. Statement of National Need

This project addresses high priority challenges that face the Fedora user community. *"Fedora has a worldwide installed user base that includes academic and cultural heritage organizations, universities, research institutions, university libraries, national libraries, and government agencies."*[1] The National Digital Platform is interested in expanding the scalability and interoperability of major open source software used by libraries and archives. Our project aims to address current performance limits identified in Fedora 4.x through research and development of new ways to perform its operations to manage large-scale collections. For decades Fedora has consistently provided an object storage framework for repositories of content and data. Fedora developers and users have created hundreds of applications, a strong community, and a rich software ecosystem of shareable code. The Fedora 4 project, now in its seventh software release, worked to redefine Fedora's conventions, the application programming interface (API) and the data model, in the age of linked data.

Looking beyond Fedora 4, the Fedora community has placed significant effort towards formalizing and specifying the Fedora application programming interface[2], designated as the Fedora 5 API. This was done to clearly define the core services expected of a Fedora repository, to stabilize the technical contract for Fedora clients, and to enable new implementations of Fedora to enter the ecosystem. The Fedora 5 API, there is no implementation, is largely based on two open standards, the W3C Linked Data Platform[3] and IETF Memento[4]. The *DRAS-TIC Fedora* team will work with the Fedora community to develop an implementation of the Fedora 5 API to address a shortcoming in the existing Fedora 4 implementation. More details on DRAS-TIC are provided in Section 2, Project Design. We discussed this project with members of the Fedora Steering Group in the fall of 2016 and held a working meeting at the Dec. 2016 CNI conference to further refine the research and development ideas. The proposal has received the very important endorsement of the Chair of Fedora Leadership Group and Steering Committee (See: http://fedorarepository.org/leadership-group and http://fedorarepository.org/steering-group). We are including this letter of support. Since then we have reached out to key universities and partners on the Fedora Development Team and cultural institutions interested in management at scale of library collections. Initial community members include 5 partners: UMD Libraries,

1 DURASPACE/Fedora home: https://wiki.duraspace.org/display/FF/Fedora+Repository+Home
2 Fedora 5 API Specification: http://fedora.info/spec/
3 Linked Data Platform 1.0 Specification: https://www.w3.org/TR/ldp/
4 Memento Specification: https://tools.ietf.org/html/rfc7089

Smithsonian Institution (Office of Research Info. Services and National Museum of American History), UIUC National Center for Supercomputing Applications (NCSA), and Georgetown U. Library. All have expressed interest and a need and we expect to reach out to these partners during the course of the project for further feedback, use cases, and guidance.

Pursuing Fedora at Scale

The storage demands of Fedora users for objects and metadata are extreme when compared to typical content management applications. Many Fedora institutions store large data files in their repositories, such as disk images, unlike the typical media objects used in publishing. They may also store millions of objects with extensive metadata attached to each. It is for this reason that the user community has an innovative history of pursuing repository scalability. The Fedora 3.x design included a modular storage layer, which allowed for the development of custom, modern storage solutions by different institutions and coalitions. These included community development of the Akubra storage layer as well as the development of an iRODS-based Fedora storage module at UNC Chapel Hill. Jansen was principal architect of this iRODS storage module and Marciano was a project advisor and iRODS researcher. In addition UCSD at that time addressed their scalability needs by implementing the Fedora 3.x API over a custom storage system of their own design. Fedora institutions like these have been major proponents of the Fedora design, applications, and community, but they have had to create custom solutions to address repository scale.

While such scaling needs are becoming more typical, the Fedora 4 software implementation was still limited by an underlying software component, ModeShape, that was designed for more typical content management and could not meet all of these demands. Importantly, Fedora 4's storage layer is limited to large storage pools for both metadata and files, a performance bottleneck. The large storage pools also limit the total storage capacity of any repository. Institutions adjust their budgets to purchase expensive high-capacity storage arrays, so that their single storage pool can grow as large as possible. Fedora 4's storage pool is a single point of failure that impacts reliability and remains the limiting factor on repository growth for the Fedora community.

Since the first release of Fedora 4, the technical team has made many strides towards better performance at scale, including a major upgrade that included an improved ModeShape 5.1 storage system. There is also a Performance and Scale Working Group that has been measuring and attempting to address Fedora performance issues across six key metrics[5]. They have performed extensive testing of various storage systems and configurations of Fedora 4.x. They have produced metrics and analysis[6] for every available storage arrangement for Fedora 4. This testing has identified three main areas of concern:

- Single server or storage pool error conditions arise after a few million storage operations, such as "timeout", "too many open files", "out of memory", and barring these others, "disk full".

[5] Fedora Performance and Scalability Test Plans: https://wiki.duraspace.org/display/FF/Performance+and+Scalability+Test+Plans#PerformanceandScalabilityTestPlans-Test1-Sizeoffiles-large

[6] These metrics and analysis are available in separate folders by configuration in their Git repository: https://github.com/fcrepo4-labs/fcrepo_perf_analysis/tree/master/dist

- Performance degrades when the underlying ModeShape folders grow larger than 1000 items. This is described as the many members problem[7], and is an instance of the "supernode" problem, discussed later in research questions.
- Performance degrades when many Fedora clients are performing simultaneous tasks, due to transaction checking overhead.

While performance of Fedora 4 has been improved, these remain the most intractable issues.

Web-Scale Technologies from Industry

The Fedora user community's needs for storage are unique among content management applications, but not unique among web-scale companies. Storing millions of objects and their metadata is not unlike storing the data and preferences for millions of users. Large companies like NetFlix, Facebook, and Google have consistently succeeded in meeting this challenge by adopting a different strategy for storage. The Fedora 4 strategy for growth is one where you plan, budget, build, and finally migrate into a single storage system that is large enough to meet expected demand. In contrast, the web-scale companies meet demand by adding more, separate storage systems to a group of similar systems, a cluster, that work together to store the data. Each system in the cluster is relatively small and cheap, so you no longer have to plan and build cost-prohibitive enterprise storage arrays. Storage cluster technologies allow for growth while maintaining a smooth, predictable cost for increases in capacity. The performance of a storage cluster is not limited by a single bottleneck system.

One of the most widely adopted and successful industry solutions for distributed data storage is called Apache Cassandra[8] and it is a free and open source project. Apache Cassandra is the most popular distributed database available and it also ranks 8th among databases overall[9]. It is part of the family of so-called NoSQL databases and is second in use and popularity to MongoDB and is in use with CERN, eBay, GitHUB, Hulu, Instagram, Netflix, Twitter, and at 1500 other companies, and scales to petabytes of storage and billions of objects. An Apache Cassandra cluster can scale up to 75,000 separate servers (perhaps the largest so far, at Apple, Inc.) and beyond. This is the technology that DRAS-TIC uses for storing files and metadata and that we think will prove a good fit for Fedora community needs. A Fedora institution might begin with a cluster of just two or three servers, and then proceed to grow this cluster incrementally as their collections grow.

Avoiding "Big Bang" Storage Planning

Many Fedora institutions have trouble predicting costs and planning storage capacity for large new collections. When the current enterprise storage system is nearly full, adding new collections may be postponed for years, until the next, higher capacity, storage system can be designed, budgeted, and built. These long planning cycles make it difficult to estimate costs for collections, which stifles local collection efforts and innovation. In contrast, using a cluster approach allows one to simply count the number of new, separate servers required to

[7] See Fedora's wiki page: https://wiki.duraspace.org/display/FF/Many+Members+Performance+Testing
[8] Details about Apache Cassandra at: https://cassandra.apache.org/
[9] See relational vs. NoSQL database rankings at: https://db-engines.com/en/ranking

store the new collection. These marginal costs of new collections are simple to calculate, and servers may be purchased and added to the existing cluster as needed. The current Fedora storage strategy can have a serious impact on collection development, budgeting, and planning long before the total size of collections have an impact on performance.

2. Project Design

The University of Maryland iSchool's Digital Curation Innovation Center, DCIC, is in a unique position to deliver a cluster-based storage strategy for the Fedora community. We have been involved in developing a cluster-based archival repository software, called DRAS-TIC [10] (Digital Repository At Scale - That Invites Computation). This approach was designed from the ground up to leverage the Apache Cassandra cluster-based storage system and provide hierarchical storage of billions of digital objects and their metadata. DRAS-TIC currently implements an open-source industry standard API called the Cloud Data Management Interface[11] (CDMI) that defines the way in which applications can create, retrieve, update, and delete datasets, resources, and metadata. It has demonstrated dramatic improvements in performance and capacity without the liabilities of expensive storage, enterprise-style storage planning, and data loss that have plagued non-cluster storage strategies. We currently manage 100 Million files with metadata and 100 Terabytes of storage in the cluster at University of Maryland. This project will work closely with the Fedora community and Steering Group to further develop DRAS-TIC into an alternate implementation of the Fedora API. This means adding the Fedora API to the existing CDMI API, or replacing the CDMI API. The project will addresses an important shortcoming in the existing Fedora implementation, towards meeting the most demanding and extreme storage requirements.

Research Questions

- **RQ1:** *Is it possible and practical to maintain both the CDMI API and the Fedora API in one software platform?*
    - Can the CDMI API be used to edit basic Fedora metadata without data loss?
- **RQ2:** *How best to use the graph database available in* DRAS-TIC *to persist and maintain linked data of the form supplied through the Fedora API?*
    - Generalize approaches to storing RDF in a graph database
    - Develop asynchronous approaches to maintaining distributed linked data, such as a "garbage collection" strategy.
- **RQ3:** *Does* DRAS-TIC*'s distributed storage offer newly distributed or decentralized modes of building software with Fedora?*
    - New software architectures: We may further advance the notion, popularized by the Hydra community, of many applications sharing a common, multi-tenanted repository.
    - Practical backup strategies for distributed data

[10] DRAS-TIC scalable repository project: https://umd-drastic.github.io/about/
[11] Opensource CDMI API Industry standard for cloud storage: https://www.snia.org/cdmi

- **RQ4:** *Does* DRAS-TIC*'s distributed storage afford new forms of shared storage and federation across institutions?*
    - Reciprocal storage mirroring across datacenters
    - Loading dock clusters
- **RQ5:** *How can we address the problem of super-nodes?* "Super-nodes" are digital objects with many more connections than other objects, such as a folder that holds a million content items? This is a common Fedora pattern that creates multiple issues for the current implementation.

Research Methods

Given the inherent complexity of distributed storage solutions, we don't feel that cluster-based approaches can be recommended to institutions without strong guidelines on cluster configuration and operations that have been proven in full scale tests. While this project proposes a technical product as one outcome, it is primarily designed to maximally explore the software implementation and cluster configuration possibilities through rapid prototyping and iteration over potential *DRAS-TIC Fedora* architectures.

There are many ways to integrate the components we have identified and indeed myriad ways to configure and use a Cassandra cluster, all of which impact performance, consistency, and reliability. Our research will include the design of graph database schemas, Cassandra table schemas, cluster tunings, consistency rules, and replication rules. We will also explore multi-datacenter replication as a way to provide of geographic and datacenter redundancy. Each candidate configuration of the system will be subjected to uniform testing and measurement, including the running of test suites from the Fedora project itself. Guided by the measures we produce through tests, we will refine and tune the cluster configuration to meet the use cases and needs identified by our partners.

The tests that are run against candidate clusters will produce metrics that will be archived as part of the project, similar to the Fedora metrics produced for Fedora 4.x by the performance and scalability group. Each of the tested configurations will remain tagged and referenceable in a public git repository, alongside complete server deployment scripts that may be used to reproduce the cluster configuration at the time of the test. These reference points will be citable in discussions with partners and can provide configuration "fork points" any institution that find a metric suitable for their particular use case.

With a consistent method for producing and archiving metrics across potential implementations, we will be freed to explore research questions and optimize for different partner use cases, trying new things without losing site of the core need of performance and scalability. The *DRAS-TIC Fedora* metrics will be compared with the Fedora 4.x metrics to determine the recommended architecture choices and trade-offs for performance at various scales.

Project Leadership

- **Richard Marciano**, professor, will act as project director. He is the founder and director of the Digital Curation Innovation Center (DCIC), and director of the Sustainable Archives and Leveraging Technologies (SALT) lab. Marciano has been the Principal Investigator for grants from IMLS, NSF, the Library of Congress, and the National Archives and Records Administration. He is currently the UMD Principal Investigator on a joint NSF/DIBBs (UIUC/NCSA & UMD) software cyberinfrastructure grant called "Brown Dog"[12], the Principal Investigator on an IMLS digital curation professional training and virtual infrastructure grant called "CurateCloud", and a collaborator on a large NSF/IIS grant on pervasive ethics for computational research (PI: Katie Shilton at the UMD iSchool) where the DRAS-TIC repository will be used as the main repository. In 2009, Marciano was the Principal Investigator on an IMLS NLG Research grant with DuraSpace called Policy-Driven Repository Interoperability (PoDRI). The principal focus of PoDRI was to investigate the feasibility of interoperability mechanisms between repositories at the policy level. At the time the research focused on the integration of an object model and a policy-aware distributed data model with Fedora and iRODS as representative open source software for each model. With recent Fedora 4 API developments, there are now renewed opportunities for making a national impact.

- **Gregory Jansen** is the principal research software architect at the Digital Curation Innovation Center at the University of Maryland, where he is developing the DRAS-TIC repository software. He is a contributor to the NCSA Brown Dog project to develop digital curation web services, including the design of scalability and performance testing of the Brown Dog framework. He has been a contributor to the Fedora 4 Technical Team, and the creator of the Curator's Workbench archival packaging software. His past work includes being a technical advisor to the Academic Preservation Trust and the technical lead on the Carolina Digital Repository at UNC Chapel Hill.

Project Support

- **The Digital Curation Innovation Center (DCIC)** at the UMD iSchool will support major project activities. Its mission is to "Integrate Education and Research" in the field of Digital Curation using Big Records and Archival Analytics. It comprises 5 labs [see: http://dcic.umd.edu/infrastructure/] including the dataCave, a peta-scale storage facility powered by NetApp, providing commercial-grade storage to support the DRAS-TIC repository. DRAS-TIC was developed in collaboration with the DCIC and Archives Analytics Solutions Ltd., a software development firm in the United Kingdom, as the result of the $10.5 million NSF "Brown Dog" grant. The DCIC launched the DRAS-TIC repository at iPRES 2016 in Oct. 2016 after the University of Maryland concluded a significant license agreement that established its clear ownership and use.

[12] NSF/DIBBs Brown Dog project: http://browndog.ncsa.illinois.edu/

- **Student Involvement:** The DCIC promotes both projects (focused on justice, human rights & cultural heritage) and cyberinfrastructure for the curation & management of digital assets at scale. Its approach has been to develop interdisciplinary student research teams to help develop new digital skills, conduct interdisciplinary research, and explore professional development opportunities at the intersection of archives, big data, and analytics through focused project themes which include: *Refugee Narratives, Community Displacement, Racial Zoning, Cyberinfrastructure for Digital Curation, Movement of People, Citizen Internment, and others.* We will create a student-centered team around this project for the two-years of the project and promote the development of repository testing and prototyping, and incorporate lessons learned in related classes.

Community Partners

In addition to the endorsement of the Chair of Fedora Leadership Group and Steering Committee, we have reached out to a number of institutions in research libraries, cultural institutions, and research organizations. We will recruit new use cases from these partners for the *DRAS-TIC Fedora* software. These institutions will provide diverse requirements, including repositories that have very large data files and repositories that store and access large amounts of linked data. We will provide assistance to these and other Fedora institutions so they can try *DRAS-TIC Fedora* for their collections and provide feedback. Initial partners include:

- *University of Maryland Libraries:*
  - Digital Systems and Stewardship
  - **Interest:** Managing Fedora in production and managing research data through the API specification.
- *Smithsonian Institution:*
  - Office of Research Info. Services (ORIS) within the Office of the CIO (OCIO)
    and
  - National Museum of American History
  - **Interest:** Curation of digital outputs using the SIdora Information Architecture which is based on Fedora (looking at genomics, hi-res multispectral 3D imaging, archaeology, museum, curation, etc.).
- *U. of Illinois at Urbana-Champaign:*
  - National Center for Supercomputing Applications (NCSA)
  - **Interest:** Establish Research Data Services (RDS) to preserve research data associated with research publications, by incorporating Brown Dog into repository technologies.
- *Georgetown University Library:*
  - Library Information Technology (LIT)
  - **Interest:** Currently using DSpace API. Interest in exploring repository scalability

Project Activities

The project will be roughly divided into three distinct phases for the purpose of start-up, exploration, and engagement respectively. Activities in earlier phases may continue in later phases as we iterate through architectures, testing apparatus, and partner use cases.

- **Phase One: Start Up (6 months)**

1. Create project structure and collaboration tools. Establish the agile software development process and related roles. Recruit student involvement.
2. Formalize communication with community partners and other stakeholders.
    - Continue to work with Fedora developers and Fedora Leadership Group and Steering Committee to ensure that this project aligns with Fedora conventions and the needs outlined above.
3. Develop the Fedora community partner relationships to identify the range of use cases and potential adopters for DRAS-TIC's capabilities.
   Explore research questions (3) and (4) with Fedora stakeholders:
    - **RQ3:** *Does* DRAS-TIC*'s distributed storage offer newly distributed or decentralized modes of building software with Fedora?*
    - **RQ4:** *Does* DRAS-TIC*'s distributed storage afford new forms of shared storage and federation across institutions?*
4. Create an easily installed *DRAS-TIC Fedora* Vagrant machine for use by all developers and to facilitate community evaluation.
5. Create a reconfigurable *DRAS-TIC Fedora* cluster for use as a performance testbed.
6. Develop server deployment and management scripts for *DRAS-TIC Fedora*, further addressing **RQ3** with regards to backup strategies.

- **Phase Two: Explore (one year)**

7. Extend and enhance Fedora's API and performance test suites to provide any missing metrics established by partner uses cases.
8. Create a first prototype of the Fedora API running on DRAS-TIC (*DRAS-TIC Fedora*) by implementing so-called CRUD operations (create, read, update and delete). *Address research question (1) and begin to evaluate (2).*
    - **RQ1:** *Is it possible and practical to maintain both the CDMI API and the Fedora API in one software platform?*
    - **RQ2:** *How best to use the graph database available in* DRAS-TIC *to persist and maintain linked data of the form supplied through the Fedora API?*
9. Setup a Continuous Integration workflow for iterative API and performance testing. This means that new *DRAS-TIC Fedora* code will automatically deploy to the testbed cluster and then trigger a fresh round of API and performance tests, capturing the results. With this test harness established, we can feel free to try new cluster and code strategies, knowing the performance will be uniformly captured.
10. ITERATE: Design, implement, and test in an iterative fashion to address partner needs (capturing outcomes through the continuous integration workflow):
    - Implement additional operations from the Fedora API as needed, suitable for a full repository and consistent with partner use cases. *Answer research questions (4) and (5):*
        - **RQ4:** *Does* DRAS-TIC*'s distributed storage afford new forms of shared storage and federation across institutions?*
        - **RQ5:** *How can we address the problem of super-nodes?*
    - Attempt software strategies made possible by DRAS-TIC's distributed storage, *in pursuit of research question (3)*:
        - **RQ3:** *Does* DRAS-TIC*'s distributed storage offer newly distributed or decentralized modes of*

building software with Fedora?

- Linked data grooming or "garbage collection" strategies that provide "eventual consistency" in the background. (Data visited and assessed outside of the moment of the API transaction.)
- Alternative persistent strategies based on the DataStax graph database or JanusGraph graph database.
- Alternative graph database schema for RDF data.

- **Phase Three: Outreach (6 months)**

11. Create an annotated Fedora API document. Clearly state which parts of the full Fedora API were implemented and why or why not. Identify any limitations to the standard Fedora API.
12. Develop a how-to guide for running a *DRAS-TIC Fedora* repository.
13. Publicize *DRAS-TIC Fedora* performance results with help from the Fedora Leaders Group. Make all metrics available to the community, alongside the code strategies and configurations that produced them.
14. Reach out to the broader Fedora community for adoption and support and present at DLF, CNI, and Open Repositories conferences.
15. If successful in finding community adoption, align *DRAS-TIC Fedora* with the Fedora project, in terms of sustainability and governance. Look to the broader Fedora community for software maintenance and future development.

3. National Impact

The proposed project will produce open source software, tested cluster configurations, documentation, and best-practice guides that will enable institutions to manage Fedora repositories with Petabyte scale collections and thus contribute to big-data ready national software infrastructure. Jansen and Marciano are participating in the IMLS-funded *"Always Already Computational: Library Collections as Data"* project (Principal Investigator: Thomas Padilla), and we will share research outcomes with the broad repository community at DLF, CNI, Open Repositories, and other upcoming venues.

Impact on Fedora Institutions

We expect that Fedora institutions and our community partners will make gains on several fronts through the *DRAS-TIC Fedora* project. Perhaps the most important gain will be the ability to smoothly increase the size of their repository collections without re-architecting storage. This smooth scaling of storage will lower costs and shorten planning cycles, allowing institutions to be more responsive to new and challenging collecting opportunities. Another impact can be better reliability, in terms of collection access and redundant datacenter storage. With Cassandra, *DRAS-TIC Fedora* can hold redundant copies of data in data centers around the world. Even if an entire data center is lost, access can remain uninterrupted, and data re-replicated to a new data center. Beyond institutional reliability, we think this opens up new possibilities for reciprocal storage arrangements between Fedora institutions.

Impact on Computational Archival Science

More broadly speaking we expect *DRAS-TIC Fedora*, by maintaining performance at scale and through distributed computation, to be a significant enabling contribution to the emerging field of Computational Archival Science (CAS).[13] Over the last year, we have convened key partners and led workshops to help shape emerging national conversations. In August 2017 we led an SAA Pop-up session on the theme of archival records in the age of big data.[14] Our working definition of CAS is:

> *Contributing to the development of the theoretical foundations of a new trans-discipline of computer and archival science*

Our current objectives are:
- Contributing to the development of the theoretical foundations of a new trans-discipline of computer and archival science
- Designing the educational foundations and delivering training in this emerging trans-discipline to support all industries and fields
- Developing a virtual and physical laboratory to test and apply scientific advances in a collaborative environment

We have recently authored:
- a Foundational Paper on **"Archives Records and Training in the Age of Big data"** to be published as a book chapter in 2017 in: *"Advances in Librarianship – Re-Envisioning the MLIS: Perspectives on the Future of Library and Information Science Education".*[15]
- a National Forum Position Paper on **"The Computational Turn in Archives & Libraries and the Notions of Levels of Computational Services,"** presented at the *"Always Already Computational: Library Collections as Data"* Forum in March 2017 at the UC Santa Barbara workshop.[16]

We will rally the CAS community as early as mid-December 2017 in Boston at our 2nd IEEE Big Data 2017 CAS Workshop[17], where the kinds of national infrastructure integration and interoperability questions explored in this proposal will be at the forefront. We will also propose a session on this theme at the 2018 Annual Meeting of the Society of American Archivists in Washington D.C.

---

13 Computational Archival Science (CAS): http://dcicblog.umd.edu/cas/
14 SAA 2016 pop-up session on CAS: https://archives2016.sched.com/event/7f9D
15 Foundational Paper on "Archives Records and Training in the Age of Big data": http://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2016/05/submission_final_draft.pdf
16 National Forum Position Paper on "The Computational Turn in Archives & Libraries..." presented at the *"Always Already Computational: Library Collections as Data"* Forum in March 2017 at the UC Santa Barbara workshop: http://dcicblog.umd.edu/cas/wp-content/uploads/sites/13/2016/05/AlwaysAlreadyComputationalPositionStatement.pdf
17 2nd IEEE Big Data 2017 CAS Workshop: http://cci.drexel.edu/bigdata/bigdata2017/index.html