## Abstract

# Improving Fedora to Work with Web-scale Storage and Services

Memory institutions around the world face a rapidly expanding need for storage and access to digital objects and metadata. The Fedora Repository has long been at the forefront of their efforts, developing software to meet the challenge, including four major versions of the Fedora Repository software. Now the Fedora leadership have put forward a bold call to the community to create new implementations of Fedora to meet emerging needs, publishing a formal API that specifies the expectations of a Fedora repository. Through our research into computational archives and through prior Fedora involvements, we have learned that a new core need is scalability, by which we mean the ability to expand storage capacity without losing performance. We believe that institutions must be able to incrementally grow a fully-functional repository as collections grow, without the need for expensive enterprise storage plans, massive data migrations, and performance limits that stem from the vertical storage strategy of previous repository implementations.

The Digital Curation Innovation Center (DCIC) at the University of Maryland's College of Information Studies (Maryland's iSchool) intends to conduct a 2-year project to research, develop, and test software architectures to improve the performance and scalability of the Fedora Repository for the Fedora community. More specifically this project will apply the new Fedora 5 application programming interface (API) to the repository software stack called DRAS-TIC to create a new Fedora implementation we are calling *DRAS-TIC Fedora*. DRAS-TIC, which stands for Digital Repository at Scale that Invites Computation, was developed over the last two years through a collaboration between UK-based storage company, Archive Analytics, and the DCIC, through funding from an NSF DIBBs (Data Infrastructure Building Blocks) grant (NCSA "Brown Dog"). DRAS-TIC leverages NoSQL industry standard distributed database technology, in the form of Apache Cassandra, to provide near limitless scaling of storage without performance degradation. With Cassandra, *DRAS-TIC Fedora* can also hold redundant copies of data in datacenters around the world. Even if an entire datacenter is lost, access can remain uninterrupted, and data re-replicated to a new datacenter. Beyond institutional reliability, we think this creates the possibility for new reciprocal storage arrangements between Fedora institutions.

To meet with this potential, DRAS-TIC will first need to be adapted to the new Fedora API and then engineered and tested to meet the performance expectations of our Fedora community partners. We have identified a range of institutional partners in the Fedora community that will work with us to develop use cases and performance expectations. As we develop and test *DRAS-TIC Fedora*, their institutional needs will guide our efforts and become our measure of success. The proposal has received the endorsement of the Fedora Leadership Group: http://fedorarepository.org/leadership-group.

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.